



# Learning hydrodynamic equations for active matter from particle simulations and experiments

Rohit Supekar<sup>a,b</sup> , Boya Song<sup>b</sup>, Alasdair Hastewell<sup>b</sup> , Gary P. T. Choi<sup>b</sup> , Alexander Mietke<sup>b,1</sup> , and Jörn Dunkel<sup>b,1</sup> 

Edited by David Weitz, Harvard University, Cambridge, MA; received April 25, 2022; accepted January 12, 2023

Recent advances in high-resolution imaging techniques and particle-based simulation methods have enabled the precise microscopic characterization of collective dynamics in various biological and engineered active matter systems. In parallel, data-driven algorithms for learning interpretable continuum models have shown promising potential for the recovery of underlying partial differential equations (PDEs) from continuum simulation data. By contrast, learning macroscopic hydrodynamic equations for active matter directly from experiments or particle simulations remains a major challenge, especially when continuum models are not known a priori or analytic coarse graining fails, as often is the case for nondilute and heterogeneous systems. Here, we present a framework that leverages spectral basis representations and sparse regression algorithms to discover PDE models from microscopic simulation and experimental data, while incorporating the relevant physical symmetries. We illustrate the practical potential through a range of applications, from a chiral active particle model mimicking nonidentical swimming cells to recent microroller experiments and schooling fish. In all these cases, our scheme learns hydrodynamic equations that reproduce the self-organized collective dynamics observed in the simulations and experiments. This inference framework makes it possible to measure a large number of hydrodynamic parameters in parallel and directly from video data.

active matter | sparse regression | coarse-graining | hydrodynamic equations

Natural and engineered active matter, from cells (1), tissues (2), and organisms (3) to self-propelled particle suspensions (4, 5) and autonomous robots (6–8), exhibits complex dynamics across a wide range of length and time scales. Predicting the collective self-organization and emergent behaviors of such systems requires extensions of traditional theories that go beyond conventional physical descriptions of nonliving matter (9–11). Due to the inherent complexity of active matter interactions in multicellular communities (12, 13) and organisms (14), or even nonequilibrium chemical (15) or colloidal (4, 5, 16) systems, it becomes increasingly difficult and inefficient for humans to formulate and quantitatively validate continuum theories from first principles. A key question is therefore whether one can utilize computing machines (17) to identify interpretable systems of equations that elucidate the mechanisms underlying collective active matter dynamics.

Enabled by recent major advances in microscopic imaging (12, 14, 18, 19) and agent-based computational modeling (20), active matter systems can now be observed and analyzed at unprecedented spatiotemporal (21–23) resolution. To infer interpretable predictive theories, the high-dimensional data recorded in experiments or simulations have to be compressed and translated into low-dimensional models. Such learned models must faithfully capture the macroscale dynamics of the relevant collective properties. Macroscale properties can be efficiently encoded through hydrodynamic variables, continuous fields that are linked to the symmetries and conservation laws of the underlying microscopic system (10, 11). Although much theoretical progress has been made in the field of dynamical systems learning over the last two decades (24–30), the inference of hydrodynamic models and their parameters from particle data has remained largely unsuccessful in practice, not least due to severe complications arising from measurement noise, inherent fluctuations, and self-organized scale-selection in active systems. Yet, extrapolating the current experimental revolution (4, 5, 12, 13, 18, 19), data-driven equation learning will become increasingly more important as simultaneous observations of physical, biological, and chemical properties of individual cells and other active units will become available in the near future (31, 32).

Learning algorithms for ordinary differential equations (ODEs) and partial differential equations (PDEs) have been proposed and demonstrated based on least-squares fitting

## Significance

Active systems, from self-propelled colloids to animal swarms, can form complex dynamical patterns as their microscopic constituents exchange energy and momentum with the environment. Reflecting this complexity, continuum models for active matter typically possess many more parameters than those of classical fluids, like water. While much progress has been made in the qualitative understanding of active pattern formation, measuring the various hydrodynamic parameters of an active system still poses major challenges. Here, we present a broadly applicable computational framework that translates video data from self-propelled particle suspensions and fish swarms into continuum equations, yielding direct estimates for previously inaccessible hydrodynamic parameters and providing predictions that agree with recent experiments.

Author contributions: J.D. designed research; R.S., B.S., A.H., G.P.T.C., and A.M. performed research; and R.S., A.H., A.M., and J.D. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2023 the Author(s). Published by PNAS. This article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](#).

<sup>1</sup>To whom correspondence may be addressed. Email: amietke@mit.edu or dunkel@mit.edu.

This article contains supporting information online at <http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2206994120/-DCSupplemental>.

Published February 10, 2023.

(24, 25), symbolic regression (26, 27), and sparse regression (28, 29) combined with weak formulations (33–36), artificial neural networks (37–41), and stability selection (30, 42). These groundbreaking studies, however, focused primarily on synthetic data from a priori known continuum models, and recent coarse-graining applications have remained limited to ODEs (43) or one-dimensional (1D) PDEs (44, 45). By contrast, it is still an open challenge to infer higher-dimensional PDE models and measure hydrodynamic coefficients directly from microscopic active matter simulations or experiments.

Here, we present a comprehensive learning framework that takes microscopic particle data as input and proposes sparse interpretable hydrodynamic models as output (Fig. 1). We demonstrate its practical potential in applications to data from simulations of nonidentical active particles and recent experimental studies of microroller suspensions (4) as well as from schooling fish (46), for which the biophysical interactions are not exactly known. The learned hydrodynamic models replicate the emergent collective dynamics seen in the experimental videos, and their predictions agree with microroller experiments performed in different geometries. Furthermore, the linear hydrodynamic coefficients identified by the algorithm agree with independent estimates obtained by analytic coarse graining and active-sound spectroscopy (4). In addition, the framework identifies previously inaccessible nonlinear coefficients, providing data-informed closure relations for hydrodynamic models (10) of active matter when analytic coarse-graining procedures fail. From a broader theoretical perspective, the analysis below demonstrates how insights from analytic coarse-graining calculations and prior knowledge of conservation laws and broken symmetries can

enhance the robustness of automated equation discovery from microscopic data. From a practical perspective, the algorithms and codes are directly applicable to imaging and tracking data generated in typical active matter experiments, offering a cost-efficient method for inferring hydrodynamic coefficients from videos.

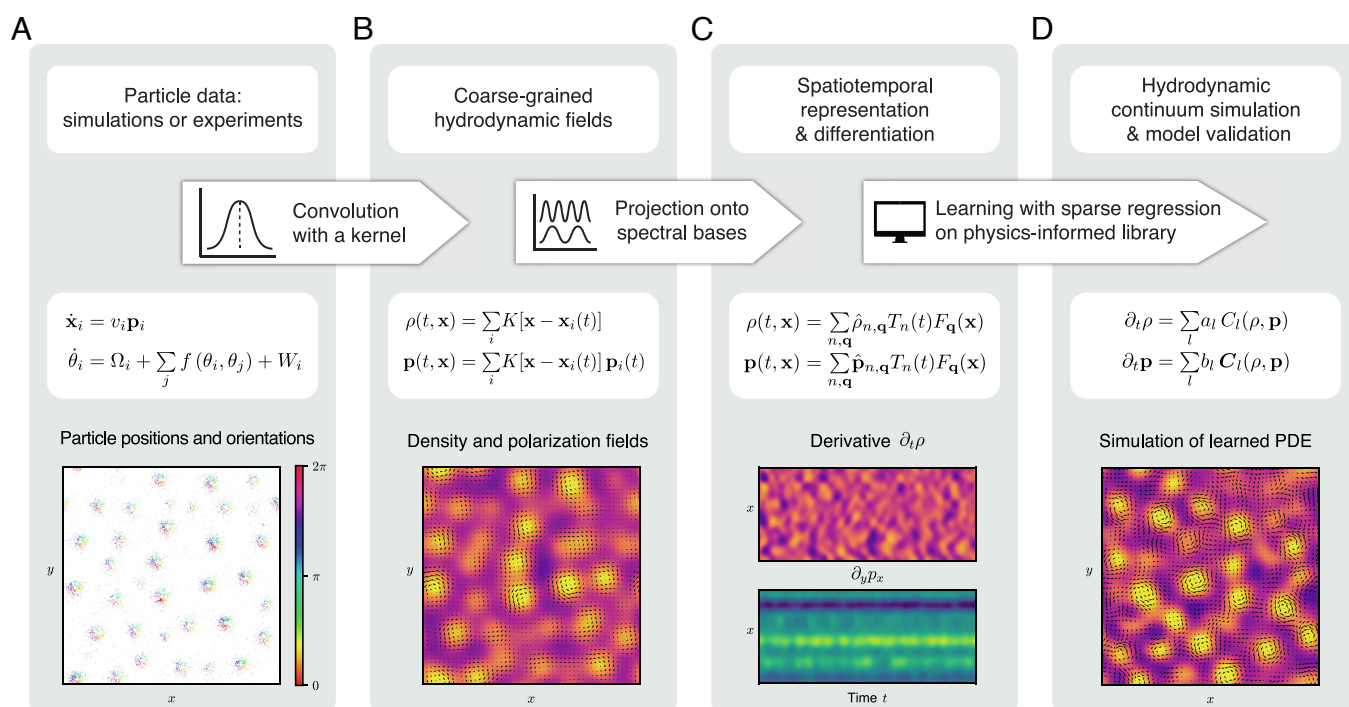
## 1. Learning Framework

**A. Active Particle Simulations.** To generate challenging test data for the learning algorithm, we simulated a 2D system of interacting self-propelled chiral particles (47–51) with broadly distributed propulsion and turning rate parameters. Microscopic models of this type are known to capture essential aspects of the experimentally observed self-organization of protein filaments (52, 53), bacterial swarms (21, 54, 55), and cell monolayers (56). In the simulations, a particle  $i$  with orientation  $\mathbf{p}_i = (\cos \theta_i, \sin \theta_i)^\top$  moved and changed orientation according to the Brownian dynamics

$$\frac{d\mathbf{x}_i}{dt} = v_i \mathbf{p}_i, \quad [1a]$$

$$\frac{d\theta_i}{dt} = \Omega_i + g \sum_{j \in \mathcal{N}_i} \sin(\theta_j - \theta_i) + \sqrt{2D_r} \eta_i, \quad [1b]$$

where  $\eta_i(t)$  denotes orientational Gaussian white noise, with zero mean and  $\langle \eta_i(t) \eta_j(t') \rangle = \delta_{ij} \delta(t - t')$ , modulated by the rotational diffusion constant  $D_r$ . The parameter  $g > 0$  determines the alignment interaction strength between particles  $i$  and



**Fig. 1.** Learning hydrodynamic models from particle simulations and experiments. (A) Inputs are time-series data for particle positions  $\mathbf{x}_i(t)$ , particle orientations  $\mathbf{p}_i(t) = (\cos \theta_i, \sin \theta_i)^\top$ , etc., measured in simulations or experiments with microscale resolution (*Active Particle Simulations*). (B) Spatial kernel coarse graining of the discrete microscopic variables provides continuous hydrodynamic fields, such as the density  $\rho(t, \mathbf{x})$  or the polarization density  $\mathbf{p}(t, \mathbf{x})$  (*Hydrodynamic Fields*). (C) Coarse-grained fields are sampled on a spatiotemporal grid and projected onto suitable spectral basis functions. Systematic spectral filtering (compression) ensures smoothly interpolated hydrodynamic fields, enabling efficient, and accurate computation of spatiotemporal derivatives (*Spatiotemporal Representation and Differentiation*). (D) Using these derivatives, a library of candidate terms  $C_l(\rho, \mathbf{p})$  and  $C_l(\rho, \mathbf{p})$  consistent with prior knowledge about conservation laws and broken symmetries is constructed. A sparse regression algorithm determines subsets of relevant phenomenological coefficients  $a_l$  and  $b_l$  (*Inference of Hydrodynamic Equations*). The resulting hydrodynamic models are sparse and interpretable, and their predictions can be directly validated against analytic coarse-graining results (*Validation and Discussion of Learned Models*) or experiments (*Learning from Experimental Data*). Bottom: Snapshots illustrating the workflow for microscopic data generated from simulations of chiral active Brownian particles in Eq. 1.

$j$  within a neighborhood  $\mathcal{N}_i$  of radius  $R$ . The self-propulsion speed  $v_i \geq 0$  and orientational rotation frequency  $\Omega_i \geq 0$  were drawn from a joint distribution  $p(v_i, \Omega_i)$  (SI Appendix, section A1). This heuristic distribution was chosen such that long-lived vortex states, similar to those observed in swimming sperm cell suspensions (57), formed spontaneously from arbitrary random initial conditions (Fig. 2A). Emerging vortices are left-handed for  $\Omega_i \geq 0$ , and their typical size is  $\sim \langle v_i \rangle_p / \langle \Omega_i \rangle_p$ , where  $\langle \cdot \rangle_p$  denotes an average over the parameter distribution  $p(v_i, \Omega_i)$ . We simulated Eq. 1 in a nondimensionalized form, choosing the interaction radius  $R$  as reference length and  $R/\langle v_i \rangle_p$  as time scale. Accordingly, we set  $R = 1$  and  $\langle v_i \rangle_p = 1$  from now on. Simulations were performed for  $N = 12,000$  particles on a periodic domain of size  $100 \times 100$  (Fig. 2A).

From a learning perspective, this model poses many of the typical challenges that one encounters when attempting to infer hydrodynamic equations from active matter experiments: spontaneous symmetry breaking and mesoscale pattern formation, microscopic parameter variability, noisy dynamics, anisotropic interactions, and so on. Indeed, similar to many experimental systems, it is not even clear a priori whether the heterogeneous active particle system described by Eq. 1 permits a description in terms of a sparse hydrodynamic continuum model, as the standard analytic coarse-graining procedure yields a physically unstable model (SI Appendix, section B1 and Fig. S5), reflecting the failure of ad hoc closure assumptions when parameters are broadly distributed. Below we will see that the learning framework is able to identify a set of hydrodynamic equations that replicate the key features of the particle simulations, including density patterns, vortex dynamics and scales, and spectral characteristics.

**B. Hydrodynamic Fields.** Given particle-resolved data, hydrodynamic fields are obtained by coarse graining. A popular coarse-graining approach is based on convolution kernels (58, 59), weight functions that translate discrete fine-grained particle densities into continuous fields, analogous to the point-spread function of a microscope. For example, given the particle positions  $\mathbf{x}_i(t)$  and orientations  $\mathbf{p}_i(t)$ , an associated particle number density field  $\rho(t, \mathbf{x})$  and polarization density field  $\mathbf{p}(t, \mathbf{x})$  can be defined by

$$\rho(t, \mathbf{x}) = \sum_i K[\mathbf{x} - \mathbf{x}_i(t)], \quad [2a]$$

$$\mathbf{p}(t, \mathbf{x}) = \sum_i K[\mathbf{x} - \mathbf{x}_i(t)] \mathbf{p}_i(t). \quad [2b]$$

The symmetric kernel  $K(\mathbf{x})$  is centered at  $\mathbf{x} = 0$  and normalized,  $\int d^2\mathbf{x} K(\mathbf{x}) = 1$ , so that the total number of particles is recovered from  $\int d^2\mathbf{x} \rho(t, \mathbf{x}) = N$ . Eqs. 2a and 2b generalize to higher tensorial density fields in a straightforward manner and can be readily adapted to accommodate different boundary conditions (SI Appendix, section A2).

We found that, in the context of hydrodynamic model learning, the coarse-graining Eqs. 2a and 2b with a Gaussian kernel  $K(\mathbf{x}) \propto \exp[-|\mathbf{x}|^2/(2\sigma^2)]$  present a useful preprocessing step that simplifies the use of fast transforms at later stages. The coarse-graining scale  $\sigma$  determines the spatial resolution of the hydrodynamic theory. In practice,  $\sigma$  must be chosen larger than the particles' mean-free path length or interaction scale to ensure smoothness of the hydrodynamic fields but also smaller than the emergent collective structures. In accordance with these requirements, we fixed  $\sigma = 5$  for the microscopic

test data from Eq. 1 (Fig. 2A and SI Appendix, Fig. S13). Interestingly, measuring the spectral entropy as a function of  $\sigma$  for both simulated and experimental data showed that coarse-grained hydrodynamic fields typically maintain only about 1% of the spectral information contained in the fine-grained particle data (SI Appendix, section E and Figs. S13 and S14).

**C. Spatiotemporal Representation and Differentiation.** A central challenge in PDE learning is the computation of spatial and temporal derivatives of the coarse-grained fields. Our framework exploits that hydrodynamic models aim to capture the long-wavelength dynamics of the slow collective modes (10). This fact allows us to project the coarse-grained fields on suitable basis functions that additionally enable sparse representations (high compression), fast transforms, and efficient differentiation. Here, we work with representations of the form

$$\rho(t, \mathbf{x}) = \sum_{n, \mathbf{q}} \hat{\rho}_{n, \mathbf{q}} T_n(t) F_{\mathbf{q}}(\mathbf{x}), \quad [3a]$$

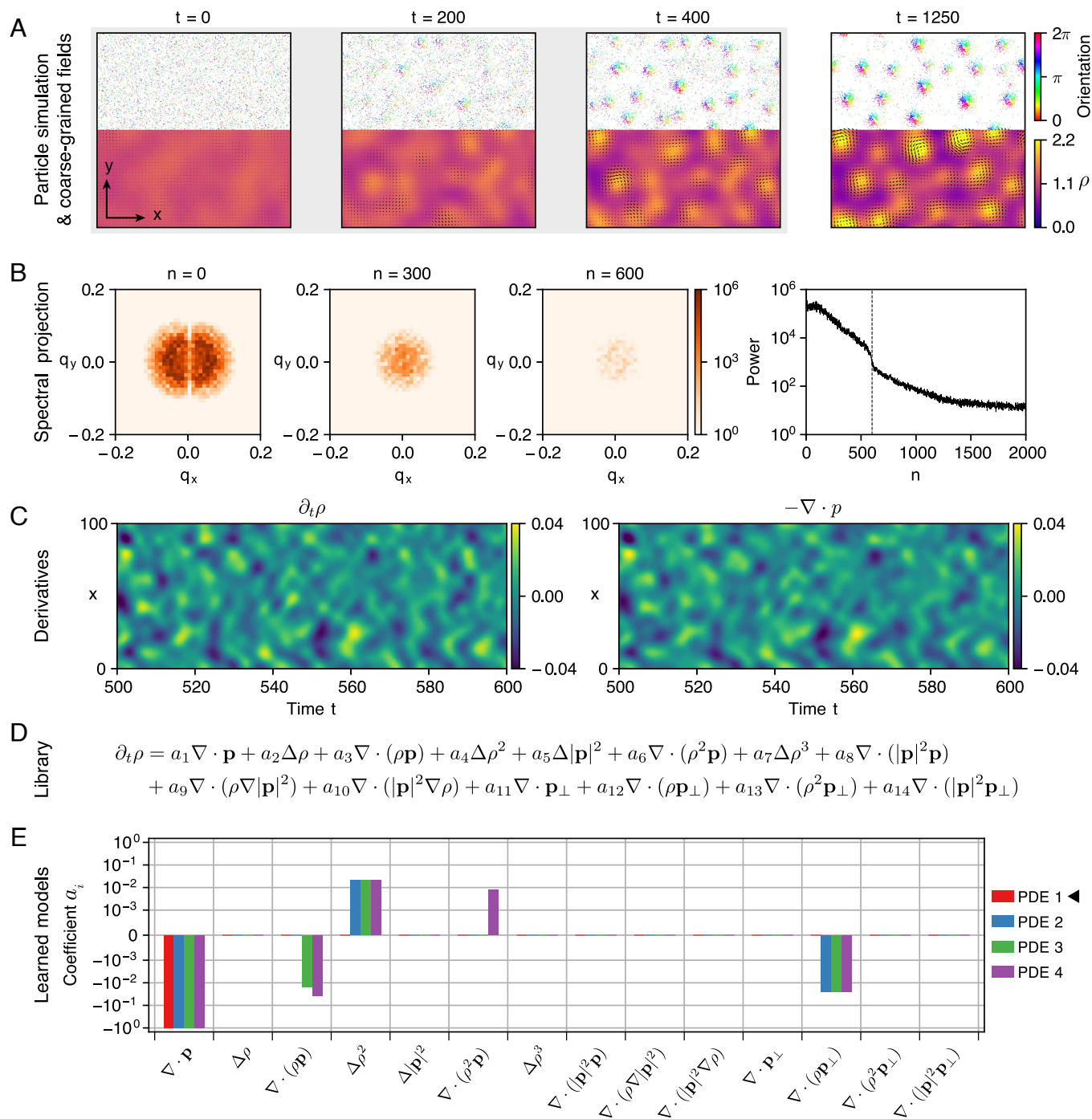
$$\mathbf{p}(t, \mathbf{x}) = \sum_{n, \mathbf{q}} \hat{\mathbf{p}}_{n, \mathbf{q}} T_n(t) F_{\mathbf{q}}(\mathbf{x}), \quad [3b]$$

where  $T_n(t)$  denotes a degree- $n$  Chebyshev polynomial of the first kind (60, 61),  $F_{\mathbf{q}}(\mathbf{x}) = \exp(2\pi i \mathbf{q} \cdot \mathbf{x})$  is a Fourier mode with wave vector  $\mathbf{q} = (q_x, q_y)^\top$ , and  $\hat{\rho}_{n, \mathbf{q}}$  and  $\hat{\mathbf{p}}_{n, \mathbf{q}}$  are complex mode coefficients (Fig. 2B). Generally, the choice of the basis functions should be adapted to the spatiotemporal boundary conditions of the microscopic data (*Learning from Experimental Data*).

The spectral representation in Eq. 3 enables the efficient and accurate computation of space and time derivatives (62). Preprocessing via spatial coarse graining (*Hydrodynamic Fields*) ensures that the mode coefficients  $\hat{\rho}_{n, \mathbf{q}}$  and  $\hat{\mathbf{p}}_{n, \mathbf{q}}$  decay fast for  $|\mathbf{q}| \gg 1/(2\pi\sigma)$  (Fig. 2B, Left). If the asymptotic decay of the mode amplitudes with the temporal mode number  $n$  is at least exponential, then deterministic PDE descriptions are sufficient, whereas algebraically decaying temporal spectra indicate that stochastic PDEs may be required to capture essential aspects of the coarse-grained dynamics. For the simulated and experimental systems considered in this work, temporal spectra were found to decay exponentially (Fig. 2B) or superexponentially (SI Appendix, Fig. S17), suggesting the existence of deterministic PDE-based hydrodynamic models. To infer such models from data, we focus on the slow hydrodynamic modes and filter out the fast modes with  $n > n_0$  by keeping only the dominant Chebyshev terms in Eq. 3. The cutoff value  $n_0$  can usually be directly inferred from a characteristic steep drop-off in the power spectrum of the data, which signals the transition to hydrodynamically irrelevant fast fluctuations (63) (Fig. 2B, Right). Choosing  $n_0$  according to this criterion yields accurate, spatiotemporally consistent derivatives as illustrated for the kymographs of the derivative fields  $\partial_t \rho$  and  $-\nabla \cdot \mathbf{p}$ , which are essential to capture mass conservation. More generally, combining kernel-based and spectral coarse graining also mitigates measurement noise, enabling a direct application to experimental data (*Learning from Experimental Data*).

**D. Inference of Hydrodynamic Equations.** To infer hydrodynamic models that are consistent with the coarse-grained projected fields from Eq. 3, we build on a recently proposed sparse regression framework (28, 29). The specific aim is to determine sparse PDEs for the density and polarization dynamics of the form





**Fig. 2.** Learning mass conservation dynamics. (A) *Top*: Time evolution of positions and orientations of 12,000 particles following the dynamics in Eqs. 2a and 2b. *Bottom*: Coarse-grained density  $\rho$  (color code) and polarization field  $\mathbf{p}$  (arrows). Starting from random initial conditions ( $t = 0$ ), a long-lived vortex pattern with well-defined handedness emerges ( $t = 1,250$ ). Training data were randomly sampled from the time window  $t \in [40, 400]$ , enclosed within the gray box. Domain size:  $100 \times 100$ . (B) Slices through the spatiotemporal power spectrum  $S_{\mathbf{q};n,\mathbf{q}} = |\mathbf{e}_{\mathbf{x}} \cdot \hat{\mathbf{p}}_{n,\mathbf{q}}|^2$  for different values of the Chebyshev polynomial order  $n \in \{0, 300, 600\}$ , corresponding to modes with increasing temporal frequencies. The rightmost panel depicts the total spatial spectral power  $\sum_{\mathbf{q}} S_{\mathbf{q};n,\mathbf{q}}$  of each Chebyshev mode  $n$ ; see Eq. 3b. The slowly decaying long tail of fast modes indicates a regime in which fluctuations dominate over a smooth signal. The cutoff  $n_0 = 600$  removes these modes, in line with the goal to learn a hydrodynamic model for the slow long-wavelength modes. (C) Kymographs of the spectral derivatives  $\partial_t \rho$  and  $-\nabla \cdot \mathbf{p}$  at  $y = 50$ , obtained from the spectrally truncated data. (D) Mass conservation in the microscopic system restricts the physics-informed candidate library to terms that can be written as divergence of a vector field. (E) Learned phenomenological coefficients  $a_i$  of PDEs with increasing complexity (decreasing sparsity) (SI Appendix, section C). PDE 1 ( $\blacktriangleleft$ ) is given by  $\partial_t \rho = a_1 \nabla \cdot \mathbf{p}$  with  $a_1 = -0.99$ . As PDE 1 is the sparsest PDE that agrees well with analytic coarse-graining results (Table 1), it is selected for the hydrodynamic model.

$$\partial_t \rho = \sum_l a_l C_l(\rho, \mathbf{p}), \quad [4a]$$

$$\partial_t \mathbf{p} = \sum_l b_l C_l(\rho, \mathbf{p}). \quad [4b]$$

Additional dynamic equations and libraries can be added to Eq. 4 if, for example, higher-rank orientational-order-parameter fields (such as a  $\mathbf{Q}$ -tensor field describing spatiotemporal nematic order; SI Appendix, section B2) are dynamically relevant and can be extracted from microscopic data. For self-propelled polar

systems, the relaxation of higher-rank hydrodynamic fields is typically fast compared to the relaxation of the polar orientation field (64). In this case, higher-rank tensorial fields are dynamically less relevant and can often be approximated by lower-rank fields and their derivatives through theoretically or empirically motivated closure relations (9, 10, 48). Accordingly, for the active particle data considered here,  $\rho$  and  $\mathbf{p}$  present a natural choice for the hydrodynamic variables in a minimal mean-field description. This rationale is supported by generic analytic coarse-graining arguments (*SI Appendix, section B*), which also suggest first-order-in-time dynamics as described by Eq. 4.

The candidate library terms  $\{C_I(\rho, \mathbf{p})\}$  and  $\{C_I(\rho, \mathbf{p})\}$  are functions of the fields and their derivatives, which can be directly evaluated using spectral representation Eq. 3 at various sample points. Eq. 4 thus define a linear system for the phenomenological coefficients  $a_I$  and  $b_I$ , and the objective is to find sparse solutions such that the resulting hydrodynamic model recapitulates the collective particle dynamics.

Hydrodynamic models for both equilibrium and nonequilibrium systems must respect the symmetries of the underlying microscopic dynamics. This requirement is a natural extension (65) of Landau-type theories for equilibrium systems, which derive hydrodynamic models from gradients of free energies that have to respect the symmetries of the underlying microscopic dynamics. However, continuum theories of nonequilibrium systems can have additional terms that are not functional derivatives of potentials, requiring more general libraries to perform model inference. Notwithstanding, prior knowledge of symmetries can greatly accelerate the inference process by placing constraints on the continuum model parameters. For example, the learning ansatz Eq. 4b already assumes global rotational invariance by using identical coefficients  $b_I$  for the  $x$  and  $y$  components of the polarization field equations. Generally, coordinate independence of hydrodynamic models demands that the dynamical fields and the library functions  $C_I$ ,  $\mathbf{C}_I$ , etc., have the correct scalar, vectorial, or tensorial transformation properties. This fact imposes stringent constraints on permissible libraries, as do microscopic conservation laws, such as particle number conservation.

**D.1. Symmetries and conservation laws: Generating a physics-informed candidate library.** Whenever prior knowledge about (broken) symmetries and conservation laws is available, it should inform the candidate library construction to ensure that the PDE learning is performed within a properly constrained model space. A useful constraint that holds in many experimental active matter systems as well as in the microscopic model in Eq. 1 arises from particle number conservation. To impose a corresponding mass conservation in the learned hydrodynamic models, we can restrict the scalar library terms  $C_I(\rho, \mathbf{p})$  in Eq. 4a to expressions that can be written as the divergence of a vector field. In this case, each term represents a different contribution to an overall mass flux, and mass conservation holds by construction for any model that will be learned. For the application considered in this work, we included fluxes up to first order in derivatives and third order in the fields (Fig. 2D). If required, such an approach can easily be generalized to other conservation laws, which then require libraries to be constructed exclusively from divergences of suitable tensors.

The active particle model in Eq. 1 describes a chiral dynamical system with intrinsic microscopic rotation rates  $\Omega_i \geq 0$ . The space of valid hydrodynamic models therefore includes PDEs in which the mirror symmetry is explicitly broken. Formally, this implies that the Levi-Civita symbol  $\epsilon_{ij}$  can be used to generate a pseudovector  $\mathbf{p}_\perp := \epsilon^\top \cdot \mathbf{p} = (-p_y, p_x)^\top$  that has to be included in the construction of the candidate libraries

$\{C_I(\rho, \mathbf{p})\}$  and  $\{\mathbf{C}_I(\rho, \mathbf{p})\}$ . The vectorial library  $\{\mathbf{C}_I(\rho, \mathbf{p})\}$  for the chiral polarization dynamics, Eq. 4b, cannot be constrained further by symmetries or conservation laws. Mechanical substrate interactions with the environment as invoked by the microscopic model in Eq. 1 and present in many active matter experiments explicitly break Galilean invariance, leading to external forces and torques whose form is not known a priori. We therefore included in Eq. 4b also vector fields that cannot be written as a divergence, such as  $\mathbf{p}_\perp$ ,  $\rho\mathbf{p}$  or  $(\mathbf{p} \cdot \nabla)\mathbf{p}$ , in our candidate library  $\{\mathbf{C}_I(\rho, \mathbf{p})\}$ .

In general, higher-order terms can be systematically constructed from the basic set of available fields and operators  $\mathcal{B} = \{\rho, \mathbf{p}, \mathbf{p}_\perp, \nabla\}$ . We illustrate the general procedure for an example library containing terms up to linear order in  $\rho$  and up to cubic order of the other terms in  $\mathcal{B}$ . The first step is to write the list of distinct rank-2 tensors

$$\mathcal{S} = \{\mathbb{I}, \mathbf{p}\mathbf{p}, \mathbf{p}\mathbf{p}_\perp, \mathbf{p}_\perp\mathbf{p}_\perp, \nabla\mathbf{p}, \nabla\mathbf{p}_\perp\}, \quad [5]$$

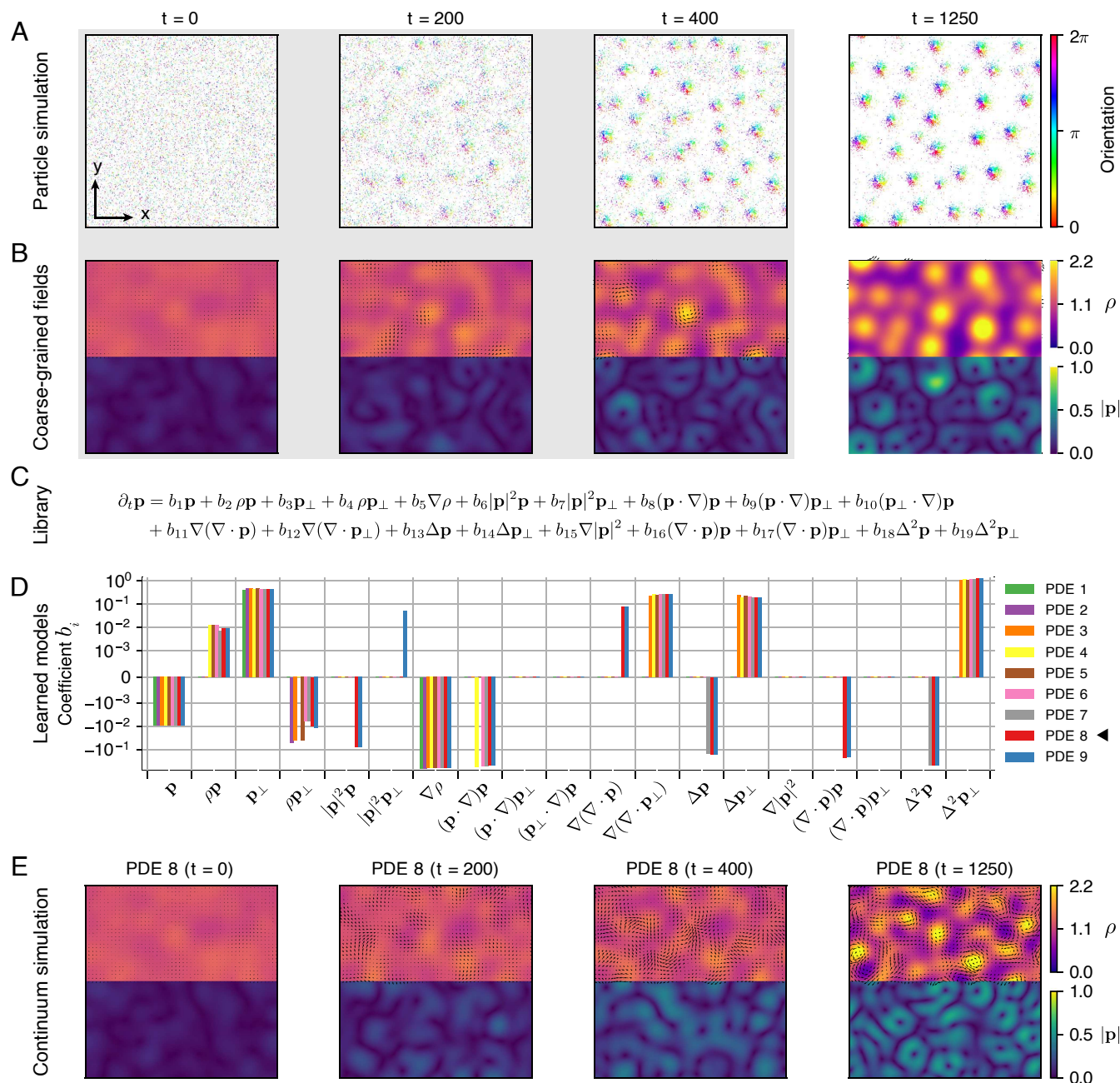
where  $s \in \{1, \rho, \nabla \cdot \mathbf{p}, \nabla \cdot \mathbf{p}_\perp\}$  represents one of the linearly independent scalars that can be formed from elements in  $\mathcal{B}$ . From any tensor  $\Sigma \in \mathcal{S}$  and its transpose, we can then generate vectorial terms  $\mathbf{C}_I$  by forming scalar products with the elements in  $\mathcal{B}$ . In particular, terms  $\nabla \cdot \Sigma$  yield possible contributions from internal stresses and torques due to alignment interactions, while  $\Sigma \cdot \mathbf{p}$  and  $\Sigma \cdot \mathbf{p}_\perp$  correspond to substrate-dependent interactions. Note that we omitted  $\epsilon$  from the set  $\mathcal{S}$ , as it yields only one additional linearly independent term  $\sim \nabla_\perp \rho$  that can be excluded for the microscopic dynamics in Eq. 1a on the basis of generic coarse-graining arguments (*SI Appendix, section B2*).

For pattern-forming systems with emergent length scale selection, the library should be extended to include Swift–Hohenberg-type (66) terms  $\Delta^2 \mathbf{p}$ ,  $\Delta^2 \mathbf{p}_\perp$ , etc. (67, 68). Such terms can stabilize small-wavelength modes and, combined with  $\Delta \mathbf{p}$  and  $\Delta \mathbf{p}_\perp$ , can give rise to patterns of well-defined length (66). The final 19-term library with linearly independent terms (*SI Appendix, section A5*) used to learn the polarization dynamics for the chiral particle model from Eq. 1 is summarized in Fig. 3C.

**D.2. Sparse model learning.** To determine the hydrodynamic parameters  $a_I$  and  $b_I$  in Eq. 4, we randomly sampled the coarse-grained fields  $\rho(t, \mathbf{x})$  and  $\mathbf{p}(t, \mathbf{x})$  and their derivatives at  $\sim 10^6$  space-time points within a predetermined learning interval (*SI Appendix, section A*). Generally, the success or failure of hydrodynamic model learning depends crucially on the choice of an appropriate space-time sampling interval. As a guiding principle, learning should be performed during the relaxation stage, when both time and space derivatives show the most substantial variation.

Evaluating Eqs. 4a and b at all sample points yields linear systems of the form  $\mathbf{U}_t = \Theta \xi$ , where the vector  $\mathbf{U}_t$  contains the time derivatives (*SI Appendix, section A3*). The columns of the matrix  $\Theta$  hold the numerical values of the library terms  $C_I(\rho, \mathbf{p})$  and  $\mathbf{C}_I(\rho, \mathbf{p})$  computed from the spectral representations in Eq. 3. The aim is to infer a parsimonious model so that the vector  $\xi$  containing the hydrodynamic parameters  $a_I$  or  $b_I$  is sparse. In this case, the corresponding PDE contains only a subset of the library terms, and we refer to the total number of terms in a PDE as its complexity.

To estimate sparse parameters  $\xi$ , we used the previously proposed sequentially thresholded least-squares (STLSQ) algorithm from SINDy (28). STLSQ first finds the least-squares estimate  $\hat{\xi} = \arg \min_{\xi} \|\mathbf{U}_t - \Theta \xi\|_2^2$ . Subsequently, sparsity of  $\hat{\xi}$  is imposed by iteratively setting coefficients below a thresholding hyperparameter  $\tau$  to zero. By construction, this regression uses data only from the bulk of the domain and



**Fig. 3.** Learning polarization dynamics. (A) Same particle dynamics as in Fig. 2A for visual reference. (B) *Top*: Coarse-grained density and polarization field as in Fig. 2A. *Bottom*: Magnitude  $|\mathbf{p}|$  of the coarse-grained polarization field. Emerging vortices ( $t=400, 1,250$ ) appear as ring-like patterns in  $|\mathbf{p}|$ . Training data were randomly sampled from the time window  $t \in [40, 400]$ , enclosed within the gray box. (C) Physics-informed candidate library (with  $b_1 = -D_T$ ) including terms constructed from  $\mathbf{p}_\perp = (-p_y, p_x)^T$ , which are allowed due to the chirality of the microscopic system. (D) Learned phenomenological coefficients  $b_i$  of PDEs with increasing complexity (SI Appendix, section C). For all PDEs, learned coefficients of the linear terms  $\mathbf{p}_\perp$  and  $\nabla \rho$  compare well with analytic predictions (Table 1 and SI Appendix, section B2). (E) Simulation of the final hydrodynamic model (PDE 8 for the polarization dynamics and PDE 1 in Fig. 2E for the density dynamics). Starting from random initial conditions ( $t = 0$ ), long-lived vortex states emerge on a similar time scale, with similar spatial patterns, and with comparable density and polarization amplitudes as in the coarse-grained microscopic model data (B). Hydrodynamic models with PDEs sparser than PDE 8 do not form stable vortex patterns.

therefore does not require information about boundary conditions. Adopting a stability selection approach (30, 42, 69, 70) in which  $\tau$  is systematically varied over a regularization path  $[\tau_{\max}, \tau_{\min}]$  (SI Appendix, section A3), we obtain candidate PDEs of increasing complexity (Figs. 2E and 3D) whose predictions need to be validated against the phenomenology of the input data.

**D.3. Performance improvements and pitfalls.** Sparse regression-based learning becomes more efficient and robust if known

symmetries or other available information can be used to reduce the number of undetermined parameters  $a_i$  and  $b_i$  in Eq. 4. Equally helpful and important is prior knowledge of the relevant time and length scales. The coarse-grained field data need to be sampled across spatiotemporal scales that contain sufficient dynamical information; oversampling in a steady-state typically prevents algorithms from learning terms relevant to the relaxation dynamics. Systems exhibiting slow diffusion time scales can pose additional challenges. For example, generic analytic coarse



graining (*SI Appendix, section B1*) shows that additive rotational noise as in Eq. 1b implies the linear term  $-D_r \mathbf{p}$  in the polarization dynamics in Eq. 4b. If the diffusive time scale  $1/D_r$  approaches or exceeds the duration of the sampling time interval, then the learned PDEs may not properly capture the relaxation dynamics of the polarization field. From a practical perspective, this is not a prohibitive obstacle, as the rotational diffusion coefficient  $D_r$  can be often measured independently from isolated single-particle trajectories (71). In this case, fixing  $-D_r \mathbf{p}$  in Eq. 4b and performing the regression over the remaining parameters produced satisfactory learning results (Fig. 3, where  $1/D_r \sim 100$  is comparable to the length of the learning interval  $t \in [40, 400]$ ).

**E. Validation and Discussion of Learned Models.** The STLSQ algorithm with stability selection proposes PDEs of increasing complexity—The final learning step is to identify the sparsest acceptable hydrodynamic model among these (Fig. 1). This can be achieved by simulating all the candidate PDEs (*SI Appendix, section A6*) and comparing their predictions against the original data and, if available, against analytic coarse-graining results (*SI Appendix, section B*).

For the microscopic particle model from Eq. 1, the sparsest learned PDE for the particle number density is  $\partial_t \rho = a_1 \nabla \cdot \mathbf{p}$  (Fig. 2E); this mass conservation equation is also predicted by analytic coarse graining (*SI Appendix, section B*). The learned coefficient  $a_1 = -0.99$  implies an effective number density flux  $-a_1 \mathbf{p} \approx \mathbf{p}$ , which agrees very well with the analytic prediction  $\langle v_i \rangle_p \mathbf{p} = \mathbf{p}$ . Additional coefficients appearing in more complex models proposed by the algorithm are at least one order of magnitude smaller than  $a_1$  (Fig. 2E). Hence, as part of a hydrodynamic description of the microscopic system in Eq. 1, we adopt the minimal density dynamics  $\partial_t \rho = a_1 \nabla \cdot \mathbf{p}$  from now on.

The sparsest learned PDE for the dynamics of the polarization field  $\mathbf{p}$  contains only three terms. However, together with the density dynamics, the resulting hydrodynamic models are either unstable or do not lead to the formation of vortex patterns. Our simulations showed that a certain level of model complexity is required to reproduce the dynamics observed in the test data. In particular, there exists a unique sparsest model (PDE 8 in Fig. 3D) for which long-lived vortex states emerge from random initial conditions. The resulting hydrodynamic model exhibits density and polarization patterns matching those observed in the original particle system (Fig. 3 A, B, and E and *SI Appendix, Movie S1*), which also form on a similar time scale. Furthermore, the learned coefficients of the linear terms  $\sim \mathbf{p}_\perp$  and  $\sim \nabla \rho$  agree well with the analytic predictions (Table 1 and *SI Appendix, section B2*). A direct comparison of temporal and spatial spectra from simulations of the learned hydrodynamic model with the coarse-grained original data shows satisfactory agreement between the characteristic length and time scales seen in each dataset (*SI Appendix, section D* and Figs. S11 and S12). Furthermore, density profiles, vortex sizes, and the disordered nature of emergent vortex patterns are also consistent between the coarse-grained particle data and the learned model (*SI Appendix, Fig. S10*), confirming that the learned model captures key features of the collective hydrodynamic modes.

The individual terms appearing in the learned hydrodynamic equations identify specific physical mechanisms that contribute to emergent pattern formation. The linear contributions are directly interpretable based on generic analytic coarse-graining arguments (*SI Appendix, section B*): The term  $b_1 \mathbf{p}$  with  $b_1 < 0$  corresponds to the lowest-order mean-field contribution of rotational diffu-

**Table 1. Parameters of the hydrodynamic model learned for the microscopic dynamics in Eq. 1 and values predicted by analytic coarse graining (*SI Appendix, section B2*)**

Term	Learned value	Analytic coarse graining
Density dynamics		
$a_1 \nabla \cdot \mathbf{p}$	$a_1 = -0.99$	$-\langle v_i \rangle_p = -1.00$
Polarization dynamics		
$b_3 \mathbf{p}_\perp$	$b_3 = 0.44$	$\langle v_i \Omega_i \rangle_p / \langle v_i \rangle_p = 0.50$
$b_5 \nabla \rho$	$b_5 = -0.60$	$-\frac{1}{2} \langle v_i^2 \rangle_p / \langle v_i \rangle_p = -0.57$

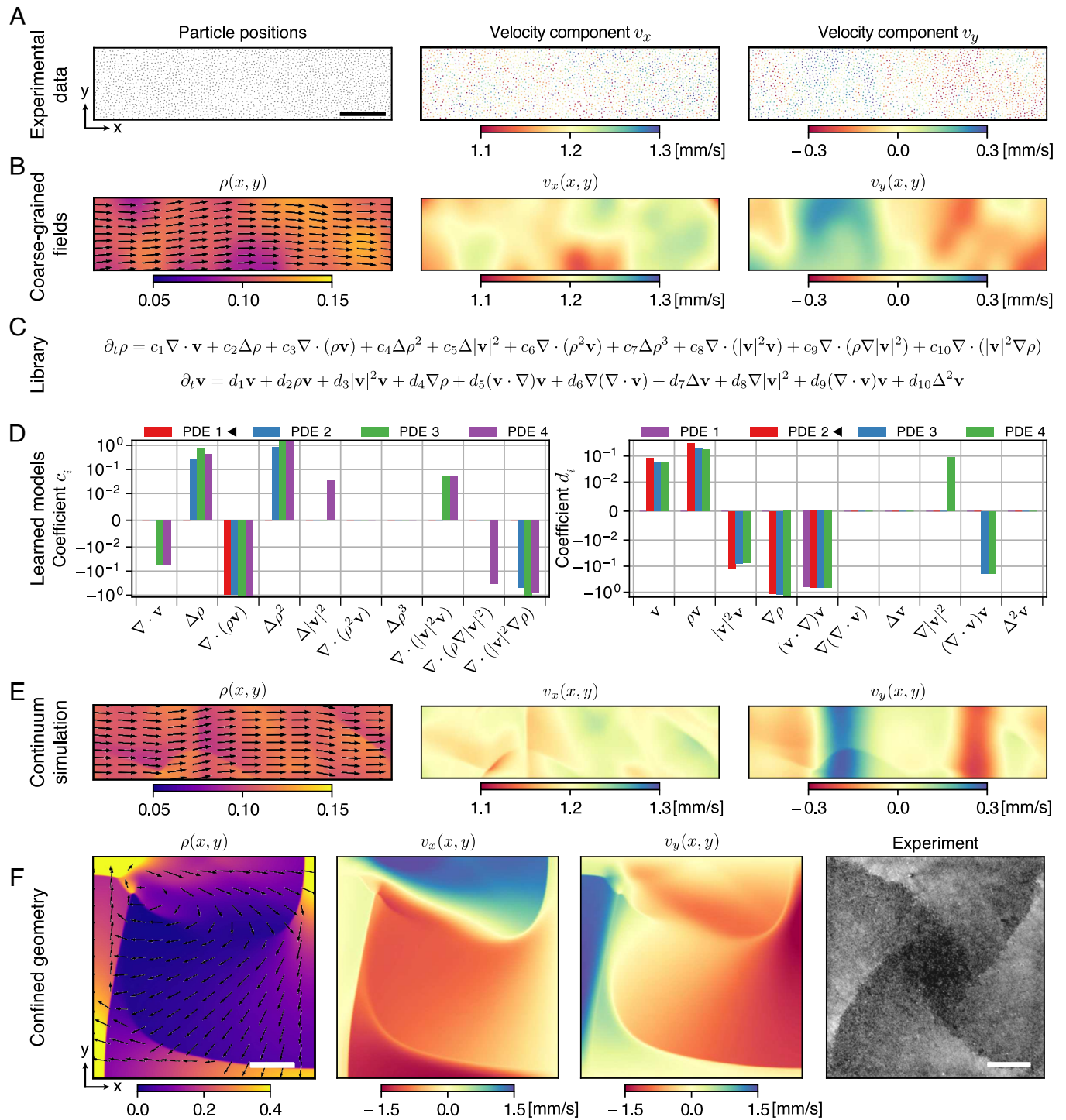
$\langle \cdot \rangle_p$  denotes averages over the distribution  $p(v_i, \Omega_i)$  of particle velocities  $v_i$  and rotation rates  $\Omega_i$  (*SI Appendix, section A1*).

sion that suppresses orientational order at long times. The chiral term  $b_3 \mathbf{p}_\perp$  with  $b_3 > 0$  drives counterclockwise rotations of the local polar field since  $\partial_t \mathbf{p} = b_3 \mathbf{p}_\perp$  is solved by the rotating vector field  $\mathbf{p} = (\cos b_3 t, \sin b_3 t)$ . This term represents the lowest-order chiral mean-field contribution to the dynamics and is a direct consequence of the active rotations  $\sim \Omega_i$  of single particles in Eq. 1b. The term  $b_7 \nabla \rho$  with  $b_7 < 0$  comes from an effective extensile isotropic stress  $\boldsymbol{\sigma} \sim -b_7 \rho \mathbb{I}$  that arises entropically in systems with moving polar particles (*SI Appendix, section B2*). The nonlinear  $\rho \mathbf{p}$  and  $|\mathbf{p}|^2 \mathbf{p}$  terms represent density-dependent polar alignment interactions, similar to ferromagnetic interactions in spin systems. Other higher-order and nonlinear terms can be identified as contributions from an effective closure relation, capturing the interplay between polar and nematic order in the particle system, or from effects of the microscopic parameter variability. (A detailed discussion is provided in *SI Appendix, section B3*.) We emphasize that, for the microscopic model in Eq. 1, standard methods for analytically deriving coarse-grained hydrodynamic equations (10, 48, 49, 55, 64, 72, 73) predict coefficients for nonlinear terms that are quantitatively different from those in our learned equations (*SI Appendix, Table SI*). Even more critically, the analytically derived model does not correctly capture vortex pattern formation, instead exhibiting locally diverging mass and polarization densities that render simulations unstable (*SI Appendix, Fig. S5*). By contrast, the learned hydrodynamic model is numerically stable and correctly reproduces the vortex formation seen in the particle simulations.

As the learning algorithm used coarse-grained field data only in the time interval  $t \in [40, 400]$ , simulation results for  $t > 400$  represent predictions of the learned hydrodynamic model (Fig. 3E). The close agreement between original data and the model simulations (Fig. 3 B and E) shows that the inference framework has succeeded in learning a previously unknown hydrodynamic description for a chiral polar active particle system with broadly distributed microscopic parameters.

## 2. Learning from Experimental Data

The inference framework can be readily applied to experimental data. We illustrate this by learning a hydrodynamic model directly from a video recorded in a recent study (4) of driven colloidal suspensions (Fig. 4A). In these experiments, an electrohydrodynamic instability enables micron-sized particles to self-propel with speeds up to a few millimeters per second across a surface. The rich collective dynamics of these so-called Quincke rollers (4, 74) provides a striking experimental realization of self-organization in active polar particle systems (10, 75, 76).



**Fig. 4.** Learning from active polar particle experiments. (A) Snapshot of particle positions and velocity components of  $\sim 2,200$  spontaneously moving Quincke rollers in a microfluidic channel (4). (Scale bar, 200  $\mu\text{m}$ .) (B) Coarse-grained density field  $\rho(t, \mathbf{x})$ , expressed as the fraction of area occupied by the rollers with diameter  $D_c = 4.8 \mu\text{m}$ , and components  $v_x(t, \mathbf{x})$  and  $v_y(t, \mathbf{x})$  of the coarse-grained velocity field ( $\sigma = 45 \mu\text{m}$ ).  $5 \times 10^5$  randomly sampled data points from  $\sim 580$  such snapshots over a time duration of 1.4 s were used for the learning algorithm. (C) Physics-informed candidate libraries for the density and velocity dynamics,  $\{\mathcal{C}_I(\rho, \mathbf{v})\}$  and  $\{\mathcal{C}_I(\rho, \mathbf{v})\}$ , respectively, Eq. 6b. These are the same libraries as shown in Figs. 2E and 3D but without the chiral terms and replacing  $\mathbf{p} \rightarrow \mathbf{v}$ . (D) Learned phenomenological coefficients  $c_i$  and  $d_i$  of the four sparsest PDEs for the density (Left) and velocity (Right) dynamics. The coefficients are nondimensionalized with length scale  $\sigma$  and time scale  $\sigma/v_0$ , where  $v_0 = 1.2 \text{ mm s}^{-1}$  is the average roller speed. PDE 1 for density dynamics corresponds to  $\partial_t \rho = c_3 \nabla \cdot (\rho \mathbf{v})$  with  $c_3 \approx -0.95$ . PDE 2 for the velocity dynamics is shown in Eq. 7b. Learned coefficients compare well with the values reported in ref. 4 (Table 2). (E) Simulation snapshot at  $t = 1.8 \text{ s}$  of the learned hydrodynamic model (PDEs marked by  $\blacktriangleleft$  in (D)) in a doubly periodic domain. Spontaneous flow emerges from random initial conditions and exhibits density and velocity fluctuations that show similar spatial patterns and amplitudes as seen in the experiments (A). (F) Simulation snapshots at  $t = 18.5 \text{ s}$  of the same hydrodynamic model as in (E) on a square domain with reflective boundary conditions. The model predicts the emergence of a vortex-like flow permeated by density shock waves. This prediction agrees qualitatively with experimental observations (Rightmost) of Quincke rollers in a  $5 \text{ mm} \times 5 \text{ mm}$  confinement with average density  $\rho_0 \approx 0.1$  (Image credits: Alexandre Morin, Delphine Geyer, and Denis Bartolo). (Scale bars, 200  $\mu\text{m}$  (simulation) and 1 mm (experiment).)



**A. Coarse Graining and Spectral Representation of Experimental Data.** To gather dynamic particle data from experiments, we extracted particle positions  $\mathbf{x}_i(t)$  from the [SI Appendix, Movie S2](#) in ref. 4, with particle velocities  $\mathbf{v}_i(t) = d\mathbf{x}_i/dt$  replacing the particle orientations  $\mathbf{p}_i(t)$  from before. This dataset captures a weakly compressible suspension of Quincke rollers in a part of a racetrack-shaped channel (Fig. 4A). We then applied the kernel coarse graining in Eq. 2 with  $\sigma = 45 \mu\text{m}$ , [SI Appendix, Fig. S14](#) to obtain the density field  $\rho$  and the velocity field  $\mathbf{v} = \mathbf{p}/\rho$ . Accounting for the nonperiodicity of the data,  $\rho$  and  $\mathbf{v}$  were projected on a Chebyshev polynomial basis, Eq. 3, in time and space (Fig. 4B). Filtering out nonhydrodynamic fast modes with temporal mode numbers  $n > n_0$ , we found that the final learning results were robust for a large range of cutoff modes  $n_0$  ([SI Appendix, section C](#)).

**B. Physics-Informed Library.** The goal is to learn a hydrodynamic model of the form

$$\partial_t \rho = \sum_l c_l \bar{C}_l(\rho, \mathbf{v}), \quad [6a]$$

$$\partial_t \mathbf{v} = \sum_l d_l \bar{\mathbf{C}}_l(\rho, \mathbf{v}), \quad [6b]$$

where  $\bar{C}_l(\rho, \mathbf{v})$  and  $\bar{\mathbf{C}}_l(\rho, \mathbf{v})$  denote library terms with coefficients  $c_l$  and  $d_l$ , respectively. The experimental Quincke roller system shares several key features with the particle model in Eq. 1, so the construction of the candidate libraries  $\{\bar{C}_l(\rho, \mathbf{v})\}$  and  $\{\bar{\mathbf{C}}_l(\rho, \mathbf{v})\}$  follows similar principles (Fig. 4C). Conservation of the particle number implies that  $\bar{C}_l$  can be written as divergences of vector fields. However, rollers do not explicitly break mirror symmetry, so chiral terms can be dropped from the  $\{\bar{\mathbf{C}}_l(\rho, \mathbf{v})\}$  library, leaving the candidate terms shown in Fig. 4C.

**C. Learned Hydrodynamic Equations and Validation.** The sparse regression algorithm proposed a hierarchy of hydrodynamic models with increasing complexity (Fig. 4D). The sparsest learned model that recapitulates the experimental observations is given by

$$\partial_t \rho = c_3 \nabla \cdot (\rho \mathbf{v}), \quad [7a]$$

$$\partial_t \mathbf{v} = d_1 \mathbf{v} + d_2 \rho \mathbf{v} + d_3 |\mathbf{v}|^2 \mathbf{v} + d_4 \nabla \rho + d_5 (\mathbf{v} \cdot \nabla) \mathbf{v}. \quad [7b]$$

Notably, Eqs. 7a and 7b contain all the relevant terms to describe the propagation of underdamped sound waves, a counterintuitive, but characteristic feature of overdamped active polar particle systems (4).

Although the finite experimental observation window and imperfect particle tracking was expected to limit the accuracy of the learned models, the learned coefficient values agree well with corresponding parameters estimated in ref. 4 by fitting a linearized Toner-Tu model to the experimental data (Table 2). The coefficient  $c_3 \simeq -0.95$  in the mass conservation equation is close to the theoretically expected value  $-1$ . The learned coefficient  $d_4$  in the velocity Eq. 7b is of similar magnitude but slightly less negative than the dispersion-based estimate in ref. 4. The learned coefficients  $d_1$ ,  $d_2$ , and  $d_3$  are described in [SI Appendix, Table SVI](#). Despite being inferred from a single video, these parameters yield a remarkably accurate prediction  $v_0(\rho_0) = \sqrt{-(d_1 + d_2 \rho_0)/d_3}$  for the typical roller speed as a function of the area fraction  $\rho_0$  ([SI Appendix, Fig. S4](#) in ref. 4 and Fig. 5). Similarly, the learned coefficient  $d_5$  of the nonlinear

**Table 2. Parameters of the learned hydrodynamic model for the Quincke roller system are close to values expected from analytic coarse graining (\*) and reported in ref. 4 for experiments performed at mean area fraction  $\rho_0 \approx 0.11$**

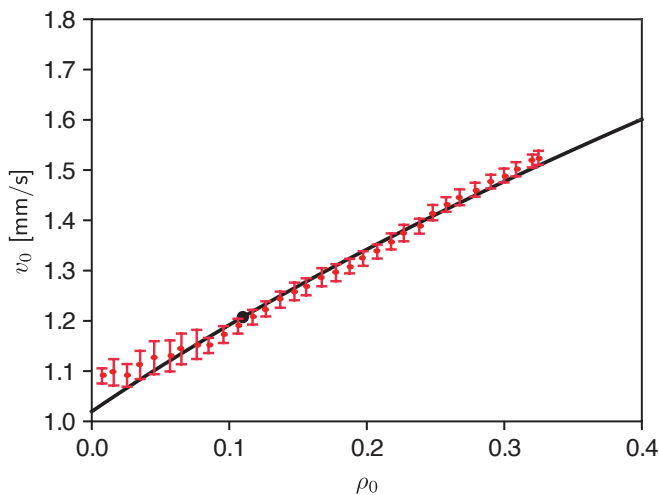
Term	Learned values	Ref. 4
Density dynamics		
$c_3 \nabla \cdot (\rho \mathbf{v})$	$c_3 = -0.95$	$-1.0^*$
Velocity dynamics		
$(d_1 + d_2 \rho) \mathbf{v}$	$\sqrt{\frac{d_1 + d_2 \rho_0}{-d_3}} = 1.21 \text{ mm/s}$	$1.20 \text{ mm/s}$
$+d_3  \mathbf{v} ^2 \mathbf{v}$	$d_4 = -1.62 \text{ mm}^2/\text{s}^2$	$-5.0 \pm 2.0 \text{ mm}^2/\text{s}^2$
$d_4 \nabla \rho$	$d_5 = -0.67$	$-0.7 \pm 0.1$
$d_5 (\mathbf{v} \cdot \nabla) \mathbf{v}$		

advective term  $\sim (\mathbf{v} \cdot \nabla) \mathbf{v}$  is in close agreement with the value reported in ref. 4. Interestingly,  $d_5 \neq -1$  reveals the broken Galilean invariance (9, 10) due to fluid-mediated roller-substrate interaction, a key physical aspect of the experimental system that is robustly discovered by the hydrodynamic model learning framework.

To validate the learned hydrodynamic model, we simulated Eq. 7 on a periodic domain comparable to the experimental observation window (Fig. 4E and [SI Appendix, section A6](#)). Starting from random initial conditions, spontaneously flowing states emerge ([SI Appendix, Movie S2](#)), even though the spontaneous onset of particle flow is not a part of the experimental data from which the model was learned. The emergent density and flow patterns are quantitatively similar to the experimentally observed ones. In particular, the learned model predicts the formation of transverse velocity bands as seen in the experiments (Fig. 4 B and E).

**D. Predicting Collective Roller Dynamics in Confinement.** Useful models must be able to make predictions for a variety of experimental conditions. At minimum, if a learned hydrodynamic model captures the most relevant physics of an active system, then it should remain valid in different geometries and boundary conditions. To confirm this for the Quincke system, we simulated Eq. 7 on a square domain using no-flux and shear-free boundary conditions ([SI Appendix, section A6](#)). Starting from random initial conditions, our learned model predicts the formation of a vortex-like flow, permeated by four interwoven density shock waves, which arise from reflections at the boundary (Fig. 4 F, Left and [SI Appendix, Movie S3](#)). Remarkably, this behavior has indeed been observed in experiments (74) in which Quincke rollers were confined within a square domain (Fig. 4 F, Right). These results demonstrate the practical potential of automated model learning for complex active matter systems. As an additional demonstration, we present in [SI Appendix, section F](#) an application of the above learning framework to recent fish schooling experiments (46). In this case, the exact nature of the underlying fish-fish interactions, which likely involve both hydrodynamic (77, 78) and visual (79, 80) cues, is not exactly known. Interestingly, the inference algorithm identifies a sparse hydrodynamic model ([SI Appendix, Table SVIII](#)) that is structurally similar to the Quincke system ([SI Appendix, Table SVI](#)), despite a vast difference in scales.

**E. Limitations and Outlook.** Any learning or inference framework is fundamentally limited by its underlying model space. In neural network (NN)-based machine learning schemes (39, 81–83), the NN architecture is prescribed by the human modeler,



**Fig. 5.** The learned model accurately predicts collective Quincke roller speeds  $v_0$  at different average area fractions  $\rho_0$ . Although Eq. 7 was learned from a single experiment (*SI Appendix, Movie S2* in ref. 4) at fixed average area fraction  $\rho_0 = 0.11$  (filled black circle), the model prediction  $v_0(\rho_0) = \sqrt{-(d_1 + d_2\rho_0)/d_3}$  (solid line) with inferred parameters  $d_1, d_2, d_3$  (*SI Appendix, Table SVI*), agrees well the experimentally measured speed values (red symbols) reported in *SI Appendix, Fig. S4* of ref. 4.

whose heuristically plausible yet ad hoc choices limit the range of predictable phenomena. In continuum dynamical systems inference (26, 28, 37, 84–86), as considered here, the spatiotemporal evolution of tensorial order-parameter fields is parameterized in terms of partial differential equations, and the range of predictable phenomena is limited by the choice of the library. Independent of whether one prefers NN-based, PDE-based, or other approaches, once the model space and its parameterization have been fixed, “learning” reduces to solving a high-dimensional (usually nonconvex) optimization problem. While NN models prioritize expressive power over interpretability, PDE-based models tend to be more easily interpretable and amenable to symmetry constraints, but their expressive power is inherently limited by the “domain knowledge” that informs library selection. An interesting approach toward reducing bias and increasing flexibility in model formulation are symbolic regression techniques (87–89) that could, in principle, discover novel classes of equations. Unfortunately, symbolic regression comes with high computational complexity, so that human-informed modeling frameworks will remain practically relevant in the physical and life sciences.

### 3. Discussion & Conclusions

Leveraging spectral representations of field observables and recent advances in the sparse PDE inference (28–30, 42), we have presented a PDE learning framework that robustly identifies hydrodynamic models for the self-organized dynamics of active matter systems. To illustrate its broad practical potential and applicability, we demonstrated the automated inference of interpretable hydrodynamic models from microscopic simulation data as well as from experimental video data for active and living systems (*SI Appendix, section F*). The underlying computational framework complements modern machine learning approaches, including model-free methods (90, 91) and others that leverage a priori known model structure to predict complex dynamics (39, 41, 59, 81), infer specific model parameters (92) or hidden fields (82), partially replace PDE models with suitably

trained neural networks (40, 93), or use them for dimensionality reduction (94, 95).

Inferring sparse hydrodynamic models from coarse-grained active matter data also complements analytic coarse-graining techniques (10, 48, 49, 55, 64, 72, 73), which generally require ad hoc moment closures to truncate infinite hierarchies of coupled mode equations (*SI Appendix, section B*). Such closures typically neglect correlations and rely on approximations that may not be valid in heterogeneous interacting active matter systems. Automated learning of hydrodynamic equations instead yields data-informed closure relations, while simultaneously providing quantitative measurements of phenomenological coefficients (viscosities, elastic moduli, etc.) from video data (92). We have shown here that this data-driven approach yields well-defined, numerically stable continuum models in situations where analytic coarse-graining methods lead to hydrodynamic equations that do not reproduce the observed patterns and instead generate locally diverging density patterns (*SI Appendix, Fig. S5*).

Successful model learning requires both good data and a good library. Good data need to sample all dynamically relevant length and time scales (84). A good library is large enough to include all hydrodynamically relevant terms and small enough to enable robust sparse regression (30). Since the number of possible terms increases combinatorially with the number of fields and differential operators, library construction should be guided by prior knowledge of global, local, and explicitly broken symmetries. From a physics perspective, using symmetry considerations as a key guiding principle to construct phenomenological models naturally builds on Ginzburg–Landau-type approaches to nonequilibrium pattern formation (65, 96). Here, this approach enabled us to infer quantitative hydrodynamic models directly from particle data, consistent with symmetry constraints arising from the microscopic dynamics. From an algorithmic perspective, physics-informed libraries ensure properly constrained model search spaces, promising a more efficient sparse regression. Equally important is the use of suitable spectral field representations—without these, an accurate evaluation of the library terms seems nearly impossible even for very-high quality data.

In view of the above successful applications, which encompassed microscopic parameter variability, explicitly noisy particle dynamics, and measurement noise in experimental data, we expect that the computational framework presented here can be directly applied to a wide variety of passive and active matter systems. Specifically, the fish schooling (46) example (*SI Appendix, section F*) demonstrates that automated model inference can yield predictive continuum models even when the biophysical particle–particle interactions are highly complex (77–80) or not yet exactly known. The computational framework presented here can likely be enhanced by combining recent advances in sparse regression (42, 97) and weak formulations (35) with statistical information criteria (85) and cross-validation (98) for model selection. Furthermore, an extension to three dimensions is conceptually and computationally straightforward: Kernel-based coarse-graining, spectral data representation, the implementation of conservation laws through suitable restrictions of library terms, and the sparse regression scheme all extend naturally to higher dimensions in a parallelizable manner. Given the rapid progress in experimental imaging and tracking techniques (12, 14, 18, 19, 46), we anticipate that many previously intractable physical and biological systems will soon find interpretable quantitative continuum descriptions that may reveal novel ordering and self-organization principles.

**Data, Materials, and Software Availability.** All study data are included in the article and/or *SI Appendix*. The codes and data required to generate Quincke Roller results are provided in this GitHub repository: <https://github.com/rohitsupekar/learning-active-matter-equations>. Due to size constraints, any other data are available from authors upon reasonable request.

**ACKNOWLEDGMENTS.** We thank Keaton Burns for helpful advice on the continuum simulations, Henrik Ronellenfisch for insightful discussions about learning methodologies, and the MIT SuperCloud for providing us access to HPC resources. We thank Tristan Walter and Iain Couzin for sharing and explaining the sunbleak data. This work was supported by a MathWorks Engineering Fellowship (R.S.), a Graduate Student Appreciation Fellowship from the MIT

Mathematics Department (B.S.), a NSF Mathematical Sciences Postdoctoral Research Fellowship (DMS-2002103, G.P.T.C.), a Longterm Fellowship from the European Molecular Biology Organization (ALTF 528-2019, A.M.), a Postdoctoral Research Fellowship from the Deutsche Forschungsgemeinschaft (Project 431144836, A.M.), a Complex Systems Scholar Award from the James S. McDonnell Foundation (J.D.), NSF Award DMS-1952706 (J.D.), and the Robert E. Collins Distinguished Scholarship Fund (J.D.).

Author affiliations: <sup>a</sup>Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; and <sup>b</sup>Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139

1. D. T. Tambe *et al.*, Collective cell guidance by cooperative intercellular forces. *Nat. Mater.* **10**, 469–475 (2011).
2. C. P. Heisenberg, Y. Bellaïche, Forces in tissue morphogenesis and patterning. *Cell* **153**, 948–962 (2013).
3. M. Tennenbaum, Z. Liu, D. Hu, A. Fernandez-Nieves, Mechanics of fire ant aggregations. *Nat. Mater.* **15**, 54–59 (2016).
4. D. Geyer, A. Morin, D. Bartolo, Sounds and hydrodynamics of polar active fluids. *Nat. Mater.* **17**, 789–793 (2018).
5. V. Soni *et al.*, The odd free surface flows of a colloidal chiral fluid. *Nat. Phys.* **15**, 1188–1194 (2019).
6. M. Rubenstein, A. Cornejo, R. Nagpal, Programmable self-assembly in a thousand-robot swarm. *Science* **345**, 795–800 (2014).
7. L. M. Nash *et al.*, Topological mechanics of gyroscopic metamaterials. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 14495–14500 (2015).
8. W. Savoie *et al.*, A robot made of robots: Emergent transport and control of a smarticle ensemble. *Sci. Robot.* **4**, eaax4316 (2019).
9. J. Toner, Y. Tu, Long-range order in a two-dimensional dynamical XY model: How birds fly together. *Phys. Rev. Lett.* **75**, 4326–4329 (1995).
10. M. C. Marchetti *et al.*, Hydrodynamics of soft active matter. *Rev. Mod. Phys.* **85**, 1143–1189 (2013).
11. F. Jülicher, S. W. Grill, G. Salbreux, Hydrodynamic theory of active matter. *Rep. Prog. Phys.* **81**, 076601 (2018).
12. R. Hartmann *et al.*, Emergence of three-dimensional order and structure in growing biofilms. *Nat. Phys.* **15**, 251–256 (2019).
13. Y. Li *et al.*, Volumetric compression induces intracellular crowding to control intestinal organoid growth via wnt/ $\beta$ -catenin signaling. *Cell Stem Cell* **28**, 63–78 (2021).
14. G. Shah *et al.*, Multi-scale imaging and analysis identify pan-embryo cell dynamics of germlayer formation in zebrafish. *Nat. Commun.* **10**, 5753 (2019).
15. N. J. Ciria, A. Benusioglio, M. Prakash, Vapour-mediated sensing and motility in two-component droplets. *Nature* **519**, 446–450 (2015).
16. W. B. Rogers, W. M. Shih, V. N. Manoharan, Using DNA to program the self-assembly of colloidal nanoparticles and microparticles. *Nat. Rev. Mater.* **1**, 16008 (2016).
17. F. Cichos, K. Gustavsson, B. Mehlig, G. Volpe, Machine learning for active matter. *Nat. Mach. Intell.* **2**, 94–103 (2020).
18. E. H. K. Stelzer, Light-sheet fluorescence microscopy for quantitative biology. *Nat. Meth.* **12**, 23–26 (2015).
19. R. M. Power, J. Huisken, A guide to light-sheet fluorescence microscopy for multiscale imaging. *Nat. Meth.* **14**, 360–373 (2017).
20. M. R. Shaebani, A. Wysocki, R. G. Winkler, G. Gompper, H. Rieger, Computational models for active matter. *Nat. Rev. Phys.* **2**, 181–199 (2020).
21. H. Jeckel *et al.*, Learning the space-time phase diagram of bacterial swarm expansion. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 1489–1494 (2019).
22. B. Qin *et al.*, Cell position fates and collective fountain flow in bacterial biofilms revealed by light-sheet microscopy. *Science* **369**, 71–77 (2020).
23. R. Hartmann *et al.*, Quantitative image analysis of microbial communities with BiofilmQ. *Nat. Microbiol.* **6**, 151–156 (2021).
24. D. P. Vallette, G. Jacobs, J. P. Gollub, Oscillations and spatiotemporal chaos of one-dimensional fluid fronts. *Phys. Rev. E* **55**, 4274–4287 (1997).
25. M. Bär, R. Hegger, H. Kantz, Fitting partial differential equations to space-time dynamics. *Phys. Rev. E* **59**, 337–342 (1999).
26. J. Bongard, H. Lipson, Automated reverse engineering of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9943–9948 (2007).
27. M. Schmidt, H. Lipson, Distilling free-form natural laws. *Science* **324**, 81–86 (2009).
28. S. L. Brunton, J. L. Proctor, J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 3932–3937 (2016).
29. S. H. Rudy, S. L. Brunton, J. L. Proctor, J. N. Kutz, Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614 (2017).
30. S. Maddu, B. L. Cheeseman, C. L. Müller, I. F. Szbalzarini, Learning physically consistent differential equation models from data using group sparsity. *Phys. Rev. E* **103**, 042310 (2021).
31. C. Linghu *et al.*, Spatial multiplexing of fluorescent reporters for imaging signaling network dynamics. *Cell* **183**, 1682–1698.e24 (2020).
32. N. Cermak *et al.*, Whole-organism behavioral profiling reveals a role for dopamine in state-dependent motor program coupling in *C. elegans*. *eLife* **9**, e57093 (2020).
33. P. A. K. Reinbold, R. O. Grigoriev, Data-driven discovery of partial differential equation models with latent variables. *Phys. Rev. E* **100**, 022219 (2019).
34. D. R. Gurevich, P. A. K. Reinbold, R. O. Grigoriev, Robust and optimal sparse regression for nonlinear PDE models. *Chaos* **29**, 103113 (2019).
35. P. A. Reinbold, D. R. Gurevich, R. O. Grigoriev, Using noisy or incomplete data to discover models of spatiotemporal dynamics. *Phys. Rev. E* **101**, 010203(R) (2020).
36. P. A. K. Reinbold, L. M. Kageorge, M. F. Schatz, R. O. Grigoriev, Robust learning from noisy, incomplete, high-dimensional experimental data via physically constrained symbolic regression. *Nat. Commun.* **12**, 3219 (2021).
37. K. Champion, B. Lusch, J. Nathan Kutz, S. L. Brunton, Data-driven discovery of coordinates and governing equations. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 22445–22451 (2019).
38. G. J. Both, S. Choudhury, P. Sens, R. Kusters, Deepmod: Deep learning for model discovery in noisy data. *J. Comput. Phys.* **428**, 109985 (2020).
39. M. Raissi, P. Perdikaris, G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).
40. C. Rackauckas *et al.*, Universal differential equations for scientific machine learning (2020).
41. S. Shankar *et al.*, “Learning non-linear spatio-temporal dynamics with convolutional neural odes” in *NeurIPS ML4PS Workshop* (2020).
42. S. Maddu, B. L. Cheeseman, I. F. Szbalzarini, C. L. Müller, Stability selection enables robust learning of differential equations from limited noisy data. *Proc. R. Soc. Lond.* **478**, 20210916 (2022).
43. J. T. Nardini, R. E. Baker, M. J. Simpson, K. B. Flores, Learning differential equation models from stochastic agent-based model simulations. *J. R. Soc. Interface* **18**, 20200987 (2021).
44. L. Felsberger, P. S. Koutsourelakis, Physics-constrained, data-driven discovery of coarse-grained dynamics. *Commun. Comput. Phys.* **25**, 1259–1301 (2019).
45. J. Bakarij, D. M. Tartakovsky, Data-driven discovery of coarse-grained equations. *J. Comput. Phys.* **434**, 110219 (2021).
46. T. Walter, I. D. Couzin, TRex, a fast multi-animal tracking system with markerless identification, and 2D estimation of posture and visual fields. *eLife* **10**, e64000 (2021).
47. F. Peruani, A. Deutsch, M. Bär, A mean-field theory for self-propelled particles interacting by velocity alignment mechanisms. *Eur. Phys. J. Spec. Top.* **157**, 111–122 (2008).
48. F. D. C. Farrell, M. C. Marchetti, D. Marenduzzo, J. Tailleur, Pattern formation in self-propelled particles with density-dependent motility. *Phys. Rev. Lett.* **108**, 248101 (2012).
49. B. Liebchen, M. E. Cates, D. Marenduzzo, Pattern formation in chemically interacting active rotors with self-propulsion. *Soft Matter* **12**, 7259–7264 (2016).
50. B. Liebchen, D. Levis, Collective behavior of chiral active matter: Pattern formation and enhanced flocking. *Phys. Rev. Lett.* **119**, 058002 (2017).
51. N. Kruk, J. A. Carrillo, H. Koeppl, Traveling bands, clouds, and vortices of chiral active matter. *Phys. Rev. E* **102**, 22604 (2020).
52. Y. Sumino *et al.*, Large-scale vortex lattice emerging from collectively moving microtubules. *Nature* **483**, 448–452 (2012).
53. L. Huber, R. Suzuki, T. Krüger, E. Frey, A. R. Bausch, Emergence of coexisting ordered states in active matter systems. *Science* **361**, 255–258 (2018).
54. H. Li *et al.*, Data-driven quantitative modeling of bacterial active nematics. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 777–785 (2019).
55. H. Chaté, Dry aligning dilute active matter. *Annu. Rev. Condens. Matter Phys.* **11**, 189–212 (2020).
56. F. Giavazzi *et al.*, Giant fluctuations and structural effects in a flocking epithelium. *J. Phys. D: Appl. Phys.* **50**, 384003 (2017).
57. I. H. Riedel, K. Kruse, J. Howard, A self-organized vortex array of hydrodynamically entrained sperm cells. *Science* **309**, 300–303 (2005).
58. A. P. Solon, J. Stenhammar, M. E. Cates, Y. Kafri, J. Tailleur, Generalized thermodynamics of phase equilibria in scalar active matter. *Phys. Rev. E* **97**, 020602 (2018).
59. E. Wallin, M. Servin, Data-driven model order reduction for granular media. *Comput. Part. Mech.* **9**, 15–28 (2022).
60. J. P. Boyd, *Chebyshev and Fourier Spectral Methods* (Courier Corporation, 2001).
61. J. C. Mason, D. C. Handscomb, *Chebyshev Polynomials* (CRC Press, 2002).
62. O. Bruno, D. Hoch, Numerical differentiation of approximated functions with limited order-of-accuracy deterioration. *SIAM J. Numer. Anal.* **50**, 1581–1603 (2012).
63. J. L. Aurentz, L. N. Trefethen, Chopping a Chebyshev series. *ACM Trans. Math. Softw.* **43** (2017).
64. E. Bertin, M. Droz, G. Grégoire, Hydrodynamic equations for self-propelled particles: Microscopic derivation and stability analysis. *J. Phys. A* **42**, 445001 (2009).
65. M. Fruchart, R. Hanai, P. B. Littlewood, V. Vitelli, Non-reciprocal phase transitions. *Nature* **592**, 363–369 (2021).
66. M. Cross, H. Greenside, *Pattern Formation and Dynamics in Nonequilibrium Systems* (Cambridge University Press, Cambridge, 2009).
67. H. H. Wensink *et al.*, Meso-scale turbulence in living fluids. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14308–14313 (2012).
68. M. James, W. J. Bos, M. Wilczek, Turbulence and turbulent pattern formation in a minimal model for active fluids. *Phys. Rev. Fluids* **3**, 061101(R) (2018).
69. N. Meinshausen, P. Bühlmann, Stability selection. *J. R. Statist. Soc. B* **72**, 417–473 (2010).



70. R. D. Shah, R. J. Samworth, Variable selection with error control: Another look at stability selection. *J. R. Statist. Soc. B* **75**, 55–80 (2013).
71. K. V. Edmond, M. T. Elsesser, G. L. Hunter, D. J. Pine, E. R. Weeks, Decoupling of rotational and translational diffusion in supercooled colloidal fluids. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17891–17896 (2012).
72. E. Bertin, A. Baskaran, H. Chaté, M. C. Marchetti, Comparison between Smoluchowski and Boltzmann approaches for self-propelled rods. *Phys. Rev. E* **92**, 042141 (2015).
73. B. Ventejou, H. Chaté, R. Montagne, X. Q. Shi, Susceptibility of orientationally ordered active matter to chirality disorder. *Phys. Rev. Lett.* **127**, 238001 (2021).
74. A. Bricard, J. B. Caussin, N. Desreumaux, O. Dauchot, D. Bartolo, Emergence of macroscopic directed motion in populations of motile colloids. *Nature* **503**, 95–98 (2013).
75. T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, O. Shochet, Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.* **75**, 1226–1229 (1995).
76. J. Toner, Y. Tu, S. Ramaswamy, Hydrodynamics and phases of flocks. *Ann. Phys. (N. Y.)* **318**, 170–244 (2005).
77. L. Ristroph, J. C. Liao, J. Zhang, Lateral line layout correlates with the differential hydrodynamic pressure on swimming fish. *Phys. Rev. Lett.* **114**, 018102 (2015).
78. A. Filella, F. M. C. Nadal, C. Sire, E. Kanso, C. Eloy, Model of collective fish behavior with hydrodynamic interactions. *Phys. Rev. Lett.* **120**, 198101 (2018).
79. D. J. G. Pearce, A. M. Miller, G. Rowlands, M. S. Turner, Role of projection in the control of bird flocks. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 10422–10426 (2014).
80. J. D. Davidson *et al.*, Collective detection based on visual information in animal groups. *J. R. Soc. Interface* **18**, 20210142 (2021).
81. D. Zhang, L. Guo, G. E. Karniadakis, Learning in modal space: Solving time-dependent stochastic PDEs using physics-informed neural networks. *SIAM J. Sci. Comput.* **42**, A639–A665 (2020).
82. M. Raissi, A. Yazdani, G. E. Karniadakis, Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science* **367**, 1026–1030 (2020).
83. J. Colen *et al.*, Machine learning active-nematic hydrodynamics. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016708118 (2021).
84. K. P. Champion, S. L. Brunton, J. N. Kutz, Discovery of nonlinear multiscale systems: Sampling strategies and embeddings. *SIAM J. Appl. Dyn. Syst.* **18**, 312–333 (2019).
85. N. M. Mangan, J. N. Kutz, S. L. Brunton, J. L. Proctor, Model selection for dynamical systems via sparse regression and information criteria. *Proc. R. Soc. A* **473**, 20170009 (2017).
86. S. L. Brunton, J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control* (Cambridge University Press, 2019).
87. M. Cranmer *et al.*, "Discovering symbolic models from deep learning with inductive biases" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, Eds. (Curran Associates, Inc., 2020), vol. 33, pp. 17429–17442.
88. S. M. Udrescu, M. Tegmark, AI Feynman: A physics-inspired method for symbolic regression. *Sci. Adv.* **6**, eaay2631 (2020).
89. P. Lemos, N. Jeffrey, M. Cranmer, S. Ho, P. Battaglia, Rediscovering orbital mechanics with machine learning arXiv [Preprint] 2022). <http://arxiv.org/abs/2202.02306>
90. J. Pathak, B. Hunt, M. Girvan, Z. Lu, E. Ott, Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Phys. Rev. Lett.* **120**, 24102 (2018).
91. S. L. Brunton, B. R. Noack, P. Koumoutsakos, Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **52**, 477–508 (2020).
92. J. Colen *et al.*, Machine learning active-nematic hydrodynamics. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016708118 (2021).
93. Y. Bar-Sinai, S. Hoyer, J. Hickey, M. P. Brenner, Learning data-driven discretizations for partial differential equations. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15344–15349 (2019).
94. A. J. Linot, M. D. Graham, Deep learning to discover and predict dynamics on an inertial manifold. *Phys. Rev. E* **101**, 062209 (2020).
95. A. J. Linot, M. D. Graham, Data-driven reduced-order modeling of spatiotemporal chaos with neural ordinary differential equations. *Chaos* **32**, 073110 (2022).
96. P. Hohenberg, A. Krekhov, An introduction to the Ginzburg-Landau theory of phase transitions and nonequilibrium patterns. *Phys. Rep.* **572**, 1–42 (2015).
97. P. Zheng, T. Askham, S. L. Brunton, J. N. Kutz, A. Y. Aravkin, A unified framework for sparse relaxed regularized regression: SR3. *IEEE Access* **7**, 1404–1423 (2019).
98. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer New York Inc., 2001).