ANNUAL REVIEWS

Annual Review of Condensed Matter Physics

Irreversibility and Biased Ensembles in Active Matter: Insights from Stochastic Thermodynamics

Étienne Fodor,[1] Robert L. Jack,[2,3] and Michael E. Cates[2]

[1]Department of Physics and Materials Science, University of Luxembourg, Luxembourg; email: etienne.fodor@uni.lu

[2]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom; email: rlj22@cam.ac.uk, m.e.cates@damtp.cam.ac.uk

[3]Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

Annu. Rev. Condens. Matter Phys. 2022. 13:215–38

First published as a Review in Advance on November 10, 2021

The *Annual Review of Condensed Matter Physics* is online at conmatphys.annualreviews.org

https://doi.org/10.1146/annurev-conmatphys-031720-032419

Keywords

self-propelled particles, nonequilibrium field theories, dissipation, entropy production, large deviations, phase transitions

Abstract

Active systems evade the rules of equilibrium thermodynamics by constantly dissipating energy at the level of their microscopic components. This energy flux stems from the conversion of a fuel, present in the environment, into sustained individual motion. It can lead to collective effects without any equilibrium equivalent, some of which can be rationalized by using equilibrium tools to recapitulate nonequilibrium transitions. An important challenge is then to delineate systematically to what extent the character of these active transitions is genuinely distinct from equilibrium analogs. We review recent works that use stochastic thermodynamics tools to identify, for active systems, a measure of irreversibility comprising a coarse-grained or informatic entropy production. We describe how this relates to the underlying energy dissipation or thermodynamic entropy production, and how it is influenced by collective behavior. Then, we review the possibility of constructing thermodynamic ensembles out of equilibrium, where trajectories are biased toward atypical values of nonequilibrium observables. We show that this is a generic route to discovering unexpected phase transitions in active matter systems, which can also inform their design.

1. INTRODUCTION

Active matter is a class of nonequilibrium systems whose components extract energy from the environment to produce an autonomous motion (1–3). Examples are found in biological systems, such as swarms of bacteria (4) and assemblies of cells (5); social systems, such as groups of animals (6) and human crowds (7); and synthetic systems, such as vibrated polar particles (8) and catalytic colloids (9). The combination of individual self-propulsion and interactions between individuals can lead to collective effects without any equivalent in equilibrium. Examples include collective directed motion, as observed in bird flocks (6), and the spontaneous formation of clusters made of purely repulsive particles without even a depletion interaction, as reported for Janus colloids in a fuel bath (9). To study these effects, minimal models have been proposed based on simple dynamical rules. Some are formulated at particle level, for instance with automaton rules (10) or by extending Langevin dynamics (11). Others work at the coarse-grained level in terms of hydrodynamic fields (12, 13). The latter can be obtained systematically by coarse-graining the particle dynamics or postulated phenomenologically. Both the microscopic and hydrodynamic approaches have successfully reproduced experimental behavior, such as the emergence of a long-ranged polar order, known as the flocking transition (14), and phase separation that occurs without any microscopic attraction, known as motility-induced phase separation (MIPS) (15).

Although active systems evade the rules of equilibrium statistical mechanics, some works have built a framework to predict their properties based on the partial applicability of thermodynamic concepts beyond equilibrium. A first approach was to map some active systems onto equilibrium ones with a similar steady state, allowing one to define effective free energies for active matter (16, 17). Other studies have extended the definitions of standard observables such as pressure (18–20), surface tension (21, 22), and chemical potential (23, 24), hoping to establish equations of state relating, for instance, pressure and density. Interestingly, one finds that the existence of such state functions cannot generically be relied upon (e.g., the pressure on a wall can depend on the type of wall) (20), highlighting the limitations of equilibrium analogies. In trying to build a thermodynamic framework for active matter, an important challenge is then to identify regimes in which it is possible to deploy equilibrium tools and to clearly distinguish these from genuine nonequilibrium regimes. In other words, how should we delineate where and when activity really matters in the emerging phenomenology? And, most importantly, can we define a systematic, unambiguous measure of the departure from equilibrium?

In passive systems, the steady state (Boltzmann) distribution involves the Hamiltonian that also drives the microdynamics. This ensures thermodynamic consistency of the dynamics, which further implies that the mechanical and thermodynamic definitions of pressure are equivalent, and precludes (real space) steady-state currents. Although such currents offer a clear nonequilibrium signature of activity, as observed, for instance, in collective motion (14), it is more challenging to distinguish, say, MIPS from standard phase separation without tracking individual particle motion (15). In particular, it is not helpful to define departure from equilibrium via deviation from an effective Boltzmann distribution in systems that are not thermodynamically consistent. Instead, the cornerstone of modern statistical mechanics, which allows one to dissociate fundamentally active and passive systems, is the reversibility of equilibrium dynamics (25). In a steady equilibrium state, forward and backward dynamics are indistinguishable, because all fluctuations exhibit time-reversal symmetry (TRS). This constraint entails other important properties, such as the fluctuation–dissipation theorem (FDT) (26), and the absence of dissipated heat at equilibrium. For active systems, the irreversibility of the dynamics then stands out as the key differentiating property that causes the violation of these and other equilibrium laws.

To quantify the breakdown of TRS, we use stochastic thermodynamics (27, 28). This framework extends standard notions of thermodynamics, such as the first and second laws, to fluctuating

trajectories. It was first developed for systems in contact with one or more reservoirs, such as energy and/or volume reservoirs, provided that each reservoir satisfies some equilibrium constraints. Such constraints do not preclude the system from operating out of equilibrium, for instance, under external fields and/or thermostats at different temperatures: In this context, the thermodynamic consistency of the dynamics yields some explicit connections among irreversibility, dissipated heat, and entropy production. In contrast, the dynamics of active systems are often formulated via phenomenological arguments, which do not follow a priori the requirements of thermodynamic consistency. A natural question is, then, Is it legitimate to extend the methods of stochastic thermodynamics to active matter? And, what can we learn from their irreversibility measures when their connection with other thermodynamic observables becomes blurred?

Another interesting direction in active matter theory is to generalize familiar thermodynamic ensembles to nonequilibrium settings. Here, the important roles of steady-state currents and other types of irreversible dynamics make it insufficient to study ensembles of configurations (characterized by a stationary measure, of which the Boltzmann distribution is an example). Instead, we must address ensembles of trajectories, as used previously to analyze fluctuation theorems (29), and other dynamical effects (30–32). Such ensembles are built similarly to the canonical ensemble of statistical mechanics, with relevant (extensive) observables coupled to conjugate (intensive) fields. In practice, one selects a dynamical observable of interest and focuses on dynamical trajectories where it has some atypical target value. The resulting biased ensembles of rare trajectories are intrinsically connected to large deviation theory (33, 34). As such, they provide insight into mechanisms for unusual and interesting fluctuations. In the passive context, they have proven useful for unveiling dynamical transitions, e.g., in kinetically constrained models of glasses (32). This leads one to ask the following question: How do the transitions in biased ensembles of active matter differ from their passive counterparts? And then, can we identify settings in which bias-induced transitions emerge in active matter that have no passive equivalent?

In what follows (Section 2), we describe how the irreversibility of active systems can be quantified, for both particle-based and hydrodynamic theories, using generalized forms of entropy production. We discuss how to relate these, in certain cases, to energy dissipation and how they allow one to identify phases and/or spatial regions in which activity comes to the fore. In Section 3, we then describe how biased ensembles provide novel insights into the emergence of collective effects in active matter. We review some useful tools of large deviation theory, such as representing dynamical bias in terms of control forces, show the utility of these methods in an active context, and discuss how they can reveal unexpected transitions in generic active systems. In Section 4, we give a brief conclusion that includes the implications of these methods for the rational design of new materials.

2. IRREVERSIBILITY AND DISSIPATION

2.1. Modeling Active Matter: From Particles to Hydrodynamic Fields

Typical particle-based dynamics start from the seminal Langevin equation. Provided that inertia is negligible, this balances the forces stemming from the heat bath (damping and thermal noise), the force deriving from a potential U, and nonconservative forces $\mathbf{f}_i$:

$$\dot{\mathbf{r}}_i = \mu\left(\mathbf{f}_i - \nabla_i U\right) + \sqrt{2\mu T}\boldsymbol{\xi}_i, \qquad 1.$$

where μ is the mobility, T the temperature of the bath, and $\boldsymbol{\xi}_i$ a set of zero-mean, unit-variance white Gaussian noises: $\langle \xi_{i\alpha}(t)\xi_{j\beta}(t')\rangle = \delta_{ij}\delta_{\alpha\beta}\delta(t-t')$, where Latin and Greek indices, respectively, refer to particle labels and spatial coordinates. Hereafter, we refer to such noise as unit white noise. The potential U describes particle interactions and/or an external perturbation applied by

the operator. The nonconservative forces $\mathbf{f}_i$ model self-propulsion, which converts a source of energy, present in the environment, into directed motion. In principle, $\mathbf{f}_i$ can include external perturbations that do not derive from a potential. When $\mathbf{f}_i = \mathbf{0}$, the system is at equilibrium with Boltzmann statistics ($\propto e^{-U/T}$).

The self-propulsion force $\mathbf{f}_i$ is often modeled as, effectively, an additional noise. In contrast with the thermal noise $\boldsymbol{\xi}_i$, it has some persistence, which captures the propensity of active particles to sustain directed motion. Typically, a persistence time τ sets the exponential decay of the two-point correlations: $\langle f_{i\alpha}(t) f_{j\beta}(0)\rangle = \delta_{ij}\delta_{\alpha\beta} f_0^2 e^{-|t|/\tau}$. In recent years, two variants have emerged as popular models of self-propelled particles. First, for active Ornstein–Uhlenbeck particles (AOUPs), the statistics of $\mathbf{f}_i$ is Gaussian, so that it can be viewed as obeying an (autonomous) Ornstein–Uhlenbeck process (17, 35):

$$\text{AOUP:} \quad \tau \dot{\mathbf{f}}_i = -\mathbf{f}_i + f_0\sqrt{2\tau}\boldsymbol{\zeta}_i, \qquad 2.$$

where the $\boldsymbol{\zeta}_i$ are unit white noises. Scaling the amplitude as $\mu f_0 = \sqrt{\mu T_a/\tau}$, the self-propulsion itself converges to a white noise source in the limit of vanishing persistence, in which regime the system reduces to passive particles at temperature $T + T_a$. Second, for active Brownian particles (ABPs), the magnitude of $\mathbf{f}_i$ is now fixed, which leads to non-Gaussian statistics, described by the following process in two dimensions (11):

$$\text{ABP:} \quad \mathbf{f}_i = f_0(\cos\theta_i, \sin\theta_i), \quad \dot{\theta}_i = \sqrt{2/\tau}\,\eta_i, \qquad 3.$$

where η_i is a unit white noise. In Equations 2 and 3, the self-propulsion is independent for each particle. This corresponds to isotropic particles, which undergo MIPS for sufficiently large persistence and density (15). In contrast, many models consider alignment interactions, yielding a collective (oriented) motion at high density and enhanced persistence (14).

To study the emergence of nonequilibrium collective effects, it is helpful to introduce the relevant hydrodynamic fields. These can be identified, and their dynamical equations found, by coarse-graining the microscopic equations of motion using standard tools of stochastic calculus (36). Another approach consists of postulating such hydrodynamic theories from phenomenological arguments, respecting any conservation laws and spontaneously broken symmetries (1). For instance, a theory of MIPS can be found by considering the dynamics of a scalar field $\phi(\mathbf{r}, t)$ that encodes the local concentration of particles (13, 37):

$$\dot{\phi} = -\nabla\cdot\mathbf{J} = -\nabla\cdot\left(-\lambda\nabla\frac{\delta\mathcal{F}}{\delta\phi} + \mathbf{J}_\phi + \sqrt{2\lambda D}\boldsymbol{\Lambda}_\phi\right), \qquad 4.$$

where λ is collective mobility, D is a noise temperature, and $\boldsymbol{\Lambda}_\phi$ is (spatiotemporal) unit white noise. The term $\mathbf{J}_\phi$ embodies contributions driving the current $\mathbf{J}$ of ϕ that cannot be derived from any free energy functional $\mathcal{F}$ as $-\lambda\nabla(\delta\mathcal{F}/\delta\phi)$. For instance, choosing a standard ϕ^4, square-gradient functional for $\mathcal{F}$ captures phase separation in equilibrium. On introducing activity, symmetry arguments require that $\mathbf{J}_\phi$ takes the form $\nabla(\nabla\phi)^2$ and/or $(\nabla\phi)(\nabla^2\phi)$ to lowest order in ϕ and its gradients (13, 37). The former term shifts the coexisting densities with respect to the equilibrium model (13), whereas the latter can lead to microphase separation, with either the vapor phase decorated with liquid droplets or the liquid phase decorated with vapor bubbles (37). Note that, although the conservative term $-\lambda\nabla(\delta\mathcal{F}/\delta\phi)$ captures all the contributions to the coarse-grained dynamics that respect TRS, it may depend on nonequilibrium parameters of the microscopic dynamics (37).

The dynamics in Equation 4 can also be coupled to a polar field $\mathbf{p}(\mathbf{r}, t)$, representing the local mean orientation of particles, to study, for instance, collective flocking motion (14):

$$\dot{\mathbf{p}} = -\frac{\delta\mathcal{F}}{\delta\mathbf{p}} + \mathbf{J}_p + \sqrt{2D}\boldsymbol{\Lambda}_p, \qquad 5.$$

where the rotational mobility is taken as unity, and $\mathbf{\Lambda}_p$ is another spatiotemporal unit white noise. As with $\mathbf{J}_\phi$, the term $\mathbf{J}_p$ represents active relaxations that cannot be written as free-energy derivatives. Note that, for the passive limit of the coupled model to respect detailed balance, the free energy $\mathcal{F}$ in Equation 5 must be the same as that in Equation 4, and the value of D in each equation must also coincide. In the absence of activity ($\mathbf{J}_\phi = \mathbf{J}_p = \mathbf{0}$), the system has Boltzmann statistics $\propto e^{-\mathcal{F}/D}$. To capture the emergence of polar order, a minimal choice is to take $\mathcal{F}$ as a $\mathbf{p}^4$ functional, with $\mathbf{J}_\phi$ proportional to $\mathbf{p}$, and $\mathbf{J}_p$ a linear combination of $\nabla\phi$, $\phi\mathbf{p}$, and $(\mathbf{p}\cdot\nabla)\mathbf{p}$ (1, 14). This can yield polar bands traveling across an apolar background, which is a key signature of aligning active particles models (10). In fact, much of the physics of flocks is retained by setting $\lambda = 0$ in Equation 4 (12). By contrast more complicated theories, describing the emergence of nematic order (14) and/or the coupling to a momentum-conserving fluid (38, 39), introduce additional hydrodynamic fields beyond ϕ and $\mathbf{p}$ as considered above.

2.2. How Far from Equilibrium Is Active Matter?

We present various approaches to define time reversal, and to quantify the corresponding TRS breakdown, in popular models of active matter. This leads us to propose several measures of the departure from equilibrium.

2.2.1. Forward and time-reversed dynamics: path-probability representation. To quantify the irreversibility of active dynamics, we compare the path probability $\mathcal{P}$ to realize a dynamical trajectory (across a time interval $[0, t]$) with that for its time-reversed counterpart (27, 28, 40). Choosing the time-reversed dynamics is a matter of subtlety (35, 39, 41–49). For the particle-based dynamics in Equation 1, exploiting the known statistics of the (unit white) noise, trajectories of position $\mathbf{r}_i$ and self-propulsion $\mathbf{f}_i$ have probability $\mathcal{P} \sim e^{-\mathcal{A}}$ (50), with action

$$\mathcal{A} = \frac{1}{4\mu T}\int_0^t \sum_i \left[\dot{\mathbf{r}}_i + \mu\left(\nabla_i U - \mathbf{f}_i\right)\right]^2 \mathrm{d}t' + \mathcal{A}_f. \qquad 6.$$

We use Stratonovich discretization, omitting a time-symmetric contribution that is irrelevant in what follows. $\mathcal{A}_f$ specifies the statistics of the self-propulsion force $\mathbf{f}_i$. For AOUPs and ABPs, $\mathbf{f}_i$ has autonomous dynamics, independent of position $\mathbf{r}_i$. One can then in principle integrate out that dynamics to obtain a reduced action $\mathcal{A}_r$. So far, the explicit expression has been derived only for AOUPs (35, 48, 51, 52):

$$\mathcal{A}_r = \int_0^t \mathrm{d}t' \int_0^t \mathrm{d}t'' \sum_i \left(\dot{\mathbf{r}}_i + \mu\nabla_i U\right)\big|_{t'} \cdot \left(\dot{\mathbf{r}}_i + \mu\nabla_i U\right)\big|_{t''} \Gamma(t' - t''). \qquad 7.$$

The kernel Γ is an even function of time t, which depends on μ, T, f_0, and τ. It reduces to $\Gamma(t) = \delta(t)/[4\mu(T + T_a)]$ for vanishing persistence ($\tau \to 0$), when the self-propulsion amplitude scales as $\mu f_0 = \sqrt{\mu T_a/\tau}$ (48), as expected for passive particles. The actions $\mathcal{A}$ and $\mathcal{A}_r$ describe the same dynamics from different viewpoints. The former tracks the realizations of position and self-propulsion, whereas the latter follows the time evolution of positions only. Thus, $\mathcal{A}_r$ treats the self-propulsion as a source of noise, without resolving its time evolution. In contrast, $\mathcal{A}$ regards $\mathbf{f}_i$ as a configurational coordinate, which might in principle be read out from the particle shape. For instance, Janus particles, which are spheres with distinct surface chemistry on two hemispheres (2, 9), generally self-propel along a body-fixed heading vector pointing from one to the other (see also **Figure 1**).

To compare forward and backward trajectories, one needs to define a time-reversed version of the dynamics. This necessarily transforms the time variable as $t' \to t - t'$, and it can also involve flipping some of the variables or fields (as detailed below). For particle-based dynamics,

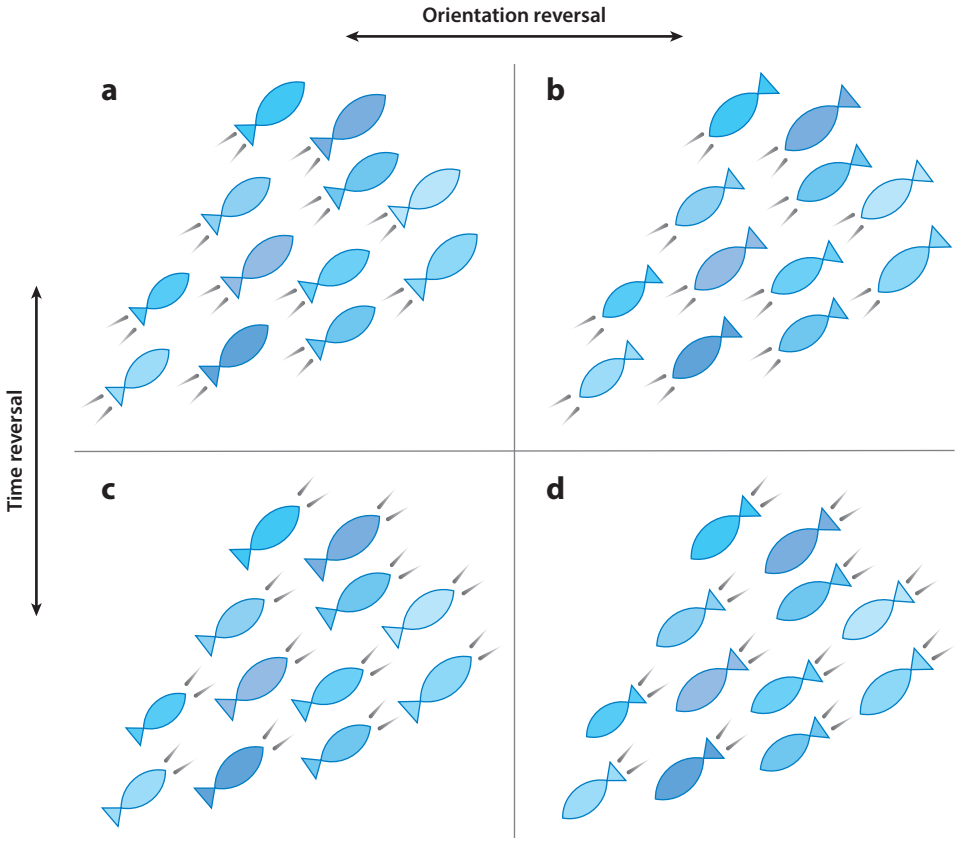

Figure 1

TR for active particles (fish). Each has an orientation (head–tail axis) and moves (swims) as indicated by the *wake lines* behind it. The TR operation may or may not reverse orientations. (*a*) Natural dynamics. (*b*) Reversing orientation but not motion: fish swim tail first along their original directions. (*c*) Reversing motion but not orientation: fish swim tail first, opposite to their original direction. (*d*) Reversing both motion and orientation: fish swim headfirst, opposite to their original direction. The situations in panels *b* and *c* occur with extremely low probability in the natural dynamics (requiring exceptional noise realizations). In dilute regimes (not shown) the case in panel *d* is equiprobable to that shown in panel *a*: Individual fish swim headfirst in both cases. For the shoals shown here, however, the case in panel *d* is less probable than that in panel *a*, because the natural dynamics has more fish at the front of the shoal: time-reversal symmetry is broken at a collective level even if orientations are reversed. Abbreviation: TR, time reversal.

time reversal of the positions $\mathbf{r}_i$ is unambiguous. Writing $\mathcal{P}$ in terms of the reduced action $\mathcal{A}_r$ in Equation 7, the time-reversed counterpart $\mathcal{A}_r^R$ reads as

Time reversal: several backward dynamics that can be defined depending on the time signature of the degrees of freedom

$$\mathcal{A}_r^R = \int_0^t \mathrm{d}t' \int_0^t \mathrm{d}t'' \sum_i \left(\dot{\mathbf{r}}_i - \mu\nabla_i U\right)\big|_{t'} \cdot \left(\dot{\mathbf{r}}_i - \mu\nabla_i U\right)\big|_{t''} \Gamma(t'-t'')\,. \qquad 8.$$

Here, we have changed variables as $\{t', t''\} \to \{t-t', t-t''\}$, and used the fact that Γ is even. For the full action $\mathcal{A}$, defining the time-reversed dynamics requires that we fix the time signature of the self-propulsion $\mathbf{f}_i$. Indeed, there are two possible reversed actions, $\mathcal{A}^{R,\pm}$, where + and − indices

refer, respectively, to even ($\mathbf{f}_i \to \mathbf{f}_i$) and odd ($\mathbf{f}_i \to -\mathbf{f}_i$) self-propulsion:

$$\mathcal{A}^{R,\pm} = \frac{1}{4\mu T}\int_0^t \sum_i \left[\dot{\mathbf{r}}_i - \mu\left(\nabla_i U \mp \mathbf{f}_i\right)\right]^2 \mathrm{d}t' + \mathcal{A}_f. \qquad 9.$$

We assumed that $\mathcal{A}_f$ is the same for forward and time-reversed dynamics, which holds for ABPs and AOUPs (but not for systems with aligning interactions; 53). $\mathcal{A}_r^R$ and $\mathcal{A}^{R,\pm}$ correspond to three different definitional choices of the time-reversed dynamics. The first of these deliberately discards the realizations of the self-propulsion forces in order to compare the forward and backward motions of particle positions. This choice is widely used for AOUPs, where particles have no heading vector. It is also the only possible choice when $T = 0$ in Equation 1 (see 35). In contrast, $\mathcal{A}^{R,\pm}$, which for each we assume that both position and self-propulsion are tracked, are usually chosen for particles with a heading vector (45, 53, 54).

For the field theories of Section 2.1, defining the time-reversed dynamics requires us to choose the time signature of each field. This choice depends on the physical system and the phases under scrutiny (Section 2.4). Scalar fields associated with a local density are clearly even (13, 39), whereas scalar fields schematically representing polarization (55), or stream functions for fluid flow (56), should generally be odd. Similarly, the vector field $\mathbf{p}$ can be chosen even (57) or odd (47), depending on whether it is viewed as the local orientation of particles (mean heading vector; see **Figure 1**) or directly as a local velocity (12). One also has to decide which fields to retain or ignore when comparing forward and backward paths. This choice is the counterpart at field level of retaining ($\mathcal{A}$) or ignoring ($\mathcal{A}_r$) the propulsive forces in a system of AOUPs. For example, in a system with ϕ and $\mathbf{p}$ variables, one can choose whether or not to keep separate track of the density current $\mathbf{J}$ alongside ϕ and $\mathbf{p}$.

In general, there are four different versions of time-reversed dynamics of Equations 4 and 5, provided that each one of the fields ϕ and $\mathbf{p}$ can be either odd or even. Often, though, constraints on the time signatures of the fields restrict these choices. To ensure that the dynamics is invariant under time reversal in the passive limit, the free energy $\mathcal{F}$ must be the same in forward and time-reversed dynamics. This ensures that odd fields can only appear as even powers in $\mathcal{F}$. Hence, if $\mathbf{p}$ appears linearly in $\mathcal{F}$, which for liquid crystal models often includes a $\mathbf{p}\cdot\nabla\phi$ or anchoring term, it must be chosen even. Further constraints arise if some of the possible noise terms are set to zero (e.g., 12) so that certain fields are deterministically enslaved to others. For instance, if $\dot{\phi} = -\nabla\cdot\mathbf{p}$ (without noise), for consistency ϕ and $\mathbf{p}$ necessarily have different signatures.

2.2.2. Distance from equilibrium: breakdown of time-reversal symmetry. For a given dynamics, once its time-reversed counterpart has been chosen, we can define systematically an irreversibility measure $\mathcal{S}$ as (27, 28, 40)

$$\mathcal{S} = \lim_{t\to\infty}\frac{1}{t}\left\langle \ln\frac{\mathcal{P}}{\mathcal{P}^R}\right\rangle = \lim_{t\to\infty}\frac{1}{t}\langle \mathcal{A}^R - \mathcal{A}\rangle. \qquad 10.$$

The average $\langle\cdot\rangle$ is taken over noise realizations. The limit of long trajectories gets rid of any transient relaxations to focus on steady-state fluctuations. Equation 10 is a cornerstone of stochastic thermodynamics (27, 28, 40). It was first established in thermodynamically consistent models (40), where $\mathcal{S}$ can be shown to be the entropy production rate (EPR) governing the dissipated heat (27, 28). In models in which $\mathcal{S}$ has this meaning, thermodynamics constrains the choice of time reversal. In Section 2.3, we return to the question of how far the thermodynamic interpretation extends to active systems. Meanwhile, $\mathcal{S}$ already offers an unambiguous measure of TRS breakdown in active matter, which we refer to as informatic entropy production rate (IEPR). Because $\mathcal{S}$ changes

Dissipated heat: the amount of energy transferred from the system to the surrounding thermostat due to internal active processes

Informatic entropy production: a quantification of the breakdown of time-reversal symmetry via the difference between forward and backward path probabilities

when the dynamics is coarse-grained, by eliminating unwanted variables, it differs when retaining ($\mathcal{A}$) or ignoring ($\mathcal{A}_r$) the propulsive forces.

In particle-based dynamics, choosing the action $\mathcal{A}_r$ so that only particle positions are tracked, the corresponding IEPR $\mathcal{S}_r$ follows from Equations 7, 8, and 10 as

$$\mathcal{S}_r = -4\mu \int_{-\infty}^{\infty} \Gamma(t) \sum_i \langle \dot{\mathbf{r}}_i(t) \cdot \nabla_i U(0) \rangle \mathrm{d}t, \qquad 11.$$

where we have again used the fact that Γ is even. The IEPR $\mathcal{S}_r$ vanishes in the absence of any potential, showing that the dynamics satisfies TRS for free active particles and also for an external harmonic potential $U \sim \mathbf{r}_i^2$ (35, 52). When $T = 0$, it reduces to $\mathcal{S}_r = \tau \langle (\sum_i \dot{\mathbf{r}}_i \cdot \nabla_i)^3 U \rangle / [2(\mu f_0)^2]$ (35). When (by using the full action $\mathcal{A}$) one follows the dynamics of both position and self-propulsion, the two possible IEPRs found from Equations 6, 9, and 10 are

$$\mathcal{S}^+ = \frac{1}{T} \sum_i \langle \dot{\mathbf{r}}_i \cdot \mathbf{f}_i \rangle, \qquad \mathcal{S}^- = \frac{\mu}{T} \sum_i \langle \nabla_i U \cdot \mathbf{f}_i \rangle. \qquad 12.$$

We have used the fact that $\frac{1}{t} \int_0^t \sum_i \dot{\mathbf{r}}_i \cdot \nabla_i U \mathrm{d}t' = \frac{U(t) - U(0)}{t}$ vanishes at large t. Substituting the expression for $\dot{\mathbf{r}}_i$ from Equation 1 into Equation 12 yields $\mathcal{S}^+ + \mathcal{S}^- = N\mu f_0^2 / T$, where N is the particle number, using that self-propulsion $\mathbf{f}_i$ and thermal noise $\sqrt{2\mu T}\boldsymbol{\xi}_i$ are uncorrelated. Therefore, in the absence of any potential U, $\mathcal{S}^-$ vanishes identically, whereas $\mathcal{S}^+$ remains nonzero. Indeed, $\mathcal{S}^-$ compares trajectories whose velocity and self-propulsion both flip on time reversal, retaining alignment (up to thermal noise) between the two: The forward and backward dynamics are indistinguishable unless potential forces intervene. In contrast, $\mathcal{S}^+$ quantifies how different trajectories are when particles move either along with or opposite to their self-propulsion (**Figure 1**), yielding the contribution $N\mu f_0^2 / T$ even when $U = 0$.

The IEPR $\mathcal{S}^+$ is lowest (and $\mathcal{S}^-$ highest) when the propulsive force $\mathbf{f}_i$ balances the interaction force $-\nabla_i U$ so that particles are almost arrested. Hence, both IEPRs are sensitive to the formation of particle clusters, which is associated with dynamical slowing-down for isotropic particles (15), and to formation of a polarized state for aligning particles (14). Note that $\mathcal{S}^+$ is proportional to the contribution of self-propulsion to pressure, known as swim pressure (19, 20). Finally, in the presence of alignment, any dynamical interaction torques appear explicitly in $\mathcal{S}$ (53). We defer further discussion on how interactions shape irreversibility, for both particle-based dynamics and field theories, to Section 2.4.

2.3. Energetics Far from Equilibrium: Extracted Work and Dissipated Heat

A major success of stochastic thermodynamics is to extend the definition of observables from classical thermodynamics to cases in which fluctuations cannot be neglected (27, 28, 40). Although first proposed for thermodynamically consistent models, this approach can be extended to some (not all) types of active matter.

2.3.1. Particle-based approach: energy transfers in microscopic dynamics. For particle dynamics, assuming that the potential U depends on the set of control parameters α_n, the work $\mathcal{W}$ produced by varying α_n during a time t is (27, 28)

$$\mathcal{W} = \int_0^t \sum_n \dot{\alpha}_n \frac{\partial U}{\partial \alpha_n} \mathrm{d}t'. \qquad 13.$$

Note that $\mathcal{W}$ is stochastic due to the thermal noise and the self-propulsion, even though the protocol $\alpha_n(t)$ is deterministic. Equation 13 relies on the precept that some external operator perturbs

the system though U, without prior knowledge of the detailed dynamics; it applies equally for active and passive systems.

The heat $\mathcal{Q}$ is now defined as the energy delivered by the system to the surrounding thermostat. For the dynamics of Equation 1, the effect of the thermostat is encoded in the damping force and the thermal noise; the fluctuating observable $\mathcal{Q}$ then follows as (27, 28)

$$\mathcal{Q} = \int_0^t \sum_i \frac{\dot{\mathbf{r}}_i}{\mu} \cdot \left(\dot{\mathbf{r}}_i - \sqrt{2\mu T}\boldsymbol{\xi}_i\right) \mathrm{d}t'. \qquad 14.$$

Substituting Equation 1 into Equation 14, and using the chain rule $\dot{U} = \sum_n \dot{\alpha}_n (\partial U/\partial \alpha_n) + \sum_i \dot{\mathbf{r}}_i \cdot \nabla_i U$, gives a relation among energy U, work $\mathcal{W}$, and heat $\mathcal{Q}$:

$$U(t) - U(0) = \mathcal{W} - \mathcal{Q} + \int_0^t \sum_i \dot{\mathbf{r}}_i \cdot \mathbf{f}_i \mathrm{d}t'. \qquad 15.$$

In passive systems, the nonconservative force $\mathbf{f}_i$ represents some intervention by the external operator (beyond changes in U), so that the term $\bar{\mathcal{W}} \equiv \int_0^t \sum_i \dot{\mathbf{r}}_i \cdot \mathbf{f}_i \mathrm{d}t'$ can be absorbed into the work $\mathcal{W}$. Then, Equation 15 is the first law of thermodynamics (FLT) (27, 28). The only possible time reversal for passive dynamics is to choose $\mathcal{S}^+$ in Equation 12, so that $\mathcal{S}^+ = \dot{\mathcal{Q}}/T$. In this context, $\mathcal{S}^+$ coincides with the thermodynamic EPR due to contact of the system with the thermostat, which connects explicitly irreversibility, entropy production, and dissipation (28). In contrast, for active dynamics, the contribution $\bar{\mathcal{W}}$ captures the energy cost to sustain the self-propulsion of particles (58, 59), which is generally distinct from both $\mathcal{W}$ and $\mathcal{Q}$. When the potential is static ($\dot{\alpha}_n = 0$), $\mathcal{S}^+ = \dot{\mathcal{Q}}/T$ still holds, yet $\mathcal{S}^+$ should not be interpreted as a thermodynamic EPR in general.

Although these considerations offer one consistent approach to the question of energy transfers for active particles, several other approaches are possible. Interestingly, one alternative definition of heat relies on replacing $\dot{\mathbf{r}}_i$ in Equation 14 by $\dot{\mathbf{r}}_i - \mu\mathbf{f}_i$. This amounts to regarding the self-propulsion as being caused by a locally imposed flow, with particle displacements evaluated in the flow frame (43, 48). It yields vanishing heat in the absence of interactions, for the same reasons as led us to zero $\mathcal{S}^-$ in Equation 12. Also, some works addressing heat engines (60–62) have redefined heat by replacing $\sqrt{2\mu T}\boldsymbol{\xi}_i$ in Equation 14 with $\sqrt{2\mu T}\boldsymbol{\xi}_i + \mu\mathbf{f}_i$, thus considering the self-propulsion $\mathbf{f}_i$ as a noise with similar status to $\sqrt{2\mu T}\boldsymbol{\xi}_i$. This yields vanishing heat when the potential is static ($\dot{\alpha}_n = 0$) by discarding all the energy dissipated by self-propulsion, allowing one to reinstate the FLT, $U(t) - U(0) = \mathcal{W} - \mathcal{Q}$. Finally, the angular diffusion of ABPs is often itself regarded as thermal. This gives an angular contribution to $\mathcal{Q}$, which is proportional to $\int_0^t \sum_i \dot{\theta}_i(\dot{\theta}_i - \sqrt{2/\tau}\eta_i)\mathrm{d}t'$, which vanishes for Equation 3 but is generally nonzero when aligning interactions are present (59). For AOUPs, interpreting the self-propulsion dynamics in terms of thermal damping and noise is less straightforward, although some studies have taken such a path (41, 42, 63).

Importantly, Equation 1 does not resolve how particles convert fuel into motion. Accordingly, Equation 14 only captures the energy dissipated by the propulsion itself, ignoring contributions from underlying, metabolic degrees of freedom. Various schematic models describe the underlying chemical reactions in thermodynamically consistent terms, maintaining them out of equilibrium by holding constant the chemical potential difference between products and reactants (43, 46, 64). For some of these models (43, 46), the time evolution of $\mathbf{r}_i$ can be mapped into Equation 1 when a chemical noise parameter is small (58). In this case, the difference between the partial and total heat, found by discarding or including reactions, is a constant that is independent of the potential U. In a more refined model, the dynamics tracks the time evolution of chemical concentration,

and the heat features explicitly the chemical current (64). Related features are discussed next, for continuum fields.

Linear irreversible thermodynamics (LIT): forces and currents that are linearly coupled, and the dissipated heat is given in terms of their product

2.3.2. Coarse-grained perspective: energy transfers at hydrodynamic level. For field theories, assuming that the free energy $\mathcal{F}$ depends on some set of control parameters α_n, the work $\mathcal{W}$ is defined by analogy with Equation 13 as

$$\mathcal{W} = \int_0^t \sum_n \dot{\alpha}_n \frac{\partial \mathcal{F}}{\partial \alpha_n} \mathrm{d}t'. \qquad 16.$$

This holds regardless of any nonequilibrium terms in Equations 4 and 5 from the case of passive to active fields. In contrast, to define heat at a hydrodynamic level, one cannot depend a priori on a physical interpretation of the various dynamical terms as we made for particles. A minimal requirement, for emergence of the FLT along the lines of Equation 15, is that $\mathcal{F}$ in Equation 16 is a genuine free energy, which stems from coarse-graining only the passive contributions of the microscopic dynamics. When theories are based solely on phenomenological arguments, this thermodynamic interpretation is absent.

In some cases, an active field theory ought to allow a thermodynamic interpretation, e.g., for the formation of membraneless organelles (65), but not when the same theory describes social systems, e.g., the demixing of animal groups. In the latter, not only active terms but $\mathcal{F}$ itself originate in behavioral rules, which do not produce any mechanical work, so that there is no meaningful FLT. In the former, to reinstate a thermodynamic framework, one can include chemical fields describing the fuel consumption underlying activity (57). This relies on linear irreversible thermodynamics (LIT) (66), as previously used to illuminate particle-based dynamics (64). LIT postulates linear relations between thermodynamic fluxes and forces, whose product determines the heat $\mathcal{Q}$. A class of active gel models was indeed first formulated in this way (67, 68). In contrast, the field theories in Equations 4 and 5 were not derived from LIT a priori, yet they can be embedded within it in a consistent manner.

To illustrate this, consider a scalar field ϕ obeying Equation 4. The conservation law $\dot{\phi} = -\nabla \cdot \mathbf{J}$ relates this to the thermodynamic flux $\mathbf{J}$, whose conjugate force is $-\nabla(\delta\mathcal{F}/\delta\phi)$. Likewise, a thermodynamic flux $\dot{n}$ describes some metabolic chemical process, with conjugate force a chemical potential difference $\Delta\mu$. Out-of-equilibrium dynamics is maintained by holding either $\dot{n}$ or $\Delta\mu$ constant (47). Considering here the latter case, LIT requires that the active current $\mathbf{J}_\phi$ be linear in $\Delta\mu$ (although, like $\mathbf{J}$, it can be nonlinear in ϕ), so that we identify $\mathbf{J}_\phi/\Delta\mu \equiv \mathbf{C}$ as an off-diagonal Onsager coefficient (57), yielding

$$[\mathbf{J}, \dot{n}] = \mathbb{L}\left[-\nabla\frac{\delta\mathcal{F}}{\delta\phi}, \Delta\mu \right] + \text{noise terms}, \qquad 17.$$

with an Onsager matrix obeying $\mathbb{L}_{\mathbf{JJ}} = \lambda\mathbf{I}$ (with $\mathbf{I}$ the d-dimensional identity); $\mathbb{L}_{\dot{n}\dot{n}} = \gamma$, a chemical mobility (such that $\dot{n} = \gamma\Delta\mu$ in the absence of ϕ dynamics); and $\mathbb{L}_{\mathbf{J}\dot{n}} = \mathbb{L}_{\dot{n}\mathbf{J}} = \mathbf{C}$. This last result encodes the famous Onsager symmetry, which stems from the underlying reversibility of LIT (25), so that the form of the active current $\mathbf{J}_\phi$ also controls the ϕ coupling in the equation for $\dot{n}$. For $\mathbf{J}_\phi = \Delta\mu\nabla[(\nabla\phi)^2]$ (13), then $\dot{n} = \gamma\Delta\mu - \nabla(\delta\mathcal{F}/\delta\phi)\cdot\nabla[(\nabla\phi)^2]$. The covariance of the noise terms in Equation 17 is set directly by $\mathbb{L}$. The off-diagonal noise depends on ϕ through $\mathbf{C}$ and thus is multiplicative. The noise terms accordingly include so-called spurious drift contributions, which ensure that the LIT dynamics reaches Boltzmann equilibrium when neither $\dot{n}$ nor $\Delta\mu$ is held constant (57).

A crucial assumption of Equation 17 is that ϕ and n are the only relevant hydrodynamic fields, so that, on the time and length scales at which they evolve, all other degrees of freedom are thermally equilibrated. Then, the heat $\mathcal{Q}$ is given in terms of the entropy production of these fields:

$\mathcal{Q} = D\ln(\mathcal{P}[\mathbf{J},\dot{n}]/\mathcal{P}^R[\mathbf{J},\dot{n}])$. The resulting thermodynamic EPR differs in form from the IEPR associated with the dynamics of ϕ alone (Equation 10). Therefore, it cannot be equated with any of the various IEPRs for differently time-reversed pure ϕ dynamics given in Section 2.2. However, because Equation 17 follows LIT, the heat can be expressed in terms of currents and forces:

$$\mathcal{Q} = \int \mathrm{d}\mathbf{r} \int_0^t \mathrm{d}t' \left(-\mathbf{J} \cdot \nabla \frac{\delta \mathcal{F}}{\delta \phi} + \dot{n}\Delta\mu \right). \qquad 18.$$

Similar expressions arise for more elaborate field theories, such as polar fields (57). Note that $\mathcal{Q}$ is stochastic, as is $\mathcal{W}$ in Equation 16, even though both are defined at hydrodynamic level. Using the chain rule $\dot{\mathcal{F}} = \sum_n \dot{\alpha}_n(\partial\mathcal{F}/\partial\alpha_n) + \int \dot{\phi}(\delta\mathcal{F}/\delta\phi)\mathrm{d}\mathbf{r}$ and the conservation law $\dot{\phi} = -\nabla \cdot \mathbf{J}$, it follows that free energy $\mathcal{F}$, work $\mathcal{W}$, and heat $\mathcal{Q}$ are related as

$$\mathcal{F}(t) - \mathcal{F}(0) = \mathcal{W} - \mathcal{Q} + \int \mathrm{d}\mathbf{r} \int_0^t \mathrm{d}t' \dot{n}\Delta\mu. \qquad 19.$$

The energy balance in Equation 19 offers the hydrodynamic equivalent of the relation among the particle-based work $\mathcal{W}$, heat $\mathcal{Q}$, and energy U in Equation 15. When neither $\dot{n}$ nor $\Delta\mu$ is maintained constant, and $\Delta\mu$ derives from a free energy ($\Delta\mu = -\delta\mathcal{F}_{\mathrm{ch}}/\delta n$), the system achieves equilibrium, so that Equation 19 reduces to the FLT with respect to the total free energy $\mathcal{F} + \mathcal{F}_{\mathrm{ch}}$. When the free-energy $\mathcal{F}$ does not change ($\dot{\alpha}_n = 0$), the heat rate equals $\int \mathrm{d}\mathbf{r}\,\langle \dot{n}\Delta\mu \rangle$, which illustrates that all the activity ultimately stems from the work done by chemical processes. At fixed $\Delta\mu$ (say), $\dot{n}$ depends on ϕ, so that $\int \mathrm{d}\mathbf{r}\,\langle \dot{n}\Delta\mu \rangle$ contains information similar to, but distinct from, Equation 10 (see discussion in Section 2.4).

2.4. Where and When Activity Matters

The various IEPRs introduced in Section 2 enable one to delineate regimes in which activity most affects the dynamics compared to equilibrium. Identifying such regimes can quantify the length and timescales where activity primarily matters and sometimes pinpoint specific locations where self-propulsion plays an enhanced role. Within this perspective, the most appropriate choice of IEPR is usually that which best reflects the dynamical symmetries of the emergent order (see **Figure 1**). By eliminating gross contributions (e.g., from propulsion pointing opposite to velocity), the right choice of IEPR can help identify dynamical features that break TRS more subtly and help unravel how activity can create effects with no equilibrium counterpart.

2.4.1. Spatial and spectral decompositions of irreversibility. If thermal noise is omitted from Equation 1, only one IEPR can be defined. This is $\mathcal{S}_r$ (see Equation 11), tracking positions only. For AOUPs with pairwise interaction $U = \sum_{i<j} V(\mathbf{r}_i - \mathbf{r}_j)$, one finds a particle-based decomposition $\mathcal{S}_r = \sum_i \sigma_i$ (69), where

$$\sigma_i = \frac{\tau}{4\mu f_0^2} \sum_j \left\langle \left[(\dot{\mathbf{r}}_i - \dot{\mathbf{r}}_j) \cdot \nabla_i \right]^3 V(\mathbf{r}_i - \mathbf{r}_j) \right\rangle. \qquad 20.$$

This shows that particle i contributes most to $\mathcal{S}_r$ when its neighbors j have a large relative velocity $\dot{\mathbf{r}}_i - \dot{\mathbf{r}}_j$. Thus, collisions between slow and fast particles are the main source of irreversibility. Interestingly, such collisions lie at the basis of the formation of particle clusters, which can potentially lead to nonequilibrium (or motility-induced) phase separation (15). For a phase-separated density profile, Equation 20 allows one to distinguish the relative contributions from particles in different spatial zones. In the dilute phase, collisions are rare so that σ_i stays small, whereas in a dense enough phase, particles barely move so that σ_i is again modest. At interfaces, collisions

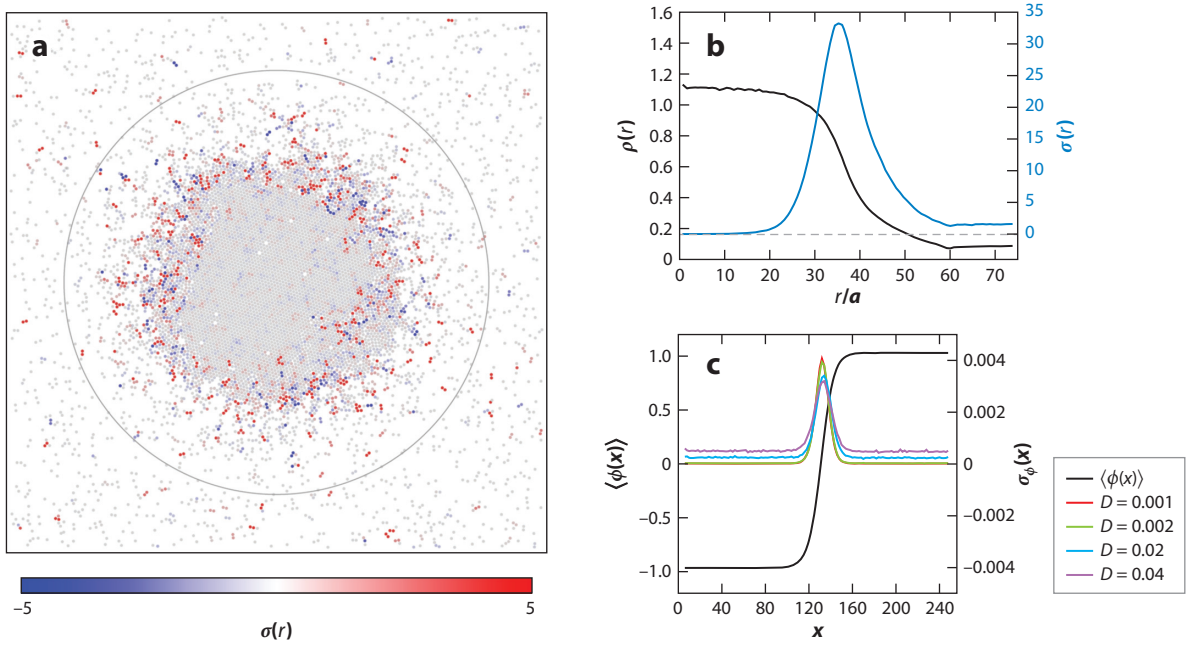

Figure 2

Spatial decomposition of the informatic entropy production rate. (*a*) Phase separation of repulsive active particles, with individual σ_i (color-coded via Equation 20) enhanced at the liquid–vapor interface. (*b*) Time-averaged profiles of particle density ρ and informatic entropy production rate density σ show that σ is small in both bulks and peaked at the interface. (*c*) Similar behavior for a conserved active scalar field ϕ. Panels *a* and *b* adapted from Reference 39; copyright 2017 American Physical Society. Panel *c* adapted with permission from Reference 69; copyright 2021 American Physical Society.

between fast and slow particles, entering respectively from dilute and dense zones, lead to locally high IEPR (see **Figure 2**).

In the presence of thermal noise, one can also study $\mathcal{S}^{\pm}$ (45). These IEPRs track the dynamics of both position and self-propulsion. The corresponding particle-based quantities, $\sigma_i^+ = (1/T)\langle \dot{\mathbf{r}}_i \cdot \mathbf{f}_i \rangle$ and $\sigma_i^- = (\mu/T)\langle \nabla_i U \cdot \mathbf{f}_i \rangle$, which take $\mathbf{f}_i$ to be even and odd, respectively, are minimal and maximal for particles at rest. This happens when $-\nabla_i U$ is equal and opposite to $\mathbf{f}_i$. In practice, σ_i^+ is low in the dense phase and high at the interface (like σ_i above), whereas σ_i^- decreases gradually from the bulk of the dense phase to the interface. Considering instead polar clusters, which emerge when aligning interactions are present, σ_i^- is now low in the dense and dilute phases, with a higher value at the interface (53). Thus, for isotropic particles it is σ_i^+ that exposes the character of the irreversibility of self-propulsion at interfaces, whereas for aligning particles σ_i^- does so. Each separately elucidates the role of interfacial self-propulsion to promote and stabilize clustering.

Turning now to field theories, we consider first the simplest scalar model of Equation 4, with ϕ a particle density that is even under time reversal. In the spirit of Landau–Ginzburg theory, the nonequilibrium forcing term $\mathbf{J}_\phi$ is generally taken as a local function of ϕ and its gradients and, hence, is also even (13, 37). The relevant measure of irreversibility reads as $\mathcal{S} = \int \sigma(\mathbf{r})\mathrm{d}\mathbf{r}$, where $\sigma = \langle \mathbf{J} \cdot \mathbf{J}_\phi \rangle / D$ can be identified as a local IEPR density (39). To lowest order, $\mathbf{J}_\phi$ contains three gradients and two fields, giving leading order (in noise) contributions where $\nabla\phi$ is large, as applies near the interface in a phase-separated profile (see **Figure 2**). This corroborates the particle-based results for active phase separation given above. (Similar principles govern various local IEPR

densities in field theories of polar active matter; 49.) In contrast to $\mathcal{S}$, the local chemical heat $\langle \dot{n}\Delta\mu\rangle$, found from Equation 18, is high throughout the bulk phases with a (bimodal) dip across the interface (57). This reflects the fact that the underlying chemical reactions fruitlessly dissipate energy throughout uniform bulk phases (where $\mathbf{C}$ vanishes, modulo noise, so they barely affect the ϕ dynamics), whereas at interfaces such reactions locally do work, against the thermodynamic force $-\nabla(\delta\mathcal{F}/\delta\phi)$, to sustain the counter-diffusive current $\mathbf{J}_\phi$.

Harada–Sasa relation: spectral decomposition of informatic entropy production across temporal frequencies and spatial modes

Another challenge is to identify the timescales on which activity dominates over thermal effects. To partially quantify this, one can define a frequency-dependent energy scale $T_{\text{eff}}(\omega) = \omega C_i(\omega)/[2R_i(\omega)]$. The autocorrelation of position reads as $C_i(\omega) = \int e^{i\omega t}\langle \mathbf{r}_i(t)\cdot\mathbf{r}_i(0)\rangle \mathrm{d}t$. The response R_i for small perturbation of the potential ($U \to U - f\hat{\mathbf{u}}\cdot\mathbf{r}_i$, where $\hat{\mathbf{u}}$ is an arbitrary unit vector) reads as $R_i(\omega) = \int \sin(\omega t)[\delta\langle \mathbf{r}_i(t)\cdot\hat{\mathbf{u}}\rangle/\delta f(0)]_{f=0}\mathrm{d}t$. In the absence of self-propulsion, T_{eff} reduces to the bath temperature T, restoring the FDT (26). The deviation $T_{\text{eff}} - T$ at small ω identifies regimes in which activity dominates, and this has been measured in living systems (70–72). Interestingly, measuring the FDT violation offers a generic route to quantify irreversibility and dissipation. In passive systems, the Harada–Sasa relation (73) explicitly connects the position autocorrelation C_i, the response function R_i, and the thermodynamic EPR, $\mathcal{S}$. For active particles, it can be generalized as (35, 74)

$$\mathcal{S}^+ = \int \sum_i \frac{\omega}{\mu T}\Big[\omega C_i(\omega) - 2TR_i(\omega)\Big]\frac{\mathrm{d}\omega}{2\pi}, \quad \mathcal{S}_r = \int \sum_i \frac{\omega}{\mu}\Big[4\mu\omega\Gamma(\omega)C_i(\omega) - 2R_i(\omega)\Big]\frac{\mathrm{d}\omega}{2\pi}. \quad 21.$$

The integrands in Equation 21 provide spectral decompositions of $\mathcal{S}^+$ and $\mathcal{S}_r$. Using models in which $\mathcal{S}^+$ is directly proportional to heat, this decomposition has been measured experimentally to provide insights into the energetics of living systems (75, 76). Notably the integrand for the coordinate-only IEPR, $\mathcal{S}_r$, is no longer directly the FDT violation. Instead the irreversibility, when evaluated from fluctuations of position only, is quantified by the violation of a modified relation between correlation and response (35). For field theories, the FDT violation can again be expressed via correlation and response functions in the Fourier domain, depending on both temporal frequencies and spatial modes. It provides a spectral decomposition of the IEPR that usefully extends the Harada–Sasa relation (39, 77). This helps identify the length and timescales primarily involved in the breakdown of TRS. When considering theories with several fields, the quantification of irreversibility typically involves FDT violations associated with each one of them (39, 47).

2.4.2. Scaling of irreversibility with dynamical parameters. For generic active dynamics, one can identify as dynamical parameters the coefficients of active terms of the dynamics. Examples include the strength of microscopic self-propulsion, $\mathbf{f}_i$, for particle-based formulations as in Equation 1, and the coefficients of nonintegrable terms in field theories, such as the leading-order contributions $\nabla[(\nabla\phi)^2]$ and $(\nabla\phi)\nabla^2\phi$ within $\mathbf{J}_\phi$ in Equation 4. Estimating how the various IEPRs depend on these activity parameters, and on temperature or density, allows one to delineate equilibrium-like regimes in which TRS is restored either exactly or asymptotically. Thus, obtaining precise scalings for IEPRs is an important step toward understanding how far active dynamics deviates from equilibrium, e.g., with a view to building a thermodynamic framework, starting with near-equilibrium cases.

Considering AOUPs in Equation 2, the persistence time is a natural parameter that controls the distance from the equilibrium limit at $\tau = 0$. This approach allows a systematic perturbative expansion in τ (at fixed T_a) for the steady state (35, 52, 69). Expanding the IEPR (here $\mathcal{S}_r$) shows the irreversibility to vanish at linear order in τ, even as the statistics differ from the Boltzmann

distribution $\sim e^{-U/T_a}$ (35). In other words, there exists a regime of small persistence in which the steady state is distinct from that of an equilibrium system at temperature T_a, yet TRS is restored asymptotically. The existence of such a regime lays the groundwork for extending standard relations of equilibrium thermodynamics.

Various works have studied the behavior of $\mathcal{S}_r$ beyond the small τ regime. They have shed light on a nonmonotonicity of $\mathcal{S}_r(\tau)$ for dense systems (78) and also for particles confined in an external potential (79). Note that, in the latter case, the dependence of $\mathcal{S}_r$ on bath temperature T is also nonmonotonic in general (52). These results suggest that departures from equilibrium can decrease with persistence and increase with thermal fluctuations. However, note that the IEPR per particle has units of inverse time (see Equation 10), so that a saturating entropy production per persistence time gives decreasing $\mathcal{S}_r(\tau)$. Turning to repulsive particles, the IEPRs $\mathcal{S}^+$ and $\mathcal{S}^-$, respectively, decrease and increase with τ, up to the onset of MIPS (53). This is consistent with their being, respectively, low and high in clustered regions, and this trend is unaffected by the rescaling $\mathcal{S}^\pm \to \mathcal{S}^\pm\tau$ (80).

Another parameter controlling nonequilibrium effects is the density ρ. In a homogeneous state, because $\mathcal{S}^-$ increases as the dynamics slows down, $\mathcal{S}^-$ increases (decreases) with ρ for isotropic (aligning) particles. In practice, some detailed scalings can be obtained for isotopic pairwise interactions of the form $U = \sum_{i<j} V(\mathbf{r}_i - \mathbf{r}_j)$. For weak interactions, namely when the amplitude of V is small compared to those of thermal noise and self-propulsion, $\mathcal{S}^-$ becomes linear in ρ (80). In general, it can be written in terms of density correlations, as follows:

$$\frac{T\mathcal{S}^-}{\mu} = \rho \int \left[(\nabla V)^2 - T\nabla^2 V\right] g_2(\mathbf{r}) \mathrm{d}\mathbf{r} + \rho^2 \iint (\nabla V(\mathbf{r})) \cdot (\nabla V(\mathbf{r}')) g_3(\mathbf{r}, \mathbf{r}') \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}', \qquad 22.$$

where $g_2 = \sum_{i\neq j} \langle \delta(\mathbf{r} - \mathbf{r}_i + \mathbf{r}_j)\rangle / (N\rho)$ and $g_3 = \sum_{i\neq j\neq k} \langle \delta(\mathbf{r} - \mathbf{r}_i + \mathbf{r}_j)\delta(\mathbf{r}' - \mathbf{r}_i + \mathbf{r}_k)\rangle / (N\rho)^2$ are two-point and three-point density correlators (81). The integrand in Equation 22 can be regarded as an integral form of the Yvon–Born–Green relation, which constrains g_2 and g_3 for passive particles (82), so that the violation of this relation provides access to $\mathcal{S}^-$ for active ones. A related form of Equation 22 has been proposed in terms of $g_2 - g_{2,\mathrm{eq}}$, where $g_{2,\mathrm{eq}}$ is the two-point correlator for passive particles with the same potential U (81, 83).

In field theories, IEPRs are generically either linear or quadratic in the activity parameters (for small parameters) depending on whether they break symmetries of the passive theory (39). The IEPR scalings with D are mechanistically revealing for both scalar (39) and polar (49) models. As previously noted, the spatial dependence of IEPR density is informative: Independent scalings can be seen in different bulk phases and at their interfaces. In general, one finds the following leading order behavior. (*a*) $\sigma \sim D^{-1}$: the deterministic part of the dynamics already breaks TRS, and very unlikely noise realizations are needed to recreate, for the reversed paths, the opposite of a deterministic forward motion. (*b*) $\sigma \sim D^0$: TRS is unbroken deterministically but violated at lowest order in fluctuations. This may require some spatial symmetry breaking at the deterministic level, so that the background fields act as a ratchet or rectifier for the fluctuations. (*c*) TRS is broken only at higher order, which gives $\sigma \sim D^1$ in all cases so far studied, although higher powers are not ruled out.

For scalar fields exhibiting bulk phase separation, the deterministic dynamics gives a stationary mean-field profile, so that there is no D^{-1} contribution. TRS is broken at leading order in fluctuations because of the interface between phases (**Figure 2**). In uniform bulk phases, irreversibility emerges only at next order, $\sigma \sim D^1$, suggesting that it arises from interactions between fluctuations. Note that active scalar models can also predict entirely new phases, whose σ scalings remain subject to investigation (37). The behavior of the thermodynamic EPR, $\langle \dot{n}\Delta\mu\rangle(\mathbf{r})/D$, found by embedding the same model within LIT and including chemical processes, is quite different (57). This

scales as D^{-1} but with a reduced amplitude in interfacial regions (see Section 2.4.1). Furthermore, for one scalar model with both conserved and nonconserved dynamics, TRS violations are gross (D^{-1}) if these two contributions to $\dot{\phi}$ are separately monitored, but perfect reversibility holds if not (84).

Stealth entropy production: entropy production that shows a nontrivial exponent even when active and passive models may share a universality class

For dynamics with a scalar density field (even under time reversal) and a polarization (even or odd) (47, 49), the IEPR scalings depend on whether the density current $\mathbf{J}$ is retained or ignored (see Equations 4 and 5, and Section 2.2); we assume the latter here. There remain two IEPRs, $\mathcal{S}^{\pm}$, according to whether $\mathbf{p}$ is even (+) or odd (−). In the phase comprising traveling bands or clusters (polarized high-density regions that propagate along $\mathbf{p}$), both of $\mathcal{S}^{\pm}$ scale as D^{-1}, because such deterministic dynamics breaks TRS. Furthermore, because the density wave has a steep front and a shallow back, flipping $\mathbf{p}$ does not restore deterministic-level TRS (**Figure 1**). In the phase of uniform $\langle \mathbf{p} \rangle$, one finds instead $\mathcal{S}^{\pm} \sim D^{0}$; similar arguments might apply now to fluctuations instead of deterministic motion.

As emphasized already, IEPRs give varying information depending on which degrees of freedom are retained and ignored. In particular, the universal properties of the critical point for active phase separation (including MIPS) can be studied by progressive elimination of degrees of freedom via the renormalization group, which could potentially be expected to give rise to emergent reversibility. Remarkably, Reference 85 has established that (*a*) the active critical point is in the same universality class as passive phase separation with TRS, but (*b*) there is no emergent reversibility, in the sense that the IEPR per space–time correlation volume does not scale toward zero at criticality (it remains constant above four dimensions, and it can even diverge below). This scenario, referred to as stealth entropy production, can be rationalized by arguing that the scaling of the IEPR has a nontrivial critical exponent in the universality class shared by active and passive systems. Because no coarse-graining can make a passive system become active, the IEPR has zero amplitude in the passive cases previously thought to define the universality class.

3. BIASED ENSEMBLES OF TRAJECTORIES

3.1. Dynamical Bias and Optimal Control

Biased ensembles generalize the canonical ensemble of equilibrium statistical mechanics from microstates to trajectories (29, 31, 86). They have proven useful in simple models of interacting particles (87–90) as well as glassy systems (32, 91) and beyond (92–94). They are constructed by biasing the value of a physical observable, which we denote here by $\mathcal{B}$. The choice of this observable depends on the physical system of interest, e.g., the heat dissipated during a dynamical trajectory or the displacement of a tagged particle.

Let X denote a dynamical trajectory of an active system, over a time period $[0, t]$. Following Section 2.2, the probability of this trajectory can be represented as $\mathcal{P}[X] = p_0[X]\mathrm{e}^{-\mathcal{A}[X]}$, where $\mathcal{A}$ is the action and p_0 is the probability of the initial condition. A biased ensemble is defined by a probability distribution over these trajectories:

$$\mathcal{P}_s[X] = \frac{1}{\mathcal{Z}(s,t)} p_0[X] \exp\left(-\mathcal{A}[X] - s\mathcal{B}[X]\right), \qquad 23.$$

where s is the biasing field and $\mathcal{Z}(s,t) = \langle \mathrm{e}^{-s\mathcal{B}} \rangle$ for normalization. Comparing with the canonical ensemble, one may identify $\mathcal{B}$ and s as an (extensive) physical observable and its conjugate (intensive) field, respectively. The field s, according to its sign, biases the distribution toward higher or lower $\mathcal{B}$. Importantly, the bias has no prejudice about the dynamical mechanism by which $\mathcal{B}$ changes: This allows the system to access fluctuation mechanisms for this quantity that might not

have been anticipated a priori (91, 93, 94). Consistent with this observation, Equation 23 can also be derived by a maximum entropy computation for trajectories with nontypical $\mathcal{B}$ (95, 96).

Optimal control: some control forces that can be added to the system to mimic biased ensembles as closely as possible

The next step is to consider very large times t, which is analogous to the thermodynamic limit of large system size in equilibrium statistical mechanics. Because $\mathcal{Z}(s)$ is analogous to a partition function, one may define a dynamical free energy density as

$$\psi(s) = \lim_{t\to\infty} \frac{1}{t} \log \mathcal{Z}(s,t). \qquad 24.$$

This construction is natural if the observable $\mathcal{B}$ scales extensively with the time t, which is assumed below; ψ is then a scaled cumulant generating function for $\mathcal{B}$ (33). In contrast to the standard ensembles of equilibrium statistical mechanics, Equation 23 does not describe practical experimental systems. However, there are several theoretical contexts in which such distributions are relevant. Examples include fluctuation theorems, in which $\mathcal{B}$ is a measure of irreversibility, whose probability distributions have symmetries related to TRS, yielding $\psi(s) = \psi(1-s)$ (28, 29, 40). Biased ensembles are also deeply connected with large deviation theory (97), applied to time-averaged quantities (33, 34), which describes rare events in which $\mathcal{B}/t$ differs significantly from its typical value. Viewing $\psi(s)$ as a dynamical free energy, its singularities can be interpreted as dynamical phase transitions. In most cases, such transitions require a limit where both t and the system size go to infinity.

Given that biased ensembles do not mimic typical experimental situations, a natural question is whether the response to the bias s can be connected with its response to some physical perturbation. Such a connection is provided by optimal control theory (98), which is also intrinsically related to large deviation theory (89, 99, 100). In fact, biased ensembles show how $\mathcal{B}$ can be reduced or increased, and the corresponding mechanisms are as close as possible to the original dynamics. We now discuss a method for identifying a physical system that mimics a given biased ensemble (34, 99–101). To reduce the level of technicality, we exclude transient behavior that happens for times close to the beginning or end of a trajectory. Biased ensembles can then be accurately reproduced by the optimally controlled system (OCS), which is also called auxiliary process (102) or driven process (101). To find the OCS, consider a wider class of controlled systems, obtained by modifying the equations of motion of the original active model. For any controlled system, the Kullback–Leibler (KL) divergence measures how its path probability $\mathcal{P}^{\rm con}[X] = p_0[X]\mathrm{e}^{-\mathcal{A}^{\rm con}[X]}$ differs from $\mathcal{P}_s$:

$$D_{\rm KL}\left(\mathcal{P}^{\rm con}||\mathcal{P}_s\right) = \int \mathcal{P}^{\rm con}[X] \log \frac{\mathcal{P}^{\rm con}[X]}{\mathcal{P}_s[X]} \mathcal{D}X, \qquad 25.$$

where $\mathcal{D}X$ denotes a path integral. The OCS has $\lim_{t\to\infty} t^{-1} D_{\rm KL}\left(\mathcal{P}^{\rm con}||\mathcal{P}_s\right) = 0$, so its differences from $\mathcal{P}_s$ are small. From Equations 23–25, one can deduce (34)

$$\lim_{t\to\infty} \frac{1}{t} \left\langle \mathcal{A}^{\rm con}[X] - \mathcal{A}[X] - s\mathcal{B}[X] \right\rangle_{\rm con} \le \psi(s), \qquad 26.$$

where the average $\langle\cdot\rangle_{\rm con}$ is computed for the controlled system. Equation 26 provides a lower bound on ψ, which becomes an equality for the OCS. Among all the processes with any given $\langle\mathcal{B}\rangle_{\rm con}$, the OCS minimizes $\langle\mathcal{A} - \mathcal{A}^{\rm con}\rangle_{\rm con}$. Hence, it provides a mechanism for achieving an atypical value of $\mathcal{B}$ while remaining as close as possible to the original dynamics.

Knowing $\mathcal{A}$, $\mathcal{A}^{\rm con}$, and $\mathcal{B}$, it suffices to maximize the bound in Equation 26 over the controlled system to obtain the OCS. The details of this computation depend on the unbiased dynamics of interest, as discussed extensively in References 34, 99, 100, and 101. In practice, the OCS usually differs from the original system in two ways: (*a*) The forces in the original model get modified

by terms that depend in simple and explicit ways on s and $\mathcal{B}$, and (b) other control forces are added as gradients of an optimal control potential $U^{\rm opt}$, which may be computed by solving either an eigenvalue problem (following Doob, as discussed in Reference 101) or a nonlinear partial differential equation (98). We emphasize that the optimal control forces do not generically appear as external forces in the dynamics; they can correspond to complicated interactions that may be long-ranged. Although exact results for $U^{\rm opt}$ are rare (103–105), Equation 26 is still useful. For example, using a controlled system with small KL divergence dramatically speeds up the numerical sampling of biased ensembles (106–109). Also, minimization of the KL divergence over simple classes of a controlled system provides useful insight into relevant fluctuation mechanisms, even without an exact solution.

3.2. Phase Transitions and Symmetry Breaking

We now examine phase transitions in biased ensembles of active particles. Interestingly, they can lead to the emergence of symmetry breaking without any equivalent in the unbiased dynamics. Furthermore, biased ensembles also open the door to proposing innovative strategies for designing active particles to achieve some target collective states.

3.2.1. Biased ensembles for active Brownian particles. To illustrate the behavior of biased ensembles, we discuss a guiding example for a system of ABPs (94, 105). Section 3.2.2 below presents biased ensembles of active systems in a wider context. **Figure 3** displays the behavior of biased ensembles for a two-dimensional system of ABPs. Trajectories are biased according to their active work; that is, $\mathcal{B} = \bar{\mathcal{W}}$ in Equation 23, with

$$\bar{\mathcal{W}}[X] = \sum_i \bar{\mathcal{W}}_i[X], \qquad \bar{\mathcal{W}}_i[X] = \frac{1}{w_0}\int_0^t \dot{\mathbf{r}}_i \cdot \mathbf{f}_i \mathrm{d}t', \tag{27.}$$

where the normalization w_0 is chosen so that $\langle \bar{\mathcal{W}}_i \rangle = t$ for noninteracting (or sufficiently dilute) particles. Physically, $\bar{\mathcal{W}}$ measures how effectively the particles' propulsive forces are converted into motion: Freely swimming particles have $(\bar{\mathcal{W}}_i/t) \approx 1$, whereas those in crowded environments are impeded by their neighbors, resulting in $(\bar{\mathcal{W}}_i/t) \approx 0$. Negative $\bar{\mathcal{W}}_i$ means that a particle's motion is opposite to its self-propulsive force, which occurs rarely. Because $\bar{\mathcal{W}}$ is closely related to the IEPR (Section 2.2), it follows a fluctuation theorem (28): The dynamical free energy of the biased ensemble obeys $\psi(w_0 - s) = \psi(s)$.

Figure 3 shows how the biasing field modifies the behavior of this system. The following discussion applies for $s < w_0/2$; the behavior for larger s can then be obtained via the fluctuation theorem. For $s > 0$, the system enters a state that is phase-separated and dynamically arrested (denoted PSA). Consistent with the arguments of Section 3.1, this happens because phase separation is the most natural (or least unlikely) mechanism for particles to collectively reduce their active work, via local crowding effects. Alternative mechanisms do exist for reduced active work; for example, dilute ABPs would have $\bar{\mathcal{W}} \approx 0$ during rare fluctuations in which the noise term $\sqrt{2\mu T}\boldsymbol{\xi}_i$ in Equation 1 is opposite to the propulsion force $\mathbf{f}_i$. Collectively, though, these trajectories have much larger action $\mathcal{A}$ than does the PSA state; and for a given value of $\bar{\mathcal{W}}$, the biased ensemble is dominated by the trajectories of least action. Hence, one observes the PSA state in the biased ensemble.

This behavior is closely related to that of biased ensembles for equilibrium systems, which can be analyzed at field-theoretic (or fluctuating hydrodynamic) level using macroscopic fluctuation theory (MFT) (110). The particle density is a locally conserved field, so large-scale density fluctuations relax on a timescale proportional to L^2 (where L is the system size). For large systems, this slow timescale decouples from particles' rapid microscopic motion, leading to a theory for the

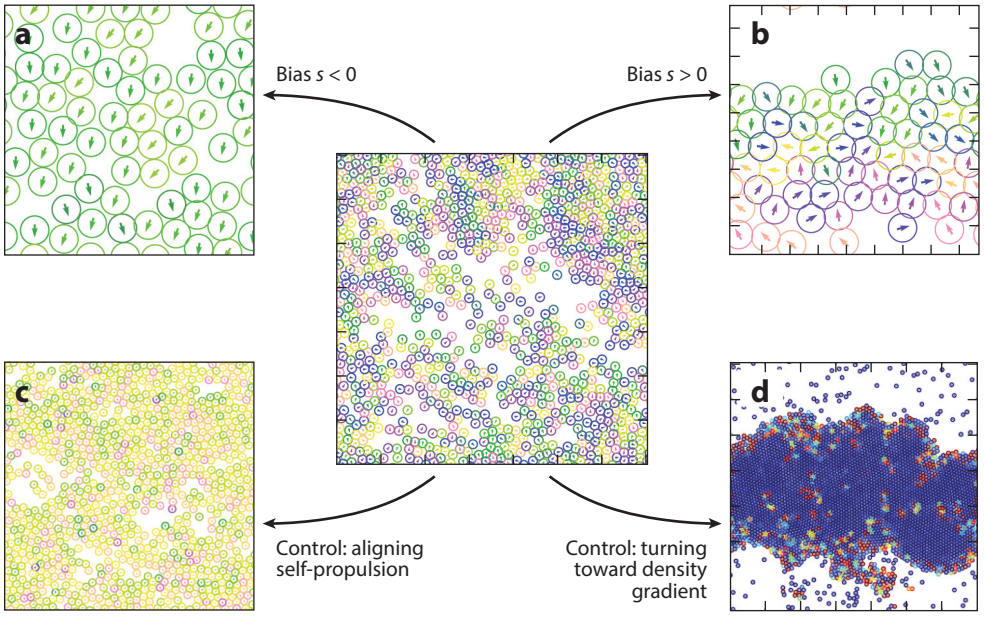

Figure 3

Behavior in biased ensembles of active Brownian particles, and comparison with controlled systems. The central panel shows a snapshot of the unbiased dynamics ($\tilde{\mathcal{W}} \simeq \langle \tilde{\mathcal{W}} \rangle$). Panels *a* and *b* show similar snapshots of collective motion and phase-separated arrested states (respectively, biased ensembles with $s < 0$ and $s > 0$) associated with atypical $\tilde{\mathcal{W}}$. Particles are colored according to their orientations. Panels *c* and *d* show how biased ensembles can be mimicked by control forces (or torques). In panel *c*, an infinite-ranged interaction favors orientational alignment; particles with similar colors are aligned with each other, as in panel *a*. In panel *d*, control torques tend to align orientations along density gradients. Particles are colored according to their control torque strength, which tends to be largest for boundary particles, facilitating phase separation. Panels *a*, *b*, and *d* adapted with permission from Reference 94; copyright 2019 American Physical Society. Panel *c* adapted from Reference 105; copyright 2021 American Physical Society.

density alone. Construction of this theory involves scaling the space–time coordinates, so that the biasing field s enters via a scaling variable $\lambda = sL^2$. For equilibrium steady states, MFT predicts dynamical phase transitions at finite λ, corresponding to a bias s of order L^{-2} (88). This argument applies also in the active case: The result is that sufficiently large systems should phase separate as soon as s is positive.

To understand this result physically, recall the connection of biased ensembles to optimal control theory (Equation 26). As well as biased systems, **Figure 3** shows those to which control forces have been applied. To mimic PSA states, a suitable control strategy is to create a macroscopic cluster, whose boundary particles have their self-propulsion forces pointing inward (similar to MIPS states). This can be achieved by applying torques to the ABP orientations, so that they point along the local density gradient, stabilizing the phase-separated cluster and increasing its density toward dynamical arrest (94). Such control forces (or torques) act mostly on the boundary particles, so that the difference between the action terms in Equation 26 scales as L (i.e., subextensive in the volume L^2 of the system). This is consistent with the irreversibility measure $(1/T)\langle \dot{\mathbf{r}}_i \cdot \mathbf{f}_i \rangle$ being high at the interface for a phase-separated profile in the unbiased dynamics (Section 2.4). Because the subextensive control forces lead to an extensive change in $\tilde{\mathcal{W}}$, one can confirm that a PSA state

is expected for all $s > 0$ (94). Note that control using torques is efficient for ABPs, because it harnesses the inherent self-propulsion to stabilize the cluster. Alternatively, applying control forces to particle positions, as in Reference 111, would result in a larger (though still subextensive) action.

Returning to **Figure 3**, one sees for $s < 0$ that the particles align their orientations to create a state of collective motion. Because ABPs move with fixed speed, this alignment suppresses particle collisions, which is a natural way to increase the active work $\bar{\mathcal{W}}$. Spontaneous breaking of rotational symmetry may be surprising, because aligning interactions are completely absent from the ABP model. However, numerical results (94, 105) show that this biased ensemble can be described rather accurately by a controlled model with mean-field (infinite-ranged) interactions among the ABP orientations.

It is clear that any field-level description of collective motion requires a polarization field in order to capture such alignment. Hence, collective motion cannot be described by a fluctuating hydrodynamic theory for the density alone, contrary to the PSA state. Also, the collective motion phase is found for biasing fields beyond a threshold of $s^* \simeq -1/\tau$, rather than at infinitesimal s as for PSA. These observations are related: In the PSA state, the response to the bias takes place by a slow (hydrodynamic) field, for which the natural scale is $s \sim L^{-2}$; for the collective motion state, the response occurs by a fast field, which requires s of order unity. In fact, responses to small biasing fields s are generically dominated by slow hydrodynamic modes (34). If the quantity $\mathcal{B}$ couples to these, it responds strongly at small s (87, 88). By contrast, for s of order unity, an accurate description of the large-scale, bias-induced atypical dynamics may require analysis of fields (like polarization in this example) that are negligible for hydrodynamic descriptions of typical trajectories.

We now consider a coarse-grained (hydrodynamic) theory that captures the collective motion in biased ensembles with low active work. This reinforces the correspondence between particle models and field-theoretic descriptions. As noted above, describing the onset of collective motion at $s \sim s^*$ requires consideration of a polarization field $\mathbf{p}$ alongside the density ϕ, as in Equations 4 and 5:

$$\mathbf{J} = v(\phi)\mathbf{p} - D_c(\phi)\nabla\phi + \sqrt{2M(\phi)}\boldsymbol{\Lambda}_\phi, \qquad \dot{\mathbf{p}} = \mathbf{F}(\phi, \mathbf{p}) + \sqrt{2D}\boldsymbol{\Lambda}_p, \qquad 28.$$

where v is self-propulsion, D_c is a (collective) diffusion constant, M is the mobility, and $\mathbf{F}$ is an effective thermodynamic force (105). To analyze biased ensembles, we need the dependence of the active work on ϕ, $\mathbf{p}$. At hydrodynamic level, it is consistent to approximate this as an integral of a local work rate w as $\bar{\mathcal{W}} \approx \int_0^t \int w(\phi, \mathbf{p})\, \mathrm{d}\mathbf{r}\, \mathrm{d}t'$. Here, w is averaged over a mesoscopic region and a time interval where density and polarization have prescribed values.

To understand collective motion, the key point is that the active work tends to be reduced by alignment (because collisions are reduced); hence, for small polarization $\mathbf{p}$: $w(\phi, \mathbf{p}) \simeq w(\phi, \mathbf{0}) - c_w(\phi)|\mathbf{p}|^2$, where $c_w > 0$ is the coupling between active work and polarization. Construction of the coarse-grained action for this minimal theory confirms that $\mathbf{p}$ is a fast field, whereas ϕ is slow (hydrodynamic). It is consistent to assume that $\mathbf{F}(\phi, \mathbf{p}) \simeq -c_F(\phi)\mathbf{p}$, where $c_F > 0$ is the strength of the restoring force toward $\mathbf{p} = \mathbf{0}$. Hence, for trajectories X with homogeneous density and small (constant) polarization $\mathbf{p}$, one has $\mathcal{P}_s[X] \sim \mathrm{e}^{-\mathcal{A}_{\mathrm{pol}}[X]}$ (see Equation 23), where $\mathcal{A}_{\mathrm{pol}} \propto \left(\frac{c_F^2}{4} + s c_w\right)|\mathbf{p}|^2 + O(|\mathbf{p}|^4)$. Minimizing the Landau-like action $\mathcal{A}_{\mathrm{pol}}$ reveals that spontaneous symmetry breaking occurs for $s < s_c$, where $s_c = -c_F^2/(4c_w) < 0$. This connects to particle-level results via the scaling estimates $c_w \sim 1$ and $\gamma c_F^2 \sim \tau^{-1}$, yielding symmetry breaking at $s_c \sim -1/\tau$ (105).

3.2.2. Generic features of biased ensembles in active systems. We now discuss the insights from the previous example in a broader context. First, note that biasing the active work can create

both orientational order and density modulations; this stems from the couplings between fluctuations of the active work and those of local alignment and density. The first of these can already be deduced via exact biased-ensemble results in a system with just two active particles (104, 105): Pairs with parallel alignment tend to have fewer collisions, so these states are promoted when biasing to higher active work (or propulsion efficiency). Broader insight stems from the optimal control interpretation of biased ensembles (Section 3.1). For ABPs biased by the active work, the collective motion at large $\bar{\mathcal{W}}$ shows that local alignment is an effective route toward efficient swimming (94, 104), suggesting a design strategy to achieve active particles with desired properties.

Similar arguments explain the coupling of active work to local density and, hence, to phase separation, because crowding reduces the particles' swimming speed, as found by fluctuating hydrodynamic arguments (94, 105, 112), perturbative calculations (80, 81), and variational computations (111, 112). Many other active systems couple alignment, density, and work by similar mechanisms; so this behavior under bias should be generic. Indeed, the results of References 80, 81, 94, 104, 105, and 111 cover a range of different active systems, including lattice models as well as AOUPs and ABPs. Dynamical arrest in biased ensembles also has a counterpart in passive glassy systems (32, 34, 91), and can be explained by general principles that couple biasing fields to metastable states (34).

All these results concern systems in which the observable $\mathcal{B}$ is defined as a sum over all particles. Instead, $\mathcal{B}$ may refer to a tagged particle within a large system (113–115). Perhaps surprisingly, results for equilibrium systems show that applying such bias to a single particle can generate a macroscopic response (116). This is because of coupling of the biased quantity to the slow (hydrodynamic) density fluctuations, as described by MFT. In active systems, the results of References 113 and 115 indicate a rich behavior for large deviations of the single-particle active work, which appears related to a coupling with collective (long-ranged) density fluctuations. However, fluctuations of the single-particle current were explained in Reference 114 by a local argument based on a density-dependent velocity.

4. CONCLUSION

The tools of stochastic thermodynamics, first introduced for thermodynamically consistent models (27, 28, 40), provide fruitful insights when analyzing the consequences of irreversibility in active dynamics. The framework has to be carefully adapted, because various thermodynamic relations do not carry over a priori. For instance, the usual connections among irreversibility, entropy production, and dissipation of energy need to be revisited (Sections 2.2 and 2.3). The FLT no longer holds in its standard form, which is a fact that strongly influences the behavior of engines using an active working substance (62). However, the IEPR can be defined without reference to the first law directly in terms of forward and backward path probabilities. This makes it a useful measure of irreversibility in active systems. In contrast with equilibrium, care is now needed in choosing which coarse-grained quantities change sign on time reversal: Different choices yield different information on how TRS is broken.

Importantly, measuring irreversibility via IEPR can provide a systematic approach for evaluating the distance of active dynamics from an equilibrium reference. This allows one to pinpoint regimes (such as specific bulk phases, or interfaces between these) where activity plays a major role, and rationalize the mechanisms whereby self-propulsion at the particle scale stabilizes collective effects with no equilibrium equivalent (Section 2.4). As reviewed above, this approach has been validated in several settings, involving either particle-based or field-level dynamics for scalar and polar materials, but much remains to be explored beyond these examples. In particular, it would be interesting to consider active nematics (14).

Recent progress in stochastic thermodynamics has opened new avenues whose implications for active matter remain largely unexplored. In particular, thermodynamic uncertainty relations (TURs) place bounds on irreversibility in the presence of currents (117). For active systems, it is not yet known how different types of emergent order affect the tightness of these bounds. Understanding this could allow delineation of regimes in which TURs are most accurate, with direct relevance for experimental studies of living systems.

Finally, biased ensembles (Section 3) offer promise for the engineering of active materials (81, 94, 104). Within this perspective, one predefines a desired macroscopic property in terms of a specific biasing observable. For a given starting material, the sampling of biased trajectories determines a strategy to tune microscopic interactions and achieve the target property. This approach is potentially also relevant when developing strategies to control the collective behavior of social and living systems, such as pedestrian crowds or cars on a freeway. Further use of biased ensembles in active matter models could therefore help establish ground rules for their wider application in systems that are far from equilibrium.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors acknowledge insightful discussions with Øyvind L. Borthne, Fernando Caballero, Timothy Ekeh, Yann-Edwin Keta, Yuting Irene Li, Tomer Markovich, Cesare Nardini, Takahiro Nemoto, Patrick Pietzonka, Sriram Ramaswamy, Udo Seifert, Thomas Speck, Julien Tailleur, Elsen Tjhung, Laura Tociu, Suriyanarayanan Vaikuntanathan, and Frédéric van Wijland. É.F. acknowledges support from an ATTRACT Grant of the Luxembourg National Research Fund. M.E.C. is funded by the Royal Society. This work was funded in part by the National Science Foundation, Grant No. NSF PHY-1748958, and the European Research Council under the Horizon 2020 Programme, Grant No. 740269.

LITERATURE CITED

1. Marchetti MC, Joanny JF, Ramaswamy S, Liverpool TB, Prost J, et al. 2013. *Rev. Mod. Phys.* 85:1143–89
2. Bechinger C, Di Leonardo R, Löwen H, Reichhardt C, Volpe G, Volpe G. 2016. *Rev. Mod. Phys.* 88:045006
3. Fodor É, Marchetti MC. 2018. *Physica A* 504:106–20
4. Elgeti J, Winkler RG, Gompper G. 2015. *Rep. Prog. Phys.* 78(5):056601
5. Saw TB, Doostmohammadi A, Nier V, Kocgozlu L, Thampi S, et al. 2017. *Nature* 544:212–16
6. Cavagna A, Giardina I. 2014. *Annu. Rev. Condens. Matter Phys.* 5:183–207
7. Bain N, Bartolo D. 2019. *Science* 363(6422):46–49
8. Deseigne J, Dauchot O, Chaté H. 2010. *Phys. Rev. Lett.* 105:098001
9. Palacci J, Sacanna S, Steinberg AP, Pine DJ, Chaikin PM. 2013. *Science* 339(6122):936–40
10. Vicsek T, Czirók A, Ben-Jacob E, Cohen I, Shochet O. 1995. *Phys. Rev. Lett.* 75:1226–29
11. Fily Y, Marchetti MC. 2012. *Phys. Rev. Lett.* 108:235702
12. Toner J, Tu Y. 1995. *Phys. Rev. Lett.* 75:4326–29
13. Wittkowski R, Tiribocchi A, Stenhammar J, Allen RJ, Marenduzzo D, Cates ME. 2014. *Nat. Commun.* 5:4351
14. Chaté H. 2020. *Annu. Rev. Condens. Matter Phys.* 11:189–212

15. Cates ME, Tailleur J. 2015. *Annu. Rev. Condens. Matter Phys.* 6:219–44
16. Tailleur J, Cates ME. 2008. *Phys. Rev. Lett.* 100:218103
17. Maggi C, Marconi UMB, Gnan N, Di Leonardo R. 2015. *Sci. Rep.* 5:10742
18. Yang X, Manning ML, Marchetti MC. 2014. *Soft Matter* 10:6477–84
19. Takatori SC, Yan W, Brady JF. 2014. *Phys. Rev. Lett.* 113:028103
20. Solon AP, Fily Y, Baskaran A, Cates ME, Kafri Y, et al. 2015. *Nat. Phys.* 11(8):673–78
21. Bialké J, Siebert JT, Löwen H, Speck T. 2015. *Phys. Rev. Lett.* 115:098301
22. Zakine R, Zhao Y, Knežević M, Daerr A, Kafri Y, et al. 2020. *Phys. Rev. Lett.* 124:248003
23. Paliwal S, Rodenburg J, van Roij R, Dijkstra M. 2018. *New J. Phys.* 20:015003
24. Guioth J, Bertin E. 2019. *J. Chem. Phys.* 150:094108
25. Onsager L. 1931. *Phys. Rev.* 37:405–26
26. Kubo R. 1966. *Rep. Prog. Phys.* 29(1):255–84
27. Sekimoto K. 1998. *Prog. Theor. Phys. Suppl.* 130:17–27
28. Seifert U. 2012. *Rep. Prog. Phys.* 75(12):126001
29. Maes C. 1999. *J. Stat. Phys.* 95(1):367–92
30. Derrida B. 2007. *J. Stat. Mech.* 2007(07):P07023
31. Lecomte V, Appert-Rolland C, van Wijland F. 2007. *J. Stat. Phys.* 127(1):51–106
32. Garrahan JP, Jack RL, Lecomte V, Pitard E, van Duijvendijk K, van Wijland F. 2007. *Phys. Rev. Lett.* 98:195702
33. Touchette H. 2009. *Phys. Rep.* 478(1):1–69
34. Jack RL. 2020. *Eur. Phys. J. B* 93:74
35. Fodor É, Nardini C, Cates ME, Tailleur J, Visco P, et al. 2016. *Phys. Rev. Lett.* 117:038103
36. Dean DS. 1996. *J. Phys. A Math. Gen.* 29(24):L613
37. Tjhung E, Nardini C, Cates ME. 2018. *Phys. Rev. X* 8:031080
38. Tiribocchi A, Wittkowski R, Marenduzzo D, Cates ME. 2015. *Phys. Rev. Lett.* 115:188302
39. Nardini C, Fodor É, Tjhung E, van Wijland F, Tailleur J, et al. 2017. *Phys. Rev. X* 7:021007. **https://doi.org/10.1103/PhysRevX.7.021007**
40. Lebowitz JL, Spohn H. 1999. *J. Stat. Phys.* 95(1):333–65
41. Mandal D, Klymko K, DeWeese MR. 2017. *Phys. Rev. Lett.* 119:258001
42. Puglisi A, Marconi UMB. 2017. *Entropy* 19(7):356
43. Speck T. 2018. *Europhys. Lett.* 123(2):20007
44. Caprini L, Marconi UMB, Puglisi A, Vulpiani A. 2018. *Phys. Rev. Lett.* 121:139801
45. Shankar S, Marchetti MC. 2018. *Phys. Rev. E* 98:020604
46. Pietzonka P, Seifert U. 2018. *J. Phys. A Math. Theor.* 51(1):01LT01
47. Dadhichi LP, Maitra A, Ramaswamy S. 2018. *J. Stat. Mech.* 2018(12):123201
48. Dabelow L, Bo S, Eichhorn R. 2019. *Phys. Rev. X* 9:021009
49. Borthne ØL, Fodor É, Cates ME. 2020. *New J. Phys.* 22(12):123012
50. Onsager L, Machlup S. 1953. *Phys. Rev.* 91:1505–12
51. Caprini L, Marconi UMB, Puglisi A, Vulpiani A. 2019. *J. Stat. Mech.* 2019(5):053203
52. Martin D, de Pirey TA. 2021. *J. Stat. Mech.* 2021(4):043205
53. Crosato E, Prokopenko M, Spinney RE. 2019. *Phys. Rev. E* 100:042613
54. Speck T. 2016. *Europhys. Lett.* 114(3):30006
55. Solon AP, Tailleur J. 2013. *Phys. Rev. Lett.* 111:078101
56. Alert R, Joanny JF, Casademunt J. 2020. *Nat. Phys.* 16(6):682–88
57. Markovich T, Fodor É, Tjhung E, Cates ME. 2021. *Phys. Rev. X* 11:021057
58. Pietzonka P, Fodor É, Lohrmann C, Cates ME, Seifert U. 2019. *Phys. Rev. X* 9:041032
59. Ekeh T, Cates ME, Fodor É. 2020. *Phys. Rev. E* 102:010101
60. Zakine R, Solon A, Gingrich T, van Wijland F. 2017. *Entropy* 19(5):193
61. Holubec V, Steffenoni S, Falasco G, Kroy K. 2020. *Phys. Rev. Res.* 2:043262
62. Fodor É, Cates ME. 2021. *Europhys. Lett.* 134(1):10003
63. Loos SAM, Klapp SHL. 2020. *New J. Phys.* 22(12):123051
64. Gaspard P, Kapral R. 2018. *J. Chem. Phys.* 148(13):134104

65. Weber CA, Zwicker D, Jülicher F, Lee CF. 2019. *Rep. Prog. Phys.* 82(6):064601
66. Groot SRD, Mazur P. 1962. *Non-Equilibrium Thermodynamics*. Amsterdam: North-Holland
67. Kruse K, Joanny JF, Jülicher F, Prost J, Sekimoto K. 2004. *Phys. Rev. Lett.* 92:078101
68. Prost J, Jülicher F, Joanny JF. 2015. *Nat. Phys.* 11:111–17
69. Martin D, O'Byrne J, Cates ME, Fodor É, Nardini C, et al. 2021. *Phys. Rev. E* 103:032607. **https://doi.org/10.1103/PhysRevE.103.032607**
70. Fodor É, Guo M, Gov NS, Visco P, Weitz DA, van Wijland F. 2015. *Europhys. Lett.* 110(4):48005
71. Ahmed WW, Fodor É, Almonacid M, Bussonnier M, Verlhac M-H, et al. 2018. *Biophys. J.* 114(7):1667–79
72. Gnesotto FS, Mura F, Gladrow J, Broedersz CP. 2018. *Rep. Prog. Phys.* 81(6):066601
73. Harada T, Sasa Si. 2005. *Phys. Rev. Lett.* 95:130602
74. Szamel G. 2019. *Phys. Rev. E* 100:050603
75. Toyabe S, Okamoto T, Watanabe-Nakayama T, Taketani H, Kudo S, Muneyuki E. 2010. *Phys. Rev. Lett.* 104:198103
76. Fodor É, Ahmed WW, Almonacid M, Bussonnier M, Gov NS, et al. 2016. *Europhys. Lett.* 116:30008
77. Seara DS, Machta BB, Murrell MP. 2021. *Nat. Commun.* 12:392
78. Flenner E, Szamel G. 2020. *Phys. Rev. E* 102:022607
79. Dabelow L, Bo S, Eichhorn R. 2021. *J. Stat. Mech.* 2021(3):033216
80. Fodor É, Nemoto T, Vaikuntanathan S. 2020. *New J. Phys.* 22:013052
81. Tociu L, Fodor É, Nemoto T, Vaikuntanathan S. 2019. *Phys. Rev. X* 9:041026
82. Hansen JP, McDonald IR. 2013. *Theory of Simple Liquids*. Oxford: Academic
83. Tociu L, Rassolov G, Fodor É, Vaikuntanathan S. 2020. arXiv:2012.10441
84. Li YI, Cates ME. 2021. *J. Stat. Mech.* 2021(1):013211
85. Caballero F, Cates ME. 2020. *Phys. Rev. Lett.* 124:240604
86. Garrahan JP, Jack RL, Lecomte V, Pitard E, van Duijvendijk K, van Wijland F. 2009. *J. Phys. A* 42(7):075007
87. Bodineau T, Derrida B. 2004. *Phys. Rev. Lett.* 92(18):180601
88. Appert-Rolland C, Derrida B, Lecomte V, van Wijland F. 2008. *Phys. Rev. E* 78(2):021122
89. Jack RL, Sollich P. 2015. *Eur. Phys. J. Spec. Top.* 224(12):2351–67
90. Dolezal J, Jack RL. 2019. *J. Stat. Mech.* 2019(12):123208
91. Hedges LO, Jack RL, Garrahan JP, Chandler D. 2009. *Science* 323(5919):1309–13
92. Garrahan JP, Lesanovsky I. 2010. *Phys. Rev. Lett.* 104(16):160601
93. Weber JK, Jack RL, Schwantes CR, Pande VS. 2014. *Biophys. J.* 107(4):974–82
94. Nemoto T, Fodor É, Cates ME, Jack RL, Tailleur J. 2019. *Phys. Rev. E* 99:022605. **https://doi.org/10.1103/PhysRevE.99.022605**
95. Simha A, Evans RML, Baule A. 2008. *Phys. Rev. E* 77(3):031117
96. Pressé S, Ghosh K, Lee J, Dill KA. 2013. *Rev. Mod. Phys.* 85(3):1115–41
97. den Hollander F. 2000. *Large Deviations*. Providence, RI: Am. Math. Soc.
98. Bertsekas DP. 2005. *Dynamic Programming and Optimal Control*, Vol. 1. Belmont, MA: Athena Sci.
99. Dupuis P, Ellis RS. 1997. *A Weak Convergence Approach to the Theory of Large Deviations*. New York: Wiley
100. Chétrite R, Touchette H. 2015. *J. Stat. Mech.* 2015(12):P12001
101. Chetrite R, Touchette H. 2015. *Ann. Henri Poincaré* 16(9):2005–57
102. Jack RL, Sollich P. 2010. *Prog. Theor. Phys. Suppl.* 184:304–17
103. Tsobgni Nyawo P, Touchette H. 2016. *Phys. Rev. E* 94:032101
104. Cagnetta F, Mallmin E. 2020. *Phys. Rev. E* 101:022130
105. Keta YE, Fodor É, van Wijland F, Cates ME, Jack RL. 2021. *Phys. Rev. E* 103:022603. **https://doi.org/10.1103/PhysRevE.103.022603**
106. Nemoto T, Bouchet F, Jack RL, Lecomte V. 2016. *Phys. Rev. E* 93(6):062123
107. Nemoto T, Jack RL, Lecomte V. 2017. *Phys. Rev. Lett.* 118(11):115702
108. Ray U, Chan GKL, Limmer DT. 2018. *Phys. Rev. Lett.* 120:210602
109. Jacobson D, Whitelam S. 2019. *Phys. Rev. E* 100:052139
110. Bertini L, De Sole A, Gabrielli D, Jona-Lasinio G, Landim C. 2015. *Rev. Mod. Phys.* 87(2):593–636

111. Whitelam S, Klymko K, Mandal D. 2018. *J. Chem. Phys.* 148(15):154902
112. GrandPre T, Klymko K, Mandadapu KK, Limmer DT. 2021. *Phys. Rev. E* 103:012613
113. Cagnetta F, Corberi F, Gonnella G, Suma A. 2017. *Phys. Rev. Lett.* 119:158002
114. GrandPre T, Limmer DT. 2018. *Phys. Rev. E* 98:060601
115. Chiarantoni P, Cagnetta F, Corberi F, Gonnella G, Suma A. 2020. *J. Phys. A Math. Theor.* 53(36):36LT02
116. Dolezal J, Jack RL. 2021. *Phys. Rev. E* 103:052132
117. Horowitz JM, Gingrich TR. 2020. *Nat. Phys.* 16(1):15–20

Annual Review of Condensed Matter Physics

Volume 13, 2022

Contents

Errata

An online log of corrections to *Annual Review of Condensed Matter Physics* articles may be found at http://www.annualreviews.org/errata/conmatphys