Data-driven discovery of active nematic hydrodynamics

Chaitanya Joshi,[1,2,*] Sattvic Ray,[3] Linnea M. Lemma,[3,1] Minu Varghese,[4,1]
Graham Sharp,[3] Zvonimir Dogic,[3,†] Aparna Baskaran,[1,‡] and Michael F. Hagan[1,§]

[1] *Department of Physics, Brandeis University, Waltham, Massachusetts 02453, USA*
[2] *Department of Physics and Astronomy, Tufts University, Medford, Massachusetts 02155, USA*
[3] *Department of Physics, University of California at Santa Barbara, Santa Barbara, California 93106, USA*
[4] *Department of Physics, University of Michigan, Ann Arbor, Michigan 48109 USA*

(Dated: December 9, 2022)

Active nematics can be modeled using phenomenological continuum theories that account for the dynamics of the nematic director and fluid velocity through partial differential equations (PDEs). While these models provide a statistical description of the experiments, the relevant terms in the PDEs and their parameters are usually identified indirectly. We adapt a recently developed method to automatically identify optimal continuum models for active nematics directly from spatio-temporal data, via sparse regression of the coarse-grained fields onto generic low order PDEs. After extensive benchmarking, we apply the method to experiments with microtubule-based active nematics, finding a surprisingly minimal description of the system. Our approach can be generalized to gain insights into active gels, microswimmers, and diverse other experimental active matter systems.

Active nematics demonstrate how energy-consuming motile constituents can self-organize into diverse non-equilibrium dynamical states [1–3]. They offer a versatile platform to advance our fundamental understanding of non-equilibrium physics and develop materials with properties that are thermodynamically forbidden in equilibrium. These twin goals require theoretical models that reveal the mechanisms underlying the emergent dynamics, and guide rational design to elicit desired spatio-temporal dynamics. Here, we combine data-driven model discovery with experiments and computational modeling to identify the most parsimonious model for an experimental realization of active nematics. Using the discovered model, we identify the relationship between key theoretical parameters, such as the magnitude of activity, and experimental control variables. The described methods can be applied to diverse active nematics ranging from shaken rods to motile cells [4–9], and other forms of active matter.

Our target is a quantitative description of microtubule-based active nematics. Being reconstituted from tunable and well-characterized components, they afford a unique opportunity to develop continuum theory models and connect these to the microscopic dynamics [10–12]. Hydrodynamic theories, built on purely symmetry considerations, have provided insight into dynamics of active nematics in general, and the microtubule-based system specifically. For example, such models have been used to describe defect dynamics [13–18], induced flows in the suspending fluid [19–21], and how confinement in planar [22–24] and curved geometries [25–27] controls defect proliferation and dynamics. These efforts employed a range of hydrodynamic models that assumed different symmetry-allowed terms, and the parameters of the model were largely undetermined. Thus, the field lacks a quantitative model and understanding of magnitudes and sources of error in existing approximations.

Data-driven approaches and machine learning have been successfully applied to study active matter [28]. However, previous studies for active nematics were limited to parameter optimization with a pre-assumed model [16, 29], or machine learning forecasting [30, 31] which, while successful, does not provide an analytical equation for the learned dynamics. To overcome these limitations, we build on the Sparse Identification of Non-linear Dynamics (SINDy) framework [32, 33] that was recently applied to particle-based simulations of active matter [34] and computational and experimental data of overdamped polar particles [35]. This method filters out the best parsimonious fit to the data from a highly generalized class of potential models. We adapt key improvements of this method [36–39] to the microtubule-based active nematics system. We then employ extensive birefringence and fluorescence measurements of microtubule alignment and PIV measurements of velocities, to identify equations governing both the orientational dynamics and the activity-driven flows. This enables direct inference of the underlying model that is rigorously supported by experimental data. In contrast to the hard-to-interpret deep neural nets generated by machine learning approaches, our method yields an optimal analytical model and estimated parameter values.

With the available alignment and velocity measurements, we seek models describing the active nematic as a single 2D fluid with nematic symmetry [12, 40]. Hence, there are two fields: the symmetric-traceless nematic tensor order parameter $\mathbf{Q} = s[\mathbf{n} \otimes \mathbf{n} - (1/2)\mathbf{I}]$ and a flow field $\mathbf{u}$, with $\mathbf{n}$ as the local orientation unit vector and s the scalar order parameter. We assume constant density and an incompressible fluid: the former is justified since the scalar order parameter captures density variations near the defects (see below); the latter is validated by numerical measurements of the divergence of the velocity field [41]. Our model then consists of 4 independent

scalar fields: Q_{xx}, Q_{xy}, u_x, u_y, and a latent variable P (pressure).

We begin by postulating the generalized form of the model. The Q-tensor dynamics takes the form common to all continuum theories of active nematics:

$$\partial_t Q_{ij} = \sum_k a_{ij}^k F_k(\mathbf{Q}, \mathbf{u}, \nabla\mathbf{Q}, \nabla\mathbf{u}, \ldots) \tag{1}$$

where the F_k's are combinations (potentially non-linear) of $\mathbf{Q}$, $\mathbf{u}$, and their spatial derivatives up to a maximum order, and the a_{ij}^k's are the corresponding phenomenological coefficients. For instance, in 2D, a well-known model for the Q-equation is [12]:

$$\partial_t\mathbf{Q} + \mathbf{u}\cdot\nabla\mathbf{Q} - \mathbf{S} = D_\mathrm{r}\mathbf{H} \tag{2}$$

where $\mathbf{S} = -(\mathbf{\Omega}\cdot\mathbf{Q} - \mathbf{Q}\cdot\mathbf{\Omega}) + \lambda\mathbf{E} - 2\lambda\mathbf{Q}(\mathbf{Q}\colon\nabla\mathbf{u})$ is the co-rotation term and $\mathbf{H} = a_2\mathbf{Q} + a_4\,\mathrm{Tr}(\mathbf{Q}^2)\mathbf{Q} + K\nabla^2\mathbf{Q}$ is the negative gradient of the liquid crystal free energy. Here, $E_{ij} = (\partial_i u_j + \partial_j u_i)/2$ and $\Omega_{ij} = (\partial_i u_j - \partial_j u_i)/2$ are the strain rate and vorticity tensors respectively, λ is the flow alignment parameter, D_r is the rotational diffusion coefficient, K is the elastic constant, and $a_2 > 0$, $a_4 < 0$ are phenomenological coefficients corresponding to the isotropic-nematic transition. (See Supplemental Material [42], which includes Refs.[30, 43–47], for further discussion.) We build a library of the terms F_k ($n = 246$) that can capture models well beyond Eq. (2). Further, we make no physics-based simplifying assumptions, e.g. translational, rotational, and Galilean invariance, for the alignment equation (Eq. 1). Hence, discovery of a model which satisfies these conditions is a test of the algorithm (see Supplemental Material [42]).

For the flow equation, the usual form assumed for model-discovery is Navier-Stokes-like, with the time-derivative on the left side and rest of the terms on the right side [33, 37–39]. However, because the active nematic is in the low Reynolds number regime [15, 20], the significance of the time-derivative term itself needs investigation. Indeed, active nematic flows have been modeled using pure Stokes [30, 48, 49], unsteady Stokes [20, 50], and full Navier-Stokes [12–14, 19, 22, 51–55] formulations. While these approaches have been compared numerically [56], there has yet to be a definitive indication of the contributions of the inertial terms for this system. Since the viscous forcing is guaranteed to exist in this regime, we assume a form

$$\nabla^2\mathbf{u} = c_0\partial_t\mathbf{u} + \sum_i c_i\mathbf{H}_i(\mathbf{Q}, \mathbf{u}, \nabla\mathbf{Q}, \nabla\mathbf{u}, \ldots) \tag{3}$$

with $\nabla\cdot\mathbf{u} = 0$, and the time-derivative on the *right hand side* so that its contribution can be evaluated. For instance, the lowest order symmetry-allowed 'active stress' in the flow equation is the well-known $-\alpha\mathbf{Q}$, with $\alpha > 0$ being the extensile 'activity' [12, 57]. In our model form, this gives a general flow equation:

$$\nabla^2\mathbf{u} = c_0\partial_t\mathbf{u} + c_1\mathbf{u}\cdot\nabla\mathbf{u} + c_2\nabla P + c_3\nabla\cdot\mathbf{Q} + \ldots$$

with the coefficient c_3 as the ratio of the activity to the viscosity, α/η.

We perform model discovery from the data as follows [58]. Setting N_x, N_y, and N_t as the number of measurements in the two spatial dimensions and time respectively, we randomly select m of the total $N_xN_yN_t$ space-time points. At each selected space-time point, we evaluate a linear system, *e.g.* for the Q_{xx} equation, $(\partial_t Q_{xx})_{m\times 1} = F_{m\times n}\cdot\vec{a}_{n\times 1}$. The derivatives are computed numerically, which amplifies noise in the data. To mitigate noise, we use two different approaches. In the *integral formulation*, for each of the m selected space-time points and n terms, we compute a local average in space and time in a small window (e.g. 5x5x5 pixels) [36]. This approach is effective for model discovery, but leads to inaccurate parameter estimates for the flow equation — since pressure is not an observable in the experiments, we must perform the operation $\hat{z}\cdot\nabla\times$ on the flow equation [33, 37], which adds one more order of derivatives, amplifying the noise. To obtain more powerful noise mitigation at the cost of additional analytical effort, we adapt a *weak formulation* of the PDE regression problem [38, 39]. Briefly, we fit the data to the weak form of Eq. (3):

$$\int_{\Omega_k} \mathbf{w}\cdot\left[\nabla^2\mathbf{u} = c_0\partial_t\mathbf{u} + c_1\mathbf{u}\cdot\nabla\mathbf{u} + \ldots\right] \tag{4}$$

By choosing an appropriate test function $\mathbf{w}$ (*s.t.* the boundary terms vanish after integration-by-parts), we can move the derivatives from the noisy experimental data to the exact test functions, and also integrate out latent variables using integration-by-parts (in this case by making $\mathbf{w}$ divergence-free, see Supplemental Material [42]). The terms included in the library are in Table S1.

Next, we seek optimal fits to these equations with the minimum number of non-zero terms, thus yielding an interpretable model that accurately describes the data but avoids overfitting. To this end, we perform Ridge regression (least-squares gives similar results), starting with all the terms in the library, and then eliminating the least important terms one-by-one to obtain a hierarchy of models [36]. Obtaining the R^2 value at each step, we plot the *optimality curve* as the logarithm of $(1 - R^2)$ as a function of the number of non-zero terms left in the model. We define the optimal number of terms n^* as the n-value at which the second derivative of the curve is highest, indicating the largest drop in $\log(1 - R^2)$.

To demonstrate the validity of our approach, we first benchmark it against data generated by numerical simulations (Fig. S1, Table S2, which includes Refs.[59, 60]). We consider two qualitatively different models for flow: one is purely Stokesian with substrate friction, and the other is unsteady Stokes flow [20, 50]. After adding synthetic noise to the simulation data, we apply the integral formulation to the alignment equation and the weak formulation to the flow equation. The framework returns

the correct equations with very small errors in the identified coefficients (Fig. S1, S2 and S3 [42]). Thus, we estimate important phenomenological parameters directly from the data, including the activity level α, bending modulus K, flow alignment coupling λ, and bulk free energy coefficients a_2 and a_4. Further benchmarking against varying window sizes and noise levels (see Supplemental Material [42]) indicates that the integral formulation benefits from high resolution, low noise data whereas the weak formulation benefits from a large amount (in space and time) of data.

Next, we perform model discovery on experimental microtubule-based active nematics (Fig. 1a, Supplemental Material [42] which includes Refs.[61–66]). Coarse-graining the director, we obtain a Q-tensor field that contains the spatially varying scalar order parameter and orientation (Fig. 1b). The low-fluorescence-intensity regions, corresponding to low microtubule density near the defect cores, are correlated with the low-scalar-order-parameter regions, thus capturing the density variation near the defects (Movie S1 [42]). This justifies the constant density assumption. The velocity is obtained from PIV analysis (Fig. 1c). We varied the ATP concentration, which determines the motor stepping speed and thus determines the structure and dynamics of active nematics. We collected the data on a field of view several vortex diameters wide (Fig. 1c) for long times ($>$ 20 velocity autocorrelation times, defined below). In addition, we acquired one more data set with higher resolution but a smaller field of view, denoted as the 'HR-SF' data [67] (see Fig. S4 [42]). Optimality curves for the alignment and flow equations respectively (Figs. 1d,e) lead to the following optimal model:

$$\begin{aligned} \partial_t \mathbf{Q} = & -\mathbf{u}\cdot\nabla\mathbf{Q} - (\mathbf{\Omega}\cdot\mathbf{Q} - \mathbf{Q}\cdot\mathbf{\Omega}) \\ & + \mathbf{E} - 2(\mathbf{Q}\colon\nabla\mathbf{u})\mathbf{Q} + K\nabla^2\mathbf{Q} \\ \eta\nabla^2\mathbf{u} = & + \alpha\nabla\cdot\mathbf{Q} + \nabla P. \end{aligned} \tag{5}$$

Note that we added the term $K\nabla^2\mathbf{Q}$ because this or an analogous term with higher order derivatives must be present for stability, discussed further below.

We arrived at this model as follows. For the alignment equation, the HR-SF data set (purple triangles in Fig. 1d) has a low error ($R^2 = 0.97$) and an abrupt shoulder that clearly defines a threshold for the optimal model. In comparison, the lower resolution data sets have larger error (see Table S3 [42]) and less distinct thresholds. Consistent with the benchmarking of numerical data described above and in Supplemental Material [42], these results show that high resolution is more important than a large field-of-view for determining the alignment equation. The threshold chosen for each data set is indicated by the tail of the corresponding arrow in Fig. 1d, and the resulting model for each data set is given in Table S3 [42]. Table S4 [42] gives the lowest-order terms beyond the threshold. For all data sets, the optimal model is dominated by flow-coupling terms, such as the convective and rotational derivatives and flow alignment. Eq. (5) corresponds to the optimal model for the HR-SF data set, and with the exception of the higher-order flow-alignment term $2(\mathbf{Q}:\nabla\mathbf{u})\mathbf{Q}$, two of the low-resolution data sets. For other data sets, there is some variability in the terms near the threshold (Tables S3 and S4), but the terms in Eq. (5) are all present near the threshold, and other near-threshold terms can be eliminated because they violate

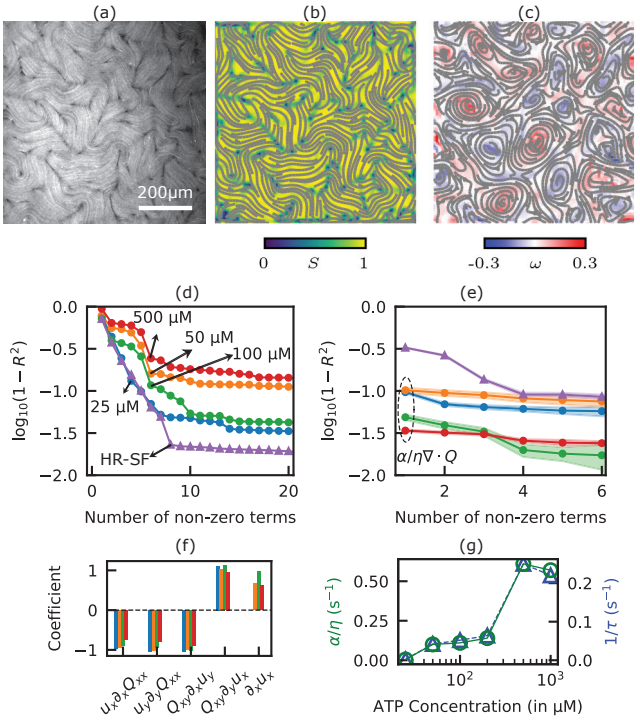

FIG. 1. Discovering active nematic hydrodynamics from experimental data. (a) A representative fluorescence image of the microtubule-kinesin active nematic at an ATP concentration of 100 μM (scale bar is 200 μm). (b) The computed director field and scalar order parameter S, and (c) the flow field and vorticity ω for the data in (a). (d) Optimality curves, $\log(1-R^2)$ vs. number of non-zero terms, for the Q_{xx} equation from the indicated data sets. The beginning of each arrow corresponds to the threshold corresponding to the highest-order term included in the optimal model. (e) Optimality curves for the weak-form flow equation. In the cases highlighted with the dashed oval, the optimal model contains only the activity term, $\nabla\cdot\mathbf{Q}$, consistent with Stokesian dynamics. In (d,e) the purple triangles correspond to the high-resolution, small field-of-view (HR-SF) data set, while the blue, orange, green, and red circles correspond to 25 μM, 50 μM, 100 μM, and 500 μM ATP respectively. (f) Values of the coefficients of key flow-coupling terms appearing in the optimal models for various ATP concentrations. The colors are the same as in (d,e). (g) The fit coefficient of the activity term, α/η, as a function of the ATP concentration (green circles). This quantity closely matches the inverse of the velocity correlation time (blue triangles), suggesting that α/η corresponds to a relevant timescale in the system.

known symmetry or conservation criteria for the system. We include the higher-order flow coupling term because the HR-SF data set has the highest statistical accuracy and because it is expected theoretically for stability of the nematic order parameter. We attribute the variability in the near-threshold terms for the low-resolution data sets to statistical inaccuracies arising from the limited experimental data and the small contributions of these terms, rather than different physics being present at different ATP concentrations. These results highlight the importance of the amount and resolution of data for accurately determining the alignment equation.

The alignment equation recovers Galilean invariance from the data: the convective and co-rotational derivatives have coefficients of ~ 1 (Fig. 1f). Furthermore, the flow alignment parameter, $\lambda \sim 1$ (Fig. 1f), is consistent with the theoretical result for the high aspect ratio $a \gg 1$ of the microtubules, $\lambda = (a^2 - 1)/(a^2 + 1) \to 1$ [68]. Importantly, the bulk liquid crystal free energy terms that stabilize nematic order (with coefficients $a_2 > 0$ and $a_4 < 0$, see Eq. (2)) are not present in the discovered model for any data set [69]. This finding supports a previous model [54] which argued that active flow alignment acts as an effective free energy that drives nematic order. These results indicate that the alignment dynamics are dominated by flow-coupling. In comparison, contributions from the free energy dissipation to the dynamics are negligible. Elastic distortion energy terms [70, 71] only appear above the threshold (see Table S4 [42]). However, a term of the form $K\nabla^2\mathbf{Q}$, which contains the elastic terms in the single constant approximation, is required for numerical stability. Moreover, the elastic terms play a key role in determining the structure of a nematic in the vicinity of defects. To understand this apparent contradiction, we compare the contributions of the distortion energy with flow coupling terms as a function of space (Movie S2 [42]). This shows that the elastic terms are small everywhere except near defects. When combined with the fact that the majority of the experimental data is far from defects due to their small core size and finite density, this is the likely reason for the absence of elastic terms in the discovered model (Fig S5).

The optimality curves for the flow equation are almost flat (Fig. 1e), showing that the active force $\alpha/\eta\nabla\cdot\mathbf{Q}$ alone balances the viscous force. Noting that this is a fit to the *weak* form of the equation, we test the strong form of the discovered equation by comparing the spatial dependence of $\nabla\times\nabla^2\mathbf{u}$ with $\alpha/\eta\nabla\times\nabla\cdot\mathbf{Q}$ and find good agreement (Movies S3 and S4 [42]). The inertial terms are absent (not appearing until $n \sim 5$), indicating that the Stokes flow approximation accurately describes the experimental active nematic. Finally, the absence of the substrate friction term $\Gamma\mathbf{u}$ indicates that the screening length $\sqrt{\eta/\Gamma}$ is larger than the typical vortex size of the flows. This result likely depends on the active nematic system and experimental conditions; for example, changing the substrate depth affects the effective friction coefficient [72]. However, the framework presented here can be applied directly to other conditions or materials.

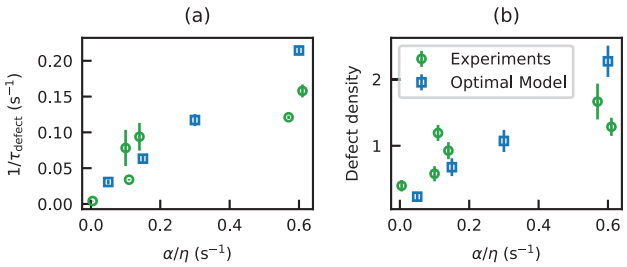

FIG. 2. Comparisons of results from simulations using the discovered optimal model against the experimental data. (a) Inverse lifetime of +1/2 defects plotted as a function of α/η for experiments at different ATP concentrations (green circles) and simulations using the optimal model performed with different values of α/η (blue squares). For the experiments, the value of α/η is obtained from the discovered optimal model at each ATP concentration. The height of the errorbars is twice the standard error of mean. (b) Defect density plotted against α/η from experiments (green circles) and simulations using the optimal model (blue squares). The density in the simulations is scaled by a constant. The height of the errorbars is twice the standard deviation.

The discovered flow equation provides a direct estimate of the scaled activity parameter α/η, an intrinsic 'active time scale' [73], as a function of the ATP concentration (Fig. 1g) [74]. Determining the relationship between activity and experimental control parameters has been a significant challenge [21]. The molecular motors that generate activity also act as passive cross-linkers between steps [75], and in a dense active nematic, forces generated by different motors are largely non-cooperative [76]. To test the estimate of α against an independent observable, we compare the active time scale to the velocity autocorrelation time τ, defined as $\bar{C}_v(\tau) = 1/e$, with the autocorrelation function $\bar{C}_v(t) = \langle\langle\mathbf{u}(\mathbf{r},t'+t)\cdot\mathbf{u}(\mathbf{r},t')\rangle_{t'} / \langle\mathbf{u}(\mathbf{r},t')\cdot\mathbf{u}(\mathbf{r},t')\rangle_{t'}\rangle_{\mathbf{r}}$. These observables closely agree at all ATP concentrations (Fig. 1g).

Finally, we test the optimal model by performing simulations of Eq. (5). For numerical stability, we include the $K\nabla^2\mathbf{Q}$ term in the $\mathbf{Q}$ equation with $K = 1$ in dimensionless simulation units. We compare the mean defect lifetime and defect density [77, 78] as a function of α/η (Fig. 2). Remarkably, the defect lifetimes for experiments and simulations align well without any fit parameters (Fig. 2a). The defect densities from experiment and simulation also match, up to a constant scaling factor [79] (Fig. 2b). The latter cannot be specified — because the terms in the discovered alignment equation all have dimensionless coefficients, we cannot directly estimate a length scale [80].

In summary, we have applied a data-driven method to identify equations governing both the orientational dynamics and the activity-driven flows of microtubule-based active liquid crystals. The optimal model is surprisingly minimal. It demonstrates that: (1) flow coupling dominates the orientational dynamics, and (2) the lowest-order active stress, proportional to the local orientational order, together with the vanishing Reynolds number limit describe the flow. This model is not only consistent with previous theoretical arguments [54], but is also less complex than most models considered in the literature. Our results also show that statistical uncertainty arising from limited experimental data impedes unequivocal identification of near-threshold terms, but suggest strategies to mitigate these effects. For example, acquiring a combination of high-resolution small field-of-view and low-resolution large field-of-view data sets would enable more accurate discovery of the alignment and flow equations respectively. Further, acquiring more data in the vicinity of defects and/or additional analysis that preferentially weights data in the vicinity of defects may identify elastic energy terms.

The identified equations enable mapping between key model parameters and experimental control variables, including the elusive relationship between the magnitude of activity and ATP concentration. Thus, our results are the first to assess the quantitative variation of phenomenological theory parameters as a function of experimental control knobs in active nematics, while *also* providing evidence for the validity of the underlying model. Through comparison of several noise reduction approaches and extensive benchmarking, we have identified an approach to model discovery which is highly robust against experimental noise. This approach can be extended to study recently developed 3D active nematic materials [49, 81, 82], complementing existing theoretical and numerical efforts [48, 52, 83–87]. It can be applied to a wide variety of active matter systems, or more broadly, to any system for which observations of dynamics can be projected onto continuous fields. This process can shed light on relationships between physical quantities or even identify new physical mechanisms.

In the final stages of this project, we learned of a complementary, concurrent work that uses symbolic regression, whose findings are generally consistent with those of our work [88].

This work was supported by the Department of Energy (DOE) DE-SC0022291. Preliminary data and analysis were supported by the National Science Foundation (NSF) DMR-1855914 and the Brandeis Center for Bioinspired Soft Materials, an NSF MRSEC (DMR-2011846). Computing resources were provided by the NSF XSEDE allocation TG-MCB090163 (Stampede and Comet) and the Brandeis HPCC which is partially supported by the NSF through DMR-MRSEC 2011846 and OAC-1920147. We thank Link Morgan for providing early experimental data for testing, Saaransh Singhal for providing the simulation data for the unsteady Stokes equation, Peter J. Foster for providing feedback on the manuscript, and Matthew S. E. Peterson, Michael M. Norton and Seth Fraden for valuable discussions. We are grateful to Matthew Golden and Roman Grigoriev for suggesting the analysis shown in Movie S3 and S4.

* chaitanya@brandeis.edu
† zdogic@physics.ucsb.edu
‡ aparna@brandeis.edu
§ hagan@brandeis.edu

[1] M. C. Marchetti, J. F. Joanny, S. Ramaswamy, T. B. Liverpool, J. Prost, M. Rao, R. A. Simha, and M. Curie, Reviews of Modern Physics **85**, 1143 (2013), arXiv:1207.2929.
[2] S. Ramaswamy, Annual Review of Condensed Matter Physics **1**, 323 (2010).
[3] J. Toner, Y. Tu, and S. Ramaswamy, Annals of Physics **318**, 170 (2005).
[4] V. Narayan, S. Ramaswamy, and N. Menon, Science **317**, 105 (2007).
[5] H. H. Wensink, J. Dunkel, S. Heidenreich, K. Drescher, R. E. Goldstein, H. Lowen, and J. M. Yeomans, Proceedings of the National Academy of Sciences **109**, 14308 (2012).
[6] S. Zhou, A. Sokolov, O. D. Lavrentovich, and I. S. Aranson, Proceedings of the National Academy of Sciences **111**, 1265 (2014).
[7] G. Duclos, C. Erlenkämper, J. F. Joanny, and P. Silberzan, Nature Physics **13**, 58 (2017), iSBN: 1745-2473 1745-2481.
[8] K. Kawaguchi, R. Kageyama, and M. Sano, Nature **545**, 327 (2017), iSBN: 1476-4687 (Electronic) 0028-0836 (Linking) Publisher: Nature Publishing Group.
[9] N. Kumar, R. Zhang, J. J. de Pablo, and M. L. Gardel, Science Advances **4**, eaat7779 (2018).
[10] T. Sanchez, D. T. N. Chen, S. J. DeCamp, M. Heymann, and Z. Dogic, Nature **491**, 431 (2012), arXiv:1301.1122.
[11] S. J. DeCamp, G. S. Redner, A. Baskaran, M. F. Hagan, and Z. Dogic, Nature Materials **14**, 1110 (2015).
[12] A. Doostmohammadi, J. Ignés-Mullol, J. M. Yeomans, and F. Sagués, Nature Communications **9**, 3246 (2018).
[13] L. Giomi, M. J. Bowick, X. Ma, and M. C. Marchetti, Phys. Rev. Lett. **110**, 228101 (2013).
[14] L. Giomi and A. DeSimone, Physical Review Letters **112**, 147802 (2014).
[15] A. Doostmohammadi, T. N. Shendruk, K. Thijssen, and J. M. Yeomans, Nature Communications **8**, 15326 (2017).
[16] A. U. Oza and J. Dunkel, New Journal of Physics **18**, 093006 (2016).
[17] D. Cortese, J. Eggers, and T. B. Liverpool, Physical Review E **97**, 022704 (2018).
[18] T. N. Shendruk, K. Thijssen, J. M. Yeomans, and A. Doostmohammadi, Physical Review E **98**, 010601 (2018), arXiv:1803.02093.
[19] S. P. Thampi, R. Golestanian, and J. M. Yeomans, Phys. Rev. E **90**, 062307 (2014).
[20] L. Giomi, Physical Review X **5**, 031003 (2015).

[21] L. M. Lemma, S. J. DeCamp, Z. You, L. Giomi, and Z. Dogic, Soft Matter **15**, 3264 (2019), publisher: The Royal Society of Chemistry.

[22] T. N. Shendruk, A. Doostmohammadi, K. Thijssen, and J. M. Yeomans, Soft Matter **13**, 3853 (2017), arXiv:1703.01531.

[23] M. M. Norton, A. Baskaran, A. Opathalage, B. Langeslay, S. Fraden, A. Baskaran, and M. F. Hagan, Phys. Rev. E **97**, 012702 (2018).

[24] T. Gao, M. D. Betterton, A.-S. Jhang, and M. J. Shelley, Physical Review Fluids **2**, 093302 (2017).

[25] R. Zhang, Y. Zhou, M. Rahimi, and J. J. de Pablo, Nature Communications **7**, 13483 (2016).

[26] P. W. Ellis, D. J. G. Pearce, Y.-W. Chang, G. Goldsztein, L. Giomi, and A. Fernandez-Nieves, Nature Physics **14**, 85 (2018).

[27] F. Alaimo, C. Köhler, and A. Voigt, Scientific Reports **7**, 1 (2017).

[28] F. Cichos, K. Gustavsson, B. Mehlig, and G. Volpe, Nature Machine Intelligence **2**, 94 (2020).

[29] H. Li, X.-q. Shi, M. Huang, X. Chen, M. Xiao, C. Liu, H. Chaté, and H. P. Zhang, Proceedings of the National Academy of Sciences **116**, 777 (2019).

[30] Z. Zhou, C. Joshi, R. Liu, M. M. Norton, L. Lemma, Z. Dogic, M. F. Hagan, S. Fraden, and P. Hong, Soft Matter **17**, 738 (2021).

[31] J. Colen, M. Han, R. Zhang, S. A. Redford, L. M. Lemma, L. Morgan, P. V. Ruijgrok, R. Adkins, Z. Bryant, Z. Dogic, M. L. Gardel, J. J. De Pablo, and V. Vitelli, Proceedings of the National Academy of Sciences **118**, e2016708118 (2021), arXiv:2006.13203.

[32] S. L. Brunton, J. L. Proctor, J. N. Kutz, and W. Bialek, Proceedings of the National Academy of Sciences of the United States of America **113**, 3932 (2016), arXiv:1509.03580.

[33] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, Science Advances **3**, e1602614 (2017).

[34] S. Maddu, Q. Vagne, and I. F. Sbalzarini, arXiv:2201.08623 [cond-mat, physics:physics] (2022), arXiv:2201.08623 [cond-mat, physics:physics].

[35] R. Supekar, B. Song, A. Hastewell, A. Mietke, and J. Dunkel, arXiv:2101.06568 [cond-mat, physics:physics] (2021), arXiv:2101.06568 [cond-mat, physics:physics].

[36] E. P. Alves and F. Fiuza, arXiv:2011.01927 [astro-ph, physics:physics] (2020), arXiv: 2011.01927.

[37] P. A. K. Reinbold and R. O. Grigoriev, Phys. Rev. E **100**, 022219 (2019).

[38] P. A. K. Reinbold, D. R. Gurevich, and R. O. Grigoriev, Phys. Rev. E **101**, 010203 (2020).

[39] P. A. K. Reinbold, L. M. Kageorge, M. F. Schatz, and R. O. Grigoriev, Nature Communications **12**, 3219 (2021).

[40] A. N. Beris and B. J. Edwards, *Thermodynamics of flowing systems with internal microstructure* (Oxford University Press, 1994).

[41] L. M. Lemma, M. M. Norton, A. M. Tayar, S. J. DeCamp, S. A. Aghvami, S. Fraden, M. F. Hagan, and Z. Dogic, Physical Review Letters **127**, 148001 (2021).

[42] (2022), see Supplemental Material at URL.

[43] G. P. G. de and J. Prost, *The physics of liquid crystals* (Clarendon Press, 1993).

[44] L. Giomi, M. J. Bowick, P. Mishra, R. Sknepnek, and M. Cristina Marchetti, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **372**, 20130365 (2014).

[45] T. Gao, R. Blackwell, M. A. Glaser, M. D. Betterton, and M. J. Shelley, Physical Review Letters **114**, 1 (2015), arXiv:1401.8059.

[46] E. J. Hemingway, P. Mishra, M. C. Marchetti, and S. M. Fielding, Soft Matter **12**, 7943 (2016).

[47] S. Ngo, A. Peshkov, I. S. Aranson, E. Bertin, F. Ginelli, and H. Chaté, Physical Review Letters **113** (2014), 10.1103/PhysRevLett.113.038302.

[48] M. Varghese, A. Baskaran, M. F. Hagan, and A. Baskaran, Physical Review Letters **125**, 268003 (2020).

[49] G. Duclos, R. Adkins, D. Banerjee, M. S. E. Peterson, M. Varghese, I. Kolvin, A. Baskaran, R. A. Pelcovits, T. R. Powers, A. Baskaran, F. Toschi, M. F. Hagan, S. J. Streichan, V. Vitelli, D. A. Beller, and Z. Dogic, Science **367**, 1120 (2020).

[50] L. Giomi, L. Mahadevan, B. Chakraborty, and M. F. Hagan, Nonlinearity **25**, 2245 (2012), arXiv:arXiv:1110.4338v1.

[51] S. Chandragiri, A. Doostmohammadi, J. M. Yeomans, and S. P. Thampi, Soft Matter **15**, 1597 (2019), arXiv:1901.06468.

[52] S. Chandragiri, A. Doostmohammadi, J. M. Yeomans, and S. P. Thampi, Physical Review Letters **125**, 148002 (2020).

[53] S. P. Thampi, R. Golestanian, and J. M. Yeomans, Physical Review Letters **111**, 118101 (2013).

[54] S. P. Thampi, A. Doostmohammadi, R. Golestanian, and J. M. Yeomans, EPL (Europhysics Letters) **112**, 28004 (2015).

[55] S. P. Thampi, A. Doostmohammadi, T. N. Shendruk, R. Golestanian, and J. M. Yeomans, Science Advances **2**, e1501854 (2016).

[56] C.-M. Koch and M. Wilczek, Physical Review Letters **127**, 268005 (2021).

[57] R. Aditi Simha and S. Ramaswamy, Physical Review Letters **89**, 058101 (2002).

[58] The codes used for this work are available under https://github.com/joshichaitanya3/actnempy.

[59] J. Zhao and Q. Wang, Journal of Scientific Computing **68**, 1241 (2016).

[60] S. P. Vanka, Journal of Computational Physics **65**, 138 (1986).

[61] M. Castoldi and A. V. Popov, Protein Expression and Purification **32**, 83 (2003).

[62] A. Hyman, D. Drechsel, D. Kellogg, S. Salser, K. Sawin, P. Steffen, L. Wordeman, and T. Mitchison, in *Methods in Enzymology*, Molecular Motors and the Cytoskeleton, Vol. 196 (Academic Press, 1991) pp. 478–485.

[63] R. Subramanian and J. Gelles, Journal of General Physiology **130**, 445 (2007).

[64] A. W. C. Lau, A. Prasad, and Z. Dogic, EPL (Europhysics Letters) **87**, 48006 (2009).

[65] R. Oldenbourg, in *Live Cell Imaging: A Laboratory Manual*, edited by R. D. Goldman and D. L. Spector (Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY) pp. 205–237.

[66] D. Garcia, Experiments in Fluids **50**, 1247 (2011).

[67] The higher resolution data has the PIV velocity measured on a 79×66 pixel grid, which is comparable to the rest of the data sets (71×71), but its field of view contains only one or two vortices (Fig. S4 [42]), thus resolving the flow field much more accurately.

[68] A. Maitra, P. Srivastava, M. Cristina Marchetti, J. S. Lintuvuori, S. Ramaswamy, and M. Lenz, Proceedings of the National Academy of Sciences of the United States of America **115**, 6934 (2018).

[69] Three data sets give (near threshold) $a_2 < 0$ (isotropic regime), with one of them containing an unstable component with $a_4 > 0$.

[70] E. Putzig, G. S. Redner, A. Baskaran, and A. Baskaran, Soft Matter **12**, 3854 (2016).

[71] N. J. Mottram and C. J. P. Newton, arXiv:1409.3542 [cond-mat] (2014), arXiv: 1409.3542.

[72] K. Thijssen, D. A. Khaladj, S. A. Aghvami, M. A. Gharbi, S. Fraden, J. M. Yeomans, L. S. Hirst, and T. N. Shendruk, Proceedings of the National Academy of Sciences **118**, e2106038118 (2021).

[73] L. Giomi, L. Mahadevan, B. Chakraborty, and M. F. Hagan, Physical Review Letters **106**, 218101 (2011).

[74] Independent measurements of the viscosity such as in [89] can be then used to estimate α, the strength of the active force.

[75] D. A. Gagnon, C. Dessi, J. P. Berezney, R. Boros, D. T.-N. Chen, Z. Dogic, and D. L. Blair, Physical Review Letters **125**, 178003 (2020).

[76] A. M. Tayar, M. F. Hagan, and Z. Dogic, Proceedings of the National Academy of Sciences **118**, e2102873118 (2021).

[77] R. D. Kamien, Reviews of Modern Physics **74**, 953 (2002).

[78] D. B. Allan, T. Caswell, N. C. Keim, C. M. van der Wel, and R. W. Verweij, "Soft-matter/trackpy: Trackpy v0.5.0," Zenodo (2021).

[79] We re-scale the defect density so that the simulation value for $\alpha/\eta = 0.3$ falls on the linear fit of the experimental values.

[80] We can force the framework to estimate dimensional parameters by constraining the regression procedure to include specific terms, while performing sparse regression on the remaining terms. For example, by forcing a term $K\nabla^2 Q_{xx}$ (see Eq. (2)), we obtain a value for the elastic modulus of $K \sim 1\,\mu\mathrm{m}^2\,\mathrm{s}^{-1}$. However, because this term has a negligible contribution to the dynamics of $\mathbf{Q}$, the quantitative accuracy of this estimate may be limited.

[81] P. Chandrakar, M. Varghese, S. A. Aghvami, A. Baskaran, Z. Dogic, and G. Duclos, Physical Review Letters **125**, 257801 (2020).

[82] B. Najma, M. Varghese, L. Tsidilkovski, L. Lemma, A. Baskaran, and G. Duclos, arXiv preprint arXiv:2112.11364 (2021), arXiv:2112.11364.

[83] D. M. Sussman and D. A. Beller, Frontiers in Physics **7**, 204 (2019).

[84] S. Mandal and M. G. Mazza, Physical Review E **99**, 063319 (2019).

[85] T. Kozhukhov and T. N. Shendruk, Science Advances **8**, eabo5788 (2022).

[86] L. J. Ruske and J. M. Yeomans, Physical Review X **11**, 021001 (2021).

[87] S. Čopar, J. Aplinc, Ž. Kos, S. Žumer, and M. Ravnik, Physical Review X **9**, 031051 (2019).

[88] M. Golden, R. Grigoriev, J. Nambisan, and A. Fernandez-Nieves, arXiv:2202.12853 [cond-mat, physics:physics] (2022), arXiv:2202.12853 [cond-mat, physics:physics].

[89] P. Guillamat, J. Ignés-Mullol, S. Shankar, M. C. Marchetti, and F. Sagués, Physical Review E **94**, 060602 (2016), arXiv:1606.05764.

Supplemental Material: Data-driven discovery of active nematic hydrodynamics

Chaitanya Joshi,[1,2,*] Sattvic Ray,[3] Linnea M. Lemma,[3,2] Minu Varghese,[4,2]
Graham Sharp,[3] Zvonimir Dogic,[3,†] Aparna Baskaran,[2,‡] and Michael F. Hagan[2,§]

[1] *Department of Physics and Astronomy, Tufts University, Medford, Massachusetts 02155, USA*
[2] *Department of Physics, Brandeis University, Waltham, Massachusetts 02453, USA*
[3] *Department of Physics, University of California at Santa Barbara, Santa Barbara, California 93106, USA*
[4] *Department of Physics, University of Michigan, Ann Arbor, Michigan 48109 USA*

(Dated: February 21, 2023)

CONTINUUM THEORY OF ACTIVE NEMATICS

The theory of active nematic suspensions builds on equilibrium nematic hydrodynamics [1, 2] and extends it to include non-equilibrium 'active stresses' in the fluid [3–13]. The most generic models of active nematics have considered the dynamics of not only the orientation and the velocity, but also of the concentration [5, 14, 15] and/or density [16]. In this work, we assume uniform density and concentration everywhere. This is a reasonable assumption for dense 2D bulk systems, and we expect that allowing density/concentration variations would only improve our results. We characterize the orientational order using the tensor order parameter $\mathbf{Q} = S[\mathbf{n}\otimes\mathbf{n} - (1/2)\mathbf{I}]$. This definition has the $\mathbf{n}\to-\mathbf{n}$ nematic symmetry built-in, and also provides a scalar order parameter $S(\mathbf{r},t)$ that represents the magnitude of the orientational order at the given location. For equilibrium systems, the free energy near the isotropic-nematic transition takes the form

$$\mathcal{F} = \frac{a_2}{2}Q^2 + \frac{a_4}{4}Q^4 + \frac{K}{2}(\nabla Q)^2 \tag{1}$$

with $a_2 = (1-\rho)$ and $a_4 = (\rho+1)/\rho^2$. The density ρ controls the transition from the isotropic ($\rho<1$) to the nematic ($\rho>1$) state. If $\rho=\rho_0>1$, the minimum of the free energy is in an ordered state, with $S_{\text{eqm}} = \sqrt{-2a_2/a_4}$. K is the elastic modulus of the nematic, where we have assumed equal moduli for splay and bend deformations. For the alignment tensor, the dynamical equations of motion are:

$$\frac{\mathrm{D}\mathbf{Q}}{\mathrm{D}t} = \partial_t\mathbf{Q} + \mathbf{u}\cdot\nabla\mathbf{Q} + (\mathbf{\Omega}\cdot\mathbf{Q} - \mathbf{Q}\cdot\mathbf{\Omega}) = \lambda\mathbf{E} + D_{\text{r}}\mathbf{H} \tag{2}$$

The left hand side corresponds to the co-moving co-rotational derivative of the Q-tensor, with $\mathbf{\Omega} = 1/2(\nabla\mathbf{u} - (\nabla\mathbf{u})^T)$ and $\mathbf{E} = 1/2(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)$. λ is the flow alignment parameter, with $\lambda\mathbf{E}$ being the lowest order contribution to the flow alignment. The next leading order term $-2\lambda(\mathbf{Q}\colon\mathbf{E})\mathbf{Q}$ [17] is also often included [18]. Lastly, $\mathbf{H} = -\delta\mathcal{F}/\delta\mathbf{Q}$ and D_{r} is the coefficient of rotational diffusion. $\mathbf{H}$ gives the free energy contribution, while all other terms come from coupling of the orientation to the underlying flow.

The corresponding fluid flow can be described by the incompressible Navier-Stokes (NS) equations. These are augmented with the (passive) back-flow σ_{p} due to the coupling to the nematic. In addition, the lowest order non-equilibrium nematic stress takes the form $-\alpha\mathbf{Q}$, where α indicates the strength of the active forces, or "activity". Adding this contribution in, we get the *active* nematic fluid equation:

$$\rho(\partial_t\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u}) = \eta\nabla^2\mathbf{u} - \nabla P + \nabla\cdot(\sigma_{\text{p}} - \alpha\mathbf{Q}) \tag{3}$$

with ρ being the density and η being the viscosity.

For thin 2D samples such as the ones considered in this Letter, Eq. (3) needs to be integrated along the height h of the confinement to obtain a quasi-2D equation. Integrating the viscous term gives an effective linear friction $-\Gamma\mathbf{u}$, with $\Gamma\sim\eta/h^2$ [10].

Since the flows are in the low-Reynolds number regime, it is common to neglect the non-linear convection term [5, 15], and often the entire inertia term [19–21]. This is the Stokes limit. The theoretical consequences of varying the advective inertia and substrate friction have been the subject of numerical studies [7, 22].

Additionally, a higher order active stress $\nabla(\mathbf{Q}\cdot\nabla\cdot\mathbf{Q})$ in 3D can give rise to a non-equilibrium active force $\mathbf{Q}\cdot\nabla\cdot\mathbf{Q}$ in 2D, which can be expected to have a similar magnitude as the primary active force $-\alpha\nabla\cdot\mathbf{Q}$ [10].

This brief sketch of active nematics models shows the wealth of information that a data-driven method like SINDy could uncover when applied to experimental systems.

For the models that we use for benchmarking calculations by numerically simulating PDEs in this Letter, we consider both the Stokes limit described above as well as the 'unsteady Stokes' limit (which retains $\partial_t\mathbf{u}$). Further, we set $\sigma_{\text{p}}=0$ as it has been observed numerically that active stresses dominate the passive stresses [18, 23]. We include the substrate friction as it adds another degree of freedom that our framework has to identify. With this, we obtain the dynamical equation for the fluid:

$$\begin{aligned} 0 &= \eta\nabla^2\mathbf{u} - \Gamma\mathbf{u} - \nabla P - \nabla\cdot(\alpha\mathbf{Q}) \quad \text{(Stokes)} \\ \rho\partial_t\mathbf{u} &= \eta\nabla^2\mathbf{u} - \nabla P - \nabla\cdot(\alpha\mathbf{Q}) \quad \text{(Unsteady Stokes)} \\ \nabla\cdot\mathbf{u} &= 0 \end{aligned} \tag{4}$$

The Q-tensor equation is unaltered by activity in these models because the active stresses arise solely due to flow coupling.

LIBRARY CREATION AND BENCHMARKING

The codes used for this work are available under https://github.com/joshichaitanya3/actnempy. Each data-set is pre-processed such that the fields Q_{xx}, Q_{xy}, u_x and u_y lie on the same $N_x \times N_y$ grid. With N_t such measurements in time, we get $N_x \times N_y \times N_t$ values for each of the field.

Integral Formulation. We begin by creating a database of terms containing the fields and their derivatives. We compute the time and space derivatives numerically using a central difference scheme. (Since $\nabla \cdot \mathbf{u} = \partial_x u_x + \partial_y u_y = 0$, we discard $\partial_y u_y$ from our database.) To form the library, we then make all multiplicative combinations of these terms with a total functional order up to f and a total gradient order up to d. We put further limits on the function and gradient order of $\mathbf{u}$ appearing in the terms. Thus, we specify two pairs, (f_t , d_t) for the overall constraint, and ($f_\mathbf{u}$, $d_\mathbf{u}$) for further constraint on the terms involving velocity. This allows us to make the computation more tractable. Motivated by theoretical models discussed in earlier sections, we use ($f_\mathrm{t} = 3, d_\mathrm{t} = 2$) and ($f_\mathbf{u} = 1, d_\mathbf{u} = 1$) for the $\mathbf{Q}$ tensor equation (Main Text Eq. (1)) and ($f_\mathrm{t} = 2, d_\mathrm{t} = 2$) for the flow equation (Main Text Eq. (3)). In this approach, we do not make any simplifying assumptions about the terms: all terms appear in an "unfolded" form, with the derivatives and inner products expanded out.

Since our data has two spatial and one time dimensions, we have a large number of data-points. Hence, we compute the library terms only on a sub-sample of the data [24]: we randomly select $m = 5000$ points from the $N_x \times N_y \times N_t$ grid and compute the local average of the terms near the points using a small (*e.g.* 5x5x5 pixels) averaging window [25].

Weak Formulation. As described in the main text, the weak formulation provides better noise mitigation than the integral formulation for the fluid flow equation due to the high-order derivatives. For the fluid flow, we assume a generalized flow equation

$$\int_{\Omega_k} \mathbf{w} \cdot \left[\nabla^2 \mathbf{u} = c_0 \partial_t \mathbf{u} + c_1 \mathbf{u} \cdot \nabla \mathbf{u} + \ldots\right]$$

Using a similar "unfolded" form for this library as for the orientation equation would generate a lot of terms, and it is not feasible to perform the integration-by-parts (IBP) required for the weak formulation for all of them. Hence, for this calculation, we create the library by hand with a judicious choice of terms, listed in Table S1. Here, $\mathbf{w}$ is a vector test function and the integration domain

Number	Term
1	$\partial_t \mathbf{u}$
2	$\mathbf{u} \cdot \nabla \mathbf{u}$
3	$\nabla \cdot \mathbf{Q}$
4	$\mathbf{Q} \cdot \nabla \cdot \mathbf{Q}$
5	$\mathbf{u}$
6	$\mathbf{Q} \cdot \mathbf{u}$
7	$\mathrm{Tr}(\mathbf{Q}^2)\mathbf{u}$

TABLE S1. Terms appearing in the library for the flow equation.

Ω_k is a rectangular box of size $2H_x \times 2H_y \times 2H_t$, centered at (x_k, y_k, t_k) [26]. This test function needs to satisfy the following criteria: (1) the boundary terms in the integration-by-parts, containing derivatives of $\mathbf{w}$, must vanish identically and (2) The 2D divergence of $\mathbf{w}$ needs to be zero everywhere for the pressure term to vanish. Condition (1) can be met by ensuring that $\mathbf{w}$ and its derivatives up to a certain order vanish at the boundary; condition (2) can be met by choosing $\mathbf{w}$ to be a curl of a scalar field. Following [26], we meet both these conditions by using

$$\mathbf{w} = \nabla \times (\psi \hat{z}), \tag{5}$$

with

$$\psi = \sin(\pi \underline{t})(\underline{x}^2 - 1)^p (\underline{y}^2 - 1)^p \tag{6}$$

and $p = 6$, where the underbar represents the rescaled variables $\underline{x} = (x - x_k)/H_x$, $\underline{x} = (y - y_k)/H_y$ and $\underline{x} = (t - t_k)/H_t$. Eq. (5) implies that $\nabla \cdot \mathbf{w} = 0$, which facilitates the elimination of pressure (see below). The functional form of Eq. (6) guarantees that $\mathbf{w}$ and its derivatives vanish at the domain boundaries given a sufficiently large value of p. This is useful for the integration by parts:

$$u_0^k = \int_{\Omega_k} \mathbf{w} \cdot \nabla^2 \mathbf{u} \, \mathrm{d}\Omega \qquad = \int_{\Omega_k} \mathbf{u} \cdot \nabla^2 \mathbf{w} \, \mathrm{d}\Omega \tag{7}$$

$$u_1^k = \int_{\Omega_k} \mathbf{w} \cdot \partial_t \mathbf{u} \, \mathrm{d}\Omega \qquad = -\int_{\Omega_k} \mathbf{u} \cdot \partial_t \mathbf{w} \, \mathrm{d}\Omega \tag{8}$$

$$u_2^k = \int_{\Omega_k} \mathbf{w} \cdot (\mathbf{u} \cdot \nabla \mathbf{u}) \, \mathrm{d}\Omega = -\int_{\Omega_k} \mathbf{u} \cdot (\mathbf{u} \cdot \nabla) \mathbf{w} \, \mathrm{d}\Omega \tag{9}$$

The activity term integrates as

$$u_3^k = \int_{\Omega_k} \mathbf{w} \cdot (\nabla \cdot \mathbf{Q}) \, \mathrm{d}\Omega = -\int_{\Omega_k} \mathbf{Q} \colon (\nabla \mathbf{w}) \, \mathrm{d}\Omega \tag{10}$$

while the pressure term integrates out to zero:

$$\int_{\Omega_k} \mathbf{w} \cdot \nabla P \, \mathrm{d}\Omega = -\int_{\Omega_k} P \nabla \cdot \mathbf{w} \, \mathrm{d}\Omega = 0 \tag{11}$$

The above indicates we need $p \geq 3$. While we use $p = 6$, values between 4-10 give similar results.

Algorithm 1: Hierarchical Ridge Regression (HRidge)

Result: $\mathbf{A}, r2s = \texttt{HRidge}(\mathbf{F}, \dot{\mathbf{X}}, \lambda_2)$
$m = \text{size}(\mathbf{F})[2]$; `// Total number of terms`
$F_n = \texttt{norm}(\mathbf{F})$; `// Column-wise norm`
$\mathbf{F}_0 = \mathbf{F}/F_n$; `// Normalize the terms for comparison`
$\mathbf{A} \leftarrow \text{Vector(size: } m \times m)$; `// Store optimal model at all hierarchies`
$r2s \leftarrow \text{Vector(size: } m)$; `// Store model accuracies at all hierarchies`
$\mathbf{a} = \text{argmin}_{\vec{a}} \left(\|\dot{\mathbf{X}} - \mathbf{F}_0 \cdot \mathbf{a}\|_2 + \lambda_2 \|\mathbf{a}\|_2 \right)$; `// Initial guess using Ridge`
$\mathbf{A}[:, m] = \mathbf{a}/F_n$; `// Store un-normalized coefficients`
$r2s[m] = \texttt{rsquared}(\mathbf{F}_0, \dot{\mathbf{X}}, \mathbf{a})$
while $len(\mathbf{a}) > 1$ **do**
  $j_0 = \text{argmin}_j |a_j|$; `// Find the position of the smallest coefficient`
  $\mathbf{a} = \mathbf{a} \setminus \{a_{j_0}\}$; `// Remove smallest coefficient`
  $\mathbf{F}_0 = \mathbf{F}_0 \setminus \mathbf{F}_0[:, j_0]$; `// Remove vector corresponding to it`
  $\mathbf{a} = \text{argmin}_{\vec{a}} \left(\|\dot{\mathbf{X}} - \mathbf{F}_0 \cdot \mathbf{a}\|_2 + \lambda_2 \|\mathbf{a}\|_2 \right)$;
  $k = len(\mathbf{a})$;
  $\mathbf{A}[:, k] = \mathbf{a}/F_n$;
  $r2s[k] = \texttt{rsquared}(\mathbf{F}_0, \dot{\mathbf{X}}, \mathbf{a})$;

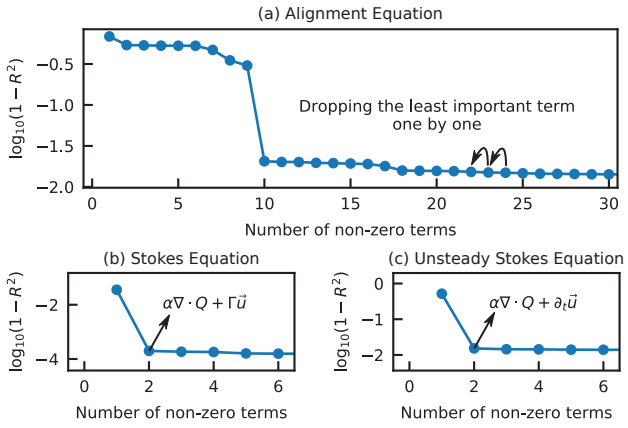

FIG. S1. Benchmarking the model discovery framework from continuum simulation data. *Optimality curves*: $\log(1-R^2)$ as a function of number of non-zero terms in the model for (a) the Q_{xx} component of Eq. (2) of the Main text, (b) the Stokes equation, and (c) the unsteady Stokes equation. (a): Starting with the full library, the least important term is eliminated one-by-one to obtain a hierarchy of models. The R^2 value deteriorates when terms used in the computational model begin to be eliminated. (b), (c): Same as (a) for the flow equation, Eq. (4) of the Main text, excluding the pressure.

The derivative on $\mathbf{Q}$ in term 4 in Table S1 cannot be fully transferred to $\mathbf{w}$ because of the non-linearity, so we integrate that term directly. Similarly, terms 5–7 are integrated directly as they do not contain any derivatives.

In each case, we perform sparse regression on the resulting linear system using our algorithm that we call 'Heirarchical Ridgre Regression (HRidge)' (see Algorithm 1). Briefly, we perform Ridge (or least-squares) regression on the normalized system and remove the term with the smallest coefficient one-by-one. We use $\lambda_2 = 10^{-5}$ for the Ridge regression, whereas $\lambda_2 = 0$ yields the least square result. See Fig. S1 for the result on the numerical data.

We benchmark these methods against varying noise levels as well as varying window sizes. To find the appropriate window size, we benchmark with a numerical data set of size $512 \times 512 \times 500$ with 5% added noise. We measure the R^2 value of the fit as well as the average % error in the coefficients in the optimal model (if found correctly) as a function of the integration window size (Fig. S3). For the integral formulation (Fig. S3 left), we find that a small window size is sufficient to mitigate the noise. For the experimental dataset results in Main Text Fig. 1, we use a window size of $25 \times 25 \times 25$ pixels, or $73\,\mu\text{m} \times 73\,\mu\text{m} \times 12\,\text{s}$. In the weak formulation, larger window sizes are needed to better sample the test function [26, 27]. We find that a window that is almost as large as the field of view in space, and ~ 5 times the velocity correlation time in the time dimension [27] works well. We perform a similar analysis with the experimental data-sets to choose the window size. To avoid self-selection, we use the largest of the correlation times from the data-sets to set the window size. In Main Text Fig. 1, we use a window size of $205 \times 205 \times 605$ pixels, or $\sim 600\,\mu\text{m} \times 600\,\mu\text{m} \times 300\,\text{s}$. As this window size is large, we take $m = 50$ measurements for the weak form.

We now investigate the same two parameters, namely the R^2 and the error in coefficients, but with varying noise levels (Fig. S4). We find that the error increases much faster with noise for the flow equation in the integral formulation, but stays low for the weak formulation. The error for the orientation equation also grows, but remains small for reasonable noise levels ($\tilde{5}\%$). These two analyses, and our results in Main Text Fig. 1 indicate that the integral formulation is sufficient for the Q-tensor equation, while the weak formulation works adequately for the flow equation.

APPENDIX C: CONTINUUM SIMULATIONS

For the Q-tensor, we use the simple form (in non-dimensionalized units)

$$\partial_t \mathbf{Q} + \nabla \cdot (\mathbf{u}\mathbf{Q}) + (\mathbf{\Omega} \cdot \mathbf{Q} - \mathbf{Q} \cdot \mathbf{\Omega}) = \lambda \mathbf{E} + \mathbf{H} \qquad (12)$$

with

$$\Omega_{ij} = \frac{1}{2}(\partial_i u_j - \partial_j u_i)$$
$$E_{ij} = \frac{1}{2}(\partial_i u_j + \partial_j u_i)$$
$$H_{ij} = (-a_2 - a_4 Q_{kl} Q_{lk}) Q_{ij} + K \partial_k \partial_k Q_{ij}$$

where λ is the flow alignment parameter, K is the elastic modulus and $a_2 < 0, a_4 > 0$ drive the system to the nematic phase. The two different flow equations used are

$$\eta \nabla^2 \mathbf{u} = \nabla P + \Gamma \mathbf{u} + \alpha \nabla \cdot \mathbf{Q} \quad \text{(Stokes)} \tag{13}$$
$$\rho \partial_t \mathbf{u} = \eta \nabla^2 \mathbf{u} - \nabla P - \alpha \nabla \cdot \mathbf{Q} \quad \text{(Unsteady Stokes)} \tag{14}$$
$$\nabla \cdot \mathbf{u} = 0 \tag{15}$$

Here, η is the viscosity, α is the activity and Γ is the substrate friction. We set $\rho = 1$ for the unsteady Stokes equation.

For the Stokes equation Eq. (13), we use a semi-implicit finite difference time stepping scheme based on a convex splitting of the nematic free energy [21, 28]. To solve the Stokes equation with incompressibility, we implement a Vanka type box smoothing algorithm on a staggered grid [21, 29]. The solution at each time-step is found using Gauss-Seidel relaxation iterations, and the rate of convergence to the solution is accelerated by using a multigrid method [21]. The simulation codes are all in-house and are written in C. We solve the equations in a square domain of size 200×200 (in simulation units) with periodic boundary conditions. We sample the evolution at a time step of 1 (in simulation units) on a rectangular grid with $\mathrm{d}x \sim 0.4$. For the Unsteady Stokes, the equations are solved using an explicit Adams-Bashforth method. we sample the data at a time-step of 0.02 on a rectangular grid with $\mathrm{d}x \sim 0.2$. The parameters used in both cases are documented in Table S2.

The equation for Q_{xx} has 10 terms:

$$\begin{aligned}\partial_t Q_{xx} = &- u_x \partial_x Q_{xx} - u_y \partial_y Q_{xx} - Q_{xy} \partial_x u_y + Q_{xy} \partial_y u_x \\ &+ \lambda \partial_x u_x - a_2 Q_{xx} - 2 a_4 Q_{xx}^3 - 2 a_4 Q_{xy}^2 Q_{xx} \\ &+ K \partial_x^2 Q_{xx} + K \partial_y^2 Q_{xx}\end{aligned} \tag{16}$$

The flow equations in the integral formulation are used for Fig. S4, and have the form of vorticity equations (obtained by taking the curl of Eq. (13) and Eq. (14)):

$$\eta \nabla^2 \omega = \Gamma \omega + \alpha \partial_x^2 Q_{xy} - 2\alpha \partial_x \partial_y Q_{xx} - \alpha \partial_y^2 Q_{xy}$$
$$\eta \nabla^2 \omega = \partial_t \omega + \alpha \partial_x^2 Q_{xy} - 2\alpha \partial_x \partial_y Q_{xx} - \alpha \partial_y^2 Q_{xy}$$

However, for the weak formulation, we obtain two terms each, after the pressure is eliminated as described earlier:

$$\eta \int_{\Omega_k} \mathbf{w} \cdot \nabla^2 \mathbf{u} = \alpha \int_{\Omega_k} \mathbf{w} \cdot \nabla \cdot \mathbf{Q} + \Gamma \int_{\Omega_k} \mathbf{w} \cdot \mathbf{u}$$
$$\eta \int_{\Omega_k} \mathbf{w} \cdot \nabla^2 \mathbf{u} = \alpha \int_{\Omega_k} \mathbf{w} \cdot \nabla \cdot \mathbf{Q} + \int_{\Omega_k} \mathbf{w} \cdot \partial_t \mathbf{u}$$

Parameter	Stokes	Unsteady Stokes
η	1	1
K	1	1
α	0.3	4.0
Γ	0.03	N.A.
a_2	-0.3	-16
a_4	1.36	32

TABLE S2. Parameters used for the Stokes and Unsteady Stokes simulations.

APPENDIX D: EXPERIMENTAL METHODS

The active nematic samples were assembled following previously established methods [30, 31]. The active mix consisted of microtubules, kinesin motor clusters, depleting agent, and an ATP regeneration system. Tubulin was purified from bovine brain, labeled with Alexa 647 dye, and polymerized in the presence of GMPCPP [32, 33]. The final concentration of polymerized microtubules in the active sample was 1.31 mg/ml. A truncated and biotinylated version of Kinesin-1 (K401-BCCP-HIS) was expressed in *E. Coli* and purified using immobilized metal affinity chromatography [34]. Motor clusters were formed by incubating 11 µL of the biotinylated kinesin (8.2 µM) with 5 µL of streptavidin (2.1 µM) on ice for 30 min in the presence of DTT (170 µM). In the active sample, this mixture was diluted to a final concentration of 140 nM kinesin and 70 nM streptavidin. Polyethylene glycol (35000 kDa, 1%) was used to induce microtubule bundling. A biochemical regeneration system consisting of adenosine triphosphate (ATP, 25 µM-1.4 mM), phosphoenol pyruvate (PEP, 26 mM), and pyruvate kinase/lactic dehydrogenase (PK/LDH) kept ATP concentration constant. Lastly, an oxygen scavenging system consisting of glucose (0.67 mg/ml), glucose oxidase (0.08 mg/ml), catalase (0.4 mg/ml), DTT (5.6 mM), and Trolox (2 mM) was used to minimize sample bleaching. The components of the active mix were combined in M2B buffer (80 mM PIPES, 1 mM EGTA, 2 mM MgCl2, pH 6.8).

Flow chambers were created with Parafilm sandwiched between a glass slide and a coverslip. The glass slide was made hydrophobic with a Rain-X coating, and the coverslip was passivated with acrylamide coating [35]. To assemble an active nematic, the chamber was first filled with HFE oil containing fluoro-surfactant (0.5% w/w, RAN Biotech), followed by the active mixture. The sample was sealed with UV glue (Norland optical adhesive). The active nematic sedimented to the oil-water interface and reached a steady state after about an hour, and was then imaged on a spinning disk confocal microscope using an Hamamatsu Orca-Fusion BT CMOS camera and 20× magnification. For each sample, a sequence of 10000 images was acquired at 2 frames/sec, except for the 25 μM ATP samples, for which 1000 frames were acquired

at 0.1 frames/sec.

High-resolution small field of view (HR-SF) data

An additional dataset was taken using a combination of LC-PolScope microscopy and dilute-labeled fluorescent MTs (see Fig. S2). With LC-PolScope microscopy, the orientation field is obtained from birefringence information of polarized light passing through the MT filaments that make up the nematic layer [31, 36]. This allows the orientation field to be measured on MTs that are not fluorescent. However, a small fraction of MTs in the sample were fluorescently labeled, and wide-field epifluorescence images were acquired simultaneously with the birefringence data. We believe that with wide-field microscopy, PIV on dilute-labeled MTs is more accurate than on fully-labeled MTs, due to the difficulty of detecting velocity in the direction of the elongated MT bundles.

Using the integral formulation with a window size of $5 \times 5 \times 5$ pixels, we obtain the equation

$$\begin{aligned}\partial_t Q_{xx} = &- (1.07\ u_x\partial_x + 1.08\ u_y\partial_y)\ Q_{xx} \\ &- (1.03\partial_x u_y - 1.04\partial_y u_x)\ Q_{xy} \\ &+ 0.99\partial_x u_x \\ &- 2Q_{xx}\{2.01 Q_{xx}\partial_x u_x \\ &\quad + Q_{xy}(1.00\partial_x u_y + 1.05\partial_y u_x)\}\end{aligned}$$

with a high R^2 value of 0.97. This provides further strong evidence for the discovered model. The flow analysis also yields a model similar to Main Text Eq. (5), but with a low R^2, due to the limitation on the window size due to the small field of view (see Fig. S3).

APPENDIX E: ANALYSIS

Orientation and velocity fields

The orientation and velocity fields were computed simultaneously on the fluorescent images obtained from spinning disk confocal microscopy. The orientation fields were measured using an in-house structure-tensor-based code (written in MATLAB) on the fluorescence images. The molecular tensor ($\mathbf{nn} - 1/2\mathbf{I}$) was computed from the orientation data, and was then coarse-grained with a Gaussian smoothing filter with $\sigma = 10$ pixels to obtain the Q-tensor. The velocity fields were measured using particle-image velocimetry implemented by the MATLAB-based PIVLab software. The fields computed by PIVLab were post-processed using a Direct Cosine Transform - Penalized Least Squares (DCT-PLS) approach that validates the raw data, replaces the spurious and missing vectors and does some smoothing. [37]. The orientation fields were computed on a high resolution grid of 1152×1152 pixels, while the velocity fields were on a coarser grid of 71×71 pixels. Therefore, both fields were interpolated on an intermediate grid of 256×256 pixels.

Defect detection and tracking

To locate the defects, we compute a map of the signed winding number $w = 1/(2\pi) \oint \nabla\theta \cdot \mathrm{d}\vec{s}$ at every point in space [11, 38] with an integration ring of radius of 5 pixels. The winding number is zero everywhere except at the defect locations [31, 39]. To eliminate spurious defects, we filter out regions with a non-zero winding number that are smaller than 60 squared pixels in area.

Once the locations of the defects are obtained, the $+1/2$ defects are tracked using the open source software Trackpy [40] using a `search_range` value of 20 pixels. The trajectories thus obtained are further filtered with a threshold of minimum three frames of survival.

Elastic terms

To investigate the elastic terms, we consider the HR-SF data set. For this data, the term $K_0\partial_x^2 Q_{xx}$ appears at $n = n^* + 1$, with $K_0 \sim 2\,\mu\mathrm{m}^2\,\mathrm{s}^{-1}$. However, the counterpart $\partial_y^2 Q_{xx}$ does not appear even at $n \sim 30$, which suggests the estimate of this coefficient is highly unreliable. As a test, we compare the contribution of contribution of the term $K_0\nabla^2 Q_{xx}$ to $\partial_t Q_{xx}$ with flow-coupling terms (see Fig. S5 *left* for the average trend vs time and Movie S2 for a spatial plot of the same). We find that the magnitude of $K_0\nabla^2 Q_{xx}$ is up to an order of magnitude smaller than the other terms. However, the elastic terms are known to play an important role in determining the structure of defects or low order regions. Therefore, we specifically compare the averaged values in regions with $S < 0.75$ (see Fig. S5 *right*). This measurement shows that the while the dynamics near the defect is dominated by convection alone, likely due to the self-propulsion of the $+1/2$ defects, the contribution of $K_0\nabla^2 Q_{xx}$ is now comparable to the other flow-coupling terms. However, the inset of Fig. S5 *right* shows that such low order regions only comprise $\lessapprox 5\%$ of the total data. This suggests that to reliably identify the elastic terms, it will be important to aquire extra data near defects.

MOVIE DESCRIPTIONS

Movie S1 Comparison of the fluorescence intensity and the scalar order parameter. *Top left* Normalized fluorescence intensity, proportional to the microtubule density. *Top right* Computed scalar order parameter S. Color goes from 0 (black) to 1 (white). The orange square in each of these two panels shows the region which is

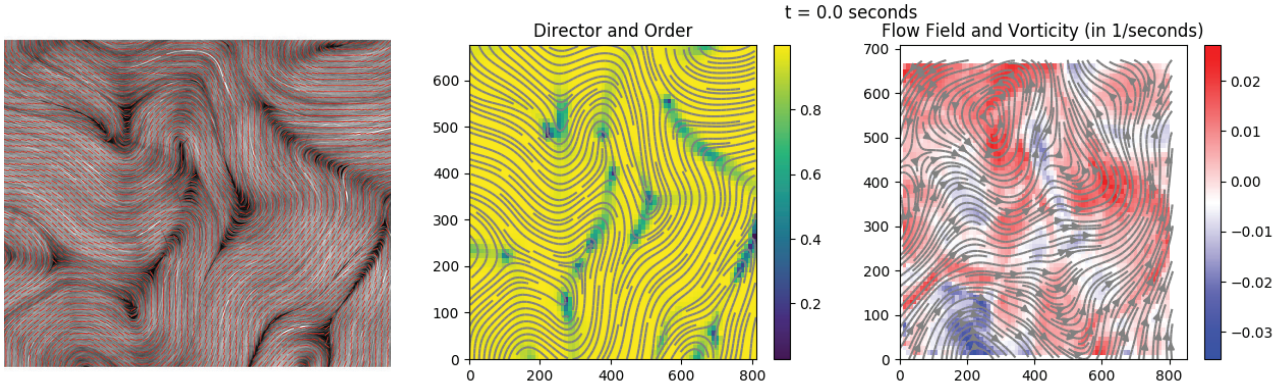

FIG. S2. A representative snapshot of the high-resolution small field of view (HR-SF) PolScope dataset. *(left)* Retardance image of the sample, with the orientation obtained using PolScope microscopy overlaid in red. *(middle)* The computed director field and scalar order parameter S, and *(right)* the flow field and vorticity ω for the frame. The lengths indicated in the right two figures are in µM.

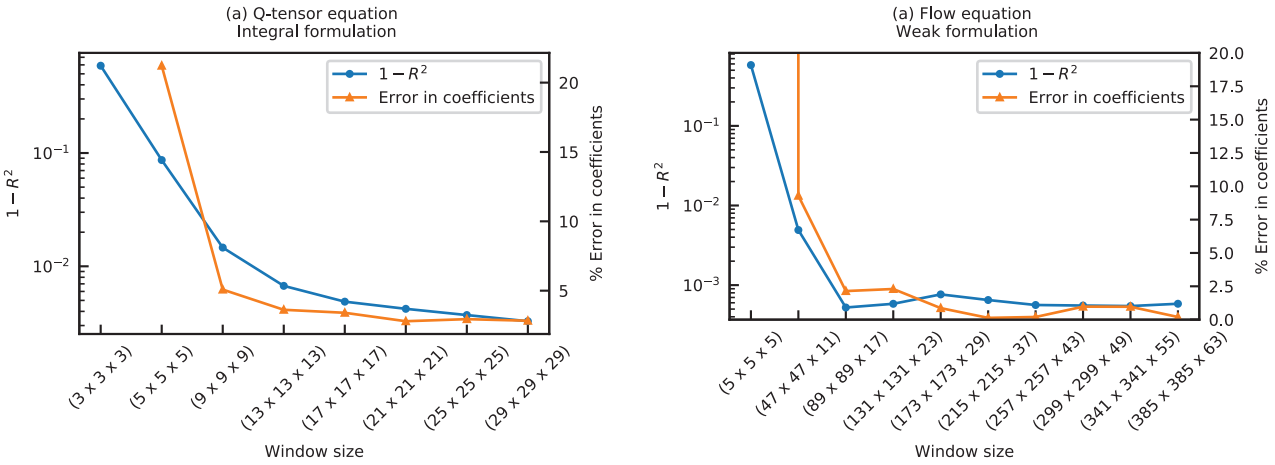

FIG. S3. Performance on simulation data as a function of window size. $(1-R^2)$ (blue circles) and average % error in coefficients (orange triangles) for the (left) Q-tensor equation using the integral formulation and (right) the flow equation using the weak formulation. The window sizes are listed in pixels, while the simulation data used had a domain size of $512 \times 512 \times 500$ pixels.

zoomed-in in the corresponding middle panel. *Mid left* Zoomed-in region showing the fluorescence signal, thresholded with a cutoff of 0.5 (black indicating less than 0.5 and white indicating more than 0.5). *Middle right* Same as middle left for S. *Bottom* Spatial correlation of the thresholded intensity and thresholded S quantifying how well S captures the low density variations near the defects.

Movie S2 Spatial map of various terms in the Q_{xx} equation for the HR-SF dataset, compared to the elastic term $K_0\nabla^2 Q_{xx}$, with K_0 found from the model at $n = n^* + 1$. The displayed terms are: $\partial_t Q_{xx}$, $\mathbf{u}\cdot\nabla Q_{xx}$ (convection), $(\mathbf{\Omega Q} - \mathbf{Q\Omega})_{xx}$ (rotation), E_{xx} (flow alignment), $(\mathbf{Q}\colon\nabla\mathbf{u})Q_{xx}$ (higher order flow alignment) and $K\nabla^2 Q_{xx}$. The colorbar is same for all the terms, showing that the elastic term is small everywhere except near defects.

Movie S3 Direct comparison of terms arising in the Stokes equation, for the 100 µM ATP data set. *Top left* Spatial map of $\alpha/\eta\nabla\times\nabla\cdot\mathbf{Q}$ computed using the smoothed Q-tensor and the α/η value obtained from our framework. *Top right* Spatial map of $\nabla\times\nabla^2\mathbf{u}$ computed using the smoothed velocity field. *Bottom* The spatial correlation of the two as a function of time. In both cases, the smoothing is performed using a Gaussian filter with a width of 10 pixels for each component of the fields. This additional smoothing reduces the noise arising from taking multiple derivatives of the experimental data.

Movie S4 Same as Movie S3 for the 500 µM ATP data set.

25 µM	50 µM	100 µM	200 µM	500 µM	HR-SF	Physical origin
$n^* = 4$	$n^* = 6$	$n^* = 6$	$n^* = 7$	$n^* = 6$	$n^* = 8$	
$R^2 = 0.88$	$R^2 = 0.84$	$R^2 = 0.88$	$R^2 = 0.72$	$R^2 = 0.75$	$R^2 = 0.97$	
$\partial_t Q_{xx} =$ $-(0.97)u_x\partial_x Q_{xx}$ $-(1.00)u_y\partial_y Q_{xx}$	$\partial_t Q_{xx} =$ $-(0.95)u_x\partial_x Q_{xx}$ $-(1.01)u_y\partial_y Q_{xx}$	$\partial_t Q_{xx} =$ $-(0.89)u_x\partial_x Q_{xx}$ $-(0.95)u_y\partial_y Q_{xx}$	$\partial_t Q_{xx} =$ $-(0.87)u_x\partial_x Q_{xx}$ $-(0.90)u_y\partial_y Q_{xx}$	$\partial_t Q_{xx} =$ $-(0.74)u_x\partial_x Q_{xx}$ $-(0.80)u_y\partial_y Q_{xx}$	$\partial_t Q_{xx} =$ $-(1.07)u_x\partial_x Q_{xx}$ $-(1.08)u_y\partial_y Q_{xx}$	Convection
$-(0.95)Q_{xy}\partial_x u_y$ $+(0.93)Q_{xy}\partial_y u_x$	$-(0.98)Q_{xy}\partial_x u_y$ $+(1.01)Q_{xy}\partial_y u_x$	$-(1.01)Q_{xy}\partial_x u_y$ $+(1.12)Q_{xy}\partial_y u_x$	$-(0.91)Q_{xy}\partial_x u_y$ $+(0.91)Q_{xy}\partial_y u_x$	$-(0.88)Q_{xy}\partial_x u_y$ $+(0.96)Q_{xy}\partial_y u_x$	$-(1.03)Q_{xy}\partial_x u_y$ $+(1.04)Q_{xy}\partial_y u_x$	Rotation
-	$+(0.66)\partial_x u_x$	$+(0.98)\partial_x u_x$	-	$+(0.63)\partial_x u_x$	$+(0.99)\partial_x u_x$	Flow alignment
-	-	$-(5.33)Q_{xx}^2\partial_x u_x$	$+(4.63)Q_{xy}^2\partial_x u_x$	-	$-(4.02)Q_{xx}^2\partial_x u_x$ $-(2.0)Q_{xx}Q_{xy}\partial_x u_y$ $-(2.10)Q_{xx}Q_{xy}\partial_y u_x$	Higher order flow alignment
-	$-(0.03)Q_{xx}$	-	$-(0.05)Q_{xx}$ $+(0.29)Q_{xx}^3$	$-(0.16)Q_{xx}$	-	Bulk free energy

TABLE S3. Optimal model for Q_{xx} for various data sets. The rows show different data sets and the columns show the values of the number of terms n^* in the optimal model, the corresponding R^2 value, and the terms appearing in the model, organized by physical origin. We omit the 1000 µM data set for which $R^2 < 0.5$.

25 µM	50 µM	100 µM	200 µM	500 µM	HR-SF	Physical origin
$n^\dagger = 8$	$n^\dagger = 11$	$n^\dagger = 11$	$n^\dagger = 9$	$n^\dagger = 8$	$n^\dagger = 14$	
$R^2 = 0.96$	$R^2 = 0.88$	$R^2 = 0.96$	$R^2 = 0.76$	$R^2 = 0.81$	$R^2 = 0.98$	
$\partial_t Q_{xx} =$	$\partial_t Q_{xx} =$	$\partial_t Q_{xx} =$	$\partial_t Q_{xx} =$	$\partial_t Q_{xx} =$	$\partial_t Q_{xx} =$	
$-(0.94)u_x\partial_x Q_{xx}$ $-(1.07)u_y\partial_y Q_{xx}$	$-(0.97)u_x\partial_x Q_{xx}$ $-(1.00)u_y\partial_y Q_{xx}$	$-(0.97)u_x\partial_x Q_{xx}$ $-(0.97)u_y\partial_y Q_{xx}$	$-(0.90)u_x\partial_x Q_{xx}$ $-(0.90)u_y\partial_y Q_{xx}$	$-(0.81)u_x\partial_x Q_{xx}$ $-(0.87)u_y\partial_y Q_{xx}$	$-(1.09)u_x\partial_x Q_{xx}$ $-(1.07)u_y\partial_y Q_{xx}$	Convection
$-(1.10)Q_{xy}\partial_x u_y$ $+(1.00)Q_{xy}\partial_y u_x$	$-(0.90)Q_{xy}\partial_x u_y$ $+(1.03)Q_{xy}\partial_y u_x$	$-(0.93)Q_{xy}\partial_x u_y$ $+(1.01)Q_{xy}\partial_y u_x$	$-(0.78)Q_{xy}\partial_x u_y$ $+(0.83)Q_{xy}\partial_y u_x$	$-(0.80)Q_{xy}\partial_x u_y$ $+(0.86)Q_{xy}\partial_y u_x$	$-(1.00)Q_{xy}\partial_x u_y$ $+(1.03)Q_{xy}\partial_y u_x$	Rotation
-	$+(0.37)\partial_x u_x$ $+(0.05)\partial_y u_x$	$+(0.93)\partial_x u_x$	-	$+(0.60)\partial_x u_x$	$+(0.96)\partial_x u_x$	Flow alignment
$+(4.25)Q_{xy}^2\partial_x u_x$ $-(2.10)Q_{xx}Q_{xy}\partial_x u_y$ $-(2.22)Q_{xx}Q_{xy}\partial_y u_x$	$+(2.62)Q_{xy}^2\partial_x u_x$	$-(3.83)Q_{xx}^2\partial_x u_x$ $-(1.97)Q_{xx}Q_{xy}\partial_x u_y$ $-(1.96)Q_{xx}Q_{xy}\partial_y u_x$	$+(4.40)Q_{xy}^2\partial_x u_x$	-	$-(3.98)Q_{xx}^2\partial_x u_x$ $-(2.07)Q_{xx}Q_{xy}\partial_x u_y$ $-(2.16)Q_{xx}Q_{xy}\partial_y u_x$	Higher order flow alignment
-	$-(0.04)Q_{xx}$ $+(0.12)Q_{xx}^3$	-	$-(0.05)Q_{xx}$ $+(0.29)Q_{xx}^3$	$-(0.16)Q_{xx}$	$+(0.007)Q_{xx}$ $-(0.027)Q_{xx}^2 Q_{xx}$ $-(0.026)Q_{xy}^2 Q_{xx}$	Bulk free energy
$-(0.93)Q_{xy}\partial_y^2 Q_{xy}$	$-(21.49)(\partial_x Q_{xy})^2$ $+(18.05)(\partial_y Q_{xy})^2$	$-(17.24)(\partial_x Q_{xy})^2$ $+(16.27)(\partial_y Q_{xy})^2$ $+(29.52)\partial_y Q_{xx}\partial_x Q_{xy}$	$-(30.43)(\partial_x Q_{xy})^2$ $+(28.13)(\partial_y Q_{xy})^2$	$-(84.98)(\partial_x Q_{xy})^2$ $+(86.29)(\partial_y Q_{xy})^2$	$+(6.39)\partial_x^2 Q_{xx}$ $-(17.97)Q_{xx}^2\partial_x^2 Q_{xx}$ $-(29.05)Q_{xy}^2\partial_x^2 Q_{xx}$	Distortion free energy

TABLE S4. Model for Q_{xx} at secondary shoulders for various data sets. The rows show different data sets and the columns show the values of the number of terms $n^\dagger$ for the secondary shoulders, the corresponding R^2 value, and the terms appearing in the model, organized by physical origin. We omit the 1000 µM data set for which $R^2 < 0.5$.

* chaitanya.joshi@tufts.edu
† zdogic@physics.ucsb.edu
‡ aparna@brandeis.edu
§ hagan@brandeis.edu

[1] G. P. G. de and J. Prost, *The physics of liquid crystals* (Clarendon Press, 1993).
[2] A. N. Beris and B. J. Edwards, *Thermodynamics of flowing systems with internal microstructure* (Oxford University Press, 1994).
[3] S. Ramaswamy, Annual Review of Condensed Matter Physics **1**, 323 (2010).
[4] M. C. Marchetti, J. F. Joanny, S. Ramaswamy, T. B. Liverpool, J. Prost, M. Rao, R. A. Simha, and M. Curie, Reviews of Modern Physics **85**, 1143 (2013), arXiv:1207.2929.
[5] L. Giomi, L. Mahadevan, B. Chakraborty, and M. F. Hagan, Nonlinearity **25**, 2245 (2012), arXiv:arXiv:1110.4338v1.
[6] L. Giomi, M. J. Bowick, P. Mishra, R. Sknepnek, and M. Cristina Marchetti, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **372**, 20130365 (2014).
[7] S. P. Thampi, R. Golestanian, and J. M. Yeomans, Phys. Rev. E **90**, 062307 (2014).
[8] T. Gao, R. Blackwell, M. A. Glaser, M. D. Betterton, and M. J. Shelley, Physical Review Letters **114**, 048101 (2015), arXiv:1401.8059.
[9] L. Giomi, Physical Review X **5**, 031003 (2015).
[10] A. Maitra, P. Srivastava, M. Cristina Marchetti, J. S. Lintuvuori, S. Ramaswamy, and M. Lenz, Proceedings of the National Academy of Sciences of the United States of America **115**, 6934 (2018).
[11] M. M. Norton, A. Baskaran, A. Opathalage, B. Langeslay, S. Fraden, A. Baskaran, and M. F. Hagan, Physical Review E **97**, 1 (2018).
[12] E. J. Hemingway, P. Mishra, M. C. Marchetti, and S. M. Fielding, Soft Matter **12**, 7943 (2016).
[13] S. Ngo, A. Peshkov, I. S. Aranson, E. Bertin, F. Ginelli, and H. Chaté, Physical Review Letters **113** (2014), 10.1103/PhysRevLett.113.038302.
[14] L. Giomi, M. J. Bowick, X. Ma, and M. C. Marchetti, Phys. Rev. Lett. **110**, 228101 (2013).
[15] L. Giomi and A. DeSimone, Physical Review Letters **112**, 147802 (2014).
[16] S. P. Thampi, A. Doostmohammadi, R. Golestanian, and J. M. Yeomans, EPL (Europhysics Letters) **112**, 28004 (2015).
[17] The second order term — $\lambda\{\mathbf{Q}\cdot\mathbf{E}+\mathbf{E}\cdot\mathbf{Q}-2/d\,\mathrm{Tr}(\mathbf{Q}\cdot\mathbf{E})\mathbf{I}\}$ [18, 21]—vanishes identically in 2D.
[18] A. Doostmohammadi, J. Ignés-Mullol, J. M. Yeomans, and F. Sagués, Nature Communications **9**, 3246 (2018).

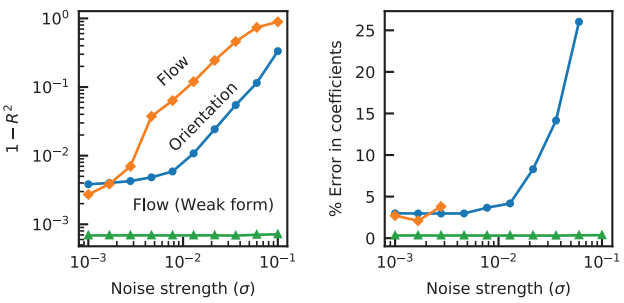

FIG. S4. Performance on simulation data as a function of added noise. (left) $(1-R^2)$ as a function of noise strength for the orientation equation (blue circles), flow equation (orange diamonds) and the flow equation in the weak form (green triangles). The first two are obtained using the weak formulation. (right) The corresponding average % error in the fitted coefficients. For the points absent on the plot, the obtained optimal model did not match with the ground truth. A $5\times5\times5$ volume was used for the integral formulation, whereas a $205\times205\times65$ window was used for the weak formulation.

[19] G. Duclos, R. Adkins, D. Banerjee, M. S. E. Peterson, M. Varghese, I. Kolvin, A. Baskaran, R. A. Pelcovits, T. R. Powers, A. Baskaran, F. Toschi, M. F. Hagan, S. J. Streichan, V. Vitelli, D. A. Beller, and Z. Dogic, Science **367**, 1120 (2020).

[20] Z. Zhou, C. Joshi, R. Liu, M. M. Norton, L. Lemma, Z. Dogic, M. F. Hagan, S. Fraden, and P. Hong, Soft Matter **17**, 738 (2021).

[21] M. Varghese, A. Baskaran, M. F. Hagan, and A. Baskaran, Physical Review Letters **125**, 268003 (2020).

[22] C.-M. Koch and M. Wilczek, Physical Review Letters **127**, 268005 (2021).

[23] S. P. Thampi, A. Doostmohammadi, T. N. Shendruk, R. Golestanian, and J. M. Yeomans, Science Advances **2**, e1501854 (2016).

[24] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, Science Advances **3**, e1602614 (2017).

[25] E. P. Alves and F. Fiuza, arXiv:2011.01927 [astro-ph, physics:physics] (2020), arXiv: 2011.01927.

[26] P. A. K. Reinbold, D. R. Gurevich, and R. O. Grigoriev, Phys. Rev. E **101**, 010203 (2020).

[27] P. A. K. Reinbold, L. M. Kageorge, M. F. Schatz, and R. O. Grigoriev, Nature Communications **12**, 3219 (2021).

[28] J. Zhao and Q. Wang, Journal of Scientific Computing **68**, 1241 (2016).

[29] S. P. Vanka, Journal of Computational Physics **65**, 138 (1986).

[30] T. Sanchez, D. T. N. Chen, S. J. DeCamp, M. Heymann, and Z. Dogic, Nature **491**, 431 (2012), arXiv:1301.1122.

[31] S. J. DeCamp, G. S. Redner, A. Baskaran, M. F. Hagan, and Z. Dogic, Nature Materials **14**, 1110 (2015).

[32] M. Castoldi and A. V. Popov, Protein Expression and Purification **32**, 83 (2003).

[33] A. Hyman, D. Drechsel, D. Kellogg, S. Salser, K. Sawin, P. Steffen, L. Wordeman, and T. Mitchison, in *Methods in Enzymology*, Molecular Motors and the Cytoskeleton, Vol. 196 (Academic Press, 1991) pp. 478–485.

[34] R. Subramanian and J. Gelles, Journal of General Physiology **130**, 445 (2007).

[35] A. W. C. Lau, A. Prasad, and Z. Dogic, EPL (Europhysics Letters) **87**, 48006 (2009).

[36] R. Oldenbourg, Live cell imaging: a laboratory manual , 205 (2005).

[37] D. Garcia, Experiments in Fluids **50**, 1247 (2011).

[38] R. D. Kamien, Reviews of Modern Physics **74**, 953 (2002).

[39] P. W. Ellis, D. J. G. Pearce, Y.-W. Chang, G. Goldsztein, L. Giomi, and A. Fernandez-Nieves, Nature Physics **14**, 85 (2018).

[40] D. B. Allan, T. Caswell, N. C. Keim, C. M. van der Wel, and R. W. Verweij, "Soft-matter/trackpy: Trackpy v0.5.0," Zenodo (2021).

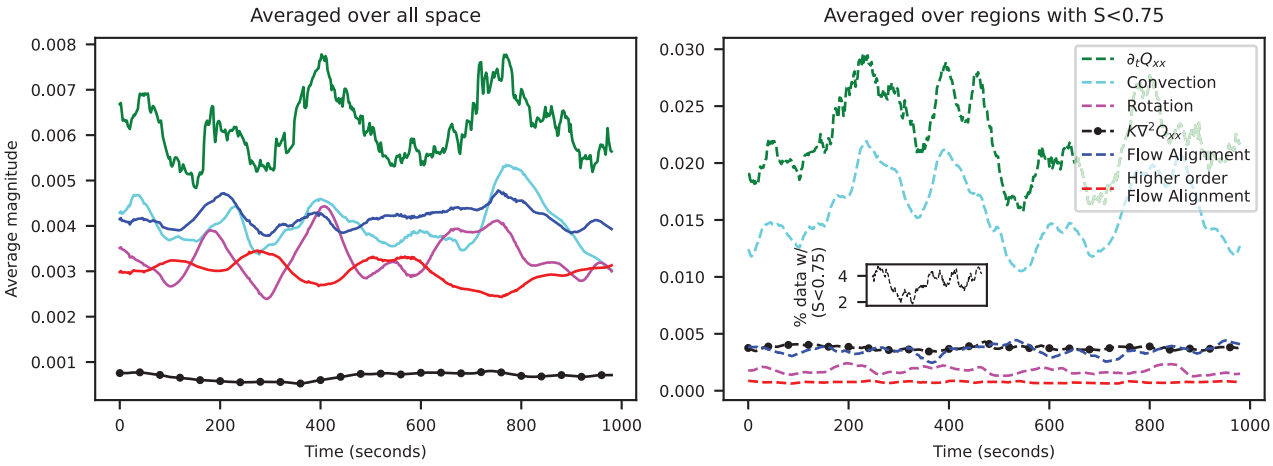

FIG. S5. Average magnitudes of various contributions to the alignment equation for the HR-SF data-set with magnitudes averaged over (left, solid lines) all space and (right, dashed lines) regions with low order ($S < 0.75$) . The value of K is obtained from the fit at $n = n^* + 1$. All quantities are rolling averages over a window of 20 seconds. (inset in right) fraction (%) of data coming from low order regions. The x axis is same as the main figure. We note that while the average magnitude of $K\nabla^2 Q_{xx}$ (black lines with circles) is much smaller than the other terms, it is comparable to flow-coupling terms near regions of distortion/low order (identified here as regions with $S < 0.75$, see right), but the fraction of data coming from these low order regions is $\lessapprox 5\%$ (see right inset), suggesting why the algorithm is not able to robustly identify the elastic term.