Understanding and Design of Non-Conservative Optical Matter Systems Using Markov State Models

Journal:	*Molecular Systems Design & Engineering*
Manuscript ID	ME-ART-05-2022-000087.R1
Article Type:	Paper
Date Submitted by the Author:	24-Jun-2022
Complete List of Authors:	Chen, Shiqi; University of Chicago, Chemistry Parker, John; University of Chicago, Chemistry Peterson, Curtis; University of Chicago, Chemistry Rice, Stuart; University of Chicago, The James Frank Institute Scherer, Norbert; University of Chicago, Chemistry Ferguson, Andrew; University of Chicago, Institute for Molecular Engineering

SCHOLARONE™
Manuscripts

An optical matter (OM) system with metal nanoparticles constituents can self-organize into various specific spatial configurations (states) in a laser beam that generates a non-conservative optical force field. In this work, we study the dynamics of OM structural reconfiguration by building Markov state models (MSMs) for those states for different beam powers. To confront the permutational invariant nature of the OM nanoparticles, we employ permutation-invariant nonlinear dimensionality reduction and spectral clustering to perform data-driven identification of the metastable states corresponding to long-lived non-equilibrium OM configurations and the transition rates between them. By constructing MSMs for various powers of the incident laser beam we construct empirical models for the relative stability of the metastable configurations and use these models to discover new beam conditions designed to preferentially stabilize particular OM configurations of interest. Our methodology presents a transferable scheme that can be used to understand, design, and control the dynamics of permutation-invariant systems with conservative or non-conservative force fields prevalent in optical and active matter systems.

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Understanding and Design of Non-Conservative Optical Matter Systems Using Markov State Models†

Shiqi Chen,[a,b] John A. Parker,[b,c] Curtis W. Peterson,[a,b] Stuart A. Rice,[a,b] Norbert F. Scherer,*[a,b] and Andrew L. Ferguson*[d]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Optical matter (OM) systems consist of nano-particle constituents in solution that, when illuminated with a laser beam, can self-organize into ordered arrays bound by electrodynamic interactions. OM systems are intrinsically non-equilibrium due to the incident electromagnetic flux and may manifest non-conservative forces and interconversion among structural isomers. Rational design of desired configurations and transitions requires quantitative understanding of the relation between the incident beam and the emergent metastable states and isomerization dynamics. We report a data-driven approach to build Markov state models appropriate to non-conservative and permutation-invariant systems. We demonstrate the approach in electrodynamics-Langevin dynamics simulations of six electrodynamically-bound nanoparticles. The Markov state models quantify the relative stability of competing metastable states and the transition rates between them as a function of incident beam power. This informs the design and testing of new beam conditions to stabilize desired nanoparticle configurations. The technique can be generalized to understand and control non-conservative and permutation-invariant systems prevalent in optical and active matter.

1 Introduction

The self-organization of nanoparticles (e.g., gold, silver, silicon, etc.) to fabricate metamaterials is a promising approach to create new functional materials[1–11]. Doing so requires knowledge of and control over the interactions between constituent elements. Liquid crystals, a well-studied example where anisotropic molecules can be manipulated to form ordered phases that can be controlled by temperature and/or external fields, have become a pillar of the information technology revolution by way of their integration into displays and other devices.[12] Given the tremendous impact of these materials for specific applications, it is highly desirable to create new classes of self-organizing materials with engineered properties.

This goal requires using and expanding the principles of physical chemistry and condensed matter physics. If we envision material formation as a kinetic assembly process, then the averaged microscopic dynamics underlying macroscopic rate laws can be inferred from ensemble (e.g., spectroscopic or far-field flux) measurements.[13–17] However, such measurements cannot provide information about individual particle motions and the forces leading to them.

Optical matter (OM) is a class of materials formed by self-organization of its particle constituents into ordered structures.[18–20] OM structures form in focused optical beams (i.e., optical traps or tweezers),[21,22] without explicit external control of particle positions. Once the nanoparticles (e.g., Au, Ag, Si, etc.) and optical beam properties are chosen, the many body electrodynamic interparticle interactions and forces established in the system generate structures. The electrodynamic interactions, termed optical binding, range from a few to many k_BT units of thermal energy, so the OM structures can undergo structural rearrangements. Fig. 1 shows the transformation of the 6-particle OM system among its three most stable structures resolved by dark-field optical microscopy microscopy.

OM systems are fundamentally non-equilibrium due to a continuous flux of optical beam power through the material. The electrodynamic interactions amongst the nanoparticles are complex but controllable. Optical beam shape, the polarization of the light, the spatial phase profile of the optical field,[23–27] as well as the nature (e.g., elemental makeup, dielectric, etc.) and shape (e.g., spheres, ellipsoids, rods, wires, cubes, platelets, etc.)[28–32] of the constituent particles can all be selected or tuned, allowing one to explore a large space of pairwise and many-body interac-

[a] Department of Chemistry, University of Chicago, Chicago, Illinois 60637.; E-mail: nfschere@uchicago.edu
[b] James Franck Institute, University of Chicago, Chicago, Illinois 60637.
[c] Department of Physics, University of Chicago, Chicago, Illinois 60637.
[d] Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637.; E-mail: andrewferguson@uchicago.edu

† Electronic Supplementary Information (ESI) available: [Supplementary Information of Building a Markov State Model for Optical Matter Systems]. See DOI: 00.0000/00000000.

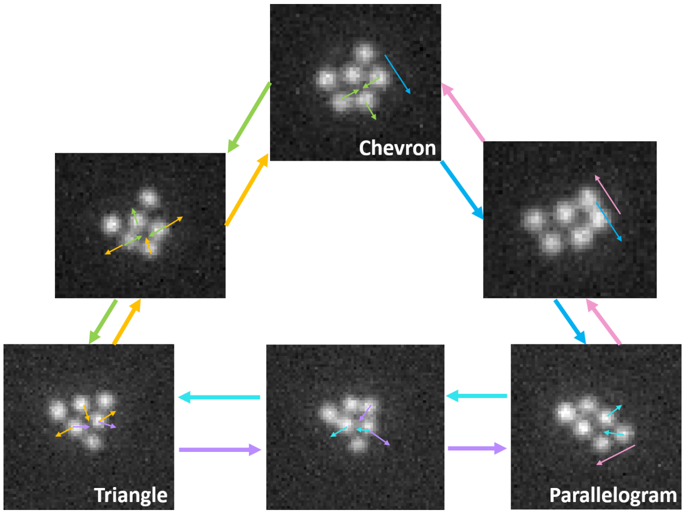

Fig. 1 Dark-field microscopy image of optical matter structures of 150 nm diameter silver particles. Six silver nanoparticles were drawn into a focused Gaussian beam (i.e., optical trap). Most nearest neighbor interparticle spacings are around 600 nm while some are closer due to near-field interactions. The chevron, triangle, and parallelogram are three commonly observed stable structures, while the other three structures correspond to transition states along the structural transformation coordinates. The colored arrows in the dark-field images sketch the structure change of the transition depicted by the arrows in the same color between the dark-field images.

tions. This large parameter space makes accessible a wide range of phenomena (e.g., dynamics of self-assembling and driven active matter, negative torque, spectroscopy of collective excitations, etc.).[32–35] Recent work has shown that the dynamic behavior of OM arrays is related to their shape and symmetry.[30,36–43] Light scattering from nanoparticle arrays can bring about unusual phenomena such as non-reciprocal forces,[40,43] negative torque,[44–46] and non-conservative forces.[47]

While understanding of how incident fields can affect the dynamics and structure of OM arrays is improving[34,48,49], and efficient computer simulation techniques exist to model OM dynamics[50–54], we still lack quantitative theoretical models of how to modulate the properties of the particles and incident light to stabilize particular desired OM states and transitions. This presents an opportunity for the development of data-driven models to learn empirical mappings from the system properties to the emergent dynamical behaviors and inform the design of steady-state and dynamic control strategies to stabilize various OM (non-ground state) structures or to drive OM isomerization.

The primary objective of the present work is to devise an analytical scheme with which we understand and control the dynamics and metastable states in numerical simulations of 6-particle OM systems. In order to achieve this, we build Markov state models (MSMs) as a powerful approach to infer long-time kinetic models from simulation trajectories.[55–59] In order to build a MSM, we have to carry out featurization and clustering analysis on the simulation trajectory and convert the configuration trajectories to trajectories of cluster labels that are the direct input of MSM construction. Due to the permutation symmetry resulting from the fungible nature of identical particles, the featurization procedure must take permutation invariance into account. Our analysis uses a permutation invariant pairwise metric that we supply as the kernel to perform nonlinear dimensionality reduction using diffusion maps[60]. We use diffusion k-means[61] to define the microstate clustering, and Robust Perron Cluster Cluster Analysis (PCCA+)[62–66] to define the macrostate clustering and build MSMs. We test the Markovianity of the macrostate MSMs using the Chapman-Kolmogorov (CK) test to verify that they are valid kinetic models of the non-equilibrium OM system and provide *post hoc* support for the use of permutationally-invariant diffusion map embeddings to identify and resolve microstates. By constructing MSMs at a variety of beam powers we quantify how the intensity of the incident light controls the relative stabilities of and transition rates between the metastable OM configurations. We then use these models to guide the design and testing of new beam conditions to preferentially stabilize desired OM states.

2 Methods

2.1 Langevin dynamics simulations of optical matter

The dynamical evolution of the OM system can be modeled by combining a finite-difference time-domain (FDTD) solution of the electrodynamic forces with a Langevin dynamics equation of motion for the particle positions.[50] However, this is insufficiently efficient to access experimental timescales. Therefore, we developed an electrodynamics-Langevin dynamics (EDLD) approach based on generalized multiparticle Mie theory (GMMT).[51,52] The resulting EDLD solver performs a numerical Verlet integration of the following Langevin equation,

$$m\frac{d^2\boldsymbol{r}}{dt^2} = \boldsymbol{F}(\boldsymbol{r},t) - \lambda_v\frac{\boldsymbol{r}}{dt} + \boldsymbol{\eta}, \qquad (1)$$

where m is the nanoparticle mass, $\boldsymbol{r}$ is its location, t is time, $\lambda_v = 6\pi\nu R$ is the friction coefficient specified by Stokes' Law where R is the nanoparticle radius and ν is the viscosity of the surrounding fluid, $\boldsymbol{\eta}$ is a stationary Gaussian random variable with zero mean and a standard deviation that satisfies the fluctuation-dissipation relation at the specified temperature, and $\boldsymbol{F}$ is the net force experienced by the particle comprising electrodynamic contributions computed from GMMT and electrostatic and interparticle interactions. The force field $\boldsymbol{F}$ computed is non-conservative due to the input power from the optical beam. The steady states reached by the system correspond to states in which the power input from the optical beam is balanced by frictional dissipation into the medium. Simulations are conducted using the MiePy software developed in the Scherer group.[53] Calculations on 28 × 2.4 GHz Intel E5-2680 v4 CPUs execute 2.5 seconds of simulation time per hour of wall clock time with a 5 μs simulation time step.

The simulation is carried out for six spherical silver nanoparticles with 150 nm diameter under several beam powers ranging from 40 mW to 90 mW. For each beam power, we simulate 100 trajectories, each of which is 100 seconds in length. Data is collected every 200 simulation time steps, so the frame-wise time step is 1 ms. We assume that the solvent medium is water (refraction index $n_b = 1.33$, viscosity $\eta = 8\times10^{-4}$ Pa·s), temperature $T = 300$ K, the beam width $w = 2.5$ μm, the wavelength $\lambda = 800$ nm. Based on our experience with previous EDLD simulations using the MiePy package,[53] we add a defocus equal to the Rayleigh range, $z = 0.5kw^2$ where $k = 2\pi n_b/\lambda$, and an electrostatic double layer potential with particle surface potential 77 mV and Debye screening length 27.6 nm according to previous theoretical work[25].

2.2 Nonlinear manifold learning

Before building the MSM and computing the transition rates between the structural states, we need to perform featurization and clustering analysis on the data set of configuration trajectories. The first step in the MSM construction pipeline is to project the simulation trajectories into their leading slow modes to define a low-dimensional embedding conducive to identification of the metastable states of the system using clustering algorithms.[56]. This is typically achieved using time lagged independent components (tICA) analysis[67] or its kernel[68] or deep[69–71] variants. However, there are technical challenges in applying these methods to systems exhibiting full permutation symmetry, such as the OM system, where all particles are identical. Therefore, in the featurization procedure, we instead use diffusion maps[60,72–74], a nonlinear manifold learning method that can generate permutation invariant coordinates for clustering. At first glance, this is possibly problematic since diffusion maps are designed to identify high-variance as opposed to slow collective modes and thus may not provide optimal embeddings for clustering into metastable states. However, the diffusion map is dynamically meaningful and the eigenfunctions are known to be identical to those of the Langevin equation for conservative systems equipped with the common Euclidean distance without permutation invariance.[75] There is no known proof that this eigenvector correspondence continues to hold for non-conservative systems with permutation symmetry, but we conjecture that the leading diffusion map eigenvectors may nevertheless provide approximations for the slow modes of the Langevin equation governing the OM dynamical evolution and may therefore offer a useful embedding for clustering the metastable states. We provide *post hoc* support for this conjecture by validating that the MSMs generated using diffusion map embeddings are valid kinetic models that pass all the numerical tests of Markov properties and convergence of implied time scales.

2.2.1 Pairwise distance calculation

Let $(x_p^{(i)}, y_p^{(i)})$ denote the 2D Cartesian coordinates of particle p in configuration i. We can calculate the distance matrix $\boldsymbol{M}^{(i)}$ for each configuration i with matrix elements,

$$M_{pq}^{(i)} = \sqrt{\left(x_p^{(i)} - x_q^{(i)}\right)^2 + \left(y_p^{(i)} - y_q^{(i)}\right)^2}. \tag{2}$$

Let $\boldsymbol{e}_k$ denote the unit column vector with the k^{th} component unity and others zero. Then the permutation-invariant distance defined between a pair of configurations i and j is,

$$d_{ij} = \min_{\boldsymbol{P} \in S_n} \sqrt{\sum_{k=1}^{N} \min_{\boldsymbol{Q}_k \in S_n} \left\| \boldsymbol{Q}_k \boldsymbol{M}^{(i)} \boldsymbol{e}_k - \boldsymbol{M}^{(j)} \boldsymbol{P} \boldsymbol{e}_k \right\|_2^2}, \tag{3}$$

where N is the number of particles, and S_n is the set of all permutation matrices so that $\boldsymbol{P}$ and $\boldsymbol{Q}_k$ are the optimal permutation matrices that minimize d_{ij}. Here, $\boldsymbol{P}$ and $\boldsymbol{Q}_k$ are $(N+1)$ independent permutation matrices to be optimized, in which $\boldsymbol{P}$ corresponds to the inter-column permutation while $\boldsymbol{Q}_k$ corresponds to the intra-column permutations for all the columns so that the norm of the difference of $M^{(i)}$ and $M^{(j)}$ is optimized over all inter-column and intra-column permutations. Then d_{ij} is a permutation-invariant pairwise distance for the configurations that serves as a kernel for the diffusion map calculations.

2.2.2 Diffusion maps

Diffusion maps are a type of non-linear manifold learning method that take the input of pairwise distances of the configurations and generate a low-dimensional non-linear subspace of the configuration space.[60,73] A brief introduction to diffusion map methodology is provided below, while full details of this method applied to colloidal self-assembly are discussed in previous work.[76–78]

First, the kernel matrix $\boldsymbol{K}$ is calculated with elements,

$$K_{ij} = \exp\left(-\frac{d_{ij}^2}{2\varepsilon^2}\right), \tag{4}$$

where d_{ij} is the permutation-invariant pairwise distance defined previously and ε is the kernel bandwidth parameter that characterizes the adjacency among the configurations. Next, $\boldsymbol{K}$ is normalized to $\tilde{\boldsymbol{K}}$ to gain correspondence to the Langevin dynamics[75,79],

$$\tilde{K}_{ij} = \frac{K_{ij}}{\sqrt{\left(\sum_k K_{ik}\right)\left(\sum_k K_{kj}\right)}}. \tag{5}$$

$\tilde{\boldsymbol{K}}$ is then used to calculate the right-stochastic Markov transition matrix (RSMTM) $\boldsymbol{T}$,

$$T_{ij} = \frac{\tilde{K}_{ij}}{\sum_{j'} \tilde{K}_{ij'}}, \tag{6}$$

with eigenvalues $\{\lambda_k\}$ and right eigenvectors $\{\psi_k\}$. Since the components of ψ_1 are all unity, $\{\psi_k\}_{k=2}^{m+1}$ is taken as the basis of the low-dimensional nonlinear configuration subspace. An appropriate value of m is identified based on a gap in the eigenvalue spectrum. Finally, we obtain,

$$\left\{\left(x_p^{(i)}, y_p^{(i)}\right)\right\}_{p=1}^{N} \longrightarrow \{\psi_k(i)\}_{k=2}^{m+1}, \tag{7}$$

which maps the Euclidean coordinates of each configuration to its corresponding diffusion map embedding. After obtaining this m-dimensional permutation-invariant reduction, configurations are clustered into microstates.

2.2.3 Nyström extension

The time and memory complexity of diffusion maps scale quadratically[80] with the number of data points n. The Nyström extension is an out-of-sample extension technique that scales linearly with n and can be used to embed a new point to a pre-existing diffusion map embedding.[81–84] In this case, we can choose n' ($n' < n$) representative data points (termed "pivots") from the trajectory, calculate the diffusion map on these n' points, and then use Nyström extension to embed the remaining $(n - n')$ points. The pivot points must cover the configuration space of the entire data set so that all points to be embedded are within ε (the kernel bandwidth) of at least one pivot point to assure that the new points are accurately interpolated[85,86]. This approach is known as pivot diffusion maps[85]. Given a new point and the n'-point diffusion map previously constructed, pivot diffusion maps compute the distance between the new point and the n' existing points $\{d_{0,j}\}_{j=1}^{n'}$ where subscript 0 denotes the new point. Next, we compute and append a new row to the kernel matrix $\boldsymbol{K}$ corresponding to the new point, apply the Langevin normalization, and then calculate the corresponding RSMTM row vector,

$$K_{0j} = \exp\left(-\frac{d_{0j}^2}{2\varepsilon^2}\right), \quad \tilde{K}_{0j} = \frac{K_{0j}}{\sqrt{\left(\sum_k K_{0k}\right)\left(\sum_k K_{kj}\right)}}, \quad T_{0j} = \frac{\tilde{K}_{0j}}{\sum_{j'} \tilde{K}_{0j'}}, \tag{8}$$

and the embedding of the new point is given by,

$$\psi_k(0) = \frac{1}{\lambda_k} \sum_{j=1}^{n'} T_{0j}\psi_k(j), \quad k = 2,3,...,m+1. \tag{9}$$

The representative set of the n' pivot points is generated to assure good coverage of the configurational phase space. First, we perform EDLD simulations for 70 mW beam power using a periodic temperature profile with a period of 7000 simulation time steps: 2000 steps at 300 K and 5000 steps at 100 K. The pairwise distances among the particles of the last configuration of every 1000-step segment are computed. The total number of degrees of freedom is $(2N-3)$, where we have N particles moving in the plane subject to two translational constraints and one rotational constraint. We impose a condition in order not to get too far away from the relevant portion of configurational space: if the number of the pairwise distances less than 1.5 optical wavelengths in the medium (i.e., 900 nm) is less than $(2N-4)$, the simulation is restarted. This is because the number of pairwise distances that are less than 1.5 optical wavelengths can be regarded as the number of particles pairs that are at the first optical binding sites of each other. Since the total number of degrees of freedom is $(2N-3)$, at most $(2N-3)$ first optical bindings can be formed. Just as formation of bonds lowers the potential energy of molecular systems, the more first optical bindings the more stable the OM structure is. If the number of first optical bindings is less than $(2N-4)$, the structure is not stable. We apply 13 temperature cycling periods in a single simulation and 100 single simulations are carried out in parallel. Next, we iterate through all the configurations in the trajectories and add them one by one to a pivot set in which all pairwise distances among the configurations are larger than 470 nm. Then, we repeat the simulation and addition of points to the pivot set 15 times before the number of points in the pivot set converges to include 545 configurations. We enrich these pivots with 19,500 configurations randomly selected from the simulation trajectories to form the terminal pivot set. We have verified that the pivot set constructed according to this procedure provides complete coverage of this space such that all remaining data points lie within ε of at least one pivot point.

2.2.4 Density-adaptive diffusion maps

Diffusion maps may not simultaneously resolve the region of configuration space with high density of points and the sparse connectivity region with low density of points. The density adaptive variant of diffusion maps was developed to address this challenge.[87] Instead of using the distance directly in the Gaussian kernel in eq. 4, this method parameterizes the kernel matrix elements as,

$$K_{ij}(\alpha) = \exp\left(-\frac{d_{ij}^{2\alpha}}{2\varepsilon^2}\right) = \exp\left(-\frac{d_{ij}^2}{2\left[\varepsilon d_{ij}^{(1-\alpha)}\right]^2}\right). \tag{10}$$

In effect, the kernel bandwidth ε is scaled according to the pairwise distance by a factor $d_{ij}^{1-\alpha}$. When $\alpha = 1$, we recover the original diffusion map with a constant kernel bandwidth. When $\alpha = 0$, the kernel bandwidth is proportional to the pairwise distance between any pair of points and K_{ij} becomes a constant value for any pair of configurations (i, j). Clearly, α should be chosen from $(0, 1]$ to make the diffusion map adapt to the density of the configuration space. In this work, we choose $\alpha = 0.1$ and $\varepsilon = 2$ nm$^{0.1}$, which is motivated by the guidelines based on the work of Wang and co-workers[87]. We provide post hoc validation that this choice of α and ε generates diffusion map embedding and clustering leading to MSMs that pass all of our numerical validations. The embedding plot of the density-adaptive diffusion map for the 50 mW beam power simulation data is shown in Fig. S1 in the ESI†. As discussed in Section 3, the diffusion map embedding provides good discrimination between the metastable macrostates.

2.3 Diffusion k-means clustering into microstates

The k-means clustering algorithm is a widely-used unsupervised clustering method.[88] Chen and Yang introduced diffusion k-means, which maximizes the within-cluster connectedness based on the diffusion distance.[61] The diffusion distance is defined as the Euclidean distance in the eigenvector space of diffusion map embedding.[60,72,73] In other words, diffusion k-means is k-means clustering applied to diffusion map embeddings. In the present work, diffusion k-means is used as the microstate clustering algorithm so the clusters generated by diffusion k-means are termed "microstates" while the clusters generated by Robust Perron Cluster Cluster Analysis (PCCA+)[64–66] are called "macrostates". The eigenvectors of a diffusion map correspond to different eigenvalues that characterize the time scale of transition between macrostates. Therefore, instead of executing k-means clustering directly on the basis set $\{\psi_k\}_{k=2}^{m+1}$, we execute it on the basis set $\{\lambda_k^{\tilde{t}}\psi_k\}_{k=2}^{m+1}$, where λ_k is the eigenvalue corresponding to ψ_k and $\tilde{t}$ is a parameter that characterizes the time scale of diffusion distances encountered in the k-means clustering[73]. As $\tilde{t}$ becomes larger, the eigenvectors with large eigenvalues become more important in the clustering, leading to merging of regions discriminated by higher order eigenvectors and discriminating microstates largely on the basis of the leading eigenvectors. (On the other hand, when $\tilde{t}$ is small or even negative, the eigenvectors with small eigenvalues become important, leading to microstates as well as macrostates connected by fast transitions and merging the regions connected by slow modes.) Here, we employ an empirical procedure to select $\tilde{t}$ such that our k-means clustering identifies the long-lived macrostates and that the results are not sensitive to the precise value of this parameter.

2.4 Clustering into macrostates and Markov state model

Each MSM is built upon a data set containing 100 trajectories that are 100 seconds long and contain 100,000 configurations for a specific optical beam power. The MSM is built using the PyEMMA software (http://www.emma-project.org/latest/).[89] We use diffusion k-means with $k = 1000$ and maximum iteration number of 200 in the microstate clustering. The microstates are clustered into macrostates using the Robust Perron Cluster Cluster Analysis (PCCA+) algorithm[64–66] that determines the stable states of the OM system. Mathematically, we construct the elements of the microstate transition matrix $\mathbf{\Gamma}$ as $\Gamma_{ij} = c_{ij}(\tau)/\sum_k c_{ik}(\tau)$, where $c_{ij}(\tau)$ are the counts of transition events between microstates i and j at a lag time of τ.[90] Due to the non-conservative nature of OM systems, $\mathbf{\Gamma}$ is not guaranteed to obey detailed balance and therefore may not possess strictly real eigenvalues and eigenvectors as required by the PCCA+ clustering algorithm employed to cluster microstates into macrostates.[90] As such, we adopt the conventional pragmatic solution of symmetrizing the count matrices under the operation $c_{ij}(\tau) \leftarrow \frac{1}{2}\left(c_{ij}(\tau) + c_{ji}(\tau)\right)$ to enforce detailed balance within $\mathbf{\Gamma}$.[90] Physically, this corresponds to collating counts over the forward and reverse trajectories, although we observe more sophisticated techniques based on likelihood maximization and Koopman reweighting also exist.[90] We perform this "reversibilization" to furnish real eigenvectors as required by PCCA+, but since the OM systems are non-equilibrium and therefore not constrained to obey detailed balance, there is a concern that the reversibilized eigenvectors may substantially deviate from, and therefore not be representative of, those of the original non-reversibilized system. We test this by computing the cosine similarity between the first 10 eigenvectors of the reversibilized and original microstate transition matrices for the 50 mW beam power and illustrate in Fig. S2 in the ESI† that they are very similar. The reversibilization procedure is justified since a PCCA+ clustering into $(n_M + 1)$ macrostates uses only the leading n_M ($n_M = 6$ in the current case) right eigenvectors. Finally, we estimate our macrostate MSM by computing count matrices and (non-reversibilized) transition matrices from our simulation trajectory data over the macrostates. Importantly, since we do not enforce detailed balance within the macrostate MSM, the microstate reversibilization procedure may be viewed purely as a means to aid in the definition of the macrostates and which has no bearing on the subsequent specification of the macrostate transition kinetics. We observe that the macrostate count matrix is naturally very close to a symmetric matrix whereas the microstate count matrix is substantially asymmetric. An analysis of the count matrix symmetry is discussed in the ESI† below Fig. S2.

There are three key hyperparameters within this protocol that must be manually selected: the parameter $\tilde{t}$ within the diffusion k-means clustering; the number of macrostates n_M within the PCCA+ clustering; and the lag time τ of the MSM. We self-consistently specify these hyperparameters by analyzing the assignment matrices and implied time scales as described below.

3 Results and Discussion

We now proceed to construct MSMs for our 6-particle OM system as a function of beam power. These MSMs represent data-driven models that we can use to predictively link beam power to the identity and stability of emergent macrostates of the OM system and can be used to guide the design of beam powers to preferentially stabilize desired microstates and transitions. We illustrate the hyperparameter tuning procedure for the MSM fitted to 6-particle OM system for a 50 mW beam power. Analogous protocols are followed for other beam powers considered.

3.1 MSM hyperparameter optimization

To execute our analysis, we must first tune the hyperparameters $(\tilde{t}, n_M, \tau)$. Fig. 2(a) shows the implied time scale (ITS) plot for the MSM constructed for the 50 mW 6-particle OM system with $\tilde{t} = 8.3$. According to this implied time scale plot, when $\tilde{t} = 8.3$, five implied time scales can be resolved at a lag time $\tau = 10$ ms, meaning that $n_M = 6$ macrostates are identified for that lag time. Fig. 2(b) shows the silhouette score[91] plotted against the number of clusters n_M, in which $n_M = 4$ has the largest score and $n_M = 6$ the second largerst. However, according to Fig. 2(a), at least four time scales are resolved at 10 ms lag time, so the number of macrostates should be greater than or equal to five. Therefore, we select $n_M = 6$.

In order to evaluate the clustering results, we compare the predictions of the MSM metastable states with physical intuition. We know that OM structures with particles on hexagonal lattice sites typically form when the incident optical trapping beam is circularly polarized.[28] Therefore, given a certain OM structure with N particles we seek the best set of N sites on a 2D hexagonal lattice (i.e., a lattice pattern such as triangle, chevron, parallelogram, etc.) that is closest to the given configuration. We then categorize the particle configurations using the corresponding lattice patterns. The details of this lattice fitting method have been reported elsewhere.[54]

We plot the Frobenius norm (F-norm) of the row-normalized assignment matrix (RNAM) for $\tau = 10$ ms with n_M and $\tilde{t}$ varying; the results are shown in Fig. S3 in the ESI†. The (i, j) entry of the assignment matrix is the number of frames that is put into the i^{th} macrostate by the clustering method while classified into the j^{th} lattice pattern by lattice fitting. This matrix displays the matching relation between the clustering result and the lattice fitting result. Then, the RNAM can be generated by dividing each row by its row sum. (Analogously, the column-normalized assignment matrix (CNAM) can be generated by dividing each column by its column sum.) We can see from Fig. S3 in the ESI† that the F-norm of the RNAMs converges as $\tilde{t}$ increases, indicating the stable identification of metastable clusters by diffusion k-means for sufficiently large $\tilde{t}$. It is clear that $\tilde{t} = 8.3$ lies within the converged region. The F-norm of the RNAM increases when n_M increases since there are more entries in RNAM, but this trend exhibits a knee at $n_M = 6$ and begins to fail to stably resolve a sufficient number of modes

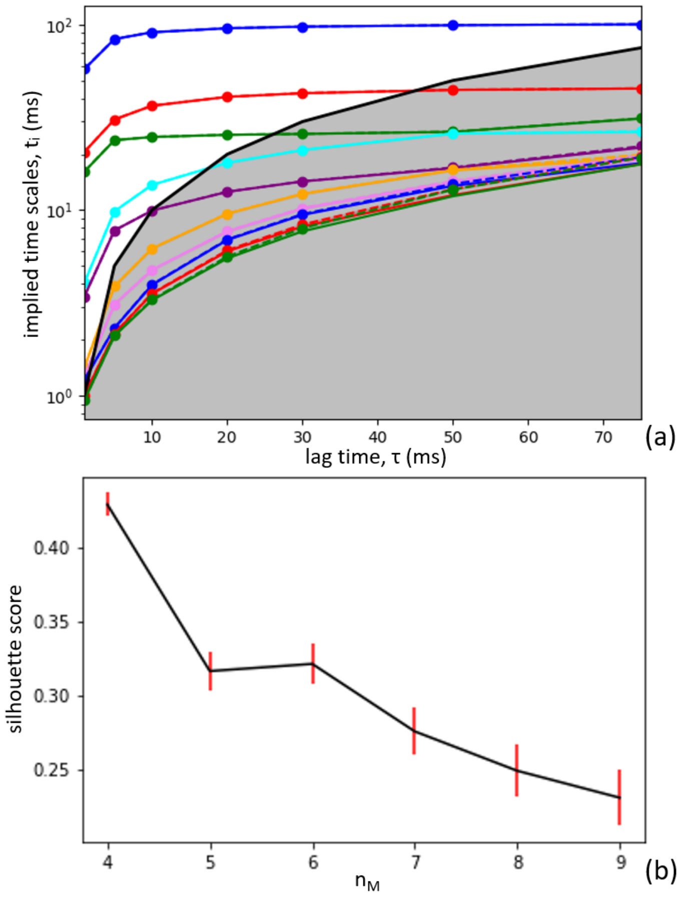

Fig. 2 Analysis used for determination of number of macrostates at $\tilde{t} = 8.3$ and $\tau = 10$ ms for the 6-particle OM system under 50 mW beam power. (a) Implied time scale plot. The shaded grey area demarcates the region where the lag time exceeds the implied time scale. Implied time scales falling into this region cannot be distinguished within the time resolution of the resulting MSM. (b) Silhouette score plot against number of clusters n_M.

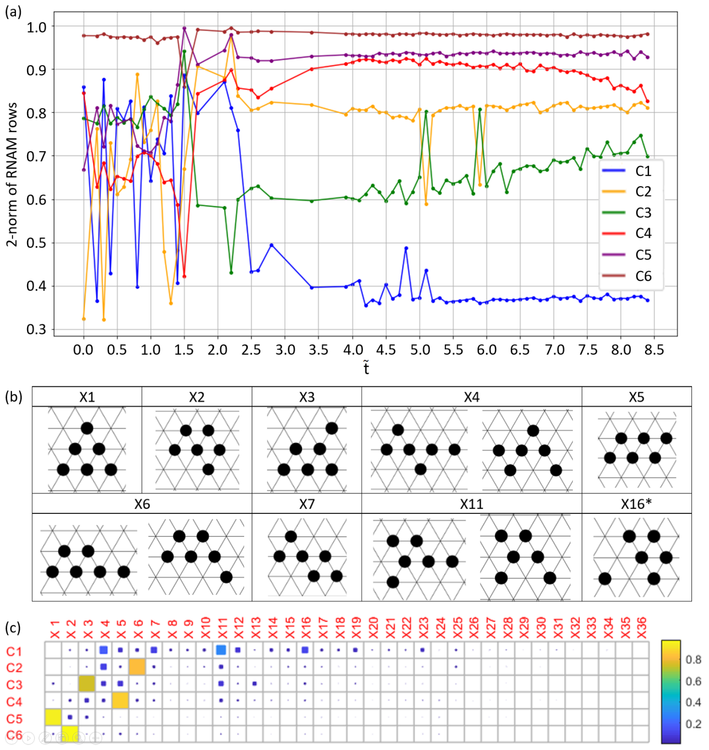

Fig. 3 Details and interpretation of the clustering result for the MSM constructed at 50 mW beam power. (a) Plot of 2-norms of rows of the row-normalized assignment matrix (RNAM) against the diffusion k-means parameter $\tilde{t}$ for $n_M = 6$ and $\tau = 10$ ms. (b) The lattice patterns of some important lattice labels. The asterisk on X16 means that it corresponds to more than two lattice patterns where one particle is separated from the other five particles that are gathered together. (c) Illustration of the RNAM illustrating the assignment probabilities of each hexagonal lattice pattern (columns, X1-36) to each of the six macrostates within the learned MSM (rows, C1-6). The pattern of matrix elements indicates that C2-6 are high-purity macrostates comprising of largely a single lattice label, whereas C1 contains a mixture of lattice labels.

for $n_M > 6$, leading us to select $n_M = 6$ for the number of clusters. From this analysis, we identify $(\tilde{t}, n_M, \tau)$ = (8.3, 6, 10 ms) is a reasonable and robust tuning of the three hyperparameters for 50 mW beam power. As a final *post hoc* validation, we return to Fig. S1 to observe that the diffusion map embedding nicely distinguishes and separates the various macrostates and that the macrostates are in good agreement with the lattice pattern labels. We follow an analogous procedure to tune the hyperparameters for the other beam powers.

3.2 Analysis of MSM macrostate clustering

Fig. 3(a) shows the plot of 2-norms of the rows of the RNAM against the parameter $\tilde{t}$ for 10 ms lag time for 50 mW beam power for each of the MSM macrostates C1-6. The closer the norm is to 1, the better the clustering agrees with lattice fitting. We see that from the 2-norms of the RNAM rows for C2-6 converge to values in excess of 0.7 as $\tilde{t}$ increases, whereas that for C1 remains at a low value of only 0.4. This indicates that five of the six clusters well agree with lattice fitting and are quite insensitive to $\tilde{t}$. Fig. 3(c) presents the RNAM indicating the assignment probabilities of each lattice pattern (X1-36) to each of the six MSM macrostates (C1-6). The important idealized nanoparticle structures on a lattice and their lattice labels are shown in Fig. 3(b), where X4, X6, and X11 correspond to two lattice patterns while X16 corresponds to many lattice patterns with one particle separated from the other five particles that are gathered compactly. The other lattice patterns are shown Fig. S4(a) in the ESI†. Macrostates C2-6 are composed largely of a single lattice pattern, whereas C1 contains a mixture of patterns. This rationalizes the trends observed in Fig. 3(a) and leads us to expect that our MSM will contain five configurationally "pure" macrostates containing structures with a single long-lived OM lattice label, and a mixed macrostate containing structures with a mixture of lattice labels that rapidly interconvert on time scales shorter than the MSM lag time. Fig. S4(b) in the ESI† displays the entries of the CNAM, showing that the five lattice patterns corresponding to the five stable macrostates are indeed not contaminated by other macrostates.

The Chapman-Kolmogorov (CK) test assesses the Markovianity of a fitted MSM and therefore determines whether or not it is a valid kinetic model[89]. The $(\tilde{t}, n_M, \tau)$ = (8.3, 6, 10 ms)

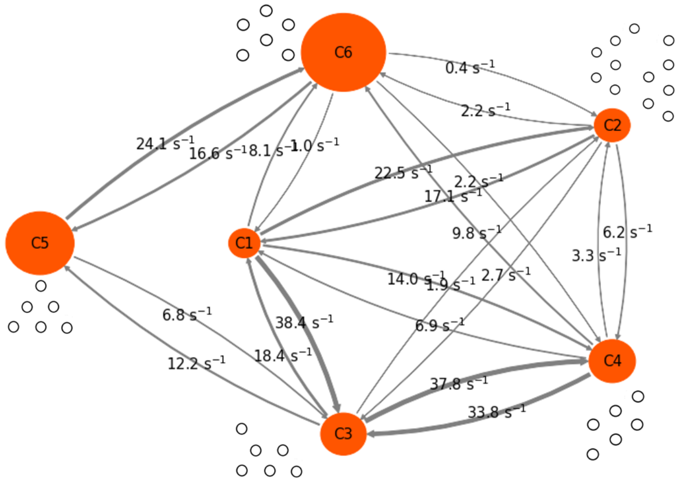

Fig. 4 The state map of the Markov state model (MSM) built for the 6-particle OM system with $\tau = 10$ ms, $n_M = 6$ and (beam power, $\tilde{t}$)=(50 mW, 8.3). The sizes of the orange circles are proportional to the probability distributions of the macrostates. The thickness of the connecting curves is in accord with the magnitude of the transition rates. Representative lattice patterns are shown next to each macrostate; C1 contains a mixture of lattice patterns too numerous to display.

macrostate MSM for a beam power of 50 mW satisfactorily passes the CK test as illustrated in Fig. S5 in the ESI†. Passing the CK test also provides *post hoc* validation of our non-canonical use of permutationally-invariant diffusion map embeddings and reversibilization of the microstate transition matrix within our MSM pipeline, demonstrating that our approach provides a satisfactory means to construct valid macrostate MSMs for non-conservative and permutational-invariant systems.

The macrostate MSM shown in Fig. 4 is the primary result of our analysis for the 50 mW beam power, and provides a wealth of interpretable and quantitative information on the metastable states and isomerization dynamics of the OM system. The sizes of the orange circles are proportional to the stationary distributions of the macrostates and the thickness of the connecting curves reflects the rate constants for transitions between macrostates. We illustrate next to each macrostate a schematic representation of the representative lattice patterns corresponding to the long-lived lattice labels contained within each macrostate. C3-6 essentially contain a single lattice pattern. C2 contains two lattice patterns that interconvert on time scales below the MSM lag time. C1 contains a mixture of lattice patterns too numerous to display. We identify the chevron (C6), triangle (C5), and parallelogram (C4) states that have been previously observed and reported in experimental studies of this 6-particle OM system (cf., Fig. 1).

3.3 Beam power dependence of the dynamics of optical matter systems

In addition to the MSM constructed for the 6-particle OM system for 50 mW beam power, we employed an analogous approach to construct MSMs for beam powers of 40, 60, 70, 80, and 90 mW. The complete set of macrostate MSMs is presented in Fig. S6 in the ESI†. By constructing MSMs over a range of beam powers we can analyze the ensemble of MSMs to extract trends in the relative stabilities of and transition rates between the various macrostates as a function of beam power.

Fig. 5 displays the changes with beam power of the stationary distribution probabilities of the six macrostates C1-6 and the rate constants of three selected macrostate-to-macrostate transitions (C4 → C3, C6 → C5, C5 → C6). Analogous plots for all 30 possible transitions are presented in Fig. S7 in the ESI†. Focusing on the five configurationally pure macrostates we see that the abundances of macrostates C2 and C3 – each corresponding to states with a single unstable (i.e., "dangling") particle – decrease as the beam power increases. Similar trends are observed for macrostate C4 corresponding to the parallelogram state. On the other hand, the abundance of the triangle macrostate C5 increases as the beam power increases, and that of the chevron, macrostate C6, first increases and then decreases with increasing beam power achieving a maximum at around 60 mW. These quantitative trends inform us that we need to further increase the beam power to stabilize triangle structure, whereas we need to decrease the beam power to further stabilize the parallelogram macrostate. We can make the chevron macrostate maximally important by tuning the beam power close to 60 mW.

The rates of the macrostate-to-macrostate transitions generally decrease when the beam power increases because the constraint exerted on the particles by the laser beam becomes larger with increasing beam power, leading to less freedom in the particle movement and thus smaller transition rate constants. There are exceptions when the beam power drops to less than or equal to 40 mW, because the constraint exerted on the particles by the laser beam is then not large enough to stabilize the OM structures for sufficiently long periods of time.

3.4 Beam power design to achieve the maximum population for the chevron state

From Fig. 5(f), we can see that the stability of the chevron pattern within macrostate C6 exhibits a non-monotonic behavior with respect to beam power. As an illustration of the value of our data-driven MSMs to inform control of the OM system, we adopt as our design goal maximal stabilization of the chevron pattern as a function of beam power. To do so, we carry out polynomial fitting of the chevron stationary distribution data for beam powers over the range 40 mW to 90 mW. Next, the Akaike information criterion (AIC) [92] is calculated for each fit, shown in the inset of Fig. 6, from which we can see that a 4th order polynomial corresponds to the smallest AIC, and is therefore the fit most supported by the data. (The maximum degree of the polynomial fit used for AIC calculation is four because AIC cannot be calculated for higher order polynomials given only six data points.) The analysis of the 4th order fit identifies a global maximum at a beam power of 62.94 mW, which represents our estimate of the beam power that maximally promotes stability of the chevron pattern within macrostate C6. To test this prediction, we carry out a simulation for a beam power of 62.94 mW, construct the corresponding macrostate MSM, and extract the stationary distribution of the C6 macrostate, which we plot as the red dot on Fig. 6. We can

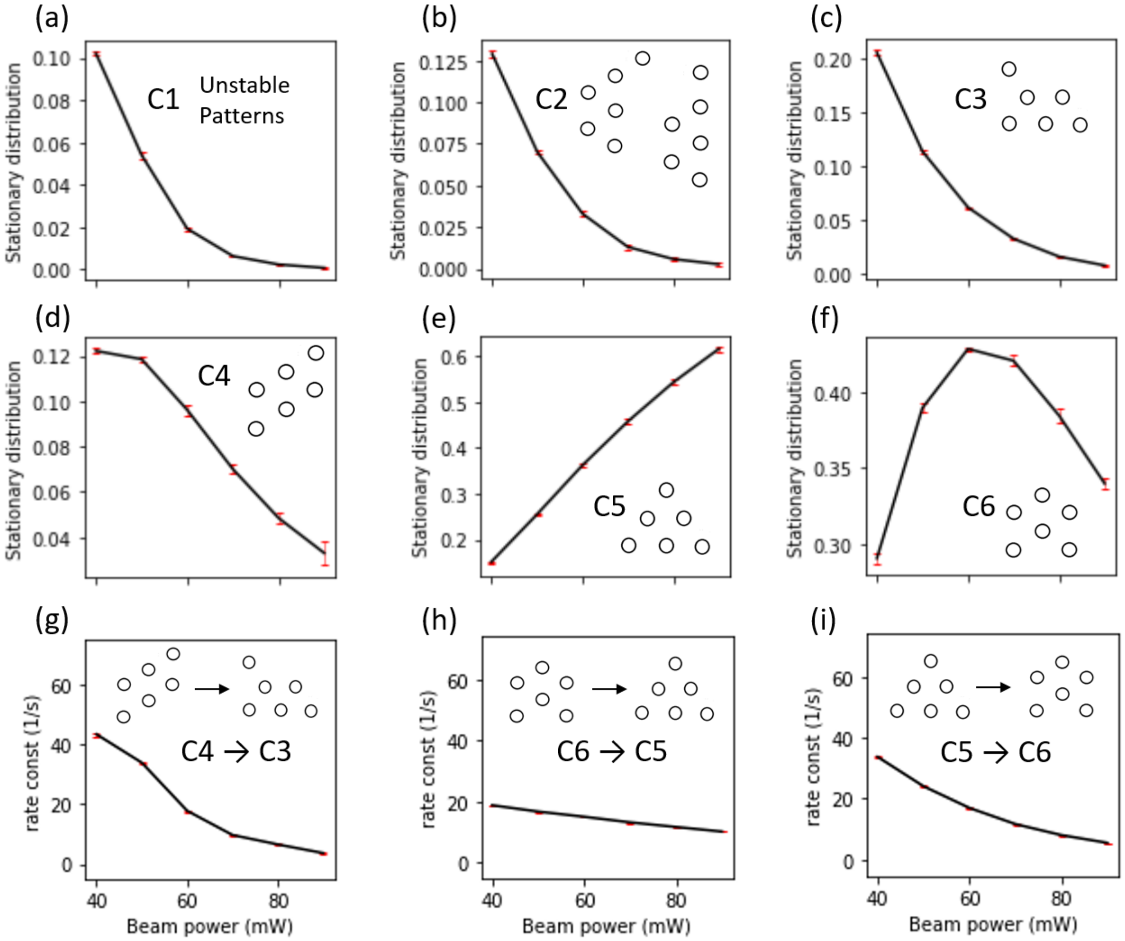

Fig. 5 Plot of (a-f) the stationary distribution probabilities for the six macrostates C1-6 and (g-i) the rate constants of 3 selected macrostate-to-macrostate transitions (C4 → C3, C6 → C5, C5 → C6) as a function of beam power for the 6-particle OM system. Error bars represent standard errors in the mean estimated by five-fold block averaging.

see that the predicted beam power indeed corresponds to a larger chevron population compared to other beam powers ranging from 40 mW to 90 mW. We could, of course, use the new 62.94 mW data point to further refine our beam power predictions by repeating this fitting and analysis approach. However, a fourth order polynomial fit to the new data predicts the maximum to lie at 62.93 mW, which is within 0.01 mW of our existing estimate of the optimal beam power. Our ensemble of MSM models enables analogous optimizations to be performed to maximally promote macrostates or transitions of interest.

4 Conclusions

We have developed a MSM construction method for non-conservative systems with permutational invariance using permutationally-symmetrized diffusion maps and reversibilized microstate transition matrix construction. We applied this approach to non-conservative OM systems to understand how the stability of the various macrostates and transition rates depend on beam power. We found that as beam power increases, the stability of most macrostates decreases while the stability of the triangle state increases and that of the chevron state first increases then decreases. A meta-analysis of our MSM models at various beam powers enables the rational control of the system via the design of beam powers to maximally promote particular self-assembled OM states or transitions. We found that the chevron macrostate reaches its maximum stability at a beam power of 62.94 mW.

The present paper represents a first proof of principle for this MSM construction method for the understanding and control of OM systems. In future work we plan to extend our analysis to additional controllable aspects of the incident beam including its phase profile and beam width in order to explore stabilization of additional self-assembled OM structures. We can also apply our analyses to more complex OM systems including those containing more particles, non-spherical particles, particles made of various other materials, or particle mixtures that have richer landscapes of self-assembled configurations. It is anticipated that this approach can help optimize the performance of optical matter machines.[53] We also anticipate that the approach may be extended to the analysis of molecular self-assembly where issues of permutational invariance and breaking of detailed balance must often be engaged in the construction of kinetic models.[93]

Future improvement of the method can include determination of dynamic, as opposed to static, control policies that can wield tighter control on the stable state and transitions through sim-

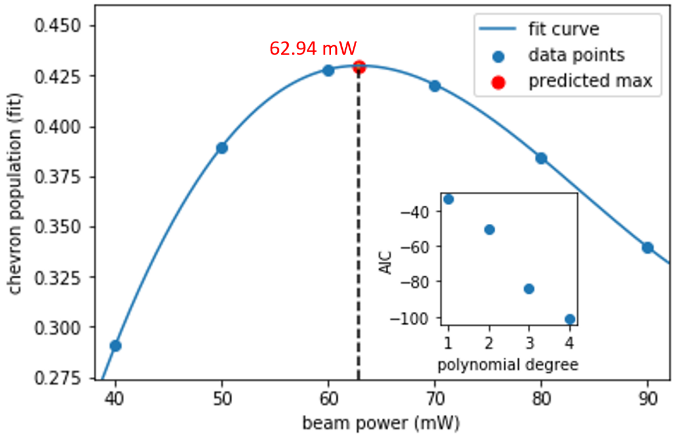

Fig. 6 Dependence of the stationary distribution of the C6 (chevron) macrostate as a function of beam power. Blue points correspond to the stationary probabilities extracted from the corresponding macrostate MSMs constructed at beam powers of 40, 50, 60, 70, 80, and 90 mW. The blue line represents the best 4th order polynomial fit to these data. The red point is the stationary distribution calculated from a MSM constructed at a beam power of 62.94 mW residing at the peak of the 4th order polynomial fit. (Inset) Scatter plot of Akaike information criterion (AIC) against polynomial degree for fits to the six initial beam powers.

ple feedback controllers that respond to the instantaneous state of the system and take the appropriate corrective action. These approaches can then be applied to construct MSMs directly from experimental as opposed to simulation data and use these models to guide experimental control strategies such as creating new stable OM structures and directing the transition of one structure to the other by on-the-fly adjustment of the beam parameters.

Author Contributions

A.L.F., S.C., and N.F.S. conceived of the study. All co-authors participated in discussions during the development of this research. S.C. performed the simulations. J.A.P. developed and implemented the EDLD simulation method. C.W.P. conducted optical trapping experiments and generated the experimental figures. S.C., N.F.S., and A.L.F. wrote the paper. S.A.R. critically revised the paper.

Conflicts of Interest

A.L.F. is a co-founder and consultant of Evozyne, Inc. and a co-author of US Patent Application 16/887,710, US Provisional Patent Applications 62/853,919, 62/900,420, and 63/314,898 and International Patent Applications PCT/US2020/035206 and PCT/US2020/050466.

Acknowledgements

The authors thank the Vannevar Bush Faculty Fellowship program sponsored by the Basic Research Office of the Assistant Secretary of Defense for Research and Engineering; Office of Naval Research (N00014-16-1-2502). This work was completed in part with resources provided by the University of Chicago Research Computing Center.

Notes and references

1 F. Shafiei, F. Monticone, K. Q. Le, X.-X. Liù, T. Hartsfield, A. Alu and X. Li, *Nat. Nanotechnol.*, 2013, **8**, 95–99.

2 S. N. Sheikholeslami, H. Alaeian, A. L. Koh and J. A. Dionne, *Nano Lett.*, 2013, **13**, 4137–4141.

3 A. Kuzyk, R. Schreiber, H. Zhang, A. O. Govorov, T. Liedl and N. Liu, *Nat. Mater.*, 2014, **13**, 862–866.

4 S. Yang, X. Ni, X. Yin, B. Kante, P. Zhang, J. Zhu, Y. Wang and X. Zhang, *Nat. Nanotechnol.*, 2014, **9**, 1002–1006.

5 Z. Qian, S. P. Hastings, C. Li, B. Edward, C. K. McGinn, N. Engheta, Z. Fakhraai and S.-J. Park, *ACS Nano*, 2015, **9**, 1263–1270.

6 J. Sharma, R. Chhabra, A. Cheng, J. Brownell, Y. Liu and H. Yan, *Science*, 2009, **323**, 112–116.

7 L. Malassis, P. Massé, M. Tréguer-Delapierre, S. P. Mornet, P. Weisbecker, V. Kravets, A. Grigorenko and P. Barois, *Langmuir*, 2013, **29**, 1551–1561.

8 J. Lee, J. H. Huh, K. Kim and S. Lee, *Adv. Funct. Mater.*, 2018, **28**, 1707309.

9 M. Kanahara, H. Satoh, T. Higuchi, A. Takahara, H. Jinnai, K. Harano, S. Okada, E. Nakamura, Y. Matsuo and H. Yabu, *Part. Part. Syst. Charact.*, 2015, **32**, 441–447.

10 U. Manna, J.-H. Lee, T.-S. Deng, J. Parker, N. Shepherd, Y. Weizmann and N. F. Scherer, *Nano Lett.*, 2017, **17**, 7196–7206.

11 A. Sánchez-Iglesias, M. Grzelczak, T. Altantzis, B. Goris, J. Perez-Juste, S. Bals, G. V. Tendeloo, S. H. Donaldson, B. F. Chmelka and J. N. Israelachvili, *ACS Nano*, 2012, **6**, 11059–11065.

12 M. Stalder and M. Schadt, *Opt. Lett.*, 1996, **21**, 1948–1950.

13 A. H. Zewail, *J. Phys. Chem. A*, 2000, **104**, 5660–5694.

14 N. F. Scherer, L. R. Khundkar, R. B. Bernstein and A. H. Zewail, *J. Chem. Phys.*, 1987, **87**, 1451–1453.

15 N. F. Scherer, D. M. Jonas and G. R. Fleming, *J. Chem. Phys.*, 1993, **99**, 153–168.

16 Y. Lee and Y. Shen, *Phys. Today*, 1980, **33**, 52–59.

17 R. D. Levine and R. B. Bernstein, *Phys. Today*, 1988, **41**, 90.

18 M. M. Burns, J.-M. Fournier and J. A. Golovchenko, *Phys. Rev. Lett.*, 1989, **63**, 1233.

19 M. M. Burns, J.-M. Fournier and J. A. Golovchenko, *Science*, 1990, **249**, 749–754.

20 S. Tatarkova, A. Carruthers and K. Dholakia, *Phys. Rev. Lett.*, 2002, **89**, 283901.

21 A. Ashkin, J. M. Dziedzic, J. E. Bjorkholm and S. Chu, *Opt. Lett.*, 1986, **11**, 288–290.

22 K. Dholakia and P.Zemánek, *Rev. Mod. Phys.*, 2010, **82**, 1767.

23 Y. Roichman, B. Sun, Y. Roichman, J. Amato-Grill and D. G. Grier, *Phys. Rev. Lett.*, 2008, **100**, 013602.

24 M. Pelton, M. Liu, H. Y. Kim, G. Smith, P. Guyot-Sionnest and N. F. Scherer, *Opt. Lett.*, 2006, **31**, 2075–2077.

25 P. Figliozzi, N. Sule, Z. yan, Y. Bao, S. Burov, S. K. Gray, S. A. Rice, S. Vaikuntanathan and N. F. Scherer, *Phys. Rev. E*, 2017, **95**, 022604.

26 J. Barton, D. Alexander and S. Schaub, *J. Appl. Phys.*, 1989, **66**, 4594–4602.

27 A. Yevick, D. J. Evans and D. G. Grier, *Philos. Trans. Royal Soc. A: Mathematical, Physical and Engineering Sciences*, 2017, **375**, 20150432.

28 Z. Yan, M. Sajjan and N. F. Scherer, *Phys. Rev. Lett.*, 2015, **114**, 143901.

29 S. E. S. Spesyvtseva and K. Dholakia, *ACS Photonics*, 2016, **3**, 719–736.

30 S. H. Simpson, P. Zemánek, O. M. Maragò, P. H. Jones and S. Hanna, *Nano Lett.*, 2017, **17**, 3485–3492.

31 S. Kuhn, A. Kosloff, B. A. Stickler, F. Patolsky, K. Hornberger, M. Arndt and J. Millen, *Optica*, 2017, **4**, 356–360.

32 D. Coursault, N. Sule, J. Parker, Y. Bao and N. F. Scherer, *Nano Lett.*, 2018, **18**, 3391–3399.

33 Z. Yan, S. K. Gray and N. F. Scherer, *Nat. Commun.*, 2014, **5**, 1–7.

34 J. Damková, L. Chvátal, J. Ježek, J. Oulehla, O. Brzobohatý and P. Zemánek, *Light Sci. Appl.*, 2018, **7**, 17135.

35 N. Sule, Y. Yifat, S. K. Gray and N. F. Scherer, *Nano Lett.*, 2017, **17**, 6548–6556.

36 J. Ng, Z. Lin, C. Chan and P. Sheng, *Phys. Rev. B*, 2005, **72**, 075130.

37 J. Taylor and G. Love, *Phys. Rev. A*, 2009, **80**, 053808.

38 S. Albaladejo, J. J. Sáenz and M. I. Marqués, *Nano Lett.*, 2011, **11**, 4597–4600.

39 L. Chvátal, O. Brzobohatý and P. Zemánek, *Opt. Rev.*, 2015, **22**, 157–161.

40 S. Sukhov, A. Shalin, D. Haefner and A. Dogariu, *Opt. Express*, 2015, **23**, 247–252.

41 V. Karásek, M. Siler, O. Brzobohatý and P. Zemaánek, *Opt. Lett.*, 2017, **42**, 1436–1439.

42 F. Nan, F. Han, N. F. Scherer and Z. Yan, *Adv. Mater.*, 2018, **30**, 1803238.

43 Y. Yifat, D. Coursault, C. W. Peterson, J. Parker, Y. Bao, S. K. Gray, S. A. Rice and N. F. Scherer, *Light Sci. Appl.*, 2018, **7**, 1–7.

44 J. Chen, J. Ng, K. Ding, K. H. Fung, Z. Lin and C. T. Chan, *Sci. Rep.*, 2014, **4**, 6386.

45 F. Han, J. A. Parker, Y. Yifat, C. W. Peterson, S. K. Gray, N. F. Scherer and Z. Yan, *Nat. Commun.*, 2018, **9**, 1–8.

46 J. Chen, S. Wang, X. Li and J. Ng, *Opt. Express*, 2018, **26**, 27694–27704.

47 S. Sukhov and A. Dogariu, *Rep. Prog. Phys.*, 2017, **80**, 112001.

48 Y. Roichman and D. G. Grier, *International Society for Optics and Photonics*, 2007, **6483**, 64830F.

49 C. W. Peterson, J. Parker, S. A. Rice and N. F. Scherer, *Nano Lett.*, 2019, **19**, 897–903.

50 N. Sule, S. A. Rice, S. Gray and N. F. Scherer, *Opt. Express*, 2015, **23**, 29978–29992.

51 Y.-I. Xu, *Appl. Opt.*, 1995, **34**, 4573–4588.

52 J. A. Lock and G. Gouesbet, *J. Quant. Spectrosc. Radiat. Transf.*, 2009, **110**, 800–807.

53 J. Parker, C. W. Peterson, Y. Yifat, S. A. Rice, Z. Yan, S. K. Gray and N. F. Scherer, *Optica*, 2020, **7**, 1341–1348.

54 S. Chen, C. W. Peterson, J. A. Parker, S. A. Rice, A. L. Ferguson and N. F. Scherer, *Nat. Commun.*, 2021, **21**, 2548.

55 U. Sengupta, M. Carballo-pacheco and B. Strodel, *J. Chem. Phys.*, 2019, **150**, 115101.

56 V. S. Pande, K. Beauchamp and G. R. Bowman, *Methods*, 2010, **52**, 99–105.

57 J. D. Chodera and F. Noé, *Curr. Opin. Struct. Biol.*, 2014, **25**, 135–144.

58 B. E. Husic and V. S. Pande, *J. Am. Chem. Soc.*, 2018, **140**, 2386–2396.

59 C. Wehmeyer, M. K. Scherer, T. Hempel, B. E. Husic, S. Olsson and F. Noé, *Living J. Comput. Mol. Sci.*, 2019, **1**, 5965.

60 R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner and S. W. Zucker, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 7426–7431.

61 X. Chen and Y. Yang, *Appl. Comput. Harmon. Anal.*, 2021, **52**, 303–347.

62 M. Weber and T. Galliat, *Konrad-Zuse-Zentrum f ur Informationstechnik Berlin*, 2002, **12**, 1–12.

63 J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill and W. C. Swope, *J. Chem. Phys.*, 2007, **126**, 155101.

64 S. Röblitz and M. Weber, *Adv. Data Anal. Classif.*, 2013, **7**, 147–179.

65 P. Deuflhard and M. Weber, *Linear Algebra Appl.*, 2005, **398**, 161–184.

66 S. Kube and M. Weber, *J. Chem. Phys.*, 2007, **126**, 024103.

67 S. Schultze and H. Grubmüller, *J. Chem. Theory Comput.*, 2021, **17**, 5766–5776.

68 C. R. Schwantes and V. S. Pande, *J. Chem. Theory Comput.*, 2015, **11**, 600–608.

69 G. Andrew, R. Arora, J. Bilmes and K. Livescu, *PMLR*, 2013, **28**, 1247–1255.

70 A. Mardt, L. Pasquali, H. Wu and F. Noé, *Nat Commun.*, 2018, **9**, 4443.

71 W. Chen, H. Sidky and A. L. Ferguson, *J. Chem. Phys.*, 2019, **150**, 214114.

72 R. R. Coifman and S. Lafon, *Appl. Comput. Harmon. Anal.*, 2006, **21**, 5–30.

73 B. Nadler, S. Lafon, R. R. Coifman and I. G. Kevrekidis, *Appl. Comput. Harmon. Anal.*, 2006, **21**, 113–127.

74 A. Singer, R. Erban, I. G. Devrekidis and R. R. Coifman, *Proc. natl. Acad. Sci. U. S. A.*, 2009, **106**, 16090–16095.

75 Z. Trstanova, B. Leimkuhler and T. Lelièvre, *Proc. R. Soc. A*, 2019, **476**, 20190036.

76 A. W. Long and A. L. Ferguson, *J. Phys. Chem. B*, 2014, **118**, 4228–4244.

77 A. W. Long, J. Zhang, S. Granick and A. L. Ferguson, *Soft Matter*, 2015, **11**, 8141–8153.

78 A. W. Long, C. L. Phillips, E. Jankowksi and A. L. Ferguson, *Soft Matter*, 2016, **12**, 7119–7135.

79 L. Boninsegna, G. Gobbo, F. Noé and C. Clementi, *J. Chem. Theory Comput.*, 2015, **11**, 5947–5960.

80 J. Hu and A. L. Ferguson, *Intelligent Data Analysis*, 2016, **20**, 637–654.

81 A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis and P. G. Debenedetti, *Chem. Phys. Lett.*, 2011, **509**, 1–11.

82 C. Fowlkes, S. Belongie, F. Chung and J. Malik, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004, **26**, 214–225.

83 S. Lafon, Y. Keller and R. R. Coifman, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, 1784–1797.

84 B. E. Sonday, M. Haataja and I. G. Kevrekidis, *Phys. Rev. E*, 2009, **80**, 031102.

85 J. Wang and A. L. Ferguson, *Macromolecules*, 2018, **51**, 598–616.

86 A. W. Long and A. L. Ferguson, *Appl. Comput. Harmon. Anal.*, 2019, **47**, 190–211.

87 J. Wang, M. A. Gayatri and A. L. Ferguson, *J. Phys. Chem. B*, 2017, **121**, 4923–4944.

88 J. B. MacQueen, in *Some methods for classification and analysis of multivariate observations*, Proc. Fifth Berkeley Sympos. Math. Statist. and Probability, 1967, pp. 281–297.

89 M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz and F. Noé, *J. Chem. Theory Comput.*, 2015, **11**, 5525–5542.

90 H. Wu, F. Nüske, F. Paul, S. Klus, P. Koltai and F. Noé, *J. Chem. Phys.*, 2017, **146**, 154104.

91 P. J. Rousseeuw, *J. Comput. Appl. Math.*, 1987, **20**, 53–65.

92 H. Akaike, *IEEE Trans. Automat. Contr.*, 1974, **19**, 716–723.

93 X. Zeng, B. Li, Q. Qiao, L. Zhu, Z.-Y. Lu and X. Huang, *Physical Chemistry Chemical Physics*, 2016, **18**, 23494–23499.