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:

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*

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

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

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

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

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

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

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

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)

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

(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*

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

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 $L^2(\mathbb{R}_+, \rho_T^L)$,

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

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

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

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.*

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

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*

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

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

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

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*

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

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

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

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 $dim(\mathcal{H}_n)\leq c_0 n$ and

$$\inf_{\varphi\in\mathcal{H}_n} \|\varphi(\cdot)-\phi(\cdot)\|^2_{L^\infty([0,R])} \leq c_1 n^{-s} \tag{13}$$

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

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

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*

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

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 $\mathcal{H}_n$,

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

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

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

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

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

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 $1 \le l \le L$, $1 \le i \le N$ and $1 \le p \le n$, 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

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

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:

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

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 $1 \le l \le L$ and $1 \le i \le N$ (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 $\mathcal{H}_{n^E}^E$, so that the candidate functions $\boldsymbol{\varphi}^E = \{\varphi_{kk'}^E\}_{k,k'=1}^K$ are expressed as $\boldsymbol{\varphi}^E(r) = \sum_{k,k'=1}^K \sum_{p=1}^{n_{k,k'}^E} \alpha_{kk',p}^E \psi_{kk',p}^E(r)$. Using the same ordering from previous discussion on the first order heterogeneous agent system, we have, for a pair (k,k') learning matrix $\Psi_{L,M}^{m,E}$ for the energy-based interaction kernel,

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

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

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

for $1 \le l \le L$, $i \in C_k$ and $\tilde{n}^A = \sum_{(k_1,k_1')<(k,k')} n_{k_1k_1'}^A$. 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 $\Psi_{L,M}^{m,\boldsymbol{v}} = \left(\Psi_{L,M}^{m,E}, \Psi_{L,M}^{m,A}\right)$, we arrive at the final formulation,

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

As usual, we solve the associated normal equations of Eq. (16) with $A_L^m := (\Psi_{L,M}^{m,\boldsymbol{v}})^\top \Psi_{L,M}^{m,\boldsymbol{v}}$ and $b_L^m := (\Psi_{L,M}^{m,\boldsymbol{v}})^\top (\mathbf{d}^{m,\boldsymbol{v}} - \vec{F}^{m,\boldsymbol{v}})$, 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 $1 \le l \le L$ and $1 \le i \le N$; then the learning matrix $\Psi_{L,M}^{m,\xi}$ is assembled similarly as

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

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

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

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, $\psi^E_{kk',p}$'s and $\psi^A_{kk',p}$'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 $\widehat{\boldsymbol{\phi}}(r)^E = \sum_{k,k'=1}^{K} \sum_{p=1}^{n^E_{kk'}} a^E_{kk',p}\psi^E_{kk',p}(r)$ and $\widehat{\boldsymbol{\phi}}(r)^A = \sum_{k,k'=1}^{K} \sum_{p=1}^{n^A_{kk'}} a^A_{kk',p}\psi^A_{kk',p}(r)$.
    if If there are ξ_i's **then**
        Construct the basis, $\psi^{\xi}_{kk',p}$'s.
        Assemble the normal equations and solve for $\mathbf{a}^{\xi}$.
        Assemble $\widehat{\boldsymbol{\phi}}(r)^{\xi} = \sum_{k,k'=1}^{K} \sum_{p=1}^{n^{\xi}_{kk'}} a^{\xi}_{kk',p}\psi^{\xi}_{kk',p}(r)$.
    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

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

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 $\left\|\Xi - \hat{\Xi}\right\|_{\text{TM}([0,\text{T}])} = \sup_{t\in[0,T]} \left\|\Xi(t) - \hat{\Xi}(t)\right\|_{\mathcal{S}}$.

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

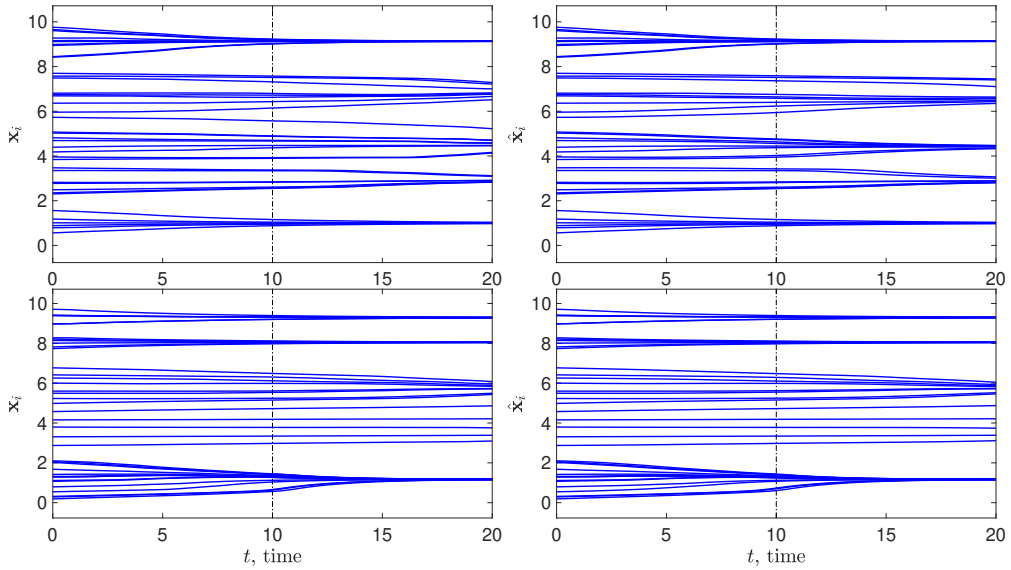


Fig. S1. (OD) Trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ obtained with ϕ and $\hat{\phi}$ respectively, for dynamics with larger $N_{\text{new}} = 4N$, over two different sets of initial conditions. We are able to accurately predict the clusters (number and location). Errors are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

Table S3

	$[0, T]$	$[T, T_f]$
mean_{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_{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_{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_{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_{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_{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.

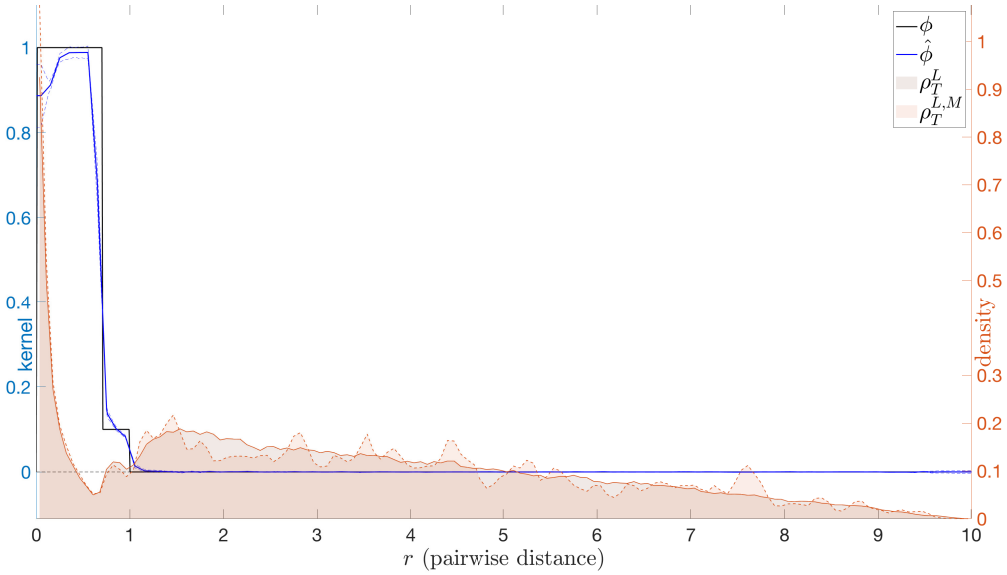


Fig. S2. (OD) Interaction kernel learned with Unif.$([-\sigma, \sigma])$ additive noise, for $\sigma = 0.1$ in the observed positions and velocities. The estimated kernels are minimally affected, mostly in regions with small ρ_T^L and near 0.

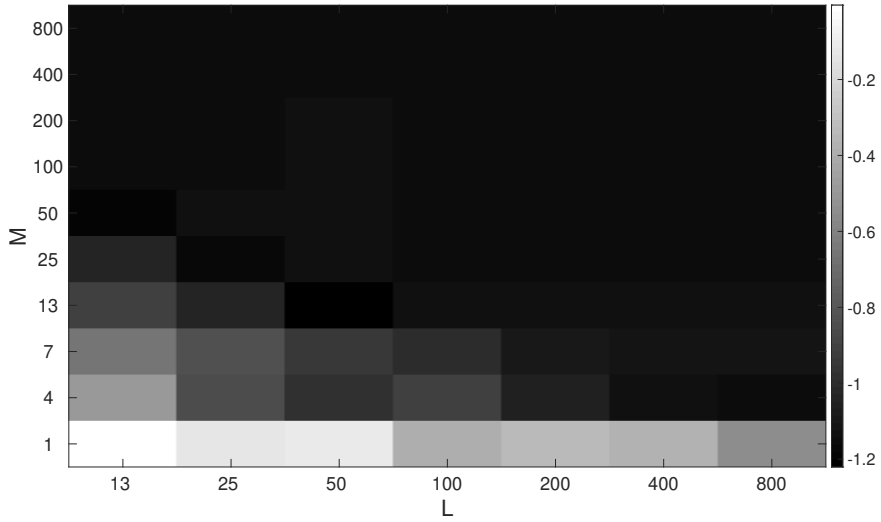
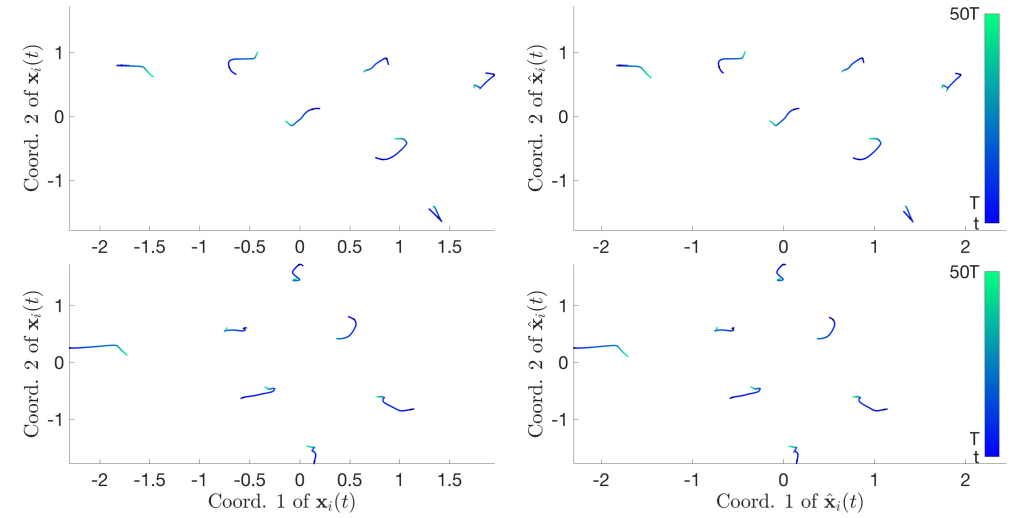
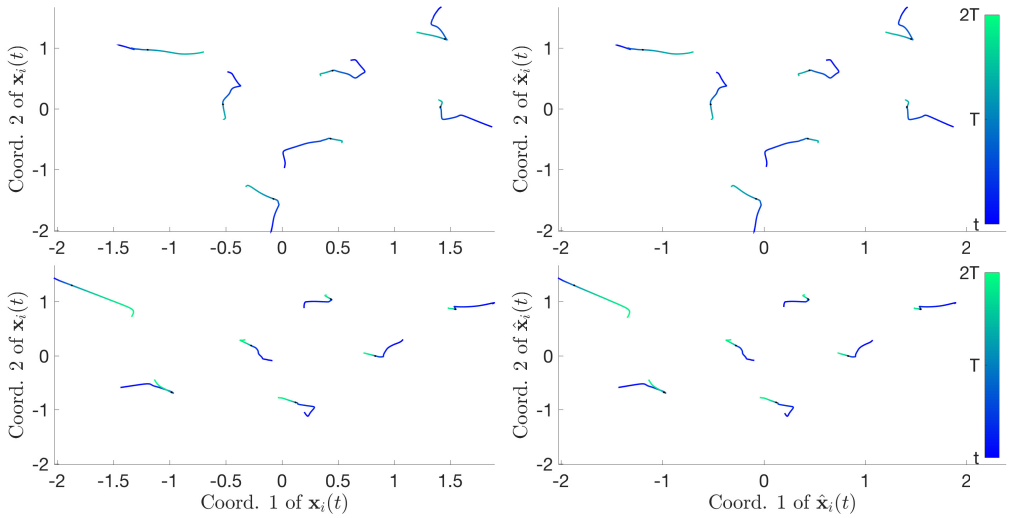


Fig. S3. (OD) Relative error, in $\log_{10}$ scale, of $\hat{\phi}$ as a function of L and M. The error decreases both in L and M, in fact roughly in the product ML, at least when M and L are not too small. $M = 1$ does not seem to suffice, no matter how large L is, due to the limited amount of "information" 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 $\{t_i\}_{i=2,\dots,L}$ 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_{IC}: Training ICs	$1.6 \cdot 10^{-3} \pm 2 \cdot 10^{-4}$	$1.7 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$
std_{IC}: Training ICs	$4.6 \cdot 10^{-4} \pm 5 \cdot 10^{-5}$	$2.1 \cdot 10^{-2} \pm 4 \cdot 10^{-3}$
mean_{IC}: Random ICs	$1.6 \cdot 10^{-3} \pm 2 \cdot 10^{-4}$	$1.7 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$
std_{IC}: Random ICs	$4.5 \cdot 10^{-4} \pm 5 \cdot 10^{-5}$	$1.9 \cdot 10^{-2} \pm 2 \cdot 10^{-3}$
mean_{IC}: Larger N	$6.2 \cdot 10^{-2} \pm 7 \cdot 10^{-3}$	$6.2 \cdot 10^{-2} \pm 2 \cdot 10^{-2}$
std_{IC}: Larger N	$8.2 \cdot 10^{-3} \pm 7 \cdot 10^{-4}$	$3.0 \cdot 10^{-2} \pm 1 \cdot 10^{-2}$
mean_{IC}: Training ICs	$3.4 \cdot 10^{-3} \pm 1 \cdot 10^{-3}$	$5.1 \cdot 10^{-3} \pm 2 \cdot 10^{-3}$
std_{IC}: Training ICs	$2.7 \cdot 10^{-3} \pm 2 \cdot 10^{-3}$	$6.6 \cdot 10^{-3} \pm 3 \cdot 10^{-3}$
mean_{IC}: Random ICs	$4.1 \cdot 10^{-3} \pm 2 \cdot 10^{-3}$	$8.7 \cdot 10^{-3} \pm 8 \cdot 10^{-3}$
std_{IC}: Random ICs	$3.6 \cdot 10^{-3} \pm 2 \cdot 10^{-3}$	$1.5 \cdot 10^{-2} \pm 2 \cdot 10^{-2}$
mean_{IC}: Larger N	$7.7 \cdot 10^{-2} \pm 1 \cdot 10^{-2}$	$6.6 \cdot 10^{-2} \pm 3 \cdot 10^{-2}$
std_{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.

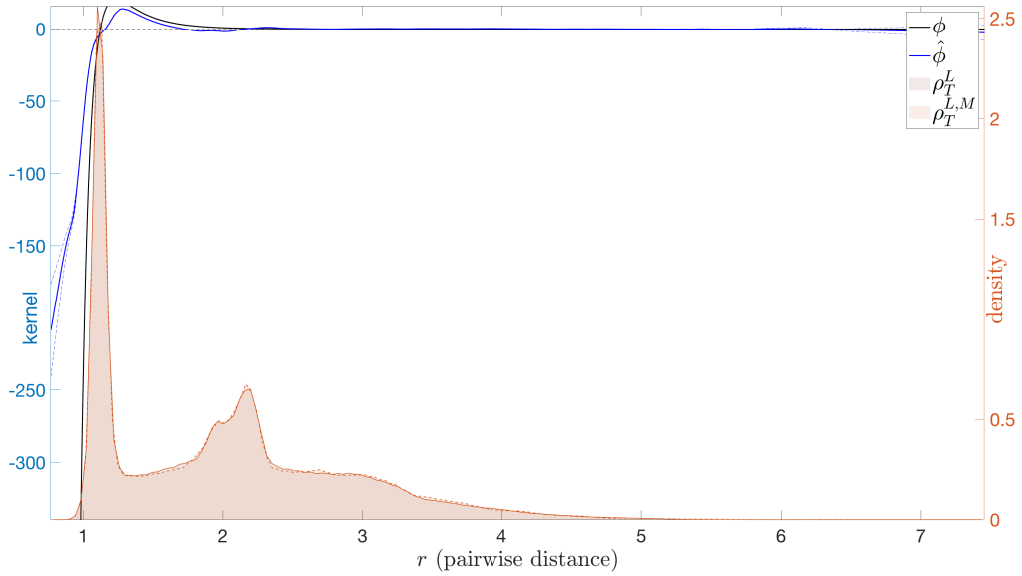


Fig. S5. (LJ) Interaction kernel learned with Unif.$([-\sigma, \sigma])$ additive noise, for $\sigma = 0.1$, in the observed positions *and observed velocities*; here $M = 500$, $L = 2000$, with all the other parameters as in Table S5.

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.

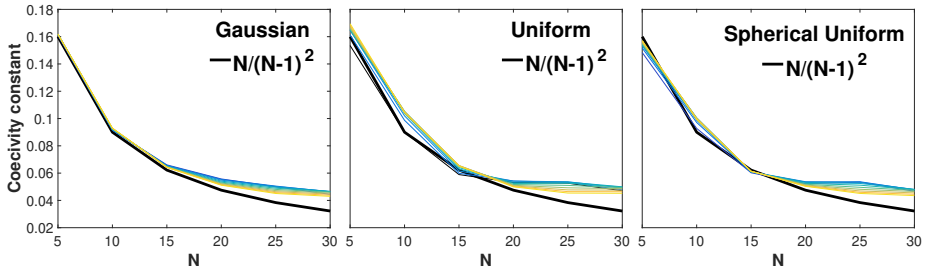


Fig. S6. (LJ) *Coercivity condition validation in 2D Lennard-Jones system with different* N. We compute the empirical coercivity constant $c_{L,N,\mathcal{H}}$ defined in Eq. (7), with $\mathcal{H}$ consisting of 200 piecewise constant basis functions with random coefficients, using $M = 131,072$ trajectories 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 ($\mathrm{PS1}^{st}$). 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

($\mathrm{PS1}^{st}$) 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.

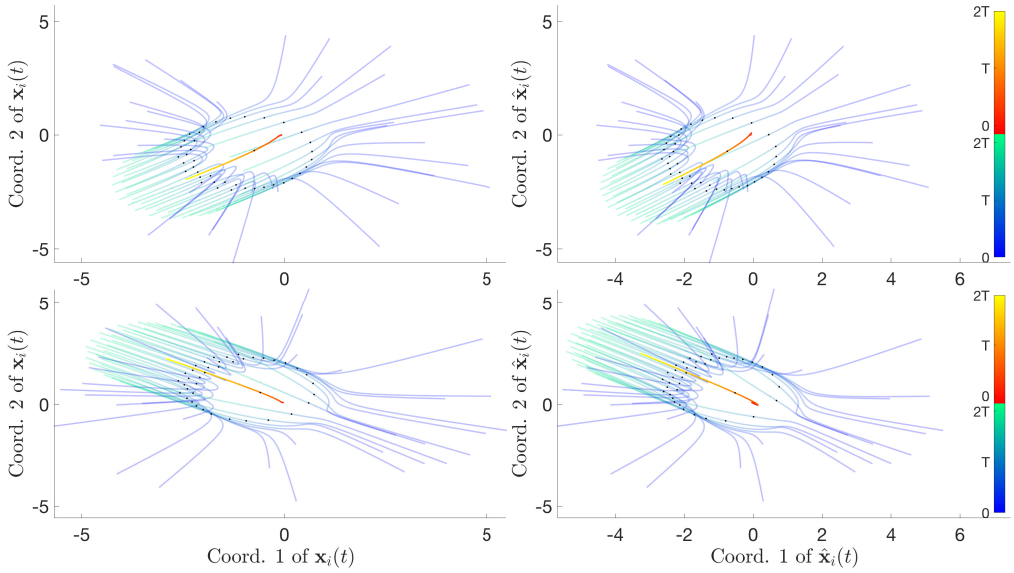


Fig. S7. (PS1st) Trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ obtained with ϕ and $\hat{\phi}$ respectively, for two randomly chosen initial conditions and evolved for N_{new} agents (with the same setup as in the case of N agents). Trajectory errors are shown in Table S11.

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

(PS1st) 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

(PS1st) 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, 1st-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.

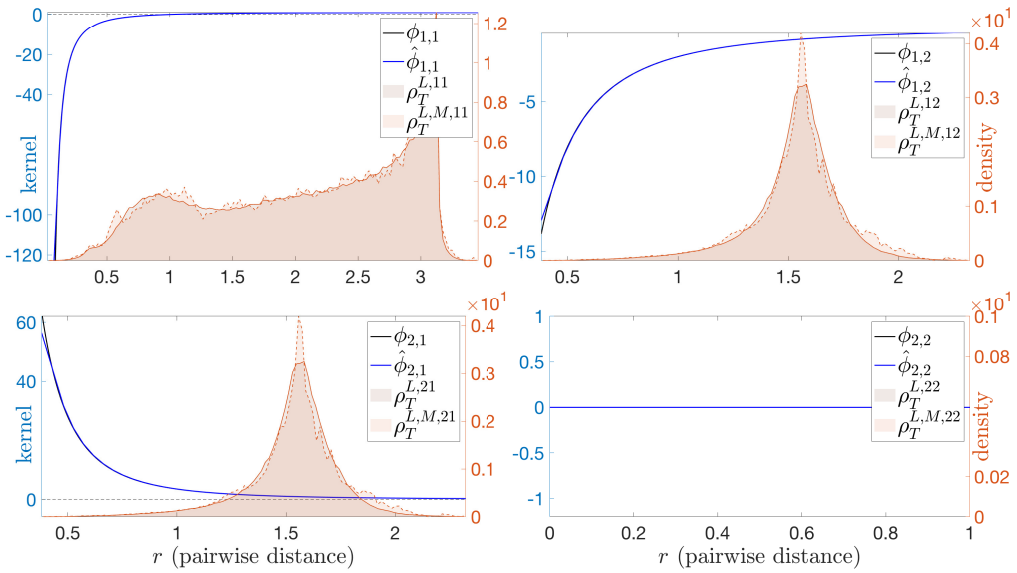


Fig. S8. (PS1st) Comparison of interaction kernels (true versus learned) when the learned kernels are generated by linear B-splines (n as in the other case considered for this system). The relative error (in $L^2(\rho_T)$ norm) for prey on prey interaction is: $6.6 \cdot 10^{-2}$; for predatory on prey: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

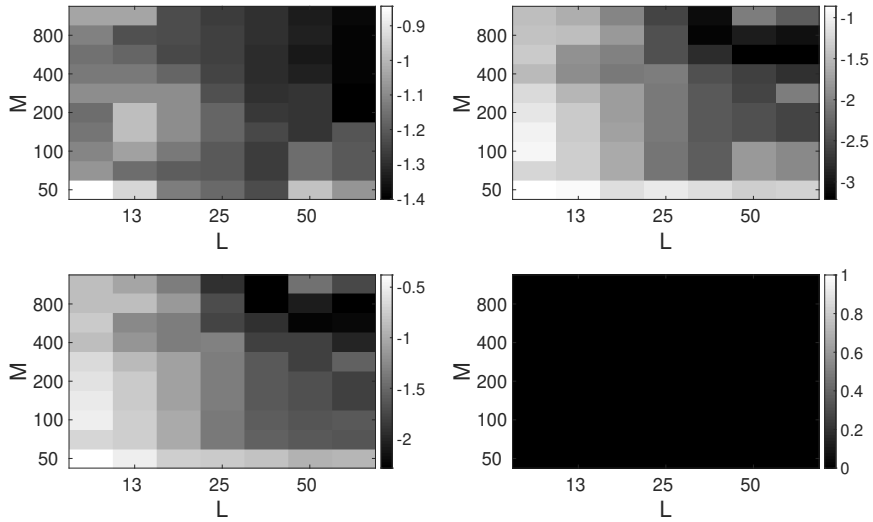


Fig. S9. (PS1) Relative error, in $\log_{10}$ scale, of $\hat{\phi}^E_{k,k'}$ (with (k,k') increasing lexicographically from top-left to bottom-right) as a function of L and M. The error decreases both in L and M, in fact roughly in the product ML. The fourth plot is an identically 0 absolute error, 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 $F_i^{\boldsymbol{v}}(\dot{\boldsymbol{x}}_i, \xi_i) = -\nu_{\xi_i}\dot{\boldsymbol{x}}_i$, 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.

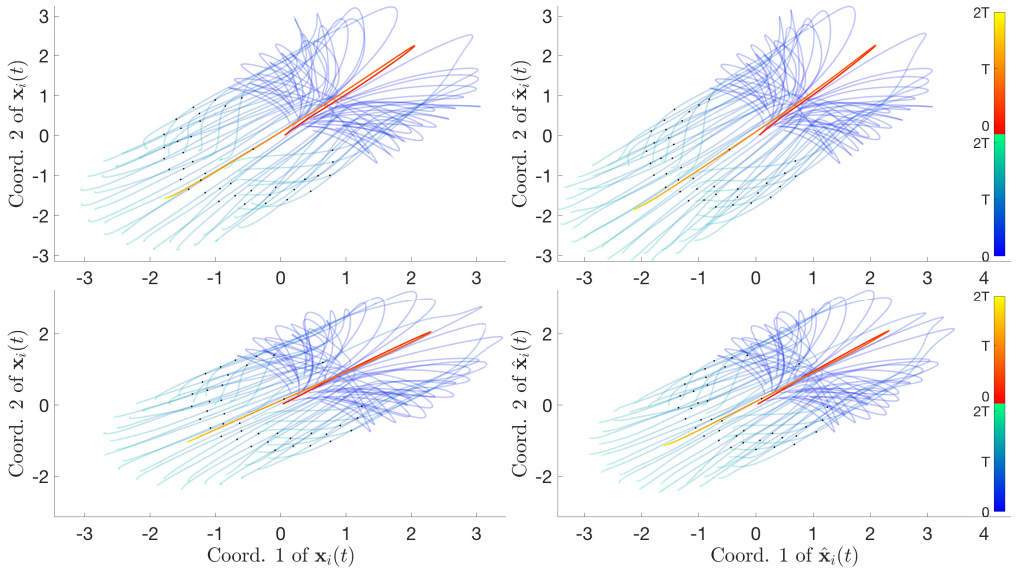


Fig. S10. (PS2nd) Trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ obtained with ϕ and $\hat{\phi}$ respectively, for two randomly chosen initial conditions and evolved for N_{new} agents (with the same setup as in the case of N agents). Trajectory errors are shown in Table S14.

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_{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_{IC}: Training ICs	$6.5 \cdot 10^{-1} \pm 2.7 \cdot 10^{-1}$	$1.2 \pm 3.7 \cdot 10^{-1}$
mean_{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_{IC}: Random ICs	$5.8 \cdot 10^{-1} \pm 1.6 \cdot 10^{-1}$	$1.2 \pm 3.1 \cdot 10^{-1}$
mean_{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_{IC}: Larger N	$1.1 \cdot 10^{-1} \pm 1.4 \cdot 10^{-2}$	$2.5 \cdot 10^{-1} \pm 5.6 \cdot 10^{-2}$

(PS2^{nd}) Trajectory Errors

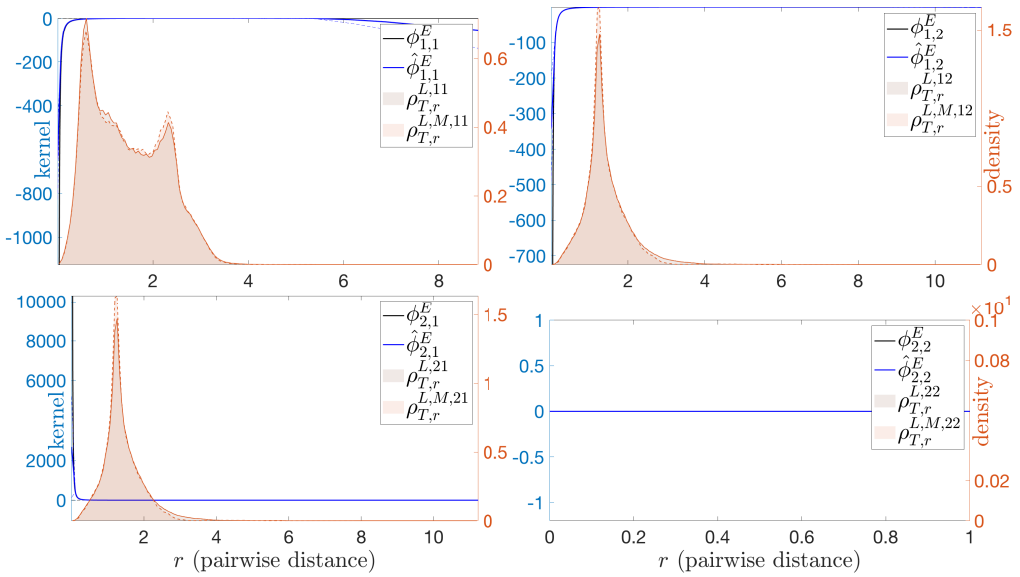


Fig. S11. (PS2^{nd}) Interaction kernels learned with Unif.($[-\sigma, \sigma]$) multiplicative noise, for $\sigma = 0.1$ in the observed positions and velocities, with parameters as in Table S12. The estimated kernels are minimally affected, mostly in regions with small ρ_T^L near 0.

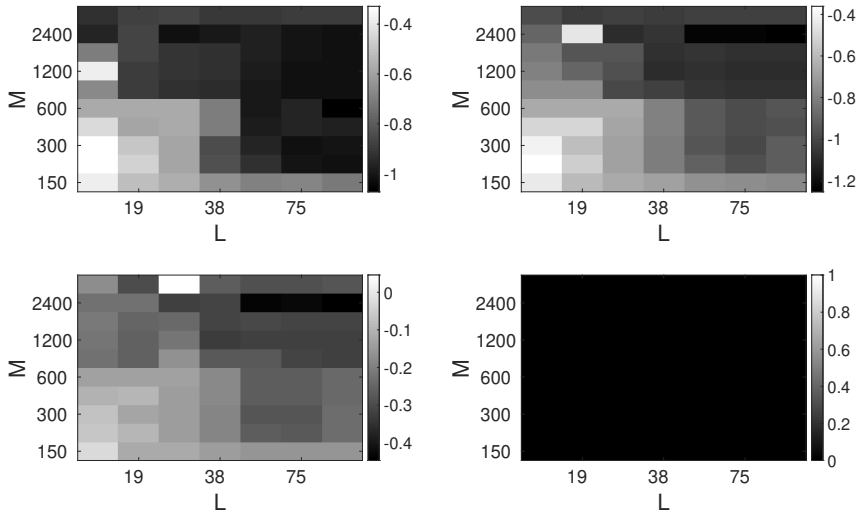


Fig. S12. (PS2) Relative error, in $\log_{10}$ scale, of $\hat{\phi}^E_{k,k'}$ (with (k, k') increasing lexicographically from top-left to bottom-right) as a function of L and M. The error decreases both in L and M, in fact roughly in the product ML (we impute the lack of monotonicity of some of the 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}	$n^{\boldsymbol{v}} = n^{\xi}$	$\deg(\psi^A_{kk'}) = \deg(\psi^{\xi}_{kk'})$
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 $L^2(\rho^L_{T,r,\dot{r}})$- and $L^2(\rho^L_{T,r,\xi})$-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.

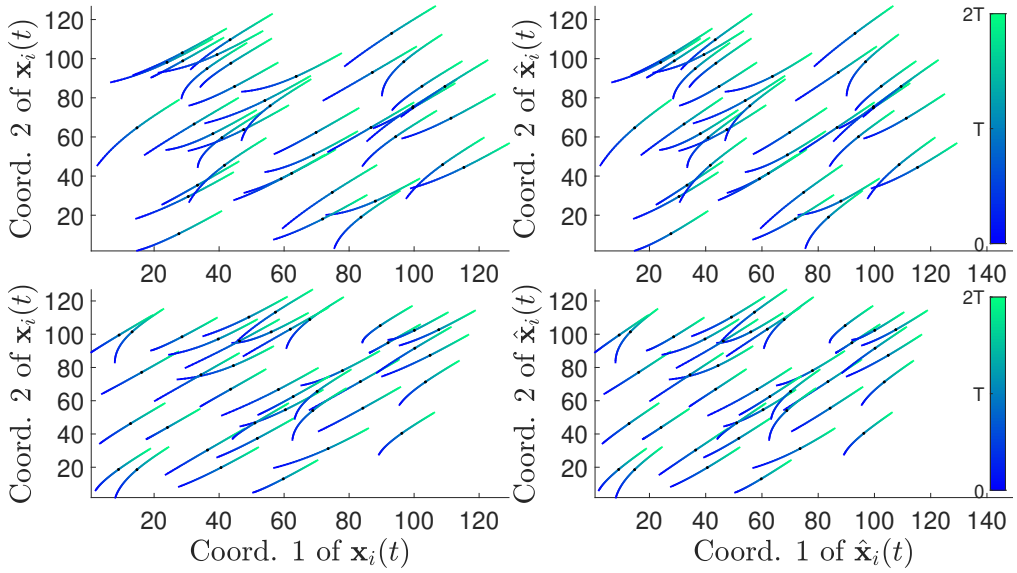


Fig. S13. (PT) Trajectories $\boldsymbol{X}(t)$ and $\widehat{\boldsymbol{X}}(t)$ obtained with true and learned interaction kernels respectively, for two randomly chosen initial conditions and evolved using the larger number of agents N_{new} (governed by the same equations as in the case of N agents).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

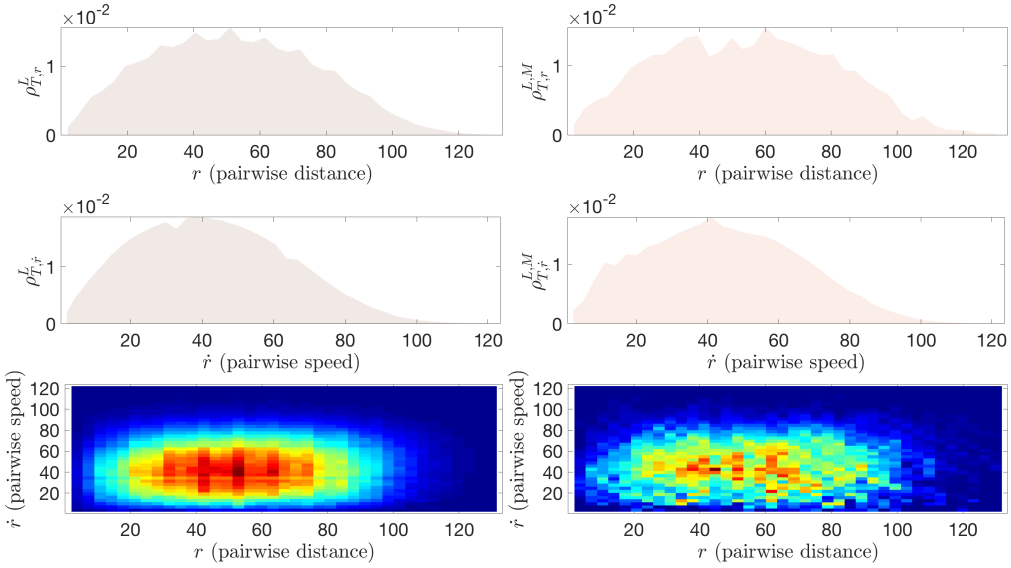
Table S17

	$[0, T]$	$[T, T_f]$
mean_{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_{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_{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_{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_{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_{IC}: Larger N	$2.9 \cdot 10^{-3} \pm 3.0 \cdot 10^{-3}$	$7.9 \cdot 10^{-3} \pm 7.0 \cdot 10^{-3}$

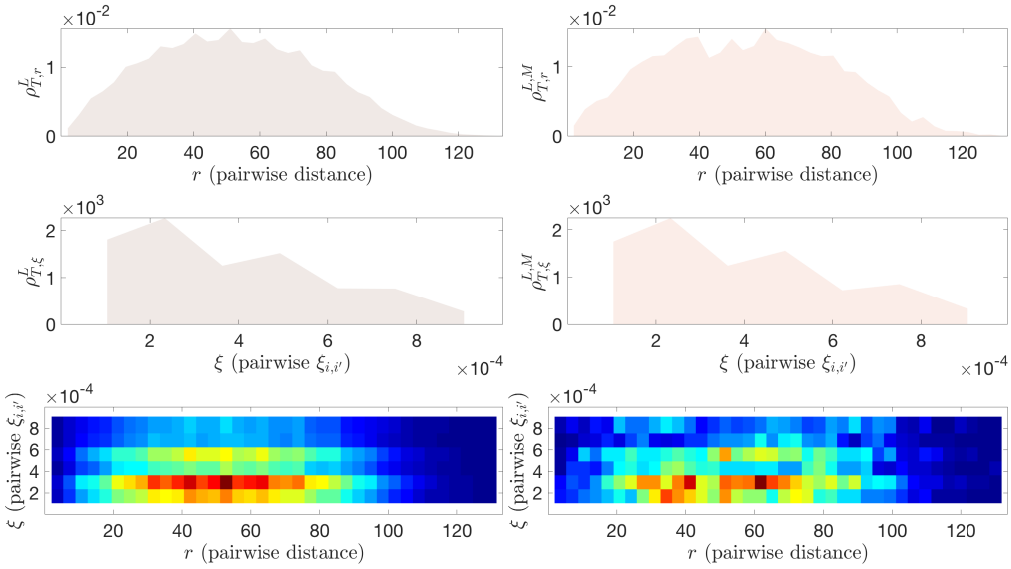
(PT) Trajectory Errors

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.

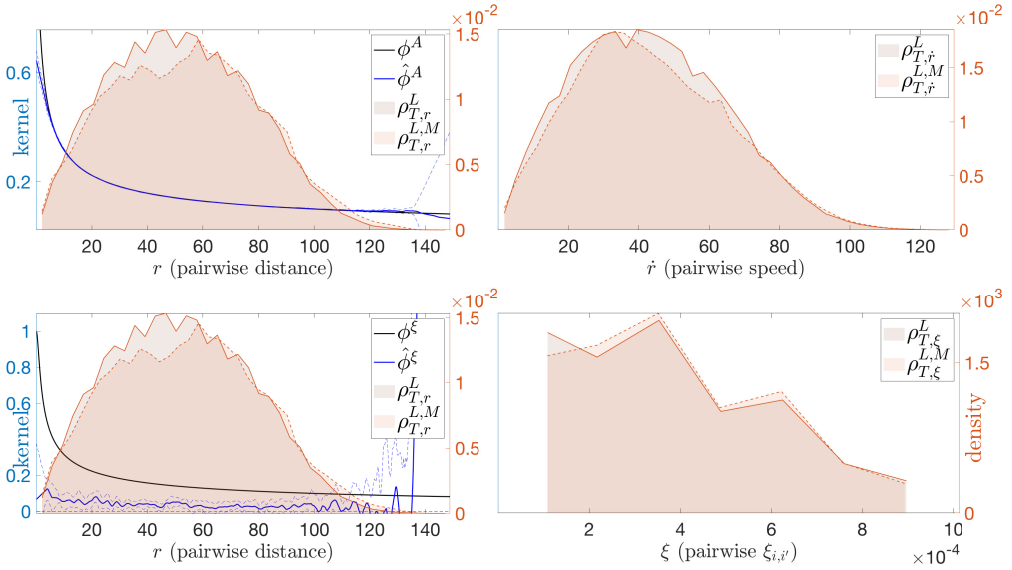


Fig. S15. (PT) Interaction kernels learned from noisy observations of positions and velocities. The noises are multiplicative, Unif.$([-\sigma,\sigma])$ with $\sigma = 0.1$ and with other parameters as in Table S15. The estimated kernel for associated with $\ddot{\boldsymbol{x}}_i$ is minimally affected, 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.

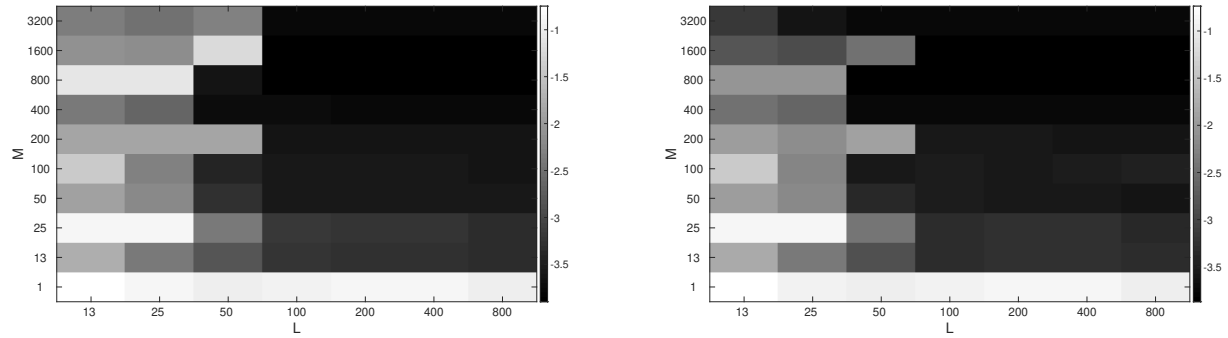


Fig. S16. (PT) Relative error, in $\log_{10}$ scale, of $\hat{\phi}^A$ (left) and $\hat{\phi}^\xi$ (right) as a function of L and M. The error decreases both in L and M, in fact roughly in the product ML. The fourth plot is an identically 0 absolute error, because both $\phi^E_{2,2}$ and its estimator 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

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

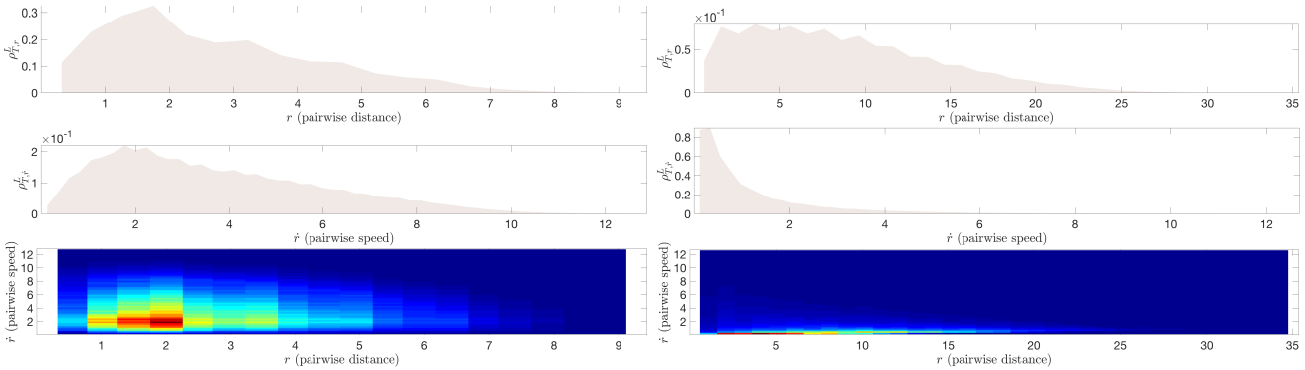
The system parameters are given in Table S18.

Table S18

d	M	L	T	$\mu_0^{\boldsymbol{X}}$	$\mu_0^{\dot{\boldsymbol{X}}}$	$n^E = n^A$	$\deg(\psi^A)$=$\deg(\psi^\xi)$
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_{IC}	$\mathbf{9.5 \cdot 10^{-3} \pm 2 \cdot 10^{-3}}$	3.9 ± 8
std_{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_{IC}	$1.6 \pm 1 \cdot 10^{-1}$	$\mathbf{1.3 \cdot 10^{-1} \pm 3 \cdot 10^{-2}}$
std_{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* 40^{th} *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.