Soft Matter

PAPER

View Article Online
View Journal | View Issue

Cite this: *Soft Matter*, 2015, **11**, 8141

Received 9th August 2015,
Accepted 24th August 2015

DOI: 10.1039/c5sm01981h

www.rsc.org/softmatter

Machine learning assembly landscapes from particle tracking data†

Andrew W. Long,[a] Jie Zhang,[a] Steve Granick[b] and Andrew L. Ferguson*[a]

Bottom-up self-assembly offers a powerful route for the fabrication of novel structural and functional materials. Rational engineering of self-assembling systems requires understanding of the accessible aggregation states and the structural assembly pathways. In this work, we apply nonlinear machine learning to experimental particle tracking data to infer low-dimensional assembly landscapes mapping the morphology, stability, and assembly pathways of accessible aggregates as a function of experimental conditions. To the best of our knowledge, this represents the first time that collective order parameters and assembly landscapes have been inferred directly from experimental data. We apply this technique to the nonequilibrium self-assembly of metallodielectric Janus colloids in an oscillating electric field, and quantify the impact of field strength, oscillation frequency, and salt concentration on the dominant assembly pathways and terminal aggregates. This combined computational and experimental framework furnishes new understanding of self-assembling systems, and quantitatively informs rational engineering of experimental conditions to drive assembly along desired aggregation pathways.

1 Introduction

Self-assembly – the spontaneous organization of building blocks into structured aggregates – is an important fabrication route for both biological[1,2] and engineered materials,[3,4] including viral capsids,[5,6] photonic crystals,[7,8] and sensing nanostructures.[9] Assembly proceeds in the high dimensional space of the location and orientation of all constituent building blocks.[10] Low-dimensional assembly landscapes or "roadmaps" providing a comprehensible, high-level description of the assembly process are of great value in revealing underlying mechanisms and developing understanding of assembly by mapping out the accessible aggregates and assembly pathways.[1,11–17] Furthermore, by quantifying how these roadmaps are influenced by experimentally controllable parameters and conditions, these descriptions can inform pathway engineering principles to steer assembly towards desired structural and/or functional aggregates.[14,17–19]

The inherently many-body nature of self-assembly means that such processes are generically expected to admit low-dimensional descriptions in a small number of collective order parameters corresponding to the dominant emergent assembly dynamics.[17–19] Computer simulations of self-assembly furnish the coordinates and orientations of all building blocks as a function of time, and dimensionality reduction techniques can discover these collective coordinates lying latent within the high dimensional trajectory.[17,20] A variety of approaches have been applied to infer collective order parameters for single-molecule dynamics, including principal components analysis (PCA),[21–23] sketch maps,[24,25] locally linear embedding (LLE),[26] Laplacian eigenmaps,[27] Isomap,[28–31] and diffusion maps.[32–37] Recently, two of us (A.W.L. and A.L.F.) developed a means to apply diffusion maps to infer collective order parameters from molecular simulations of many-body systems, and used this technique to compute the assembly landscape for self-assembling patchy colloids.[17] In this work, we apply this approach to infer self-assembly landscapes from experimental particle tracking data, and quantify how these landscapes change as a function of experimental conditions. To the best of our knowledge, this represents the first time that collective order parameters and assembly landscapes have been inferred directly from experimental data.

The particular system we consider is the nonequilibrium self-assembly of Janus colloids, micron sized spheres whose hemispheres possess different surface chemistries.[38–40] These highly tunable building blocks can be induced to self-assemble into diverse aggregate structures on experimentally measurable time scales, presenting an ideal system to study the effect of particle anisotropy and experimental conditions on the assembly of clusters, chains, helices, sheets, discs, and tubes.[39,41–45]

[a] *Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA. E-mail: alf@illinois.edu; Fax: +1-217-333-2736; Tel: +1-217-300-2354*

[b] *Center for Soft and Living Matter, Ulsan National Institute of Science and Technology, 50 UNIST-gil, Ulju-gun, Ulsan 689-798, South Korea*

† Electronic supplementary information (ESI) available: Five figures pertaining to assembly landscapes, three tables showing the various experimental conditions considered, and two movies showing representative assembly trajectories. See DOI: 10.1039/c5sm01981h

In particular, we are interested in the nonequilibrium self-assembly of metallodielectric Janus particles under an applied AC electric field.[46,47] The oscillating field induces a differential dipole moment between metallic and dielectric hemispheres leading to anisotropic interactions between colloids and induced motions through reverse induced-charge electrophoresis (rICEP) at high field frequencies. This process of rICEP motion is incompletely understood,[48] hindering the development of physics-based models and design principles. An attractive feature of our data-driven machine learning approach to discover collective order parameters and assembly roadmaps is that it requires no knowledge of the underlying physics of the system. Accordingly, we can infer empirical assembly roadmaps directly from experimental data without a complete understanding of the underlying physics, and the pathways and collective dynamics discovered by our analysis can inform improved understanding of the system.

The structure of this paper is as follows. In Section 2, we describe the experimental details of the Janus particle synthesis and self-assembly, and the computational details of the machine learning algorithm. In Section 3, we describe the application of our approach to nonequilibrium self-assembly of two metallodielectric Janus particle systems: (i) the templated assembly of Janus pinwheels and other branched structures directed by passive linker particles, and (ii) the spontaneous self-assembly of long Janus particle chains and loops. In each case we demonstrate that our approach reveals the underlying self-assembly pathways, and furnishes quantitative design rules linking the attainable terminal aggregates to the AC frequency, electric field strength, and salt concentration. In Section 4 we present our conclusions and outlook for future work.

2 Materials & methods

2.1 Janus particle tracking experiments

Janus particles are synthesized by depositing 20 nm of titanium and then 20 nm of SiO_2 vertically on a submonolayer of 3 μm silica particles (Tokuyama) using an electron-beam evaporator. After being washed with isopropyl alcohol and deionized water, Janus particles are sonicated down from the substrate to deionized water. For the templated assembly experiments with binary mixtures of Janus and "linker" particles, Janus particles and untreated silica particles are mixed in a 10 : 1 ratio. NaCl stock solution is added to the particle suspension to prepare 0.01 mM and 0.1 mM NaCl solutions, respectively. The particle suspensions are sandwiched between two ITO coated coverslips (SPI Supplies) separated by a 120 μm-thick spacer (GraceBio SecureSeal) with a 9 mm hole in the center to confine the fluid. An AC electric field is applied to the sample cell using a function generator (Agilent 33522A). The sample cell is imaged with a 40× air objective on an inverted microscope (Axiovert 200). Microscopic images and videos are taken with a CMOS camera (Edmund Optics 5012M GigE). A schematic of our experimental setup is given in Fig. 1a. Representative movies of the templated and homogenous self-assembly experiments are presented in Movies S1 and S2 (ESI†).

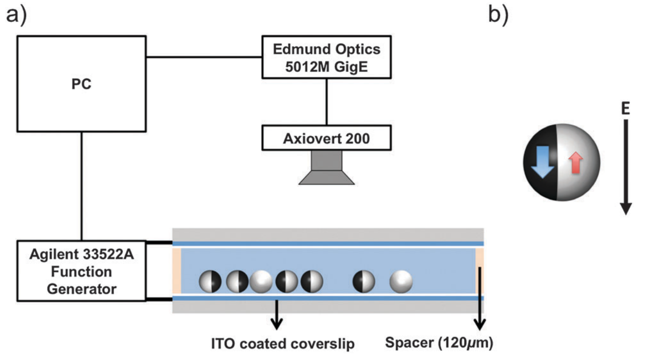

Fig. 1 Janus particle self-assembly under an applied AC electric field. (a) Schematic of the Janus particle tracking experimental setup comprising both Janus particles and passive linker particles. (b) Under an applied electric field, the Janus interface aligns parallel with the field direction and a dipole moment is induced in the metallic (left) and dielectric (right) hemispheres of the Janus particle. The magnitude of the dipole moment is greater in the metallic hemisphere due to an increased polarizability of the thin metallic layer, and aligns with the direction of the applied field. The dipole moment of the dielectric hemisphere aligns antiparallel with the applied field.

Particle positions are determined using a circular Hough transform,[49] providing the 2D coordinates of all particle centers in a given image or video frame. We use these positions to identify the distinct particle clusters in a frame. First, we define particles as being "bonded" to one another if their center of mass separation is less than $r_{\mathrm{cutoff}} = 1.15D$, where $D = 3$ μm is the particle diameter, to form a complete binary interaction graph at a particular instance in time, $\mathbf{K}$, where $\mathbf{K}(p,q) = 1$ if particles p and q are bonded and $\mathbf{K}(p,q) = 0$ if not bonded. Using Tarjan's algorithm,[50] we identify distinct subgraphs, $\mathbf{G}$, defining connected clusters of particles within the complete graph. We aggregate all of these clusters from the video frames across all of our experimental trajectories to form a library of clusters $\{\mathbf{G}_i\}$. We have verified that our results are robust to choices of bonding cutoffs in the range $[1.10D, 1.25D]$.

2.2 Janus particle self-assembly

Under the influence of the perpendicular AC electric field, the Janus particles align their interface parallel to the field and an induced dipole moment develops in each hemisphere as illustrated in Fig. 1b. The net interaction between two Janus particles, A and B, at a separation, r, and relative orientation of the particle directors normal to the Janus interface, θ, is the sum of the four distinct dipole–dipole interactions between each of the two dipoles, i, in particle A and the two dipoles, j, in particle B,

$$\begin{aligned} U_{\mathrm{AB}}(r,\theta) &= \sum_{i,j} u_{ij}\left(r_{ij}(r,\theta)\right), \\ u_{ij}\left(r_{ij}\right) &= \frac{\mathrm{Re}\left(\alpha_i\alpha_j\right)E^2\left(1-3\cos^3\beta\right)}{4\pi\varepsilon {r_{ij}}^3}, \end{aligned} \tag{1}$$

where $i = \{\mathrm{metal_A},\mathrm{dielectric_A}\}$, $j = \{\mathrm{metal_B},\mathrm{dielectric_B}\}$, $r_{ij}(r,\theta)$ is the orientation and separation dependent dipole–dipole separation,

α_k is the complex polarizability of the Janus hemisphere composed of material k, E is the electric field strength, β is the angle of the interface ($\beta = \pi/2$ for a Janus particle), and ε is the permittivity of the solution. The differing polarizability of the metallic and dielectric hemispheres gives rise to an orientation-dependent attractive or repulsive interaction between particles. To calculate the induced dipole of a dielectric particle coated with metal on one hemisphere, we assume the induced dipole moment of each hemisphere is half of a spherical particle of the same diameter and material. For a spherical colloidal particle, the complex polarizability is computed as $\alpha = 4\pi\varepsilon KR^3$, where K is the complex dipole coefficient, $R = 1.5$ μm is the particle radius, and $\varepsilon/\varepsilon_0 = 80$, where ε_0 is the vacuum permittivity. K is sensitive to the electric field frequency, particle surface charge and ionic strength in the solution. We compute the complex dipole coefficient of a metallic sphere using the analytical solution in ref. 51, considering the effect of the protective SiO_2 coating in reducing the electric field outside the metal coating. For a negatively charged silica sphere, we employ the approximate analytical solution in ref. 52.

Above a particular transition frequency of the applied electric field, this difference in polarization across the Janus interface also leads to particle motion by reverse induced charge electrophoresis (rICEP) perpendicular to the field in the direction of the metallic hemisphere.[48] The underlying physics of this phenomena is not well understood, limiting our capacity to predict the effect of frequency, field strength, and salt concentration on the particle velocities. Instead, we empirically quantify the ballistic velocities from a quadratic fit of particle mean squared displacements at short times using the particle tracking algorithm developed by Crocker and Grier.[53]

2.3 Machine learning of assembly roadmaps

The particle tracking data recorded during the self-assembly experiments described in Section 2.1 contains all of the information on the self-assembly mechanisms and terminal aggregates attainable under different experimental conditions, but it is exceedingly challenging to resolve these mechanisms and terminal states by visual inspection alone.[42] Visualization can be supported by tracking cluster size distributions,[18,19] or tracking the evolution of the evolving clusters using canonical shape descriptors,[54] but such intuitive, coarse-grained descriptors are typically not coincident with the emergent collective order parameters that govern assembly.

The difficulty in parsing the particle tracking data is largely a question of dimensionality. The orientation and location of a single Janus particle in the plane oriented parallel to the external electric field is given by three numbers specifying its 2D location and its rotational orientation. The assembly trajectory for a collection of N Janus particles residing in the plane, resides in a $3N$-dimensional phase space. For even a modest number of particles, it is extremely challenging to identify within this high-dimensional space the underlying assembly pathways and accessible aggregates that lie buried within this data. Despite existing in this high-dimensional phase space, self-assembly is an inherently multi-body process that depends on cooperative interparticle interactions. This coupling of building block degrees of freedom generally results in a separation of time scales, such that the long time evolution of the system is governed by a relatively small number of collective modes.[34,55] Extracting these slow collective modes permits construction of a low-dimensional subspace capturing the important dynamical features of self-assembly, and the existence of such low-dimensional subspaces – frequently of dimensionality as low as 2–3 – has been borne out in molecular simulations of polymer dynamics, protein folding, and colloidal self-assembly.[17,29,35,56,57]

Nonlinear learning offers a means to systematically extract this low-dimensional subspace – the so-called intrinsic manifold[35] – from the high-dimensional data, and in doing so reveal the important collective modes, assembly pathways, and accessible aggregates. In contrast to linear techniques (*e.g.*, principal components analysis[21]), nonlinear approaches are more powerful and flexible in determining the potentially nonlinear combinations of the particle degrees of freedom comprising the collective modes.[20,29,56] We recently reported a new approach rendering diffusion maps applicable to many-body systems, and demonstrated its capacity to infer assembly mechanisms of patchy colloids from molecular dynamics simulation trajectories.[17] In this work, we apply this technique to particle tracking trajectories to infer assembly pathways and the accessible terminal aggregates directly from experimental data. Full details of our methodology are given in ref. 17, but we briefly sketch the approach below. We note that our methodology requires neither that the number nor identity of particles remain constant in each frame of the trajectory, and the interaction potentials between particles need not be known. Rather the approach is a data-driven one that infers a low-dimensional description of the assembly process from empirical observations of the diversity of structural aggregates within the system.

2.3.1 Cluster distance metric. We start by compiling a library of all clusters observed in our particle tracking trajectories using the procedure described in Section 2.1. Each distinct cluster observed in the particle tracking movies is represented by its underlying bonding network, yielding a binary adjacency matrix $\mathbf{G}$ where $\mathbf{G}(p,q) = 1$ denotes the presence, and $\mathbf{G}(p,q) = 0$ the absence of a bond between particles p and q. We then compute distances between all pairs of graphs in our library, $(\mathbf{G}_i,\mathbf{G}_j)$, using graph matching to identify the pseudo-optimal permutation, $\mathbf{H}$, of particle labels between these two clusters such that the adjacency matrices are maximally similar. Mathematically this corresponds to finding the $\mathbf{H}$ that minimizes the Frobenius norm between $\mathbf{G}_i' = \mathbf{H}\mathbf{G}_i\mathbf{H}^\mathrm{T}$ and $\mathbf{G}_j$, for clusters with different numbers of particles, the smaller graph is augmented by a number of rows and columns of zero to bring it up to the same size as the larger when computing alignment. This optimization is strongly polynomial, requiring an approximate solution algorithm for even moderately sized networks.[58,59] We adopt a greedy search procedure based on an adaptation of the IsoRank algorithm[60] that preserves local bond connectivity that we detail in ref. 17.

Given the optimal particle label permutation between two graphs, we define their dissimilarity based on the separation between analogous pairs of bonded particles. For each cluster we construct the matrix $\boldsymbol{\chi}$, where $\boldsymbol{\chi}(p,q)$ is the real space distance between particles p and q if $\boldsymbol{\chi}(p,q) < r_{\text{cutoff}}$, and is zero otherwise. We define the structural dissimilarity of clusters i and j as,

$$\mathbf{d}_{ij} = \sum_{p} \sum_{q>p} \left\| \chi_i{}'(p,q) - \chi_j(p,q) \right\|, \tag{2}$$

where $\boldsymbol{\chi}_j$ contains the distances between the bonded particles in cluster j, and $\boldsymbol{\chi}_i{}'$ the distances between the corresponding particles in cluster i under the permutation defined by $\mathbf{H}$. Ghost particles are added as necessary to bring the clusters to equal size, which corresponds to the addition of a number of imaginary non-bonded particles to the smaller cluster.[17] In the case of the templated assembly of Janus pinwheels directed by passive linker particles, $\boldsymbol{\chi}$ corresponds to distances between particle centers. In the homogeneous self-assembly of Janus particles into chains and loops it corresponds to the distance between metallic face centers – the midpoint between the particle center and metallic surface normal to the Janus interface – which provides information on relative particle orientations that is critical in distinguishing different loop and chain architectures.

The distance metric $\mathbf{d}_{ij}$ provides a good measure of the structural dissimilarity of clusters that encapsulates both the breaking and forming of bonds between different cluster sizes and architectures, and deviations in bond lengths within a single architecture.[17] We observe that the definition of structural distances based on graph matching of the underlying networks surmounts the difficulties associated with the absence of a spatially invariant real space basis in which to compare clusters of different numbers of indistinguishable particles translating and rotating through space.

2.3.2 Diffusion map dimensionality reduction. Having computed structural distances between all pairs of clusters, we apply diffusion maps[20,32–34,61] to infer the collective order parameters driving self-assembly in which to construct a low-dimensional assembly landscape or "roadmap" of assembly. In a geometrical sense, the diffusion map seeks to identify a low-dimensional subspace – the intrinsic manifold – within the 3N-dimensional space of particle coordinates to which the important assembly dynamics are effectively restrained.[17] In a temporal sense, it seeks a small number of slowly collective modes defining a slow subspace to which the remaining degrees of freedom are effectively slaved.[17] The diffusion map was first developed by Coifman and coworkers,[32–34,61] and we have detailed its application to molecular and colloidal systems in ref. 17, 20 and 35. In brief, we model transitions between different cluster configurations in the configurational phase space as a Markov process with hopping probabilities based on pairwise structural proximity. We then identify the important collective order parameters as the slowest relaxation modes of the Markov chain. We first form the matrix $\mathbf{A}$ by convoluting the pairwise distances with a Gaussian kernel, $\mathbf{A}_{ij} = \exp(-\mathbf{d}_{ij}{}^2/2\varepsilon)$, where ε is a soft-thresholding bandwidth that limits transitions to between structurally similar cluster configurations residing in the same neighborhood of the high dimensional space. The Gaussian kernel is the infinitesimal generator of a diffusion process, and by forming this convolution we model a discrete diffusion process over the clusters residing in the high-dimensional space.[32] An appropriate value of ε is specified using the automated procedure detailed in ref. 62.

We then form the diagonal matrix $\mathbf{D}$ as, $\mathbf{D}_{ii} = \sum_k \mathbf{A}_{ik}$, such that the matrix product $\mathbf{M} = \mathbf{D}^{-1}\mathbf{A}$ is a right-stochastic Markov matrix describing a random walk over the clusters in configurational space. $\mathbf{M}$ can be diagonalized into a set of right eigenvectors, $\{\vec{\Psi}_i\}$, and associated eigenvalues $\{\lambda_i\}$, where, by the Markov property, $\vec{\Psi}_1 = \vec{1}$ and $\lambda_1 = 1$. These eigenvectors are the discrete analog of the eigenfunctions of the Fokker–Planck equation describing the collective harmonic modes describing the time evolution of the probability density over the data.[32,33] Large eigenvalues correspond to slow relaxation modes of the Markov process and small eigenvalues correspond to fast modes. A gap in the eigenvalue spectrum is indicative of a separation of time scales, wherein the slow collective modes govern the long time evolution of the system to which the remaining modes effectively couple as noise.[34,35,55] Systematic techniques exist to infer the existence and location of a spectral gap.[35,63]

For an eigenvalue spectrum possessing a gap after $\lambda_{(k+1)}$, we can locate each experimentally observed cluster on the k-dimensional intrinsic manifold spanned by $\{\vec{\Psi}_i\}_{i=2}^{k+1}$ (recalling that $\vec{\Psi}_1 = \vec{1}$ is the trivial all-ones vector associated with the steady-state distribution) by forming the k-dimensional diffusion map embedding of the ith cluster into the ith component of the top k non-trivial eigenvectors,

$$\text{cluster}_i \rightarrow (\vec{\Psi}_2(i), \vec{\Psi}_3(i), \ldots, \vec{\Psi}_{k+1}(i)) \tag{3}$$

A limitation of diffusion maps is that the methodology does not furnish an explicit mapping between the high dimensional coordinate space and the slow collective modes spanning the low dimensional embedding, making it a challenge to assign physical interpretability to these modes. Methods exist to sieve pools of candidate physical variables to find good combinations approximating the leading eigenvectors,[64,65] but the resultant functions can themselves be so complicated as to obscure a transparent physical interpretation. Indeed, the existence of a simple physical characterization is not guaranteed for complex many-body problems involving the interaction of many degrees of freedom.[17,20] In this work we assist in the physical interpretation of the modes by coloring the diffusion map embeddings with simple physical "bridge" variables that show good correlation with the eigenvectors spanning the embedding, including cluster size and average cluster network path length.

Under the mild assumptions that (i) the dynamics of the system may be modeled as a diffusion process, and (ii) that our measure of pairwise similarity between clusters is an appropriate measure of the short-time diffusive motions, then Euclidean distances in diffusion map space correspond to diffusion distances in the original high-dimensional space measuring the time required for one cluster to dynamically evolve into another.[17,20,32,35,61] That we observe the emergence of a small number of slow collective modes above a spectral gap in the eigenvalue spectrum suggests that the system dynamics can be effectively modeled in the Mori–Zwanzig formalism as a set of coupled stochastic differential equations in the slow modes to which the fast modes couple as noise,[55,66] providing *post hoc* support that the system may be modeled as a diffusion process. Secondly, we have shown in ref. 17 that our structural distance metric provides a good measure of structural remodeling on short time scales, and is therefore expected to capture short time diffusive motions. The dynamic interpretability conveyed to Euclidean distances in the diffusion map is a powerful property that allows us to interpret structural transitions over the slowest dynamical modes given by the diffusion map, resolving both local and global deviations in the underlying configurational space.

2.3.3 Effective free energy landscapes. By collecting histograms over the embeddings we construct effective free energy landscapes describing the relative probabilities of the various cluster architectures in the experimental trajectories. We construct these landscapes using the standard relationship from (equilibrium) statistical mechanics, $\beta\hat{F}(\vec{\xi}) = -\ln\hat{P}(\vec{\xi}) + C$, where $\beta = 1/k_BT$ is the inverse temperature, k_B is Boltzmann's constant, T is the absolute temperature, $\vec{\xi}$ is a k-dimensional vector specifying a point on the intrinsic manifold spanned by the vectors $(\vec{\Psi}_2,\vec{\Psi}_3,\ldots,\vec{\Psi}_{k+1})$, $\hat{P}(\vec{\xi})$ is a histogram approximation to the probability density of single particles on the manifold at $\vec{\xi}$, weighting each point in the manifold by the number of particles belonging to the corresponding cluster, $\hat{F}(\vec{\xi})$ is an effective per particle free energy at $\vec{\xi}$, and C is an arbitrary constant that we specify such that the free energy of an isolated monomer defines the zero of free energy. Since the experimental self-assembly trajectories were conducted in the presence of an external driving force (the oscillating AC electric field), they are inherently out of equilibrium, and therefore $\hat{F}$ cannot be interpreted as a true thermodynamic free energy, but rather an effective free energy that is best interpreted as a convenient representation of the likelihood to find a particle in a particular cluster architecture. By computing these effective free energy landscapes upper different conditions – salt concentration, electric field strength, AC frequency – we use these effective free energy landscapes, along with the dynamic interpretability of the diffusion map embeddings, to link experimental control parameters to changes in the relative propensities of different cluster architectures and mechanistic pathways over the assembly landscape.

3 Results & discussion

3.1 Templated Janus "pinwheel" assembly

Mixtures of metallodielectric Janus particles with passive dielectric particles in a 10 : 1 ratio were subjected to AC frequencies of 70 kHz–11 MHz and applied AC field strengths of 250–833 V cm^{-1} at a salt concentration of [NaCl] = 0.1 mM (*cf.* Section 2.1). A total of 28 experiments were conducted over this range at the particular field strengths and frequencies listed in Table S1 (ESI†). At each set of conditions, the system was allowed to attain steady state with respect to the applied AC field by waiting for the particle velocity distribution to stabilize, typically occurring approximately 8 seconds after initial application of the field. The transient portion of each particle tracking trajectory was rejected from our analysis such that the aggregates and assembly pathways extracted by our analysis correspond to those produced by the dynamical assembly and disassembly of clusters at steady state.

Within the ensemble of particle trajectories from all 28 experiments, we identified a total of 3 403 918 clusters (including free monomers) belonging to 21 708 distinct cluster architectures. Since the size of the matrices used to perform the diffusion mapping scale quadratically with cluster number, we make the analysis computationally tractable by retaining a random subset of cluster from each unique cluster architecture to generate a reduced ensemble of 60 926 clusters. We apply diffusion maps using a kernel bandwidth of $\varepsilon = \exp(10)$ determined using the approach in ref. 62. We note that particle identity – Janus or linker – is not easily identifiable in these experiments, and as such our diffusion map analysis operates purely on the geometric cluster architectures rather than the type of particles constituting these clusters. We employ the L-method[63] to identify a gap in the eigenvalue spectrum after λ_3 (Fig. S1, ESI†), implying an effective dimensionality of two and motivating the construction of two-dimensional diffusion map embeddings in the top two non-trivial eigenvectors $\{\vec{\Psi}_2,\vec{\Psi}_3\}$. To preserve the experimentally observed cluster distribution, we project into the diffusion map embedding the remaining $(3\,403\,918 - 60\,926) = 3\,342\,992$ clusters using the Nyström extension.[67–69] By analyzing all systems simultaneously, we construct a single composite diffusion map defining a common basis within which to compare the distribution of clusters at different experimental conditions by restricting the ensemble to the clusters extracted from particular experimental trajectories.

We present the two-dimensional composite diffusion map in Fig. 2. Each point corresponds to a particular cluster observed in one of the experimental particle tracking trajectories. To assist in visual discrimination of different cluster architectures, we color each point according to the average path length between pairs of particles in the cluster bonding network as a coarse-grained measure of cluster size and connectivity. We also visualize representative clusters to illustrate the cluster architectures populating different regions of the intrinsic manifold constituting the assembly landscape. The landscape reveals four distinct quadrants defining different aggregation states. The lower left quadrant is populated primarily by free monomers residing at $(\Psi_2 \approx -0.135, \Psi_3 \approx -0.185)$ and possessing an average path length of zero. Tracing a pathway up to the upper left quadrant corresponds to the formation of relatively small, dense cluster architectures residing in the vicinity of $(\Psi_2 \approx -0.135, \Psi_3 \approx -0.165)$ and possessing a small average path length (high network connectivity). The assembly pathway linking the monomers to the lower right quadrant corresponds to the formation of spinning "pinwheels" in the vicinity of $(\Psi_2 \approx -0.120, \Psi_3 \approx -0.180)$ and possessing long average path lengths (low network connectivity) reflecting the presence of three chains bound to a

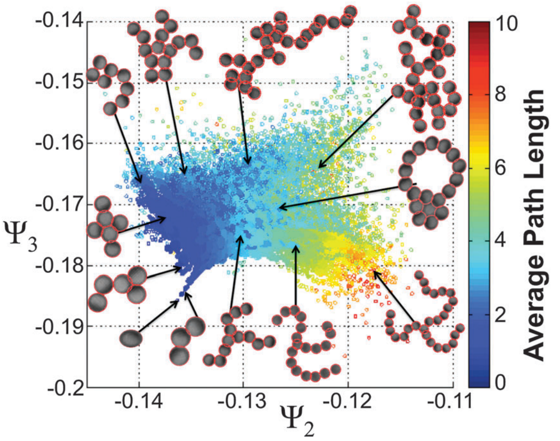

Fig. 2 Composite diffusion map embedding for the self-assembly of a mixture of active Janus particles with passive linker particles in the top two collective modes [Ψ_2,Ψ_3] furnished by the diffusion map. Each point represents one of the 3 403 918 clusters observed in the 28 experiments conducted over a range of AC frequencies and electric field strengths (*cf.* Table S1, ESI†). To aid in visualization, points are colored by the average path length between pairs of particles in the cluster bonding network, and representative aggregates superposed onto the manifold.

central linker particle. Finally, the assembly pathway leading to the upper right quadrant corresponds to the assembly of "archipelago" type structures in the region of ($\Psi_2 \approx -0.125$, $\Psi_3 \approx -0.160$), corresponding to clusters with locally dense packings connected by spanning chains.

By restricting the clusters projected into the composite diffusion map embedding to those observed under particular ranges of experimental conditions, we can determine the impact of experimentally controllable variables on the self-assembly behavior. Due to the wide range of AC frequencies investigated in this work, we bin the frequency distribution into two classes: low (70–300 kHz) and high (0.6–5 MHz). We reject the very high frequency regime (7–11 MHz) from this analysis due to poor characterization of the field strength arising from the capacitive behavior of the experimental setup. Similarly, we bin electric field strengths into two classes, low (250–500 V cm^{-1}) and high (583–833 V cm^{-1}). By binning our data, we accumulate more data within each class and improve the statistical robustness of our analysis, but elect to restrict our investigation to the high electric field strength regime due to relatively poor sampling under low electric field conditions (constituting just 9% of all observed particles combined over both frequency regimes). In Fig. 3a and b, we present the intrinsic manifold corresponding to each of the two frequency bins at high electric field strengths of 583–833 V cm^{-1}. In Fig. 3c and d, we present the analogous effective free energy landscapes, $\hat{F}(\Psi_2,\Psi_3)$, illustrating the relative probability of observing a single particle in a particular cluster configuration. Since each frequency bin contains data from multiple experimental trajectories, we average the probability distributions, $\hat{P}(\Psi_2,\Psi_3)$, extracted from each individual run, then compute the effective free energy as described in Section 2.3.3. Bootstrap estimates reveal the uncertainty in any bin of the effective free energy landscape to be less than $0.3k_BT$, with the largest errors occurring in regions of high effective free energy (Fig. S2, ESI†). At low frequencies (Fig. 3c), the dominant low effective free energy pathway connects free monomers to local effective free energy minima in the lower right quadrant at ($\Psi_2 \approx -0.120$, $\Psi_3 \approx -0.180$) containing pinwheels of characteristic size $N \sim 16$. The dense cluster architectures in the upper left quadrant and archipelago structures in the upper right quadrant are relatively disfavored, residing at higher effective free energies. We quantify the relative prevalence of the various cluster architectures by reporting on the figure the mass fraction of clusters projected into each quadrant of the landscape. At high frequencies (Fig. 3d), the size of the intrinsic manifold shrinks, reflecting the disappearance of the larger aggregates of size $N \gtrsim 12$ residing around the periphery of the landscape under these conditions. The topography of the effective free energy surface becomes flatter, attenuating the depth of the local minima within the pinwheel architectures. Similarly by considering the mass fraction of particles existing in the different assembly regimes as we shift from low to high AC frequencies, we see an increase in fraction of particle mass for small clusters (91 to 95%), a substantial reduction in mass fraction of pinwheels (7 to 3%), and small increases in the mass fraction of archipelago (0.7 to 0.8%) and dense cluster architectures (0.8 to 1.1%).

The nonequilibrium self-assembly is governed by a balance of hydrodynamics, dipole–dipole interactions, and rICEP motion. The decrease in volume of the intrinsic manifold at high frequencies discovered by the diffusion map is consistent with the frequency dependence of the Janus hemisphere polarizabilities and particle velocity under induced rICEP motion. We illustrate in Fig. 4a the frequency dependence of the real portion of the product of the hemisphere polarizabilities (*cf.* eqn (1)), and thus the relative magnitude of the different dipolar interactions between particles at [NaCl] = 0.1 mM. In the low frequency regime, $f \leq 300$ kHz, the metal–metal and metal–dielectric dipolar interactions are attractive, giving rise to more variety in energetically favorable configurations and enabling a larger volume of the intrinsic manifold to be explored. Upon increasing the frequency, metal–metal interactions become repulsive, limiting the possible structures that can be formed, favoring primarily head-to-tail configurations, and shrinking the manifold. In Fig. 4b, we present the frequency dependence of the ballistic velocity of a Janus particle moving under rICEP. These data show the particle velocity to increase steeply between the low and high frequency regimes. The velocity peaks at 5500 kHz, then decreases steadily to 44 MHz, beyond which there is a precipitous drop-off due to capacitive breakdown in our experimental system. We suggest that the combination of an increase in metal–metal repulsions and elevated particle velocity at high frequency serves to inhibit the formation of large clusters, as metal–silica attachments are disfavored and chain-like structures such as pinwheels shear off weakly bonded particles (*cf.* Movie S1, ESI†).

In sum, by applying our machine learning algorithm to experimental particle tracking data, we have extracted a two-dimensional assembly landscape revealing the presence of

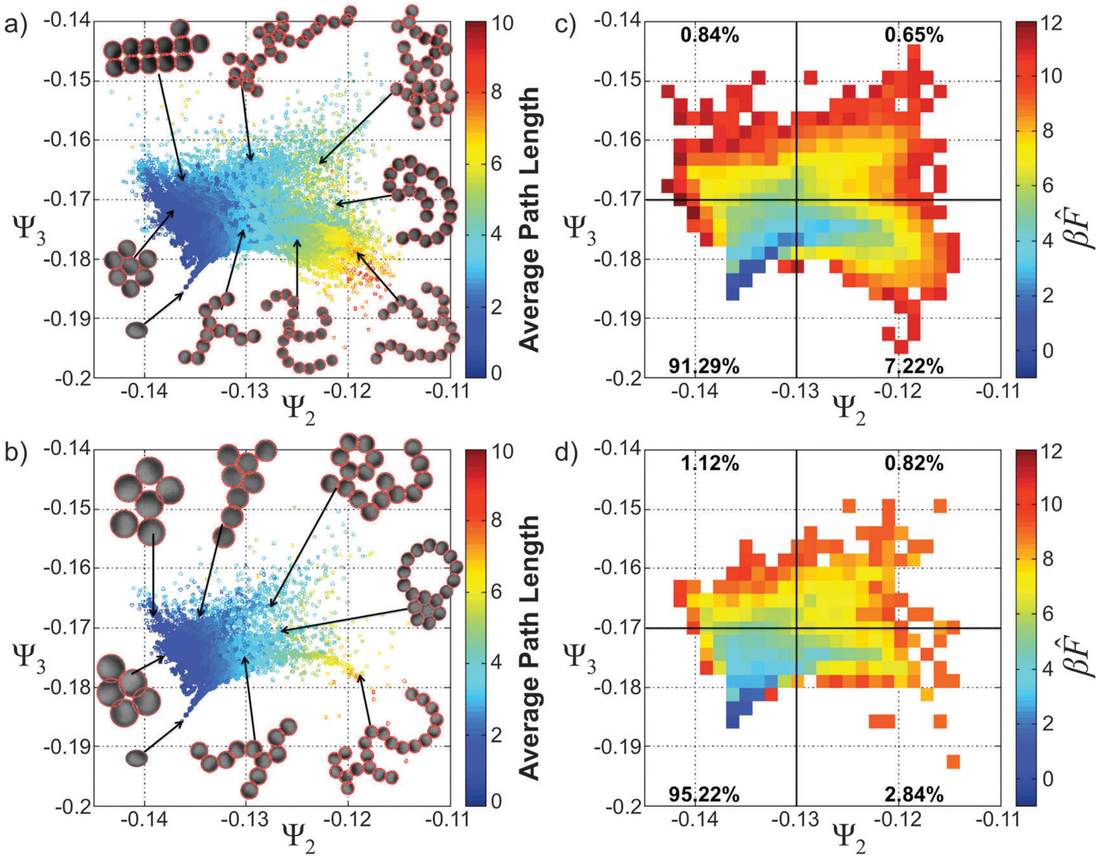

Fig. 3 Restriction of the composite diffusion map intrinsic manifold for the self-assembly of a mixture of active Janus particles with passive linker particles in Fig. 2 to the cluster ensembles extracted at (a) low (70–300 kHz) and (b) high (600 kHz–5 MHz) AC frequencies, *f*, at electric field strengths, *E*, of 583–833 V cm^{-1}. Points are colored by the average path length between pairs of particles in the cluster bonding network, and representative clusters superposed onto the manifolds. Effective free energy landscapes in the (c) low, and (d) high frequency regimes. The four percentages listed on panels (c) and (d) denote the mass fraction of particles residing within that quadrant.

three distinct families of aggregates. We have also determined how the landscape changes as a function of the frequency of the applied AC electric field and quantified the relative prevalence of the different cluster architectures and assembly pathways in good agreement with the physical understanding of the frequency response of the interparticle attractions and velocity. The assembly landscape provides quantitative insight into the relative prevalence of different cluster architectures, and provides a roadmap to tune experimental conditions to favor the assembly of desired aggregates. Specifically, low AC frequencies (70–300 kHz) preferentially favor the assembly of three-armed pinwheels of ~16 particles, whereas high frequencies (600 kHz–5 MHz) inhibit the formation of large clusters of $N \gtrsim 12$ particles.

3.2 Tunable Janus chain formation

We studied the self-assembly of homogeneous ensembles of metallodielectric Janus particles over AC frequencies of 70 kHz–11 MHz, electric field strengths of 167–833 V cm^{-1}, and NaCl concentrations of 0.01 mM and 0.1 mM. A total of 537 experiments were conducted over this full parameter space, with specific conditions considered listed in Tables S2 and S3 (ESI†) for [NaCl] = 0.01 mM and [NaCl] = 0.1 mM, respectively. The transient portion of each particle tracking trajectory was rejected such that the assembly dynamics of each system was studied at steady state. Analysis of all 537 experiments revealed a total of 739 246 clusters (including free monomers) belonging to 338 distinct architectures. A small number of clusters comprising more than 25 particles were very rarely observed, constituting just under 0.04% of the observed structures and appearing as extreme outliers in our diffusion map embeddings. It is known that the presence of rare observations disconnected from the bulk of the data can compromise the resolution of the diffusion map embedding,[56] in this case causing us to lose discriminatory power to resolve architectures and pathways at small cluster sizes of $N \lesssim 20$. Accordingly we followed our previously described "deislanding" approach to eliminate these rarely observed cluster aggregates from our analysis.[56] We applied diffusion maps to a subsampled ensemble of 34 151 clusters over the unique cluster architectures, employing a kernel bandwidth of ε = exp(5.5).[62] A gap in the eigenvalue spectrum after λ_3 (Fig. S3, ESI†), led us to construct two-dimensional diffusion map embeddings in $\{\vec{\Psi}_2,\vec{\Psi}_3\}$.[63] The remaining (739 246 − 34 151) = 705 095 were projected into the embedding using the Nyström extension.

We present in Fig. 5 the two-dimensional composite diffusion map embedding with points colored by cluster size.

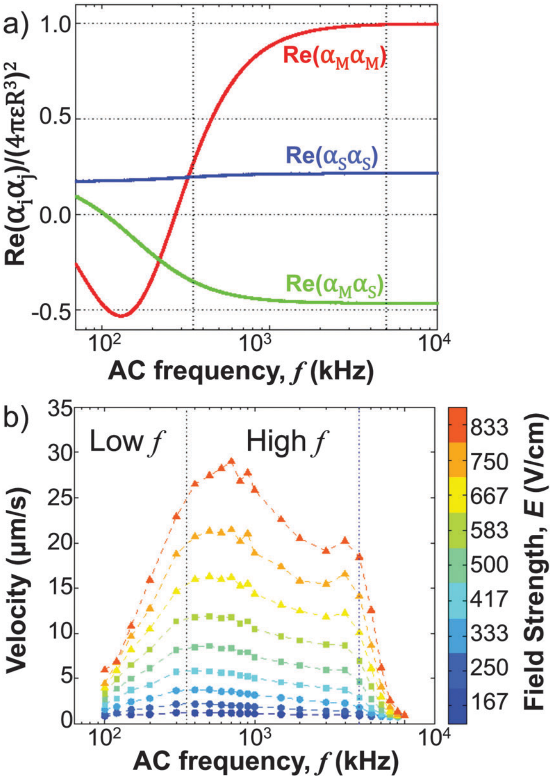

Fig. 4 Dependence of particle polarization and rICEP velocity upon AC frequency, f, at [NaCl] = 0.1 mM. (a) Dimensionless frequency dependent real portion of the product of polarizabilities of metal–metal (red), metal–silica (green), and silica–silica (blue) interactions with R = 1.5 μm and $\varepsilon/\varepsilon_0$ = 80, where ε_0 is the vacuum permittivity (*cf.* eqn (1)). (b) Ballistic velocity profile for Janus particles under dilute particle concentrations in a 0.1 mM NaCl solution as a function of AC frequency, f, at different electric field strengths, E. Consistent with the binning of the field strengths in the main text, circles correspond to low (167–333 V cm^{-1}), squares to intermediate (417–583 V cm^{-1}), and triangles to high (667–833 V cm^{-1}) field strengths. The vertical lines delineate the low (70–300 kHz) and high (400 kHz–5 MHz) frequency regimes.

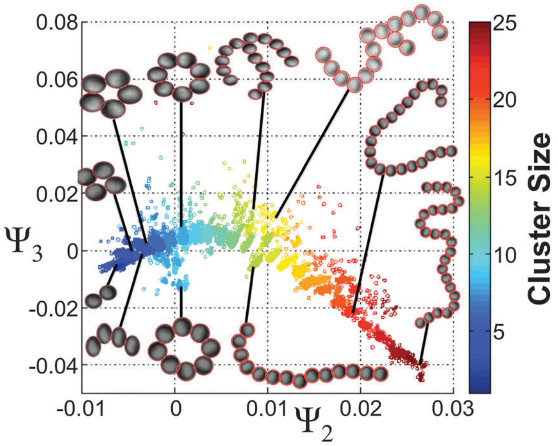

Fig. 5 Composite diffusion map embedding for the self-assembly of active Janus particles in the top two collective modes [Ψ_2,Ψ_3] furnished by the diffusion map. Each point represents one of the 739 246 clusters observed in the 537 experiments conducted over a range of AC frequencies, electric field strengths, and NaCl concentrations (*cf.* Tables S2 and S3, ESI†). To aid in visualization, points are colored by the number of particles in the cluster, and representative aggregates superposed onto the manifold.

Our analysis reveals a quasi one-dimensional assembly landscape. Advancing left to right along the principal axis of the manifold corresponds to the aggregation of progressively longer one-dimensional chains of Janus particles. Excursions perpendicular to the manifold correspond to the synthesis of Janus rings formed by a chain wrapping back on itself to "bite" its own tail, ejecting the excess particles that are pinched off in the formation of the ring. (We provide an illustration of this interesting process over the manifold in Fig. S4 (ESI†), showing the fragmentation of a 12-mer chain in this manner to form a 6-mer ring and a 6-mer chain.) Progressing further still from the manifold leads to a sparsely populated outer corona comprising of branched clusters typically formed by the collision of linear chains (*cf.* Movie S2, ESI†).

To quantify the impact of electric field strength, AC frequency, and salt concentration upon the self-assembly process, we again binned the experimental data into different regimes: (i) low (50–300 kHz), intermediate (400–800 kHz), high (900 kHz–3 MHz) AC frequency, f, (ii) low (167–333 V cm^{-1}), intermediate (417–583 V cm^{-1}), and high (667–833 V cm^{-1}) electric field strength, E, and (iii) low (0.01 mM) and high (0.1 mM) NaCl concentration, [NaCl]. We again neglect trends in the very high frequency regime due to capacitive effects in our experimental setup precluding accurate control of the field strength. We present in Fig. 6 the effective free energy landscapes in the nine different E–f regimes at low salt concentration. At high salt concentrations, we observe dramatically suppressed assembly behavior due to increased electrostatic screening from counterions in the solution, only observing significant aggregation in the high frequency regime where the induced dipoles are largest (*cf.* Fig. 4a). Accordingly, we relegate the [NaCl] = 0.1 mM data to Fig. S5 (ESI†), and all subsequent discussions pertain to [NaCl] = 0.01 mM unless otherwise noted.

From Fig. 6, we can readily infer how varying the frequency and field strength affects assembly. At low frequencies, self-assembly is limited to predominantly monomers and transient small aggregate formations. Higher frequencies stabilize the formation of Janus chains of varying sizes, with low field strengths producing chains of ∼8 particles while intermediate and high field strengths generate chains of size ∼10–15 particles. At high frequency, a larger range of structures are stabilized in the intermediate and high field regimes, corresponding to chains of ∼20–25 particles, as well as ring and branched structures of various sizes.

By partitioning the quasi-one dimensional assembly process into the different regions indicated in Fig. 7, we quantify the mass fraction of different chain (regions I, III, IV) and non-chain (regions II, IV, VI) aggregates to guide the design of experimental conditions to favor the assembly of desired aggregates. For example, if we are only concerned with maximizing the yield of intermediate length chains (region III), we should assemble the particles at high f – intermediate E to maximize

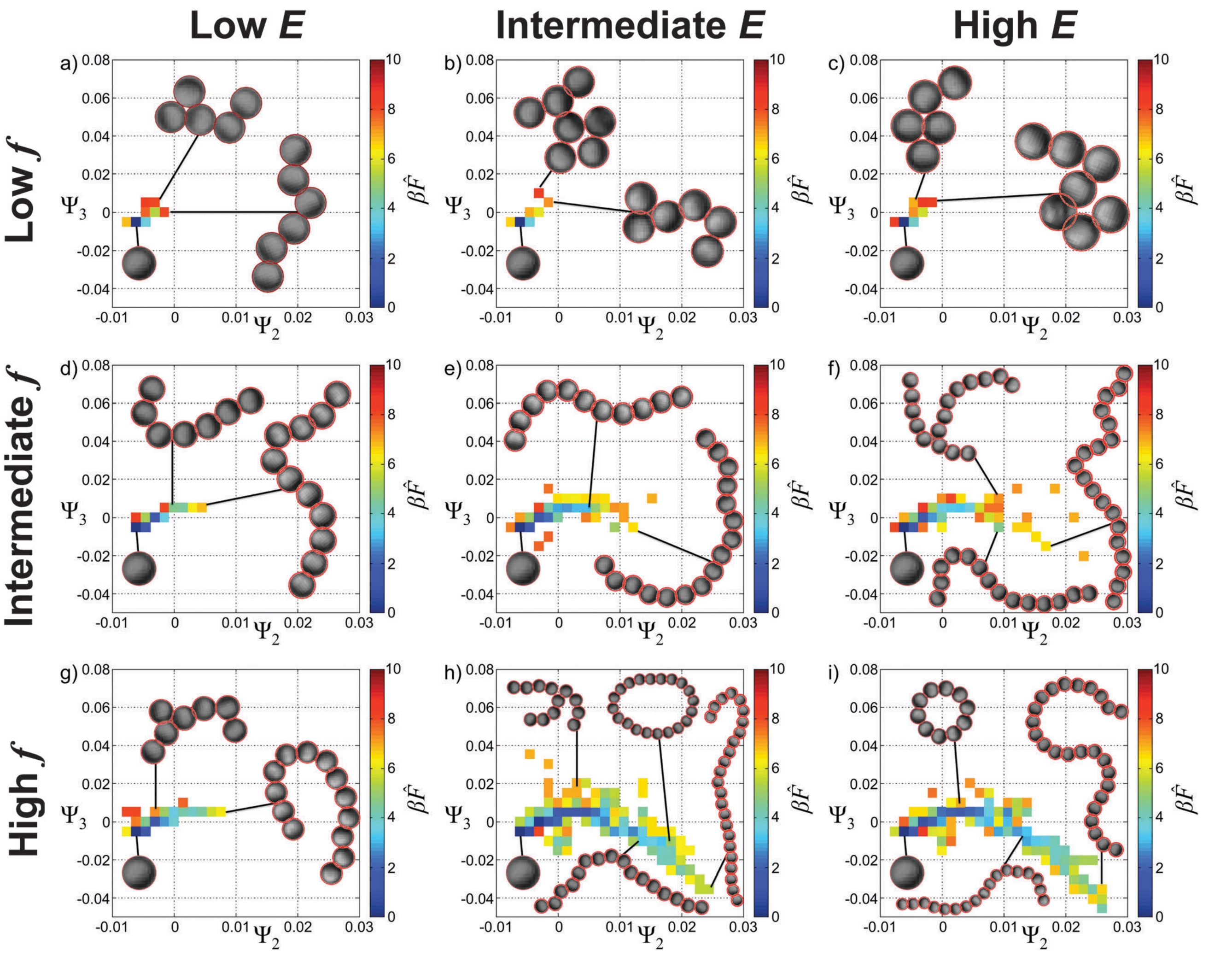

Fig. 6 Effective free energy landscapes for the self-assembly of active Janus particles at low (0.01 mM) NaCl concentration at different applied AC electric field strengths, *E*, and frequencies, *f*. Columns partition the electric field strength into low (167–333 V cm^{-1}), intermediate (417–583 V cm^{-1}), and high (667–833 V cm^{-1}) regimes. Rows split the AC frequency of the applied field into low (50–300 kHz), intermediate (400–800 kHz), and high (900 kHz–3 MHz) regimes.

the mass of the system residing in this region at 36.6%, with 6.0% residing in the non-chain regions II/IV/VI. If instead we wished to form intermediate length chains, but also limit formation of non-chain structures, we could shift to intermediate *f* – intermediate *E* (17.5% in region III, 1.9% in II/IV/VI), high *f* – low *E* (9.1%, 0.5%), or intermediate *f* – low *E* (3.7%, 0.3%) depending on our tolerance for non-chain aggregates.

By studying the changes to the assembly landscape as a function of the experimental conditions, we can relate these changes to the underlying particle behaviors. We first consider the dependence of assembly behavior upon AC frequency. From Fig. 6, we observe that larger clusters form at higher frequencies. This behavior can be understood from the dependencies upon AC frequency of the polarizability and velocity of Janus particles at [NaCl] = 0.01 mM presented in Fig. 8. The real portions of the product of the hemisphere polarizabilities are a weakly increasing function of frequency below 300 kHz, remaining effectively constant over the intermediate and high frequency range (*cf.* eqn (1)). Only the metal–dielectric interactions are favorable, consistent with the observation of primarily chain-like aggregates. The particle velocity, however, is highest in the low frequency regime, dropping sharply within the intermediate and high frequency regimes Fig. 8b. Particles in the low frequency regime therefore experience weaker interparticle attractions and higher kinetic energies, restricting self-assembly to predominantly monomers and dimers. We observe the assembly of heavier aggregates at intermediate and high frequencies due to an increase in attractive potential between particles and a decrease in particle velocity.

We now consider the dependence of assembly behavior upon electric field strength. From Fig. 6 and 7, the assembly landscape shows that increasing field strength beyond the low field strength regime leads to the formation of larger clusters at intermediate and high AC frequencies. At low frequencies, only small aggregate sizes are observed independent of field strength. By eqn (1), the interparticle attraction increases as the square of

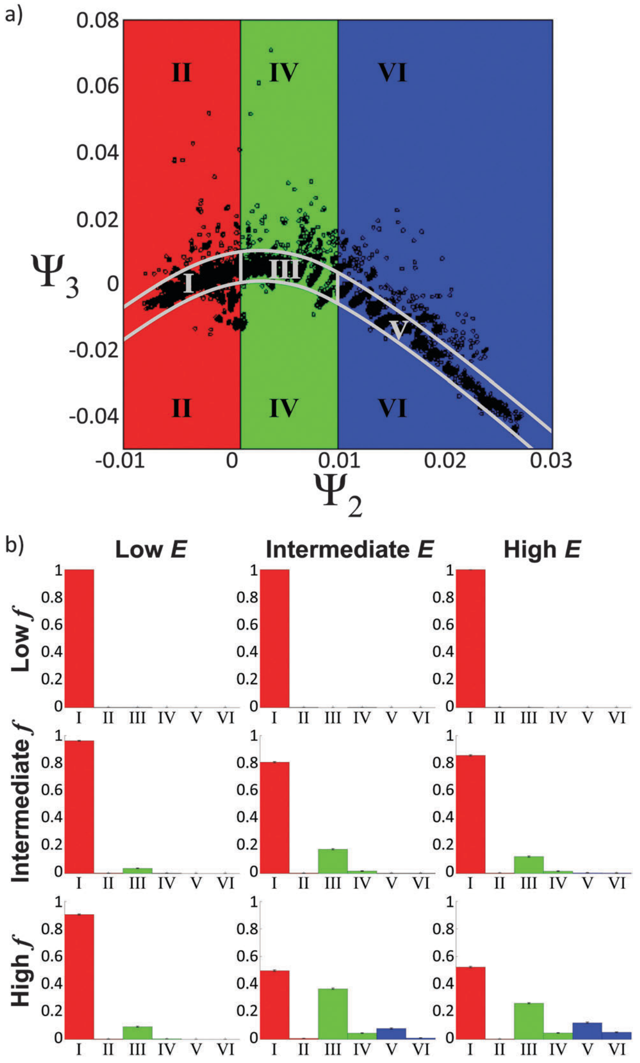

Fig. 7 Mass fraction of different chain architectures as a function of electric field strength and AC frequency for the self-assembly of active Janus particles. (a) Partitioning of diffusion map space into 6 distinct regimes corresponding to small (I and II), medium (III and IV), and large (V and VI) clusters, inside (I, III and V) and outside (II, IV and VI) of the linear chain envelope. (b) Fraction of system mass located inside these 6 distinct regions of self-assembly as a function of experimental conditions. Columns partition the electric field strength, *E*, into low (167–333 V cm^{-1}), intermediate (417–583 V cm^{-1}), and high (667–833 V cm^{-1}) regimes. Rows split the AC frequency of the applied field into low (50–300 kHz), intermediate (400–800 kHz), and high (900 kHz–3 MHz) regimes.

the electric field strength, E^2. Similarly from Fig. 8c, particle velocities also scale approximately as E^2. We observe a critical field strength for the assembly of aggregates of size $N \gtrsim 8$ residing between the low and intermediate field strength regimes. We suggest that this may be due to weak interparticle interactions and low particle velocities resulting in a reduced likelihood for the particles to come into contact range and bind. At intermediate and high frequencies, shifting from intermediate to high field strengths yields a small increase in the stability of larger aggregate structures. Although assembly behavior is heavily suppressed under the high salt conditions, we

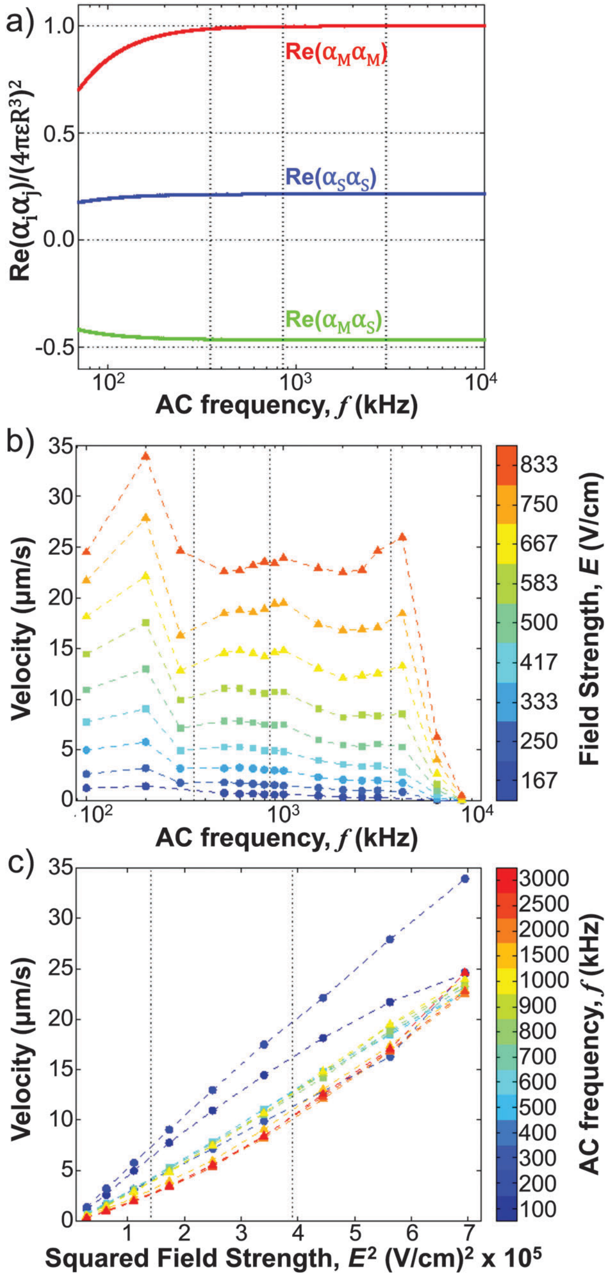

Fig. 8 Dependence of particle polarization and rICEP particle velocity upon AC frequency and electric field strength at [NaCl] = 0.01 mM. (a) Dimensionless real portion of product of hemispheric polarizabilities as a function of AC frequency, *f*, with *R* = 1.5 μm and $\varepsilon/\varepsilon_0$ = 80, where ε_0 is the vacuum permittivity (*cf.* eqn (1)). (b) Dependence of Janus particle ballistic velocities as a function of AC frequency, *f*, at different electric field strengths, *E*. Consistent with the binning of the field strengths in the main text, circles correspond to low (167–333 V cm^{-1}), squares to intermediate (417–583 V cm^{-1}), and triangles to high (667–833 V cm^{-1}) field strengths. The vertical lines selected to separate the low (70–300 kHz), intermediate (400–800 kHz), and high (900 kHz–3 MHz) frequency regimes. (c) Dependence of Janus particle ballistic velocities as a function of squared electric field strengths, E^2, at different AC frequencies, *f*. Consistent with the binning of the field strengths in the main text, circles correspond to low (70–300 kHz), squares to intermediate (400–800 kHz), and circles to high (900 kHz–3 MHz) frequency regimes. The vertical lines selected to separate the low (167–333 V cm^{-1}), intermediate (417–583 V cm^{-1}), and high (667–833 V cm^{-1}) field strengths.

observe a similar trend at [NaCl] = 0.1 mM in the high frequency regime (Fig. S5, ESI†). The observation of markedly larger aggregates at high field strengths demonstrates that despite similar E^2 scaling of velocity and attraction, the balance of these two effects results in an elevated aggregation propensity at larger values of E. We are quick to note that more detailed analysis and understanding of rICEP and electrohydrodynamic motion, may be necessary to fully account for the observed trends.[48,70,71] Nevertheless, a complete understanding of the physics underpinning the system behavior is not required by our data-driven approach, which can inform understanding of assembly by empirically quantifying assembly behaviors from particle tracking data.

In sum, we have used our machine learning framework to construct two-dimensional embeddings capturing the self-assembly of metallodielectric Janus particles, and mapped the regions of configuration space accessible under various salt concentrations, AC frequencies, and electric field strengths. The manifolds generated provide new insights for our understanding of self-assembly in this system, and provide a roadmap showing how to control assembly behavior by manipulating experimental conditions. Low AC frequencies suppress assembly at all field strengths and salt concentrations, resulting in essentially only free monomers. To produce chain aggregates, we can move into the intermediate frequency regime, where for low salt conditions we can generate chains spanning from ∼8 particles under a low applied field to ∼10–15 particles in the intermediate and high field strength regimes. Finally, moving into the high frequency regime, where form diverse architectures including rings and ramified structures, as well as chains in excess of ∼20 particles.

4 Conclusions

We have presented an approach to infer low-dimensional roadmaps of self-assembly by analyzing experimental particle tracking trajectories using diffusion maps. The variables spanning the low-dimensional embeddings discovered by the diffusion map correspond to the slow collective modes governing the long-time dynamics of assembly, and reveal the range of accessible aggregates and assembly pathways without requiring prior knowledge of the underlying physics governing particle aggregation or motion. The effective free energy landscapes constructed over these embeddings reveal the relative prevalence and stability of different cluster architectures. By recovering these landscapes under different experimental conditions, we can quantify the impact of experimentally controllable parameters on the topography of the free energy surface. The findings from this data-driven approach may be compared and rationalized with existing theoretical and experimental understanding of the system, or used to help infer the properties of novel or poorly characterized systems. Furthermore, these landscapes furnish quantitative understanding of the system response to externally imposed conditions, and can guide the tuning of these conditions to favor desired cluster architectures.

We have demonstrated our methodology in applications to the nonequilibrium assembly of Janus particles subjected to an oscillating electric field that drives interparticle attractions and self-propelled particle motion. In an application to the templated assembly of active Janus particles and passive linker particles, our approach revealed the existence of an effectively two-dimensional embedding comprising three different assembly routes leading to dense disordered clusters, three-armed pinwheels, and extended archipelago topologies. We then demonstrated that the relative stability of these various architectures could be tuned by manipulating the AC frequency of the applied field, and showed that this behavior was in good accord with the physical understanding of the particle response to the AC field. In a second application to homogeneous ensembles of active Janus particles, our method discovered a quasi one-dimensional projection mapping the range of accessible cluster architectures. The principal axis of the manifold traced the assembly of progressively longer linear chains of Janus particles, with departures perpendicular to the manifold corresponding to the formation of rings and branched chains of particles. We quantified the relative stability of the various aggregates as a function of electric field strength, AC frequency, and salt concentration, showing how these three control parameters can be simultaneously manipulated to favor the assembly of clusters of desired size and architecture. By relating these empirical findings to our understanding of the underlying particle physics, we found good agreement for our model of assembly behavior as a function of these experimental controls.

By integrating experimental particle tracking technology with sophisticated machine learning tools, this work presents a new approach to infer assembly pathways and attainable aggregates directly from experimental data, providing insight into the interplay of different experimentally controllable parameters on assembly behavior, and informing the rational design of conditions favoring the assembly of desired structures. In future work, we intend to apply our approach to distinguishable multicomponent systems, where we can utilize particle identity information to improve the mapping procedure and identify novel assembly behavior as a function of the different component concentrations. We also intend to ''close the feedback loop'' to establish an iterative design process wherein the insights and design rules discovered by machine learning are used to inform rational redesign of particle properties to favor assembly of desired aggregates. We anticipate that with continued development this approach will improve our understanding and control of self-assembly processes, and help forge a powerful new pathway to rationally engineer novel self-assembling materials with desired structure and function.

Acknowledgements

The computational material is based upon work supported by a National Science Foundation CAREER Award to A.L.F. (Grant No. DMR-1350008). The experiments (J.Z.) were supported by the U.S. Department of Energy, Division of Materials Science, under Award DEFG02-02ER46019. S.G. acknowledges office support from the Institute for Basic Science, Project Code IBS-R020-D1.

References

1 G. M. Whitesides and B. Grzybowski, *Science*, 2002, **295**, 2418–2421.

2 J. Sticht, M. Humbert, S. Findlow, J. Bodem, B. Muller, U. Dietrich, J. Werner and H.-G. Krausslich, *Nat. Struct. Mol. Biol.*, 2005, **12**, 671–677.

3 S. C. Glotzer and M. J. Solomon, *Nat. Mater.*, 2007, **6**, 557–562.

4 Q. Meng, Y. Kou, X. Ma, Y. Liang, L. Guo, C. Ni and K. Liu, *Langmuir*, 2012, **28**, 5017–5022.

5 M. F. Hagan and D. Chandler, *Biophys. J.*, 2006, **91**, 42–54.

6 A. T. Da Poian, A. C. Oliveira and J. L. Silva, *Biochemistry*, 1995, **34**, 2672–2677.

7 H. Ning, A. Mihi, J. B. Geddes, M. Miyake and P. V. Braun, *Adv. Mater.*, 2012, **24**, OP153–OP158.

8 K. A. Arpin, M. D. Losego, A. N. Cloud, H. Ning, J. Mallek, N. P. Sergeant, L. Zhu, Z. Yu, B. Kalanyan, G. N. Parsons, G. S. Girolami, J. R. Abelson, S. Fan and P. V. Braun, *Nat. Commun.*, 2013, **4**, 1–8.

9 H. Zhang, R. Wu, Z. Chen, G. Liu, Z. Zhang and Z. Jiao, *CrystEngComm*, 2012, **14**, 1775–1782.

10 S. C. Glotzer, M. J. Solomon and N. A. Kotov, *AIChE J.*, 2004, **50**, 2978–2985.

11 S. Whitelam, E. H. Feng, M. F. Hagan and P. L. Geissler, *Soft Matter*, 2009, **5**, 1251–1262.

12 M. J. Solomon, *Curr. Opin. Colloid Interface Sci.*, 2011, **16**, 158–167.

13 E. Jankowski and S. C. Glotzer, *J. Phys. Chem. B*, 2011, **115**, 14321–14326.

14 E. Jankowski and S. C. Glotzer, *Soft Matter*, 2012, **8**, 2852–2859.

15 W. L. Miller and A. Cacciuto, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2009, **80**, 021404.

16 W. L. Miller and A. Cacciuto, *J. Chem. Phys.*, 2010, **133**, 234108.

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

18 A. W. Wilber, J. P. K. Doye, A. A. Louis, E. G. Noya, M. A. Miller and P. Wong, *J. Chem. Phys.*, 2007, **127**, 085106.

19 A. W. Wilber, J. P. K. Doye, A. A. Louis and A. C. F. Lewis, *J. Chem. Phys.*, 2009, **131**, 175102.

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

21 I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, 2nd edn, 2002.

22 G. G. Maisuradze, A. Liwo and H. A. Scheraga, *J. Mol. Biol.*, 2009, **385**, 312–329.

23 P. I. Zhuravlev, C. K. Materese and G. A. Papoian, *J. Phys. Chem. B*, 2009, **113**, 8800–8812.

24 M. Ceriotti, G. A. Tribello and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 13023–13028.

25 G. A. Tribello, M. Ceriotti and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 5196–5201.

26 S. T. Roweis and L. K. Saul, *Science*, 2000, **290**, 2323–2326.

27 M. Belkin and P. Niyogi, *Neural Comput.*, 2003, **15**, 1373–1396.

28 J. B. Tenenbaum, V. De Silva and J. C. Langford, *Science*, 2000, **290**, 2319–2323.

29 P. Das, M. Moll, H. Stamati, L. E. Kavraki and C. Clementi, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 9885–9890.

30 E. Plaku, H. Stamati, C. Clementi and L. E. Kavraki, *Proteins: Struct., Funct., Bioinf.*, 2007, **67**, 897–907.

31 H. Stamati, C. Clementi and L. E. Kavraki, *Proteins: Struct., Funct., Bioinf.*, 2010, **78**, 223–235.

32 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.

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

34 R. Coifman, I. Kevrekidis, S. Lafon, M. Maggioni and B. Nadler, *Multiscale Model. Simul.*, 2008, **7**, 842–864.

35 A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti and I. G. Kevrekidis, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 13597–13602.

36 M. A. Rohrdanz, W. Zheng, M. Maggioni and C. Clementi, *J. Chem. Phys.*, 2011, **134**, 124116.

37 R. A. Mansbach and A. L. Ferguson, *J. Chem. Phys.*, 2015, **142**, 105101.

38 P. G. de Gennes, *Rev. Mod. Phys.*, 1992, **64**, 645–648.

39 S. Jiang, Q. Chen, M. Tripathy, E. Luijten, K. S. Schweizer and S. Granick, *Adv. Mater.*, 2010, **22**, 1060–1071.

40 S. Granick, S. Jiang and Q. Chen, *Phys. Today*, 2009, **62**, 68–69.

41 L. Hong, A. Cacciuto, E. Luijten and S. Granick, *Langmuir*, 2008, **24**, 621–625.

42 Q. Chen, J. K. Whitmer, S. Jiang, S. C. Bae, E. Luijten and S. Granick, *Science*, 2011, **331**, 199–202.

43 Q. Chen, S. C. Bae and S. Granick, *Nature*, 2011, **469**, 381–384.

44 A. Walther and A. H. E. Müller, *Chem. Rev.*, 2013, **113**, 5194–5261.

45 A. Walther and A. H. E. Müller, *Soft Matter*, 2008, **4**, 663–668.

46 S. Gangwal, O. J. Cayre, M. Z. Bazant and O. D. Velev, *Phys. Rev. Lett.*, 2008, **100**, 058302.

47 S. Gangwal, O. J. Cayre and O. D. Velev, *Langmuir*, 2008, **24**, 13312–13320.

48 S. Wang, F. Ma, H. Zhao and N. Wu, *ACS Appl. Mater. Interfaces*, 2014, **6**, 4560–4569.

49 T. Atherton and D. Kerbyson, *Image Vision Comput.*, 1999, **17**, 795–803.

50 R. Tarjan, *SIAM J. Comput.*, 1972, **1**, 146–160.

51 P. Garca-Sánchez, Y. Ren, J. J. Arcenegui, H. Morgan and A. Ramos, *Langmuir*, 2012, **28**, 13861–13870.

52 V. Shilov, A. Delgado, F. Gonzalez-Caballero and C. Grosse, *Colloids Surf., A*, 2001, **192**, 253–265.

53 J. C. Crocker and D. G. Grier, *J. Colloid Interface Sci.*, 1996, **179**, 298–310.

54 A. S. Keys, C. R. Iacovella and S. C. Glotzer, *J. Comput. Phys.*, 2011, **230**, 6438–6463.

55 R. Zwanzig, *Nonequilibrium Statistical Mechanics*, Oxford University Press, USA, 2001.

56 A. L. Ferguson, S. Zhang, I. Dikiy, A. Z. Panagiotopoulos, P. G. Debenedetti and A. J. Link, *Biophys. J.*, 2010, **99**, 3056–3065.

57 P. G. Bolhuis, C. Dellago and D. Chandler, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 5877–5882.

58 G. W. Klau, *BMC Bioinf.*, 2009, **10**, S59.

59 M. Zaslavskiy, F. Bach and J.-P. Vert, *Bioinformatics*, 2009, **25**, 1259–1267.

60 R. Singh, J. Xu and B. Berger, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 12763–12768.

61 B. Nadler, S. Lafon, R. R. Coifman and I. G. Kevrekidis, *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference (Neural Information Processing)*, The MIT Press, 2006, pp. 955–962.

62 R. Coifman, Y. Shkolnisky, F. Sigworth and A. Singer, *IEEE Trans. Image Process.*, 2008, **17**, 1891–1899.

63 S. Salvador and P. Chan, Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on, 2004, pp. 576–584.

64 A. Ma and A. R. Dinner, *J. Phys. Chem. B*, 2005, **109**, 6769–6779.

65 B. Peters and B. L. Trout, *J. Chem. Phys.*, 2006, **125**, 054108.

66 J. Xing and K. S. Kim, *J. Chem. Phys.*, 2011, **134**, 1–11.

67 C. T. Baker and C. Baker, *The numerical treatment of integral equations*, Clarendon Press, Oxford, 1977, vol. 13.

68 C. R. Laing, T. A. Frewen and I. G. Kevrekidis, *Nonlinearity*, 2007, **20**, 2127.

69 B. E. Sonday, M. Haataja and I. G. Kevrekidis, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2009, **80**, 031102.

70 M. Trau, D. A. Saville and I. A. Aksay, *Langmuir*, 1997, **13**, 6375–6381.

71 W. D. Ristenpart, I. A. Aksay and D. A. Saville, *J. Fluid Mech.*, 2007, **575**, 83–109.