REVIEW ARTICLE
https://doi.org/10.1038/s42256-020-0146-9

nature machine intelligence

Machine learning for active matter

Frank Cichos[1], Kristian Gustavsson[2], Bernhard Mehlig[2] and Giovanni Volpe[2]

The availability of large datasets has boosted the application of machine learning in many fields and is now starting to shape active-matter research as well. Machine learning techniques have already been successfully applied to active-matter data—for example, deep neural networks to analyse images and track objects, and recurrent nets and random forests to analyse time series. Yet machine learning can also help to disentangle the complexity of biological active matter, helping, for example, to establish a relation between genetic code and emergent bacterial behaviour, to find navigation strategies in complex environments, and to map physical cues to animal behaviours. In this Review, we highlight the current state of the art in the application of machine learning to active matter and discuss opportunities and challenges that are emerging. We also emphasize how active matter and machine learning can work together for mutual benefit.

The past decade has seen a significant increase in the amount of experimental data gathered in the natural sciences, as well as in the computer resources used to store and analyse these data. Prompted by this development, data-driven machine-learning methods are beginning to be widely used in the natural sciences, for example in condensed-matter physics[1,2], microscopy[3,4], fluid mechanics[5] and biology[6].

During the past few years, active-matter research[7] has also begun to successfully employ machine-learning approaches. This is not surprising, because active-matter systems, like complex systems in general, tend to exhibit many more degrees of freedom than conservation laws, causing fundamental challenges for mechanistic models. This was realized already about 50 years ago, when chaos theory emerged as the basis for statistical descriptions of complex systems—causing a paradigm shift from mechanistic to probabilistic interpretations of experimental data, highlighting the fundamental difficulty in predicting the dynamics of non-linear complex systems[8]. Furthermore, standard statistical-physics approaches are most easily applied to systems in thermodynamical equilibrium, while active matter often features far-from-equilibrium dynamics.

Data-driven machine-learning methods offer unprecedented opportunities for active-matter research, where the availability of a large amount of data is matched by the lack of simple statistical-physics models. For example, data-driven methods can be used to obtain model-free predictions, which can be extraordinarily useful for many purposes (weather forecasting is an important example). But will machine learning initiate new trends in active-matter research? Will this lead to ground-breaking insight and applications? More fundamentally, how can machine learning contribute to our understanding of active matter? Can this help us identify unifying principles and systematize active matter?

In this Review, after a brief introduction to active matter, we illustrate the potential of machine-learning methods in active-matter research by describing the most successful recent machine-learning applications in this field. Then we discuss the main opportunities and, most importantly, the principal challenges for future applications.

Active matter

The term active matter was coined to describe natural and artificial systems (Fig. 1a) made of active particles that draw energy from their local environment to perform mechanical work[7]. Natural systems, from molecular motors, to cells and bacteria, to fish, birds and other organisms, are intrinsically out of thermodynamic equilibrium as they convert chemical energy. Their biochemical networks and sensory systems are optimized by evolution to perform specific tasks: in the case of motile microorganisms, for example to cope with ocean turbulence, to navigate along chemical gradients, and more generally to follow specific strategies in foraging[9–12] (Fig. 1b). Artificial active matter covers a similar size range from self-propelling artificial molecules and microparticles, whose development is just in its infancy, to macroscopic robots, which consume energy from sources such as heat and electricity.

Active-matter research is concerned with understanding how macroscopic spatio-temporal collective patterns may emerge, driven by energy conversion from the smallest to the largest scales, mediated by physical interactions (Fig. 1c). Dense systems of bacteria, for example, develop active turbulence at length scales where only laminar flows are expected from the underlying physical laws[13,14]. Cells grow into tissues and sometimes may form tumours. Dense filaments and motor proteins, which are the structural building blocks of cells, develop active nematic structures with new physical properties[15]. The onset of such collective behaviours is also observed in artificial systems where increased energy input above a threshold density drives a phase transition to an aggregated state[16,17].

In dilute systems, when active particles have no direct physical interaction with each other, natural systems have evolved sensing capabilities, which allow them to gain information about their environments or to communicate. This introduces a new dimension with entirely different challenges and importance in other fields such as ecology[18,19]. In schools of fish and flocks of birds or midges, individuals exchange information as part of their behaviour to self-organize into a collective state[20]. The underlying behavioural rules are often hard to identify but could be extremely useful for the application in robots swarms[21–23] or for information-based structure formation in microscopic systems[24].

Experimental active-matter studies provide the testing grounds for new non-equilibrium descriptions, which are by necessity often computational. They are either based on hypothesized mechanistic models for local interactions, coarse-grained hydrodynamic approximations[25] or basic fluctuation theorems[26]. The question is

[1]Peter Debye Institute for Soft Matter Physics, Faculty of Physics and Earth Sciences, Leipzig University, Leipzig, Germany. [2]Department of Physics, University of Gothenburg, Gothenburg, Sweden. e-mail: cichos@physik.uni-leipzig.de; kristian.gustafsson@physics.gu.se; bernhard.mehlig@physics.gu.se; giovanni.volpe@physics.gu.se

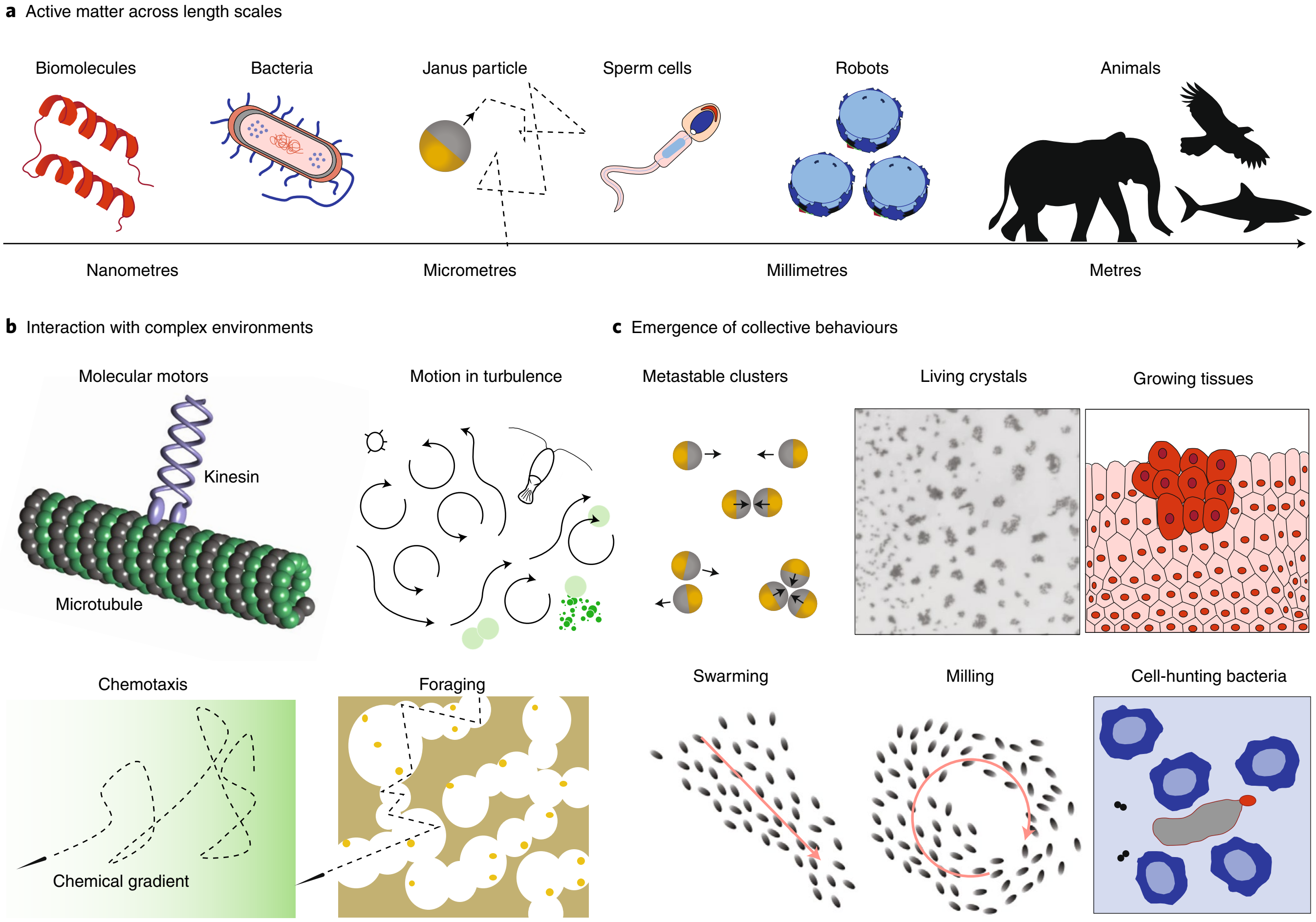


Fig. 1 | Active-matter systems and phenomena. a, Examples of active particles range in size from micrometres to metres (for example, biomolecular motors, motile bacteria, sperm cells, artificial microscopic particles, fish, birds, mammals and robots). **b**, Active particles react to environmental signals and optimize their behaviour to reach certain goals—for example, biomolecular motors move along microtubules, microorganisms swim in turbulent flows, motile cells respond to chemotactic gradients, and animals look for food (foraging). **c**, Interactions between active particles may lead to complex collective behaviours, such as the growth of metastable clusters of particles, and to the emergence of collective dynamics such as swarming and milling.

often how local energy input and physical interactions determine the macroscopic spatio-temporal patterns. Answers are sought by simulation using molecular dynamics[27], Monte Carlo methods[28], cellular automata[29] or numerical solutions of hydrodynamical equations[30]. Data-driven machine-learning methods offer radically different, and partly complementary, opportunities. The reasoning is reversed: instead of asking how spatio-temporal patterns emerge from given microscopic interactions, the goal is to determine the fundamental principles that govern the spatio-temporal dynamics directly from the data, obtained from experiments or simulations.

Machine learning for active matter

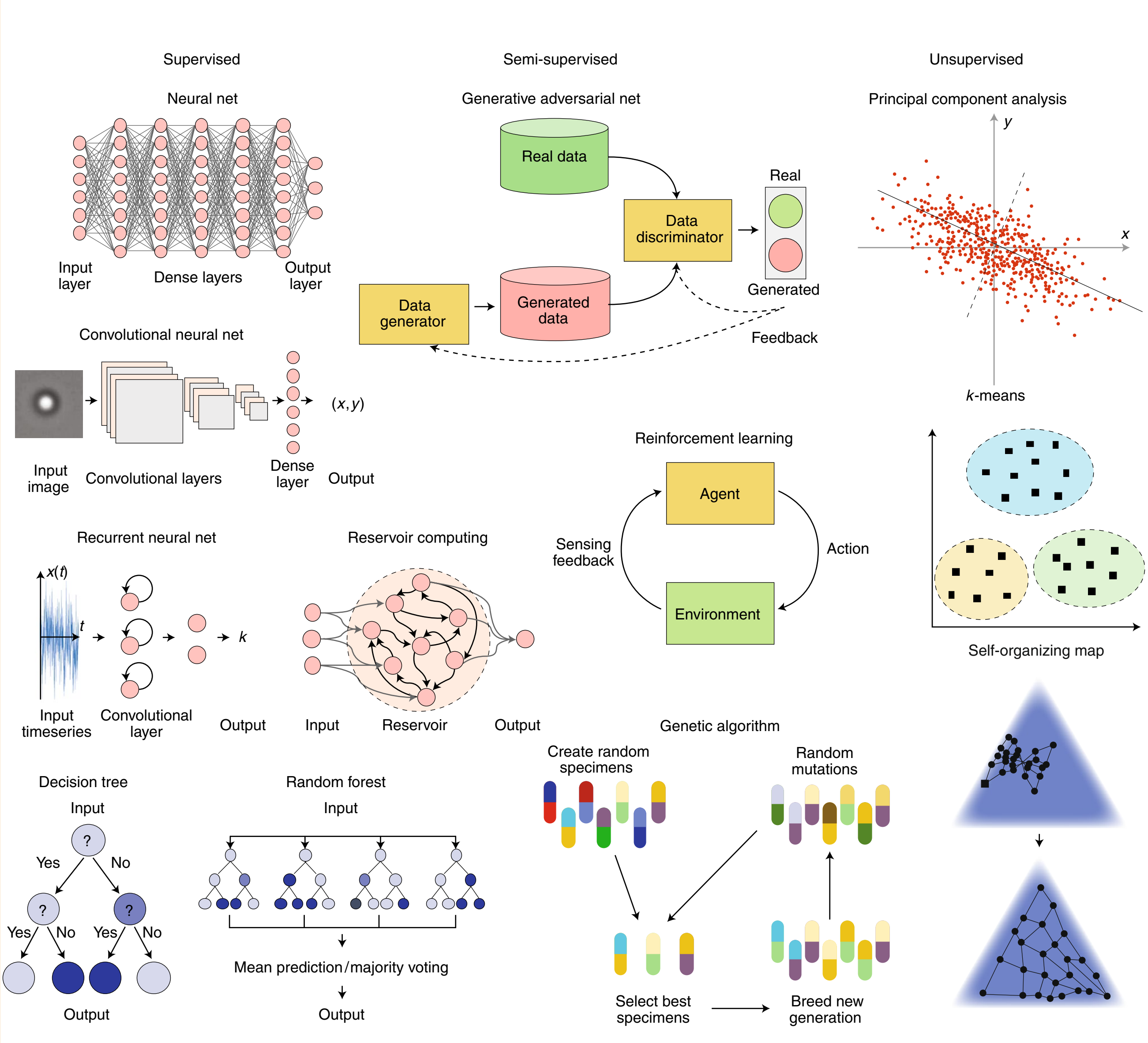
During the past 70 years, the formulation and improvement of machine-learning algorithms has benefited from the understanding of biological systems[31]. Recently, this development has accelerated substantially[1,2,5,6], sparked by the the availability of large amounts of training data that has led to a tremendous success of neural-net-based algorithms in image recognition and classification[32] (Box 1). One usually distinguishes between supervised, semi-supervised and unsupervised methods. Supervised-learning models are trained on labelled datasets annotated with the correct classifications (targets). Semi-supervised methods rely on partial target information for learning. Unsupervised methods do not require training on annotated data, while they uncover hidden relations in high-dimensional data that are not easily discernible. The application of these techniques in active-matter research can be grouped into four different fields.

Data acquisition and analysis. The most important and common use of machine learning in active-matter research is in the analysis and classification of experimental data using supervised learning models (usually neural nets)[33–37] (Box 1). In fact, most active-matter experiments are performed using video microscopy, which provides large, high-quality training datasets that also cover less likely experimental conditions, which are ideally suited to image analysis with supervised machine-learning methods[38].

Supervised learning based on convolutional nets (Box 1) has enhanced video microscopy by improving its spatial resolution[39], extending its depth of field[40] and retaining objects in focus[41]. More generally, detecting and following microscopic active particles in video-microscopy recordings poses enormous challenges, especially for heterogeneous samples with multiple species, varying contrast and low signal-to-noise ratio. Recently, convolutional nets were shown to improve particle imaging velocimetry[33], to localize particles in holographic microscopy[34], and to track particles[36]. Convolutional nets outperform conventional centroid-based algorithms in the analysis of video-microscopy recordings of microscopic particles and motile cells (Fig. 2a). They have also been used to track thousands of individual honey bees in a hive,

Box 1 | Overview of machine-learning methods

Supervised learning builds on labelled datasets containing the inputs as well as the properties (targets) the algorithm is trained to learn. Supervised models are often based on neural nets[38]: networks of non-linear computation units (artificial neurons) connected by weights. These weights are iteratively adjusted (trained) until the neural net learns to associate the correct target to each input. Deep neural nets with many layers of neurons have great potential, but deeper nets are more prone to training instabilities[99]. Convolutional nets are particularly well-suited for image analysis[99,100] and recurrent nets for time-series prediction[101]. Apart from neural nets, decision trees and random forests (ensembles of decision trees) are frequently used models for supervised learning[102]. Unsupervised learning, by contrast, does not require training on labelled data, but exploits redundancy in the input data[38] to learn to compress data, identify likely input patterns, and find patterns in the data using non-linear projections[103] or self-organized maps[104]. Several common statistical-analysis methods fall into this category: principal-component analysis[105], *k*-means clustering[106] and other clustering algorithms[107]. Semi-supervised methods, such as reinforcement learning[78,108–111], learn from partially labelled data, or from incomplete feedback in the form of penalty or reward. Genetic algorithms[112] are inspired by the genetic-sequence evolution through drift, mutation and recombination; penalty and reward are modelled on natural selection. Reservoir computing[46] uses recurrent nets where only the output neurons are trained, while the others form a reservoir of recurrently connected neurons. Adversarial pairs of neural nets[113] train each other to generate synthetic data that is very difficult to distinguish from authentic data.



obtaining the data necessary to analyse how their social structure is reflected in their collective dynamics[35], an important question also for other species, such as flocking jackdaws[42] and even humans[43].

Often the study of active-matter systems requires time-series analysis, a task that lends itself to supervised machine-learning approaches (Box 1).

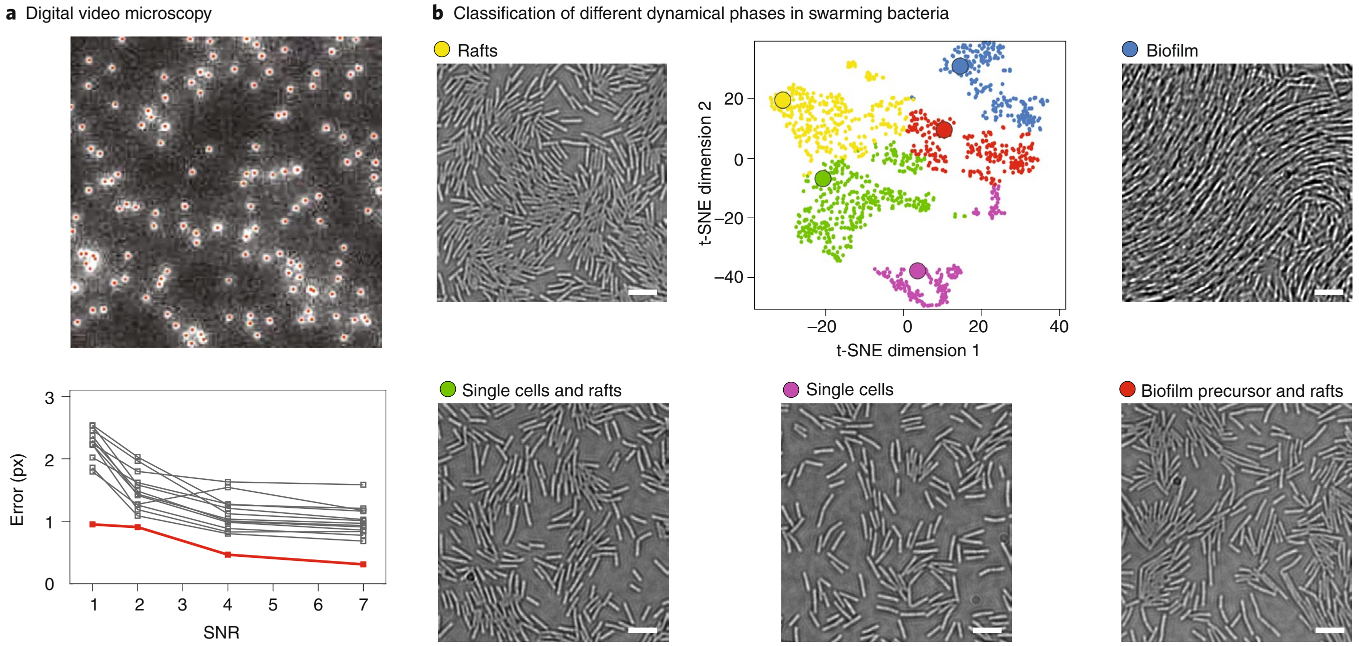


Fig. 2 | Data acquisition and analysis. a, Convolutional nets (red line) outperform other particle-tracking methods (grey lines) to detect microscopic particles (red marker in top panel), especially in the presence of noise and under poor illumination conditions. SNR, signal to noise ratio. **b**, Classification of different dynamical phases in swarming bacteria using stochastic neighbourhood embedding (t-SNE) and *k*-means clustering: single cells, rafts, biofilm, single cells and rafts, and biofilm precursor with rafts. Scale bars, 10 μm. Figure adapted with permission from ref. [37], OSA (**a**); and ref. [44], PNAS (**b**).

Unsupervised machine-learning methods (Box 1) can automatically categorize large amounts of video-microscopy data on the collective motion of swarming agents. Figure 2b illustrates how a variant of *k*-means clustering manages to identify different dynamical phases of swarming bacteria, which, combined with a numerical model analysis, explains the observed swarming behaviours[44]. Unsupervised learning will also be used in the Human Cell Atlas project, which aims to classify all human cell types based on their molecular profiles[45].

Data-driven models. The dynamics of dense active matter results from the interaction between many, often microscopic, parts. While microscopic motion typically appears random, well-defined spatiotemporal patterns may emerge at meso- and macroscopic scales. Such patterns are notoriously difficult to describe, understand and predict from first principles. Semi-supervised learning can tackle this problem using time-series data to construct numerical models for the spatio-temporal dynamics. For example, reservoir computing (Box 1) managed to forecast complex spatio-temporal patterns, which may exist in chemical, biological and physical systems, directly from the input data[46] (Fig. 3a). As another example, motile cells, bacteria or artificial active particles may exhibit anomalous diffusion[7]. Their subdiffusive and superdiffusive dynamics have been classified and characterized using recurrent nets[47] (Fig. 3b) and random forests[48], determining the value of the anomalous diffusion exponent and its temporal fluctuations, which is essential to discover the mechanisms that generate motility, and determine anisotropic and heterogeneous motility patterns[49,50].

More generally, one can infer underlying models from time-series data by symbolic regression using genetic algorithms[51]. Sparse regression ensures that the model has as few fitting parameters as possible[52]. This method has achieved some success in finding partial differential equations from spatial time-series data[53]. An efficient way to obtain a model from high-dimensional time-series data is to reduce the dimensionality by projection—that is, to find a low-dimensional surrogate model that is easier to handle and analyse, but still describes the main features of the original data. The best projection is often not simply a spatial one, but perhaps a projection onto the relevant modes[54]. Machine learning appears to be ideally suited to solve this problem. For example, it has already been used to solve the difficult task of elucidating the intricate three-dimensional spatio-temporal patterns associated with turbulent flows[55]. Furthermore, machine learning can provide invaluable help to infer dynamic information and underlying models from static information[56,57].

Navigation and search strategies. Motility, navigation and search strategies are interesting for physics, biology, ecology and robotics. Like foraging animals, active particles can navigate and search complex environments[58,59]. To understand how evolution shaped search strategies of small motile organisms, one can use reinforcement learning to identify optimal and alternative strategies[60] (Fig. 4a). A challenge is that many active-matter systems are suspensions, presenting the agents with a fluctuating environment. For example, motile plankton must cope with ocean turbulence[61]. A recent proof-of-principle study demonstrated how reinforcement learning finds good strategies for navigation in a steady flow[62] (Fig. 4b), and for point-to-point navigation in complex fluid environments[63–65]. Reinforcement learning can be used to find strategies for marine probes to target certain oceanic regions of interest[64], and also yields fundamental insight into how birds soar in thermal updrafts guided by cues from the turbulent air flow[66], enabling gliders to soar in such updrafts[67] (Fig. 4c).

Collective dynamics in interacting populations. Groups of animals often feature organized collective behaviours, from swarms of insects to flocks of birds[68]. Swarming provides several benefits to the individuals: it can reduce the risk of predation, increase the opportunities for feeding, provide chances for reproduction, and reduce energy consumption by optimizing hydrodynamical interactions in schools of fish or flocks of birds[58,68]. However, swarming requires different navigation skills compared to moving alone[69]. Developing such skills entails a cost, so the corresponding strategies can only emerge through

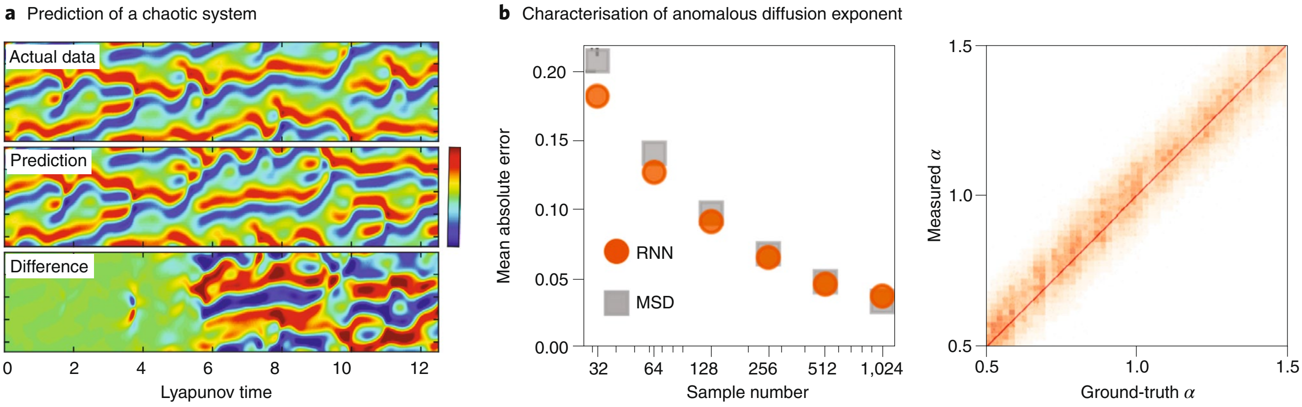


Fig. 3 | Data-driven models. a, Dynamics of a spatio-temporally chaotic system (upper panel), dynamics predicted using reservoir computing (middle panel), and their difference (lower panel). The prediction is accurate up to approximately six Lyapunov times. **b**, Recurrent nets (RNN) allow the characterization of the anomalous diffusion exponent α better than the standard mean square displacement (MSD), especially for very few sample point numbers (left panel), even in situations where standard methods fail, for example for irregularly sampled and intermittent processes (the right panel shows the accuracy of the prediction for an irregularly sampled trajectory). Figure adapted with permission from ref. [46], APS (**a**); and ref. [47], APS (**b**).

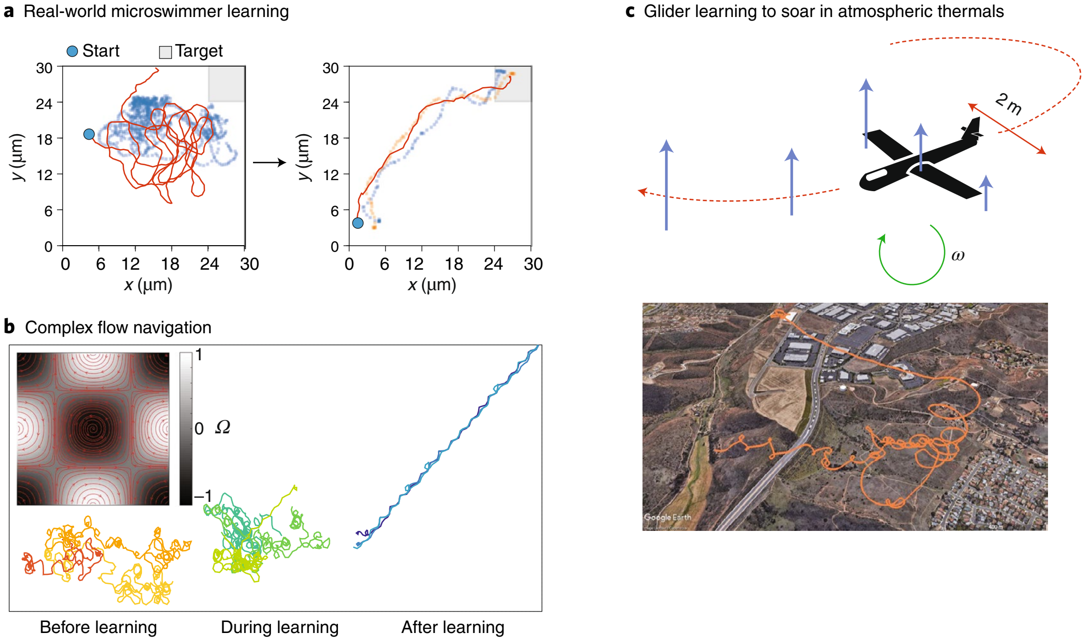

a Real-world microswimmer learning
Start
Target
30
24
18
12
6
0
y (µm)
x (µm)
b Complex flow navigation
1
0 Ω
−1
Before learning
During learning
After learning
c Glider learning to soar in atmospheric thermals
2 m
ω
Google Earth

Fig. 4 | Navigation and search strategies. a, Real-world microswimmer learning. Reinforcement learning allows artificial microswimmers to explore real-world environments and efficiently reach a specified target, in this example the shaded square. The two plots show typical trajectories before (left) and after (right) training. **b**, Complex flow navigation. Reinforcement learning allows one to efficiently find optimal (non-intuitive) navigation strategies in simulated complex flows. This example shows how a simulated microswimmer learns to swim in the vertical direction in a tessellation of counter-rotating vortices (inset), guided only by the local flow vorticity Ω. Trajectories before (left), during (middle) and after (right) training are shown. **c**, Learning how to soar. Reinforcement learning provides an efficient way to train a glider to autonomously navigate atmospheric thermals, using the available navigational cues, namely gradients in vertical wind velocities (indicated by the length of the blue arrows) along the glider and across its wings. The orange line in the lower panel shows one resulting flight path. Figure adapted with permission from ref. [60] (**a**); ref. [62], APS (**b**); and ref. [67], Springer Nature Ltd (**c**).

evolution if they lead to significant gains. Multi-agent reinforcement learning allows one to determine under which circumstances collective sensing may emerge, where a group may sense scalar gradients (turbulence intensity, food concentration, or light intensity), despite the fact that individuals can only measure scalars[70–73].

Machine learning offers opportunities far beyond categorizing the different dynamical phases of swarming (Fig. 2b), permitting the exploration of the yet-unknown physical mechanisms underlying such advanced behaviours. For example, reinforcement learning[74] and deep reinforcement learning[75] have been used to find optimal swimming strategies that minimize drag and energy consumption in a simulated school of fish (Fig. 5a).

Opportunities and challenges

The recent success of machine-learning approaches in active-matter research provides a glimpse into possible future applications. Naturally, these opportunities come with challenges, primarily the fact that many machine-learning methods are effectively black-box models that cannot provide the interpretability that is expected in the natural sciences. In the following we summarize the most

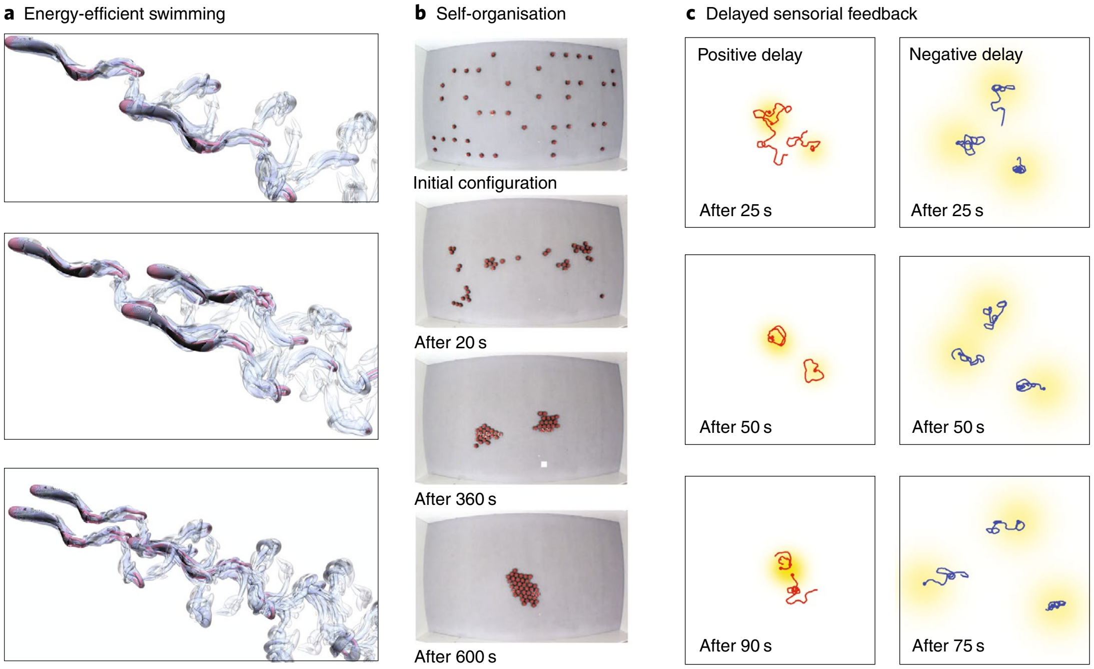


Fig. 5 | Collective dynamics in interacting populations. a, Energy-efficient swimming. Numerical simulations of energy efficient swimming strategies from deep reinforcement learning to exploit the wake of leading swimmers with one leader and one follower (top panel), one leader and two followers (middle panel), and two leaders and one follower (bottom panel). **b**, Self-organized aggregation in a robot swarm. The behaviour is inferred from the observation of aggregating robots. **c**, Delayed sensorial feedback. A group of three phototactic robots, which emit a radially decaying light intensity around themselves and adjust their speed as a function of the sensed light intensity, aggregate into a dynamic cluster if their sensorial delay is positive (left panels) and segregate if it is negative (right panels). The trajectories show a period of 10 s preceding the time indicated on the plot, and the dots indicate the final position of the robots. Figure adapted with permission from ref. [75], PNAS (**a**); ref. [80], Springer (**b**); and ref. [70], APS (**c**).

far-reaching opportunities and the most significant challenges for machine learning in active-matter research.

Opportunities. We identified a number of so-far-unexplored research directions with great potential.

Improvement of data acquisition and analysis. Until now, active-matter research has used standard machine-learning models for image processing (for example, convolutional nets for particle tracking[33–37]) and for time series analysis (recurrent nets for data classification[44,45]). Going forward, suitable neural-net architectures for analysing more complex, multimodal data will require more targeted, hybrid approaches[76].

The use of machine-learning methods in data acquisition and analysis will go along with improved feedback control of experimental setups and protocols. For example, machine-learning control could improve temperature stabilization when actuating artificial active particles, or material deposition when microfabricating such particles. Machine learning may also help to find optimal parameters in real time during experiments, and support data acquisition and analysis methods that dynamically adapt to a time-varying signal (for example, due to the presence of drifts in the experimental setup).

Inspiration from swarm robotics. Active-matter research should seek analogies with the problems studied in swarm robotics[21,77], where genetic algorithms[21,22] (Box 1) and reinforcement learning[23] have been applied for some time. Recent trends in swarm-robotics research are deep reinforcement learning[78], model-based behaviour trees evolved using genetic programming that result in human-readable output[79], and machine-learning methods based on generative adversarial networks[80] (Box 1 and Fig. 5b). Engineering the interplay between system length and time scales, correlated noises and sensorial delays, it is even possible to tune the macroscopic swarm dynamics[70,71,81–83] (Fig. 5c). These methods developed in swarm-robotics research can also have important implications for biological active-matter systems.

Systematiation of active matter. On a more fundamental level, machine learning can help to tame the diversity of active matter. There are many different active-matter systems (Fig. 1): dense or dilute, granular or fluid, biological or artificial, microscopic or macroscopic. These systems are subject to quite different interaction mechanisms, such as hydrodynamic, electrostatic, adhesive, chemical or just steric interactions. One of the main challenges is to determine the fundamental underlying similarities and differences. Machine learning can help by using its ability to mine information from large experimental and numerical datasets. For example, the governing equations of physical phenomena have been recently discovered from large datasets using sparse regression[52,53].

A possible route towards this systematization goal is to understand and classify how different interaction mechanisms are related to macroscopic spatio-temporal patterns observed in many different active-matter systems. Active particles use energy for locomotion and drive the system at the smallest scales; they may also use local information to find suitable strategies, resulting in macroscopic patterns or collective behaviours. Often, we do not know what these local, microscopic interactions look like. Devising models for such a diverse range of emergent behaviours is no simple task. This has been mostly attempted by proposing heuristic mechanistic rules for the dynamics at the individual level, such as matching velocities, avoiding collisions or forming a centred group[68]. Instead, machine-learning methods can suggest possible local interaction mechanisms from the data. Reinforcement learning, for example, can help to find candidate strategies, either for an individual or for the population as a whole, to reach certain goals. As we have seen in the previous section, similar algorithms have been widely used in robotics[22,23] and research along these lines has recently started also in models of biological active matter[62,73]. The next steps will be to develop more complex strategies, which work in realistic environments and respond to environmental cues.

Insight into biological active matter. During evolution, biological species have developed sensorial networks to interact with their environment. Active matter, machine learning and biology can join forces to decode how this sensorial information is used to determine the interaction strategies. Reinforcement learning and simple neural nets are particularly well-suited to compare the importance of different sensorial inputs[61]. For example, consider how birds survive and navigate local air turbulence. Reinforcement learning has been recently employed to identify strategies for the birds to cope with turbulence[66]. However, this pioneering work tested only a limited discrete set of possible signals and cues that inform the birds about their environment, which do not necessarily reflect what the birds actually sense.

Many species rely on chemotaxis to find potential mating partners or prey[84]. This is straightforward in a quiescent environment where the strategy is simply to climb the concentration gradient. But what is the best strategy when there is flow, as, for example, in nutrient-rich upwelling regions in a turbulent flow? This problem becomes still more challenging when the chemical signal is intermittent or in unsteady flows. Different search strategies have been proposed, found using standard algorithmic approaches[58,59]. Machine-learning methods can improve on these results and find strategies that are even better at specific tasks, and are adapted to complex, time-varying environments.

Evolution of biological active matter. The morphology, functionality and behaviour of biological active matter evolved as a tradeoff between benefit, cost and risk. Evolution is an optimization problem with multiple cost functions, which may even vary over time. Optimizing one cost function may frustrate or promote other important goals. A far-reaching, fundamental question is how these strategies, morphologies and functions have evolved in dynamical environments, and to what extent the morphological features of an organism are themselves part of its computing power[85]. Leading further, one should compare strategies that are optimal for a single individual with strategies that are optimal for a swarm. Are there swarming strategies that are equally beneficial for the individuals, but cost less (for example, require fewer sensory inputs, or a smaller number of motility functions)? Are there circumstances where strategies evolve that are beneficial for the group as a whole, but that damage the individual? How does the behaviour of a species depend on other species it interacts with (for example, predators)? When it comes to swarms of motile organisms, there are fundamental open questions concerning the role of leadership, hierarchy and other social structure in collective decision making. This has been studied using mechanistic models[68], but many open questions that can be addressed with machine learning remain, such as what changes when we take into account that the particles can adapt their behaviour in response to sensory inputs. In the long run, it may even be possible to use these insights into biological evolution to create evolvable artificial active matter, where machine-learning algorithms optimize the morphology and behaviour of artificial active matter to optimize some goals such as particle self-assembly, collective swarming or targeted drug delivery.

Active matter with embodied intelligence. In the conventional approaches to applying machine learning to microscopic active matter discussed until now, the machine-learning part is performed on a computer, which provides the computational power and speed to execute the machine-learning algorithms. An interesting avenue to explore in the future is to develop artificial active materials that are smart enough to carry out basic computation and adaption without referring to in silico training procedures. This requires new versatile ways to incorporate signal inputs, signal processing and memory storage into microscopic materials and agents that can then act truly autonomously. For example, such approaches may build on the networks of chemical reactions involving macromolecules and active components[86]. These possibilities are recently exploited in the development of DNA computing, where neural nets are implemented by DNA strand displacement cascades[87].

Challenges. In general, machine learning is used either as a tool or a model, and this applies to active-matter research too. In data processing, forecasting, projection and optimization, learning the final result is usually more important than the internal processing of the algorithm. One might think that the interpretation of the results is quite unproblematic under these circumstances, but it is worth noting a number of important caveats and challenges. Catastrophic forgetting[88] may limit what a network can learn, and there is much speculation about which factors affect the ability of the network to generalize[38], yet there are few definite mathematical results (see ref. [89] for an exception). A potentially even more serious problem is that convolutional nets tend to be quite certain in their classification, even when they are wrong[90]. It is a question of ongoing research how to know when and why the network fails to classify correctly.

Many machine-learning tools work as black boxes[91], which is problematic when using machine learning to generate a simple model of a complex system that can be intuitively understood and therefore generalized to other parameter values or situations. Reinforcement learning has the definite advantage that its results are interpretable, but there are nevertheless a number of caveats specific to active-matter research. Reinforcement-learning approaches often use simulation data instead of empirical data (see refs. [22,60,67] for exceptions), because it is usually easier to generate the required large datasets using simulations. In some cases, the necessary information may not even be accessible experimentally, as in many living active-matter systems. Strategies found from simulation data usually work less well in the real world—this is the so-called reality gap[21,22,92]. Research on how to best avoid or overcome the reality gap is ongoing[93].

Therefore, some caution must be taken when incorporating machine-learning models into the scientific process of understanding active matter. Guidelines for how to apply machine learning have been compiled, taking into account the specific issues arising in different fields[93–96]. Below we give guidelines suitable for initial application of machine learning in active-matter systems.

1. The performance of machine-learning models must be benchmarked against other known and commonly used approaches. This serves as a sanity check, which establishes a baseline to beat. For example, ref. [37] demonstrates how a neural-network-based particle-detection algorithm outperforms conventional particle-tracking methods (Fig. 2a). A second example is reinforcement learning, where the crucial question is whether the strategies found by reinforcement learning outperform previously known strategies (Fig. 4b)[62].
2. Machine-learning models with simpler architectures and fewer parameters are better[97], not only because they are less prone to overfitting, but also because their results are usually easier to interpret. For example, there is no need for deep neural nets or random forests when a simpler clustering algorithm manages to classify high-dimensional data accurately and efficiently (Fig. 3b)[44].
3. Usually the input data must be preprocessed. This appears to be at odds with a central dogma of the deep-learning revolution,

namely to avoid feature engineering before applying machine-learning methods. However, preprocessing is essential to minimize the risk of overfitting to spurious correlations that might be present in large datasets[98], and because machine-learning methods have difficulties coping with distorted inputs[38].

4. It is important to avoid applying machine-learning models outside the range of input data for which they have been trained. This is obvious because extrapolation is a more complex and risky operation than interpolation. However, in the case of machine-learning models, a danger is that extrapolation occurs unintentionally and in an uncontrolled fashion. For example, recurrent nets may correctly predict an anomalous diffusion exponent in a certain range, but fail for data with smaller or larger exponents than the training set[47]. More fundamentally, it is not known how the method would perform on data that is generated by different models from those in the training set. In a similar vein, the spatio-temporal predictions of reservoir computing for chaotic systems[46] are likely to fail for extreme initial conditions, outside the training set (Fig. 3a).
5. The most important point, in our opinion, is that one should strive to use physics-informed machine-learning models, even though it is not yet clear how to do this in general. These are machine-learning models that take into account the physical properties of the system under investigation. For example, they could enforce the conservation laws and symmetries that characterize the physical system under study. This will have the added benefit of allowing simpler machine-learning models that are also easier to understand, as well as allowing the evaluation of how the model changes as different symmetries or interactions are implemented.

Potential benefits for machine learning. Active-matter systems may also serve as a tool to advance machine learning mostly by providing well-controlled reference systems:

1. Active-matter systems in the lab can generate very large high-quality datasets corresponding to complex but controllable physical phenomena. These datasets can be used to train machine-learning models and gain theoretical insights into how they work and what the best architectures are. Furthermore, these datasets can work as suitable benchmarks to test alternative machine-learning models.
2. Thanks to their microscopic nature, many active-matter systems can be observed on multiple time and length scales, providing direct access to both their microscopic dynamics and the resulting macroscopic behaviours. Active colloidal particles undergo Brownian motion and/or turbulent diffusion, which, together with their intrinsic motility and external forces, may result in complex patterns at different scales. Video microscopy gives access to such colloidal motion at time and length scales ranging from microseconds to hours and from nanometres to millimetres. Thus, the study of active-matter systems may provide an ideal model system to connect microscopic and macroscopic dynamics. This may help to develop machine-learning algorithms that predict the dynamics of systems on long time scales when few samples of their short-time dynamics and local interactions are available.
3. Thanks to the fact that we know the underlying physics for specific systems, several active-matter systems are ideally suited to explore how training data can be simulated in the most efficient and reliable way. For example, it is possible to accurately simulate the hydrodynamic interactions between colloidal active particles, even though it is very computationally expensive. These accurate numerical datasets can be used as a testbed to study how we can generate sufficient training data for real-world applications, where it might be quite expensive, time-consuming and not straightforward to do so, such as in microscopy and virtual tissue staining.

Conclusions

Even though the application of machine learning to active-matter research is still in its infancy, the range of current applications already indicates what machine learning can do for future active-matter research. Besides some standard ways of enhancing data analysis or controlling experimental conditions, machine learning will permit new insights into the functioning of biological systems and the process of how and why these functions evolved. As living active matter has often been the conceptual inspiration for machine learning, we expect that applying machine learning to active matter will quickly enhance machine learning as well, and eventually also lead to artificial active matter that embodies fundamental machine-learning functionalities. Machine learning and active matter thus seem to be bound in a vivid, growing, far-reaching, synergetic relationship with mutual benefits.

Received: 12 August 2019; Accepted: 6 January 2020;
Published online: 14 February 2020

References

1. Mehta, P. et al. A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.* **810**, 1–124 (2019).
2. Das Sarma, S., Deng, D. L. & Duan, L. M. Machine learning meets quantum physics. *Phys. Today* **72**, 48–54 (2019).
3. Waller, L. & Tian, L. Machine learning for 3D microscopy. *Nature* **523**, 416–417 (2015).
4. Barbastathis, G., Ozcan, A. & Situ, G. On the use of deep learning for computational imaging. *Optica* **6**, 921–943 (2019).
5. Brunton, S. L., Noack, B. R. & Koumoutsakos, P. Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **52**, 477–508 (2020).
6. Webb, S. Deep learning for biology. *Nature* **554**, 555–557 (2018).
7. Bechinger, C. et al. Active particles in complex and crowded environments. *Rev. Mod. Phys.* **88**, 045006 (2016).
8. Ruelle, D. *Chance and Chaos* (Princeton Univ. Press, 1991).
9. Gustavsson, K., Berglund, F., Jonsson, P. & Mehlig, B. Preferential sampling and small-scale clustering of gyrotactic microswimmers in turbulence. *Phys. Rev. Lett.* **116**, 108104 (2016).
10. Sengupta, A., Carrara, F. & Stocker, R. Phytoplankton can actively diversify their migration strategy in response to turbulent cues. *Nature* **543**, 555–558 (2017).
11. Durham, W. M., Kessler, J. O. & Stocker, R. Disruption of vertical motility by shear triggers formation of thin phytoplankton layers. *Science* **323**, 1067–1070 (2009).
12. Durham, W. M. et al. Turbulence drives microscale patches of motile phytoplankton. *Nat. Commun.* **4**, 2148 (2013).
13. Yeomans, J. M. Nature's engines: active matter. *Europhys. News* **48**, 21–25 (2017).
14. Urzay, J., Doostmohammadi, A. & Yeomans, J. M. Multi-scale statistics of turbulence motorized by active matter. *J. Fluid Mech.* **822**, 762–773 (2017).
15. Doostmohammadi, A., Ignés-Mullol, J., Yeomans, J. M. & Sagués, F. Active nematics. *Nat. Commun.* **9**, 3246 (2018).
16. Palacci, J., Sacanna, S., Steinberg, A. P., Pine, D. J. & Chaikin, P. M. Living crystals of light-activated colloidal surfers. *Science* **339**, 936–940 (2013).
17. Buttinoni, I. et al. Dynamical clustering and phase separation in suspensions of self-propelled colloidal particles. *Phys. Rev. Lett.* **110**, 238301 (2013).
18. Charlesworth, H. J. & Turner, M. S. Intrinsically motivated collective motion. *Proc. Natl Acad. Sci. USA* **116**, 15362–15367 (2019).
19. Strandburg-Peshkin, A. et al. Visual sensory networks and effective information transfer in animal groups. *Curr. Biol.* **23**, R709–R711 (2013).
20. Attanasi, A. et al. Information transfer and behavioural inertia in starling flocks. *Nat. Phys.* **10**, 691–696 (2014).
21. Trianni, V. *Evolutionary Swarm Robotics* (Springer, 2008).
22. Doncieux, S., Bredeche, N., Mouret, J.-B. & Eiben, A. E. G. Evolutionary robotics: what, why, and where to. *Front. Robot. AI* https://doi.org/10.3389/frobt.2015.00004 (2015).
23. Bayındır, L. A review of swarm robotics tasks. *Neurocomputing* **172**, 292–321 (2016).
24. Khadka, U., Holubec, V., Yang, H. & Cichos, F. Active particles bound by information flows. *Nat. Commun.* **9**, 3864 (2018).
25. Marchetti, M. C. et al. Hydrodynamics of soft active matter. *Rev. Mod. Phys.* **85**, 1143–1189 (2013).

26. Falasco, G., Pfaller, R., Bregulla, A. P., Cichos, F. & Kroy, K. Exact symmetries in the velocity fluctuations of a hot brownian swimmer. *Phys. Rev. E* **94**, 030602 (2016).
27. Frenkel, D. & Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications* (Elsevier, 2001).
28. Rosenbluth, M. N. Genesis of the Monte Carlo algorithm for statistical mechanics. *AIP Conf. Proc.* **690**, 22–30 (2003).
29. Wolfram, S. Cellular automata as models of complexity. *Nature* **311**, 419–424 (1984).
30. Lauga, E. & Powers, T. R. The hydrodynamics of swimming microorganisms. *Rep. Prog. Phys.* **72**, 096601 (2009).
31. Floreano, D. & Mattiussi, C. *Bio-inspired Artificial Intelligence: Theories, Methods, and Technologies* (MIT Press, 2008).
32. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
33. Rabault, J., Kolaas, J. & Jensen, A. Performing particle image velocimetry using artificial neural networks: a proof-of-concept. *Meas. Sci. Tech.* **28**, 125301 (2017).
34. Hannel, M. D., Abdulali, A., O'Brien, M. & Grier, D. G. Machine-learning techniques for fast and accurate feature localization in holograms of colloidal particles. *Opt. Express* **26**, 15221–15231 (2018).
35. Boenisch, F. et al. Tracking all members of a honey bee colony over their lifetime using learned models of correspondence. *Front. Robot. AI* **5** (2018).
36. Newby, J. M., Schaefer, A. M., Lee, P. T., Forest, M. G. & Lai, S. K. Convolutional neural networks automate detection for tracking of submicron-scale particles in 2D and 3D. *Proc. Natl Acad. Sci. USA* **115**, 9026–9031 (2018).
37. Helgadottir, S., Argun, A. & Volpe, G. Digital video microscopy enhanced by deep learning. *Optica* **6**, 506–513 (2019).
38. Mehlig, B. Artificial neural networks. Preprint at https://arxiv.org/abs/1901.05639 (2019).
39. Rivenson, Y. et al. Deep learning microscopy. *Optica* **4**, 1437–1443 (2017).
40. Wu, Y. et al. Extended depth-of-field in holographic imaging using deep-learning-based autofocusing and phase recovery. *Optica* **5**, 704–710 (2018).
41. Pinkard, H., Phillips, Z., Babakhani, A., Fletcher, D. A. & Waller, L. Deep learning for single-shot autofocus microscopy. *Optica* **6**, 794–797 (2019).
42. Ling, H. et al. Behavioural plasticity and the transition to order in jackdaw flocks. *Nat. Commun.* **10**, 5174 (2019).
43. Ouellette, N. T. Flowing crowds. *Science* **363**, 27–28 (2019).
44. Jeckel, H. et al. Learning the space-time phase diagram of bacterial swarm expansion. *Proc. Natl Acad. Sci. USA* **116**, 1489–1494 (2019).
45. Regev, A. et al. The human cell atlas. *eLife* **6**, e27041 (2017).
46. Pathak, J., Hunt, B., Girvan, M., Lu, Z. & Ott, E. Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Phys. Rev. Lett.* **120**, 024102 (2018).
47. Bo, S., Schmidt, F., Eichhorn, R. & Volpe, G. Measurement of anomalous diffusion using recurrent neural networks. *Phys. Rev. E* **100**, 010102(R) (2019).
48. Muñoz-Gil, G., Garcia-March, M. A., Manzo, C., Martín-Guerrero, J. D. & Lewenstein, M. Single trajectory characterization via machine learning. *New J. Phys.* **22**, 013010 (2020).
49. Dehkharghani, A., Waisbord, N., Dunkel, J. & Guasto, J. S. Bacterial scattering in microfluidic crystal flows reveals giant active Taylor–Aris dispersion. *Proc. Natl Acad. Sci. USA* **116**, 11119–11124 (2019).
50. Borgnino, M. et al. Alignment of nonspherical active particles in chaotic flows. *Phys. Rev. Lett.* **123**, 138003 (2019).
51. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009).
52. Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl Acad. Sci. USA* **113**, 3932–3937 (2016).
53. Rudy, S. H., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614 (2017).
54. Cvitanović, P. Recurrent flows: the clockwork behind turbulence. *J. Fluid Mech.* **726**, 1–4 (2013).
55. Fonda, E., Pandey, A., Schumacher, J. & Sreenivasan, K. R. Deep learning in turbulent convection networks. *Proc. Natl Acad. Sci. USA* **116**, 8667–8672 (2019).
56. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl Acad. Sci. USA* **115**, E2467–E2476 (2018).
57. Pearce, P. et al. Learning dynamical information from static protein and sequencing data. *Nat. Commun.* **10**, 5368 (2019).
58. Viswanathan, G. M., Da Luz, M. G. E., Raposo, E. P. & Stanley, H. E. *The Physics of Foraging: An Introduction to Random Searches and Biological Encounters* (Cambridge Univ. Press, 2011).
59. Volpe, G. & Volpe, G. The topography of the environment alters the optimal search strategy for active particles. *Proc. Natl Acad. Sci. USA* **114**, 11350–11355 (2017).
60. Muiños-Landin, S., Ghazi-Zahedi, K. & Cichos, F. Reinforcement learning of artificial microswimmers. Preprint at https://arxiv.org/abs/1803.06425 (2018).
61. Kiørboe, T. *A Mechanistic Approach to Plankton Ecology* (Princeton Univ. Press, 2008).
62. Colabrese, S., Gustavsson, K., Celani, A. & Biferale, L. Flow navigation by smart microswimmers via reinforcement learning. *Phys. Rev. Lett.* **118**, 158004 (2017).
63. Yoo, B. & Kim, J. Path optimization for marine vehicles in ocean currents using reinforcement learning. *J. Mar. Sci. Tech.* **21**, 334–343 (2015).
64. Biferale, L., Bonaccorso, F., Buzzicotti, M., Leoni, P. C. D. & Gustavsson, K. Zermelo's problem: optimal point-to-point navigation in 2D turbulent flows using reinforcement learning. *Chaos* **29**, 103138 (2019).
65. Schneider, E. & Stark, H. Optimal steering of a smart active particle. *Europhys. Lett.* **127**, 34003 (2019).
66. Reddy, G., Celani, A., Sejnowski, T. J. & Vergassola, M. Learning to soar in turbulent environments. *Proc. Natl Acad. Sci. USA* **113**, E4877–E4884 (2016).
67. Reddy, G., Wong-Ng, J., Celani, A., Sejnowski, T. J. & Vergassola, M. Glider soaring via reinforcement learning in the field. *Nature* **562**, 236–239 (2018).
68. Vicsek, T. & Zafeiris, A. Collective motion. *Phys. Rep.* **517**, 71–140 (2012).
69. Berdahl, A. M. et al. Collective animal navigation and migratory culture: from theoretical models to empirical evidence. *Phil. Trans. R. Soc. B* **373**, 20170009 (2018).
70. Mijalkov, M., McDaniel, A., Wehr, J. & Volpe, G. Engineering sensorial delay to control phototaxis and emergent collective behaviors. *Phys. Rev. X* **6**, 011008 (2016).
71. Leyman, M., Ogemark, F., Wehr, J. & Volpe, G. Tuning phototactic robots with sensorial delays. *Phys. Rev. E* **98**, 052606 (2018).
72. Volpe, G. & Wehr, J. Effective drifts in dynamical systems with multiplicative noise: a review of recent progress. *Rep. Prog. Phys.* **79**, 053901 (2016).
73. Palmer, G. & Yaida, S. Optimizing collective fieldtaxis of swarming agents through reinforcement learning. Preprint at https://arxiv.org/abs/1709.02379 (2017).
74. Gazzola, M., Tchieu, A. A., Alexeev, D., de Brauer, A. & Koumoutsakos, P. Learning to school in the presence of hydrodynamic interactions. *J. Fluid Mech.* **789**, 726–749 (2016).
75. Verma, S., Novati, G. & Koumoutsakos, P. Efficient collective swimming by harnessing vortices through deep reinforcement learning. *Proc. Natl Acad. Sci. USA* **115**, 5849–5854 (2018).
76. Donahue, J. et al. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* 2625–2634 (IEEE, 2015).
77. Bierbach, D. et al. Insights into the social behavior of surface and cave-dwelling fish (*Poecilia mexicana*) in light and darkness through the use of a biomimetic robot. *Front. Robot. AI* **5**, 3 (2018).
78. Hüttenrauch, M., Šošić, A. & Neumann, G. Deep reinforcement learning for swarm systems. *J. Mach. Learn. Res.* **20**, 1–31 (2019).
79. Jones, S., Winfield, A. F., Hauert, S. & Studley, M. Onboard evolution of understandable swarm behaviors. *Adv. Intell. Sys.* **1**, 1900031 (2019).
80. Li, W., Gauci, M. & Groß, R. Turing learning: a metric-free approach to inferring behavior and its application to swarms. *Swarm Intell.* **10**, 211–243 (2016).
81. Halloy, J. et al. Social integration of robots into groups of cockroaches to control self-organized choices. *Science* **318**, 1155–1158 (2007).
82. Rubenstein, M., Cornejo, A. & Nagpal, R. Programmable self-assembly in a thousand-robot swarm. *Science* **345**, 795–799 (2014).
83. Palmer, S. E., Marre, O., Berry, M. J. & Bialek, W. Predictive information in a sensory population. *Proc. Natl Acad. Sci. USA* **112**, 6908–6913 (2015).
84. R., S. & J. R., S. Ecology and physics of bacterial chemotaxis in the ocean. *Microbiol. Mol. Biol. Rev.* **76**, 792–812 (2012).
85. Zahedi, K. & Ay, N. Quantifying morphological computation. *Entropy* **15**, 1887–1915 (2013).
86. Bray, D. Protein molecules as computational elements in living cells. *Nature* **376**, 307–312 (1995).
87. Qian, L., Winfree, E. & Bruck, J. Neural network computation with dna strand displacement cascades. *Nature* **475**, 368–372 (2011).
88. Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl Acad. Sci. USA* **114**, 3521–3526 (2017).
89. Abbe, E. & Sandon, C. Provable limitations of deep learning. Preprint at https://arxiv.org/abs/1812.06369 (2019).
90. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. 31st Int. Conf. Advances in Neural Information Processing Systems* 6402–6413 (NIPS, 2017).
91. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
92. Jakobi, N., Husbands, P. & Harvey, I. In *Advances in Artificial Life* (eds Morán, F. et al.) 704–720 (Springer, 1995).

93. Birattari, M. et al. Automatic off-line design of robot swarms: a manifesto. *Front. Robot. AI* https://doi.org/10.3389/frobt.2019.00059 (2019).
94. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55**, 78 (2012).
95. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Min.* **10**, 35 (2017).
96. Nichols, J. A., Chan, H. W. H. & Baker, M. A. B. Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophys. Rev.* **11**, 111–118 (2019).
97. Hand, D. J. Classifier technology and the illusion of progress. *Stat. Sci.* **21**, 1–14 (2006).
98. Smith, G. *The AI Delusion* (Oxford Univ. Press, 2018).
99. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
100. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning internal representations by error propagation. *Nature* **323**, 533–536 (1986).
101. Lipton, Z. C., Berkowitz, J. & Elkan, C. A critical review of recurrent neural networks for sequence learning. Preprint at https://arxiv.org/abs/1506.00019 (2015).
102. Ho, T. K. Random decision forests. In *Proc. 3rd Int. Conf. Document Analysis Recognition* **Vol. 1**, 278–282 (IEEE, 1995).
103. Oja, E. A simplified neuron model as a principal component analyzer. *J. Math. Biol.* **15**, 267–273 (1982).
104. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybernetics* **43**, 59–69 (1982).
105. Bengio, Y., Courville, A. & Pascal, V. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
106. Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **31**, 651–666 (2010).
107. Xu, D. & Tian, Y. A comprehensive survey of clustering algorithms. *Ann. Data Sci.* **2**, 165–193 (2015).
108. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 2018).
109. Neftci, E. O. & Averbeck, B. B. Reinforcement learning in artificial and biological systems. *Nat. Mach. Intell.* **1**, 133–143 (2002).
110. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
111. Foerster, J. N., Assael, I. A., de Freitas, N. & Whiteson, S. Learning to communicate with deep multi-agent reinforcement learning. In *Proc. 30th Int. Conf. Neural Information Processing Systems* 2137–2145 (NIPS, 2016).
112. Davis, L. *Handbook of Genetic Algorithms* (Van Nostrand Reinhold, 1991).
113. Goodfellow, I. et al. Generative adversarial nets. In *Proc. 27th Int. Conf. Neural Information Processing Systems* 2672–2680 (NIPS, 2014).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to F.C., K.G., B.M. or G.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.