Probing the theoretical and computational limits of dissipative design

Shriram Chennakesavalu and Grant M. Rotskoff

COLLECTIONS

Paper published as part of the special topic on 2021 JCP Emerging Investigators Special Collection

ARTICLES YOU MAY BE INTERESTED IN

Learning intermolecular forces at liquid–vapor interfaces
The Journal of Chemical Physics 155, 164101 (2021); https://doi.org/10.1063/5.0067565

Thermodynamics and kinetics of crystallization in deeply supercooled Stillinger–Weber silicon
The Journal of Chemical Physics 155, 194502 (2021); https://doi.org/10.1063/5.0069475

Reinforcement learning of rare diffusive dynamics
The Journal of Chemical Physics 155, 134105 (2021); https://doi.org/10.1063/5.0057323
Probing the theoretical and computational limits of dissipative design

Shriram Chennakesavalu and Grant M. Rotskoff

AFFILIATIONS
Department of Chemistry, Stanford University, Stanford, California 94305, USA

Note: This paper is part of the 2021 JCP Emerging Investigators Special Collection.

Author to whom correspondence should be addressed: rotskoff@stanford.edu

ABSTRACT
Self-assembly, the process by which interacting components form well-defined and often intricate structures, is typically thought of as a spontaneous process arising from equilibrium dynamics. When a system is driven by external nonequilibrium forces, states statistically inaccessible to the equilibrium dynamics can arise, a process sometimes termed direct self-assembly. However, if we fix a given target state and a set of external control variables, it is not well-understood (i) how to design a protocol to drive the system toward the desired state nor (ii) the cost of persistently perturbing the stationary distribution. In this work, we derive a bound that relates the proximity to the chosen target with the dissipation associated with the external drive, showing that high-dimensional external control can guide systems toward target distribution but with an inevitable cost. Remarkably, the bound holds arbitrarily far from equilibrium. Second, we investigate the performance of deep reinforcement learning algorithms and provide evidence for the realizability of complex protocols that stabilize otherwise inaccessible states of matter.

Published under an exclusive license by AIP Publishing. https://doi.org/10.1063/5.0067695

I. INTRODUCTION
Designing molecular materials that robustly and autonomously assemble into specific, targeted mesoscale structures remains a central challenge in a variety of fields, from materials science to biology. The canonical approach to this design problem is to engineer components with specific molecular interactions that stabilize a thermodynamic ground state corresponding to the target, an inverse approach pioneered in Refs. 9–11. Advances in tunable materials, such as patchy particles, DNA coated colloids, and DNA origami, have enabled the realization of highly specific, directional interactions among the constituent elements and subsequent efforts to optimize these interactions. However, this paradigm is by no means general—in fact, in many instances, self-assembly is driven not by unique and addressable interactions but rather by weak and nonspecific ones. Here, we explore a distinct approach, one based on nonequilibrium external control of a dynamically assembling system, as opposed to designing interactions among the components of a system.

Tailoring self-assembly to produce materials with exotic or desirable properties has long been a goal in the molecular sciences. Early computational work on this topic focused on the inverse design of interaction potentials that produced materials with, for example, target radial distribution functions, densities, and bandgaps. Parameterizing flexible interaction potentials that consistently achieve the desired states is nevertheless challenging. The widespread adoption of machine learning techniques in scientific computing has led many to revisit this problem using more sophisticated numerical representations of optimizable potential energy functions. These techniques have led to significant advances, allowing both for more complex interactions and new computational approaches, increasing the set of structures accessible to designer self-assembly.

At molecular scales, it is often impossible to alter the nature of an interaction without fundamentally changing the molecular components as well, potentially disrupting the biological or chemical function. Rather than viewing self-assembly as a thermodynamic process in which the ultimate structure is determined by a minimum of the free energy, we examine the dynamics of assembly trajectories and ask whether an external agent can perturb the assembling components so as to maximize the yield of a target structure. Because the magnitude of fluctuations at the nanoscale is comparable to the
size of the system itself, the task of determining an effective external protocol for control resembles a stochastic optimal control problem. Just as interaction design is inherently limited by the constituent materials, the precision of control is dictated by the external fields that couple to a given system and the spatiotemporal resolution with which we can reasonably alter these fields. Moreover, an external control approach to directed self-assembly presents new computational challenges: stochastic optimal control problems are typically formulated as high-dimensional partial differential equations, which cannot be solved either analytically or numerically for nontrivial systems. Here, we instead pose the design problem as the optimization of a Markov decision process,\textsuperscript{21} which is, in turn, amenable to deep reinforcement learning algorithms. Of course, these complicated high-dimensional problems have also benefited from advances in deep learning, enabling the optimization of very high-dimensional feedback protocols for essentially arbitrary physical systems.

In this paper, we investigate the theoretical and computational limits of nonequilibrium control in the context of two minimal models of molecular self-assembly. We establish theoretically a relationship between the cost of a protocol, as measured by the entropy production in the medium, and the fidelity with which a target structure can be produced, akin to bounds that have been established for nonequilibrium growth processes.\textsuperscript{22} We then explore the capabilities of deep reinforcement learning algorithms to control the quenched cluster-size distribution of a system of particles using a feedback thermal annealing protocol as well as the steady-state cluster-size distribution of nonequilibrium actively driven colloids. Taken together, our theoretical and computational results emphasize that high-dimensional control can target assembly outcomes with high precision but with an inescapable cost.

II. THE DISSIPATIVE COST OF HIGH-FIDELITY CONTROL

Consider a physical system with coordinates $x \in \Omega \subset \mathbb{T}^d$ evolving according to the overdamped Langevin equation

$$dX_t = b(X_t)dt + \sqrt{2D_d}dW_t, \tag{1}$$

where $b$ is a nonequilibrium drift, $D = k_B T/\mu$ is the diffusion coefficient, and $W_t$ is a Wiener process in $\mathbb{R}^d$. We assume that $X_t$ is ergodic so that there exists a unique stationary probability density $\rho_{ss} : \Omega \rightarrow \mathbb{R}^n$.

Our goal is to develop a feedback-guided, external control protocol $u_t$ that pushes the steady-state distribution toward a specified target. At present, we focus on external driving that can be represented as a force, not a noise term, though we consider both regimes in the subsequent numerical examples. We assume that the external control can be implemented as a spatially varying external force $u_t$ leading to the controlled stochastic differential equation (SDE)

$$dX_t^u = [b(X_t^u) + u_t(X_t^u)]dt + \sqrt{2D_d}dW_t, \tag{2}$$

which, in turn, has an associated steady-state density $\rho_u$. While the most generic design task requires tuning $u$ to coincide with a target steady-state distribution $\rho_\ast$, it is not clear how to specify a target density function for a large interacting particle system, as we typically characterize these systems instead by some low-dimensional observable. This more limited description requires setting some target average value of a given observable $f : \Omega \rightarrow \mathbb{R}$. Let us denote the target value of $f$ by $f_\ast$. The optimal controller then solves the minimization problem

$$u_\ast = \arg\min_u |\mathbb{E}_uf - f_\ast|, \tag{3}$$

where $\mathbb{E}_u$ denotes the expectation over the controlled process (2) and $\mathbb{E}_uf \equiv f_\ast$ denotes the target value of $f$, which we view as the expectation over the unknown target distribution (cf. Appendix A for a detailed discussion).

In some cases, the chosen observable $f$ might not be informative about the system. However, while we seek to carry out this minimization for a particular choice of $f$, if we instead allow $f$ to vary and solve the minimax problem to find the controller that minimizes the mean discrepancy over all functions $g$,\textsuperscript{23} then the objective is the Kantorovich–Rubenstein dual formulation of the Wasserstein-1 distance.\textsuperscript{24} This metric quantifies the distance between the target distribution and the steady-state distribution of the controlled process,

$$W_1(\rho_{ss}, \rho_\ast) = \max_{g \in \mathcal{U}} \min_u |\mathbb{E}_uf - \mathbb{E}_ug|, \tag{4}$$

The Wasserstein distance is an optimal transport distance on probability distributions, measuring the cost to reallocate mass from one distribution to another. Applications of optimal transport distances have become widespread in data analysis and machine learning. See the work of Peyré and Cuturi\textsuperscript{25} for an applied perspective. For the fixed observable of interest, the mean discrepancy is bounded above by the Wasserstein distance, which, in turn, is bounded by the Kullback–Liebler (KL) divergence or relative entropy

$$\min_u |\mathbb{E}_uf - f_\ast| \leq W_1(\rho_{ss}, \rho_\ast) \leq C\sqrt{2D_kL(\rho_\ast \| \rho_{ss})}. \tag{5}$$

The first inequality follows from an application of dual formulation of the total variation distance and subsequently an application of Pinsker’s inequality. Interestingly, this upper bound has a direct physical interpretation in terms of the increase in entropy production of medium when controlling the trajectory to alter the steady-state distribution. Minimizing the Kullback–Liebler divergence between the nonequilibrium distribution of the controlled process and that of the target steady-state requires driving the system away from the uncontrolled steady state $\rho_{ss}$. Because the KL divergence satisfies the triangle inequality, $D_{KL}(\rho_\ast \| \rho_{ss}) \leq D_{KL}(\rho_\ast \| \rho_u) + D_{KL}(\rho_u \| \rho_{ss})$, minimizing $D_{KL}(\rho_\ast \| \rho_{ss})$ will increase $D_{KL}(\rho_u \| \rho_{ss})$. To see this, we note that without control, $\rho_u = \rho_{ss}$. When two probability distributions are equal, the KL divergence between them is zero. So if we assume that $D_{KL}(\rho_u \| \rho_{ss}) > 0$, which is the minimal assumption to make the control problem nontrivial, then note that without
control, $D_{\text{KL}}(\rho_u \| \rho_u) = 0$. The optimization that we wish to solve is one in which we try to minimize $D_{\text{KL}}(\rho_u \| \rho_u)$, ideally so that it vanishes. If the control succeeds in changing the distribution at all, by the positivity of the $D_{\text{KL}}$, the term $D_{\text{KL}}(\rho_u \| \rho_u)$ will increase. If we further succeed in bringing $\rho_u$ closer to $\rho_u$ as measured by the KL divergence, then the triangle inequality implies that $D_{\text{KL}}(\rho_u \| \rho_u)$ must increase by at least $D_{\text{KL}}(\rho_u \| \rho_u) - D_{\text{KL}}(\rho_u \| \rho_u)$, if we assume that this last quantity is positive, meaning that the control has decreased the KL divergence to the target relative to the case without control.

Interestingly, this divergence between the controlled and uncontrolled steady states has thermodynamic interpretation due to the fluctuation theorem for nonequilibrium steady states. In particular, the KL divergence between the controlled and uncontrolled steady states can be bounded by the entropy production in the medium in the long time limit, which can be interpreted as a cost.

Persistently driving the steady state $\rho_u$ from its controlled steady state $\rho_u$ requires energetic input, and such cost have been examined for Markov jump processes and in the contexts related to the design of physical systems. References 22 and 29 arrive at analogous expressions that relate the entropy production in the environment to the divergence between an equilibrium distribution and the distribution driven away from equilibrium by the control. The bound we derive relates directly to the measured observable $f$ and its deviation from the target value. We provide further discussion of this relation and details of the argument above in Appendix A.

### III. REINFORCEMENT LEARNING FOR HIGH-DIMENSIONAL CONTROL PROTOCOLS

A priori, finding the optimal $u$ requires both detailed information about the dynamics of the system and the ability to implement complex interactions. Fortunately, the minimization problem (3) is also amenable to model-free reinforcement learning, an approach we explore here. Moreover, the infimum over protocols $u : \mathbb{R}^d \to \mathbb{R}^d$ requires minimizing with respect to arbitrary many-body force functions that are unlikely to be realizable in experimental settings. We pursue a more pragmatic approach by building the constraints of control directly into our protocol $u$; these constraints are imposed by restricting to a fixed class of experimentally accessible protocols.

We drive the system so that the observable of interest, a cluster-size distribution throughout this paper, matches an externally specified target. We defined the observable by first introducing a map $h : \mathbb{R}^d \to \mathbb{N}^n$, which counts the number of clusters of size $k$ and stores that number in the $k$th entry of an $n$-dimensional vector where $n$ is the total number of particles in the system and hence the maximum cluster size. We then define the normalized histogram of cluster sizes for the configuration $\rho(h(x))$ and measure the discrepancy between this histogram and the target using the Kullback–Leibler divergence,

$$C(x) = D_{\text{KL}}(\rho(h(x)) \| \rho^\ast).$$

This empirical distribution implicitly depends on the external controller $u$ through the sampled state $x$. Importantly, this cost functional makes evaluation of the loss function independent of the particle dynamics, which, in turn, allows for model-free reinforcement learning. Using a cluster-size distribution rather than the average cluster size ensures that the observable robustly describes the target state even when the region of control is large.

Because the objective (8) does not depend explicitly on an unspecified path measure, it is a tractable target for optimization. We consider control functions $u$, depicted schematically in Fig. 1, in which the external control drives a system locally with a fixed spatial patterning. While the protocol is not an arbitrary many-body force, in our case, it remains high-dimensional and $u$ may be a complicated function of the configuration $x$. Despite this complexity, neural networks offer a robust, high-dimensional function representation, which we exploit in our representation of $u$. The steady-state $\rho_u$ distribution depends on the dynamics of the system, so direct, gradient-based optimization of (8) is challenging. In our setup, the duration of the period between protocol updates is
sufficiently long that the gradients of the control parameters become too small to meaningfully optimize the objective by backpropagating through the dynamics. There is also a conceptual reason for employing a model-free optimization algorithm— in experimental settings, we cannot require precise knowledge of the microscopic dynamics of the system to design an external protocol.

Because the optimal \( u \) is time-independent by construction, we use a time-local representation of the joint dynamics of the system and the controller, also known as a Markov decision process. Optimization problems of this type are the basic framework for reinforcement learning and have been studied extensively in the machine learning and control literature. Within this framework, maintaining the optimal steady-state distribution requires incorporating information about the expected future divergence from the target distribution, as measured by \( C \). For a fixed protocol \( u \), the cumulative expected future divergence from the target distribution is an expectation over the dynamics of the system, starting from a given state \( x_t \), which is

\[
\hat{C}(x_t, u) = \mathbb{E}_u \sum_{k=0}^{\infty} y^k C(x_{t+k+1}),
\]

where \( y < 1 \) is a so-called discount factor that ensures that the sum is convergent and gives additional weight to temporally proximate states. Some reinforcement learning algorithms, e.g., policy gradient, use (9) as a direct target for optimization, but the lack of time locality makes gradient-based optimization challenging when the dynamics occurs over long time scales.

Deep reinforcement learning algorithms based on Q-learning lift the expected cost \( \hat{C} \) so that it depends on a given state-action pair \((x_t, a_t)\) and protocol \( u \). We assume that the system evolves with \( u(x_t) = a_t \) constant for a fixed duration \( s \in [t, t+\tau] \) so that \( x_{t+s} \) is obtained by solving (2) with the initial condition \( x_t \); this means that the feedback protocol has a time lag and, in practice, we choose \( \tau \) to be sufficiently short that changes in the cost were minimal. The lifted cost functional, conventionally called \( Q \), quantifies the future cost assuming that the action \( a_t \) is taken at time \( t \) explicitly,

\[
Q^u(x_t, a_t) = \mathbb{E}_u \left[ C(x_{t+s}; a_t) + \sum_{k=1}^{\infty} y^k C(x_{t+s+k+1}) \right],
\]

(10)

In (10), the expectation is carried out over a trajectory initialized at \( x_t \) and subject to the action \( a_t \) until time \( \tau \) and subsequently...
using the protocol $u$. The optimal next action for a given protocol $u$ is then simply $\text{argmin}_a Q^u(x_t, a)$. Employing Q-learning requires estimating $Q$ usually with a value iteration algorithm; traditionally, this has been carried out using a tabular representation of $Q$, meaning that every state-action pair must be visited by the dynamics in order to provide an accurate representation. Of course, if the state or action space is high-dimensional, this is infeasible. Recently, alternative strategies that represent $Q$ as a deep neural network have shown promise in a variety of contexts. Importantly, when $Q$ is represented as a neural network, high-dimensional state and action spaces become tractable. We employ a variant of deep Q-learning $^{15}$ that uses two deep neural networks to approximate the $Q$ function, called double Q-learning, which helps avoid minimization bias in the estimate of the minimizer; we discuss the details of this approach in Appendix B.

IV. CONTROLLING CLUSTER SIZES IN ACTIVE COLLOIDS

To test this reinforcement learning approach to dissipative design, we first considered a model colloidal particle system actively driven by externally controlled light sources. Models of self-propelled or active matter evince rich phase behavior and have rapidly become canonical models for pattern formation out of equilibrium. In these systems, when the Péclet number is sufficiently large, the mobility depends strongly on the local density, which results in nonequilibrium phase separation. $^{34}$ This phenomenon, called motility induced phase separation (MIPS), requires energy consumption and is largely independent of the inter-particle interactions. $^{35,36}$

Because the activity can be externally modulated with simple controls, for example, by selectively illuminating a portion of the system with light of variable intensity, it offers a natural example for control. Reinforcement learning has shown some success in controlling active systems: recently, Falk et al. $^{39}$ examined enhancing transport properties using low-dimensional external protocols optimized with an actor-critic model. The externally modulated activity leads to clustered states, but when this activity is turned off, the particles diffuse apart and the clusters disintegrate. The limitations of this control will necessarily limit the steady-state distributions that can be accessed. In turn, this means that excess dissipation associated with the feedback protocol may enhance control, but it is possible that some of the energy expended goes to waste.

To explore the limits of an activity-inducing dissipative external control, we sought to maintain a steady-state distribution consisting of clusters of particles much smaller than the macroscopic aggregate that forms when there is constant activity. We specified a target distribution $\rho_\ast$ of cluster sizes as in (8) using three distinct target distributions, all with identical mean and variance. The distribution of cluster sizes is discrete, so we first tested a binomial distribution with $n = 20, p = 3/4$. Because this distribution decays rapidly in the tails, we also tested a gamma distribution $\Gamma(k, \theta)$ with shape parameter $k = 25$ and scale parameter $\theta = 3/5$ and computed the corresponding probability mass function by integrating the probability density over the bins. Finally, to remove any effects due to asymmetry of the target distribution, we used a Gaussian target distribution with $\mu = 20, \sigma^2 = 9$. See Figs. 4–9 in Appendixes C and D.

We used deep Q-learning to optimize an external protocol that controlled the intensity of the activity over a spatial grid, as depicted in Fig. 1. Both the cost function and the corresponding action were evaluated locally; that is, the cluster-size distribution was computed over a given region and then the activity was chosen to minimize the estimated $Q$ function, given the observed state. We tested this approach with grids of increasing resolution, corresponding to increasingly fine-tuned spatial control. The optimization was carried out for at least 20 “episodes” of 350 decisions, until the relative entropy between the instantaneous distribution and the target had converged. $^{40}$ Once the protocol had converged, we computed the relative entropy and the mean cluster size over a collection of long test trajectories, shown in Figs. 1(a) and 1(b). When the number of control regions is small (e.g., $4 \times 4$), perturbations to the activity are not localized enough to prevent the formation of large clusters. As a result, the mean cluster size is substantially larger than the target. On the other hand, the benefits of high-resolution control diminish as the number of regions becomes very large ($48 \times 48$). In effect, once a single region can accommodate only a few clusters of the target size, protocols of increasing resolution achieve the same outcome. Furthermore, these high-dimensional protocols come with an additional computational or practical burden.

Protocols with sufficient spatial resolution to execute local control perform well (Fig. 1), leading to mean values for the average cluster size that are close to the target. The variance of the empirical distribution of the steady state under the optimized protocol is larger than the target, emphasizing that our relatively coarse external control is still limited. The timescale over which a cluster diffusively disintegrates in the absence of activity exceeds the time required to form a cluster. Because the decision period has a fixed duration and activity will favor the formation of large clusters, there is a relative abundance of clusters with a size that exceeds the mean. These clusters also contribute to variance in the left tail of the distribution because dissolving them creates a large number of free particles. Nevertheless, as shown in the schematic of Fig. 1, the typical clusters are in line with the target distribution and the formation of macroscopic clusters is always avoided.

We examine the generic relationship between the entropy production and fidelity of control by computing the total entropy production for increasing the resolutions of control. To do so, we compute the entropy production rate per particle, $^{41}$

$$\langle \sigma_m \rangle = \frac{1}{T} \int_0^T \left\langle \left[ b(X^i_t) + u_i(X^i_t) \right] \cdot \sigma \, dX^i_t \right\rangle,$$  \hspace{1cm} (11)$$

where $\langle \cdot \rangle$ denotes an average over particles in the system (indexed in the expression by $i$) and $\cdot$ is the time-symmetric Stratonovich product. This quantity is $\Delta S_m/(NT)$, where $N$ is the number of particles and $T$ is the duration of the trajectory. We averaged this quantity over trajectories with the optimized protocol for a duration of 3000 s for the points shown in Fig. 2. In general, we found that more control regions led to better control as measured by the cost function (8). Figure 2 plots the average entropy production rate $\langle \sigma_m \rangle$ as a function of the KL divergence from the target; each point is averaged over 100 realizations of the optimized feedback protocol.
FIG. 2. The entropy production rate, as defined in Eq. (11), of the controlled system plotted against the fidelity of control as measured by the KL divergence from the target.

As the number of control regions increases from 4 × 4 up to 24 × 24, the fidelity increases but with a clear increase in dissipation, providing evidence of the utility of the bound (5). At the highest resolutions, diminishing returns become evident because the typical cluster size becomes comparable to the region itself—in this regime, the cost function is essentially exactly the mean discrepancy in (3). For other target cluster-size distributions (see Appendix C), we observed a similar trend. We emphasize that the generic trade-off between dissipation and fidelity is not necessarily a monotonic relationship, as observed for different cost functions in Appendix C because not all dissipative dynamics will lead to improvement of the cost.

V. FEEDBACK-GUIDED ANNEALING

While thermal annealing has a long history for macroscopic systems, repeated annealing cycles are an important part of the preparation of a wide variety of nanoscopic materials, from thin films to DNA origami. Annealing is limited as a mechanism for control because there is essentially only one parameter that can be tuned, the rate at which the temperature is decreased. This process will eventually find a global free energy minimum, meaning that the ultimate structure is determined entirely by the thermodynamic properties of the system. Indeed, there has been significant focus on designing interaction potentials that lead to specific structural motifs in a variety of contexts.

We examine an alternative paradigm that exercises more localized control with measurement-guided feedback to design an annealing schedule. Rather than globally tuning a temperature, we locally update the temperature on a grid; see Fig. 3. The specific temperature that the protocol prescribes depends on the local configuration. While the annealing process itself is substantially more complicated in this framework, the approach could be used with arbitrary materials and does not require the realization of highly specific interactions, which can be enormously challenging to engineer in nanoscale systems.

To assess the prospects of this feedback-guided annealing procedure, we studied a minimal example of cluster formation using a 2D Lennard-Jones system. At low temperatures, the free energy minimum corresponds to a single cluster, but at low densities, an instantaneous temperature quench from a high-temperature state will yield a kinetically trapped configuration that consists of small clusters.

We sought to control the distribution of these intermediate clusters by optimizing a thermal annealing function using reinforcement learning. We fixed a target cluster-size distribution, chosen to be a gamma distribution with variance $\sigma^2 = 1$ and a mean $\mu = 4$, shown as a dashed line in Fig. 3. We trained an external controller that used the local cluster-size distribution to determine

FIG. 3. (a) The target distribution of cluster sizes $\rho^\star$ and the empirical distribution obtained using the optimized annealing protocol denoted $\hat{\rho}$. While there is an over-representation of isolated particles, as in the case of the active particle system, the model of the distribution and the tail are well-matched by using the protocol. (b) The evolution of the absolute deviation of the mean instantaneous cluster size $\mu_k(\tau)$ from the mean target size $\mu_k^\star$ for increasing resolution of temperature control. (c) The evolution of the discrepancy between the instantaneous distribution and the target as a function of time for increasing resolution of temperature control.
the subsequent temperature within each region of control. With only coarse control, the annealing reliably produced distributions of cluster sizes with a mean value close to the target [Fig. 3(c)]. However, high-resolution control was required to yield a distribution close—as measured by the KL divergence—to the target distribution. The empirical cluster-size distribution, averaged over 1000 annealing trajectories, is shown in Fig. 3(a) for a 15 \times 15 grid. While there is an over-representation of isolated particles by small amount, the distribution is remarkably close to the target. For fewer regions, the KL divergence [Fig. 3(d)] is substantially larger and the distributions (cf. Appendix D) differ markedly from the target.

Because the external control goes beyond the regime that we treat theoretically in Sec. II, while our analysis extends to systems in which the external control is represented as a drift function \( u \), this system relies on changes to the diffusion tensor, which requires a substantially different mathematical treatment, a topic we plan to explore in future work.

VI. CONCLUSION

Experimental advances enabling high-resolution external control create new opportunities to produce materials with exotic properties. In this work, we seek to address several fundamental questions about the “realizability” of high-fidelity control protocols. In doing so, we derive a general bound that establishes a trade-off between the fidelity of control and the cost of implementing it as measured by the entropy production in the medium.

In many applications, the external fields that could be modulated with high-resolution will be imprecise. That is, we will not necessarily be able to tune a field directly conjugate to the observable of interest. We do not a priori know how to choose observables that robustly approximate the optimal transport distance, which is a significant topic for future work. Despite the arguably “coarse” control we have—we at least compared with the case of interaction design—we nevertheless find that the optimization is successful at producing states with well-defined target mean cluster sizes. That is, our results demonstrate that with appropriate objective functions, reinforcement learning algorithms can identify protocols that closely match target structures even without highly specific interactions. Moreover, though we do not directly design the protocols to show the relation between dissipation and fidelity, we observe a trend consistent with the prediction throughout.

The approach we have pursued raises many additional questions. Foremost among these perhaps is the tightness of the bound (5) for a given observable, a question we hope to examine in model systems in future work. Many other reinforcement learning strategies could be deployed on the systems we studied, including approaches based on policy gradient. Some of these approaches may enable model-free control for more complicated systems, leading to insights about driven self-assembly in complex environments.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to declare.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request. All simulations were conducted using OpenMM 7.5.1 and PyTorch 1.9.0. Our code is available at https://github.com/rotskoff-group/dissipative-design.

APPENDIX A: RELATION BETWEEN COARSE-GRAINED CONTROL AND DISSIPATION

Let \( x \in \Omega \subset \mathbb{R}^d \) denote a configuration on the system, taking coordinates on the torus due to periodic boundary conditions. We set

\[
C = \sup_{x,y \in \Omega} |x - y|, \tag{A1}
\]

the diameter on this compact space. Because it is impractical to directly specify the target distribution \( \rho_s : \Omega \to \mathbb{R} \) for a high-dimensional interacting particle system, we instead specify the target value of a given observable \( f : \mathbb{R}^d \to \mathbb{R} \) and denote this target value by \( f_s \). For technical reasons described below, we assume that \( f \) is Lipschitz continuous (essentially meaning that its derivative remains bounded) with Lipschitz constant \( K = 1 \); any value of \( K \) could be used, and consequently, the bound derived below will have a prefactor of \( CK \). We assume that the dynamics of the system is governed by an overdamped Langevin equation controlled by an external force \( u \), that is,

\[
dX_t^u = [b(X_t^u) + u(X_t^u)]dt + \sqrt{2D}dW_t, \tag{A2}
\]

as discussed in the main text. We also assume that the resulting dynamics is ergodic so that the process (A2) relaxes to a unique stationary distribution \( \rho_s \).

Naturally, the objective function for the controller seeks to match the time-averaged value of the observable \( f \) with the target value \( f_s \). That is, we want to find the optimal controller \( u^* \) that solves

\[
u^* = \arg \min_u \mathbb{E}_u f - f_s \tag{A3}
\]

over all control functions \( u \). Here, the notation \( \mathbb{E}_u \) denotes an expectation

\[
\mathbb{E}_u f = \int_{\Omega} f(x) \rho_u^s(x) dx = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(X_t^u) dt, \tag{A4}
\]

where \( X_t^u \) solves (A2). Although we do not know the functional form of the target steady-state density, we assume that it is “implementable” in the sense that

\[
f_s = \mathbb{E}_u f = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(X_t^u) dt, \tag{A5}
\]

where

\[
dX_t^u = b_s(X_t^u)dt + \sqrt{2D}dW_t, \tag{A6}
\]

\[
dX_t^u = b_s(X_t^u)dt + \sqrt{2D}dW_t, \tag{A6}
\]
for some unknown \( b_a \).

In the approach we take, the optimal controller depends on the choice of observable. Some choices of \( f \) may not be particularly informative about the system—for example, the bulk density could remain fixed for steady-state distributions with differing microscopic structures. However, issues with uninformative observables could be overcome by allowing any 1-Lipschitz function \( g : \Omega \to \mathbb{R} \) and carrying out the “adversarial” optimization

\[
\max \min_{\xi \in \mathcal{G}} |E_u g - E_* g| \tag{A7}
\]

so that the optimal controller must minimize the mean discrepancy for every observable. This stringent requirement actually amounts to minimizing a measure of the distance between the probability distribution with density \( \rho_u \) and the unspecified target distribution \( \rho_* \). In fact, the maximization coincides with the Kantorovich–Rubenstein dual formulation of the Wasserstein-1 optimal transport distance\(^{24} \)

\[
W_1(\rho_u, \rho_*) = \inf_{\Pi(\rho_u, \rho_*)} \int_{\Omega \times \Omega} |x - y| d\Pi(x, y), \tag{A8}
\]

where \( \Pi(\rho_u, \rho_*) \) denotes all distributions with marginals \( \rho_u \) and \( \rho_* \). This so-called integral probability metric is related to other measures of “distance” between probability distribution\(^{14} \) and is upper bounded by the total variation distance

\[
W_1(\rho_u, \rho_*) \leq C|\rho_u - \rho_*|_{TV}, \tag{A9}
\]

which is proved by using the coupling definition of the total variation distance and noting that the expected distance on the coupling that realizes the infimum in \((A8)\) must be less than the maximum distance between points in \( \Omega \) multiplied by the expected fraction of unequal points. Via Pinsker’s inequality, this bound can be extended to the KL divergence between the two distributions,

\[
W_1(\rho_u, \rho_*) \leq C|\rho_u - \rho_*|_{TV} \leq C\sqrt{2D_{KL}(\rho_u || \rho_*)}. \tag{A10}
\]

The KL divergence, or relative entropy, is a well-studied object in nonequilibrium statistical mechanics because it provides a measure of the entropic distance between distributions.\(^{26,27} \) The KL divergence is not a distance because it is not symmetric, but it does satisfy the triangle inequality,

\[
D_{KL}(\rho_u || \rho_*) \leq D_{KL}(\rho_u || \rho_*) + D_{KL}(\rho_*) || \rho_u). \tag{A11}
\]

Because our goal is minimizing \( D_{KL}(\rho_u || \rho_*) \), a successful optimization will increase \( D_{KL}(\rho_* || \rho_u) \). This is due to the fact that, in the absence of control, \( \rho_u = \rho_* \). Let us assume that the stationary distribution of the nonequilibrium steady state with the protocol fixed (that is, all activity values are not varying in time) is given by

\[
\rho_{u*}(x) = e^{-\phi(x_u)}, \tag{A12}
\]

and the function \( \phi \) is generally referred to as a nonequilibrium potential; we cannot typically access it directly as the distribution is non-Boltzmann. For a nonequilibrium system subject to the external control protocol \( u \), the entropy production fluctuation theorem states that at a particular time \( t' > 0 \), \(^{22} \)

\[
\frac{\rho_u(X_{t'})}{\rho^*_u(X_{t'})} = \left\{ e^{-\Delta \phi - \Delta s_m} \right\}, \tag{A13}
\]

where \( \Delta \phi \) is the difference in nonequilibrium potential, \( \Delta s_m \) is the entropy production of the medium, and the notation \( \left\{ \right\} \) refers to averaging in the forward ensemble. Because at \( t = 0 \) we assume that the system is in the uncontrolled steady-state distribution \( \rho_{u*} \), choosing \( t' = t \) implies that

\[
\log \frac{\rho_u(X_t)}{\rho^*_u(X_t)} \geq -\Delta \phi - \Delta s_m, \tag{A14}
\]

where we have used Jensen’s inequality to obtain the bound. Now, because \( \rho^*_u(X_t) = \rho_{u*}(X_t) \) by assumption, we see that

\[
D_{KL}(\rho_u || \rho_*) \leq (\Delta \phi + \Delta s_m). \tag{A15}
\]

In the long time limit, the difference in nonequilibrium potential (which does not grow with time as it is a steady-state quantity) will be dominated by the \( \Delta s_m \), which is time-extensive. As a result, \( \Delta \phi + \Delta s_m \approx \Delta s_m \), which provides a natural upper bound on the steady-state “cost” of driving the system away from its original distribution \( \rho_u \) for a duration \( T \). Bounds of this type have previously appeared in the literature, especially in the context of driving systems away from equilibrium;\(^{25,29} \) note that our bound remains valid if \( \rho_{u*} \) is a nonequilibrium steady state. To relate the fidelity of control to this cost, we compare the empirical KL divergence of the controlled distribution with respect to the target distribution to the entropy production rate of the controlled dynamics.

**APPENDIX B: COMPUTATIONAL DETAILS FOR DEEP Q-LEARNING**

Q-learning is a model-free, off-policy reinforcement learning algorithm used to estimate the optimal state-action value function \( Q^* \). Because Q-learning is model-free, we can avoid incorporating the dynamics of the system into the optimization framework, which may be unknown in experimental systems.

In many RL algorithms, we assume that the environment is a Markov decision process; in our case, this means we consider a dynamics specified by states \( x \in \mathbb{R}^d \), actions of the external protocol \( a \in [0, a_{\text{max}}]_{\text{per cycle}} \), and a cost function \( C \). In deep Q-learning,\(^{35} \) we represent the state-action value function, \( Q \), using a deep neural network and optimize this network using the Bellman dynamic programming principle. For a deterministic policy \( u \), the Bellman equation reads

\[
Q^*(x_t, a_t) = \mathbb{E}_u \left[ C(x_t, a_t) + \sum_{k=1}^{\infty} \gamma^k C(x_{t+k}, u_{k+1}) \right] \tag{B1}
\]
In this expression, the cost is evaluated for a state \( x_{t+1} \) after taking action \( a_t \) and subsequently following the protocol \( u \). In practice, we use a second target neural \( Q' \) network to estimate the predicted value of \( Q' (x_{t+1}, \text{if}(x_{t+1})) \) to ensure more stable updates. We update the weights of the target network to match the weights of the \( Q \) network at a rate \( \tau \). Finally, we use experience replay, a technique in which we store past experiences in a replay buffer \( \mathcal{R} \) and sample from the buffer during update steps. Each experience here is a tuple of \((x_t, a_t, c_t, x_{t+1})\) that was observed within each “grid” (i.e., a spatial region) of control.

Our state space \( \mathcal{S} \) represents the normalized empirical cluster-size distribution over some sample space \( \Omega \), which was specified differently for each of our target distributions, as described in the main text. We found that this representation worked the best

**ALGORITHM 1.** Clipped double Q-learning training.

1: Initialize replay buffer \( \mathcal{R} \) to capacity \( N \)
2: Randomly initialize \( Q_1 \) and \( Q_2 \) networks with weights \( \theta^{(1)} \) and \( \theta^{(2)} \)
3: Initialize target networks \( Q'_1 \) and \( Q'_2 \) with weights \( \theta'^{(1)} \leftarrow \theta^{(1)} \) and \( \theta'^{(2)} \leftarrow \theta^{(2)} \)
4: for \( e = 0 \ldots M \) do
5: \hspace{1em} Initialize state \( X_0 \) for episode \( e \)
6: \hspace{1em} for \( t = 0 \ldots T \) do
7: \hspace{2em} for each \( x_t \) in \( X_t \) do
8: \hspace{3em} if \( x_t \) is empty and \( e \geq e_{\text{expl}} \) then
9: \hspace{4em} Select \( a_t = a_{\text{max}} \) and execute action \( a_t \)
10: \hspace{4em} continue
11: \hspace{3em} end if
12: \hspace{3em} if \( e < e_{\text{expl}} \) then
13: \hspace{4em} Select a random action \( a_t \) from \( \mathcal{A} \)
14: \hspace{3em} else
15: \hspace{4em} Select \( a_t = \arg\min_a Q_1 (x_t, a) \)
16: \hspace{4em} end if
17: \hspace{3em} Execute action \( a_t \) and observe next state \( x_{t+1} \) and cost \( c_t = C \)
18: \hspace{3em} if \( x_t \) or \( x_{t+1} \) is empty then
19: \hspace{4em} continue
20: \hspace{3em} end if
21: \hspace{3em} Store transition \((x_t, a_t, c_t, x_{t+1})\) in \( \mathcal{R} \)
22: \hspace{1em} end for
23: \hspace{1em} Sample random batch of transitions \((x_t, a_t, c_t, x_{t+1})\) from \( \mathcal{R} \)
24: \hspace{1em} Set \( y_j = c_j + \gamma \max_{a_j} \min_{i} Q'_1 (x_{t+1}, a) \)
25: \hspace{1em} Update \( Q_{1,2} \) by minimizing \( L_{1,2} = (y_j - Q_{1,2} (x_t, a))^2 \)
26: \hspace{1em} Update the target networks \( \theta'^{(1,2)} \leftarrow (1 - \tau) \theta'^{(1,2)} + \tau \theta^{(1,2)} \)
27: \hspace{1em} end for
28: \hspace{1em} end for

**TABLE I.** Relevant parameters for Q-learning training.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Active colloids</th>
<th>LJ system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q-learning rate</td>
<td>( 3 \times 10^{-4} )</td>
<td>( 3 \times 10^{-4} )</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
<td>Adam</td>
</tr>
<tr>
<td>Target update rate (( \tau ))</td>
<td>( 5 \times 10^{-3} )</td>
<td>( 5 \times 10^{-3} )</td>
</tr>
<tr>
<td>Batch size</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Discount factor (( \gamma ))</td>
<td>0.9</td>
<td>0.95</td>
</tr>
<tr>
<td>Number of hidden layers</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>( e_{\text{expl}} )</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>( T )</td>
<td>350</td>
<td>150</td>
</tr>
<tr>
<td>( a_{\text{high}} )</td>
<td>1.5</td>
<td>1.0</td>
</tr>
<tr>
<td>( A )</td>
<td>([0.0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5] )</td>
<td>([0.01, 0.25, 1.0] )</td>
</tr>
</tbody>
</table>
compared to other possible state representations, such as images of the system configurations. Each state $x^i_t$ was the normalized cluster-size distribution within a grid of the system $X_t$. This allowed us to utilize local information about the system to spatially control the system.

$Q$-learning methods are known to suffer from underestimation bias. This bias arises during training because we are using $\min_a Q^\pi(x,a)$ as an estimate for $\min_a Q^\star(x,a)$. By Jensen’s inequality, we have

$$E \min_a Q(x,a) \leq \min_a E Q(x,a) = \min_a Q^\star(x,a). \quad (B2)$$

Underestimation bias results in suboptimal actions, which will have an artificially lower state-action value, being selected as the optimal action. One approach to mitigating this bias is to use clipped double $Q$-learning, where we maintain two estimates of $Q^\star$: $Q_1$ and $Q_2$. We then use the maximum of the two estimates provided by these networks as an upper bound of the estimate of the state-action value. With this estimate, we update both $Q_1$ and $Q_2$. Finally, we update the weights of the target networks $Q_1'$ and $Q_2'$ toward the weights of $Q_1$ and $Q_2$ at a rate $\tau$. The parameters used for the Q-learning algorithm (Algorithm 1) are summarized in Table 1.

Reinforcement learning algorithms are constrained by a trade-off between exploration and exploitation. In order to determine an optimal policy $\pi$, it is necessary to exploit our current estimates of the state-action value. However, it is necessary to sufficiently explore the state-action space in order to improve our current estimates of the state-action value. One common approach to this

![FIG. 4. Active colloids. Summary of results for Binomial target distribution. Histograms of cluster sizes for $4 \times 4$ (top left) $12 \times 12$ (top right) $24 \times 24$ (center left) and $48 \times 48$ (center right). Absolute deviation of mean cluster size from mean target size (bottom left), KL cost function (bottom center), and average entropy production rate as a function $\Delta_t$ (bottom right).](image-url)
dilemma is to use an $\epsilon$-greedy search, where a random action is selected with probability $\epsilon$ during training to promote exploration. In our approach, we instead always randomly select an action during the first $e_{\text{explore}}$ episodes of training. After the first $e_{\text{explore}}$ episodes, we decided to train our state-action value function using a greedy approach (i.e., selecting the optimal action that minimizes the cost). We used this approach because of the small size of our action space $A$ and because the decisions made in each grid were considered to be independent experiences. This allowed us to more thoroughly explore the action space during the first $e_{\text{explore}}$ episodes, especially when using higher-resolutions of control.

When using higher-resolutions of control, it is possible for a grid to not contain any particles. During training, if a grid did not contain any particles at time $t$ or $t+1$, it was not included in the replay buffer $R$. While training with the greedy approach (i.e., if the episode $e \geq e_{\text{explore}}$), we set the action to be $a_{\text{max}}$ if the grid did not contain any particles at some time $t$. For our Lennard-Jones system, $a_{\text{max}}$ represented the highest possible temperature in order to ensure that any free particle that entered the region would continue to diffuse until it reached a region with more particles. For our active matter system, $a_{\text{max}}$ represented the highest possible activity to similarly ensure that any free particle that entered the region would continue self-propulsion until it was able to form clusters. Finally, when a cluster was located in multiple grids, each of these grids was considered to contain the entire cluster. This effectively allows neighboring grids that shared a cluster.

![Graphs showing cluster size distributions and performance metrics for different grid resolutions.](image-url)

**FIG. 5.** Active colloids. Summary of results for gamma target distribution. Histograms of cluster sizes for $4 \times 4$ (top left), $12 \times 12$ (top right), $24 \times 24$ (center left), and $48 \times 48$ (center right). Absolute deviation of mean cluster size from mean target size (bottom left), KL cost function (bottom center), and average entropy production rate as a function of $D_{KL}$ (bottom right).
to “communicate,” which was especially important during the high-resolution control cases.

APPENDIX C: ACTIVE COLLOIDS

We modeled our active colloids based on the experimental system in Ref. 38. In this system, colloidal activity can be modulated by blue light. In addition to self-propulsion, these particles exhibit an attractive interaction due to phoretic and osmotic effects when activated by blue light. Results for the active system are shown in Figs. 4–7.

We work in normalized units of the particle diameter and the self-propulsion velocity. The parameters of the model are summarized in Table II. The model consists of a purely repulsive interaction

\[ U_{\text{WCA}}(r) = \begin{cases} 
0, & r > r_{\text{cut}} \\
4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right] + \epsilon, & r \leq r_{\text{cut}}
\end{cases} \]  

(C1)

with \( r_{\text{cut}} = 2^{1/6} \sigma \). In addition to the hard sphere repulsion, there is an

![Graphs and images showing results for Gaussian target distribution. Histograms of cluster sizes, absolute deviation of mean cluster size from mean target size, KL cost function, and average entropy production rate as a function of \( D_{KL} \).]

FIG. 6. Active colloids. Summary of results for Gaussian target distribution. Histograms of cluster sizes for 4 × 4 (top left), 12 × 12 (top right), 24 × 24 (center left), and 48 × 48 (center right). Absolute deviation of mean cluster size from mean target size (bottom left), KL cost function (bottom center), and average entropy production rate as a function of \( D_{KL} \) (bottom right).
attractive force induced by the activity that models hydrodynamic effects,

\[ U_{\text{attract}}(x_i, x_j) = \frac{\sqrt{A_i A_j}}{||x_i - x_j||^2}, \]  

(C2)

where the coefficient \( A_i \) is determined by the instantaneous value of the activity

\[ A_i = \alpha(x_i)^2 A_0 \]  

(C3)

FIG. 7. Active colloids. We compared the predicted actions with a control resolution of 48 \( \times \) 48 between the optimized protocols for Gaussian, binomial, and gamma target distributions. The columns from left to right are the predicted actions for the Gaussian, binomial, and gamma target distributions. From top to bottom, we consider a large cluster, smaller clusters, and a free particle. The optimized protocols for the Gaussian and gamma target distributions favor larger clusters, by selecting a higher activity, compared to that of the binomial target distribution, which selects lower activities, ensuring the breakup of the cluster.
TABLE II. Parameters for the active colloid system.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value (normalized units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_r$</td>
<td>0.125</td>
</tr>
<tr>
<td>$D_t$</td>
<td>0.041667</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>0.5</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>1</td>
</tr>
<tr>
<td>$A_0$</td>
<td>0.87</td>
</tr>
<tr>
<td>$\Delta t$</td>
<td>0.00005</td>
</tr>
</tbody>
</table>

and $A_0$ is a constant. Here, $\alpha(x_i)$ is the predicted “action” by the RL algorithm.

In addition to the conservative force that arises from these potential terms, there is an active force

$$F_{\text{active}}(x) = (\alpha(x) \cos(\theta), \alpha(x) \sin(\theta), 0), \quad (C4)$$

where $\theta$ is a vector of particle directions that itself has a purely diffusive dynamics

$$d\theta_t = \sqrt{2D_r}dW_t, \quad (C5)$$

and $A_0$ is a constant. Here, $\alpha(x_i)$ is the predicted “action” by the RL algorithm.

In addition to the conservative force that arises from these potential terms, there is an active force

$$F_{\text{active}}(x) = (\alpha(x) \cos(\theta), \alpha(x) \sin(\theta), 0), \quad (C4)$$

where $\theta$ is a vector of particle directions that itself has a purely diffusive dynamics

$$d\theta_t = \sqrt{2D_r}dW_t, \quad (C5)$$
APPENDIX D: LENNARD-JONES SYSTEM

To investigate feedback-guided thermal annealing, we modeled a system of colloidal particles that interact via Lennard-Jones interactions,

\[
U_{\text{LJ}}(\mathbf{x}_i, \mathbf{x}_j) = 4\epsilon \left[ \left( \frac{\sigma}{|\mathbf{x}_i - \mathbf{x}_j|} \right)^{12} - 2 \left( \frac{\sigma}{|\mathbf{x}_i - \mathbf{x}_j|} \right)^{6} \right].
\]  

(D1)

Equation of motion:

\[
d\mathbf{X}_i = -\nabla U_{\text{LJ}}(\mathbf{X}_i)dt + \sqrt{2\beta_i^{-1}}(\mathbf{X}_i)dW_i.
\]  

(D2)

The parameters of the model are summarized in Table III. As described, we update the temperature of each grid of the system based on our reinforcement learning approach. The \(\beta_i^{-1}\) for a particle depends on where the particle is at the beginning of a decision and is not changed during the duration of a decision. Because our decision length is only 0.25 s, our particle will not, on average, diffuse between grids within a decision even for our highest resolution of control. At the beginning of the next decision, we instantaneously update the temperature of each grid and subsequently update the \(\beta_i^{-1}\) for each particle depending on the temperature of the grid in which it is located. Results for the Lennard-Jones system are shown in Figs. 8 and 9.

FIG. 9. Lennard-Jones system. We visualized the predicted actions with a control resolution of 15 × 15 of the optimized protocols for a gamma target distribution. For a cluster size at the mean of the target distribution (left), the predicted action is a low temperature, stabilizing the cluster. For a free particle (center), the predicted action is a medium temperature, allowing the cluster to diffuse around the system. For a large cluster (right), the predicted action is a high temperature, resulting in the cluster breaking apart.

REFERENCES


23The set of functions \( g \) satisfies technical assumptions detailed in Appendix A.


40It is difficult to directly compare the duration of training because the replay buffer grows more quickly for systems with more regions of control.


