Nonparametric inference of interaction laws in systems of agents from trajectory data

Fei Lu[a,b,c], Ming Zhong[b], Sui Tang[a], and Mauro Maggioni[a,b,c,d,1]

[a]Department of Mathematics, Johns Hopkins University, Baltimore, MD 21218; [b]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218; [c]Institute for Data Intensive Engineering and Science, Johns Hopkins University, Baltimore, MD 21218; and [d]Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, MD 21218

Edited by Bin Yu, University of California, Berkeley, CA, and approved June 3, 2019 (received for review December 26, 2018)

Inferring the laws of interaction in agent-based systems from observational data is a fundamental challenge in a wide variety of disciplines. We propose a nonparametric statistical learning approach for distance-based interactions, with no reference or assumption on their analytical form, given data consisting of sampled trajectories of interacting agents. We demonstrate the effectiveness of our estimators both by providing theoretical guarantees that avoid the curse of dimensionality and by testing them on a variety of prototypical systems used in various disciplines. These systems include homogeneous and heterogeneous agent systems, ranging from particle systems in fundamental physics to agent-based systems that model opinion dynamics under the social influence, prey–predator dynamics, flocking and swarming, and phototaxis in cell dynamics.

data-driven modeling | dynamical systems | agent-based systems

1. Introduction

Systems of interacting agents arise in a wide variety of disciplines, including Physics, Biology, Ecology, Neurobiology, Social Sciences, and Economics (e.g., refs. 1–4 and references therein). Agents may represent particles, atoms, cells, animals, neurons, people, rational agents, opinions, etc. The understanding of agent interactions at the appropriate scale in these systems is as fundamental a problem as the understanding of interaction laws of particles in Physics.

How can laws of interaction between agents be discovered? In Physics, vast knowledge and intuition exist to formulate hypotheses about the form of interactions, inspiring careful experiments and accurate measurements, that together lead to the inference of interaction laws. This is a classical area of research, dating back to at least Gauss, Lagrange, and Laplace (5), that plays a fundamental role in many disciplines. In the context of interacting agents at the scale of complex organisms, there are fewer controlled experiments possible and few "canonical" choices for modeling the interactions. Different types and models of interactions have been proposed in different scientific fields and fit to experimental data, which in turn may suggest new modeling approaches, in a model–data validation loop. Often, the form of governing interaction laws is chosen a priori, within perhaps a small parametric family, and the aim is often to reproduce only qualitatively, and not quantitatively, some of the macroscopic features of the observed dynamics, such as the formation of certain patterns.

Our work fits at the boundary between statistical/machine learning and dynamical systems, where equations are estimated from observed trajectory data, and inference takes into account assumptions about the form of the equations governing the dynamics. Since the past decade, the rapidly increasing acquisition of data, due to decreasing costs of sensors and measurements, has made the learning of large and complex systems possible, and there has been an increasing interest in inference techniques that are model-agnostic and scalable to high-dimensional systems and large datasets.

We establish statistically sound, dynamically accurate, computationally efficient techniques* for inferring these interaction laws from trajectory data. We propose a nonparametric approach for learning interaction laws in particle and agent systems, based on observations of trajectories of the states (e.g., position, opinion, etc.) of the systems, on the assumption that the interaction kernel depends on pairwise distances only, unlike recent efforts that either require feature libraries or parametric forms for such interactions (6–10), or aim at identifying only the type of interaction from a small set of possible types (11–13). We consider a least-squares (LS) estimator, classical in the area of inverse problems (dating back to Legendre and Gauss), suitably regularized and tuned to the learning of the interaction kernel in agent-based systems.

The unknown is the interaction kernel, a function of pairwise distances between agents of the systems. While the values of this function are not observed, in contrast to the standard regression problems, we are able to show that our estimator converges at an optimal rate as if we were in the 1D regression setting. In particular, the learning rate has no dependency on the dimension of the state space of the system, therefore avoiding any curse of dimensionality, and making these estimators well-suited for the modern high-dimensional data regime. It may be easily extended to a variety of complex systems; here, we consider first- and second-order models, with single and multiple types of agents, and with interactions with simple environments. We demonstrate with examples that the theoretical guarantees on the performance of the estimator make it suitable for testing hypotheses on underlying models of interactions,

Significance

Particle and agent-based systems are ubiquitous in science. The complexity of emergent patterns and the high dimensionality of the state space of such systems are obstacles to the creation of data-driven methods for inferring the driving laws from observational data. We introduce a nonparametric estimator for learning interaction kernels from trajectory data, scalable to large datasets, statistically optimal, avoiding the curse of dimensionality, and applicable to a wide variety of systems from Physics, Biology, Ecology, and Social Sciences.

Author contributions: F.L. and M.M. designed research; F.L., M.Z., S.T., and M.M. performed research; F.L., M.Z., S.T., and M.M. analyzed data; and F.L., M.Z., S.T., and M.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The software package implementing the proposed algorithms can be found on https://github.com/MingZhongCodes/LearningDynamics.

[1] To whom correspondence may be addressed. Email: mauromaggionijhu@icloud.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1822012116/-/DCSupplemental.

*The software package implementing the proposed algorithms can be found on https://github.com/MingZhongCodes/LearningDynamics.

assisting an investigator in choosing among different possible (nonparametric) models.

Finally, our estimator is constructed with algorithms that are computationally efficient (with complexity $O(LN^2M)$ when the interaction kernel is Lipschitz; *SI Appendix*, section 2F) and may be implemented in a streaming fashion: It is, therefore, well-suited for large datasets.

2. Learning Interaction Kernels

We start with a model that is used in a wide variety of interacting agent systems [e.g., physical particles or influence propagation in a population (14, 15)]: Consider $N > 1$ agents $\{\boldsymbol{x}_i\}_{i=1}^N$ in $\mathbb{R}^d$, evolving according to the system of ordinary differential equations (ODEs)

$$\dot{\boldsymbol{x}}_i(t) = \frac{1}{N}\sum_{i'=1}^{N}\phi(\|\boldsymbol{x}_{i'}(t)-\boldsymbol{x}_i(t)\|)(\boldsymbol{x}_{i'}(t)-\boldsymbol{x}_i(t)), \qquad \textbf{[1]}$$

where $\dot{\boldsymbol{x}}_i(t) = \frac{d}{dt}\boldsymbol{x}_i(t)$; $\|\cdot\|$ is the Euclidean norm, and $\phi : \mathbb{R}_+ \to \mathbb{R}$ is the interaction kernel. In other words, every agent's velocity is obtained by superimposing the interactions with all of the other agents, each weighted in a way dependent on the distance to the interacting agent. In a prototypical example—e.g., arising in particle systems (Section 2B) and flocking systems—the interaction kernel may be negative for small distances, inducing repulsion, and attractive for large distances. Let $\boldsymbol{X} := (\boldsymbol{x}_i)_{i=1}^N \in \mathbb{R}^{dN}$ be the state vector for all of the agents, $\boldsymbol{r}_{ii'}(t) := \boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)$ and $r_{ii'}(t) := \|\boldsymbol{r}_{ii'}(t)\|$. The evolution Eq. **1** is the gradient flow for the potential energy $\mathcal{U}(\boldsymbol{X}(t)) := \frac{1}{2N}\sum_{i\neq i'}\Phi(r_{ii'}(t))$, with $\phi(\cdot) = \Phi'(\cdot)/\cdot$. The function $\phi(\cdot)\cdot$ reappears naturally below, the fundamental reason being its relationship with $\mathcal{U}$ and Φ. Our observations are positions along trajectories: $\boldsymbol{X}_{\mathrm{tr}} := \{\boldsymbol{X}^m(t_l)\}_{l=1,m=1}^{L,M}$, with $0 = t_1 < \ldots < t_L = T$ being the times at which observations occur, and m indexing M different trajectories. Velocities $\dot{\boldsymbol{X}}^m(t_l)$ are approximated by finite differences. The M initial conditions (ICs) $\boldsymbol{X}_0^m := \boldsymbol{X}^m(0)$ are drawn independently at random from a probability measure μ_0 on $\mathbb{R}^{dN}$.

Our goal is to infer, in a nonparametric fashion, the interaction kernel ϕ, by constructing an estimator $\hat{\phi}$ from training data. A fundamental statistical problem that involves estimating a function is regression: Given samples $(z_i, g(z_i))_{i=1}^n$, with the z_i's independent and identically distributed (i.i.d.) samples from an (unknown) measure ρ_Z in $\mathbb{R}^D$, and g a suitably regular (say, Hölder s) unknown function $\mathbb{R}^D \to \mathbb{R}$, one constructs an estimator $\hat{g}$ such that $\|\hat{g} - g\|_{L^2(\rho_Z)} \lesssim n^{-\frac{s}{2s+D}}$, with high probability (over the z_i's). This rate is optimal in a minimax sense (16), and its dramatic degradation with D is a manifestation of the curse of dimensionality. Upon rewriting Eq. **1** as $\dot{\boldsymbol{X}} = \mathbf{f}_\phi(\boldsymbol{X})$, our observations (with either approximated or directly observed velocities) resemble those needed for regression if we thought of $Z = \boldsymbol{X}$ as a random variable, and $g = \mathbf{f}_\phi$. However, our observations are not i.i.d. samples of $\boldsymbol{X}$ with respect to any probability measure, the lack of independence being the most glaring aspect. If we nevertheless pursued this line of thought, we would be hit with the curse of dimensionality in trying to learn the target function $g = \mathbf{f}_\phi$ on the state space $\mathbb{R}^{dN}$, leading to a rate $n^{-O(1/dN)}$ for regression. This renders this approach useless in practice as soon as, say, $dN \geq 20$. A direct application of existing approaches (e.g., refs. 6–8), developed for low-dimensional systems, go in this direction, These works would try to ameliorate this curse of dimensionality by requiring $\mathbf{f}_\phi$ to be well-approximated by a linear combination of a small number of functions in a known large dictionary. While such dictionaries may be known for specific problems, they are usually not given in the case of complex, agent-based systems. Finally, such dictionaries typically grow dramatically in size with the dimension (here, dN), and existing guarantees that avoid the curse of dimensionality require further, strong assumptions on the measurements or the dynamics.

We proceed in a different direction, aiming for the flexibility of a nonparametric model while exploiting the structure of the system in Eq. **1**. The target function ϕ depends on just one variable (pairwise distance), but it is observed through a collection of nonindependent linear measurements (the left-hand side of Eq. **1**), at locations $r_{ii'}^m(t_l) = \|\boldsymbol{x}_{i'}^m(t_l) - \boldsymbol{x}_i^m(t_l)\|$, with coefficients $\boldsymbol{r}_{ii'}^m(t_l) = \boldsymbol{x}_{i'}^m(t_l) - \boldsymbol{x}_i^m(t_l)$, as in the right-hand side of Eq. **1**. When the t_l's are equidistant in time, we consider an estimator minimizing the empirical error functional

$$\mathcal{E}_{L,M}(\varphi) := \frac{1}{LMN}\sum_{l,m,i=1}^{L,M,N}\left\|\dot{\boldsymbol{x}}_i^m(t_l) - \mathbf{f}_\varphi(\boldsymbol{x}^m(t_l))_i\right\|^2, \qquad \textbf{[2]}$$

$$\widehat{\phi} = \widehat{\phi}_{L,M,\mathcal{H}} := \underset{\varphi\in\mathcal{H}}{\arg\min}\,\mathcal{E}_{L,M}(\varphi), \qquad \textbf{[3]}$$

where $\mathcal{H}$ is a hypothesis space of functions $\mathbb{R}_+ \to \mathbb{R}$, of dimension n (we will choose n dependent on M). We introduce a natural probability measure ρ_T on $\mathbb{R}_+$ adapted to the dynamics: It can be thought of as an "occupancy" measure, in the sense that for any interval I, $\rho_T(I)$ is the probability (over the random ICs distributed according to μ_0) of seeing a pair of agents with a distance between them being a value in I, averaged over the time interval $[0, T]$; see Eq. **4** for a formal definition.

We measure the performance of $\hat{\phi}$ in terms of the error $\|\hat{\phi}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T)}$. Theorem (Thm.) 3.3, our main result, will bound this error by $\tilde{O}(M^{-s/(2s+1)})$ if ϕ is Hölder s: This is the optimal exponent for learning ϕ if we were in the (more favorable) 1D regression setting! We therefore completely avoid the curse of dimensionality. In fact, we show under some rather general assumptions that not only the rate, but even the constants in the bound are independent of N, making the bounds essentially dimension-free. It is crucial that ρ_T has wide support in order for the error to be informative. When the system is ergodic, we expect ρ_T to have a large support for large T, as the system explores its ergodic distribution. However, many deterministic systems of interest may reach a stationary state (as in the cases of the Lennard–Jones or opinion dynamics, to be considered momentarily), in which case ρ_T becomes highly concentrated on a finite set for large T: In these cases, it may be more relevant to consider T small compared with the relaxation time.

We are also interested in whether trajectories $\boldsymbol{X}(t)$ of the true system are well-approximated by trajectories $\widehat{\boldsymbol{X}}(t)$ of the system governed by the interaction kernel $\hat{\phi}$, on both the "training" time interval $[0, T]$ and after time T. Proposition (Prop.) 3.4 below bounds $\sup_{t\in[0,T']}\|\widehat{\boldsymbol{X}}(t) - \boldsymbol{X}(t)\|$ in terms of $\|\hat{\phi}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T)}$, at least for T' not too large; this further validates the use of $L^2(\rho_T)$. We will report on this distance for both $T' = T$ and $T' > T$ ("prediction" regime).

Finally, while the error $\|\hat{\phi}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T)}$ is unknown in practice (since ϕ is unknown), our results give guarantees on its size, which in turn imply guarantees on accuracy of trajectory predictions. Proxies for the error on trajectories, for example, by holding out portions of trajectories during the training phase, may be derived from data. These measures of error may be used to test and validate different models of the dynamics: Too large an error with one model may invalidate it and suggest that a different one (e.g., second vs. first order or multiple vs. single agent types) should be used (Section 5).

A. Different Sampling Regimes and Randomness. The total number of observations is (number of ICs)× (number of temporal observations in $[0, T]) = M \times L$, each in $\mathbb{R}^{dN}$. We will consider several regimes:

Many short time trajectories. T is small, L is small (e.g., $L = 1$), and M is large (many ICs sampled from μ_0);

Single large time trajectory. T is large (even comparable to the relaxation time of the system if applicable), L is large, and $M = 1$ (or very small);

Intermediate time scale. T, L and M are all not small, but none is very large, corresponding to multiple "medium"-length trajectories, with several different ICs.

Randomness is injected via the ICs, and in our main results in Section 3, the sample size will be M. If the system is ergodic, the regimes above are partially related to each other, at least when the ICs are sampled from the ergodic distribution μ_{erg}. Indeed, at times much larger than the mixing time T_{mix}, the state of the system becomes indistinguishable from a random sample of μ_{erg}, and we may interpret the subsequent part of the trajectory as a new trajectory with that IC. The M observed trajectories of length $T \gg T_{\text{mix}}$ are then equivalent to $M \times T/T_{\text{mix}}$ trajectories of length T_{mix}, to which our results apply. In regimes when M is very small or μ_0 is very concentrated, there is little randomness: The problem is close to a fixed-design inverse problem, which is solvable if the dynamics produces different-enough pairwise distances.

B. Example: Interacting Particles with the Lennard–Jones Potential. We illustrate the learning procedure on a particle system with $N = 7$ particles in $\mathbb{R}^2$, interacting according to Eq. **1** with $\phi(r) = \Phi'_{LJ}(r)/r$, where $\Phi_{LJ}(r) := 4\epsilon\left((\sigma/r)^{12} - (\sigma/r)^6\right)$ is the Lennard–Jones potential, consisting of a strong near-field repulsion and a long-range attraction. The system converges quickly to equilibrium configurations, which often consist of ordered, crystal-like structures. This example is challenging for various reasons: the Interaction kernel is unbounded, has unbounded support, and equilibrium is reached quickly, reducing the amount of information in trajectories. *SI Appendix*, section 3B contains a detailed description of the experiments. Fig. 1 demonstrates that the estimators approximate the true kernel well in different sampling regimes and that the trajectories of the true system are well-approximated by those of the learned system both in the "training" interval ($[t_0, T]$) and in the "prediction" interval ($[T, 50T]$ and $[T, 2T]$ respectively for the two regimes). We also show, as a simple example of transfer learning, that we can use the interaction kernel learned on the system with N particles to accurately predict trajectories of a system with $4N$ particles.

The rate of decay of the estimation error is close to the optimal rate in Thm. 3.3 (Fig. 2); this is a consequence of two factors: the use of an empirical approximation to ρ_T^L and

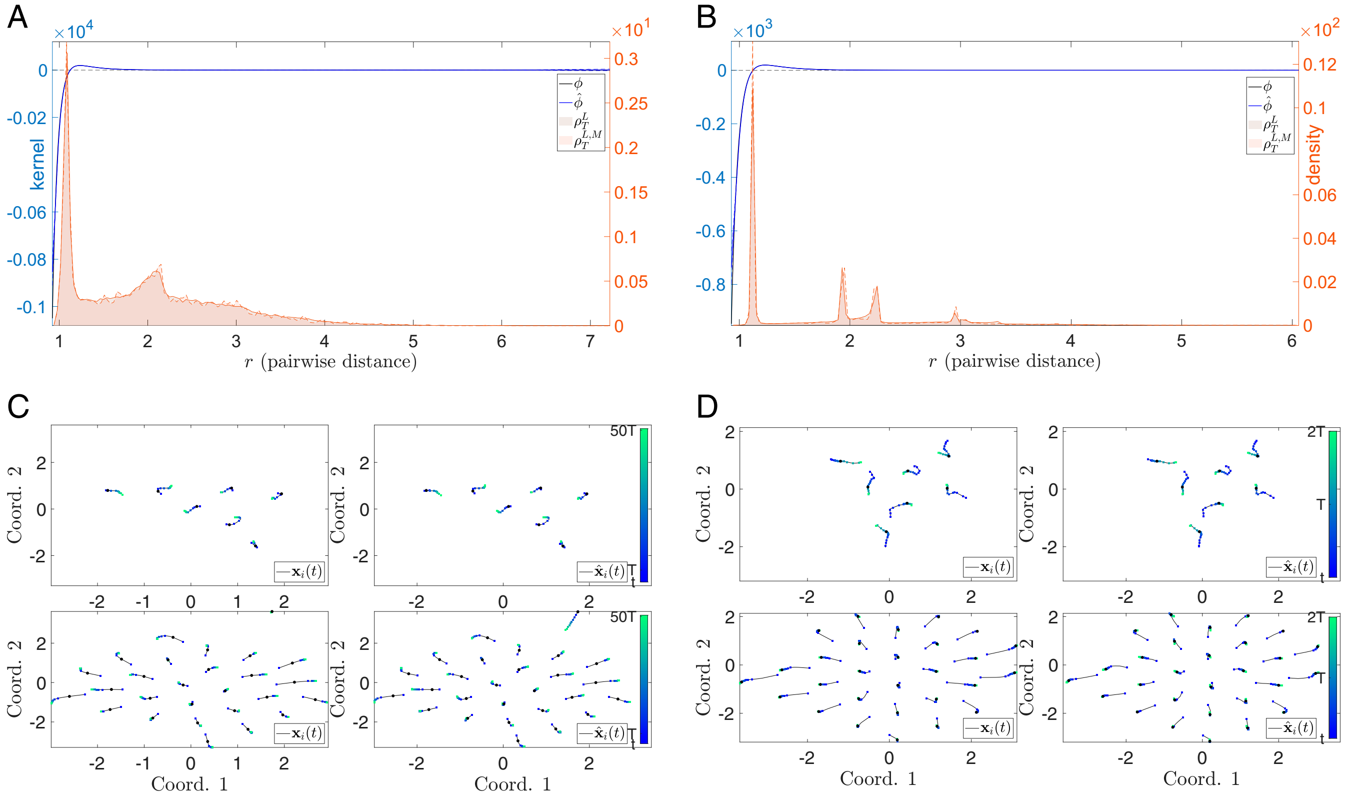


Fig. 1. Interaction kernel estimation and trajectory prediction for the Lennard–Jones system. (*A* and *B*) Estimators $\hat{\phi}$ (in blue) of the true interaction kernel ϕ (in black) in two sampling regimes: many short-time trajectories (*A*) and a few large-time trajectories (*B*). The proposed nonparametric estimators perform extremely well—the means and SDs of the relative $L^2(\rho_T^L)$ errors are $6.6 \cdot 10^{-2} \pm 5.0 \cdot 10^{-3}$ and $7.2 \cdot 10^{-2} \pm 1.0 \cdot 10^{-2}$, respectively, over 10 independent learning runs. The SD (dashed) lines on the estimated kernel are so small to be barely visible. In both cases, we superimpose histograms of ρ_T^L (estimated from a large number of trajectories, outside of training data) and $\rho_T^{L,M}$ (estimated from the M training data trajectories; *SI Appendix*, Eq. 5). The estimators belong to a hypothesis space $\mathcal{H}_n$ of piecewise linear functions with equidistant knots and yield accurate estimators in $L^2(\rho_T^L)$. Note that we observe the dynamics starting from a suitable $t_0 > 0$, due to the singularity of Lennard–Jones kernel at $r = 0$. See *SI Appendix*, section 3B for details about the setup and results. (*C* and *D*) The true and predicted trajectories for the N-particle system (*Upper*) and a $4N$-particle system (*Lower*) with interaction kernels learned on the N-particle system, for randomly sampled ICs. *C* and *D* show true and predicted trajectories for systems with interaction kernels learned in *A* and *B*, respectively. The blue-to-green color gradient indicates the movement of particles in time (see color scales on the side). We achieve small errors in predicting the trajectories in all cases, even when we transfer the interaction kernel learned on an N-particle system to predict trajectories of a system with $4N$ particles. Coord., coordinates.

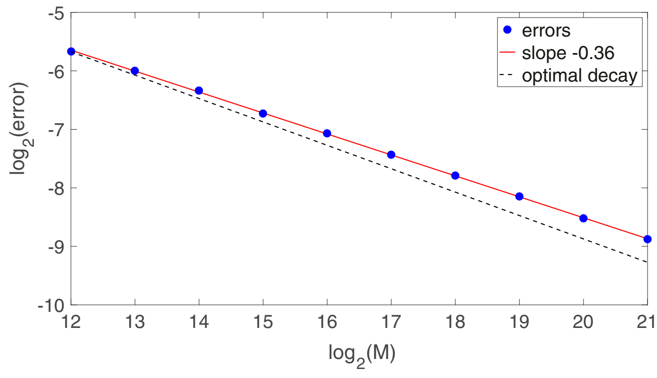


Fig. 2. Learning rate in M for the Lennard–Jones system. The estimation error in $L^2(\rho_T^L)$ decays at rate 0.36, close to the optimal rate 0.4 for admissible kernels; Thm. 3.3.

the blowup at 0 of Φ_{LJ}, which is not an admissible kernel as in Thm. 3.3 (see *SI Appendix*, Section 3B for a detailed discussion).

Fig. 3 shows the behavior of the error of the estimators as both L and M are increased. It indicates that a single long trajectory may not contain enough "information" to learn the kernel, at least for deterministic systems approaching a steady state. It also shows the behavior predicted by Thm. 3.3—namely, for each fixed L the error decreases as M increases.

3. Learning Theory

We introduce an error functional based on the structure of the dynamical system $\dot{\boldsymbol{X}} = \mathbf{f}_\phi(\boldsymbol{X})$, whose minimizer will be our estimator of the interaction kernel ϕ. We consider kernels in the admissible set $\mathcal{K}_{R,S} := \{\phi \in C^1(\mathbb{R}_+) : \ \mathrm{supp}(\phi) \subset [0,R], \sup_{r\in[0,R]} |\phi(r)| + |\phi'(r)| \leq S\}$, for some $R, S > 0$. The boundedness of ϕ and ϕ' ensures the global well-posedness of the system in Eq. **1**. The restriction $\mathrm{supp}(\phi) \subset [0,R]$ models the finite range of interaction between agents, and it may be relaxed to $\phi \in W^{1,\infty}(\mathbb{R}_+)$ with a suitable decay.

A. Probability Measures Adapted to the Dynamics. To measure the quality of the estimator of the interaction kernel ϕ, we introduce two probability measures on $\mathbb{R}_+$, the space of pairwise distances $r_{ii'}^m(t_l) = \|\boldsymbol{x}_{i'}^m(t_l) - \boldsymbol{x}_i^m(t_l)\|$. We consider the expectation of the empirical measure of pairwise distances, for continuous and discrete time observations, respectively:

$$\rho_T(r) := \frac{1}{\binom{N}{2} T} \int_{t=0}^{T} \mathbb{E}_{\boldsymbol{X}_0 \sim \mu_0} \left[\sum_{i,i'=1, i<i'}^{N} \delta_{r_{ii'}(t)}(r)\, dt \right], \quad \textbf{[4]}$$

$$\rho_T^L(r) := \frac{1}{\binom{N}{2} L} \sum_{l=1}^{L} \mathbb{E}_{\boldsymbol{X}_0 \sim \mu_0} \left[\sum_{i,i'=1, i<i'}^{N} \delta_{r_{ii'}(t_l)}(r) \right]. \quad \textbf{[5]}$$

The expectations are over the ICs, with distribution μ_0. The measure ρ_T is intrinsic to the dynamical system, dependent on μ_0 and the time scale T, and independent of the observation data. ρ_T^L depends also on the sampling scheme $\{t_l\}_{l=1}^L$ in time. Both are Borel probability measures on $\mathbb{R}_+$ (*SI Appendix*, Lemma 1.1), measuring how much regions of $\mathbb{R}_+$ on average (over the observed times and ICs) are explored by the system. Highly explored regions are where the learning process ought to be more accurate, as they are populated by more "samples" of pairwise distances. We will measure the estimation error of our estimators in $L^2(\rho_T)$ or $L^2(\rho_T^L)$.

We report here on the analysis in the discrete-time observation case, most relevant in practice, with ρ_T^L; the arguments, however, also apply to continuous-time observations, with ρ_T.

B. Learnability: The Coercivity Condition. A fundamental question is the learnability of the kernel, i.e., the convergence of the estimator $\hat{\phi}_{L,M,\mathcal{H}}$ defined in Eq. **3** to the true kernel ϕ as the sample size increases (i.e., $M \to \infty$) and $\mathcal{H}$ increases in a suitable way. The following condition, similar to the one introduced in ref. 17 for studying the mean field limit ($N \to \infty$), ensures learnability and well-posedness of the estimation.

Definition 3.1. The dynamical system in Eq. **1**, with IC sampled from μ_0 on $\mathbb{R}^{dN}$, satisfies the **coercivity condition** on a set $\mathcal{H}$ if there exists a constant $c_{L,N,\mathcal{H}} > 0$ such that for all $\varphi \in \mathcal{H}$ with $\varphi(\cdot)\cdot \in L^2(\rho_T^L)$,

$$c_{L,N,\mathcal{H}} \|\varphi(\cdot)\cdot\|^2_{L^2(\rho_T^L)} \leq \frac{1}{NL} \sum_{l,i=1}^{L,N} \mathbb{E} \left\| \frac{1}{N} \sum_{i'=1}^{N} \varphi(r_{ii'}(t_l)) \boldsymbol{r}_{ii'}(t_l) \right\|^2 . \quad \textbf{[6]}$$

The coercivity condition ensures learnability, by implying the uniqueness of minimizer of $\mathcal{E}_{L,\infty}(\varphi) := \mathbb{E}[\mathcal{E}_{L,M}(\varphi)]$ and, eventually, the convergence of estimators through a control of the error of the estimator in $L^2(\rho_T^L)$ (*SI Appendix*, Thm. 1.2 and Prop. 1.3). Thm. 3.1 proves that the coercivity condition holds under suitable hypotheses, even independently of N; numerical tests suggest that it holds generically over larger classes of interaction kernels and distributions of ICs, for large L, and as long as ρ_T^L is not degenerate (*SI Appendix*, Fig. S6). Finally, $c_{L,N,\mathcal{H}}$ also controls the condition number of the matrix in the LS problem yielding the estimator (see *SI Appendix*, Prop. 2.1 for details).

We prove that coercivity holds when μ_0 is exchangeable (i.e., the distribution is invariant under permutation of components), Gaussian, and $L = 1$. Numerical tests (*SI Appendix*, Fig. S6) suggest that the coercivity condition holds true for a larger class of interaction kernels, for various initial distributions including Gaussian and uniform distributions, and for large L, as long as ρ_T^L is not degenerate. We conjecture that the coercivity condition holds true in much greater generality (but not always!), leaving a detailed investigation to future work.

Theorem 3.1. *Suppose $L = 1$, $N > 1$ and assume that the distribution of $\boldsymbol{X}(t_1) = (\boldsymbol{x}_1(t_1), \ldots, \boldsymbol{x}_N(t_1))$ is exchangeable Gaussian with $\mathrm{cov}(\boldsymbol{X}_i) - \mathrm{cov}(\boldsymbol{X}_i, \boldsymbol{x}_{i'}) = \lambda I_d$ for a constant $\lambda > 0$. Then, the*

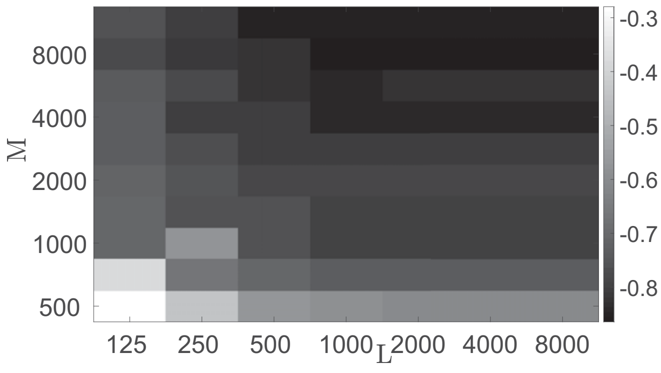


Fig. 3. The relative error of the estimated kernel as a function of M, L for the Lennard–Jones system. The relative error, in $\log_{10}$ scale, of $\hat{\phi}$ decreases both in L and M, in fact, roughly in the product ML, at least when M and L are not too small. $M = 1$ does not seem to suffice, no matter how large L is, due to the limited amount of "information" contained in a single trajectory.

coercivity condition holds true with $c_{L,N,\mathcal{H}} = \frac{(N-1)(N-2)}{N^2} c_{\mathcal{H}} + \frac{N-1}{N^2}$*, where* $c_{\mathcal{H}}$ *is independent of* N*, is positive for any compact* $\mathcal{H} \subset L^2(\rho_T^L)$*, and is zero for* $\mathcal{H} = L^2(\rho_T^L)$*.*

In this setting, the analysis of the coercivity constant $c_{L,N,\mathcal{H}}$ is based on the exchangeability of the initial distribution of the agents and relates coercivity to a positive integral kernel:

Lemma 3.2. *Let* X, Y, Z *be exchangeable Gaussian random vectors in* $\mathbb{R}^d$ *with* $\mathrm{cov}(X) - \mathrm{cov}(X, Y) = \lambda I_d$ *for a constant* $\lambda > 0$*. Suppose* $L = 1$*. Then, there is a positive definite integral kernel* $\mathcal{K}(r, s) : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ *such that for any* $g \in L^2(\rho_T^L)$

$$\mathbb{E}\left[g(|X-Y|)g(|X-Z|)\langle X-Y, X-Z\rangle\right] = \iint g(r)rg(s)s\mathcal{K}(r,s)\,drds,$$

where $\rho_T^L(r) \propto r^{d-1}e^{-r^2/3}$*, since* $L = 1$*. Therefore, there exists* $c_{\mathcal{H}} \geq 0$*, depending only on* $\mathcal{H} \subset L^2(\rho_T^L)$*, such that for* $g \in \mathcal{H}$

$$\iint g(r)rg(s)s\mathcal{K}(r,s)\,drds \geq c_{\mathcal{H}} \|g(\cdot)\cdot\|^2_{L^2(\rho_T^L)},$$

and $c_{\mathcal{H}} > 0$ *if* $\mathcal{H}$ *is compact in* $L^2(\rho_T^L)$*.*

We conclude that under the assumptions of Thm. 3.1, if $\mathcal{H}$ is compact, then $c_{L,N,\mathcal{H}}$ is bounded below uniformly in N.

C. Optimal Rates of Convergence. The classical bias–variance trade-off in statistical estimation guides the selection of a hypothesis space $\mathcal{H}$, whose dimension will depend on M, the number of observed trajectories. On the one hand, $\mathcal{H}$ should be large so that the bias (distance between the true kernel ϕ and $\mathcal{H}$) is small; on the other hand, $\mathcal{H}$ should be small so that variance of the estimator is small. In the extreme case where $\mathcal{H} = \mathcal{K}_{R,S}$, the bias is 0, the variance of the estimator dominates, and we obtain the bound $\mathbb{E}[\|\widehat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T^L)}] \leq CM^{-1/4}$ (*SI Appendix*, Prop. 1.5). In fact, significantly better rates may be achieved for regular ϕ's:

Theorem 3.3. *Assume that* $\phi \in \mathcal{K}_{R,S}$*. Let* $\{\mathcal{H}_n\}_n$ *be a sequence of subspaces of* $L^\infty([0,R])$*, with* $\dim(\mathcal{H}_n) \leq c_0 n$ *and* $\inf_{\varphi \in \mathcal{H}_n} \|\varphi - \phi\|_{L^\infty([0,R])} \leq c_1 n^{-s}$*, for some constants* $c_0, c_1, s > 0$*. Assume that the coercivity condition holds on* $\mathcal{H} := \overline{\cup_{n=1}^{\infty} \mathcal{H}_n}$*. Such a sequence exists, for example, if* ϕ *is* s*-Hölder regular, and can be chosen so that* $\mathcal{H}$ *is compact in* $L^2(\rho_T^L)$*. Choose* $n_* = (M/\log M)^{1/(2s+1)}$*. Then, there exists a constant* $C = C(c_0, c_1, R, S)$ *such that*

$$\mathbb{E}\left[\|\widehat{\phi}_{L,M,\mathcal{H}_{n_*}}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T^L)}\right] \leq \frac{C}{c_{L,N,\mathcal{H}}}\left(\frac{\log M}{M}\right)^{\frac{s}{2s+1}}. \quad [7]$$

The rate [i.e., the exponent $s/(2s+1)$] we achieve is *optimal*: It coincides with the minimax rate in the classical regression setting, where one can observe directly noisy values of an s-Hölder regression function at the sample points. We obtain this optimal rate, even if we do not observe the values $\{\phi(r_{ii'}^m(t_l))\}_{l,i,i',m}$, but a "mixture" of them in the observed trajectory data. Many choices of $\{\mathcal{H}_n\}$ are consistent with the requirements in the theorem, e.g., splines on increasingly finer grids, or band-limited functions with increasing frequency limits. These choices affect the constants in Eq. **7**, the computational complexity of computing $\widehat{\phi}_{L,M,\mathcal{H}_{n_*}}$, but not the rate in M. While the rate is independent of the dimension dN of the state space, the constant may depend on d and N via $c_{L,N,\mathcal{H}}$. However, we expect that under rather general conditions, beyond those in Thm. 3.1, $c_{L,N,\mathcal{H}}$ is, in fact, lower-bounded independently of N for any compact subset $\mathcal{H}$ of $L^2(\rho_T^L)$ and is a fundamental property of the mean field limit ($N \to \infty$) of the system.

One shortcoming of our result is that the rate is not a function of the total number of observations, which is $O(LN^2M)$ (we have $LN^2/2$ pairwise distances for each of the M trajectories), but only of M, the number of *random* samples. Numerical experiments (see Fig. 3 and similar experiments for the other systems, reported in *SI Appendix*) suggest that the estimator improves as L increases, at least to a point, limited by the "information" in a single trajectory. Comparing to ref. 17, where the mean field limit $N \to \infty$, $M = 1$, is studied, we see the rates in ref. 17 seem no better than $N^{-1/d}$, i.e., they are cursed by dimension. So are sparsity-based inference techniques such as those in refs. 6–8, 11, and 18, which also require a good dictionary of template functions, are not nonparametric (at least in the form therein presented), and lack performance guarantees, except in some cases under stringent assumptions.

Our work here may be compared with the classical parameter estimation problem for the ODE models (19–22), where one is interested in estimating the vector parameter $\boldsymbol{\theta}$ in the ODE model $\dot{\boldsymbol{X}} = \boldsymbol{f}(\boldsymbol{X}(t), t, \boldsymbol{\theta})$ from the observation of a single noisy trajectory. Our error functional, in spirit, is the same with the gradient-matching method (also called the two-stage method) used in the parameter-estimation problems (23–27). A challenging problem is the identifiability of $\boldsymbol{\theta}$. We refer the reader (28) for the statistical analysis and (29) (and references therein) for a comprehensive survey of this topic. However, the problem and approach we considered here are different from the

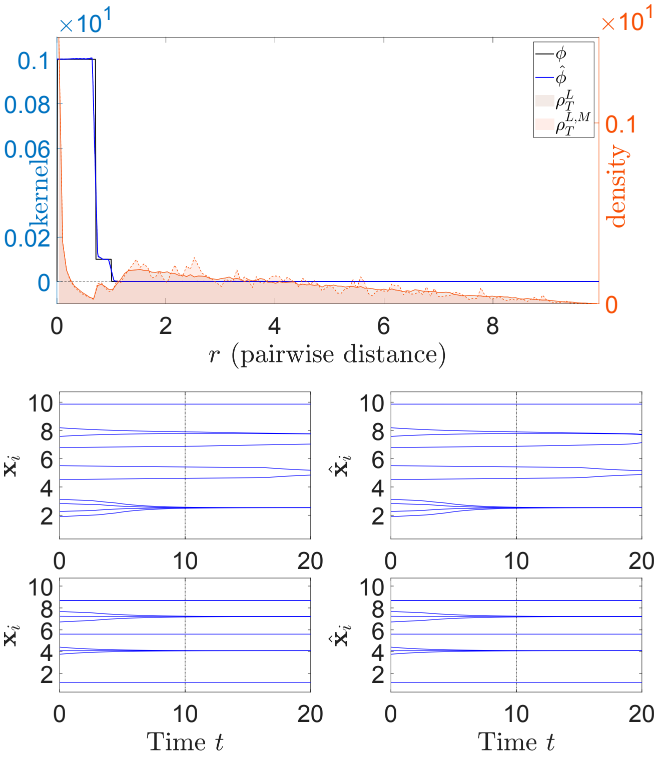


Fig. 4. Opinion dynamics. (*Upper*) Comparison between true and estimated interaction kernel, together with histograms for ρ_T^L and $\rho_T^{L,M}$. The mean and SD of the relative error for the interaction kernel are $1.6 \cdot 10^{-1} \pm 2.3 \cdot 10^{-3}$ over 10 independent learning runs. The SD lines (in dashed lines) on the estimated kernel are so small to be barely visible. (*Lower*) Trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ obtained with ϕ and $\hat{\phi}$, respectively, for an IC in the training data (top row) and an IC randomly chosen (bottom row). The black dashed vertical line at $t = T$ divides the "training" interval $[0, T]$ from the "prediction" interval $[T, T_f]$ (which in this case, $T_f = 2T$). We achieve small errors in all cases, in particular predicting number and location of clusters for large time.

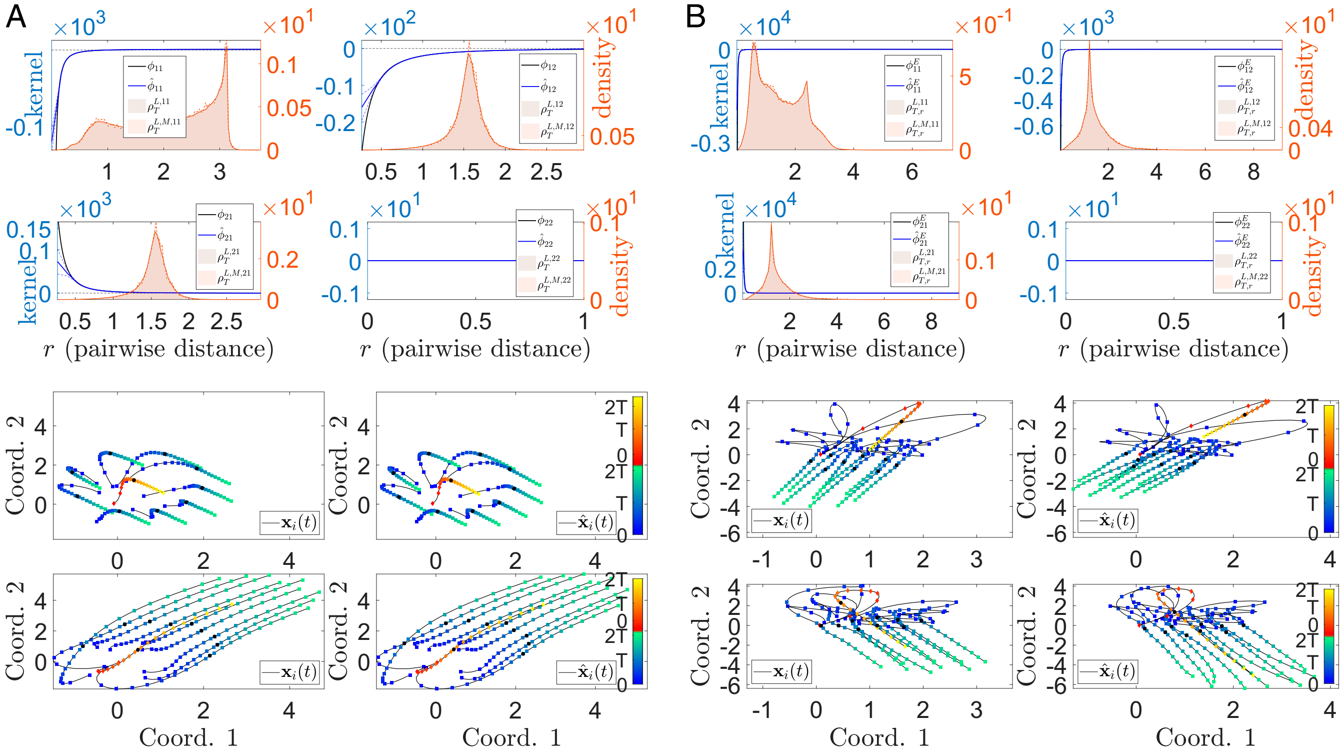


Fig. 5. Estimation of interaction kernels and trajectory prediction for predator–swarm first- and second-order systems. Results for the first-order (*A*) and second-order (*B*) predator–swarm systems, as described in Sections 4 and 5, are shown. For each system (corresponding to each column), *Upper* represents $\phi_{k,k'}$ and $\widehat{\phi}_{k,k'}$, superimposed with the histograms of ρ_T^L (estimated from a large number of trajectories, outside of training data) and $\rho_T^{L,M}$ (estimated from the *M* training data trajectories; *SI Appendix*, Eq. 5). *Lower* shows trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ of the corresponding (original and estimated) systems, evolved from the same ICs as the training data (third row) and newly sampled ICs (fourth row), over both the training time interval $[0, T]$ and in the future $[T, 2T]$ (color bars; the black dots in the trajectories correspond to $t = T$). For trajectories generated by the predator–swarm system, red-to-yellow lines indicate the movement of predators, whereas the blue-to-green lines indicate the movement of prey. The color gradients indicate time; see the color scales on the side of the plots. The estimators $\hat\phi_{k,k'}$ perform extremely well: with negligible differences in the regions with large ρ_T^L and with possibly larger errors in regions with small ρ_T^L (where the SDs over 10 independent learning runs become visible). The $L^2(\rho_T^L)$ errors of the estimators are reported numerically in *SI Appendix*, section 3. Note that they are truncated to a constant while preserving continuity, when there are no samples (e.g., *r* near 0 or *r* very large). The measure ρ_T^L is quite smooth but can have interesting features; $\rho_T^{L,M}$ is typically a noisy version of ρ_T^L. The trajectories of the estimated system are typically good approximations to those of the original system, on both ICs in the training data and newly sampled ICs. The error of the estimated trajectories increases with time, as expected, albeit it still typically excellent also in the "prediction" time interval $[T, 2T]$, showing that the bounds in Prop. 3.4, while sharp in general, may be overly pessimistic in some practical cases. Some slightly larger errors are present in some trajectories, e.g., when prey and predators get much closer to each other than they did in the training data. Coord., coordinates.

parameter-estimation problem in several aspects. First of all, our state variable $\boldsymbol{X}$ enters into the domain of the ϕ (via its "projection" onto pairwise distance), while the parameter vector $\boldsymbol{\theta}$ is decoupled from the state variable $\boldsymbol{X}$. Moreover, our estimator is nonparametric—i.e., the goal is to estimate a function ϕ (a vector infinite dimensions) instead of a finite-dimensional vector $\boldsymbol{\theta}$ of parameters. Finally, we establish identifiability conditions for ϕ from the perspective that the observations are i.i.d. trajectories with random ICs, in contrast with the identifiability of $\boldsymbol{\theta}$ from observations along a fixed single trajectory with i.i.d. noise. We would like to mention the different, but related, problem of inferring potentials from ground states and unstable modes (for example, ref. 30), as well as recent results on existence and properties of ground states for systems with nonlocal interactions (31).

D. Trajectory-Based Performance Measures. It is important not only that $\widehat{\phi}$ is close to ϕ, but also that the dynamics of the system governed by $\widehat{\phi}$ approximate well the original dynamics. The error in prediction may be bounded trajectory-wise by a continuous-time version of the error functional and bounded in average by the $L^2(\rho_T)$ error of the estimated kernel (further evidence of the usefulness of ρ_T):

Proposition 3.4. *Assume* $\widehat{\phi}(\|\cdot\|)\cdot \in \mathrm{Lip}(\mathbb{R}^d)$, *with Lipschitz constant* C_{Lip}. *Let* $\widehat{\boldsymbol{X}}(t)$ *and* $\boldsymbol{X}(t)$ *be the solutions of systems with kernels* $\widehat{\phi}$ *and* ϕ, *respectively, started from the same IC. Then, for each trajectory*

$$\sup_{t\in[0,T]} \|\widehat{\boldsymbol{X}}(t)-\boldsymbol{X}(t)\|^2 \le 2Te^{8T^2C_{\mathrm{Lip}}^2}\int_0^T \left\|\dot{\boldsymbol{X}}(t)-\mathbf{f}_{\hat{\phi}}(\boldsymbol{X}(t))\right\|^2 dt,$$

and on average with respect to the distribution μ_0 *of ICs*:

$$\mathbb{E}_{\mu_0}\left[\sup_{t\in[0,T]}\|\widehat{\boldsymbol{X}}(t)-\boldsymbol{X}(t)\|\right] \le C\sqrt{N}\|\hat{\phi}(\cdot)\cdot-\phi(\cdot)\cdot\|_{L^2(\rho_T)},$$

where the measure ρ_T *is defined in Eq.* **4** *and* $C = C(T, C_{\mathrm{Lip}})$.

4. Extensions: Heterogeneous Agent Systems, First and Second Order

The method proposed extends naturally to a large variety of interacting agent systems arising in a multitude of applications (4), including systems with multiple types of agents, driven by second-order equations, and including interactions with an environment. For detailed discussions of related topics on self-organized dynamics, we refer the readers to refs. 3 and 32–35 and the recent surveys (36, 37).

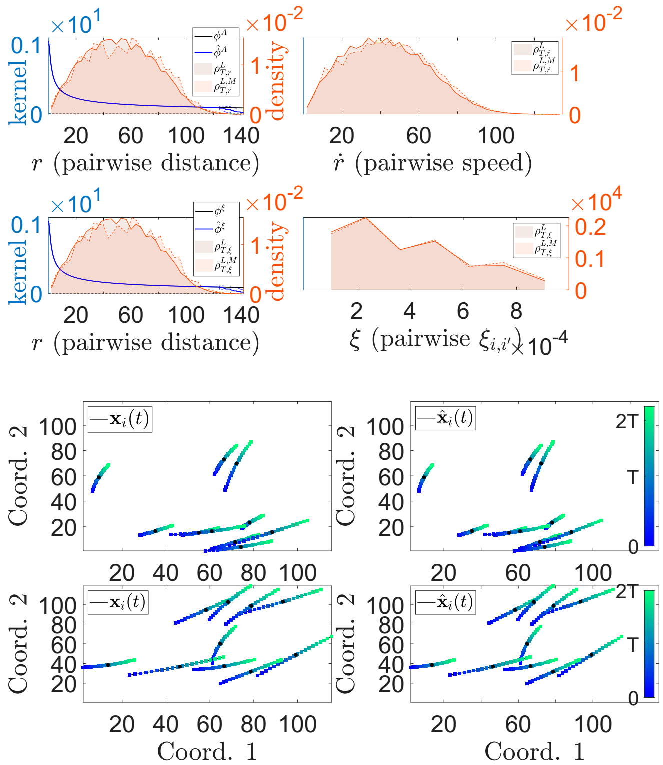


Fig. 6. Estimation of interaction kernels (*Upper*) and trajectory prediction (*Lower*) for the Phototaxis system. Results for the Phototaxis systems, as described in Sections 4 and 5, are shown. (*Upper*) *Left* represents ϕ^A vs. $\hat\phi^A$ (top row), and ϕ^ξ vs. $\hat\phi^\xi$ (bottom row), superimposed with the histograms of $\rho^L_{T,r}$ and, respectively, $\rho^{L,M}_{T,r}$. *Right* shows the comparison of the marginal distributions, $\rho^L_{T,\dot r}$ vs. $\rho^{L,M}_{T,\dot r}$ and $\rho^L_{T,\xi}$ vs. $\rho^{L,M}_{T,\xi}$. (*Lower*) *Left* represents the trajectories generated from true interaction kernels, whereas *Right* shows the trajectories generated by the estimated kernels, generated from training IC data (top row) and from a new random IC (bottom row). In this system, the interaction kernels ϕ^A and ϕ^ξ are the same; the corresponding estimators $\hat\phi^A$ and $\hat\phi^\xi$ are both learned accurately, but note that they are being learned from two different sets of data, $(r,\dot r)$ and (r,ξ), respectively. In both cases, data are scarce or missing for large values r, leading to estimators tapering to 0 faster than the true interaction kernels. However, despite the undesired tail-end behavior of our estimators, the estimators perform extremely well in regenerating the trajectories. See *SI Appendix*, section 3 for more details. Coord., coordinate.

A. First-Order Heterogeneous Agents Systems. Let the agents be divided into K disjoint sets $\{C_k\}_{k=1}^K$ ("types"), with different interaction kernels for each ordered pair of types:

$$\dot{\boldsymbol{x}}_i(t) = \sum_{i'=1}^N \frac{1}{N_{\mathrm{k}_{i'}}}\phi_{\mathrm{k}_i\mathrm{k}_{i'}}(r_{ii'}(t))\boldsymbol{r}_{ii'}(t), \tag{8}$$

where k_i is the index of the type of agent i—i.e., $i\in C_{\mathrm{k}_i}$; $N_{\mathrm{k}_{i'}}$ is the number of agents in type $C_{\mathrm{k}_{i'}}$; $\boldsymbol{r}_{ii'} = \boldsymbol{x}_{i'} - \boldsymbol{x}_i$ and $r_{ii'} = \|\boldsymbol{r}_{ii'}\|$; $\phi_{kk'}:\mathbb{R}_+\to\mathbb{R}$ is the interaction kernel governing how agents in type $C_{k'}$ influence agents in type C_k. As usual we let $\boldsymbol{X} := (\boldsymbol{x}_i)_{i=1}^N \in \mathbb{R}^{dN}$ be the vector describing the state of the system. We assume that the interaction kernels $\phi_{\mathrm{k}_i\mathrm{k}_{i'}}$'s are the only unknown factors in the model; in particular, we know the sets C_k's (i.e., the type of each agent is known). The goal is to infer the interaction kernels $\phi_{kk'}$ from observations $\{\boldsymbol{X}^m(t_l)\}_{l,m=1}^{L,M}$ with $0 = t_1 < \ldots < t_l = T$ and with the ICs $\boldsymbol{X}^m(0) = \boldsymbol{X}_0^m$ randomly sampled from μ_0.

Let $\mathbf{f}_{\boldsymbol{\phi}}(\boldsymbol{X}^m)\in\mathbb{R}^{dN}$ be the vectorization of the right hand sides of Eq. **8**, and $\boldsymbol{\phi} = (\phi_{kk'})_{k,k'=1}^K$. Dropping from the notation of quantities that are assumed known, we rewrite the equations for the dynamics in Eq. **8** as $\dot{\boldsymbol{X}}^m = \mathbf{f}_{\boldsymbol{\phi}}(\boldsymbol{X}^m)$. We use an error functional similar to Eq. **2**, with a weighted norm, to define the estimators:

$$\widehat{\boldsymbol{\phi}} := \underset{\boldsymbol{\varphi}\in\mathcal{H}}{\operatorname{argmin}}\ \frac{1}{ML}\sum_{m=1,l=1}^{M,L}\left\|\dot{\boldsymbol{X}}^m(t_l) - \mathbf{f}_{\boldsymbol{\varphi}}(\boldsymbol{x}^m(t_l))\right\|_{\mathcal{S}}^2, \tag{9}$$

where $\boldsymbol{\varphi} = (\varphi_{kk'})_{k,k'=1}^K$, $\widehat{\boldsymbol{\phi}} = (\hat\phi_{kk'})_{k,k'=1}^K$ and $\|\boldsymbol{X}\|_{\mathcal{S}}^2 := \sum_{i=1}^N \frac{1}{N_{\mathrm{k}_i}}\|\boldsymbol{x}_i\|^2$. The weighted norm $\|\cdot\|_{\mathcal{S}}^2$ is introduced so that, when different types of agents have significantly different cardinalities (e.g., a large number of preys vs. a single predator), the error functional will take into suitable consideration the least numerous type. Otherwise, only the interaction kernel of the most numerous type of agents would be accurately learned. Other more general weighting strategies may be considered, with minimal changes to the algorithm.

The generalization of ρ_T^L in Eq. **5** (similarly for ρ_T) to the heterogeneous-agent case is the family, indexed by ordered pairs $\{(k,k')\}_{k,k'\in\{1,\ldots,K\}}$, of probability measures on $\mathbb{R}_+$

$$\rho_T^{L,kk'}(r) = \frac{1}{LN_{kk'}}\sum_{l=1}^L \mathbb{E}_{\mathbf{X}_0\sim\mu_0}\sum_{i\in C_k, i'\in C_{k'}, i\neq i'} \delta_{r_{ii'}(t_l)}(r), \tag{10}$$

where $N_{kk'} = N_k N_{k'}$ when $k\neq k'$ and $N_{kk'} = \binom{N_k}{2}$ when $k = k'$ (for $N_k > 1$, otherwise there is no interaction kernel to learn). The error of an estimator, $\hat\phi_{kk'}$, will be measured by $\left\|\hat\phi_{kk'}(\cdot)\cdot - \phi_{kk'}(\cdot)\cdot\right\|_{L^2(\rho_T^{L,kk'})}$.

While this case requires learning multiple interaction kernels, it turns out that the learning theory developed for the single-type agent systems can be generalized, and the estimator in Eq. **9** still achieves optimal rates of convergence, and a similar control on the error of predicted trajectories can be obtained.

B. Second-Order Heterogeneous Agent Systems. Here, we focus on a broad family of second-order multitype agent systems (not included, even when rewritten as first-order systems, in the family discussed above). We consider systems with K types of agents:

$$\begin{cases} m_i\ddot{\boldsymbol{x}}_i = F_i^{\boldsymbol{v}}(\dot{\boldsymbol{x}}_i,\xi_i) + \displaystyle\sum_{i'=1}^N \frac{1}{N_{\mathrm{k}_{i'}}}\left(\phi^E_{\mathrm{k}_i\mathrm{k}_{i'}}(r_{ii'})\boldsymbol{r}_{ii'} + \phi^A_{\mathrm{k}_i\mathrm{k}_{i'}}(r_{ii'})\dot{\boldsymbol{r}}_{ii'}\right) \\ \dot\xi_i = F_i^\xi(\xi_i) + \displaystyle\sum_{i'=1}^N \frac{1}{N_{\mathrm{k}_{i'}}}\phi^\xi_{\mathrm{k}_i\mathrm{k}_{i'}}(r_{ii'})\xi_{ii'}, \end{cases} \tag{11}$$

for $i = 1,\ldots,N$. Here $\mathrm{k}_i\in\{1,\ldots,K\}$ is the type of agent i, $\xi_i\in\mathbb{R}$ is a variable modeling the agent's response to the environment (e.g., food/light source), $\xi_{ii'} = \xi_{i'} - \xi_i$, and m_i, N_k, mass of agent i and number of agents of type k; $F_i^{\boldsymbol{v}}$, F_i^ξ, noncollective influences on $\dot{\boldsymbol{x}}_i$ and ξ_i; and $\phi^E_{kk'}$, $\phi^A_{kk'}$, $\phi^\xi_{kk'}$, energy-, alignment-, and $\xi-$type interaction kernels.

Note that here each agent is influenced by a weighted sum of different influences over agents of different types, leading to a rich family of models (including but not limited to prey–predator, leader–follower, and cars–pedestrian models). Using vector notation, let $\mathbf{f}_{\boldsymbol{\phi}^E}(\boldsymbol{X}^m)$ and $\mathbf{f}_{\boldsymbol{\phi}^A}(\boldsymbol{X}^m,\dot{\boldsymbol{X}}^m)\in\mathbb{R}^{dN}$ be the collection of the energy and alignment induced interaction terms respectively, and $\mathcal{F}^{\boldsymbol{v}}(\dot{\boldsymbol{X}}^m,\Xi^m)_i = F_i^{\boldsymbol{v}}(\dot{\boldsymbol{x}}_i,\xi_i)$ (similar setup for $\mathcal{F}^\xi(\Xi^m)$ and $\mathbf{f}_{\boldsymbol{\phi}^\xi}(\boldsymbol{X}^m,\Xi^m)$) we can rewrite the equations as:

$$\begin{cases} \ddot{\boldsymbol{X}}^m = \mathcal{F}^{\boldsymbol{v}}(\dot{\boldsymbol{X}}^m,\Xi^m) + \mathbf{f}_{\boldsymbol{\phi}^E}(\boldsymbol{X}^m) + \mathbf{f}_{\boldsymbol{\phi}^A}(\boldsymbol{X}^m,\dot{\boldsymbol{X}}^m) \\ \dot\Xi^m = \mathcal{F}^\xi(\Xi^m) + \mathbf{f}_{\boldsymbol{\phi}^\xi}(\boldsymbol{X}^m,\Xi^m), \end{cases} \tag{12}$$

Table 1. Model selection: First- vs. second-order

System	Learned as first order	Learned as second order
First-order system	**0.01 ± 0.002**	1.6 ± 1.1
Second-order system	1.7 ± 0.3	**0.2 ± 0.06**

The table shows the mean and SD of the errors of estimated trajectories, over $M = 250$ train-test runs, with random ICs in each case. Small errors, consistent with our theory that the errors are on a scale of $M^{-2/5}$, indicate a correct model. The order is correctly identified in each case (highlighted in bold).

where $\boldsymbol{\phi}^E = \{\phi^E_{kk'}\}$, $\boldsymbol{\phi}^A = \{\phi^A_{kk'}\}$ and $\boldsymbol{\phi}^\xi = \{\phi^\xi_{kk'}\}$, with $k, k' = 1, \ldots, K$. We assume that the interaction kernels are the only unknowns in the model, to be estimated from the observations $\{\boldsymbol{X}^m(t_l), \dot{\boldsymbol{X}}^m(t_l), \Xi^m(t_l)\}_{l,m=1}^{L,M}$, with M ICs $\boldsymbol{X}^m_0 := \boldsymbol{X}^m(0)$, $\dot{\boldsymbol{X}}^m_0 := \dot{\boldsymbol{X}}^m(0)$, and $\Xi^m_0 := \Xi^m(0)$ sampled independently from $\mu_0^{\boldsymbol{X}}$, $\mu_0^{\dot{\boldsymbol{X}}}$, and μ_0^{Ξ}, respectively. With $\ddot{\boldsymbol{X}}^m(t_l)$ approximated by finite difference, we construct estimators similar to those in Eq. **2**

$$(\widehat{\boldsymbol{\phi}}^E, \widehat{\boldsymbol{\phi}}^A) := \underset{\boldsymbol{\varphi}^E, \boldsymbol{\varphi}^A \in \mathcal{H}^{\boldsymbol{v}}}{\operatorname{argmin}} \frac{1}{ML} \sum_{m,l=1}^{M,L} \Big\| \ddot{\boldsymbol{X}}^m(t_l) - \mathcal{F}^{\boldsymbol{v}}(\dot{\boldsymbol{X}}^m(t_l), \Xi^m(t_l)) - \mathbf{f}_{\boldsymbol{\varphi}^E}(\boldsymbol{X}^m(t_l)) - \mathbf{f}_{\boldsymbol{\varphi}^A}(\boldsymbol{X}^m(t_l), \dot{\boldsymbol{X}}^m(t_l)) \Big\|^2_{\mathcal{S}}, \qquad [13]$$

and the interactions acting on the auxiliary variable ξ_i can be obtained separately as

$$\widehat{\boldsymbol{\phi}}^\xi := \underset{\boldsymbol{\phi}^\xi \in \mathcal{H}^\xi}{\arg\min} \frac{1}{ML} \sum_{m=1,l=2}^{M,L} \|\dot{\Xi}^m_l - \mathcal{F}^\xi(\Xi^m_l) - \mathbf{f}_{\boldsymbol{\phi}^\xi}(\boldsymbol{X}^m_l, \Xi^m_l)\|^2_{\mathcal{S}},$$

where $\dot{\Xi}^m_l = \dot{\boldsymbol{X}}^m(t_l)$, $\boldsymbol{X}^m_l = \boldsymbol{X}^m(t_l)$, $\Xi^m_l = \Xi^m(t_l)$, $\widehat{\boldsymbol{\phi}}^\xi = \{\hat{\phi}^\xi_{kk'}\}_{k,k'=1}^K$, and the state space norm $\|\cdot\|_{\mathbb{S}}$ is defined similarly to the first-order case. Here, we are using a vectorized notation for $\boldsymbol{\varphi}^E, \boldsymbol{\varphi}^A$, $\mathcal{H}^{\boldsymbol{v}}$ (a suitable product hypothesis space). To measure performance, for each pair (k, k'), we define a probability measure on $\mathbb{R}_+ \times \mathbb{R}_+$

$$\rho_T^{kk'}(r, \dot{r}) = \frac{1}{TN_{kk'}} \int_{t=0}^{T} \mathbb{E} \sum_{i \in C_k, i' \in C_{k'} i \neq i'} \delta_{r_{ii'}(t), \dot{r}_{ii'}(t)}(r, \dot{r})\, dt,$$

and another probability measure on $\mathbb{R}_+ \times \mathbb{R}_+$,

$$\rho^{L,kk'}_{T,r,\xi}(r, \xi) = \frac{1}{LN_{kk'}} \sum_{l=1}^{L} \mathbb{E} \sum_{i \in C_k, i' \in C_{k'}, i \neq i'} \delta_{r_{ii'}(t_l), \xi_{ii'}(t)}(r, \xi),$$

where the expectation is with respect to ICs distributed according to $\mu_0^{\boldsymbol{X}} \times \mu_0^{\dot{\boldsymbol{X}}} \times \mu_0^{\Xi}$, and we let $\dot{r} = \|\dot{\boldsymbol{r}}\|$ (with abuse of notation), $\xi_{ii'}(t) = \big|\xi_{i'}(t) - \xi_i(t)\big|$, $N_{kk'} = N_k N_{k'}$ if $k \neq k'$ and $N_{kk'} = \binom{N_k}{2}$ if $k = k'$ (and $N_k > 1$, as there is no kernel to learn if $N_k = 1$). Let $\rho^{kk'}_{T,r}$ be the marginal of $\rho^{kk'}_T$ with respect to r. We will measure the errors for $\hat{\phi}^E_{kk'}(r)r$, $\hat{\phi}^A_{kk'}(r)\dot{r}$, and $\widehat{\boldsymbol{\phi}}^\xi_{kk'}(r)\xi$ in $L^2(\rho^{kk'}_{T,r})$, $L^2(\rho^{kk'}_T)$, and $L^2(\rho^{kk'}_{T,r,\xi})$, respectively.

The algorithm to construct the estimator in Eq. **13** generalizes that for the first-order single-type agent systems, and involves a LS problem with a structured matrix with K^2 vertical bands indexed by (k, k'), accommodating the estimators for the interaction kernels. Note that such an LS problem takes into account, as it should, the dependencies in learning the various interaction kernels, all at once.

We note that while of course the second-order system may be written as a first-order system in the variables $\boldsymbol{x}_i$ and $\boldsymbol{v}_i = \dot{\boldsymbol{x}}_i$; even when $F^{\boldsymbol{v}}_i \equiv 0$ and $\phi^A_{\mathrm{k}_i, \mathrm{k}_{i'}} \equiv 0$, the resulting equations for $(\boldsymbol{x}_i, \boldsymbol{v}_i)$ are different from those governing the first-order systems considered above in Eq. **8**.

5. Examples

We consider the learning of interaction kernels and the prediction of trajectories for three canonical categories of examples of self-organized dynamics (see *SI Appendix*, section 3 for details).

Opinion Dynamics These are first-order ODE systems with a single type of agent, with bounded, discontinuous, compactly supported, and attraction-only interaction kernels. They model how the opinions of people influence each other and how consensus is formed based on different kinds of influence functions (refs. 14, 15, and 38 and references therein).

Predator–Swarm System We consider a first-order system with a single predator and a swarm of prey, with the interaction kernels (prey–prey, predator–prey, and prey–predator) similar to Lennard–Jones kernels (with appropriate signs to model attractions and repulsions). Different chasing patterns arise depending on the relative interaction strength of predator–prey vs. prey–predator interactions. We also consider a second-order

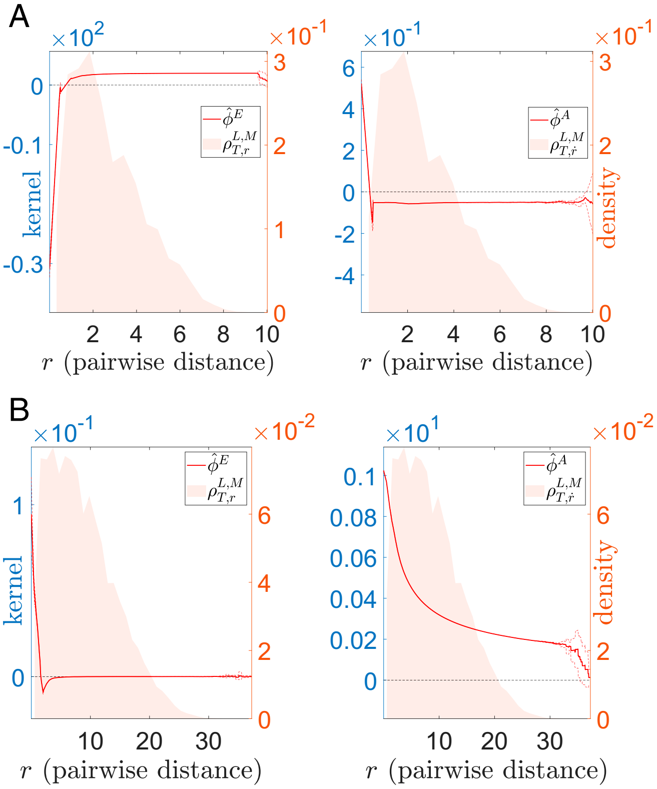


Fig. 7. Model selection: energy-based vs. alignment-based. The estimated interaction kernels for an energy-based model (*A*) and an alignment-based model (*B*). For each model, we compute two estimators: an energy-based interaction kernel $\hat{\phi}^E$ and an alignment-based interaction kernel $\hat{\phi}^A$. Our estimators correctly identify the type of model in each case: The $L^2(\rho^L_{T,r})$ norm of $\hat{\phi}^E$ is significantly larger than that of $\hat{\phi}^A$ (means and SDs: **18.8 ± 0.4** vs. 6.5 ± 0.3) for the energy-based model, and the $L^2(\rho^L_{T,r})$ norm of $\hat{\phi}^A$ is larger than that of $\hat{\phi}^E$ (means and SDs: **27.6 ± 0.7** vs. $2.4 \cdot 10^{-2} \pm 0.1$) for the alignment-based model. Note that the *y* axes are on very different scales.

predator–swarm system, with the collective interaction acting on accelerations, leading to even richer dynamics and chasing patterns (e.g., refs. 39–41).

Phototaxis This is a second-order ODE system with a single type of agents interacting in an environment, modeling phototactic bacteria moving toward a far-away fixed light source. The response of the bacteria to the light source is represented in the auxiliary variable ξ_i as the excitation level for each bacteria i (e.g., refs. 42–44). Another example which we do not pursue here is the Vicsek model (45), which fits perfectly in our model upon choosing $\xi_i = \theta_i$ (θ_i: moving direction of agent i).

In our experiments, we report the measure $\rho_T^{L,M}$ estimated from the training data, our estimator, and similarly in the case of noisy observations; we measure performance in terms of (relative) $L^2(\rho_T^L)$ error of the kernel estimators and of distance between true trajectories $\boldsymbol{X}(t)$ and estimated trajectories $\widehat{\boldsymbol{X}}(t)$, on both the "training" interval $[0, T]$ (where observations were given) and in the future $[T, 2T]$ (predictions). See Prop. 3.4, where the bounds may be overly pessimistic, especially for systems tending to stable configurations. Our estimator performs extremely well in all these examples: The interaction kernels are accurately estimated, and the trajectories are accurately predicted. We refer the reader to Fig. 4 for the results of the opinion dynamics, Fig. 5 for the results of the predator–swarm dynamics, Fig. 6 for the results of the phototaxis, and *SI Appendix*, section 3 for further details on the setup for the experiments and a comprehensive report of all of the results, as well as a detailed description of the final algorithm and its computation complexity in *SI Appendix*, section 2.

Model Selection and Transfer Learning. We also consider the use of our method for model selection, where the theoretical guarantees on learning the interaction kernels and on predicting trajectories are used to decide between different models for the dynamics. We consider two examples of model selection, to test whether: (i) a second-order system is driven by energy-based or alignment-based interactions; or (ii) a heterogeneous agent system is driven by first- or second-order ODEs. For each of them, we construct two estimators assuming either case and then select models according to the performance of the estimators in predicting trajectories. See Table 1 and Fig. 7 for results and discussions and *SI Appendix*, section 3E for details.

As a simple example of transfer learning, we use the interaction kernel learned on a system with N agents to accurately predict trajectories of the same type of system but with more agents ($4N$ in our simulations); the interaction kernel acts as a sort of "latent variable" that seamlessly enables transfer across such related systems. In *SI Appendix*, section 3, we report the corresponding results, for all of the systems considered (see, however, Fig. 1 for the Lennard–Jones system).

Noisy Observations. Our estimators appear robust under observation noise, namely, if the observed positions and derivatives are corrupted by noise. Fig. 8 demonstrates the kernel estimation and trajectory prediction for the first-order predator–swarm system when only noisy observations are available. Similar results (reported in *SI Appendix*, section 3) are obtained in all of the other systems considered.

Choice of the Basis of the Hypothesis Space. Our learning approach is robust to the choice of hypothesis space $\mathcal{H}$, as long as the coercivity condition is satisfied by $\mathcal{H}$ (or the sequence $\mathcal{H}_n$). Additionally, different well-conditioned bases may be used in $\mathcal{H}$ to compute the projection onto $\mathcal{H}$, implying, together with the coercivity condition, a control of the condition number of the LS problem (*SI Appendix*, Prop. 2.1). To demonstrate this numerically, we compare the B-splines linear basis with the piecewise polynomial basis on the 1^{st}-order predator–swarm system, with results shown in *SI Appendix*, Fig. S8.

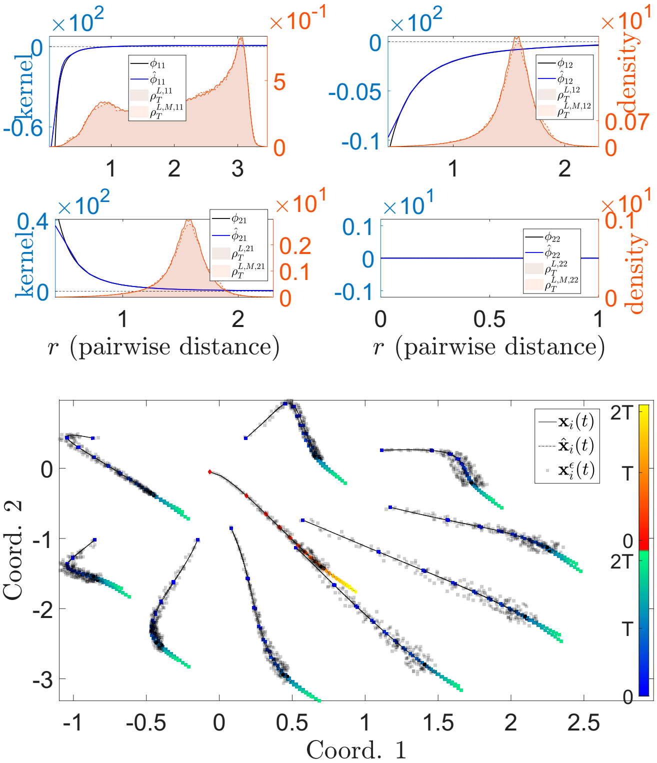


Fig. 8. Kernel estimation for PS1st from noisy observations. (*Upper*) Interaction kernels learned with Unif.($[-\sigma, \sigma]$) multiplicative noise with $\sigma = 0.1$ in the observed positions and velocities, with parameters as in *SI Appendix*, Table S9. The estimated kernels are minimally affected and only in regions with small ρ_T^L. (*Lower*) One of the observed trajectories before and after being perturbed by noise. The solid lines represent the true trajectory, the dashed semitransparent lines represent the noisy trajectory used as training data (together with noisy observations of the derivative), and the dashed-dotted lines are the predicted trajectory learned from the noisy trajectory.

6. Discussion and Conclusion

We proposed a nonparametric estimator for learning interaction kernels from observations of agent systems, implemented by computationally efficient algorithms. We applied the estimator to several classes of systems, including first- and second-order, with single- and multiple-type agents, and with simple environments. We have also considered observation data from different sampling regimes: many short-time trajectories, a single large-time trajectory, and intermediate time scales.

Our inference approach is nonparametric, does not rely on a dictionary of hypotheses (such as in refs. 6–8), exploits the structure of dynamics, and enjoys optimal rates of convergence (which we proved here for first-order systems), independent of the dimension of the state space of the system. Having techniques with solid statistical guarantees is fundamental in establishing trust in data-driven models for these systems and in using them as an aide to the researcher in formulating and testing conjectures about models underlying observed systems. In this vein, we presented two examples of model selection, showing that our estimators can reliably identify the order of a system and identify whether a system is driven by energy- or alignment-type interactions.

We expect further generalizations to the case of stochastic dynamical systems and to the cases of more general interaction kernels that depend on more general types of interaction between agents, beyond pairwise, distance-based interactions. Other future directions include (but are not limited to) a better understanding of learnability, model selection based on the theory, learning from partial observations, and learning reduced models for large systems.

ACKNOWLEDGMENTS. We thank the reviewers for comments, which led to significant improvements to the paper; and Prof. Massimo Fornasier, Prof. Pierre-Emmanuel Jabin, Prof. Yannis Kevrekidis, Prof. Nathan Kutz, Prof. Yaozhong Hu, and Dr. Cheng Zhang for discussions. We thank Duke University and the Maryland Advanced Research Computing Center for access to computing equipment. This work was supported by National Science Foundation Grants DMS-1708602, ATD-1737984, IIS-1546392, DMS-1821211, and IIS-1837991; Air Force Office of Scientific Research Grant AFOSR-FA9550-17-1-0280; and S.T. is grateful for support from the American Mathematical Society-Simons Travel grant.

1. J. A. Carrillo, Y. Choi, S. Perez, "A review on attractive–repulsive hydrodynamics for consensus in collective behavior" in *Active Particles*, N. Bellomo, P. Degond, E. T, Eds. (Birkhäuser, Cham, Switzerland, 2017), Vol 1, pp. 259–298.
2. T. Kolokolnikov, H. Sun, D. Uminsky, A. Bertozzi, Stability of ring patterns arising from two-dimensional particle interactions. *Phys. Rev. E* **84**, 015203(R) (2011).
3. T. Vicsek, A. Zafeiris, Collective motion. *Phys. Rep.* **517**, 71–140 (2012).
4. Y. Shoham, K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundation* (Cambridge University Press, Cambridge, UK, 2009).
5. S. M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900* (Harvard Univ Press, Cambridge, MA, ed. 1, 1986).
6. H. Schaeffer, R. Caflisch, C. D. Hauck, S. Osher, Sparse dynamics for partial differential equations. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6634–6639 (2013).
7. S. Brunton, J. Proctor, J. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 3932–3937 (2016).
8. G. Tran, R. Ward, Exact recovery of chaotic systems from highly corrupted data. *Multi Model Simul.* **15**, 1108–1129 (2017).
9. W. Bialek *et al.*, Statistical mechanics for natural flocks of birds. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 4786–4791 (2012).
10. Y. Li, J. Wu, R. Tedrake, J. B. Tenenbaum, A. Torralba, Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. arXiv:1810.01566 (3 October 2018).
11. M. Ballerini *et al.*, Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1232–1237 (2008).
12. R. Lukeman, Y. X. Li, L. Edelstein-Keshet, Inferring individual rules from collective behavior. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12576–12580 (2010).
13. Y. Katz, K. Tunstrom, C. Ioannou, C. Huepe, I. Couzin, Inferring the structure and dynamics of interactions in schooling fish. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18720–8725 (2011).
14. U. Krause, A discrete nonlinear and non-autonomous model of consensus formation. *Commun. Part. Differ. Equation* **2000**, 227–236 (2000).
15. I. Couzin, J. Krause, N. Franks, S. Levin, Effective leadership and decision-making in animal groups on the move. *Nature* **433**, 513–516 (2005).
16. L. Györfi, M. Kohler, A. Krzyzak, H. Walk, *A Distribution-Free Theory of Nonparametric Regression* (Springer, New York, NY, 2002).
17. M. Bongini, M. Fornasier, M. Hansen, M. Maggioni, Inferring interaction rules from observations of evolutive systems I: The variational approach. *Math. Mod. Methods Appl. Sci.* **27**, 909–951 (2017).
18. H. Schaeffer, G. Tran, R. Ward, Extracting high-dimensional dynamics from limited data. *SIAM J. Appl. Math.* **78**, 3279–3295 (2017).
19. N. J. Brunel, Parameter estimation of ODE's via nonparametric estimators. *Electron. J. Stat.* **2**, 1242–1267 (2008).
20. H. Liang, H. Wu, Parameter estimation for differential equation models using a framework of measurement error in regression models. *J. Am. Stat. Assoc.* **103**, 1570–1583 (2008).
21. J. Cao, L. Wang, J. Xu, Robust estimation for ordinary differential equation models. *Biometrics* **67**, 1305–1313 (2011).
22. J. O. Ramsay, G. Hooker, D. Campbell, J. Cao, Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69**, 741–796 (2007).
23. R. Bellman, R. S. Roth, The use of splines with unknown end points in the identification of systems. *J. Math. Anal. Appl.* **34**, 26–33 (1971).
24. J. M. Varah, A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Comput.* **3**, 28–46 (1982).
25. J. O. Ramsay, Principal differential analysis: Data reduction by differential operators. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 495–508 (1996).
26. M. Pascual, S. P. Ellner, Linking ecological patterns to environmental forcing via nonlinear time series models. *Ecology* **81**, 2767–2780 (2000).
27. J. Timmer, H. Rust, W. Horbelt, H. Voss, Parametric, nonparametric and parametric modelling of a chaotic circuit time series. *Phys. Lett. A* **274**, 123–134 (2000).
28. H. Miao, X. Xia, A. S. Perelson, H. Wu, On identifiability of nonlinear ode models and applications in viral dynamics. *SIAM Rev.* **53**, 3–39 (2011).
29. J. Ramsay, G. Hooker, *Dynamic Data Analysis: Modeling Data with Differential Equations* (Springer Series in Statistics, Springer, New York, NY, 2018).
30. J. von Brecht, D. Uminsky, On soccer balls and linearized inverse statistical mechanics. *J. Nonlinear Sci.* **22**, 935–959 (2012).
31. R. Simione, D. Slepčev, I. Topaloglu, Existence of ground states of nonlocal-interaction energies. *J. Stat. Phys.* **159**, 972–986 (2015).
32. F. Cucker, J. G. Dong, A general collision-avoiding flocking framework. *IEEE Trans. Automat. Contr.* **56**, 1124–1129 (2011).
33. F. Cucker, E. Mordecki, Flocking in noisy environments. *J. Math. Pure Appl.* **89**, 278–296 (2008).
34. G. Grégoire, H. Chaté, Onset of collective and cohesive motion. *Phys. Rev. Lett.* **92**, 025702 (2004).
35. J. Ke, J. W. Minett, C. P. Au, W. S. Y. Wang, Self-organization and selection in the emergence of vocabulary. *Complexity* **7**, 41–54 (2002).
36. J. A. Carrilo, Y. P. Choi, M. Haurray, "The derivation of swarming models: Mean-field limit and Wasserstein distances" in *Collective Dynamics from Bacteria to Crowds: An Excursion Through Modeling, Analysis and Simulation*, A. Muntean, F. Toschi, Eds. (CISM International Centre for Mechanical Sciences Courses and Lectures, Springer, Wien, Austria, Vol. 553, 2014), pp. 1–46.
37. J. A. Carrilo, M. Fornasier, G. Toscani, F. Vecil, "Particle, kinetic, and hydrodynamic models of swarming" in *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences, Modeling and Simulation in Science, Engineering and Technology*, G. Naldi, L. Pareschi, G. Toscani, N. Bellom, Eds. (Springer, Birkhäuser Boston, MA, 2010), pp. 297–336.
38. S. Mostch, E. Tadmor, Heterophilious dynamics enhances consensus. *SIAM Rev.* **56**, 577–621 (2014).
39. Y. Chen, T. Kolokolnikov, A minimal model of predator-swarm interactions. *J. R. Soc. Interf.* **11**, 20131208 (2013).
40. J. Jeschke, R. Tollrian, Prey swarming: Which predators become confused and why? *Anim. Behav.* **74**, 387–393 (2007).
41. M. Zheng, Y. Kashimori, O. Hoshino, K. Fujita, T. Kambara, Behavior pattern (innate action) of individuals in fish schools generating efficient collective evasion from predation. *J. Theor. Biol.* **235**, 13–167 (2005).
42. S. Ha, D. Levy, Particle, kinetic and fluid models for phototaxis. *Discrete Contin. Dyn. Syst. Ser. B* **12**, 77–108 (2009).
43. J. M. Skerker, H. C. Berg, Direct observation of extension and retraction of type IV pili. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 6901–6904 (2001).
44. D. Bhaya, A. Takahashi, A. R. Grossman, Light regulation of type IV pilus-dependent motility by chemosensor-like elements in synechocystis PCC6803. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 7540–7545 (2001).
45. T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, O. Shochet, Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.* **75**, 1226–1229 (1995).

PNAS
www.pnas.org

Supplementary Information for

Nonparametric inference of interaction laws in systems of agents from trajectory data

Fei Lu, Ming Zhong, Sui Tang, Mauro Maggioni

Mauro Maggioni.
E-mail: mauromaggionijhu@icloud.com

This PDF file includes:

Supplementary text
Figs. S1 to S17
Tables S1 to S22
References for SI reference citations

www.pnas.org/cgi/doi/10.1073/pnas.1822012116

Supporting Information Text

1. Learning Theory

Consider the problem of estimating the interaction kernel $\phi : \mathbb{R}_+ \to \mathbb{R}$ of the dynamical system as follows

$$\dot{\boldsymbol{x}}_i(t) = \frac{1}{N}\sum_{i'=1}^{N} \phi(\|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|\,)(\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t))\,, \qquad [1]$$

from observations of discrete-time trajectories and derivatives, $\{\boldsymbol{X}^m(t_l)\}$ and $\{\dot{\boldsymbol{X}}^m(t_l)\}$ with $0 = t_1 < \cdots < t_L = T$ and $m = 1, \ldots, M$. We let $\boldsymbol{X} := (\boldsymbol{x}_i)_{i=1}^N \in \mathbb{R}^{dN}$ be the state space variable. The initial conditions $\boldsymbol{X}_0^m := \boldsymbol{X}^m(0)$ are sampled independently from a probability measure μ_0 on $\mathbb{R}^{dN}$.

Such a system can also be described as the gradient flow $\dot{\boldsymbol{X}} = \mathbf{f}_\phi(\boldsymbol{X}) = \nabla\mathcal{U}(\boldsymbol{X})$ of the potential energy $\mathcal{U}(\boldsymbol{X}) = \frac{1}{2N}\sum_{i,i'}\Phi(\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|)$, with the function $\Phi : \mathbb{R}_+ \to \mathbb{R}$ satisfying $\Phi'(r) = \phi(r)r$. Therefore, the estimation of ϕ is equivalent to the estimation of Φ'. As we will see later, the function $\phi(\cdot)\cdot$ appears naturally in assessing the quality of approximation of estimators of ϕ, the fundamental reason being the relationship with the potential involving Φ.

We restrict our attention to kernels in the *admissible set*

$$\mathcal{K}_{R,S} := \{\phi \in W^{1,\infty} : \mathrm{supp}(\phi) \in [0,R], \sup_{r\in[0,R]} \Big[|\phi(r)| + |\phi'(r)|\Big] \le S\} \qquad [2]$$

for some $R, S > 0$. The boundedness of ϕ and its derivative ensures the existence and uniqueness of a global solution to initial value problems of the first order system Eq. (1), and the continuous dependence of the solution on the initial condition. The restriction $\mathrm{supp}(\phi) \subset [0,R]$ represents the finite range of interaction between particles, and this restriction may be replaced by functions with unbounded support but with a suitable decay on $\mathbb{R}_+$.

We shall construct an error functional based on the special structure of the dynamical system $\dot{\boldsymbol{X}} = \mathbf{f}_\phi(\boldsymbol{X})$, taking advance of the form of the dependency of the right-hand side $\mathbf{f}_\phi$ on the interaction kernel ϕ. This learning procedure deviates from standard regression in two aspects: (i) the values of the interaction kernel are not observed, and cannot be explicitly estimated from the observations of the state variables; (ii) the observations of the independent variable of the interaction kernel, given by the pairwise distance between the agents, though abundant, are not independent and may be redundant.

We would also like to stress the importance of using a carefully chosen measure on the pairwise distance space, so as to account for both the randomness from the initial conditions and the evolution of the dynamical system, and to reflect the (relative) abundances of pairwise distances. Our analysis shows that the expectation of the empirical measure of the pairwise distances is a natural choice, and it is closely related to the coercivity condition, the other fundamental ingredient which ensures learnability and convergence of the estimators.

A. The Error functional and estimators.

Given the structure of the first order system Eq. (1), we consider the error functional

$$\mathcal{E}_{L,M}(\varphi) := \frac{1}{MN}\sum_{l,m,i=1}^{L,M,N} w_l \Big\|\dot{\boldsymbol{x}}_i^m(t_l) - \mathbf{f}_\varphi(\boldsymbol{x}^m(t_l))_i\Big\|^2\,, \qquad [3]$$

where $\{w_l\}_{l=1}^L$ is a normalized set of weights ($w_l > 0$ and $\sum_{l=1}^L w_l = 1$), and define an estimator

$$\widehat{\phi}_{L,M,\mathcal{H}} := \underset{\varphi\in\mathcal{H}}{\arg\min}\, \mathcal{E}_{L,M}(\varphi), \qquad [4]$$

where $\mathcal{H}$ is a suitable class of functions that will be referred as hypothesis space. Natural choices of weights $\{w_l\}$ may be chosen to be all equal to $1/L$, as in the case of equi-spaced t_l's, which is what we considered throughout the paper, and is consistent with the definition of ρ_T^L and its use in measuring the performance of the estimator in $L^2(\rho_T^L)$. However, if one wished to measure the performance in a different L^2 space, one could choose the weights differently. A distinguished choice would be $L^2(\rho_{\mathrm{Lebesgue}})$, in which case one may choose $w_l = 1/(t_{l+1} - t_l)$, for $l = 1, \ldots, L-1$ (and change all the summations involving l to stop at $L-1$ instead of L). Other choices of weights corresponding to other quadrature rules are also be possible.

Note that the error functional is quadratic in φ and bounded below by 0, therefore the minimizer exists for any finite dimensional convex hypothesis spaces $\mathcal{H}$. We can always truncate this minimizer so that it is bounded above by S, the upper bound of the functions in the admissible set $\mathcal{K}_{R,S}$, and this truncated estimator behaves similarly to the estimator obtained by assuming that the functions in $\mathcal{H}$ are uniformly bounded. In fact, such truncation can only reduce the error. Hence, without loss of generality, we assume $\mathcal{H}$ to be a compact set in the L^∞ norm.

Our objectives are measuring the quality of approximation of the estimator and finding the hypothesis spaces for which the optimal rate of convergence of $\hat{\phi}$ to the true interaction kernel ϕ is achieved.

B. Measures on the pairwise distance space. We introduce a probability measure on $\mathbb{R}_+$, to define a suitable function space that contains all the estimators and the true interaction kernel, and to provide a norm to assess the accuracy of the estimators. We let

$$\mathbf{r}_{ii'}(t) = \boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t), \text{ and } r_{ii'}(t) = \|\boldsymbol{x}_{i'}(t) - \boldsymbol{x}_i(t)\|.$$

Note that the independent variable of the interaction kernel is the pairwise distances $r^m_{ii'}(t)$, which can be computed from the observed trajectories. It is natural to start from the empirical measure of pairwise distances

$$\rho_T^{L,M}(r) = \frac{1}{\binom{N}{2}LM} \sum_{l,m=1}^{L,M} \sum_{i,i'=1,i<i'}^{N} \delta_{r^m_{ii'}(t_l)}(r)\,, \qquad [5]$$

which tends, as $M \to \infty$, using the law of large numbers, to ρ_T^L defined in (5) in the main text. When trajectories are observed continuously in time, the counterpart of ρ_T^L is the measure defined in (5). We now establish basic properties of these measures:

Lemma 1.1. *For each $\phi \in \mathcal{K}_{R,S}$ defined in* Eq. (2)*, the measures ρ_T^L and ρ_T defined in (5) and (4) in the main text are Borel probability measures on $\mathbb{R}_+$. They are absolutely continuous with respect to the Lebesgue measure provided that μ_0 is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{dN}$.*

C. Learnability: the coercivity condition. A fundamental question is the learnability of the true interaction kernel, i.e. the well-posedness of the inverse problem of kernel learning. Since the estimators $\hat{\phi}_{L,M,\mathcal{H}}$ always exists for suitably chosen hypothesis spaces $\mathcal{H}$ (e.g. compact sets), learnability is equivalent to the convergence of the estimator $\hat{\phi}_{L,M,\mathcal{H}}$ to the true kernel ϕ as the sample size increases (i.e. $M \to \infty$) and as the hypothesis space grows. To ensure such a convergence, one would naturally wish: (i) that the true kernel ϕ is the unique minimizer of the expectation of the error functional (by the law of large numbers)

$$\mathcal{E}_{L,\infty}(\varphi) := \lim_{M\to\infty} \mathcal{E}_{L,M}(\varphi) = \frac{1}{LN} \sum_{l,i=1}^{L,N} \mathbb{E}\left[\Big\| \frac{1}{N} \sum_{i'=1}^{N} (\varphi - \phi)\,(r_{ii'}(t_l))\boldsymbol{r}_{ii'}(t_l) \Big\|^2 \right]; \qquad [6]$$

(ii) that the error of the estimator, in terms of a metric based on the $L^2(\rho_T^L)$ norm, can be controlled by the discrepancy between the empirical error functional and its limit.

Note that $\mathcal{E}_{L,\infty}(\varphi) \geq 0$ for any φ and that $\mathcal{E}_{L,\infty}(\phi) = 0$. Furthermore, Eq. (6) reveals that $\mathcal{E}_{L,\infty}(\varphi)$ is a quadratic functional of $\varphi - \phi$, and we have, by Jensen's inequality,

$$\mathcal{E}_{L,\infty}(\varphi) \leq \frac{(N-1)^2}{N^2} \|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|^2_{L^2(\rho_T^L)}\ .$$

This inequality suggests the above weighted $L^2(\rho_T^L)$ norm as a metric on the error of the estimator that we wish to be controlled. Therefore, as long we as can bound the limit error functional from below by $\|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|^2_{L^2(\rho_T^L)}$, we can conclude that ϕ is the unique minimizer of $\mathcal{E}_{L,\infty}(\cdot)$ and that the estimators converge to ϕ. This suggests the following coercivity condition:

Definition 1.1 (Coercivity condition)**.** *We say that the dynamical system defined in* Eq. (1) *together with the probability measure μ_0 on $\mathbb{R}^{dN}$, satisfies the coercivity condition on $\mathcal{H}$ with a constant $c_{L,N,\mathcal{H}} > 0$, if*

$$c_{L,N,\mathcal{H}}\|\varphi(\cdot)\cdot\|^2_{L^2(\rho_T^L)} \leq \frac{1}{NL} \sum_{i,l=1}^{L,N} \mathbb{E}\left[\Big\| \frac{1}{N} \sum_{i'=1}^{N} \varphi(r_{ii'}(t_l))\boldsymbol{r}_{ii'}(t_l) \Big\|^2 \right] \qquad [7]$$

for all $\varphi \in \mathcal{H}$ such that $\varphi(\cdot)\cdot \in L^2(\rho_T^L)$, with the measure ρ_T^L defined in (4) in the main text, and the expectation being with respect to initial conditions distributed according to μ_0.

The above inequality is called a coercivity condition because that it implies coercivity of the bilinear functional $\langle\!\langle \cdot,\cdot \rangle\!\rangle$ on $L^2(\mathbb{R}_+, \rho_T^L)$,

$$\langle\!\langle \varphi_1, \varphi_2 \rangle\!\rangle := \frac{1}{LN} \sum_{l,i=1}^{L,N} \mathbb{E}\left[\left\langle \frac{1}{N} \sum_{j=1}^{N} \varphi_1(r_{ji}(t_l))\boldsymbol{r}_{ij}(t_l), \frac{1}{N} \sum_{j=1}^{N} \varphi_2(r_{ji}(t_l))\boldsymbol{r}_{ij}(t_l) \right\rangle \right], \qquad [8]$$

as Eq. (7) may be rewritten as

$$c_{L,N,\mathcal{H}}\, \|\varphi(\cdot)\cdot\|^2_{L^2(\mathbb{R}_+,\rho_T^L)} \leq \langle\!\langle \varphi, \varphi \rangle\!\rangle.$$

The coercivity condition plays a key role in the learning of the interaction kernel. It ensures learnability by ensuring the uniqueness of minimizer of the expectation of the error functional, and by guaranteeing convergence of estimators through a control of the error of the estimator on every compact convex hypothesis space $\mathcal{H}$ in $L^2(\rho_T^L)$. To see this, apply the coercivity inequality to $\varphi - \phi$, to obtain

$$c_{L,N,\mathcal{H}}\, \|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|^2_{L^2(\mathbb{R}_+,\rho_T^L)} \leq \mathcal{E}_{L,\infty}(\varphi). \qquad [9]$$

From the facts that $\mathcal{E}_{L,\infty}(\varphi) \geq 0$ for any φ and that $\mathcal{E}_{L,\infty}(\phi) = 0$, we can conclude that the true kernel ϕ is the unique minimizer of the $\mathcal{E}_{L,\infty}(\varphi)$. Furthermore, the coercivity condition enables us to control the error of the estimator, on every compact convex hypothesis space in $L^2(\rho_T^L)$, by the discrepancy of the error functional (see Proposition 1.3), therefore guaranteeing convergence of the estimator.

Theorem 1.2. *Let $\mathcal{H}_n$ be a sequence of compact convex subsets of $L^\infty([0,R])$ such that*

$$\inf_{\varphi\in\mathcal{H}_n}\|\varphi(\cdot)\cdot-\phi(\cdot)\cdot\|_{L^2(\rho_T^L)}\to 0$$

as $n\to\infty$. Assume that the coercivity condition holds on $\cup_{n=1}^\infty\mathcal{H}_n$. Then the estimator $\widehat{\phi}_{L,M,\mathcal{H}_n}$ defined in Eq. (4) *converges to the true kernel in $L^2(\rho_T^L)$ almost surely as n,M approaches infinity, i.e.*

$$\lim_{n\to\infty}\lim_{M\to\infty}\|\widehat{\phi}_{L,M,\mathcal{H}_n}(\cdot)\cdot-\phi(\cdot)\cdot\|_{L^2(\rho_T^L)}=0,\ \textit{almost surely.}$$

The above theorem follows from the next proposition.

Proposition 1.3. *Let $\mathcal{H}$ be a compact convex subset of $L^2(\rho_T^L)$ and assume the coercivity condition holds true on $\mathcal{H}$. Then the functional $\mathcal{E}_{L,\infty}$ defined in* Eq. (6) *admits a unique minimizer*

$$\widehat{\phi}_{L,\infty,\mathcal{H}}=\arg\min_{\varphi\in\mathcal{H}}\mathcal{E}_{L,\infty}(\varphi), \qquad [10]$$

in $L^2(\rho_T^L)$. Furthermore, for all $\varphi\in\mathcal{H}$

$$\mathcal{E}_{L,\infty}(\varphi)-\mathcal{E}_{L,\infty}(\widehat{\phi}_{L,\infty,\mathcal{H}})\geq c_{L,N,\mathcal{H}}\|\varphi(\cdot)\cdot-\widehat{\phi}_{L,\infty,\mathcal{H}}(\cdot)\cdot\|^2_{L^2(\rho_T^L)}. \qquad [11]$$

D. Optimal rate of convergence of the estimator. We now turn to the rate of convergence of the estimator.

Theorem 1.4. *Let the true kernel $\phi\in\mathcal{K}_{R,S}$, and let $\mathcal{H}\subset L^\infty([0,R])$ be compact convex and bounded above by $S_0\geq S$. Assume that the coercivity condition in* Eq. (7) *holds. Then for any $\epsilon>0$, we have*

$$c_{L,N,\mathcal{H}}\|\widehat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot-\phi(\cdot)\cdot\|^2_{L^2(\rho_T^L)}\leq 2\inf_{\varphi\in\mathcal{H}}\|\varphi(\cdot)\cdot-\phi(\cdot)\cdot\|^2_{L^\infty([0,R])}+2\epsilon \qquad [12]$$

with probability at least $1-\delta$, provided that

$$M\geq\frac{1152S_0^2R^2}{c_{L,N,\mathcal{H}}\epsilon}\big(\log(\mathcal{N}(\mathcal{H},\frac{\epsilon}{48S_0R^2}))+\log(\frac{1}{\delta})\big),$$

where $\mathcal{N}(\mathcal{H},\eta)$ is the η-covering number of $\mathcal{H}$ under the ∞-norm.

We discuss first the implications of this theorem on the choice of hypothesis space in view of obtaining optimal rates of convergence of our estimator. The proof of the theorem will be presented at the end of this section. In practice, given a set of M trajectories, we would like to chose the best finite-dimensional hypothesis space $\mathcal{H}$ to minimize the error of the estimator. There are two competing issues. On one hand, we would like the hypothesis space $\mathcal{H}$ to be large so that the bias $\inf_{\varphi\in\mathcal{H}}\|\varphi-\phi\|^2_{L^\infty([0,R])}$ is small. On the other hand, we would like to keep $\mathcal{H}$ to be small so that the covering number $\mathcal{N}(\mathcal{H},\epsilon/48S_0R^2)$, and therefore the variance of the estimator is small. This is the classical bias-variance trade-off in statistical estimation. Inspired from approximation methods in regression (1–3) , the following proposition quantifies the effect of hypothesis spaces on the rate of convergence of the estimator.

Proposition 1.5. *Assume that the coercivity condition holds with a constant $c_{L,N,\mathcal{H}}$, and recall $\widehat{\phi}_{L,M,\mathcal{H}}$ defined in* Eq. (4) *is a minimizer of the empirical error functional over a hypothesis space $\mathcal{H}$.*
(a) For $\mathcal{H}=\mathcal{K}_{R,S}$, there exists a constant $C=C(S,R)$ such that

$$\mathbb{E}[\|\widehat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot-\phi(\cdot)\cdot\|_{L^2(\rho_T^L)}]\leq\frac{C}{c_{L,N,\mathcal{H}}}M^{-\frac{1}{4}}.$$

(b) Assume that $\mathcal{H}_n$ is a sequence of finite dimensional spaces of $L^\infty([0,R])$ such that dim$(\mathcal{H}_n)\leq c_0n$ and

$$\inf_{\varphi\in\mathcal{H}_n}\|\varphi(\cdot)-\phi(\cdot)\|^2_{L^\infty([0,R])}\leq c_1n^{-s} \qquad [13]$$

for all n for some constants $c_0,c_1,s>0$, then by choosing $n=n_*:=(M/\log M)^{\frac{1}{2s+1}}$, we have

$$\mathbb{E}[\|\widehat{\phi}_{L,M,\mathcal{H}_{n_*}}(\cdot)\cdot-\phi(\cdot)\cdot\|_{L^2(\rho_T^L)}]\leq\frac{C}{c_{L,N,\mathcal{H}}}\left(\frac{\log M}{M}\right)^{\frac{s}{2s+1}},$$

where $C=C(c_0,c_1,R,S)$.

It is interesting to compare this rate with those in the mean field regime, where the regime $N\to\infty$ (with $M=1$, $L\to\infty$) was studied: the rates implied by (4) they seem to be no better than $N^{-1/d}$, i.e. they are cursed by the dimension d, even if the problem is fundamentally that of estimating a 1-dimensional function. It would be interesting to understand whether that rate is optimal for this problem in the mean-field regime ($N\to\infty$), or if in fact, the results in the present work lead to sharper, dimension-independent bounds in the mean-field limit as well.

The proof of Thm. 1.4 is based on this technical Proposition:

Proposition 1.6. *Assume the coercivity condition holds true and let $\mathcal{H} \subset L^\infty([0,R])$ be compact convex, bounded above by S_0. Let*

$$\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi) := \mathcal{E}_{L,\infty}(\varphi) - \mathcal{E}_{L,\infty}(\widehat{\phi}_{L,\infty,\mathcal{H}}) \quad , \quad \mathcal{D}_{L,M,\mathcal{H}}(\varphi) := \mathcal{E}_{L,M}(\varphi) - \mathcal{E}_{L,M}(\widehat{\phi}_{L,\infty,\mathcal{H}}),$$

where $\widehat{\phi}_{L,\infty,\mathcal{H}}$ is the minimizer of $\mathcal{E}_{L,\infty}(\cdot)$ over $\mathcal{H}$. Then for all $\epsilon > 0$ and $0 < \alpha < 1$, we have

$$P\left\{ \sup_{\varphi \in \mathcal{H}} \frac{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi) - \mathcal{D}_{L,M,\mathcal{H}}(\varphi)}{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi) + \epsilon} \geq 3\alpha \right\} \leq \mathcal{N}\left(\mathcal{H}, C_1\alpha\epsilon\right) e^{-C_2\alpha^2 M\epsilon}$$

where $C_1 = \frac{1}{8S_0R^2}$ and $C_2 = \frac{-c_{L,N,\mathcal{H}}}{32S_0^2R^2}$.

***Proof of the Theorem* 1.4** . Put $\alpha = \frac{1}{6}$ in Proposition 1.6. We know that, with probability at least

$$1 - \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{48S_0R^2}\right) e^{-\frac{c_{L,N,\mathcal{H}}M\epsilon}{1152S_0^2R^2}},$$

we have

$$\sup_{\varphi \in \mathcal{H}} \frac{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi) - \mathcal{D}_{L,M,\mathcal{H}}(\varphi)}{\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi) + \epsilon} < \frac{1}{2},$$

and therefore, for all $\varphi \in \mathcal{H}$,

$$\frac{1}{2}\mathcal{D}_{L,\infty,\mathcal{H}}(\varphi) < \mathcal{D}_{L,M,\mathcal{H}}(\varphi) + \frac{1}{2}\epsilon.$$

Taking $\varphi = \widehat{\phi}_{L,M,\mathcal{H}}$, we have

$$\mathcal{D}_{L,\infty,\mathcal{H}}(\widehat{\phi}_{L,M,\mathcal{H}}) < 2\mathcal{D}_{L,M,\mathcal{H}}(\widehat{\phi}_{L,M,\mathcal{H}}) + \epsilon\,.$$

But $\mathcal{D}_{L,M,\mathcal{H}}(\widehat{\phi}_{L,M,\mathcal{H}}) = \mathcal{E}_{L,M}(\widehat{\phi}_{L,M,\mathcal{H}}) - \mathcal{E}_{L,M}(\widehat{\phi}_{L,\infty,\mathcal{H}}) \leq 0$ and hence by Proposition 1.3 we have

$$c_{L,N,\mathcal{H}}\|\widehat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot - \widehat{\phi}_{L,\infty,\mathcal{H}}(\cdot)\cdot\|^2_{L^2(\rho_T^L)} \leq \mathcal{D}_{L,\infty,\mathcal{H}}(\widehat{\phi}_{L,M,\mathcal{H}}) < \epsilon.$$

Therefore,

$$\begin{aligned}\|\widehat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot - \phi(\cdot)\cdot\|^2_{L^2(\rho_T^L)} &\leq 2\|\widehat{\phi}_{L,M,\mathcal{H}}(\cdot)\cdot - \widehat{\phi}_{L,\infty,\mathcal{H}}(\cdot)\cdot\|^2_{L^2(\rho_T^L)} + 2\|\widehat{\phi}_{L,\infty,\mathcal{H}}(\cdot)\cdot - \phi(\cdot)\cdot\|^2_{L^2(\rho_T^L)}\\ &\leq \frac{2}{c_{L,N,\mathcal{H}}}(\epsilon + \inf_{\varphi\in\mathcal{H}}\|\varphi(\cdot)\cdot - \phi(\cdot)\cdot\|^2_\infty),\end{aligned}$$

where the last inequality follows from the coercivity condition and by the definition of $\widehat{\phi}_{L,\infty,\mathcal{H}}$(see Eq. (10)). Given $0 < \delta < 1$, we see we need M large enough so that

$$1 - \mathcal{N}(\mathcal{H}, \frac{\epsilon}{48S_0R^2})e^{-\frac{c_{L,N,\mathcal{H}}M\epsilon}{1152S_0^2R^2}} \geq 1 - \delta\,.$$

The conclusion follows. □

E. Trajectory-based Performance Measures. After having established results on the convergence rate of our estimator, we turn to control the accuracy of trajectories predicted when using the estimated interaction kernel, evolved from initial conditions both in and outside of the training data. Trajectory-based measurements of accuracy are interesting because (a) they provide a quantitative assessment on the quality of the approximated dynamics, (b) while the true interactions kernels are typically not known, and so the accuracy of the estimated interaction kernel may not be evaluated, trajectories are known, and may be used to perform model validation and cross-validation for parameter selection (if needed).

The next Proposition shows that the error in prediction is (i) bounded trajectory-wise by a continuous time version of the error functional, and (ii) bounded in the mean squared sense by the mean squared error of the estimated interaction kernel.

Proposition 1.7. *Let $\widehat{\phi}$ be an estimator of the true interaction kernel ϕ. Suppose that the function $\widehat{\phi}(||\cdot||)\cdot$ is Lipschitz continuous on $\mathbb{R}^d$, with Lipschitz constant C_{Lip}. Denote by $\widehat{\boldsymbol{X}}(t)$ and $\boldsymbol{X}(t)$ the solutions of the systems with interaction kernels $\widehat{\phi}$ and ϕ respectively, starting from the same initial condition. Then we have*

$$\sup_{t\in[0,T]}\|\widehat{\boldsymbol{X}}(t) - \boldsymbol{X}(t)\|^2 \leq 2Te^{8T^2C_{\text{Lip}}^2}\int_0^T \left\|\dot{\boldsymbol{X}}(t) - \mathbf{f}_{\hat{\varphi}}(\boldsymbol{X}(t))\right\|^2 dt$$

for each trajectory, and on average with respect to the initial distribution μ_0,

$$\mathbb{E}_{\mu_0}[\sup_{t\in[0,T]}\|\widehat{\boldsymbol{X}}(t) - \boldsymbol{X}(t)\|^2] \leq C(T, C_{\text{Lip}})\sqrt{N}\|\widehat{\varphi}(\cdot)\cdot - \phi(\cdot)\cdot\|^2_{L^2(\rho_T)}$$

for a constant $C(T, C_{\text{Lip}})$, where the measure ρ_T is as in Eq. (4) in the main text.

2. Algorithm

We start from describing the algorithm in its simplest form, for learning first order system with homogeneous agents; we then move to first order systems with heterogeneous agents, and finish with the second order systems with heterogeneous agents.

A. First Order Homogeneous Agent Systems. Recall that we would like to estimate the interaction kernel ϕ of the N-agent system in Eq. (1) from M independent trajectories $\{\boldsymbol{x}_i^m(t_l), \dot{\boldsymbol{x}}_i^m(t_l)\}_{i=1,l=1,m=1}^{N,L,M}$ with $t_l = \frac{lT}{L}$. We obtain an estimator by minimizing the discrete empirical error functional, over all φ in a hypothesis space $\mathcal{H}_n$,

$$\mathcal{E}_{L,M}(\varphi) = \frac{1}{LMN} \sum_{l,m,i=1}^{L,M,N} \left\| \dot{\boldsymbol{x}}_i^m(t_l) - \sum_{i'=1}^{N} \frac{1}{N} \varphi(r_{i,i'}^m(t_l)) \boldsymbol{r}_{i,i'}^m(t_l) \right\|^2 . \qquad [14]$$

When only the positions can be observed, we assume that T/L is sufficiently small so that we can accurately approximate the velocity $\dot{\boldsymbol{x}}_i^m(t_l)$ by finite differences, for example

$$\dot{\boldsymbol{x}}_i^m(t_l) \approx \Delta \boldsymbol{x}_i^m(t_l) = \frac{\boldsymbol{x}_i^m(t_l) - \boldsymbol{x}_i^m(t_{l-1})}{t_l - t_{l-1}}, \quad \text{for } 1 \leq l \leq L,$$

where we assumed t_0 is also observed. The error of the backward difference approximation is of order $O(T/L)$, leading to a $O(T/L)$ bias in the estimator. Therefore, for simplicity, we assume in the theoretical discussion that follows that the velocity $\dot{\boldsymbol{x}}_i^m(t_l)$ is observed.

First, we set the hypothesis space $\mathcal{H}_n$ to be the span of $\{\psi_p\}_{p=1}^n$, a set of linearly independent functions on $[0, R]$. It is natural to use an orthonormal basis of $\mathcal{H}_n$ in $L^2(\rho_L^T)$ for efficient computations. If the true interaction kernel is known to be uniformly smooth, a global basis (e.g. Fourier) may be used. Since our admissible set is in $W^{1,\infty}$, we shall use a local basis consisting of piecewise polynomial functions on a partition of increasingly finer intervals. The partitions will be on the interval $[R_{min}, R_{max}]$, where R_{min} and R_{max} are minimal and maximal values of r such that the empirical density $\rho_{L,M}^T(r)$ of the pairwise distances $\{r_{i,i'}^m(t_l)\}$ is greater than a threshold.

Next, we minimize the empirical error functional over $\mathcal{H}_n$ to obtain an estimator. To simplify notation, for each m, we denote

$$\mathbf{d}^m := \left(\dot{\boldsymbol{x}}_1^m(t_2), \ldots, \dot{\boldsymbol{x}}_N^m(t_2); \ldots; \dot{\boldsymbol{x}}_1^m(t_L) \ldots \dot{\boldsymbol{x}}_N^m(t_L)\right) \qquad [15]$$

a column vector in $\mathbb{R}^{LNd}$; and denote

$$\Psi_L^m(li, p) := \sum_{i'=1}^{N} \frac{1}{N} \psi_p(r_{i,i'}^m(t_l)) \boldsymbol{r}_{i,i'}^m(t_l) \in \mathbb{R}^d \,,$$

for $1 \leq l \leq L$, $1 \leq i \leq N$ and $1 \leq p \leq n$, and refer it as the learning matrix Ψ_L^m. Here and in what follows, the index li denotes, with some abuse of notation, the double-index (l, i) mapped (in any fixed way) bijectively onto a one-dimensional array. Then we can rewrite the empirical error functional as

$$\mathcal{E}_{L,M}(\varphi) = \mathcal{E}_{L,M}(\mathbf{a}) = \frac{1}{LNM} \sum_{m=1}^{M} \|\mathbf{d}^m - \Psi_L^m \mathbf{a}\|_{\mathbb{R}^{LNd}}^2 \,.$$

Our estimator is the minimizer of $\mathcal{E}_{L,M}(\mathbf{a})$ over $\mathbb{R}^n$. This is a Least Squares problem, and we solve for the minimizer from the normal equations

$$\underbrace{\frac{1}{M} \sum_{m=1}^{M} A_L^m}_{A_{L,M}} \mathbf{a} = \frac{1}{M} \sum_{m=1}^{M} b_L^m \,, \qquad [16]$$

where the trajectory-wise regression matrices are

$$A_L^m := \frac{1}{LN} (\Psi_L^m)^T \Psi_L^m \quad \text{and} \quad b_L^m := \frac{1}{LN} (\Psi_L^m)^T \mathbf{d}^m.$$

We emphasize that the above regression is ready to be computed in parallel: we can compute simultaneously the matrices A_L^m and b_L^m for different trajectories. The size of the matrices A_L^m is $n \times n$, and there is no need to read and store all the data at once, thereby dramatically reducing memory usage.

B. Well-conditioning from coercivity. We show next that the coercivity condition implies that $A_{L,M}$ is well-conditioned and positive definite for large M. More specifically, the coercivity constant provides a lower bound on the smallest singular value of $A_{L,M}$, provided the basis for the hypothesis space is well-conditioned (e.g. orthonormal), therefore enabling control of the condition number of the regularized problem.

Recall the bilinear functional $\langle\langle\cdot,\cdot\rangle\rangle$ defined in Eq. (8).

Proposition 2.1. *Assume that the coercivity condition holds on $\mathcal{H}_n \subset L^\infty([0,R])$ with $c_{L,N,\mathcal{H}} > 0$. Let $\{\psi_1, \cdots, \psi_n\}$ be a basis of $\mathcal{H}_n$ such that*

$$\langle \psi_p(\cdot)\cdot, \psi_{p'}(\cdot)\cdot\rangle_{L^2(\rho_T^L)} = \delta_{p,p'}, \|\psi_p\|_\infty \le S_0 \qquad [17]$$

and $A_{L,\infty} = \left(\langle\langle \psi_p, \psi_{p'}\rangle\rangle\right)_{p,p'} \in \mathbb{R}^{n\times n}$. Then the smallest singular value of $A_{L,\infty}$ satisfies

$$\sigma_{\min}(A_{L,\infty}) \ge c_{L,N,\mathcal{H}}\,.$$

Moreover, $A_{L,\infty}$ is the a.s. limit of $A_{L,M}$ in Eq. (16)*. Therefore, for large M, the smallest singular value of $A_{L,M}$ satisfies*

$$\sigma_{\min}(A_{L,M}) \ge 0.9 c_{L,N,\mathcal{H}}$$

with probability at least $1 - 2n\exp(-\frac{c_{L,N,\mathcal{H}}^2 M}{200n^2c_1^2 + \frac{10c_{L,N,\mathcal{H}}c_1}{3}n})$, where $c_1 = R^2S_0^2 + 1$.

Proof. For each $\mathbf{a} \in \mathbb{R}^n$,

$$\mathbf{a}^T A_{L,\infty}\mathbf{a} = \langle\langle \sum_{p=1}^n a_p\psi_p, \sum_{p=1}^n a_p\psi_p\rangle\rangle \ge c_{L,N,\mathcal{H}} \Big\| \sum_{p=1}^n a_p\psi_p(\cdot)\cdot \Big\|^2_{L^2(\rho_T^L)} = c_{L,N,\mathcal{H}}\|\mathbf{a}\|^2\,.$$

This proves the desired bound on the smallest singular value.

Going back to the case of finite M: by the law of large numbers, the matrix $A_{L,M} = \sum_{m=1}^M A_L^m$ converges to $A_{L,\infty} = \mathbb{E}[A_L^m]$ as $M \to \infty$. Hence if the sample size M is large enough, then we apply the matrix Bernstein inequality to get the probability estimates for the event that $\sigma_{min}(A_{L,M})$ is bounded below by $0.9c_{L,N,\mathcal{H}}$. □

Remark 2.2. *Proposition 2.1 highlights the importance of choosing basis functions to be linearly independent in $L^2(\rho_T^L)$ instead of in $L^\infty([0,R])$ for the hypothesis space $\mathcal{H}_n$ (orthonormality can be easily obtained through Gram-Schmidt orthogonalization if the functions are linearly independent). To see this, consider a set of basis functions consisting of piecewise polynomials that are supported on a partition of the interval $[0,R]$. These functions are linearly independent in $L^\infty([0,R])$, but can be linearly dependent in $L^2(\rho_T^L)$ if some of the partitioned intervals have zero probability under the measure ρ_T^L. This would lead to an ill-conditioned normal matrix $A_{L,\infty}$. This issue can deteriorate in practice when the unknown ρ_T^L is replaced by the empirical measure $\rho_T^{L,M}$. In this work we use piecewise polynomials on a partition of the support of $\rho_T^{L,M}$, which are orthogonal in $L^2(\rho_T^{L,M})$.*

C. First Order Heterogeneous Agent Systems. For these systems the empirical error to be minimized is as in (9) in the main text:

$$\frac{1}{LM}\sum_{l,m,i=1}^{L,M,N}\frac{1}{N_{k_i}}\left\|\dot{\boldsymbol{x}}_i^m(t_l) - \sum_{i'=1}^N \frac{1}{N_{k_{i'}}}\varphi_{k_ik_{i'}}(r_{i,i'}^m(t_l))\boldsymbol{r}_{i,i'}^m(t_l)\right\|^2,$$

over all possible $\boldsymbol{\varphi} = \{\varphi_{kk'}\}_{k,k'=1}^K \in \mathcal{H}$. Here $\boldsymbol{r}_{i,i'}(t_l)$ and $r_{i,i'}(t_l)$ are as in Eq. (14). When given observation data, $\{\boldsymbol{x}_i^m(t_l)\}_{i=1,m=1,l=1}^{N,M,L}$, but no derivative information, we approximate the derivatives using backward differencing scheme for $1 \le l \le L$ (assuming observations at t_0); in either case we assemble the derivative vector $\mathbf{d}$ similarly to Eq. (15), but with the normalization

$$\mathbf{d}^m(li) = N_{k_i}^{-1/2}\Delta\boldsymbol{x}_i^m(t_l) \in \mathbb{R}^d.$$

Proceeding analogously to the homogeneous agent case, we search for $\varphi_{kk'}$ in a $n_{kk'}$-dimensional hypothesis space $\mathcal{H}_{n_{kk'}}$, with basis $\{\psi_{kk',p}\}_{p=1}^{n_{kk'}}$, and write $\varphi_{kk'}(r) = \sum_{k,k'=1}^K\sum_{p=1}^{n_{kk'}} a_{kk',p}\psi_{kk',p}(r)$ for some vector of coefficients $(a_{kk',p})_{p=1}^{n_{kk'}}$. For the learning matrix Ψ_L^m, we will divide the columns into K^2 regions, each region indexed by the pair (k,k'), with $k,k' = 1,\cdots,K$. We adopt the usual lexicographic partial ordering on these pairs. The columns of Ψ_L^m corresponding to (k,k') are given by

$$\Psi_L^m(li, \tilde{n}_{kk'} + p) = N_{k_i}^{-1/2}\sum_{i'\in C_{k'}}\frac{1}{N_{k'}}\psi_{kk',p}(r_{i,i'}^m(t_l))\boldsymbol{r}_{i,i'}^m(t_l) \in \mathbb{R}^d,$$

for $i \in C_k$ and $1 \le l \le L$, and $\tilde{n}_{kk'} = \sum_{(k_1,k_1')<(k,k')} n_{k_1k_1'}$. We define

$$\mathbf{a} = \left(a_{11,1},\ldots,a_{11,n_{11}};\ldots;a_{KK,1},\ldots,a_{KK,n_{KK}}\right) \in \mathbb{R}^{d_0}$$

with $d_0 = \sum_{k,k'=1}^K n_{k,k'}$, to arrive at Eq. (16)

D. Second Order Heterogeneous Agent Systems. The learning problems of inferring the interactions of the $\dot{\boldsymbol{x}}_i$'s and ξ_i's can be de-coupled. We start with the inference of the interactions on $\dot{\boldsymbol{x}}_i$'s. Let the observations of the second order heterogeneous agent system be $\{\boldsymbol{x}_i^m(t_l), \dot{\boldsymbol{x}}_i^m(t_l), \xi_i^m(t_l)\}_{l,i,m=1}^{L,N,M}$. Let $\boldsymbol{v}_i^m = \dot{\boldsymbol{x}}_i^m$. As usual, if velocities and/or accelerations are not observed, they are approximated by a finite-difference (in time) scheme, for example

$$\Delta\boldsymbol{v}_i^m(t_l) = \frac{\boldsymbol{v}_i^m(t_l) - \boldsymbol{v}_i^m(t_{l-1})}{t_l - t_{l-1}} \quad , \quad \Delta\xi_i^m(t_l) = \frac{\xi_i^m(t_l) - \xi_i^m(t_{l-1})}{t_l - t_{l-1}},$$

for $1 \leq l \leq L$ and $1 \leq i \leq N$ (assuming observations also at t_0). For the data corresponding to the m^{th} initial condition, we assemble the external influence (from interaction with the environment) vector $\vec{F}^{m,\boldsymbol{v}}$ as:

$$\vec{F}^{m,\boldsymbol{v}}(li) = N_{k_i}^{-1/2} F^{\boldsymbol{v}}(\boldsymbol{v}_i^m(t_l), \xi_i^m(t_l)) \in \mathbb{R}^d,$$

and the approximated derivative of $\boldsymbol{v}_i$'s as

$$\mathbf{d}^{m,\boldsymbol{v}}(li) = N_{k_i}^{-1/2} m_i \Delta\boldsymbol{v}_i^m(t_l) \in \mathbb{R}^d.$$

We use a finite dimensional subspace $\mathcal{H}^E_{n^E}$, so that the candidate functions $\boldsymbol{\varphi}^E = \{\varphi^E_{kk'}\}_{k,k'=1}^K$ are expressed as $\boldsymbol{\varphi}^E(r) = \sum_{k,k'=1}^K \sum_{p=1}^{n^E_{k,k'}} \alpha^E_{kk',p}\psi^E_{kk',p}(r)$. Using the same ordering from previous discussion on the first order heterogeneous agent system, we have, for a pair (k,k') learning matrix $\Psi^{m,E}_{L,M}$ for the energy-based interaction kernel,

$$\Psi^{m,E}_{L,M}(li, \tilde{n}^E + p) = N_{k_i}^{-1/2} \sum_{i' \in C_{k'}} \frac{1}{N_{k'}} \psi^E_{kk',p}(r^m_{i,i'}(t_l))\boldsymbol{r}^m_{i,i'}(t_l),$$

for $1 \leq l \leq L$, $i \in C_k$ and $\tilde{n}^E = \sum_{(k_1,k_1')<(k,k')} n^E_{k_1k_1'}$. The construction of the alignment-based learning matrix $\Psi^{m,A}_{L,M}$ is analogous:

$$\Psi^{m,A}_{L,M}(li, \tilde{n}^A + p) = N_{k_i}^{-1/2} \sum_{i' \in C_{k'}} \frac{1}{N_{k'}} \psi^A_{kk',p}(r^m_{i,i'}(t_l))\boldsymbol{r}^m_{i,i'}(t_l),$$

for $1 \leq l \leq L$, $i \in C_k$ and $\tilde{n}^A = \sum_{(k_1,k_1')<(k,k')} n^A_{k_1k_1'}$. We put all the α's together into $\mathbf{a}^E$ and $\mathbf{a}^A$, and further grouping them into one big vector, $\mathbf{a}^{\boldsymbol{v}} = \begin{pmatrix}\mathbf{a}^E \\ \mathbf{a}^A\end{pmatrix}$ and $\Psi^{m,\boldsymbol{v}}_{L,M} = \left(\Psi^{m,E}_{L,M}, \Psi^{m,A}_{L,M}\right)$, we arrive at the final formulation,

$$\frac{1}{LM}\sum_{m=1}^M \left\| \mathbf{d}^{m,\boldsymbol{v}} - \vec{F}^{m,\boldsymbol{v}} - \Psi^{m,\boldsymbol{v}}_{L,M}\mathbf{a}^{\boldsymbol{v}} \right\|^2_{\mathbb{R}^{LNd}}.$$

As usual, we solve the associated normal equations of Eq. (16) with $A^m_L := (\Psi^{m,\boldsymbol{v}}_{L,M})^\top \Psi^{m,\boldsymbol{v}}_{L,M}$ and $b^m_L := (\Psi^{m,\boldsymbol{v}}_{L,M})^\top(\mathbf{d}^{m,\boldsymbol{v}} - \vec{F}^{m,\boldsymbol{v}})$, reducing the system size from $(MLNd) \times (n^E + n^A)$ to $(n^E + n^A)^2$.

For the inference of the interactions on ξ_i's, we let

$$\vec{F}^{m,\xi}(li) = N_{k_i}^{-1/2} F^\xi(\xi_i^m(t_l)) \quad \text{and} \quad \mathbf{d}^{m,\xi}(li) = N_{k_i}^{-1/2}\Delta\xi_i^m(t_l),$$

for $1 \leq l \leq L$ and $1 \leq i \leq N$; then the learning matrix $\Psi^{m,\xi}_{L,M}$ is assembled similarly as

$$\Psi^{m,\xi}_{L,M}(li, \tilde{n}^\xi + p) = N_{k_i}^{-1/2} \sum_{i' \in C_{k'}} \frac{1}{N_{k'}} \psi^\xi_{kk',p}(r^m_{i,i'}(t_l))\boldsymbol{r}^m_{i,i'}(t_l),$$

for $1 \leq l \leq L$, $i \in C_k$, and $\tilde{n}^\xi = \sum_{(k_1k_1')<(k,k')} n^\xi_{k_1,k_1'}$. We then arrive at the Least Squares problem

$$\frac{1}{LM}\sum_{m=1}^M \left\| \mathbf{d}^{m,\xi} - \vec{F}^{m,\xi} - \Psi^{m,\xi}_{L,M}\mathbf{a}^\xi \right\|^2_{\mathbb{R}^{LNd}}$$

and solve it from the associated normal equations.

E. The Final Algorithm. Given observation data, $\{\boldsymbol{x}_i^m(t_l)$ and $\dot{\boldsymbol{x}}_i^m(t_l)$ and/or $\xi_i^m(t_l)\}_{l,i,m=1}^{L,N,M}$, we use the Algorithm 1 to find the estimators for the interaction kernels.

Algorithm 1 Learning Interaction Kernels from Observations

```
Input: $\{\boldsymbol{x}_i^m(t_l)$ and/or $\dot{\boldsymbol{x}}_i^m(t_l)$ and/or $\xi_i^m(t_l)\}_{l,i,m=1}^{L,N,M}$.
Output: estimators for the interaction kernels.
if First Order System then
    Find out the maximum interaction radii $R_{kk'}$'s.
    Construct the basis, $\psi_{kk',p}$'s.
    Assemble the normal equations (16) (in parallel) and solve for $\mathbf{a}$.
    Assemble $\widehat{\boldsymbol{\phi}}(r) = \sum_{k,k'=1}^{K}\sum_{p=1}^{n_{kk'}} a_{kk',p}\psi_{kk',p}(r)$.
else if Second Order System then
    Find out the maximum interaction radii $R_{kk'}$'s.
    Construct the basis, $\psi_{kk',p}^E$'s and $\psi_{kk',p}^A$'s.
    Assemble the normal equations (16) (in parallel), solve for $\mathbf{a}^{\boldsymbol{v}}$, and partition it to $\mathbf{a}^E$ and $\mathbf{a}^A$.
    Assemble $\widehat{\boldsymbol{\phi}}(r)^E = \sum_{k,k'=1}^{K}\sum_{p=1}^{n_{kk'}^E} a_{kk',p}^E\psi_{kk',p}^E(r)$ and $\widehat{\boldsymbol{\phi}}(r)^A = \sum_{k,k'=1}^{K}\sum_{p=1}^{n_{kk'}^A} a_{kk',p}^A\psi_{kk',p}^A(r)$.
    if If there are $\xi_i$'s then
        Construct the basis, $\psi_{kk',p}^\xi$'s.
        Assemble the normal equations and solve for $\mathbf{a}^\xi$.
        Assemble $\widehat{\boldsymbol{\phi}}(r)^\xi = \sum_{k,k'=1}^{K}\sum_{p=1}^{n_{kk'}^\xi} a_{kk',p}^\xi\psi_{kk',p}^\xi(r)$.
    end if.
end if.
```

F. Computational Complexity. The computational complexity is driven by the construction and solution of the least squares problem in Algorithm 1. Though the observation data $\{\boldsymbol{x}_i^m(t_l), \dot{\boldsymbol{x}}_i^m(t_l), \xi_i^m(t_l)\}_{l,i,m=1}^{L,N,M}$ requires an array of size $MLN(2d+1)$, the linear system to be solved, i.e. the system consisting of normal equations, is only of size $n^E + n^A$; in the case of choosing the optimal basis, n^E and n^A behave like $\mathcal{O}(M^{\frac{1}{2s+1}})$. When the system of the normal equations is ill-conditioned or ill-posed, a truncated singular value decomposition will be used, which performs a singular value decomposition of the matrix $A_{L,M}$, and keeps those singular values which are above a (preset) threshold, then assemble an approximated matrix with the truncated singular value matrices.

Furthermore, since the M trajectories are independent, we can construct $\Psi^{m,E}$ and other related quantities for each trajectory at a time (which can be done in a parallel environment with two communication needed, one to send/receive the maximum interaction radii's, and the other to send/receive A_L^m and b_L^m in the normal equations after they are built on the master core), each requires a total memory of $LNd(n^E + n^A) + LNd + LNd$, which is $\mathcal{O}(LNd)$, since $n^E + n^A \ll LNd$.

The computing time of the algorithm depends heavily on the time to assemble normal equations from M trajectories, which is $\mathcal{O}((n^E + n^A)^2 LN^2)$; solving the final linear system requires time $\mathcal{O}((n^E + n^A)^3) = \mathcal{O}(M^{\frac{3}{2s+1}})$ in the worst case, for example when using a highly stable truncated singular value decomposition solver.

Therefore, the algorithm is effective at inferring the interactions from a wide variety of systems; the results will be discussed in the next section.

3. Examples

We consider here four important examples of self-organized dynamics: the opinion dynamics, the particle system with the Lennard-Jones potential, the predator-swarm system and the phototaxis dynamics. We describe here in detail how the numerical simulations are set up for each of these examples. In all but the Lennard-Jones system, we set up the experiments using the parameters as shown in Table S1. We consider the regime with a rather small number of observations in terms of both M and L to emphasize that our technique can achieve good results even when a relatively small number of samples is given.

Table S1

N	# Trials	$M_{\rho_T^L}$	$[0, T_f]$
10	10	2000	$[0, cT]$

Parameters used in all the examples but the Lennard-Jones system. Here the observation time T is system-specific. $c = 2$ in all examples unless otherwise specified.

We use a large number $M_{\rho_T^L}$ (in particular, $M_{\rho_T^L} \gg M$) of independent trajectories (not to be used elsewhere) to obtain an accurate approximation of the unknown probability measure ρ_T^L in (4) in the main text. In what follows, to keep the notation from becoming cumbersome, we denote by ρ_T^L this empirical approximation to ρ_T^L. We run the dynamics over the time $[0, T_f]$

with M different initial conditions (drawn from the dynamics-specific probability measure μ_0), and the observations consist of the state vector, with no derivative information, at L equidistant time samples in the time interval $[0, T]$. We report the relative (i.e. normalized by the norm of the true interaction kernel) error of our estimators in the $L^2(\rho_T^L)$ norm. In the spirit of Proposition (3.4) in the main text, we also report on the error on trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ generated by the system with the true interaction kernel and with the learned interaction kernel, on both the "training" time interval $[0, T]$ and on a "prediction" time interval $[T, T_f]$ ($T_f = 2T$ unless otherwise specified), with both the same initial conditions as those used for training, and on new initial conditions (sampled according to the specified measure μ_0). The trajectory error will be estimated using M trajectories (we report mean and standard deviation of the error). We run a total of 10 independent learning trials and compute the mean and standard deviation of the corresponding estimators, their errors, and the trajectory errors just discussed. Since each learning trial generates different mean and standard deviation of the trajectory errors over different Initial Conditions (ICs), we also report the mean and standard deviation over the 10 learning trials for mean$_{IC}$ and std$_{IC}$.

All ODE systems are evolved using **ode15s** in MATLAB® with a relative tolerance at 10^{-5} and absolute tolerance at 10^{-6}. We choose the finite-dimensional hypothesis space $\mathcal{H}_n$ (with n chosen differently in each example, based on sample size) as the span of either piecewise constant or piecewise linear functions on n intervals forming a uniform partition of $[0, R_{k,k'}]$, where $R_{k,k'}$ is the maximum observed pairwise distance between agents of type k' and agents in type k for $t \in [0, T]$.

Learning results are showcased in Fig. 5 in the main text . The first one compares the learned interaction kernel(s) to the true interaction kernel(s) (with mean and standard deviation over the total number of learning trials) with the background showing the comparison of ρ_T^L (computed on $M_{\rho_T^L}$ trajectories, as described above) and $\rho_T^{L,M}$ (generated from the observed data consisting of M trajectories). The second plot compares the true trajectories (evolved using the true interaction law(s)) and learned trajectories (evolved using the learned interaction law(s)) over two different sets of initial conditions – one taken from the training data, and one new, randomly generated from μ_0. The third plot compares the true trajectories and the trajectories generated with the estimated interaction kernel, but for a different system with number of agents $N_{\text{new}} = 4N$, again over two different sets of randomly chosen initial conditions. Measurements of performance are also shown alongside the figures: ($L^2(\rho_T^L)$ errors, trajectory errors, etc. Let $\boldsymbol{X}(t)$ and $\hat{\boldsymbol{X}}(t)$ be two sets of continuous-time trajectories; the max-in-time error is defined as

$$\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}\right\|_{\mathrm{TM([0,T])}} = \sup_{t\in[0,T]} \left\|\boldsymbol{X}(t) - \hat{\boldsymbol{X}}(t)\right\|_{\mathcal{S}} . \qquad [18]$$

For second order systems with the auxiliary environment variable ξ_i's, we are also interested in the trajectories of ξ_i, for which we may use $\left\|\Xi - \hat{\Xi}\right\|_{\mathrm{TM([0,T])}} = \sup_{t\in[0,T]} \left\|\Xi(t) - \hat{\Xi}(t)\right\|_{\mathcal{S}}$.

Finally, for each example we consider adding noise to the observations: in the case of additive noise the observations are $\{(\boldsymbol{X}^m(t_l) + \eta_{1,l,m}, \dot{\boldsymbol{X}}^m(t_l)) + \eta_{2,l,m}\}_{l=1,m=1}^{L,M}$, while in the case of multiplicative noise they are $\{(\boldsymbol{X}^m(t_l) \cdot (1 + \eta_{1,l,m}), \dot{\boldsymbol{X}}^m(t_l)) \cdot (1 + \eta_{2,l,m})\}_{l=1,m=1}^{L,M}$, where in both cases $\eta_{1,l,m}$ and $\eta_{2,l,m}$ are i.i.d. samples from a distribution modeling noise, which we will pick to be Unif.($[-\sigma, \sigma]$). Note that in both these cases velocities are part of our observations, since with noise added in the position the inference of velocities becomes problematic due to the amplification of the noise that a simple finite difference scheme would incur.

Finally, for several examples we also report the behavior of the relative error of the estimator as a function of the number of samples L in time and of the number of trajectories M. We observe the decrease in error as L increases, which is expected but is not captured by the estimate in Thm. (3.3) in the main text. These plots are qualitatively the same for all the experiments.

We devote the next sections to the various examples, discussing setups particular to each example and corresponding results.

A. Opinion Dynamics. Modeling using self-organized dynamics has seen successful applications in studying and analyzing how the opinions of people influence each other and how consensus is formed based on different kinds of influence functions. We refer to these systems as opinion dynamics. We consider the first order model in Eq. (1), and the interaction kernel defined as

$$\phi(r) = \begin{cases} 1, & 0 \le r < \frac{1}{\sqrt{2}}, \\ 0.1, & \frac{1}{\sqrt{2}} \le r < 1, \\ 0, & 1 \le r. \end{cases}$$

In this context $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is sometimes referred to as the scaled influence function, modeling the change of each agents' opinion by relative differences in the opinions of the other agents. Here $\boldsymbol{x}_i \in \mathbb{R}^d$ is the vector opinions of agent i. Here $\|\cdot\|$ can be taken as the normal Euclidean norm, but other metrics depending on the problem at hand may be used as well, with no changes in our definitions and constructions. The time-discretization of this system is referred to as the classical Krause model for opinion dynamics. With the specific ϕ above, there is only attraction present in the system, the opinions of the agents merge into clusters, with the number of clusters significantly smaller than the number of agents. This clustering behavior severely reduces the amount of effective samples of pairwise distance observable at large times. We consider the system and test parameters given in Table S2.

Table S2

d	M	L	T	μ_0	n	deg(ψ)
1	50	200	10	$\mathcal{U}([0,10]^2)$	200	0

(OD) Parameters for the system

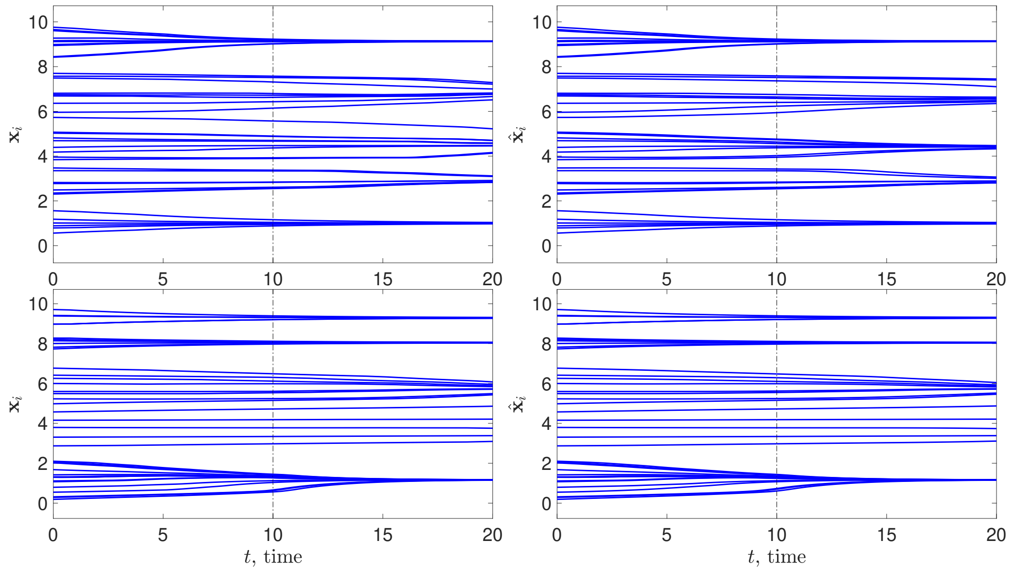


Fig. S1. (OD) Trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ obtained with ϕ and $\hat{\phi}$ respectively, for dynamics with larger $N_{\text{new}} = 4N$, over two different sets of initial conditions. We are able to accurately predict the clusters (number and location). Errors are reported in Table S3.

Table S3

	$[0, T]$	$[T, T_f]$
mean$_{\text{IC}}$: Training ICs	$3.5 \cdot 10^{-2} \pm 8.1 \cdot 10^{-3}$	$4.8 \cdot 10^{-2} \pm 1.4 \cdot 10^{-2}$
std$_{\text{IC}}$: Training ICs	$5.2 \cdot 10^{-2} \pm 1.3 \cdot 10^{-2}$	$7.6 \cdot 10^{-2} \pm 2.7 \cdot 10^{-2}$
mean$_{\text{IC}}$: Random ICs	$3.2 \cdot 10^{-2} \pm 7.4 \cdot 10^{-3}$	$4.6 \cdot 10^{-2} \pm 1.2 \cdot 10^{-2}$
std$_{\text{IC}}$: Random ICs	$5.0 \cdot 10^{-2} \pm 1.7 \cdot 10^{-2}$	$7.2 \cdot 10^{-2} \pm 2.7 \cdot 10^{-2}$
mean$_{\text{IC}}$: Larger N	$3.1 \cdot 10^{-2} \pm 2.0 \cdot 10^{-3}$	$7.3 \cdot 10^{-2} \pm 4.1 \cdot 10^{-3}$
std$_{\text{IC}}$: Larger N	$2.1 \cdot 10^{-2} \pm 2.1 \cdot 10^{-3}$	$6.1 \cdot 10^{-2} \pm 4.2 \cdot 10^{-3}$

(OD) Trajectory Errors: ICs used in the training set (first two rows), new IC"s randomly drawn from μ_0 (second set of two rows), for ICs randomly drawn for a system with $4N$ agents (last two rows). Means and std's are over 10 learning runs.

Fig.S1 shows the comparison between the estimated interaction kernel $\hat{\phi}$ (as the mean over learning trials) and the true one, ϕ. We obtain a faithful approximation of the true interaction kernel, including near the discontinuity and the compact support. Our estimator also performs well near 0, notwithstanding that information of $\phi(0)$ is lost due to the structure of the equations, that have terms of the form $\phi(0)\vec{0} = \vec{0}$. The same figure also compares the trajectories generated by the system governed by ϕ and that governed by $\hat{\phi}$. Table S3 reports the max-in-time error for those trajectories. We also test the robustness to noise, by adding noise to the observations of both positions and velocities, as described above: the estimated kernel is shown in Figure S2. Figure S3 shows the behavior of the error of the estimator as both L and M are increased.

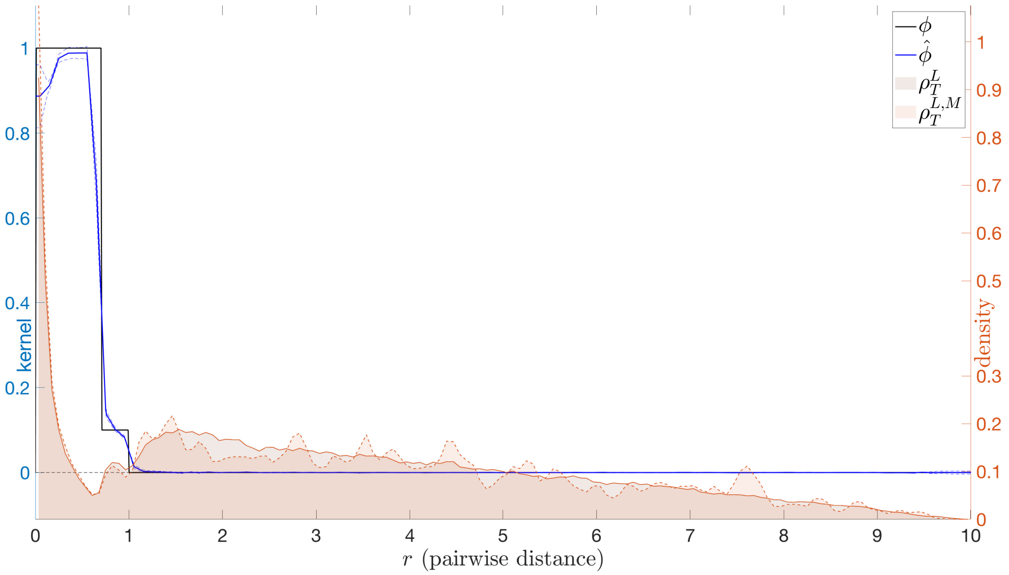


Fig. S2. (OD) Interaction kernel learned with Unif.$([-\sigma, \sigma])$ additive noise, for $\sigma = 0.1$ in the observed positions and velocities. The estimated kernels are minimally affected, mostly in regions with small ρ_T^L and near 0.

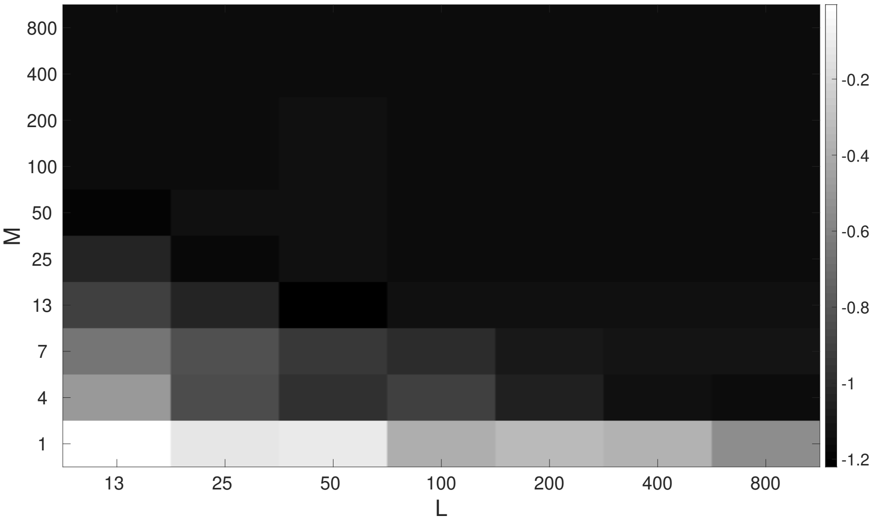
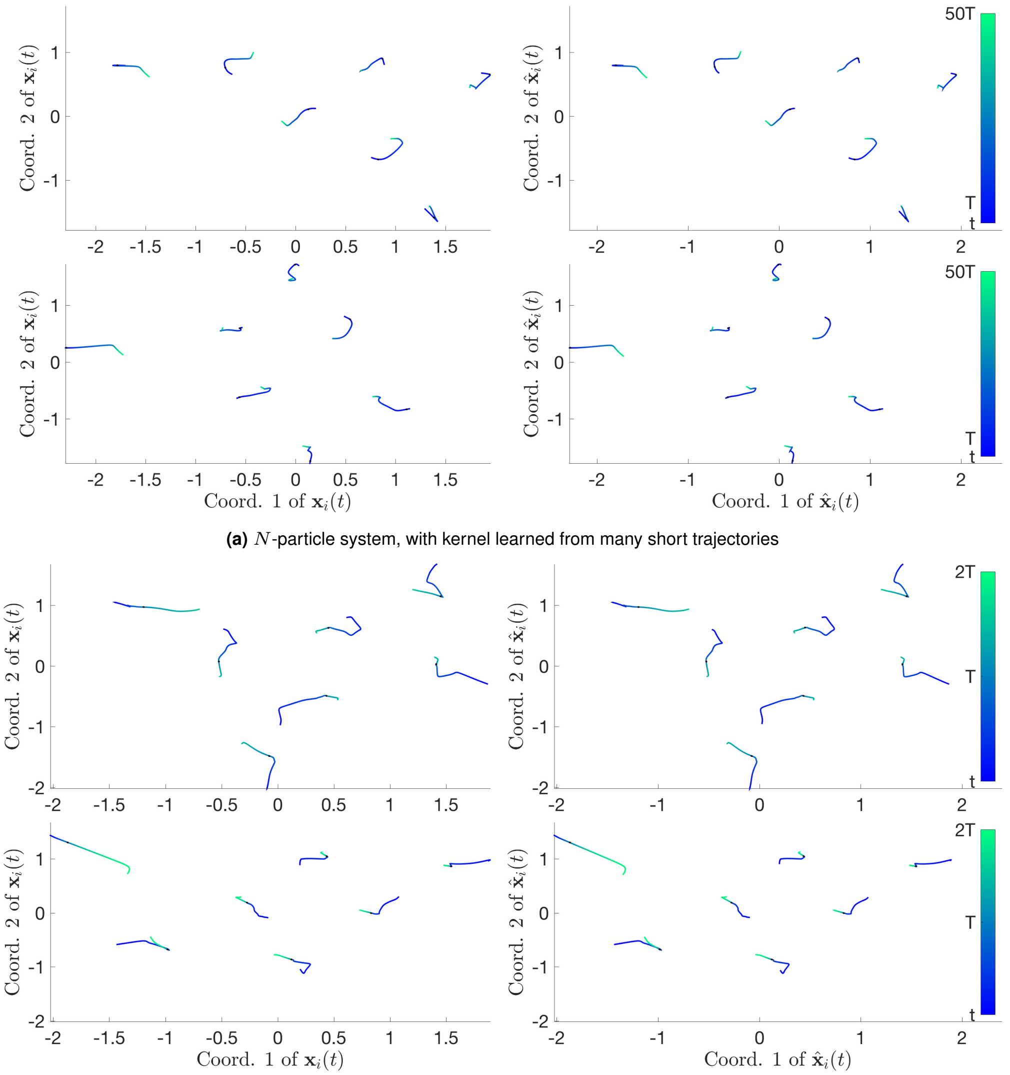


Fig. S3. (OD) Relative error, in $\log_{10}$ scale, of $\hat{\phi}$ as a function of L and M. The error decreases both in L and M, in fact roughly in the product ML, at least when M and L are not too small. $M = 1$ does not seem to suffice, no matter how large L is, due to the limited amount of "information" contained in a single trajectory.



(a) N-particle system, with kernel learned from many short trajectories

(b) N-particle system, with kernel learned from a few long trajectories

Fig. S4. (LJ) (a) and (b)presents trajectories $\boldsymbol{X}(t)$ (left) and $\widehat{\boldsymbol{X}}(t)$ (right) obtained with ϕ and $\hat{\phi}$ respectively, for initial conditions in the training dataset (top) and randomly sampled initial conditions (bottom). The time T is as in Table S5. Trajectory errors for all cases are reported in Table S7.

B. Interacting Particles in Lennard-Jones Potential. The expression of the Lennard-Jones potential is

$$\Phi(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] = \epsilon \left[\left(\frac{r_m}{r}\right)^{12} - 2\left(\frac{r_m}{r}\right)^{6} \right]$$

where ϵ is the depth of the potential well, σ is the finite distance at which the inter-particle potential is zero, r is the distance between the particles, and r_m is the distance at which the potential reaches its minimum. At r_m, the potential function has the value $-\epsilon$. The r^{-12} term describes Pauli repulsion at short ranges due to overlapping electron orbitals, and the r^{-6} term describes attraction at long ranges (van der Waals force, or dispersion force). We set $\epsilon = 10$ and $\sigma = 1$ in our simulations.

In the experiments, whose results are represented in Fig. 1 in the main text, the distribution μ_0 for the M i.i.d. initial conditions is a standard Gaussian vector in $\mathbb{R}^{2N}$. In this Lennard-Jones interacting system, one has to be careful in choosing the observation time interval. Since the minimum distance between the particles at initial configurations is very close to 0 with high probability, the particles have very large velocities (e.g. $\sim 10^{22}$) due to the singularity of the interaction kernel at 0. This obstruction made the learning algorithm infeasible since our algorithm is for learning bounded kernels. Therefore, we chose an observation time starting from a suitable time t_0, small but positive. On the other side of the training time interval, since the system evolves to equilibrium configurations very quickly, we observe the dynamics up to a time T which is a fraction of the equilibrium time. In each sampling regime, we observe the dynamics at discrete times $\{t_i\}_{i=2,\ldots,L}$ and then use the standard finite difference method to obtain a faithful approximation of velocities of agents.

Table S4

N	d	μ_0	# Trials	$M_{\rho_T^L}$	$[t_0, T_f]$	deg($\psi_{kk'}$)
7	2	$N(0, I_{2N})$	10	2000	$[t_0, cT]$	1

(LJ) Parameters used in Lennard-Jones system

Table S5

	M	L	n	$[t_0, T]$	c
Many short traj.	200	91	600	$[0.001, 0.01]$	50
Single long traj.	20	4991	600	$[0.001, 0.5]$	2

(LJ) Observation parameters for the Lennard-Jones system

Table S4 and Table S5 summarize the parameters used for the two regimes: many short-time trajectories, and a single large-time trajectory. In the first regime, the randomness of initial conditions enables the agents to explore large regions of state space, and in the space of pairwise distance, in a short time. In the second regime, the large-time dynamics plays a fundamental role in driving the pairwise distance between agents to cover areas of interest.

Table S6

	Many short trajectories	a few long trajectories
Rel. Err. for $\hat{\phi}$	$6.6 \cdot 10^{-2} \pm 5 \cdot 10^{-3}$	$7.2 \cdot 10^{-2} \pm 1 \cdot 10^{-2}$

(LJ) Relative error of the estimator for the Lennard-Jones system

The estimator belongs to a piecewise linear function space $\mathcal{H}_n$ of dimension $n = 600$. As reported in Fig.1 of the main text, the estimated interaction kernel $\hat{\phi}$ approximates the true interaction kernel ϕ well in the regions where ρ_T^L (and ρ_T) is large, i.e. regions with an abundance of observed values of pairwise distances to reconstruct the interaction kernel. The dependency on T of ρ_T^L, and of the space $L^2(\rho_T^L)$ (see (5) in the main text) used for learning, is rather pronounced, as may be seen from the histogram visualization also in Fig. 1. As usual we also compare trajectories $\widehat{\boldsymbol{X}}(t)$ generated by the system with the estimated interaction kernel learned with trajectories $\boldsymbol{X}(t)$ generated by the original system, given the same initial conditions at t_0, both on the learning interval $[t_0, T]$ and on larger time intervals $[t_0, cT]$. Figure S4 provides a visualization of such trajectories. Visualization of the corresponding systems with a larger number of agents N_{new} can be found in Figure 1 of the main text. We report the estimation errors of the interaction kernel and the trajectory errors in Tables S6 and S7.

Table S6 shows the mean and standard deviations of the relative $L^2(\rho_T)$ errors of the kernel estimators in 10 different simulations. We report the relative errors of trajectory prediction in SI Sec.3B.

Table S7

	$[t_0, T]$	$[T, T_f]$
mean$_{\text{IC}}$: Training ICs	$1.6 \cdot 10^{-3} \pm 2 \cdot 10^{-4}$	$1.7 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$
std$_{\text{IC}}$: Training ICs	$4.6 \cdot 10^{-4} \pm 5 \cdot 10^{-5}$	$2.1 \cdot 10^{-2} \pm 4 \cdot 10^{-3}$
mean$_{\text{IC}}$: Random ICs	$1.6 \cdot 10^{-3} \pm 2 \cdot 10^{-4}$	$1.7 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$
std$_{\text{IC}}$: Random ICs	$4.5 \cdot 10^{-4} \pm 5 \cdot 10^{-5}$	$1.9 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$
mean$_{\text{IC}}$: Larger N	$6.2 \cdot 10^{-2} \pm 7 \cdot 10^{-3}$	$6.2 \cdot 10^{-2} \pm 2 \cdot 10^{-2}$
std$_{\text{IC}}$: Larger N	$8.2 \cdot 10^{-3} \pm 7 \cdot 10^{-4}$	$3.0 \cdot 10^{-2} \pm 1 \cdot 10^{-2}$
mean$_{\text{IC}}$: Training ICs	$3.4 \cdot 10^{-3} \pm 1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3} \pm 2 \cdot 10^{-3}$
std$_{\text{IC}}$: Training ICs	$2.7 \cdot 10^{-3} \pm 2 \cdot 10^{-3}$	$6.6 \cdot 10^{-3} \pm 3 \cdot 10^{-3}$
mean$_{\text{IC}}$: Random ICs	$4.1 \cdot 10^{-3} \pm 2 \cdot 10^{-3}$	$8.7 \cdot 10^{-3} \pm 8 \cdot 10^{-3}$
std$_{\text{IC}}$: Random ICs	$3.6 \cdot 10^{-3} \pm 2 \cdot 10^{-3}$	$1.5 \cdot 10^{-2} \pm 2 \cdot 10^{-2}$
mean$_{\text{IC}}$: Larger N	$7.7 \cdot 10^{-2} \pm 1 \cdot 10^{-2}$	$6.6 \cdot 10^{-2} \pm 3 \cdot 10^{-2}$
std$_{\text{IC}}$: Larger N	$1.5 \cdot 10^{-2} \pm 1 \cdot 10^{-2}$	$5.7 \cdot 10^{-2} \pm 3 \cdot 10^{-2}$

(LJ) Trajectory Errors for Many Short Trajectories Learning (top) and Single Large Time Trajectories Learning (bottom)

We also test the convergence of our estimator as $M \to \infty$: we choose the parameters for observations and learning as in Table S8. It is important that we choose the dimension n of hypothesis space to be dependent on M, as dictated by Thm. (3.3) in the main text. Also, in this experiment (and this experiment only!) we observe the true derivatives (instead of approximating them by finite differences of positions), as those would introduce a bias term that does not vanishes unless L also increased with n.

Table S8

$[t_0, T]$	L	$\log_2(M)$	n
$[0.001, 0.01]$	10	12 : 21	$64(M/\log M)^{0.2}$

(LJ) Observation parameters in the plot of convergence rate

We obtain a decay rate for for $\|\hat{\phi}(\cdot)\cdot - \phi(\cdot)\cdot\|_{L^2(\rho_T^L)}$ around $M^{-0.36}$, which is close to the theoretical optimal learning rate $M^{-0.4}$ – see Fig. 2 in the main text. We impute this (small) difference to the singularity of the Lennard-Jones interaction kernel at 0, which makes this interaction kernel not admissible in the our learning theory.

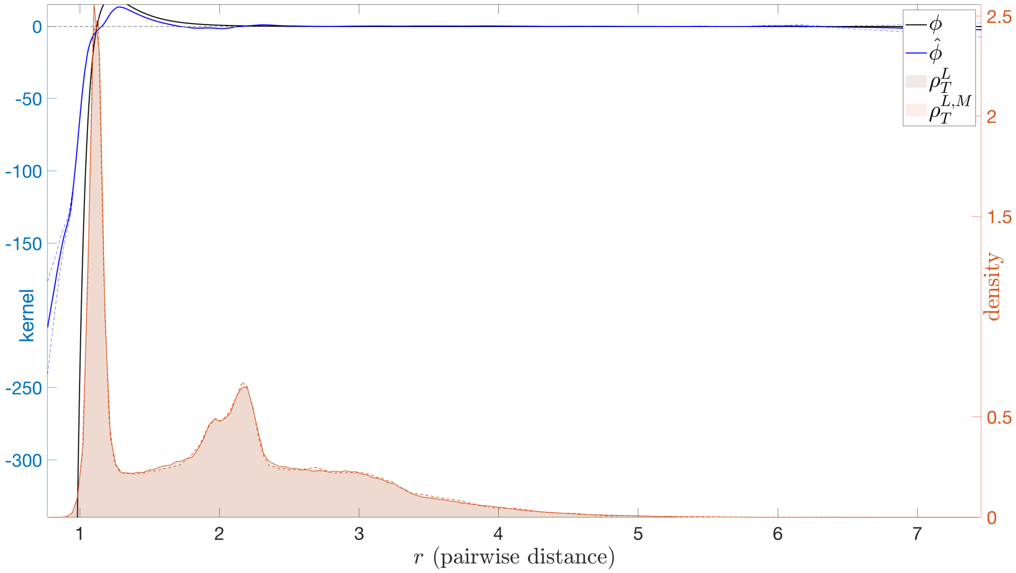


Fig. S5. (LJ) Interaction kernel learned with Unif.$([-\sigma, \sigma])$ additive noise, for $\sigma = 0.1$, in the observed positions *and observed velocities*; here $M = 500$, $L = 2000$, with all the other parameters as in Table S5.

However, the singularity of the Lennard-Jones interaction kernel at 0 forces the particles close to each other to be repel each other. Also, the system evolves rapidly to a steady-state, and the particles only explore a bounded region due to the large range attraction. Therefore, to obtain a well-supported non-degenerate measure ρ_T^L, we should make observations on a time interval that avoids reaching either the singularity of the interaction kernel or the steady-state. The restriction of the Lennard-Jones interaction kernel to the support of ρ_T^L is bounded and smooth, and hence our learning theory applies and we achieve an almost optimal rate of learning in the numerical experiments. The estimated interaction kernel with noisy observation is visualized in Figure S5.

Finally, Fig.S6 reports numerical validations of the coercivity condition in Definition 1.1 for this system. We consider the number of agents N ranging from 5 to 30, three different initial distributions μ_0, and observations on different time intervals. The coercivity constants computed by Monte Carlo sampling are close to the theoretical lower bound in all these cases.

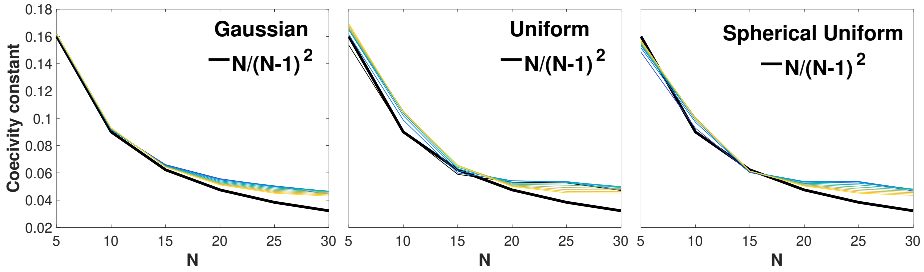


Fig. S6. (LJ) *Coercivity condition validation in 2D Lennard-Jones system with different* N. We compute the empirical coercivity constant $c_{L,N,\mathcal{H}}$ defined in Eq. (7), with $\mathcal{H}$ consisting of 200 piecewise constant basis functions with random coefficients, using $M = 131,072$ trajectories with initial conditions drawn from μ_0. Three initial distributions for μ_0 are tested: the standard Gaussian vector in $\mathbb{R}^{2N}$ (left), the uniform distribution on $[-0.5, 0.5]^{2N}$ (middle), and the uniform distributions on the unit spheres in $\mathbb{R}^{2N}$ (right). Ten different lengths of trajectories are considered (represented in each figure by the colored curves above the black curve, the theoretical lower bound of $c_{L,N,\mathcal{H}}$): each with the same initial time $t_1 = 0.001$, but the end time t_L ranges from 0.0059 to 0.0509 with a uniform time gap 10^{-4}. In all these ten sampling regimes (all are short time periods), the coercivity constant is around $\frac{N-1}{N^2}$, matching the theoretical lower bound in Thm. 3.1for one time step. We also note that $c_{L,N,\mathcal{H}}$ appears to not go to 0 as N increases, consistent with the conjecture that in rather great generality $c_{L,N,\mathcal{H}}$ stays bounded away from 0 independently of N.

C. Predator-Swarm system. There is an increasing amount of literature in discussing models of self-organized animal motion (5–15). Even more challenging is modeling interactions between agents of multiple types, in complex and emergent physical and social phenomena (11, 16–19). We consider here a representative heterogeneous agent dynamics: a Predator-Swarm system with a group of preys and a single predator, governed by either a first order or a second order system of ODE's. The intensity of interaction(s) between the single predator and group of preys can be tuned with parameters, determining dynamics with various interesting patterns (from confusing the predator with fast preys, to chase, to catch up to one prey). Since there is one single predator in the system, there is no predator-predator interaction to be learned. The interaction kernels (prey-prey, predator-prey) have both short-range repulsion to prevent the agents to collide, and long-range attraction to keep the agents in the flock. Because of the strong short-range repulsion, the pairwise distances stay bounded away from $r = 0$. We will see that these difficulties, similar to those confronted with the Lennard-Jones interaction kernel, do not prevent us from learning the interactions kernels.

In our notation for the heterogeneous system, the set C_1 corresponds to the set of preys, and C_2 to the set consisting of the single predator.

Predator-Swarm, 1^{st} order (PS1^{st}). We start from the first order system. It is a special case of the first order heterogeneous agent systems we considered, with the following interaction kernels:

$$\phi_{1,1}(r) = 1 - r^{-2}, \quad \phi_{1,2}(r) = -2r^{-2}, \quad \phi_{2,1}(r) = 3r^{1.5}, \quad \phi_{2,2}(r) \equiv 0.$$

The simulation parameters are given in Table S9.

Table S9

d	N_1	N_2	M	L	T
2	9	1	50	200	5

$n_{1,1}$	$n_{1,2} = n_{2,1}$	$n_{2,2}$	deg($\psi_{kk'}$)	Preys $\mu_0^{\boldsymbol{X}}$	Pred. $\mu_0^{\boldsymbol{X}}$
360	120	64	$[1, 1; 1, 0]$	Unif. on ring $[0.5, 1.5]$	Unif. on disk at 0.1

(PS1^{st}) System parameters for first order Predator-Swarm system

In the first column of Fig. 5 in the main text, we show the comparison of the learned interaction kernels versus the true interaction kernels (with $\rho_T^{L,kk'}$ and $\rho_T^{L,M,kk'}$ shown in the background), and the comparison of true and learned trajectories over two different sets of initial conditions.

As is shown in the top left a portion (4 sub-figures) of Fig. 5 in the main text, we are able to match faithfully all four learned interactions to their corresponding true interactions over the range of ρ_T when the pairwise distance data is abundant. We are not able to learn the interaction kernels for r close to 0, demonstrated by the larger area of uncertainty (surrounded by the dashed lines) towards 0: first, the prey-to-prey interaction is preventing preys colliding into each other; second, in the case of chasing predators, the preys are able to push away the predator. The predator-to-prey and prey-to-predator interactions are learned over the same set of pairwise distance data, however, we are able to learn the details of the two interaction kernels, and judging from the learned interaction kernels, they are not simply negative of each other. The predator-to-predator interaction simply is learned as a zero function, even though there is no pairwise distance data of a predator to a different predator. Errors in their corresponding $L^2(\rho_T^{L,kk'})$ norms are reported in Table S10.

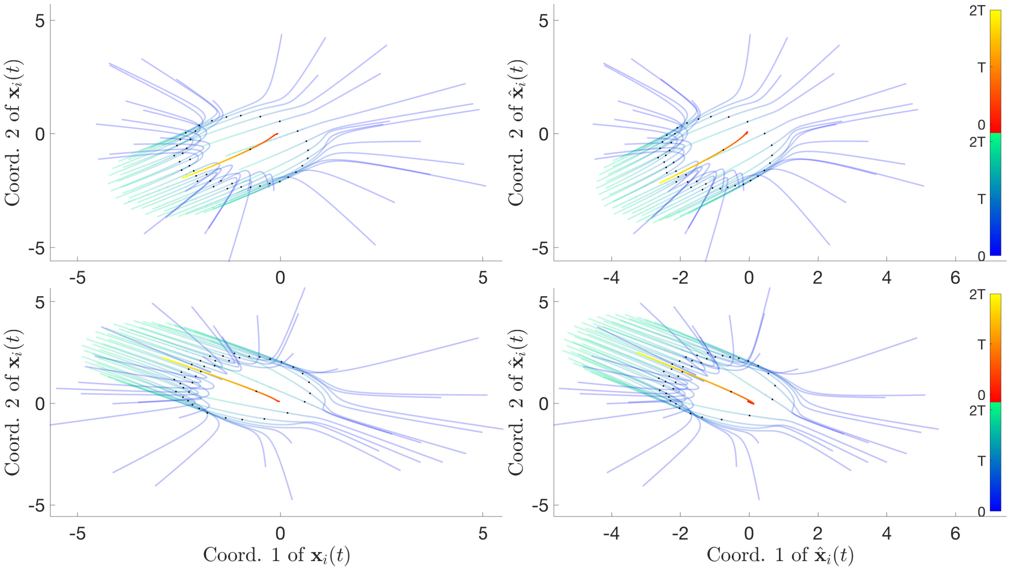


Fig. S7. (PS1^{st}) Trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ obtained with ϕ and $\hat{\phi}$ respectively, for two randomly chosen initial conditions and evolved for N_{new} agents (with the same setup as in the case of N agents). Trajectory errors are shown in Table S11.

The trajectory comparisons are shown in the bottom left portion (4 sub-figures) of Fig. 5 in the main text. We use color changing lines to indicate the movement of agents in time: with the blue-to-green lines attached to preys and the red-to-yellow line for the predator). The black dot on the trajectories indicate the position of the agents at time $t = T$, and it shows the time divide: the first half of the time, $[0, T]$, is used for learning; and the second half of the time, $[T, T_f]$, is used for prediction.

And the first row of 2 sub-figures show the comparison of the trajectories over the initial condition taken from training data, it shows (visually) no major difference between the two, except one of the prey-trajectory, is having a bigger loop in the learned trajectories. The second row of 2 sub-figures compares the trajectories from a randomly chosen initial condition (outside of the training set). We are able to predict the movement of the predator in the learned trajectories, and movement of most preys. In Fig. S7 we compare the true and predicted trajectories over a corresponding system a dynamics but with a larger number N_{new} of agents. Table S11 reports the max-in-time error Eq. (18) in the trajectories in all cases considered. We consider the effect of adding noise to observations, with results visualized in Fig. 8 of the main text.

Table S10

Rel. Err. for $\hat{\phi}_{1,1}$	$5.6 \cdot 10^{-2} \pm 1.1 \cdot 10^{-3}$
Rel. Err. for $\hat{\phi}_{1,2}$	$6.6 \cdot 10^{-3} \pm 2.4 \cdot 10^{-3}$
Rel. Err. for $\hat{\phi}_{2,1}$	$2.7 \cdot 10^{-2} \pm 8.9 \cdot 10^{-3}$
Abs. Err. for $\hat{\phi}_{2,2}$	0

(PS1^{st}) Estimator Errors

Table S11

	$[0, T]$	$[T, T_f]$
mean$_{\text{IC}}$: Training ICs	$4.2 \cdot 10^{-2} \pm 1.0 \cdot 10^{-2}$	$1.1 \cdot 10^{-1} \pm 3.0 \cdot 10^{-2}$
std$_{\text{IC}}$: Training ICs	$7.2 \cdot 10^{-2} \pm 5.6 \cdot 10^{-2}$	$1.9 \cdot 10^{-1} \pm 1.4 \cdot 10^{-1}$
mean$_{\text{IC}}$: Random ICs	$3.8 \cdot 10^{-2} \pm 1.4 \cdot 10^{-2}$	$9.5 \cdot 10^{-2} \pm 3.2 \cdot 10^{-2}$
std$_{\text{IC}}$: Random ICs	$5.5 \cdot 10^{-2} \pm 6.2 \cdot 10^{-2}$	$1.4 \cdot 10^{-1} \pm 1.4 \cdot 10^{-1}$
mean$_{\text{IC}}$: Larger N	$4.2 \cdot 10^{-1} \pm 1.7 \cdot 10^{-1}$	3.1 ± 4.6
std$_{\text{IC}}$: Larger N	$1.7 \cdot 10^{-1} \pm 9.6 \cdot 10^{-2}$	15.8 ± 27.4

(PS1^{st}) Trajectory Errors

We show numerically that our learning approach is robust to the choice of hypothesis space, as predicted by the theory, by testing on the Predator-Swarm, 1^{st}-order system with the B-splines basis. Results are shown in Fig. S8. Note that the estimators perform similarly in comparison with Fig. 8 of the main text and are consistent with the error statistics in Table S11, in both of which the hypothesis space uses piece-wise polynomial basis.

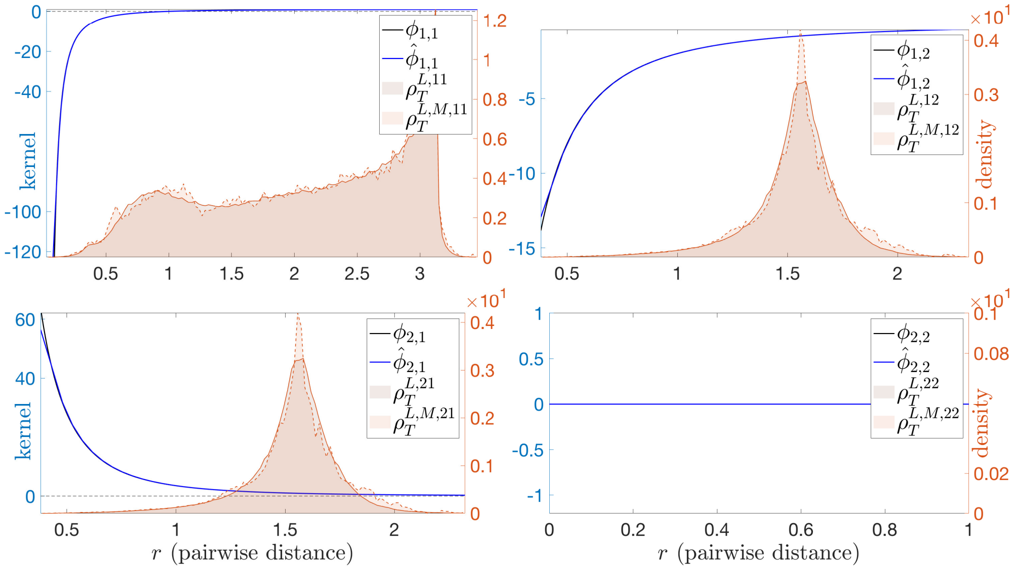


Fig. S8. (PS1st) Comparison of interaction kernels (true versus learned) when the learned kernels are generated by linear B-splines (n as in the other case considered for this system). The relative error (in $L^2(\rho_T)$ norm) for prey on prey interaction is: $6.6 \cdot 10^{-2}$; for predatory on prey: $6.1 \cdot 10^{-3}$; for prey on predator: $3.6 \cdot 10^{-2}$; and finally for predator on predator: 0.

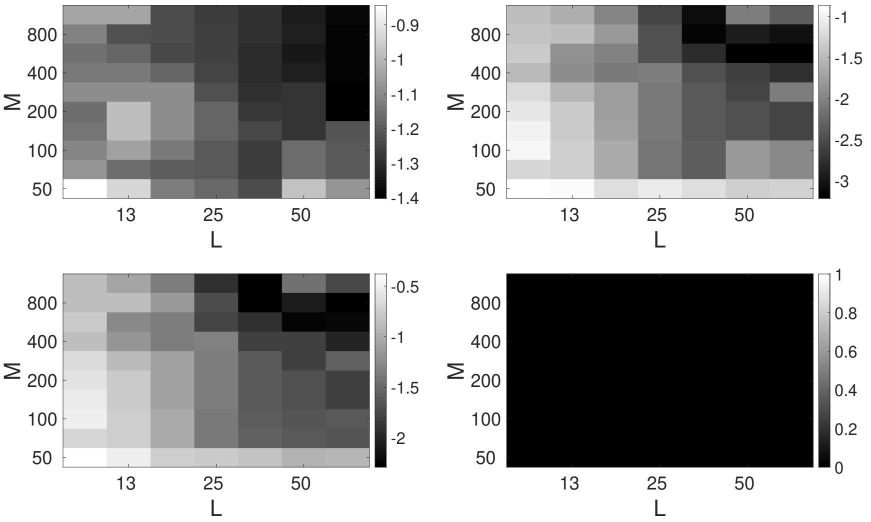


Fig. S9. (PS1) Relative error, in $\log_{10}$ scale, of $\hat{\phi}^E_{k,k'}$ (with (k,k') increasing lexicographically from top-left to bottom-right) as a function of L and M. The error decreases both in L and M, in fact roughly in the product ML. The fourth plot is an identically 0 absolute error, because both $\phi^E_{2,2}$ and its estimator are identically 0, since there is only one predator. Note $M \gg 1$ seems to be needed for accurate inference of the interaction kernels, regardless of how large L is: the trajectories explored for small M do not explore enough configuration to enable estimation, suggesting that the limit $M \to +\infty$ considered in this work is of fundamental importance, at least for non-ergodic systems.

Predator-Swarm, 2^{nd}-order (PS2nd). The second order Predator-Swarm system is a special case of the second order system which is considered in this paper, without alignment-based interactions and without environment variables ξ_i's, similar to the Cucker-Dong model of repulsion-attraction (20) and D'Orsogna-Bertozzi model for modeling fish school formation (5, 6) without the non-collective forcing term. The energy-based interactions are

$$\phi_{1,1}(r) = 1 - r^{-2}, \quad \phi_{1,2}(r) = -r^{-2}, \quad \phi_{2,1}(r) = 1.5r^{-2.5}, \quad \phi_{2,2}(r) \equiv 0.$$

The non-collective change on $\dot{\boldsymbol{x}}_i$ is $F^{\boldsymbol{v}}_i(\dot{\boldsymbol{x}}_i, \xi_i) = -\nu_{\mathcal{k}_i}\dot{\boldsymbol{x}}_i$, where the friction constants are type-based and $\nu_k = 1$ for all $k = 1, \cdots, K$; and the mass of each agent is $m_i = 1$ for all $i = 1, \cdots, N$. We consider the system and test parameters given in table S12 (the initial velocity of preys and predator are fixed at $0 \in \mathbb{R}^2$).

Table S12

d	N_1	N_2	M	L	T
2	9	1	150	300	10

$n_{1,1}$	$n_{1,2} = n_{2,1}$	$n_{2,2}$	deg($\psi^E_{kk'}$)	Preys $\mu_0^{\boldsymbol{X}}$	Pred. $\mu_0^{\boldsymbol{X}}$
1620	540	180	$[1, 1; 1, 0]$	Unif. on $[0.1, 1]^2$	Unif. on $[0, 0.08]^2$

(PS2nd) System Parameters

Note that the two dynamics, predator-prey 1^{st} order and predator-prey 2^{nd} order, use a similar set of interaction kernels, however, the resulting dynamics are significantly different from each other, as demonstrated in both the distribution of pairwise distance data and in the trajectories.

In the middle column of Fig. 5 in the main text, we show the comparison of the learned interaction kernels versus the true interaction kernels (with $\rho_{T,r}^{L,kk'}$ and $\rho_{T,r}^{L,M,kk'}$ shown in the background), and the comparison of true and learned trajectories over two different sets of initial conditions. Similar observations to those for the 1^{st} order system apply here. Errors of the estimators in the $L^2(\rho_T^{L,kk'})$ norms are reported in Table S13. The test on trajectories (bottom middle portion (4 sub-figures) of Fig. 5 in the main text) shows visually the accuracy of the predicted trajectories, quantified by the numerical report in Table S14. We also compare in Fig. S10 the true and learned trajectories over a corresponding system with N_{new} agents. We consider the effect of adding noise to observations, with results visualized in Figure S11. Figures S9 and S12 show the behavior of the error of the estimator (for systems (PS1st) and (PS2nd) respectively) as both L and M are increased.

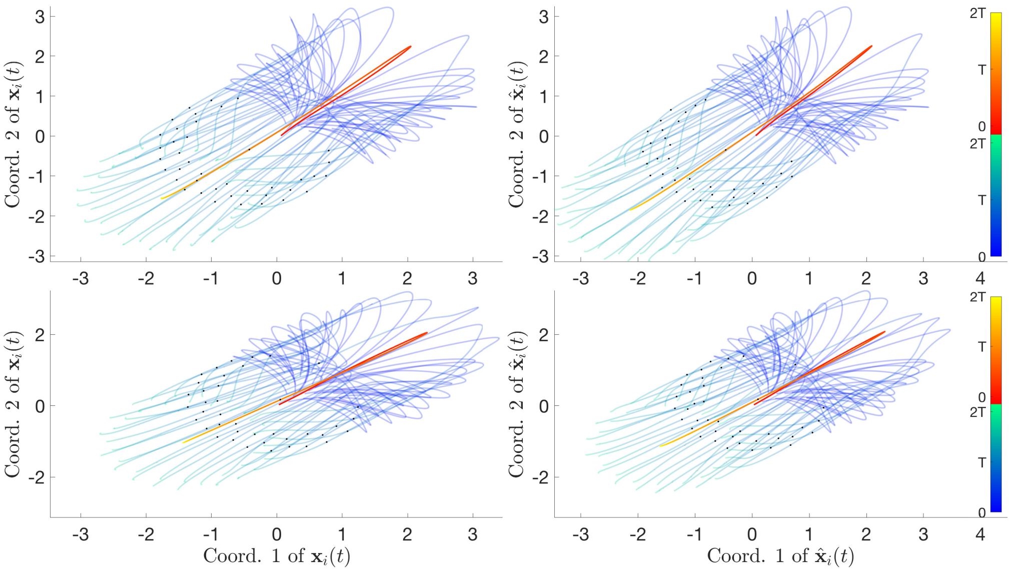


Fig. S10. (PS2nd) Trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ obtained with ϕ and $\hat{\phi}$ respectively, for two randomly chosen initial conditions and evolved for N_{new} agents (with the same setup as in the case of N agents). Trajectory errors are shown in Table S14.

Table S13

Rel. Err. for $\hat{\phi}^E_{1,1}$	$1.5 \cdot 10^{-1} \pm 5.0 \cdot 10^{-2}$
Rel. Err. for $\hat{\phi}^E_{1,2}$	$1.3 \cdot 10^{-1} \pm 1.1 \cdot 10^{-2}$
Rel. Err. for $\hat{\phi}^E_{2,1}$	$7.1 \cdot 10^{-1} \pm 3.8 \cdot 10^{-1}$
Abs. Err. for $\hat{\phi}^E_{2,2}$	0

(PS2nd) Estimator Errors

Table S14

	$[0, T]$	$[T, T_f]$
mean$_{\text{IC}}$: Training ICs	$3.5 \cdot 10^{-1} \pm 1.2 \cdot 10^{-1}$	$7.9 \cdot 10^{-1} \pm 2.1 \cdot 10^{-1}$
std$_{\text{IC}}$: Training ICs	$6.5 \cdot 10^{-1} \pm 2.7 \cdot 10^{-1}$	$1.2 \pm 3.7 \cdot 10^{-1}$
mean$_{\text{IC}}$: Random ICs	$3.5 \cdot 10^{-1} \pm 1.2 \cdot 10^{-1}$	$8.0 \cdot 10^{-1} \pm 2.3 \cdot 10^{-1}$
std$_{\text{IC}}$: Random ICs	$5.8 \cdot 10^{-1} \pm 1.6 \cdot 10^{-1}$	$1.2 \pm 3.1 \cdot 10^{-1}$
mean$_{\text{IC}}$: Larger N	$2.0 \cdot 10^{-1} \pm 3.0 \cdot 10^{-2}$	$4.6 \cdot 10^{-1} \pm 1.2 \cdot 10^{-1}$
std$_{\text{IC}}$: Larger N	$1.1 \cdot 10^{-1} \pm 1.4 \cdot 10^{-2}$	$2.5 \cdot 10^{-1} \pm 5.6 \cdot 10^{-2}$

(PS2nd) Trajectory Errors

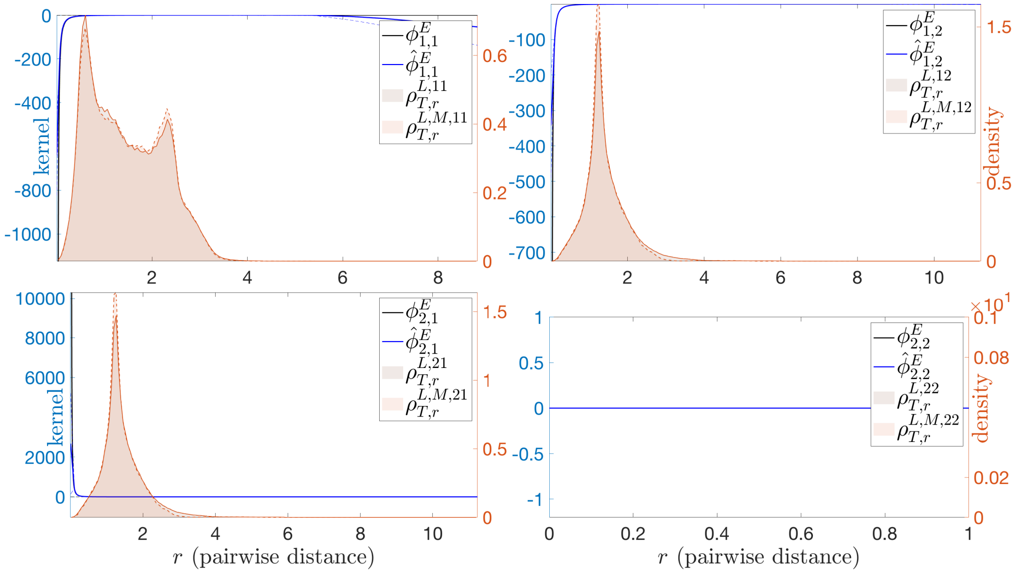


Fig. S11. (PS2nd) Interaction kernels learned with Unif.($[-\sigma, \sigma]$) multiplicative noise, for $\sigma = 0.1$ in the observed positions and velocities, with parameters as in Table S12. The estimated kernels are minimally affected, mostly in regions with small ρ^L_T near 0.

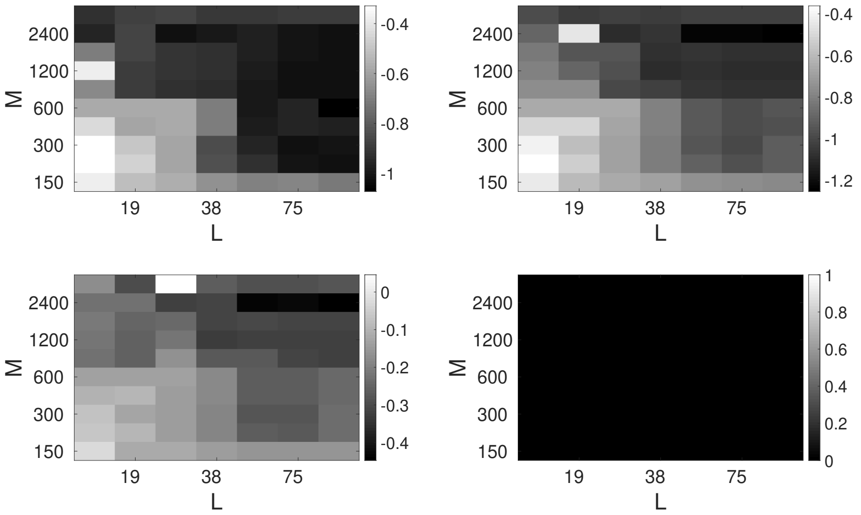


Fig. S12. (PS2) Relative error, in $\log_{10}$ scale, of $\hat\phi^E_{k,k'}$ (with (k, k') increasing lexicographically from top-left to bottom-right) as a function of L and M. The error decreases both in L and M, in fact roughly in the product ML (we impute the lack of monotonicity of some of the entries in the plots to the variance in the results). The fourth plot is an identically 0 absolute error, because both $\phi^E_{2,2}$ and its estimator are identically 0, since there is only one predator. Note $M \gg 1$ seems to be needed for accurate inference of the interaction kernels, regardless of how large L is: the trajectories explored for small M do not explore enough configuration to enable estimation, suggesting that the limit $M \to +\infty$ considered in this work is of fundamental importance, at least for non-ergodic systems.

D. Phototaxis Dynamics. Second order models have been widely used in describing self-organized human motion (21–23), synthetic agent (robots, drones, etc.) behavior (24–27), and bacteria/cell aggregation and motility (28–31). A step further in accurately model reality is to consider models with responses of agents to their surrounding environment or the spread of emotion among agents within a system. Such phenomena appear in a variety of applications, including modeling of emergency evacuation, crowded pedestrian dynamics, bacteria movement toward certain food sources (28–36). We choose here a system modeling the dynamics of phototactic bacteria towards a fixed light source. This system extends the Cucker-Smale system (9, 37, 38) with an extra auxiliary variable ξ_i modeling the response (called excitation level) of individual bacteria to the light source. The dynamics is known to lead to flocking (all bacteria moving in the same direction) within a rather short amount time, due to the interaction kernel having a long interaction range and the effect of light entering the dynamics uniformly. This system is within our family of the second order systems, with homogeneous agents and no energy-induced interaction kernel. The alignment-based interaction kernels acting on $\dot{\boldsymbol{x}}_i$ and ξ_i are the same:

$$\phi^{\boldsymbol{v}}(r) = \phi^{\xi}(r) = (1 + r^2)^{-\frac{1}{4}}.$$

The non-collective change on $\dot{\boldsymbol{x}}_i$ is given by

$$F_i^{\boldsymbol{v}}(\dot{\boldsymbol{x}}_i, \xi_i) = I_0(\boldsymbol{v}_{\text{term}} - \dot{\boldsymbol{x}}_i)(1 - \gamma(\xi_i; \xi_{\text{cr}})),$$

where $I_0 = 0.1$ is the light intensity, $\boldsymbol{v}_{\text{term}} = (60, 0)$ is the terminal velocity (light source at infinity), $\xi_{\text{cr}} = 0.3$ is the critical excitation level (when the light effect activates the bacteria), and $\gamma(\cdot)$ is the smooth cutoff function

$$\gamma(\xi; \xi_c) = \begin{cases} 1, & 0 \le \xi < \xi_c, \\ \frac{1}{2}(\cos(\frac{\pi}{\xi_c}(\xi - \xi_c) + 1), & \xi_c \le \xi < 2\xi_c, \\ 0, & 2\xi_c \le \xi. \end{cases}$$

Here ξ_c is a a threshold constant. The non-collective change on ξ_i is given by

$$F_i^{\xi}(\xi_i) = I_0\gamma(\xi_i; \xi_{\text{cp}}),$$

where $\xi_{\text{cp}} = 0.6$ is the maximum excitation level of light effect on the bacteria. The system parameters are summarized in Table S15.

Table S15

d	M	L	T
2	50	200	0.25

$\mu_0^{\boldsymbol{X}} = \mu_0^{\dot{\boldsymbol{X}}}$	μ_0^{Ξ}	$n^{\boldsymbol{v}} = n^{\xi}$	deg($\psi_{kk'}^{A}$) = deg($\psi_{kk'}^{\xi}$)
Unif. on $[0, 100]^2$	Unif. on $[0, 0.001]^2$	400	1

(PT) Parameters for Phototaxis Dynamics

In the right column of Fig. 5 in the main text, we show the comparison of the learned interaction kernels $\hat{\phi}^A$ and $\hat{\phi}^{\xi}$ versus the true interaction kernels, as well as the comparison of true and learned trajectories over two different sets of initial conditions. We are able to accurately learn the interaction kernels $\hat{\phi}^A$ and $\hat{\phi}^{\xi}$ over the support of ρ_T when pairwise distance data is abundant. When the pairwise distance data becomes scarce towards the two ends of the interaction interval $[0, R]$, we are able to faithfully capture the behavior of ϕ at $r = 0$; the errors are larger near the upper end $r = R$, where the data is extremely scarce. Crucially, we recover faithfully the interactions between the agents and their environment. Estimation errors in the appropriate $L^2(\rho_{T,r,\dot{r}}^L)$- and $L^2(\rho_{T,r,\xi}^L)$-norms are reported in Table S16. A case with noisy observation is also investigated and shown in Fig. S15. Trajectory errors are shown in Table S17. We also compare in Fig. S13 the true and learned trajectories for a corresponding system a dynamics with larger N.

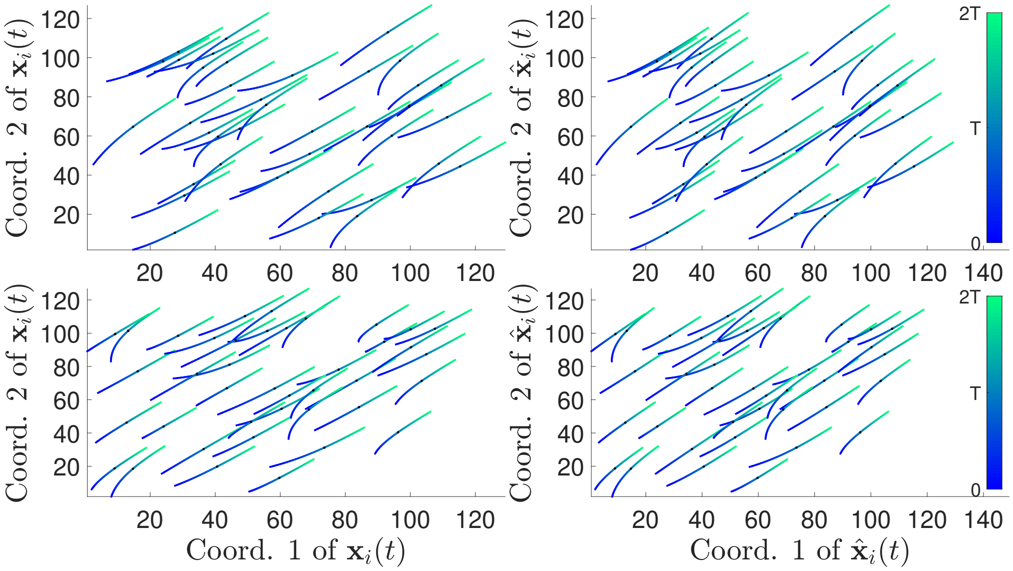


Fig. S13. (PT) Trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ obtained with true and learned interaction kernels respectively, for two randomly chosen initial conditions and evolved using the larger number of agents N_{new} (governed by the same equations as in the case of N agents). Trajectory errors are shown in Table S17.

Table S16

Rel. Err. for $\hat{\phi}^A$	$9.4 \cdot 10^{-3} \pm 5.2 \cdot 10^{-3}$
Rel. Err. for $\hat{\phi}^\xi$	$8.2 \cdot 10^{-3} \pm 5.0 \cdot 10^{-3}$

(PT) Estimator Errors

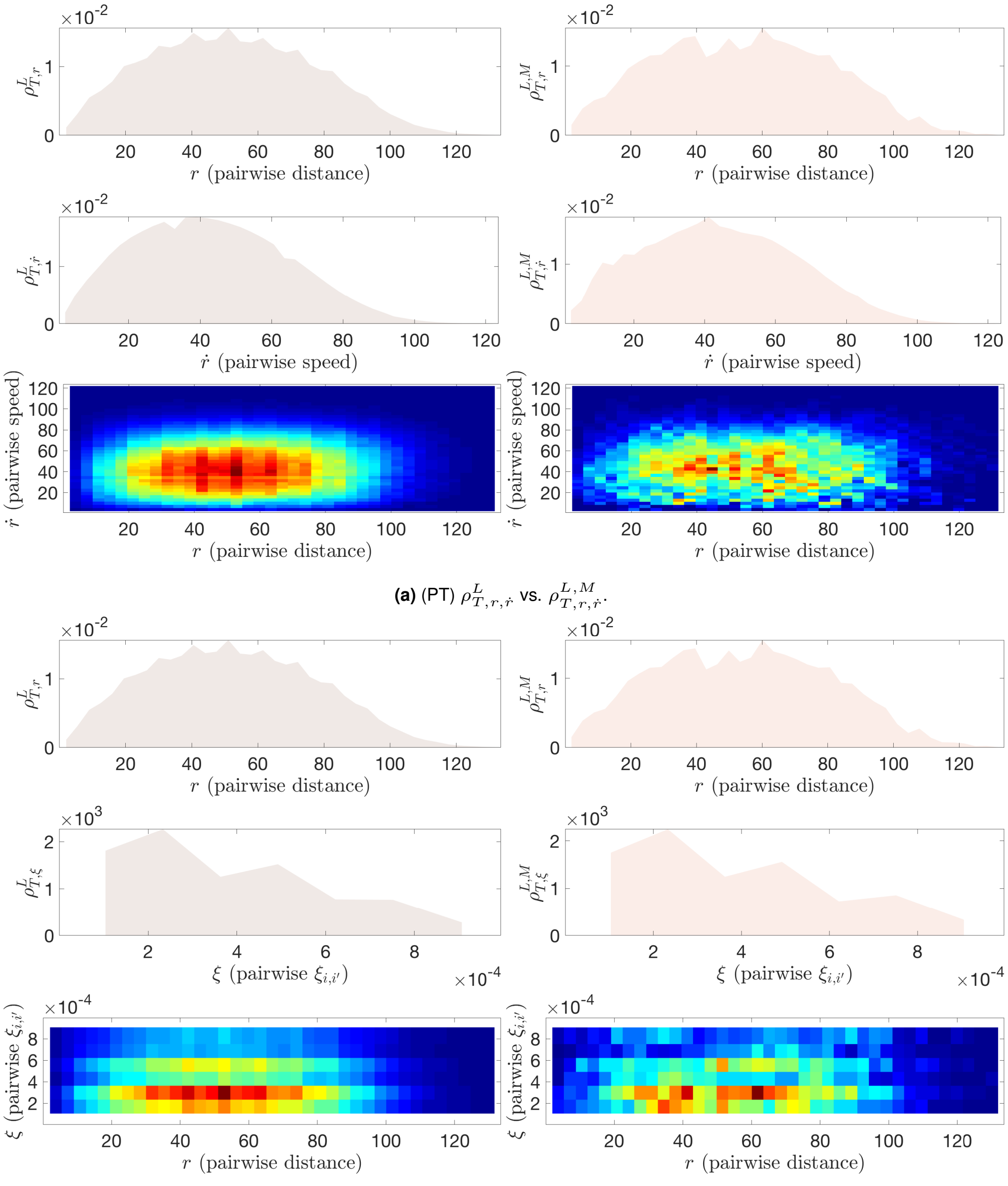
Table S17

	$[0, T]$	$[T, T_f]$
mean$_{\text{IC}}$: Training ICs	$1.6 \cdot 10^{-3} \pm 5.7 \cdot 10^{-5}$	$6.5 \cdot 10^{-3} \pm 9.1 \cdot 10^{-4}$
std$_{\text{IC}}$: Training ICs	$3.1 \cdot 10^{-4} \pm 4.8 \cdot 10^{-5}$	$8.1 \cdot 10^{-3} \pm 3.9 \cdot 10^{-3}$
mean$_{\text{IC}}$: Random ICs	$1.8 \cdot 10^{-3} \pm 8.0 \cdot 10^{-4}$	$7.3 \cdot 10^{-3} \pm 3.2 \cdot 10^{-3}$
std$_{\text{IC}}$: Random ICs	$1.5 \cdot 10^{-3} \pm 3.4 \cdot 10^{-3}$	$1.1 \cdot 10^{-2} \pm 1.2 \cdot 10^{-2}$
mean$_{\text{IC}}$: Larger N	$4.2 \cdot 10^{-3} \pm 1.6 \cdot 10^{-3}$	$8.4 \cdot 10^{-3} \pm 3.8 \cdot 10^{-3}$
std$_{\text{IC}}$: Larger N	$2.9 \cdot 10^{-3} \pm 3.0 \cdot 10^{-3}$	$7.9 \cdot 10^{-3} \pm 7.0 \cdot 10^{-3}$

(PT) Trajectory Errors

x

Finally we display, in Fig. S14a and S14b, the two joint distributions $\rho^L_{T,r,\dot{r}}$ and $\rho^L_{T,r,\xi}$, used to define the appropriate L^2-norms for measuring the performance of $\hat{\phi}^A$ and $\hat{\phi}^\xi$. We also calculated the ℓ^1 distance between the joint distribution $\rho^L_{T,r,\dot{r}}$ and the product of its marginals, and it is $1 \cdot 10^{-1}$. For the ℓ^1 distance between $\rho^L_{T,r,\xi}$ and the product of its marginals, it is $7 \cdot 10^{-2}$. For the empirical distributions (over 10 learning trials), the ℓ^1 distance for $\rho^{L,M}_{T,r,\dot{r}}$ and the product of its marginal is $7 \cdot 10^{-1} \pm 1 \cdot 10^{-2}$; whereas the ℓ^1 distance of $\rho^{L,M}_{T,r,\xi}$ to the product of its marginals is $3.7 \cdot 10^{-1} \pm 7 \cdot 10^{-3}$.



(a) (PT) $\rho^{L}_{T,r,\dot r}$ vs. $\rho^{L,M}_{T,r,\dot r}$.

(b) (PT) $\rho^{L}_{T,r,\xi}$ vs. $\rho^{L,M}_{T,r,\xi}$.

Fig. S14. (PT) Density plots for the various ρ^{L}_{T} measures.

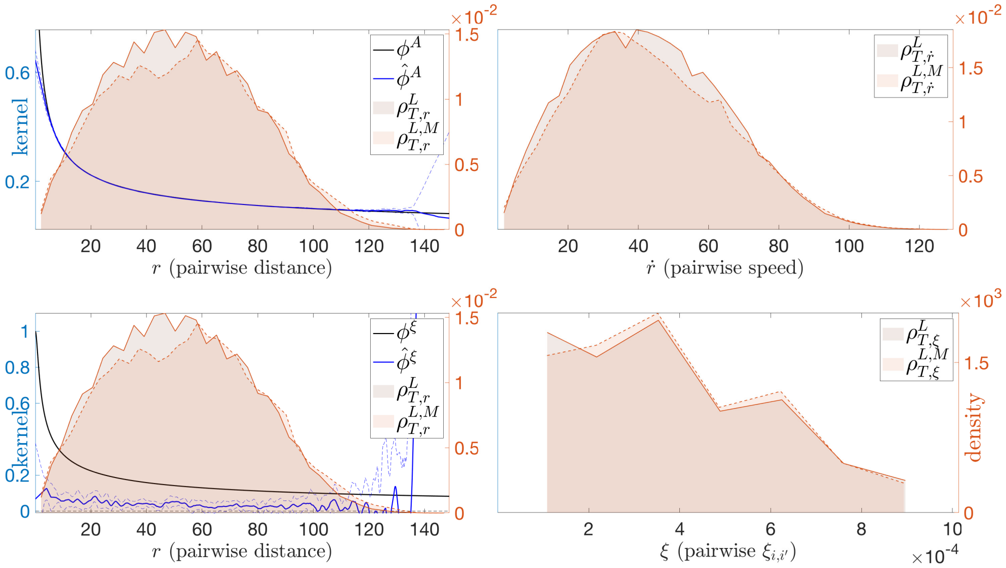


Fig. S15. (PT) Interaction kernels learned from noisy observations of positions and velocities. The noises are multiplicative, Unif.$([-\sigma,\sigma])$ with $\sigma = 0.1$ and with other parameters as in Table S15. The estimated kernel for associated with $\dot{\boldsymbol{x}}_i$ is minimally affected, mostly in regions with small ρ_T^L; the additive noise is on a scale far great then that on ξ_i hence severely affects the learning result on the interaction kernel on ξ_i.

Figure S16 shows the behavior of the error of the estimators as both L and M are increased.

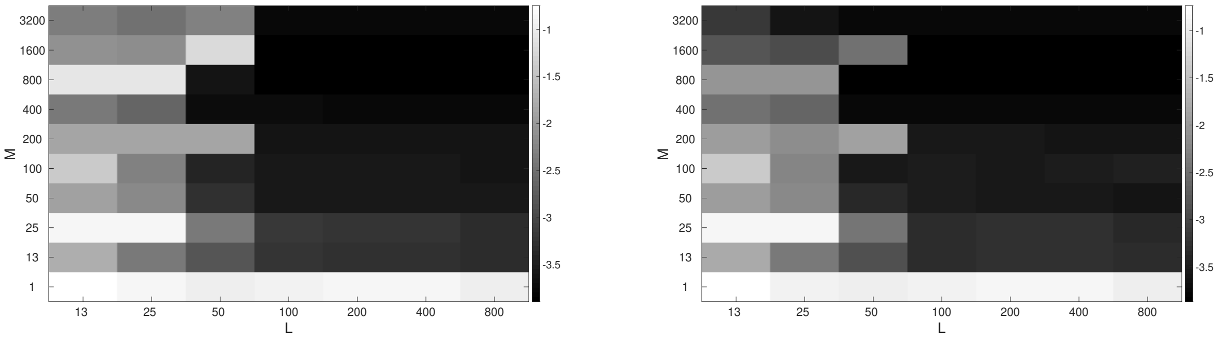


Fig. S16. (PT) Relative error, in $\log_{10}$ scale, of $\hat\phi^A$ (left) and $\hat\phi^\xi$ (right) as a function of L and M. The error decreases both in L and M, in fact roughly in the product ML. The fourth plot is an identically 0 absolute error, because both $\phi^E_{2,2}$ and its estimator are identically 0, since there is only one predator. Note $M \gg 1$ seems to be needed for accurate inference of the interaction kernels, regardless of how large L is: the trajectories explored for small M do not explore enough configuration to enable estimation, suggesting that the limit $M \to +\infty$ considered in this work is of fundamental importance, at least for non-ergodic systems.

E. Model Selection. Our learning approach can be used to identify the model of the system from the observation data. We consider here two different scenarios of model selection: one is identifying the type – energy-based vs. alignment-based – of interaction kernels from a second order system driven by only one type of interaction kernel; the other is to identify the order of the system from a heterogeneous dynamics.

Model Selection: energy-based vs. alignment-based interactions. We consider a special case of the second order homogeneous agent dynamics, given as either

$$\ddot{\boldsymbol{x}}_i = \sum_{i'=1}^{N} \frac{1}{N}\phi^E(r_{ii'})\boldsymbol{r}_{ii'} \quad \text{or} \quad \ddot{\boldsymbol{x}}_i = \sum_{i'=1}^{N} \frac{1}{N}\phi^A(r_{ii'})\dot{\boldsymbol{r}}_{ii'},$$

with the (unknown) interaction kernels defined as

$$\phi^E(r) = 2 - \frac{1}{r^2} \quad \text{and} \quad \phi^A(r) = \frac{1}{(1+r^2)^{0.25}}.$$

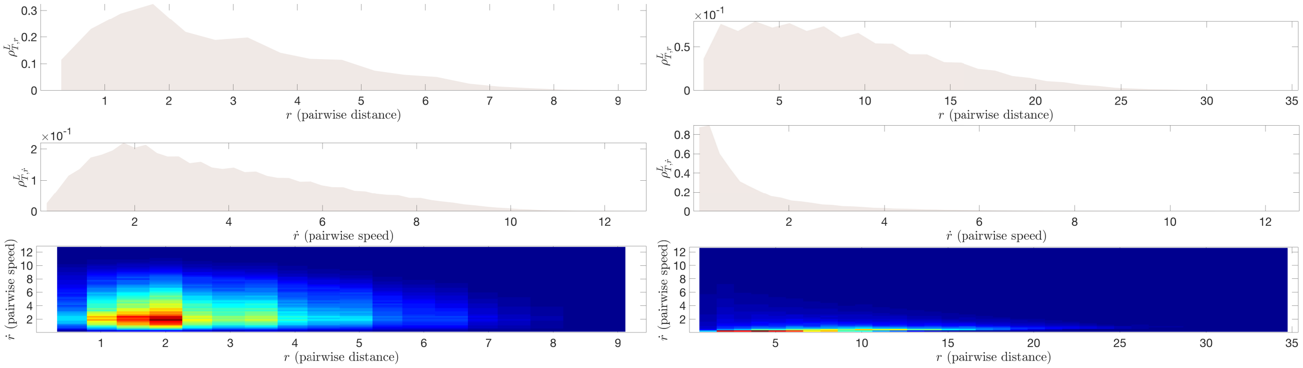
The system parameters are given in Table S18.

Table S18

d	M	L	T	$\mu_0^{\boldsymbol{X}}$	$\mu_0^{\dot{\boldsymbol{X}}}$	$n^E = n^A$	deg(ψ^A)=deg(ψ^ξ)
2	200	200	10	Unif. on ring $[0.5, 1]$	$\mathcal{U}([0, 10]^2)$	800	1

(MS1 and 2) Test Parameters

Given the observation data from either system (ϕ^E- or ϕ^A-driven), we proceed to learn the interaction kernels as usual, i.e. as if the dynamics were generated with both energy-based and alignment-based interaction kernels present. Results are shown in Fig. 7 in the main text. The two sub-figures on the left show the learned interaction kernels $\hat{\phi}^E$ and $\hat{\phi}^A$ from a purely energy-based system: $\hat{\phi}^A$ is small in the appropriate norm, while $\hat{\phi}^E$ is large (and a good approximation to ϕ^E): the estimators can therefore detect this is an energy-driven system. In the two sub-figures on the right, we display the analogous results corresponding to learning the interaction kernels for an alignment-based system. We obtain (almost) 0 for the norm of $\hat{\phi}^E$. The reason why the $L^2(\rho^L_{T,r,\dot{r}})$ norm of $\hat{\phi}^A$ (from the first case) is not as close to 0 as the $L^2(\rho^L_{T,r})$ norm of the $\hat{\phi}^E$ (from the second case) lies in the difference in the joint distribution of the two cases, see Figures S17a and S17b. To further investigate the properties of the joint distributions (and also to differentiate the two dynamics), we calculated the ℓ^1 distance of the respective joint distributions to the product and their marginals. For MS1, the ℓ^1 distance (over 10 learning trials) between the joint distribution $\rho^{L,M}_{T,r,\dot{r}}$ and the product of its marginals is $1.3 \cdot 10^{-1} \pm 3.8 \cdot 10^{-3}$. For MS2, the ℓ^1 distance (over 10 learning trials) between the joint distribution $\rho^{L,M}_{T,r,\dot{r}}$ and the product of its marginals is $4.6 \cdot 10^{-1} \pm 3.4 \cdot 10^{-3}$.



(a) (MS1) Joint distribution of $\rho^L_{T,r,\dot{r}}$.

(b) (MS2) Joint distributions of $\rho^L_{T,r,\dot{r}}$.

Fig. S17. (MS1 and 2) Density plots for the various ρ^L_T measures.

Model Selection: first order vs. second order. We consider two different heterogeneous agent systems, one first order and one second order, with the order of the system unknown to the estimator. The observations are in the time interval $[0, T]$, and in this case $T_f = T$. We first consider the first order heterogeneous agent system

$$\dot{\boldsymbol{x}}_i = \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \phi_{k_i k_{i'}}(r_{ii'})\boldsymbol{r}_{ii'},$$

with

$$\phi_{1,1}(r) = 1 - r^{-2}, \quad \phi_{1,2}(r) = -2r^{-2}, \quad \phi_{2,1}(r) = 3.5r^{-3}, \quad \phi_{2,2}(r) \equiv 0,$$

and the type information setup similar to that of the Predator-Swarm first order system (detailed in Sec.3C). For the second scenario, we consider the data generated by the following second order heterogeneous agent dynamics,

$$\ddot{\boldsymbol{x}}_i = -\dot{\boldsymbol{x}}_i + \sum_{i'=1}^{N} \frac{1}{N_{k_{i'}}} \phi^E_{k_i k_{i'}}(r_{ii'})\boldsymbol{r}_{ii'},$$

with

$$\phi_{1,1}(r) = 1 - r^{-2}, \quad \phi_{1,2}(r) = -r^{-2}, \quad \phi_{2,1}(r) = 1.5r^{-2.5}, \quad \phi_{2,2}(r) \equiv 0,$$

and the type information setup similar to that of the Predator-Swarm second order system (details shown in Sec.3C). The parameters for both systems are given in Tables S19 and S20.

Table S19

d	M	L	T
2	250	250	1

n	Deg($\psi_{kk'}$)	Prey $\mu_0^{\boldsymbol{X}}$	Pred. $\mu_0^{\boldsymbol{X}}$
$[298, 150; 150, 2]$	$[1, 1; 1, 0]$	Unif. on ring $[0.5, 1.5]$	Unif. on disk at 0.1

(MS3) Test Parameters

Table S20

d	M	L	T
2	250	250	1

n	deg($\psi^E_{kk'}$)	Prey $\mu_0^{\boldsymbol{X}}$	Pred. $\mu_0^{\boldsymbol{X}}$
$[298, 150; 150, 2]$	$[1, 1; 1, 0]$	$\mathcal{U}([0.1, 1]^2)$	$\mathcal{U}([0, 0.07]^2)$

(MS4) Test Parameters

With the order of the ODE system and the interaction kernels being the missing information, we construct estimators for the interaction kernels in two ways: first assuming a first order system, then assuming a second order system (without non-collective forcing). We then generate predicted trajectories using the learned interaction kernels, and the same initial conditions as in the training data. Next, we calculate the trajectory max-in-time error, obtaining the results in Table 1 of the main text (shown as the mean of the trajectory error plus or minus standard deviation of the error over 10 runs). As indicated by the trajectory error statistics, the predicted trajectories with smaller error indicate the correct order of the true underlying system in both cases. Details on the statistics of the trajectory errors are reported in Tables S21 and S22. In each, the column with smaller values (within both mean and standard deviation of the trajectory errors) corresponds the correct order of the system.

Table S21

	Learned as 1^{st} order	Learned as 2^{nd} order
mean$_{\text{IC}}$	$\mathbf{9.5 \cdot 10^{-3} \pm 2 \cdot 10^{-3}}$	3.9 ± 8
std$_{\text{IC}}$	$\mathbf{1.8 \cdot 10^{-2} \pm 1.1 \cdot 10^{-2}}$	$48 \pm 1 \cdot 10^2$

(MS3) Trajectory Errors

Table S22

	Learned as 1^{st} order	Learned as 2^{nd} order
mean$_{\text{IC}}$	$1.6 \pm 1 \cdot 10^{-1}$	$\mathbf{1.3 \cdot 10^{-1} \pm 3 \cdot 10^{-2}}$
std$_{\text{IC}}$	$9.4 \cdot 10^{-1} \pm 2 \cdot 10^{-1}$	$\mathbf{2.0 \cdot 10^{-1} \pm 5 \cdot 10^{-2}}$

(MS4) Trajectory Errors

References

1. Cucker F, Smale S (2002) On the mathematical foundations of learning. *Bull Amer Math Soc* 39(1):1–49.
2. Binev P, Cohen A, Dahmen W, DeVore R, Temlyakov V (2005) Universal algorithms for learning theory part i: piecewise constant functions. *J Mach Learn Res* 6(Sep):1297–1321.
3. DeVore R, Kerkyacharian G, Picard D, Temlyakov V (2006) Approximation methods for supervised learning. *Found Comput Math* 6(1):3–58.
4. Bongini M, Fornasier M, Hansen M, Maggioni M (2017) Inferring interaction rules from observations of evolutive systems I: The variational approach. *Math Mod Methods Appl Sci* 27(05):909–951.
5. Carrillo JA, D'Orsogna MR, Panferov V (2009) Double Milling in self-propelled swarms from kinetic theory. *Kinet Relat Mod* 2(2):363 – 378.
6. Chuang Y, D'Orsogna M, Marthaler D, Bertozzi A, Chayes L (2007) State transition and the continuum limit for the 2D interacting, self-propelled particle system. *Physica D* 232:33 – 47.
7. Cristiani E, Piccoli B, Tosin A (2010) Modeling self-organization in pedestrians and animal groups from macroscopic and microscopic viewpoints in *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences, Modeling and Simulation in Science, Engineering and Technology*, eds. Naldi G, Pareschi L, Toscani G, Bellomo N. (Springer, Birkhäuser Boston), pp. 337 – 364.

8. Couzin I, Franks N (2002) Self-organized lane formation and optimized traffic flow in army ants. *Proc R Soc Lond* B 270:139 – 146.
9. Cucker F, Smale S (2007) Emergent behavior in flocks. *IEEE Trans Automat Contr* 52(5):852.
10. Niwa H (1994) Self-organizing dynamic model of fish schooling. *J Theor Biol* 171:123 – 136.
11. Parrish JK, Edelstein-Keshet L (1999) Complexiy, pattern, and evolutionary trade-offs in animal aggregation. *Science* 284:99 – 101.
12. Parrish J, Viscido S, Gruenbaum D (2002) Self-organized fish schools: An examination of emergent properties. *Biol Bull* 202:296 – 305.
13. Romey W (1996) Individual differences make a difference in the trajectories of simulated schools of fish. *Ecol Model* 92:65 – 77.
14. Toner J, Tu Y (1995) Long-range order in a two-dimensional dynamical xy model: How birds fly together. *Phys Rev Lett* 75:4326 – 4329.
15. Yates C, et al. (2009) Inherent noise can facilitate coherence in collective swarm motion. *Proc Natl Acad Sci USA* 106:5464 – 5469.
16. Escobedo R, Muro C, Spector L, Coppinger RP (2014) Group size, individual role differentiation and effectiveness of cooperation in a homogeneous group of hunters. *J R Soc Interface* 11:20140204.
17. Cohn H, Kumar A (2009) Algorithmic design of self-assembling structures. *Proc Natl Acad Sci USA* 106:9570 – 9575.
18. Nowak MA (2006) Five rules for the evolution of cooperation. *Science* 314:1560 – 1563.
19. Fryxell JM, Mosser A, Sinclair ARE, Packer C (2007) Group formation stabilizes predator-prey dynamics. *Nature* 449:1041 – 1043.
20. Cucker F, Dong JG (2014) A conditional, collision-avoiding, model for swarming. *Discrete Continuous Dyn Syst* 43(3):1009 – 1020.
21. Cristiani E, Piccoli B, Tosin A (2011) Multiscale modeling of granular flows with application to crowd dynamics. *Multi Model Simul* 9(1):155 – 182.
22. Cucker F, Smale S, Zhou D (2004) Modeling language evolution. *Found Comput Math* 4(5):315 – 343.
23. Short MB, et al. (2008) A statistical model of criminal behavior. *Math Models Methods Appl Sci* 18(suppl.):1249 – 1267.
24. Chuang Y, Huang Y, D'Orsogna M, Bertozzi A (2007) Multi-vehicle flocking: scalability of cooperative control algorithms using pairwise potentials. *IEEE Intern Conf Robotics and Automation* pp. 2292 – 2299.
25. Leonard N, Fiorelli E (2001) Virtual leaders, artificial potentials and coordinated control of groups. *Proc 40th IEEE Conf Decision Contr* pp. 2968 – 2973.
26. Pera L, Gómez G, Elosegui P (2009) Extension of the Cucker-Smale control law to space flight formations. *J Guid Control Dyn* 32:527 – 537.
27. Sugawara K, Sano. M (1997) Cooperative acceleration of task performance: Foraging behavior of interacting multi-robots system. *Physica D* 100:343 – 354.
28. Camazine S, et al. (2001) *Self-organization in Biological Systems*, Princeton studies in complexity. (Princeton University Press, Princeton).
29. Evelyn FK, Lee AS (1970) Initiation of slime mold aggregation viewed as an instability. *J Theor Biol* 26 3:399–415.
30. Koch AL, White D (1998) The social lifestyle of myxobacteria. *BioEssays* 20(12):1030–1038.
31. Perthame B (2007) *Transport Equations in Biology*, Frontiers in Mathematics. (Birkhäuser Basel).
32. Moussaid M, Helbing D, Theraulaz G (2011) How simple rules determine pedestrian behavior and crowd disasters. *Proc Natl Acad Sci USA* 108(17):6884 – 6888.
33. Durupinar F, Gudukbar U, Aman A, Badler NI (2015) Psychological Parameters for Crowd Simulation: From Audiences to Mobs. *IEEE Trans Vis Comput Graph* 21:1 – 15.
34. Bosse T, Duell R, Memon ZA, Treur J, van der Wal CN (2009) A multi-agent model for mutual absorption of emotions in *European council on modeling and simulation*, ECMS 2009.
35. Bosse T, Hoogendoorn M, Klein MCA, Treur J, van der Wal CN (2011) Agent-based analysis of patterns in crowd behaviour involving contagion of mental states in *Modern Approaches in Applied Intelligence: 24th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2011, Syracuse, NY, USA, June 28 – July 1, 2011, Proceedings, Part II*, eds. Mehrotra KG, Mohan CK, Oh JC, Varshney PK, Ali M. (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 566–577.
36. Lin J, Luckas TA (2015) A particle swarm optimization model of emergency airplane evacuation with emotion. *Net Het Media* 10:631 – 646.
37. Cucker F, Smale S (2007) On the mathematics of emergence. *Jpn J Math* 2(1):197 – 227.
38. Ha SY, Ha T, Kim JH (2010) Emergent behavior of a Cucker-Smale type particle model with nonlinear velocity couplings. *IEEE Trans Automat Contr* 55(7):1679 – 1683.