PNAS RESEARCH ARTICLE | APPLIED PHYSICAL SCIENCES

Deep learning probability flows and entropy production rates in active matter

Nicholas M. Boffi[a,1] and Eric Vanden-Eijnden[a]

Edited by Michael F. Hagan, Brandeis University, Waltham, MA; received October 18, 2023; accepted May 1, 2024, by Editorial Board Member Chris Jarzynski

Active matter systems, from self-propelled colloids to motile bacteria, are characterized by the conversion of free energy into useful work at the microscopic scale. They involve physics beyond the reach of equilibrium statistical mechanics, and a persistent challenge has been to understand the nature of their nonequilibrium states. The entropy production rate and the probability current provide quantitative ways to do so by measuring the breakdown of time-reversal symmetry. Yet, their efficient computation has remained elusive, as they depend on the system's unknown and high-dimensional probability density. Here, building upon recent advances in generative modeling, we develop a deep learning framework to estimate the score of this density. We show that the score, together with the microscopic equations of motion, gives access to the entropy production rate, the probability current, and their decomposition into local contributions from individual particles. To represent the score, we introduce a spatially local transformer network architecture that learns high-order interactions between particles while respecting their underlying permutation symmetry. We demonstrate the broad utility and scalability of the method by applying it to several high-dimensional systems of active particles undergoing motility-induced phase separation (MIPS). We show that a single network trained on a system of 4,096 particles at one packing fraction can generalize to other regions of the phase diagram, including to systems with as many as 32,768 particles. We use this observation to quantify the spatial structure of the departure from equilibrium in MIPS as a function of the number of particles and the packing fraction.

machine learning | scientific computing | active matter | stochastic thermodynamics

Significance

A long-standing challenge in statistical mechanics has been to quantify the nonequilibrium nature of active systems such as swarming bacteria, schooling fish, and flocking birds. The entropy production rate characterizes how far from equilibrium a system resides, while probability currents describe nonequilibrium transport processes that occur even at statistical steady state. Computation of these quantities has proved difficult, as they depend on the system's unknown high-dimensional probability density. Here, we introduce a broadly applicable deep learning framework that fuses recent advances in generative modeling with stochastic thermodynamics, yielding access to this canonically intractable density. Applying the method to phase-separated active systems, we find that nonequilibrium currents and the entropy production rate are dominated by interfacial contributions, confirming recent theoretical predictions.

Active matter systems are driven out of equilibrium by a continuous injection of energy at the microscopic scale of the constituent particles (1–3). The nonequilibrium nature of their dynamics manifests itself in the breakdown of time-reversal symmetry (TRS), which can be quantified by the global rate of entropy production (EPR) (4–7), and by the presence of probability currents at statistical steady state (8–10). Despite their wide recognition as quantities of fundamental importance, computing either the global EPR or the magnitude of the probability current has remained a long-standing challenge. At a fundamental level, both are defined via the microscopic density for the system (11, 12), which is generically unknown outside of a few simplistic cases due to its high dimensionality and its complexity (13).

The global EPR can in principle be computed directly from the microscopic equations of motion (14–18) by making use of the Crooks fluctuation theorem (19). However, this leads to a single number, which fails to quantify where TRS breaks down spatially in the system, and fails to reveal which particles are responsible. This issue can be addressed for active matter field theories, where a similar approach leads to a local, spatially dependent definition of the EPR (20, 21), but only after a coarse-graining of the microscopic dynamics. In general, methods based on the Crooks fluctuation theorem require a suitable definition of a time-reversed dynamics, which has been debated in the literature (22–24). The use of a time reversal can be avoided via the stochastic thermodynamics definition of the "entropy of the system" (7, 11), but doing so requires the logarithm of the system's microscopic density, which is unavailable outside of the simplest cases. An orthogonal approach makes use of data compression algorithms to compute the global (25) or local EPR (26), but these methods are only valid asymptotically in the limit of infinite system size, and it is difficult to understand what they compute away from this limit. Several methods have also been developed to infer a global measure of the probability current (27–29), but thus far have been restricted to low-dimensional systems. For a detailed coverage of the use of the EPR and steady-state currents to quantify nonequilibrium effects in active matter, we refer the reader to ref. 2.

Author affiliations: [a]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012

The authors declare no competing interest.

This article is a PNAS Direct Submission. M.F.H. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: boffi@cims.nyu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2318106121/-/DCSupplemental.

Published June 11, 2024.

Here, building upon recent advances in generative modeling (30–34), we tackle the challenging problem of estimating spatially local probability currents and entropy production rates (EPR) directly from their microscopic definitions. To this end, we develop a machine learning method that estimates the gradient of the logarithm of the system's probability density, which can be characterized as a solution to the many-body stationary Fokker–Planck equation (FPE). This quantity, known as the score function (30, 31), enters the definition of both the probability current and the EPR. We show how the method naturally decomposes the global EPR or probability current into microscopic contributions from the individual particles and their degrees of freedom, which enables us to identify spatial structure in the breakdown of TRS. To validate the accuracy of the learned solution, we develop diagnostics based on invariants of the stationary FPE that can be verified a posteriori.

We apply the method to several model systems involving active swimmers: two swimmers on the torus, where we can visualize the EPR and the probability flow across the entire phase space, a system of 64 swimmers in a harmonic trap, and a system of 4,096 swimmers undergoing motility-induced phase separation (MIPS) (35). For the MIPS system, we learn using a spatially local architecture that does not depend on the total number of swimmers, and we show that it can be extended to systems of up to 32,768 swimmers at values of the packing fraction that differ from those seen during training. Despite the high dimensionality of these latter examples, our approach provides us with a microscopic description of both the current and the local EPR. Importantly, this enables us to visualize the contributions of the individual particles directly without any need for averaging. We use this property to confirm theoretical predictions about the spatial features of entropy production in MIPS, such as concentration on the interface between the dilute and condensed phases (20, 26). Our main contributions can be summarized as follows:

1. We revisit the framework of stochastic thermodynamics and show how signatures of nonequilibrium behavior and lack of time reversibility, such as the probability current and the EPR, can be related to the score of the system's stationary probability density.
2. We show how to use machine learning tools from the field of generative modeling to estimate the score function from microscopic data (Fig. 1). To approximate this high-dimensional function accurately, we develop a transformer neural network architecture that incorporates spatial locality and permutation symmetry. This enables transferability to systems with differing numbers of particles or packing fractions than seen during training.
3. We illustrate the usefulness of the approach on systems involving active particles undergoing MIPS, where we show that the method can quantify the EPR at the individual particle level as a function of the activity, number of particles, or packing fraction. We confirm that entropy is dominantly produced at the interface between the cluster and the gas.

These contributions continue in a line of work that seeks to apply methods based on machine learning to high-dimensional problems in scientific computing (36–40), applied mathematics (34, 41–49), and the physical sciences (50–54). In particular, considerable research effort has been spent designing machine learning methods to compute solutions of the many-body Schrödinger equation (55–58); our work can be seen as an extension of this research effort to classical statistical mechanics and stochastic thermodynamics.

Stochastic Thermodynamics

Active Swimmers. As an application of our approach, we consider a suspension of N self-propelled particles in $d = 1$ or $d = 2$

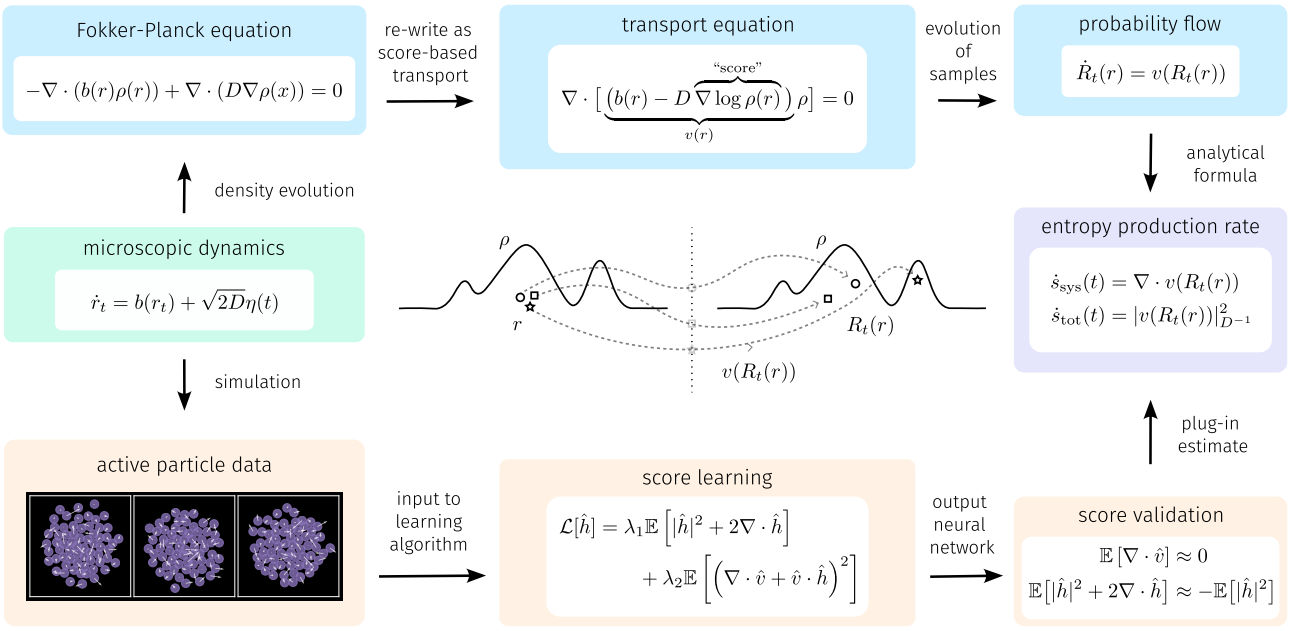

Fig. 1. Method overview. (Green) The starting point for our approach is a microscopic dynamics describing the evolution of a set of interacting active particles. (Purple) The target is estimation of several definitions of the EPR of the system, which we will accomplish by means of the probability flow. (Blue) Mathematically, our approach is built on viewing the system from the perspective of dynamical transport of measure. The microscopic stochastic dynamics induces an FPE for the high-dimensional density describing the configuration of the system. This FPE is equivalent to a transport equation that depends on the unknown "score" $\nabla \log \rho$ of the solution. The characteristics of this equation obey a probability flow ordinary differential equation, which gives immediate access to the EPR. (*Center*) Illustration of nonequilibrium transport of measure at stationarity. (Orange) Algorithmically, our method approximates the unknown score by machine learning over a dataset of microscopic particle data. The learned approximation can be validated a posteriori by checking invariants of the stationary FPE and can be plugged in directly to the definition of the EPR to obtain an estimate.

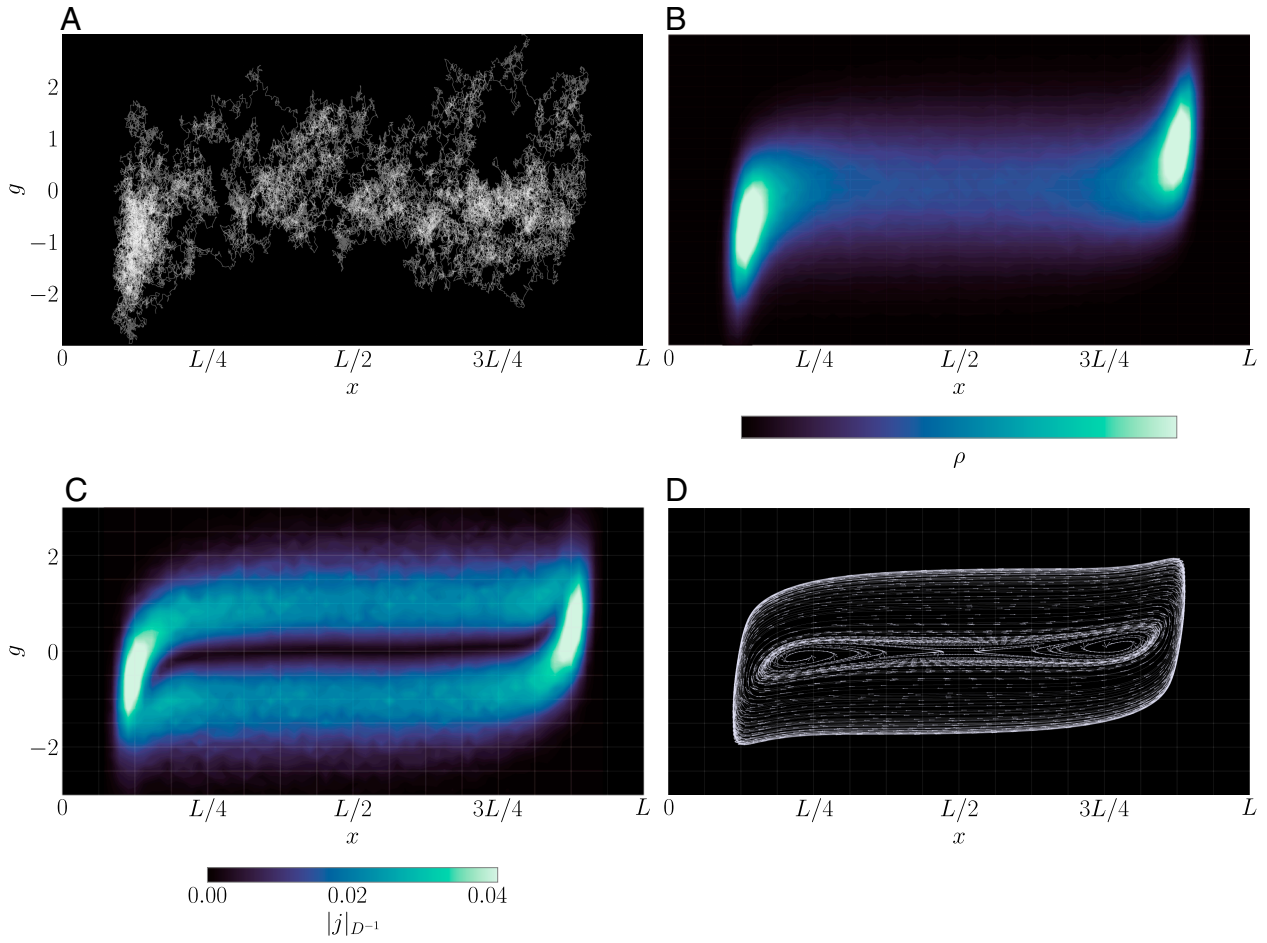

Fig. 2. Stochastic dynamics and probability flows. (*A*) Individual stochastic trajectories of Eq. **1** for $N = 2$ and $d = 1$ in the variables $x_t = x_t^2 - x_t^1$ and $g_t = g_t^2 - g_t^1$, with periodic boundary conditions on $[0, L]$. The trajectories (x_t, g_t) tend to accumulate in two clusters corresponding to situations where particle 1 is just in front of particle 2 or vice versa. This occurs because one particle catches up to the other in a typical trajectory (since either $|g_t^1| > |g_t^2|$ or $|g_t^1| < |g_t^2|$), but does not pass over it due to the short-range repulsive force between them. Random transitions between these modes occur when the magnitudes of $|g_t^1|$ and $|g_t^2|$ change order. (*B*) Stationary probability density function ρ of (x_t, g_t) confirming the metastability observed in (*A*): ρ is the solution of the stationary FPE (Eq. **3**). (*C*) Visualization of the (diffusion-weighted) norm of the probability current j, defined in Eq. **4**, over the phase space. The current is concentrated in the two modes but is also nonzero along transition pathways between them. (*D*) Phase portrait of the probability flow Eq. **7**. Similar to the stochastic trajectories in (*A*), the flow lines preserve the density ρ in (*B*), but are deterministic and interpretable, highlighting limit cycles within and between the two clusters. A movie of these limit cycles in a frame with one particle fixed is available at this link.

dimensions with translational degrees of freedom $x_t^i \in \mathbb{R}^d$ and orientational degrees of freedom $g_t^i \in \mathbb{R}^d$. Their dynamics is given by the so-called active Ornstein–Uhlenbeck or Gaussian colored-noise model (15, 23, 59–62):

$$\begin{aligned} \dot{x}_t^i &= \mu \sum_{j \neq i} f(x_t^i - x_t^j) + v_0 g_t^i + \sqrt{2\epsilon}\, \eta_x^i(t), \\ \dot{g}_t^i &= -\gamma g_t^i + \sqrt{2\gamma}\, \eta_g^i(t). \end{aligned} \qquad [1]$$

In Eq. **1**, μ is the mobility, $f(x)$ is a short-range repulsive potential force whose specific form will be specified later, and $v_0 \geq 0$ is the self-propulsion speed of the particles. $\eta_x^i(t)$ and $\eta_g^i(t)$ are independent white-noise sources, i.e., Gaussian processes with mean zero and covariances given by $\langle \eta_x^i(t) \eta_x^j(t') \rangle = \langle \eta_g^i(t) \eta_g^j(t') \rangle = \delta(t - t') \delta_{i,j} Id$. The parameter $\epsilon \geq 0$ sets the scale of the thermal noise (and need not be small), while $\gamma > 0$ tunes the persistence timescale of the self-propulsion. The orientational degrees of freedom g_t^i introduce an active noise term with a finite correlation time $1/\gamma$ into the translational dynamics for r_t^i; the presence of this correlated noise drives the system out of equilibrium for any $v_0 \neq 0$ and $\gamma < \infty$.

Trajectories of Eq. **1** are shown in Fig. 2*A*, where we consider $N = 2$ particles in dimension $d = 1$ on the interval $[0, L]$ with periodic boundary conditions. By translation invariance, we can define $x = x^2 - x^1$ and $g = g^2 - g^1$ to reduce dimensionality, which allows us to visualize the entire phase space. We will use this low-dimensional system as a running illustrative example, while our main results consider Eq. **1** in higher-dimensional situations with up to $N = 32{,}768$ particles in $d = 2$ dimensions.

General Microscopic Description. Since the tools that we introduce to study Eq. **1** are transportable to other nonequilibrium systems, it is convenient to view these equations as an instance of the generic stochastic differential equation (SDE) for $r_t \in \Omega$

$$\dot{r}_t = b(r_t) + \sqrt{2D}\, \eta(t), \qquad [2]$$

where $b(r)$ denotes the deterministic drift, D denotes the diffusion tensor (assumed to be symmetric and positive semidefinite but not necessarily invertible), and $\eta(t)$ is a white noise

process. Eq. **1** can be cast into the form of Eq. **2** by setting $r_t = (r_t^1, \ldots, r_t^N)$ with $r_t^i = (x_t^i, g_t^i) \in \mathbb{R}^{2d}$ for $i = 1, \ldots, N$ (so that $\Omega = \mathbb{R}^{2Nd}$), along with proper identification of $b(r)$ and D. For simplicity, we focus on drifts $b(r)$ that are independent of time, along with diffusion tensors D that are constant in both space and time. Importantly, we study systems that may not respect detailed balance, so that $b(r) \neq -D\nabla U(r)$ for some potential $U(r)$.

Many-Body FPE. The probability density function ρ_t of the solution r_t to Eq. **2** satisfies a many-body FPE that can be written as a continuity equation

$$\partial_t \rho_t(r) + \nabla \cdot j_t(r) = 0, \qquad [3]$$

where we have defined the probability current $j_t(r)$

$$j_t(r) = b(r)\rho_t(r) - D\nabla\rho_t(r). \qquad [4]$$

We study systems that have reached statistical steady state, so that $\rho_t(r) = \rho(r)$ and $j_t(r) = j(r)$. Then, Eq. **3** reduces to $\nabla \cdot j(r) = 0$ with $j(r) = b(r)\rho(r) - D\nabla\rho(r)$. Since we do not assume that the system is in detailed balance, its stationary density ρ and current j are in general unknown. In particular, the system can sustain a nonequilibrium stationary current $j \neq 0$. We visualize the stationary density ρ and the steady-state current j for our low-dimensional illustrative system in Fig. 2 *B* and *C*, respectively.

Current Velocity and Probability Flow. At stationarity, assuming that $\rho(r) > 0$ everywhere in Ω, we may rewrite Eq. **3** as a time-independent transport equation

$$0 = \nabla \cdot (v(r)\rho(r)), \qquad [5]$$

where v is the current (mean local) velocity field (11, 63) defined as

$$v(r) = j(r)/\rho(r) = b(r) - D\nabla \log \rho(r). \qquad [6]$$

The current velocity v is a fundamental object, and we will show that various definitions of the EPR can be computed from it (64, 65). It contains strictly more information than ρ alone, because it is always possible to construct an equilibrium system with the same ρ. Calculation of the EPR requires access to the steady state currents captured by v, which arise through an interplay between both the system's stationary density and structural information about its dynamics.

To gain access to v without explicit knowledge of ρ, we will develop a learning algorithm that estimates the high-dimensional $\nabla \log \rho$ *from data from the SDE in Eq. 2:* $\nabla \log \rho$ is known as the Hyvärinen "score" function in the machine learning literature (30). In addition to enabling the computation of various definitions of the EPR, v allows us to directly interrogate the flow of probability in the system. To do so, we may study the characteristics of Eq. **5** via solution of the ordinary differential equation (ODE)

$$\dot{R}_t(r) = v(R_t(r)), \qquad R_{t=0}(r) = r. \qquad [7]$$

We refer to Eq. **7** as the probability flow equation, as it describes the transport of samples in phase space according to the probability current j. In particular, at stationarity, the flow map $R_t(r)$ leaves the density ρ invariant, so that the density of $R_t(r)$ is ρ when r is drawn randomly from ρ. This means that for any observable $A(r)$, we have

$$\forall t \in \mathbb{R} \; : \quad \int_\Omega A(R_t(r))\rho(r)dr = \int_\Omega A(r)\rho(r)dr. \qquad [8]$$

We stress that for $j \neq 0$, transport can occur even at stationarity; the condition in Eq. **8** ensures that this transport preserves ρ. We visualize the phase portrait of Eq. **7** for our low-dimensional illustrative example in Fig. 2*D*. The resulting ordered limit cycles may be contrasted with trajectories of the equivalent stochastic dynamics (Eq. **2**) in Fig. 2*A*; despite their striking qualitative differences, both leave ρ invariant.

EPR

In this work, we are primarily interested in nonequilibrium systems, and we will study how their nonequilibrium dynamics arises spatially from nonzero v and $\nabla \cdot v$. To this end, we now relate $\nabla \cdot v$ and $|v|^2$ to several definitions of the EPR.

Gibbs Entropy and System EPR. Given the stationary solution ρ to Eq. **3**, the Gibbs entropy of the system is defined as

$$S_{\text{sys}} = -\int_\Omega \log \rho(r)\rho(r)dr. \qquad [9]$$

At stationarity, S_{sys} is time independent and hence must be preserved by the dynamics. To see how this occurs at the level of the individual degrees of freedom, following Seifert (7, 11), we can study the evolution of the *stochastic entropy of the system* along trajectories of the SDE 7

$$\tilde{s}_{\text{sys}}(t) = -\log \rho(r_t). \qquad [10]$$

Taking the time derivative of Eq. **10** gives

$$\dot{\tilde{s}}_{\text{sys}}(t) = -\nabla \log \rho(r_t) \circ \dot{r}_t, \qquad [11]$$

where $\circ$ denotes the Stratonovich product. The quantity defined in Eq. **11** is a stochastic function of time that can be evaluated along any trajectory. To obtain a deterministic function of r that conveys local information about the EPR, we can take the expectation of Eq. **11** conditioned on the event $r_t = r$ (7, 11). The current velocity defined in Eq. **6** can be expressed in terms of this conditional expectation as $v(r) = \langle \dot{r}(t) | r(t) = r \rangle$ (*SI Appendix*), so that $\dot{s}_{\text{sys}}(r) = \langle \dot{\tilde{s}}_{\text{sys}}(t) | r_t = r \rangle$ is given by

$$\dot{s}_{\text{sys}}(r) = -\nabla \log \rho(r) \cdot v(r) = \nabla \cdot v(r), \qquad [12]$$

where the last equality follows from Eq. **5** after division by $\rho > 0$.

The function defined in Eq. **12** is referred to as the (local) system EPR: it is visualized over the phase space of our low-dimensional illustrative example in Fig. 3, which highlights alternating regions of system entropy production and consumption in the two modes. Because S_{sys} is a constant of motion at stationarity, we arrive at the condition

$$\dot{S}_{\text{sys}} = \int_\Omega (\nabla \cdot v(r))\, \rho(r)dr = 0. \qquad [13]$$

Later, we will make use of Eq. **13** as a quantitative test to measure convergence of our learning algorithm. To understand how $\dot{s}_{\text{sys}}$ is distributed spatially in systems with a high-dimensional phase space, we may decompose $\nabla \cdot v$ into a *local* sum of contributions

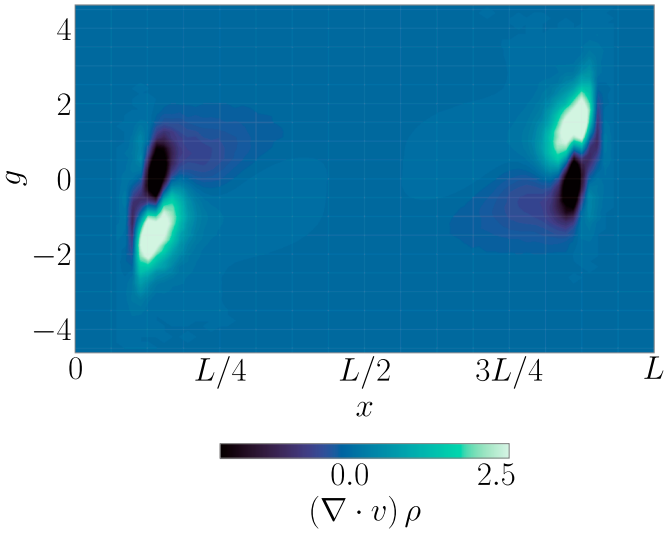

Fig. 3. System EPR. Visualization of $\nabla \cdot v(r)$ across the phase space for Eq. **1** with $N = 2$ and $d = 1$ in the variables $x_t = x_t^2 - x_t^1$ and $g_t = g_t^2 - g_t^1$. The system EPR at a point r in phase space can be written as $\dot{s}_{\text{sys}}(r) = \nabla \cdot v(r)$, so that $\nabla \cdot v(r)$ gives insight into how entropy is generated locally by the system. Even though $\mathbb{E}_\rho[\nabla \cdot v] = 0$ at stationarity, $\nabla \cdot v \neq 0$ pointwise when the system is out of equilibrium. Here, system entropy is produced locally when the two particles collide, and released when they separate.

from individual particles using $v(r) = (v^1(r), \ldots, v^N(r))$ to obtain

$$\nabla \cdot v(r) = \sum_{i=1}^{N} \nabla_i \cdot v^i(r), \qquad [14]$$

where ∇_i denotes the gradient with respect to r^i.

Total EPR. Assuming that D is invertible, we can use Eq. **6** written as $\nabla \log \rho = D^{-1}(v - b)$ to decompose Eq. **12** as

$$\dot{s}_{\text{sys}}(r) = \underbrace{|v(r)|^2_{D^{-1}}}_{\dot{s}_{\text{tot}}(r)} - \underbrace{b(r) \cdot D^{-1} v(r)}_{\dot{s}_{\text{m}}(r)}. \qquad [15]$$

The quantity $\dot{s}_{\text{tot}}$ is nonnegative and can be identified as the (local) total EPR (7, 11, 66). The quantity $\dot{s}_{\text{m}}$ is of indefinite sign and can be identified as the (local) EPR of the medium (4, 11, 19, 22, 67, 68). Similar to Eq. **14**, assuming that D is made of N diagonal blocks D_i, we may decompose

$$\dot{s}_{\text{tot}}(r) = |v(r)|^2_{D^{-1}} = \sum_{i=1}^{N} |v^i(r)|^2_{D_i^{-1}}, \qquad [16]$$

into local contributions from individual particles.

Global EPR. The global EPR is defined as the Kullback–Leibler divergence between the forward and reverse path measures (4, 19)

$$\dot{S}_{\text{tot}} = \frac{1}{T} \left\langle \log \left(\frac{\mathcal{P}(\phi_T)}{\mathcal{P}^{\text{R}}(\phi_T)} \right) \right\rangle, \qquad [17]$$

where $T > 0$ is arbitrary and where $\phi_T = \{r_t\}_{0 \le t \le T}$ denotes a path of the SDE 2 with the initial condition drawn from ρ. $\mathcal{P}$ denotes the path measure of ϕ_T, $\mathcal{P}^{\text{R}}$ denotes the path measure of a reverse-time path, and the angular brackets denote an average over ϕ_T drawn from $\mathcal{P}$. The global EPR can be challenging to compute because it requires a choice of reverse-time dynamics (to set $\mathcal{P}^{\text{R}}$), and the correct choice has been a subject of debate (22–24, 62, 69, 70). Interestingly, there is a way to construct a reverse-time dynamics such that $\dot{S}_{\text{tot}}$ can be written as an expectation of $\dot{s}_{\text{tot}}(r)$ over ρ, but it again requires knowledge of $v(r)$. This reverse-time dynamics is the SDE whose solutions have the same statistical properties as the solutions to Eq. **2** played in reverse (71). It reads (*SI Appendix*)

$$\dot{r}_t^{\text{R}} = b(r_t^{\text{R}}) - 2v(r_t^{\text{R}}) + \sqrt{2D}\,\eta(t), \qquad [18]$$

where $\eta(t)$ is the same Gaussian white noise process as in the forward SDE 2. It is easy to check that the stationary density of Eq. **18** is also ρ. By a standard path integral argument (72) or an application of the Girsanov theorem (73), when D is invertible, we may compute (*SI Appendix*)

$$\dot{S}_{\text{tot}} = \int_\Omega |v(r)|^2_{D^{-1}} \rho(r) dr, \qquad [19]$$

which, by Eq. 8, may be understood as the total EPR $\dot{s}_{\text{tot}}$ averaged over r drawn from ρ. Eq. **19** highlights that a system is at equilibrium if and only if $v = 0$, so that $\dot{S}_{\text{tot}} = 0$. An analogous relation first appeared in ref. 11 for driven colloidal particles, where the definition of the reverse-time process is less ambiguous than for the active systems we study here.

Learning Algorithm

Score. The expressions for v, the system EPR, and the total EPR depend on the score $\nabla \log \rho(r)$, which is a high-dimensional function we typically do not have access to. In this section, we develop a machine learning algorithm to approximate it: a graphical summary of the method is given in Fig. 1. In addition to providing access to v, and therefore to the total and system EPRs, $\nabla \log \rho$ has the important advantage that it is independent of the normalizing constant of ρ, which is typically unknown and intractable. This enables us to exploit expressive function classes that need not represent normalized probability distributions.

Score Matching. The score $\nabla \log \rho$ can be shown to be the unique minimizer of the loss

$$\mathcal{L}_{\text{sm}}[\hat{h}] = \mathbb{E}_\rho \left[|\hat{h}|^2 + 2\nabla \cdot \hat{h} \right],$$
$$\nabla \log \rho = \underset{\hat{h}}{\text{argmin}}\, \mathcal{L}_{\text{sm}}[\hat{h}], \qquad [20]$$

where $\mathbb{E}_\rho$ denotes expectation over ρ. Eq. **20** is known as the "score matching" loss in the machine learning literature (30). For the reader's convenience, we provide a derivation of this loss and demonstrate the uniqueness of its minimizers in *SI Appendix*.

Exploiting the Stationary FPE. While a useful loss function, Eq. **20** is valid for *any* data distribution and does not make use of the fact that ρ solves the stationary FPE [**3**]; it is therefore agnostic to the underlying physics. Intuitively, exploiting our prior knowledge that ρ solves Eq. **3** should impose additional structure that can be leveraged to improve the quality of the learned score. As written, Eq. **3** is an equation for ρ, while we are interested in estimating $\nabla \log \rho$. Dividing by ρ yields a nonlinear equation for the score

$$\nabla \cdot v + v \cdot \nabla \log \rho = 0. \qquad [21]$$

Eq. **21** may be used to construct a physics-informed loss based on the squared residual (39, 53)

$$\mathcal{L}_{\mathsf{FPE}}[\hat{h}] = \mathbb{E}_\rho\left[\left(\nabla\cdot\hat{v}+\hat{v}\cdot\hat{h}\right)^2\right], \qquad [22]$$

where $\hat{v}(r) = b(r) - D\hat{h}(r)$. We propose minimization of the composite loss

$$\mathcal{L}[\hat{h}] = \lambda_1\mathcal{L}_{\mathsf{sm}}[\hat{h}] + \lambda_2\mathcal{L}_{\mathsf{FPE}}[\hat{h}], \qquad [23]$$

which consists of both the physics-agnostic score matching loss $\mathcal{L}_{\mathsf{sm}}$ and the physics-informed loss $\mathcal{L}_{\mathsf{FPE}}$. In our experiments, we find best performance incorporating both terms, and we set $\lambda_1 = \lambda_2 = 1$ throughout unless otherwise indicated.

Empirical Loss. In practice, we minimize an empirical approximation of Eq. **23**

$$\hat{\mathcal{L}}[\hat{h}] = \frac{\lambda_1}{n}\sum_{\alpha=1}^{n}\left(|\hat{h}(r_\alpha)|^2 + 2\nabla\cdot\hat{h}(r_\alpha)\right) + \frac{\lambda_2}{n}\sum_{\alpha=1}^{n}\left(\nabla\cdot\hat{v}(r_\alpha)+\hat{v}(r_\alpha)\cdot\hat{h}(r_\alpha)\right)^2 \qquad [24]$$

over a dataset of samples $\{r_\alpha\}_{\alpha=1}^n$ with each $r_\alpha \sim \rho$. We can generate such a dataset by simulating the SDE in Eq. **2** with a numerical integration scheme like the Euler–Maruyama method. To make the optimization computationally tractable for high-dimensional systems of particles, we can perform the estimation of $\hat{h}(r)$ over an expressive parametric class of functions such as a class of neural networks, and can use a first-order optimization scheme such as Adam (74) to optimize the parameters. To increase the diversity of the dataset, we can take steps of the SDE 1 between steps of the optimization algorithm, which is similar to online learning and helps prevent overfitting.

Quantitative Validation. There are several metrics that we can use to verify the accuracy of the learned approximation $\hat{h}$ to $\nabla\log\rho$. The loss in Eq. **22** is exactly the squared residual for the stationary score-based FPE [**21**] and hence provides a quantitative measure of how well the learned score satisfies its governing equation. At optimality, Eq. **20** satisfies $\mathcal{L}_{\mathsf{sm}}[\nabla\log\rho] = -\mathbb{E}_\rho\left[|\nabla\log\rho|^2\right]$ (*SI Appendix*); deviation from this relation also provides a measure of convergence. Last, we can verify the constraint Eq. **13** to ensure that the global EPR is a constant of the motion.

Neural Network Architecture

Permutation Symmetry. An important ingredient in our learning algorithm is a proper choice of the neural network used to estimate $\nabla\log\rho$. One guiding principle that can be used for physical problems is to build the symmetries of the system into the network (50). In addition to its conceptual motivation, this approach has been shown to be statistically advantageous (75). In Eq. **1**, the most relevant symmetry group is permutation invariance among the particles, which generates complex multimodal structure in the stationary density ρ. Generically, all the configurations generated by permutations will not be present in a given dataset, and this makes it crucial to use a representation of $\nabla\log\rho(r)$ where the permutation invariance is built in.

Invariance and Equivariance. Permutation invariance at the level of ρ gives rise to permutation *equivariance* at the level of $\nabla\log\rho$. In a numerical implementation, we can choose to parameterize $\log\rho$ and take its gradient or to parameterize $\nabla\log\rho$ directly. While it seems physically natural to parameterize $\nabla\log\rho$ as a gradient field, state-of-the-art results in diffusion-based generative modeling directly parameterize the score without this added constraint (31, 76, 77), and we follow this approach here. Doing so reduces the number of gradients that must be computed via automatic differentiation during training and tends to improve performance.

Transformers. Perhaps the most natural way to proceed is to employ an architecture that can directly learn the relevant order of the interactions in the system. The transformer architecture (78) has emerged as a powerful tool for learning complex interactions in language (79, 80) and is built upon operations (self-attention and token-wise mappings) that are naturally permutation equivariant. In addition to language modeling, transformers currently achieve state of the art in image classification (81, 82) and have been applied to problems such as protein structure prediction (83) and quantum chemistry (84). Yet, to our knowledge, they have not been used to study interacting particle systems in active matter and stochastic thermodynamics. Transformers also have the advantage that their attention maps can be inspected a posteriori for insight into the interactions learned by the model (85).

We introduce a transformer architecture that learns interactions between embeddings of the particle positions x^i and orientations g^i (Fig. 4). The output of a series of transformer encoder blocks consisting of self-attention, LayerNorm (86), and particle-wise multilayer perceptrons (MLPs) is decoded by an additional particle-wise MLP to obtain the score $\hat{h}$. For large numbers of interacting particles (as we will study in the MIPS system), we introduce a modification of this architecture that exploits a spatially local ansatz to define the score $\hat{h}^i$ at

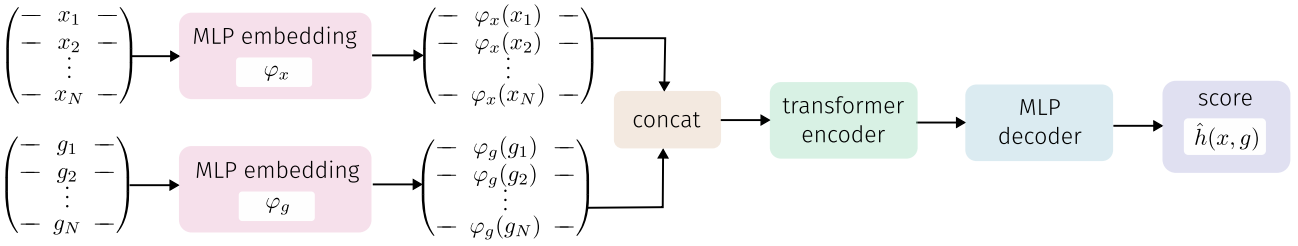

Fig. 4. Network architecture. Depiction of the transformer architecture introduced in this work. The particle positions and orientations are fed into separate multilayer perceptrons (MLPs) that embed the input into a latent space of higher dimensionality. The embeddings are concatenated particle-wise and fed into a transformer encoder block (see *SI Appendix* for further details), where multiple layers of multihead attention modules learn relevant interactions between particles. The output of the encoder block is decoded by a shared MLP applied to each particle state to obtain the score.

the level of the individual particles. Remarkably, in addition to a large gain in memory efficiency, this architecture enables transfer learning to datasets with a larger number of particles, where we find physically meaningful predictions without any additional training. We provide an overview of the relevant features of the transformer architecture, including more detail on the constituent elements of the encoder blocks in *SI Appendix*.

Active Swimmers in a Trap

Dynamics. As a first application of our method, we now consider a system of $N = 64$ interacting active Ornstein–Uhlenbeck particles in a harmonic trap, similar to what was studied by Martin et al. (15); this gives rise to a 256-dimensional many-body FPE in Eq. **3**. In this case, the system under study is given by Eq. **1** with $\epsilon = 0.1$, $\gamma = 0.1$, $\mu = 1$, and where f is a conservative force governed by the many-body potential

$$\Phi(x) = \frac{A}{2}\sum_{i=1}^{N}|x_i|^2 + \frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{N} V(x_i - x_j). \qquad [25]$$

Above, $A = 0.05$ and $V(x) = \frac{k}{2}(2a - |x|)^2\,\Theta(2a - |x|)$ with $a = 1$ the particle radius, $k = 2.5$, and where Θ denotes the Heaviside step function. To make the system more amenable to learning, we smooth the force slightly to avoid the hard cutoff mediated by Θ (see *SI Appendix* for details). Due to the presence of the trap, the system assembles into an active cluster with a dense core and a motile boundary that is similar to the phase separation observed in MIPS. The trap makes the presence of these features more robust to variations in the parameters γ, ϵ, and v_0, which enables us to study the structure of the total and system EPRs as a function of the activity v_0 at fixed persistence γ and bath temperature ϵ. Localizing the cluster also allows us to perform spatial averaging of the EPR, which connects our results with the field-theoretic approach developed in ref. 20. We stress, however, that our approach does not require this spatial averaging, and we will remove the trap when we study MIPS.

EPR Decomposition. The contribution of each particle to the system EPR $\nabla_i \cdot v^i(r) = \nabla_{x^i}\cdot v_x^i(r) + \nabla_{g^i}\cdot v_g^i(r)$ can be decomposed additively into local contributions from the orientational degrees of freedom g^i and the translational degrees of freedom x^i to understand how they independently contribute to the system EPR. Similarly, the contribution to the total EPR decomposes as $|v^i(r)|^2_{D_i^{-1}} = \frac{1}{\gamma}|v_g^i(r)|^2 + \frac{1}{\epsilon}|v_x^i(r)|^2$. In the following, we make use of these decompositions to isolate further how the total and system EPRs are built up from individual particle contributions.

Orientational Contributions. We first focus on the contribution of the orientational degrees of freedom to the total EPR $|v_g|$ and the system EPR $\nabla_g \cdot v_g$ (Figs. 5 and 6, *Top*). For $v_0 = 0$, the system is at equilibrium and hence the EPR vanishes. As v_0 is increased, the system becomes increasingly nonequilibrium, and spatial structure begins to emerge in the EPR. Consistent with theoretical predictions (20), we find that both quantities concentrate on the boundary of the cluster. The contribution to the EPR of the system $\nabla_g \cdot v_g$ increases smoothly with radial distance from the center of the cluster. The contribution to the total entropy production $|v_g|$ is similar but is more dominated by a few outliers on the boundary.

In *SI Appendix*, Figs. S5 and S6, we visualize the translational contributions $|v_x|$ and $\nabla_x \cdot v_x$, as well as the EPR $|v|$ and the system EPR $\nabla \cdot v$. $|v_x|$ displays similar features to $|v_g|$ with slightly lower contrast between the core and the boundary, so that $|v|$ also displays concentration on the interface. We find that $\nabla_x \cdot v_x \approx -\nabla_g \cdot v_g$, which causes $\nabla \cdot v$ to roughly vanish pointwise per-particle. Small-scale fluctuations in the particles are present around zero, which together average so that $\mathbb{E}_\rho[\nabla \cdot v] \approx 0$ as required by stationarity. We find high accuracy as measured by the residual of the score-based FPE in Eq. **21** and the stationarity condition Eq. **13** (*SI Appendix*, Fig. S7).

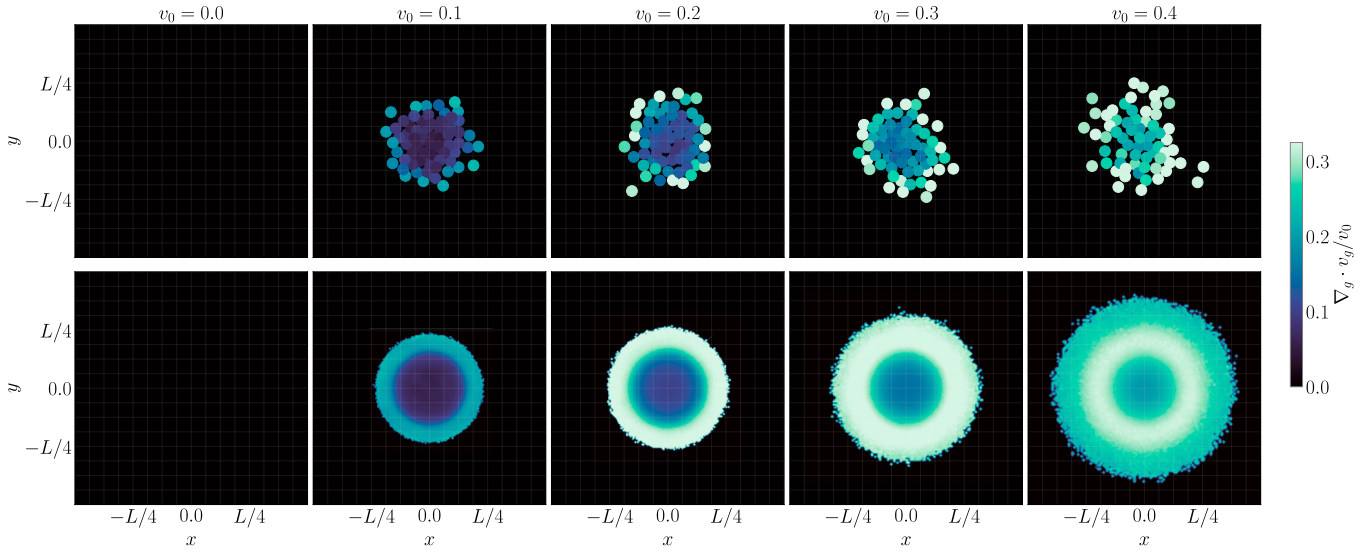

Fig. 5. 64 swimmers in a harmonic trap: system EPR. (*Top*) The contribution of the per-particle orientational degrees of freedom to the system EPR $\nabla_{g^i} \cdot v_g^i$ as a function of the activity v_0, visualized directly on the particles. For $v_0 = 0$, the system is at equilibrium and the network learns that the system EPR vanishes. As v_0 increases, nonequilibrium effects emerge, and the particles on the boundary display the highest contribution to the EPR. (*Bottom*) A spatial map visualizing the typical contribution of a particle at position (x, y) to the system EPR, obtained by averaging the data in the top row over many system snapshots. The map highlights the role of interfacial contributions and displays a prominent ring at the boundary of the cluster.

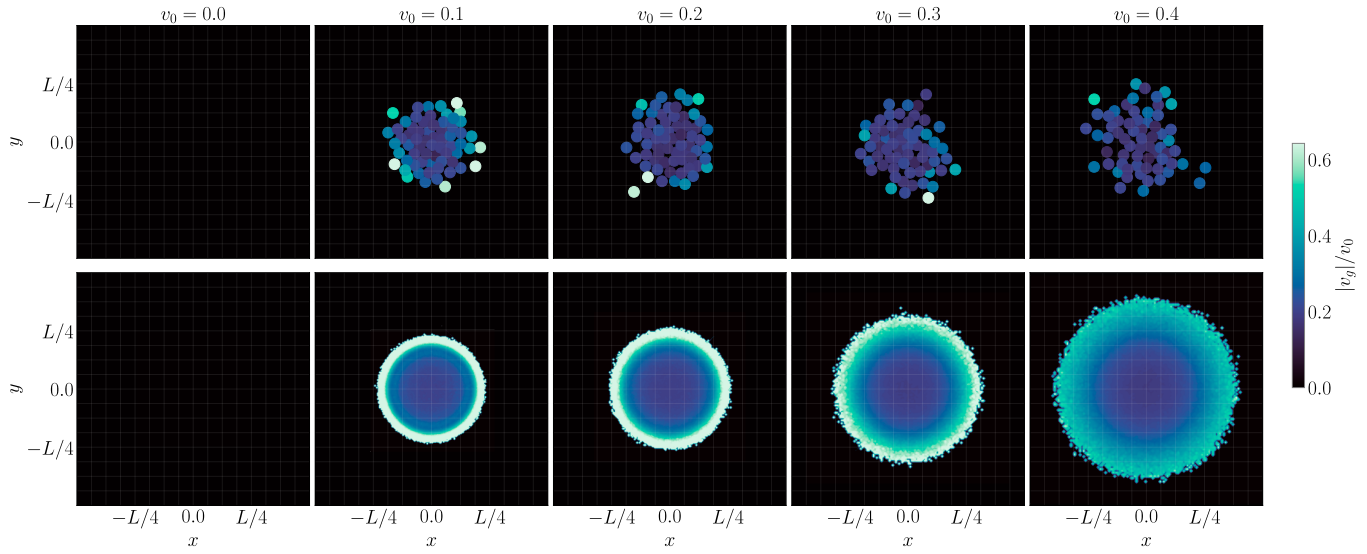

Fig. 6. 64 swimmers in a harmonic trap: total EPR. (*Top*) The contribution of the per-particle orientational degrees of freedom to the total EPR $|v_g^i|^2$ as a function of the activity v_0, visualized directly on the particles. As in Fig. 5, the network learns that the system is at equilibrium for $v_0 = 0$, and the total EPR vanishes. As v_0 increases, the total EPR is dominated by outlier contributions from particles on the edge of the cluster. (*Bottom*) A spatial map visualizing the typical contribution of a particle at position (x, y) to the total EPR, obtained by averaging the data in the top row over many system snapshots. The map distills the signal present in the outliers in the top row and displays a concentrated ring of entropy production at the interface.

A Spatial Map of Entropy Production. The presence of the trap constrains the shape and location of the cluster, which facilitates averaging in space and in time. To build up a spatial map of the entropy production, we discretize space into a 256×256 grid. We can then sum the particle-wise contributions $\nabla_{g^i} \cdot v_g^i$ and $|v_g^i|^2$ in each grid cell over a dataset of samples, normalizing by the number of particles that appear in each cell in the dataset. The result is a spatial map that describes the typical value of the EPR a particle would attain at a given spatial position. We visualize these spatial maps in Figs. 5 and 6 (*Bottom*), where we find a distinct ring of entropy production at the boundary of the core.

Attention Map. An advantage of the transformer architecture is that we can visualize the *attention map*, which gives us insight into which *other* particles are used to compute the score of a given particle. We employed attention rollout (85) to propagate the flow of attention across all heads from layer to layer (*SI Appendix*). The result reveals that the network learns a physically intuitive spatially local attention pattern, where each particle is primarily influenced by its nearest neighbors (*SI Appendix*, Fig. S8). Interestingly, the interactions are still significantly longer range than those present in the interaction potential for the system.

Motility-Induced Phase Separation

We now consider a system of $N = 4{,}096$ interacting particles undergoing motility-induced phase separation, given by Eq. **1** with $v_0 = 0.025$, $\mu = 1$, $\gamma = 10^{-4}$, $\epsilon = 0$, and with periodic boundary conditions. In this case, the corresponding many-body FPE is 16,384-dimensional, and its solution poses a formidable challenge. Because we consider the athermal, hypoelliptic setting with $\epsilon = 0$, the velocity field defined in Eq. **6** only depends on the score in the g variables $\nabla_g \log \rho$. To target $\nabla_g \log \rho$ directly, we consider only the score matching loss Eq. **20** and set $\lambda_2 = 0$ in the combined loss Eq. **23**. The resulting loss decouples into equivalent losses for $\nabla_x \log \rho$ and $\nabla_g \log \rho$, while the physics-informed loss Eq. **22** couples the two scores, so they cannot be learned independently. Because $\epsilon = 0$, the diffusion tensor D is no longer invertible, and the connection between v and the total EPR $\dot{s}_{\text{tot}}(r) = |v(r)|^2_{D^{-1}}$ breaks down. By contrast, the identity $\dot{s}_{\text{sys}}(r) = \nabla \cdot v(r)$ still holds, and v remains a fundamental object that describes the transport of the particles according to the probability current. For further details, including an overview of a variant of the denoising score matching loss function (87) that we use to reduce computational expense, along with a discussion of the technical issues that arise for $\epsilon = 0$, see *SI Appendix*. We learn $\nabla_g \log \rho$ on a single dataset with $N = 4{,}096$ particles and with a packing fraction $\phi = 0.5$, but make use of an architecture that enables transfer learning to higher values of N and other values of ϕ, as we now describe.

Network Architecture. The large number of particles makes it computationally intractable to use the same transformer architecture we used for $N = 64$: the self-attention mechanism has time and memory complexities that scale as $\mathcal{O}(N^2)$, which quickly both become prohibitive for large N. Nevertheless, *SI Appendix*, Fig. S8 shows that the learned attention map is spatially local in the $N = 64$ case, and we expect the same behavior to hold true for the MIPS system. To exploit spatial locality, we developed a transformer architecture defined at the level of an individual particle, which restricts the attention mechanism to a local neighborhood (*SI Appendix*). This approach has the additional advantage of increasing the effective size of the dataset because there are many distinct local neighborhoods in a given snapshot of the system. As the size of the attention window is increased, the architecture used for $N = 64$ is recovered. Because we define the transformer at the single particle level, our network can be extended to systems with larger N or with a different packing fraction ϕ, which we also demonstrate in this section.

System EPR and Magnitude of the Probability Flow. Fig. 7*A* shows the MIPS cluster for reference, while Fig. 7 *B* and *C* display the orientational contributions to the magnitude of the probability flow $|v_g|$ and the system EPR $\nabla_g \cdot v_g$, respectively. Both quantities are visualized as individual particle contributions without any averaging in space or in time. Consistent with the results for $N = 64$ from the previous section, we find that the

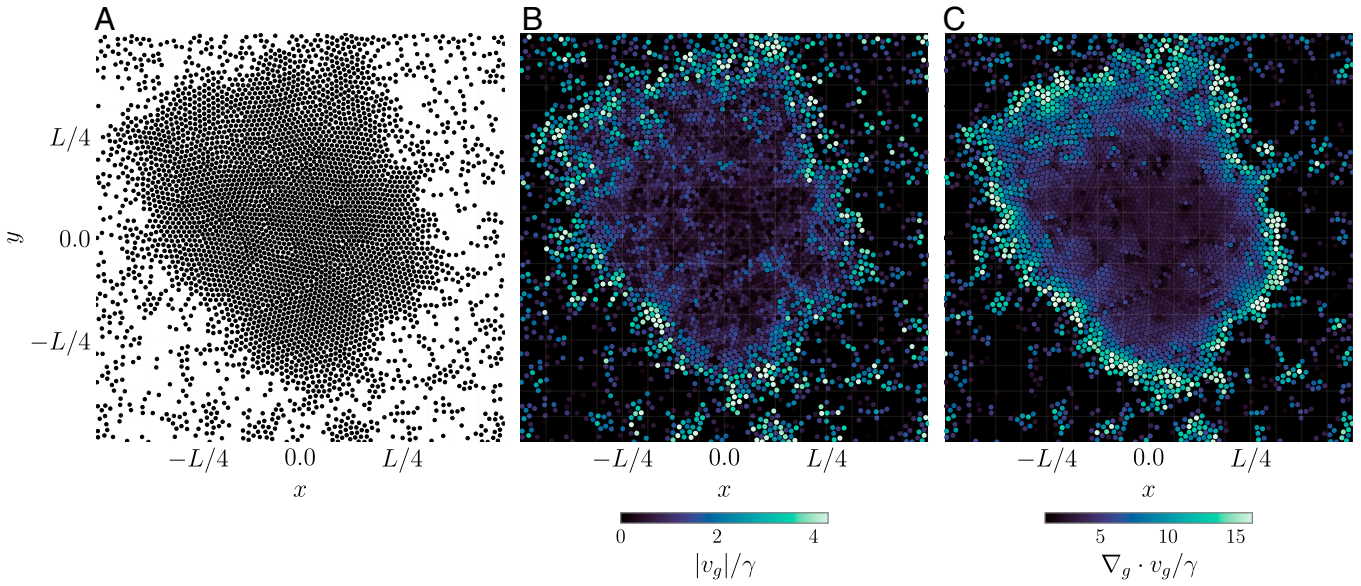

Fig. 7. Motility-induced phase separation. (*A*) Reference depiction of the MIPS cluster. (*B* and *C*) Particle-wise orientational contributions to the total EPR $|v_g^i|^2$ and the system EPR $\nabla_{g^i} \cdot v_g^i$. Both quantities concentrate on the boundary of the cluster, with sporadic contributions throughout the dilute phase when particles collide. $|v_g^i|$ vanishes in the center of the cluster, indicating a nontrivial phase dependence in the velocity field. A movie visualizing the evolution of these quantities along stochastic trajectories can be found at this link.

dominant source of entropy production is at the interface between the gas and solid phases. There are also pockets of entropy production spread sporadically throughout the gas in regions with particle–particle collisions. A movie of a stochastic trajectory, colored as in Fig. 7, can be found at this link (we recommend downloading the movie for high-resolution viewing).

Transfer Learning Toward Larger Systems. Because our neural network architecture is defined at the particle level, it is agnostic to the number of particles N in the system. This means that we can take a single network trained with $N = 4{,}096$ and investigate whether it can make reasonable predictions for higher values of N without any retraining. Physically, because the possible local environments for a given particle should be roughly independent of N for N sufficiently large, we expect our learned network to generalize beyond the training data. In Fig. 8, we show predictions of $|v_g|$ and $\nabla_g \cdot v_g$ as a function of N, ultimately scaling up to 32,768 particles. As the number of particles increases, the cluster becomes more well defined, and the signals in the probability flow and the EPR seen for $N = 4{,}096$ become increasingly high resolution. These results highlight the remarkable fact that the local environment learned with $N = 4{,}096$—where dataset generation is significantly cheaper—can be used to make predictions about systems with a larger number of interacting particles.

Probability Flow. We can use our ability to scale to larger N to investigate the probability flow near the thermodynamic limit. In Fig. 9*A*, we visualize the directionality of v_g^i, and we compare it to g^i in Fig. 9*B*. The result reveals a surprising set of observations: v_g^i vanishes in the solid phase, typically points outward at the solid edge of the interface, and typically points inward at the gaseous edge of the interface. While it follows by force-balance that $|v_x^i|$ must vanish in the solid, an equivalent for $|v_g^i|$ is nontrivial. This is highlighted when contrasting the particle-wise values of $|v_g^i|$ with those of $|g^i|$. Unlike the $|v_g^i|$, the $|g^i|$ appear random, and by eye, uncorrelated with their phase. This is consistent with the fact that their dynamics is decoupled from the translational degrees of freedom in Eq. **1**.

Together, these observations reveal a simple picture for the probability flow. Particles in the solid are frozen with $v_x^i = v_g^i = 0$. Free particles in the gas have $v_x^i = v_0 g^i$ and $v_g^i = 0$. At particle–particle collisions, v_g^i becomes nonzero, and entropy is produced. These events are mostly concentrated at the interface between the gas and the solid, where there are particles both exiting and entering the cluster, but also occur sporadically throughout the gas.

Packing Fraction Transfer. In addition to transferring to larger values of N, we can investigate the ability of the learned network to transfer to other regimes of the phase diagram, so long as the system parameters defining the particles are fixed. For example, because the score is defined at the level of the local neighborhood of each particle, and because these local environments should be statistically similar in some regions of the phase diagram, we expect the score to be able to transfer to other packing fractions ϕ. In Fig. 10, we show that a single network trained on a dataset of $N = 4{,}096$ particles with $\phi = 0.5$ can make physically consistent predictions for a range of values from $\phi = 0.01$ to $\phi = 0.9$ on a dataset with $N = 8{,}192$ particles. For very low packing fraction, there are few particle–particle collisions and no cluster, and the EPR is essentially zero everywhere. As the packing fraction increases, particle–particle collisions begin to occur, so that pockets of entropy production become spread throughout the gas. As a cluster forms for intermediate ϕ, the EPR becomes concentrated at the interface. As ϕ increases further, the cluster becomes dominant, again leaving few regions of entropy production. Analogous figures for $N = 4{,}096$, $N = 16{,}384$, and $N = 32{,}768$ are shown in *SI Appendix*, Figs. S8–S10.

Discussion

In this work, we demonstrated the capability of machine learning algorithms to learn the EPR and the probability flow of complex interacting particle systems, even in high-dimensional scenarios

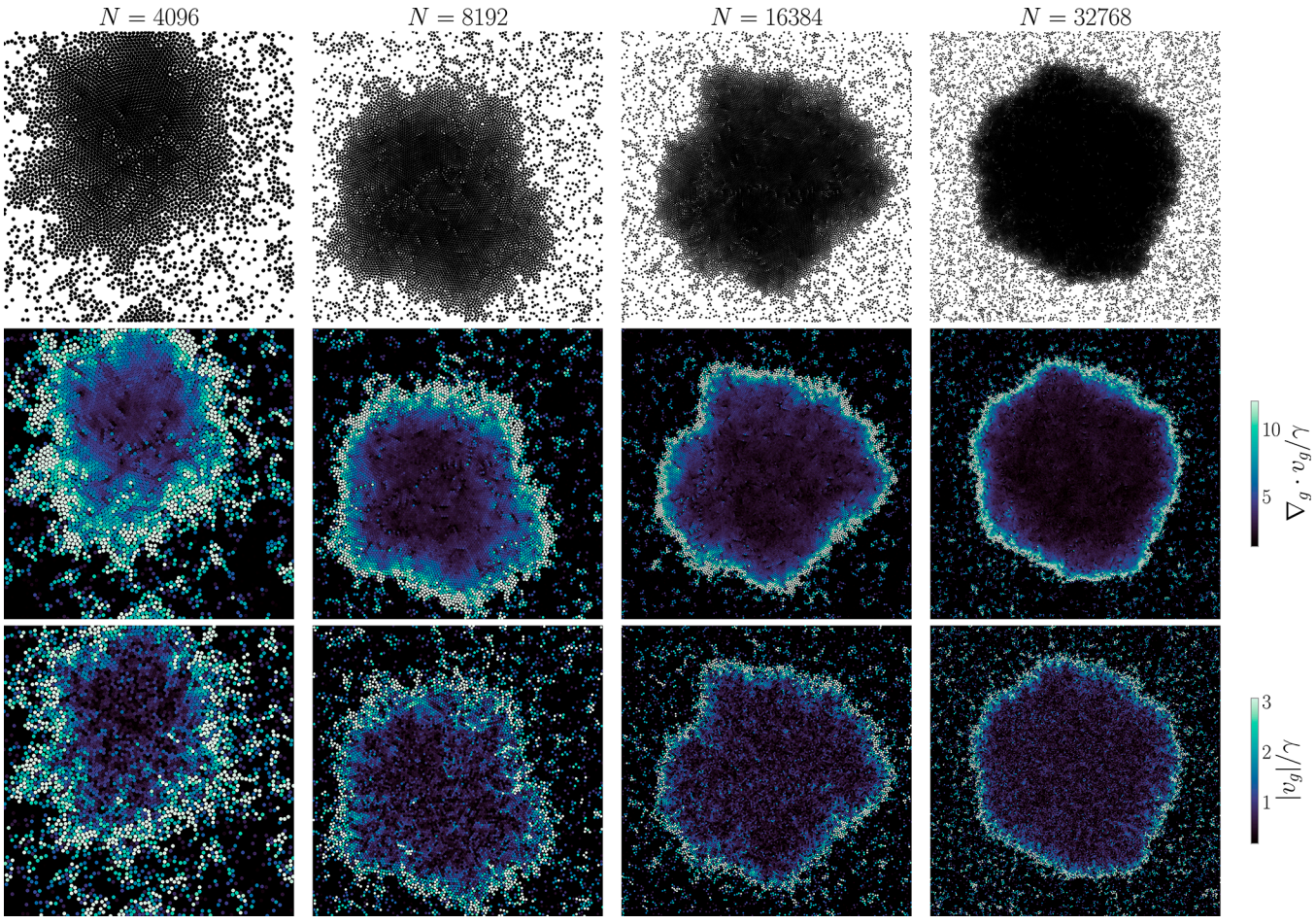

Fig. 8. Motility-induced phase separation: transfer learning toward larger systems. Our network architecture is defined at the level of individual particles and depends only on local neighborhoods. This enables us to extend the learned solution from the $N = 4{,}096$ training set to larger values of N. We consider values of N up to 8× larger and find physically consistent predictions in all cases. Movies visualizing the evolution of the EPR along stochastic trajectories can be found at the following links: $N = 8{,}192$, $N = 16{,}384$, $N = 32{,}768$.

typically plagued by the curse of dimensionality. In addition to uncovering structure in the EPR, we highlighted that a network trained with a given number of particles N and a fixed packing fraction ϕ can generalize to other values of N and ϕ. As a result, our method paves the way to investigating questions about active systems in the thermodynamic limit.

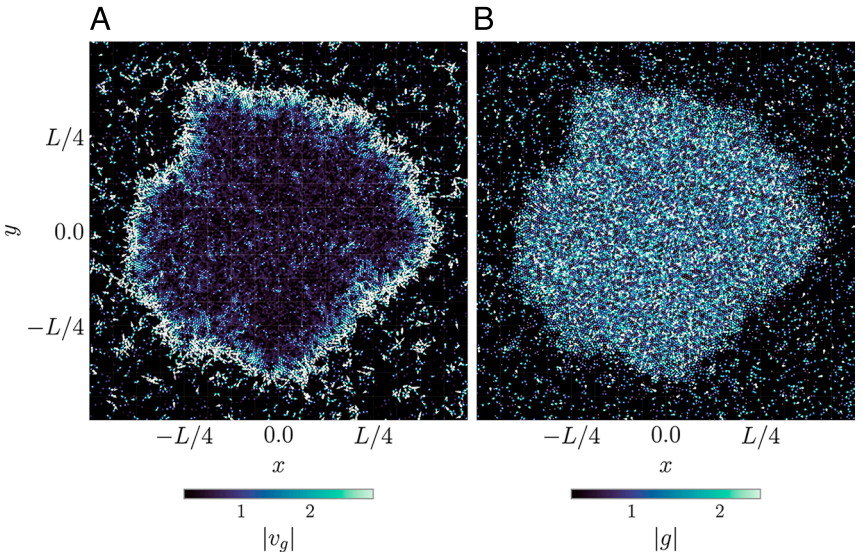

Fig. 9. Motility-induced phase separation: probability flow. Particle values of v_g^i (*A*) and g^i (*B*), with directionality visualized as arrows for $N = 16{,}384$. v_g^i vanishes in the interior of the solid but points outward near the solid side of the interface. There is also a layer of particles pointing both inward and outward directly at the interface, corresponding to particles that are exiting and leaving the cluster. The values of g^i appear random and do not have a clear phase dependence by eye, in contrast with v_g^i.

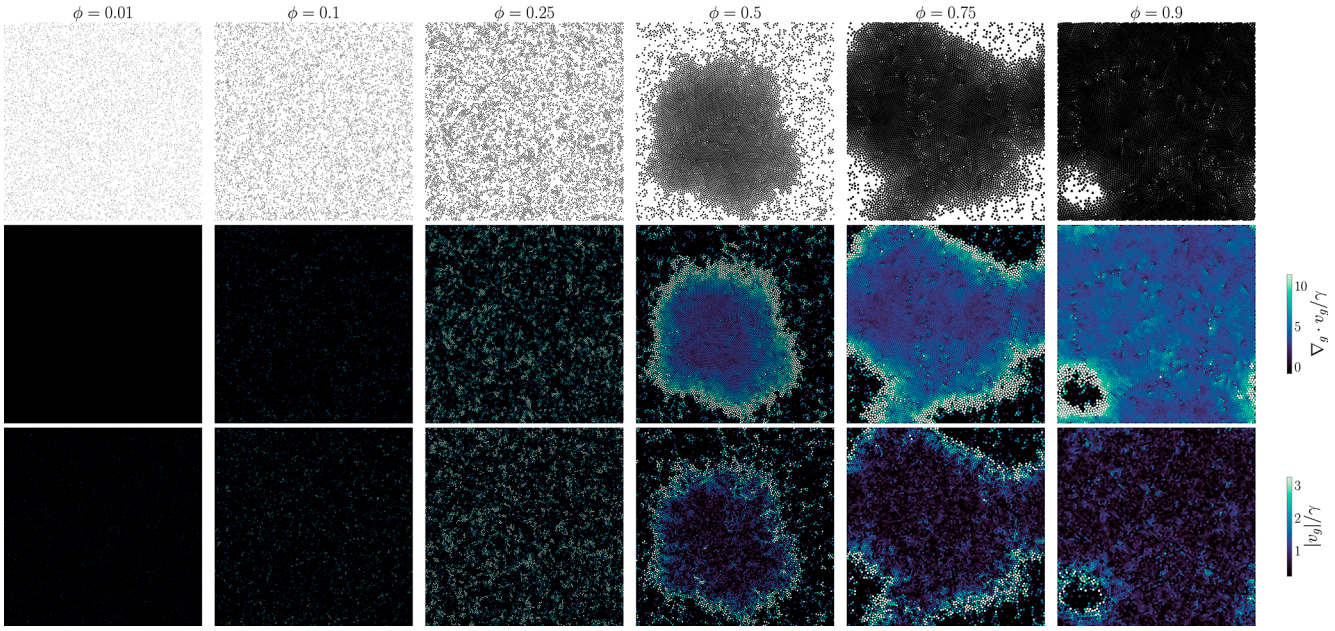

Fig. 10. Motility induced phase separation: transfer learning to other packing fractions. Similar to Fig. 8, we find that the learned solution generalizes to other packing fractions ϕ. Here, we vary ϕ at resolution $N = 8{,}192$ by varying the size L of the box; for presentation, we rescale the results to the same square. The solution identifies contributions to the EPR from particle–particle collisions in the gaseous phase and at the interface of the gaseous and solid phases. For low and high packing fraction, the EPR diminishes, as the system becomes dominated by a gas or a solid.

Physically, we focused on active particles without alignment interactions. A natural extension of this work is to consider more complex models such as the Vicsek model (88). Numerically, we considered transformer architectures based on standard self-attention modules, but could likely scale to larger systems with less local interactions by incorporating recent advances such as FlashAttention (89, 90).

Materials and Methods

All implementation and algorithmic details are included in either the main text or *SI Appendix*. In particular, we include a review of stochastic thermodynamics, derivations of the EPRs, and a derivation of the reverse-time stochastic dynamics. We also include a review of score-matching algorithms and transformer neural networks. Finally, we provide fine-grained details about neural network training, architectures, initialization, and further results.

Data, Materials, and Software Availability. Code data have been deposited in Github (91).

ACKNOWLEDGMENTS. We thank Grant Rotskoff, Michael Albergo, and Stephen Tu for many useful discussions. N.M.B. is partially supported by the National Science Foundation under Award DMS-2134216. E.V.-E. is supported by the National Science Foundation under Awards DMR-1420073, DMS-2012510, and DMS-2134216, by the Simons Collaboration on Wave Turbulence, Grant No. 617006, and by a Vannevar Bush Faculty Fellowship.

Author contributions: N.M.B. and E.V.-E. designed research; performed research; N.M.B. wrote the code and trained the neural networks; and N.M.B. and E.V.-E. wrote the paper.

1. M. C. Marchetti *et al*., Hydrodynamics of soft active matter. *Rev. Mod. Phys.* **85**, 1143–1189 (2013).
2. E. Fodor, R. L. Jack, M. E. Cates, Irreversibility and biased ensembles in active matter: Insights from stochastic thermodynamics. *Annu. Rev. Condens. Matter Phys.* **13**, 215–238 (2022).
3. J. O'Byrne, Y. Kafri, J. Tailleur, F. van Wijland, Time irreversibility in active matter, from micro to macro. *Nat. Rev. Phys.* **4**, 167–183 (2022).
4. J. L. Lebowitz, H. Spohn, A Gallavotti–Cohen-type symmetry in the large deviation functional for stochastic dynamics. *J. Stat. Phys.* **95**, 333–365 (1999).
5. J. M. R. Parrondo, C. V. d. Broeck, R. Kawai, Entropy production and the arrow of time. *New J. Phys.* **11**, 073008 (2009).
6. C. Jarzynski, Equalities and inequalities: Irreversibility and the second law of thermodynamics at the nanoscale. *Annu. Rev. Condens. Matter Phys.* **2**, 329–351 (2011).
7. U. Seifert, Stochastic thermodynamics, fluctuation theorems, and molecular machines. *Rep. Prog. Phys.* **75**, 126001 (2012).
8. T. Bodineau, B. Derrida, Cumulants and large deviations of the current through non-equilibrium steady states. *C. R. Phys.* **8**, 540–555 (2007).
9. C. Battle *et al*., Broken detailed balance at mesoscopic scales in active biological systems. *Science* **352**, 604–607 (2016).
10. L. Bertini, A. De Sole, D. Gabrielli, G. Jona-Lasinio, C. Landim, Macroscopic fluctuation theory. *Rev. Mod. Phys.* **87**, 593–636 (2015).
11. U. Seifert, Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Phys. Rev. Lett.* **95**, 040602 (2005).
12. A. Gomez-Marin, J. M. R. Parrondo, C. V. d. Broeck, The "footprints" of irreversibility. *Europhys. Lett.* **82**, 50002 (2008).
13. P. Pietzonka, U. Seifert, Entropy production of active particles and for particles in active baths. *J. Phys. A: Math. Theor.* **51**, 01LT01 (2018).
14. E. Fodor *et al*., How far from equilibrium is active matter? *Phys. Rev. Lett.* **117**, 038103 (2016).
15. D. Martin *et al*., Statistical mechanics of active Ornstein–Uhlenbeck particles. *Phys. Rev. E* **103**, 032607 (2021).
16. T. Speck, Active Brownian particles driven by constant affinity. *Europhys. Lett.* **123**, 20007 (2018).
17. S. Shankar, M. C. Marchetti, Hidden entropy production and work fluctuations in an ideal active gas. *Phys. Rev. E* **98**, 020604 (2018).
18. L. P. Dadhichi, A. Maitra, S. Ramaswamy, Origins and diagnostics of the nonequilibrium character of active systems. *J. Stat. Mech: Theory Exp.* **2018**, 123201 (2018).
19. G. E. Crooks, Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E* **60**, 2721–2726 (1999).
20. C. Nardini *et al*., Entropy production in field theories without time-reversal symmetry: Quantifying the non-equilibrium character of active matter. *Phys. Rev. X* **7**, 021007 (2017).
21. O. L. Borthne, E. Fodor, M. E. Cates, Time-reversal symmetry violations and entropy production in field theories of polar active matter. *New J. Phys.* **22**, 123012 (2020).
22. R. Chetrite, K. Gawedzki, Fluctuation relations for diffusion processes. *Commun. Math. Phys.* **282**, 469–518 (2008).
23. D. Mandal, K. Klymko, M. R. DeWeese, Entropy production and fluctuation theorems for active matter. *Phys. Rev. Lett.* **119**, 258001 (2017).
24. L. Caprini, U. M. B. Marconi, A. Puglisi, A. Vulpiani, Comment on "Entropy production and fluctuation theorems for active matter". *Phys. Rev. Lett.* **121**, 139801 (2018).
25. S. Martiniani, P. M. Chaikin, D. Levine, Quantifying hidden order out of equilibrium. *Phys. Rev. X* **9**, 011031 (2019).
26. S. Ro *et al*., Model-free measurement of local entropy production and extractable work in active matter. *Phys. Rev. Lett.* **129**, 220601 (2022).
27. J. Li, J. M. Horowitz, T. R. Gingrich, N. Fakhri, Quantifying dissipation using fluctuating currents. *Nat. Commun.* **10**, 1666 (2019).

28. S. Otsubo, S. Ito, A. Dechant, T. Sagawa, Estimating entropy production by machine learning of short-time fluctuating currents. *Phys. Rev. E* **101**, 062106 (2020).
29. S. Otsubo, S. K. Manikandan, T. Sagawa, S. Krishnamurthy, Estimating time-dependent entropy production from non-equilibrium trajectories. *Commun. Phys.* **5**, 1–10 (2022).
30. A. Hyvarinen, Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **6**, 695–709 (2005).
31. Y. Song *et al*., Score-based generative modeling through stochastic differential equations. arXiv [Preprint] (2021). https://arxiv.org/abs/2011.13456 (Accessed 18 October 2023).
32. M. S. Albergo, E. Vanden-Eijnden, Building normalizing flows with stochastic interpolants. arXiv [Preprint] (2023). https://arxiv.org/abs/2209.15571 (Accessed 18 October 2023).
33. M. S. Albergo, N. M. Boffi, E. Vanden-Eijnden, Stochastic interpolants: A unifying framework for flows and diffusions. arXiv [Preprint] (2023). https://arxiv.org/abs/2303.08797 (Accessed 18 October 2023).
34. N. M. Boffi, E. Vanden-Eijnden, Probability flow solution of the Fokker–Planck equation. *Mach. Learn. Sci. Technol.* **4**, 035012 (2023).
35. M. E. Cates, J. Tailleur, Motility-induced phase separation. *Annu. Rev. Condens. Matter Phys.* **6**, 219–244 (2015).
36. F. Noe, S. Olsson, J. Kohler, H. Wu, Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365**, eaaw1147 (2019).
37. Y. Khoo, J. Lu, L. Ying, Solving for high dimensional committor functions using artificial neural networks. arXiv [Preprint] (2018). https://arxiv.org/abs/1802.10275 (Accessed 18 October 2023).
38. Y. Khoo, J. Lu, L. Ying, Solving parametric PDE problems with artificial neural networks. *Eur. J. Appl. Math.* **32**, 421–435 (2021).
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. 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).
41. M. Gabrie, G. M. Rotskoff, E. Vanden-Eijnden, Adaptive Monte Carlo augmented with normalizing flows. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2109420119 (2022).
42. 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).
43. B. Lusch, J. N. Kutz, S. L. Brunton, Deep learning for universal linear embeddings of nonlinear dynamics. *Nat. Commun.* **9**, 4950 (2018).
44. S. H. Rudy, S. L. Brunton, J. L. Proctor, J. N. Kutz, Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614 (2017).
45. N. Nusken, L. Richter, Solving high-dimensional Hamilton–Jacobi–Bellman PDEs using neural networks: Perspectives from the theory of controlled diffusions and measures on path space. *Partial Differ. Equations Appl.* **2**, 48 (2021).
46. W E, J Han, A Jentzen,, Algorithms for solving high dimensional PDEs: From nonlinear Monte Carlo to machine learning. *Nonlinearity* **35**, 278 (2021).
47. W E, B Yu,, The Deep Ritz method: A deep learning-based numerical algorithm for solving variational problems. *Commun. Math. Stat.* **6**, 1–12 (2018).
48. N. Kovachki *et al*., Neural operator: Learning maps between function spaces with applications to PDEs. *J. Mach. Learn. Res.* **24**, 1–97 (2023).
49. Z. Li *et al*., Fourier neural operator for parametric partial differential equations. arXiv [Preprint] (2020). https://arxiv.org/abs/2010.08895 (Accessed 18 October 2023).
50. G. Carleo *et al*., Machine learning and the physical sciences. *Rev. Mod. Phys.* **91**, 045002 (2019).
51. S. L. Brunton, B. R. Noack, P. Koumoutsakos, Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **52**, 477–508 (2020).
52. R. Supekar *et al*., Learning hydrodynamic equations for active matter from particle simulations and experiments. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2206994120 (2023).
53. G. E. Karniadakis *et al*., Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
54. D. Kochkov *et al*., Machine learning-accelerated computational fluid dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2101784118 (2021).
55. G. Carleo, M. Troyer, Solving the quantum many-body problem with artificial neural networks. *Science* **355**, 602–606 (2017).
56. J. Hermann, Z. Schatzle, F. Noe, Deep-neural-network solution of the electronic Schrodinger equation. *Nat. Chem.* **12**, 891–897 (2020).
57. Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby, T. F. Miller III, OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **153**, 124111 (2020).
58. D. Pfau, J. S. Spencer, A. G. d. G. Matthews, W. M. C. Foulkes, Ab-initio solution of the many-electron Schrödinger equation with Deep Neural Networks. *Phys. Rev. Res.* **2**, 033429 (2020).
59. M. Paoluzzi, C. Maggi, U. Marini Bettolo Marconi, N. Gnan, Critical phenomena in active matter. *Phys. Rev. E* **94**, 052602 (2016).
60. G. Szamel, Self-propelled particle in an external potential: Existence of an effective temperature. *Phys. Rev. E* **90**, 012111 (2014).
61. E. Flenner, G. Szamel, L. Berthier, The nonequilibrium glassy dynamics of self-propelled particles. *Soft Matter* **12**, 7136–7149 (2016).
62. L. Caprini, U. M. B. Marconi, A. Puglisi, A. Vulpiani, The entropy production of Ornstein–Uhlenbeck active particles: A path integral method for correlations. *J. Stat. Mech: Theory Exp.* **2019**, 053203 (2019).
63. E. Nelson, *Dynamical Theories of Brownian Motion* (Princeton University Press, 1967).
64. E. Aurell, K. Gawedzki, C. Mejia-Monasterio, R. Mohayaee, P. Muratore-Ginanneschi, Refined second law of thermodynamics for fast random processes. *J. Stat. Phys.* **147**, 487–505 (2012).
65. S. Chennakesavalu, G. M. Rotskoff, Unified, geometric framework for nonequilibrium protocol optimization. *Phys. Rev. Lett.* **130**, 107101 (2023).
66. S. Pigolotti, I. Neri, E. Roldan, F. Julicher, Generic properties of stochastic entropy production. *Phys. Rev. Lett.* **119**, 140604 (2017).
67. U. M. B. Marconi, A. Puglisi, C. Maggi, Heat, temperature and Clausius inequality in a model for active Brownian particles. *Sci. Rep.* **7**, 46496 (2017).
68. A. Puglisi, U. Marini Bettolo Marconi, Clausius relation for active particles: What can we learn from fluctuations. *Entropy* **19**, 356 (2017).
69. L. Dabelow, S. Bo, R. Eichhorn, Irreversibility in active matter systems: Fluctuation theorem and mutual information. *Phys. Rev. X* **9**, 021009 (2019).
70. V. Y. Chernyak, M. Chertkov, C. Jarzynski, Path-integral analysis of fluctuation theorems for general Langevin processes. *J. Stat. Mech: Theory Exp.* **2006**, P08001 (2006).
71. B. D. Anderson, Reverse-time diffusion equation models. *Stochastic Processes Their Appl.* **12**, 313–326 (1982).
72. L. Onsager, S. Machlup, Fluctuations and irreversible processes. *Phys. Rev.* **91**, 1505–1512 (1953).
73. B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications* (Springer, ed. 6, 2003).
74. D. P. Kingma, J. Ba, ADAM: A method for stochastic optimization. arXiv [Preprint] (2017). https://arxiv.org/abs/1412.6980 (Accessed 18 October 2023).
75. A. Bietti, L. Venturi, J. Bruna, On the sample complexity of learning with geometric stability. arXiv [Preprint] (2021). https://arxiv.org/abs/2106.07148 (Accessed 18 October 2023).
76. Y. Song, C. Durkan, I. Murray, S. Ermon, Maximum likelihood training of score-based diffusion models. arXiv [Preprint] (2021). https://arxiv.org/abs/2101.09258 (Accessed 18 October 2023).
77. T. Karras, M. Aittala, T. Aila, S. Laine, "Elucidating the design space of diffusion-based generative models" in *Advances in Neural Information Processing Systems 35* (NeurIPS 2022), S. Koyejo *et al*., Eds. (Curran Associates, Inc., 2022), pp. 26565–26577.
78. A. Vaswani *et al*., Attention is all you need. arXiv [Preprint] (2017). https://arxiv.org/abs/1706.03762 (Accessed 18 October 2023).
79. M. Zaheer *et al*., Big bird: Transformers for longer sequences. arXiv [Preprint] (2021). https://arxiv.org/abs/2007.14062 (Accessed 18 October 2023).
80. Q. Wang *et al*., Learning deep transformer models for machine translation. arXiv [Preprint] (2019). https://arxiv.org/abs/1906.01787 (Accessed 18 October 2023).
81. N. Parmar *et al*., Image transformer. arXiv [Preprint] (2018). https://arxiv.org/abs/1802.05751 (Accessed 18 October 2023).
82. A. Dosovitskiy *et al*., An image is worth 16x16 words: Transformers for image recognition at scale. arXiv [Preprint] (2021). https://arxiv.org/abs/2010.11929 (Accessed 18 October 2023).
83. A. Chandra, L. Tünnermann, T. Löfstedt, R. Gratz, Transformer-based deep learning for predicting protein properties in the life sciences. *eLife* **12**, e82819 (2023).
84. I. von Glehn, J. S. Spencer, D. Pfau, A self-attention ansatz for ab-initio quantum chemistry. arXiv [Preprint] (2023). https://arxiv.org/abs/2211.13672 (Accessed 18 October 2023).
85. S. Abnar, W. Zuidema, Quantifying attention flow in transformers. arXiv [Preprint] (2020). https://arxiv.org/abs/2005.00928 (Accessed 18 October 2023).
86. J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization. arXiv [Preprint] (2016). https://arxiv.org/abs/1802.05751 (Accessed 18 October 2023).
87. P. Vincent, A connection between score matching and denoising autoencoders. *Neural Comput.* **23**, 1661–1674 (2011).
88. 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).
89. T. Dao, D. Y. Fu, S. Ermon, A. Rudra, C. Ré, FlashAttention: Fast and memory-efficient exact attention with IO-awareness. arXiv [Preprint] (2022). https://arxiv.org/abs/2205.14135 (Accessed 18 October 2023).
90. T. Dao, FlashAttention-2: Faster attention with better parallelism and work partitioning. arXiv [Preprint] (2023). https://arxiv.org/abs/2307.08691 (Accessed 18 October 2023).
91. N. M. Boffi, E. Vanden-Eijnden, Public code repository for "Deep learning probability flows and entropy production rates in active matter." Github. https://github.com/nmboffi/active_pflows. Deposited 1 April 2024.