Unsupervised machine learning for detection of phase transitions in off-lattice systems. II. Applications ⊘

R. B. Jadrich; B. A. Lindquist; W. D. Piñeros; D. Banerjee; T. M. Truskett

Check for updates

CHORUS

View Online   Export Citation   CrossMark

Unsupervised machine learning for detection of phase transitions in off-lattice systems. II. Applications

R. B. Jadrich,[1] B. A. Lindquist,[1] W. D. Piñeros,[2] D. Banerjee,[1] and T. M. Truskett[1,3,a]

[1]*McKetta Department of Chemical Engineering, University of Texas at Austin, Austin, Texas 78712, USA*
[2]*Department of Chemistry, University of Texas at Austin, Austin, Texas 78712, USA*
[3]*Department of Physics, University of Texas at Austin, Austin, Texas 78712, USA*

We outline how principal component analysis can be applied to particle configuration data to detect a variety of phase transitions in off-lattice systems, both in and out of equilibrium. Specifically, we discuss its application to study (1) the nonequilibrium random organization (RandOrg) model that exhibits a phase transition from quiescent to steady-state behavior as a function of density, (2) orientationally and positionally driven equilibrium phase transitions for hard ellipses, and (3) a compositionally driven demixing transition in the non-additive binary Widom-Rowlinson mixture. *Published by AIP Publishing.* https://doi.org/10.1063/1.5049850

I. INTRODUCTION

Principal component analysis (PCA) is a simple and widely used unsupervised machine learning tool for dimensionality reduction.[1–3] Perhaps the most common application of PCA is for the lossy compression of images. One popular demonstration is the analysis of facial images, leading to the aptly named "eigenfaces" that capture collective attributes of facial structure.[1,2] Only a subset of the eigenfaces—much fewer than the naïve dimensionality of the problem—are required to recover the salient aspects of facial images by simple linear combination. Another routine use is in natural language processing, where PCA is employed to shrink the data dimensionality down from the large number of words appearing in a data set or in a dictionary.[1,2] Use of the resultant lower dimensional representation greatly improves the development of predictive models to classify text documents.

The combined power and simplicity of PCA has made it a popular tool in the biological and physical sciences as well. For example, DNA microarray data are routinely treated with PCA to reduce the high dimensionality of the problem in order to identify unique gene expression states across various experimental conditions.[2,4] Furthermore, PCA is commonly leveraged to extract dominant collective modes in simulations of proteins, referred to as "Essential Dynamics" in that field.[5,6] More recently, various spin models from statistical physics have been investigated via PCA and other machine learning methods.[7–14] These studies have demonstrated the ability of machine learning tools to detect and quantify phase transitions by the autonomous construction of an order parameter (OP).

The aforementioned work on phase transitions in spin models served as motivation for this two-part series of papers. In the first manuscript (henceforth referred to as Paper I[52]),

we developed guidelines for the utilization of PCA[1–3] to detect phase transitions in off-lattice, particle-based systems. We also demonstrated that PCA can readily identify the freezing transitions in hard disks and hard spheres, as well as liquid-gas phase separation in a binary mixture of patchy particles with complementary attractions. In developing and evaluating this approach, we initially focused on phase transitions that were equilibrium in nature and could be identified on the basis of features reflecting the positional degrees of freedom of the particles.

Here, we seek to generalize the formalism developed in Paper I[52] to assess its utility for detecting phase transitions in a broader class of systems. Examples include equilibrium systems with (1) anisotropic particles leading to orientational as well as positional ordering[15,16] and (2) compositional degrees of freedom that can induce demixing.[17] We also address active or driven matter, which exhibits phase transitions whose detection and characterization cannot generally be facilitated based on arguments from equilibrium statistical mechanics.[18–27]

We propose several numerical encoding schemes (i.e., feature vector representations) for data describing particle configurations in these systems to detect their phase transitions with PCA. We find that prior knowledge of the phase transition is not required to construct a useful feature vector; consideration of the properties of the model system at hand is sufficient. However, we also show that by performing PCA on several choices for the feature vector, one can gain physical insights into the nature of the phase transition.

The balance of the manuscript is organized as follows. In Sec. II, considerations for constructing features for the detection of phase transitions in off-lattice systems using PCA are presented. The model systems analyzed in this work and the corresponding simulation details for each model are also provided. Section III is divided into three subsections, each dedicated to a different model system. The first, Sec. III A, describes a study of the random

---

a)Electronic mail: truskett@che.utexas.edu

    149, 194110-1

organization model, which exhibits a nonequilibrium phase transition between a quiescent state and a dynamically evolving steady state as a function of increasing density.[28–31] Section III B addresses the fluid-nematic (orientationally driven) and the nematic-solid (positionally driven) phase transitions that occur upon densification of hard ellipses.[32–35] Finally, in Sec. III C, compositional demixing in the Widom-Rowlinson (WR) model—a binary mixture where unlike particles interact via excluded volume effects but like particles are noninteracting[36–38]—is explored. Concluding remarks are presented in Sec. IV.

II. METHODS

A. Feature construction

Features (f_i) are scalar quantities that inform a machine learning algorithm about some aspect of the system being studied.[1,2] Here, we denote a general vector of m features as

$$f \equiv \left[f_1, f_2, \ldots, f_m \right]^T, \tag{1}$$

where T indicates a transpose. Feature vectors provide a numerical encoding for the separate realizations (or measurements) contained in the data set ($\mathcal{D}$).

When possible, features should reflect any known constraints; for physics problems, these include invariance to translation and rotation.[39–42] Such constraints can be easily encoded via the use of internal coordinates (e.g., interparticle distances or relative angular orientations) as features. Here, we compute pairwise quantities $g_\beta^{(\alpha)}$ that are in reference to a probe particle (α) and a corresponding particle in its environment (β). A feature vector built from information considering n_P probe particles (each with n_{NN} corresponding environmental particles) can be represented as

$$
\begin{aligned}
f &= \left[g_1^T, \ g_2^T, \ \ldots, \ g_{n_P}^T \right]^T, \\
g_\alpha^T &\equiv \left[g_1^{(\alpha)}, \ g_2^{(\alpha)}, \ \ldots, \ g_{n_{NN}}^{(\alpha)} \right],
\end{aligned}
\tag{2}
$$

where the full vector of vectors g_α^T corresponding to each probe particle α is implicitly flattened to form one contiguous feature vector (block matrix notation).

Within the above mathematical framework, there is no unique choice for the selection of either the probe particles or the neighboring particles that define their environment. Once a collection of probe and corresponding environment particles are chosen, we also must specify how the resultant pairwise quantities ($g_\beta^{(\alpha)}$) are assigned to the α and β indices in Eq. (2). So we do not have to compute properties with respect to every particle in the simulation box, and we select n_P probe particles at random. For the corresponding environmental particles, we use physical intuition as a guide by assuming that the distance between the probe particle and a given environmental particle $r_\beta^{(\alpha)}$ will influence the manner in which the associated feature $g_\beta^{(\alpha)}$ reports on a given phase transition. As a result, we use a distance-based criterion to determine which particles comprise the environment for a given probe (e.g., the first twenty nearest neighbors or every tenth nearest neighbor), hence our use of n_{NN} to denote the number of environmental particles.

Similarly, we assign the index β on the basis of interparticle distance so that

$$r_1^{(\alpha)} \leq r_2^{(\alpha)} \leq \cdots \leq r_{n_{NN}}^{(\alpha)}. \tag{3}$$

The assignment of a probe particle to a given α is less intuitive and could be model-dependent; however, random assignment is always a possibility, and, as we discuss below, the results obtained from that initial assignment can in some cases help identify a superior assignment scheme for the probe particles.

In principle, n_P could be as large as the number of particles in the simulation, N, and n_{NN} could have a maximum value of $N - 1$. Since the total feature vector size [in relation to Eq. (1)] is $m = n_P \times n_{NN}$, the preceding choices would yield a feature vector of length $N(N - 1)$. For most systems of interest, PCA for feature vectors of this size would be computationally infeasible. Therefore, practical implementation of PCA using particle-based coordinate data requires sensible choices for n_P and n_{NN} that we describe in Subsections II B 1–II B 3.

Finally, we refer to features where $g_\beta^{(\alpha)}$ are physically motivated quantities as "intuited" features (f_I). In Paper I,[52] we showed that f_I do not necessarily approximate white noise in the disordered reference state (here, the ideal gas) limit and therefore may possess correlations that could obscure the detection of a phase transition via PCA. Arriving at corrected features (f_C) that are linearly decorrelated when applied to an ideal gas reference data set ($\mathcal{D}_0$) is accomplished by deriving a PCA whitening transformation[43] ($f_I \rightarrow f_C$) that satisfies $\langle f_C f_C^T \rangle_{\mathcal{D}_0} = I$, where I is the unit matrix and $\langle \ldots \rangle_{\mathcal{D}_0}$ is an average over the reference data.

B. Models

We provide a brief description of each model examined in this work as well as the relevant phase transition(s) below. We then specify the form of the associated feature vectors within the framework provided by Eq. (2). Finally, we describe the simulation protocols used to generate the configurations on which the PCA is performed. Throughout, N denotes the number of particles in a two-dimensional (2D) periodically replicated simulation cell of area A, $\rho = N/A$ is the number density, and $\eta = \rho \pi \sigma^2 / 4$ is the packing fraction. For the remainder of the paper, we set $\sigma = 1$, implicitly rendering all length and volume units dimensionless with respect to σ.

1. Random organization model

In one variant of the Random Organization (RandOrg) model, a circular particle of diameter σ is defined as active if it overlaps with any other particle.[30,31] For a given configuration, all active particles are simultaneously given random displacements; all other particle positions are unaltered. Particle positions are initialized at random, from which the above procedure is repeated until either (1) a so-called absorbing state is reached where no particle overlaps are present (lower densities) or (2) a steady-state is reached where the fraction of active particles fluctuates about some non-zero value (higher densities).

Given that the RandOrg model comprises identical, radially symmetric particles, features that explicitly encode positional packing correlations around tagged particles are an obvious first choice to try. Specifically, we utilize mean subtracted interparticle distances as our features,

$$g_\beta^{(\alpha)} = r_\beta^{(\alpha)} - \langle r_\beta^{(\alpha)} \rangle_\mathcal{D}. \tag{4}$$

Furthermore, while the model is technically single-component to the extent that there are no immutable labels associated with the particles, multiple particle types (active and inactive) are created on-the-fly due to the dynamics prescribed by the model. To capture emergent inhomogeneity with respect to the particle environment, it is critical to utilize multiple probe particles as prescribed by Eq. (2). In the present work, we use a fixed feature length of $m = n_P \times n_{NN} = 400$ and examine the effects of co-varying n_{NN} and n_P.

Within the above approach, there is still the question of how to assign the probe particles to specific values of α in Eq. (2). One valid, though perhaps not particularly informative, choice is to randomly order the probe particles such that α assignment does not encode any information. In Sec. III A, we demonstrate how performing PCA with this choice produces results that suggest a more informative sorting scheme, where probe particles are assigned to the index α on the basis of their first NN distance, i.e., $r_1^{(1)} \leq r_1^{(2)} \leq \cdots \leq r_1^{(n_P)}$.

We note two additional technical points regarding PCA for the RandOrg model. First, as we increase n_P, the magnitude of the OP grows in a nonlinear fashion. We can collapse OPs onto the same scale by dividing by the square root of the explained variance of their dimension, a procedure equivalent to data "whitening" discussed in Paper I[52] in the context of correcting the physics-motivated features. We also find that OPs obtained from both dimensional (preserving units of distance) and nondimensionalized features (dividing raw distances by $\rho^{-1/D}$, where D is the dimension) accurately detect the phase transition of the RandOrg model. However, as we demonstrate, the PCA-derived OP using the former convention shows behavior that is more strikingly reminiscent of the classical OP for this system.

To generate the configuration data required to construct the above features, we performed 2D simulations in a square box with $N = 1000$ particles and employed a maximum displacement of 0.25σ in both the x and y directions for the active particles. The length of the simulations varied with proximity to the critical point characteristic of the transition between an absorbing and a steady state. At densities below the critical point, an individual simulation ended when the absorbing state is reached; however, critical slowing down impacts the simulation length required to achieve that state.[29,30] We used a maximum of 10^5 simulation steps for densities below the critical point. For higher densities, the fraction of active particles decreased from the initial random state before fluctuating about a steady-state. The number of simulation steps was chosen to be at least twice as long as the initial relaxation time scale, ranging from 10^3 steps (at the highest densities) to 10^5 steps (just past the critical point). We performed 10^3 separate simulations, using only the last frame from the simulation in the PCA. Values for ρ ranging from 0.38 to 0.63 were simulated

in increments of 0.005. From the simulation data, we computed 25 feature vectors from each simulation snapshot, where the probe particles were selected at random. Within a single feature vector, probe particles are selected without replacement; however, a particular probe particle can appear in multiple feature vectors.

2. Hard ellipses

Densification of hard ellipses bears similarity to the freezing of hard disks studied in Paper I[52] but with added complexity derived from particle-shape anisotropy. In addition to ordering on the center-of-mass positional level, quasi-long ranged orientational ordering is possible, yielding the so-called nematic phase.[15,16] Two obvious pairwise properties to compute from the configurational data of hard ellipses are the center-of-mass distances and the relative orientations of the ellipses. With respect to the former, we use the positional features with a single probe particle as employed in Paper I[52] for hard disks and spheres. This form is equivalent to the feature vector defined by Eqs. (2)–(4) for the case where $n_P = 1$ and $n_{NN} = N - 1$, where N is the number of particles. Subsequently, the size of the feature vector is reduced by only including every 10th NN distance after the first feature in the final feature vector. The pairwise distances are normalized with respect to the mean interparticle spacing $l \equiv \rho^{-1/D}$, where ρ is the number density and D is the spatial dimensionality, to remove trivial density scaling.

For the latter case of relative orientations, we still employ one probe and index its environmental particles on the basis of NN sorting [Eqs. (2) and (3)]; however, we use a measure of relative pair orientations in place of pair distances in the feature vector. Defining $\delta\theta_\beta^{(\alpha)}$ as the angular difference between the probe and environmental particles assigned to indices α and β, respectively, we employ features of the form

$$g_\beta^{(\alpha)} = |\cos(\delta\theta_\beta^{(\alpha)})| - \langle|\cos(\delta\theta_\beta^{(\alpha)})|\rangle_\mathcal{D}. \tag{5}$$

That is, $g_1^{(1)}$ quantifies the relative orientation of the single probe particle with its closest NN ellipse. From the sorted list defined by the combination of Eqs. (2), (3), and (5), only every 10th NN is included in the feature vector, as was performed for the positional features above.

The feature vectors used as input to PCA were collected from Monte Carlo simulations of hard ellipses carried out at a constant particle number and volume using the HOOMD-blue software package.[44–46] The box shape was chosen to approximate a square by an appropriate distribution of triangular lattice cells with an aspect ratio of $\sqrt{3}\kappa$, where $\kappa = b/a$ is the ratio of the semi-major (b) and semi-minor (a) elliptical axes, respectively. (Here, we set the lengthscale as $2a = 1$.) Specifically, given the number of cells in the y direction, n_y, the number of cells in the x direction is chosen as $n_x = \text{round}(\sqrt{3}\kappa)$. For ellipses with $\kappa = \{3, 4, 6, 9\}$, we chose $n_y = \{17, 15, 12, 10\}$ which yielded the total number of particles $N = \{2992, 3120, 3000, 3120\}$, respectively. For each step, the move type (rotation or translation) was selected at random with equal probability. The maximum degree of translation and rotation per move were independently scaled to yield an ~25%

acceptance rate for efficient phase sampling. Density ranges were chosen to span the isotropic, nematic, and solid phases. For ellipses with $\kappa = \{3, 4, 6, 9\}$, we chose $\eta = \{0.6-0.9, 0.55-0.9, 0.4-0.9, 0.3-0.9\}$, respectively. A typical run proceeded as follows. A system of N hard ellipses was started from an ideal triangular lattice at maximum packing fraction and expanded to a target η value. Next, the range of translational and rotation move sizes was optimized using 50 iterations of 100 steps to achieve the targeted acceptance ratio, where a step is equal to an HOOMD-blue "time step," or approximately four sweeps over all particles. Then, the system was equilibrated for 6×10^6 steps, and data were collected every 6000 steps from an additional 6×10^6 step production run. From each frame, 30 feature vectors were constructed, where the probe particles were selected without replacement within a given frame.

3. Widom-Rowlinson model

The Widom-Rowlinson (WR) model[36] is composed of a binary mixture of A and B particles where like pairs (A-A or B-B) are non-interacting and unlike pairs (A-B) interact isotropically via a hard-core repulsion with diameter σ. Upon densification, the WR model compositionally demixes to form separate A- and B-rich phases. The resulting phase transition can straightforwardly be used to model compositional demixing; however, by integrating out the coordinates of one of the species, a model for liquid-gas coexistence can be obtained. In this work, we study the symmetric WR model where the number of A and B particles is equal.

The full specification of an individual particle in the WR model requires knowledge of both its type (A or B) and its position, yielding two obvious quantities to include in the feature construction. Instead of directly encoding the particle type as a categorical variable, we use particle type information to modify the assignments of the α and β indices. We construct NN positional features as prescribed by [Eqs. (2)–(4)], but we only include distances between pairs of A particles in the feature vector. We use a single probe particle ($n_P = 1$) with $n_{NN} = 1200$ nearest neighbors for the environmental descriptors. By neglecting one of the WR components, we construct features that explicitly leverage both compositional and positional information. Finally, we normalize the distances in the same fashion as the ellipse positional features described in Sec. II B 2.

For the production of the configuration data required to construct feature vectors for PCA, the HOOMD-Blue hard-particle Monte Carlo integrator[44–46] was used to perform the simulations of the WR model in a square box for $N = 4096$ particles in 2D. Equilibrium samples were generated at number densities spanning both the mixed and ordered phases as follows. After compressing the final configuration from simulation at the previous density, the system was equilibrated for 10^7 steps, followed by a production run of 10^7 steps, from which data were collected every 10^3 steps, for a total of 10^4 snapshots per density. A step is equivalent to a HOOMD-blue "time step" as defined in Sec. II B 2. Simulations were run from $\rho = 0.064$ to 3.82 in increments of 0.064. From each frame, a single feature vector was constructed.

III. RESULTS AND DISCUSSION

Prior to examining the PCA results for the above models, we explain the general interpretation of the quantities that result from PCA below. For the features constructed according to the protocols described in Sec. II, PCA discovers a set of orthogonal axes—the principle components (PCs)—that are constructed in succession so as to maximize the data variance projected along each new axis. In this work, we monitor the relative explained variance of the PCs, denoted as λ_i for the ith PC; by convention, the PCs are sorted so that $\lambda_i \geq \lambda_{i+1}$. A comparatively large value for λ_1 indicates that the information content of the features has been effectively concentrated into a single dimension: the first PC.

Of particular relevance to interpreting the PCA results is the projection of the feature vectors along the PCs: the PC score, denoted p_i for the ith PC. Given that the first PC contains the largest explained variance, we evaluate the use of p_1 as an OP-like quantity to report on the phase transition of interest.[47] This strategy amounts to coalescing as much "information" (i.e., variance) as possible from the high-dimensional feature vector $\boldsymbol{f}_C$ into the scalar p_1.

Since each p_1 is associated with a single feature vector, we define two quantities that are averaged over a given state-point $\mathcal{S}$ (here, the state points are densities): $P_1 \equiv \langle p_1 \rangle_{\mathcal{S}}$ and the associated standard deviation, $\sigma_1 \equiv \sqrt{\langle p_1^2 \rangle_{\mathcal{S}} - \langle p_1 \rangle_{\mathcal{S}}^2}$. When P_1 and σ_1 are plotted as a function of density, phase transitions will generally be indicated by a sigmoid in the former and a peak in the latter metric. The maximum of σ_1 is useful as an operational definition of the phase transition point when P_1 is slowly varying—as can be the case for very continuous phase transitions. This is not to say that P_1 is useless in such cases though. In fact, if the dimensionality reduction is successful—as gauged by the presence of dominant principal components—the behavior of P_1 is always revealing of the transition. σ_1 should be viewed as akin to a susceptibility of the order parameter, which can be useful as a supplemental metric.

Throughout, P_1 is compared to an appropriate classical OP for the transition under consideration. The utility of the PCA derived OP is gauged by the quantitative closeness of the estimated transition points and not by the overall similarity between the two OPs. For hard disks, in Paper I,[52] as well as for the random organization model in this paper, PCA can yield, at least very closely, the functional form of the classical OP. However, such similarity is not required for the PCA to inform the location of the phase transition. Multiple distinct metrics may be capable of quantifying any given phase transition. Our goal in this work is to explore the capability of PCA to develop a useful order parameter for the transition, not necessarily to replicate a classical analog.

The final relevant quantities from the PCA calculation are the PCs themselves—the weights that relate the features and the PC scores. As described in Sec. II A, the intuited features ($\boldsymbol{f}_I$) are transformed in a corrected representation ($\boldsymbol{f}_C$), the latter of which are input into the PCA calculation. Since the values comprising $\boldsymbol{f}_I$ are straightforward to interpret physically, we explore the relationship between the PC scores and $\boldsymbol{f}_I$ (instead of $\boldsymbol{f}_C$). As described in Paper I,[52] it is possible to

TABLE I. Common PCA variable definitions.

λ_i	Relative (fractional) explained variance captured by the ith PC, ranging between 0 and 1, where larger values are indicative of greater importance.
p_i	The ith PC score. Mathematically, this is the projection of a feature vector along the ith PC. PCs offer a new coordinate system with information concentrated along the earlier (smaller index) PCs.
P_i	Average of the ith PC score, p_i, over data from a state point ($\mathcal{S}$). P_1 serves as the OP-like quantity to report on phase transitions.
σ_i	Standard deviation of the ith PC score, p_i, at a state point ($\mathcal{S}$). This is used as an effective "susceptibility" to locate the phase transition by identifying the maximum value.
$\boldsymbol{q}_i$	Vector of weights that quantify the relevance of each feature to the ith PC.

write down a linear relationship between the scalar p_i and the vector $\boldsymbol{f}_I$ via the following dot product:

$$p_i = \boldsymbol{q}_i^T \boldsymbol{f}_I, \tag{6}$$

where $\boldsymbol{q}_i$ is the desired vector of weights that map $\boldsymbol{f}_I$ to p_i. Examination of these weights reveals which physically meaningful quantities are particularly relevant to the phase transition.

In summary, we consider (1) the effectiveness of the dimensionality reduction via λ_1, (2) the low-dimensional (OP-like) representation of the data via quantities that depend on p_1 (P_1 and σ_1), and (3) the relative importance of the physical quantities that comprise $\boldsymbol{f}_I$ via the weights $\boldsymbol{q}_i$. For convenience, we summarize the above notation in Table I.

A. Random organization model

The RandOrg model was developed to understand the transition from reversible to irreversible dynamics that occurs upon increasing either the applied periodic shear or the density of a material.[29] In the first incarnation of the RandOrg model, an initial configuration was sheared and any particles overlapping with others as a result of deforming the simulation box were defined as active particles. Only the active particles were given a random displacement after which the simulation box was restored to its original geometry. At sufficiently low combinations of density and applied shear, a quiescent "absorbing" state eventually results, where shear does not generate further particle overlaps and there are no longer active particles. However, at greater densities and/or shear rates, shearing the system will always generate some overlaps. The reversible-to-irreversible transition reflects a state where the onset of particle collisions upon shearing prevents the system from returning to its original state when the shear is removed.[28,29]

A modified version of the RandOrg model, where shear is not included, has also been studied.[30,31] Instead, as described in Sec. II B 1, initial particles are placed at random; active particles correspond to overlapping particles. This model possesses the same type of transition from an absorbing state at sufficiently low densities to an evolving steady-state containing a non-zero number of active particles at higher densities, while being technically simpler to implement. Figure 1(a) shows the

fraction of active particles f_A in this version of the RandOrg model as a function of number density ρ. Two simulation configurations below ($\rho = 0.5$) and above ($\rho = 0.51$) the critical point are shown in Figs. 1(b) and 1(c), respectively.

Like hard disks, the RandOrg model comprises identical, radially symmetric particles, for which distance-based features are a sensible choice. Therefore, we first employed the feature vector developed in Paper I[52] for hard disks—sorted nearest-neighbor (NN) distances associated with a single probe particle [Eqs. (2)–(4) where $n_P = 1$]. However, this construction of the feature vector did not produce a satisfactory OP for either P_1 or σ_1. Use of a feature vector for which $n_P = 1$ likely fails because, above the critical point, the system always has two effective particle types—active and inactive [see Fig. 1(c)]. Therefore, the environment of a single particle is not an accurate representation of the simulation box as a whole at higher densities.

As described in Sec. II A, distance-based feature vectors can naturally incorporate multiple probe particles (and
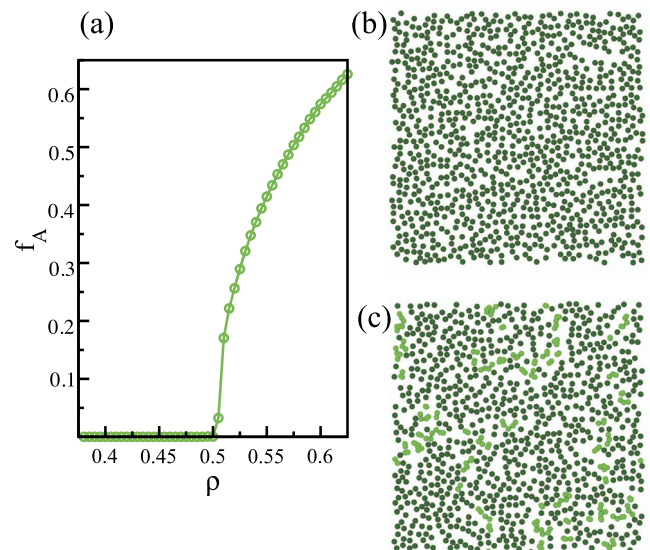

FIG. 1. (a) Fraction of active particles, f_A, in the RandOrg model as a function of number density, ρ. [(b) and (c)] Simulation snapshots (b) below ($\rho = 0.5$) and (c) above ($\rho = 0.51$) the critical density. Active particles are shown in lighter green in panel c.

their corresponding neighbors). As with the NN distances for a given probe particle, one must decide how to order the probe particles inside the feature vector. In the absence of any information about the nature of a given phase transition, a first choice might be to randomly order the probe particles. In Fig. 2, we show results for the first PC using a feature vector constructed from 40 probe particles ($n_P = 40$), each of which is encoded via its first 10 NN distances ($n_{NN} = 10$). In principle, the weights associated with each probe particle should be identical since there is no physics-based interpretation for their ordering in the feature vector. Indeed in Fig. 2(c), we find a repeating pattern for every 10 weights: the first NN distance ($r_1^{(\alpha)}$) component weight for each probe is large in magnitude. The relative uniformity of the first NN distance weights, compared to the noisiness in the larger NN distances, indicates that the $r_1^{(\alpha)}$ values are informative to the PCA.

The corresponding OP is shown in Fig. 2(a) and bears striking resemblance to the standard OP shown in Fig. 1(a). Indeed, by arbitrarily shifting and scaling the PC score, we find that f_A essentially overlaps with the PCA-deduced OP. It seems that the repeating unit in the component weights is able to distinguish between overlapping and non-overlapping particles and therefore can report on the relative amounts of active and inactive particles at a given value of ρ.

We can use the above component weights to intelligently devise a better sorting scheme for the probe particles. From Fig. 2(c), it is clear that $r_1^{(\alpha)}$ is a highly weighted contribution to the feature vector; therefore, we performed a separate PCA calculation with the same values for n_P and n_{NN} while
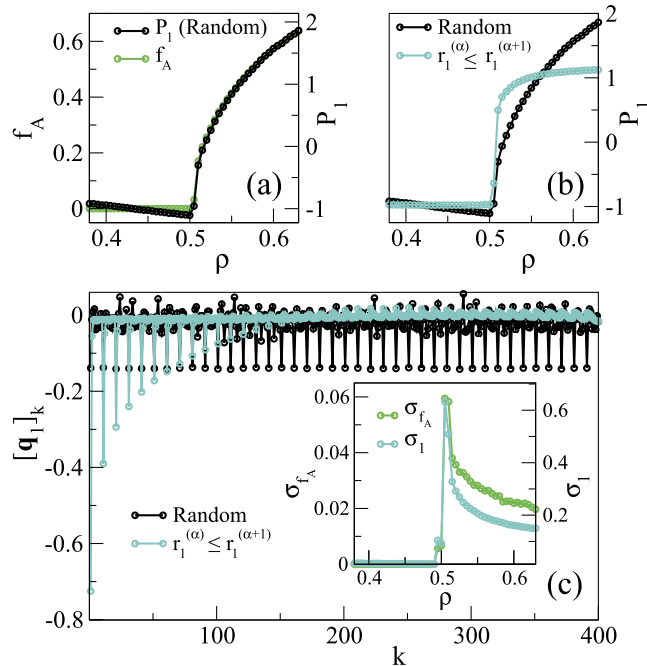
sorting the probe particles so that $r_1^{(\alpha)} \leq r_1^{(\alpha+1)}$. Because we have sorted the probe particles on the basis of a physically meaningful descriptor, the symmetry among probe particles is broken and the probe particles with the closest NNs (i.e., those assigned to lower α) are weighted more heavily than other probe particles [Fig. 2(c)]. Moreover, the associated OP is significantly sharper at the phase transition, essentially giving a binary classification into absorbing states and dynamic steady states on the basis of the OP [Fig. 2(b)].

The approximate location of the phase transition can be determined by the lowest density at which f_A is non-zero, which occurs at $\rho = 0.505$ for the simulation protocol described in Sec. II B 1. This value coincides with the sudden increase in P_1 from a flat line for the sorted features shown in Fig. 2(b). The standard deviation in both f_A (σ_{f_A}) and the p_1 associated with the NN sorted features (σ_1), shown in the inset of Fig. 2(c), both possess an obvious spike at $\rho = 0.505$, indicating that the maximum value in σ_1 is also an accurate reporter on the location of the phase change.

With the above sorting scheme in hand, we vary both n_P and n_{NN} while keeping the length of the feature vector fixed at $m = n_P \times n_{NN} = 400$. As n_P increases and therefore n_{NN} decreases, the PCA-deduced OP sharpens into a sigmoidal curve that separates quiescent absorbing states from diffusive steady-states; see Fig. 3(a). Conversely when $n_P = 1$ (as was the case for Paper I[52]), P_1 cannot detect the transition. Correspondingly, when features constructed with more probe particles are used, the explained variance associated with the first PC increases dramatically [Fig. 3(b)]. The preceding trend is monotonic—there is no value to include more than the nearest interparticle distance per probe particle at constant m, an indication of the local character of the phase transition in the RandOrg model.

For the above series of PCA calculations, the first 80 component weights are plotted in Fig. 3(c). When n_P is small, the
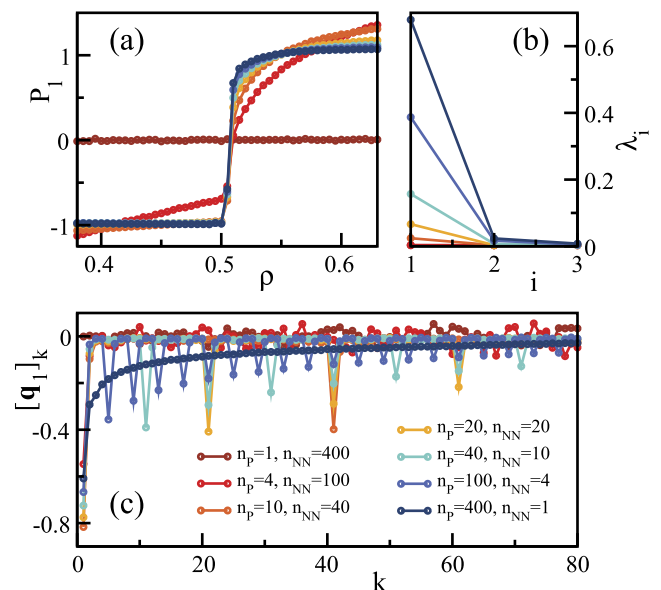

FIG. 2. (a) The PCA-deduced OP P_1 (with probe particles sorted randomly in the feature vector) as a function of number density ρ compared to the conventional OP (the fraction of active particles, f_A) for the RandOrg model. (b) Comparison of P_1 with probe particles sorted randomly (black) versus according to their first NN distance so that $r_i^{(\alpha)} \leq r_1^{(\alpha+1)}$ (blue). (c) Component weights, $[\boldsymbol{q}_1]_k$, as a function of feature dimension k for the two PC scores shown in panel (b). (Inset) Standard deviations of f_A ($\sigma(f_A)$) and of p_1 or the PC for an individual feature sorted by the first NN distance, (σ_1) as a function of ρ.

FIG. 3. (a) PCA-deduced OP P_1 of the RandOrg model as a function of number density ρ for different numbers of probe particles n_P and corresponding nearest neighbors n_{NN}, respectively. (b) Corresponding explained variance λ_i for the first three PCs and (c) the first 80 PC weights $[\boldsymbol{q}_1]_k$.

components appear to be largely random, but as n_P is increased, the components develop more structuring. For each probe particle, the component weights associated with $r_1^{(\alpha)}$ have a much larger weight than that of the rest of the features, reinforcing the importance of the first NN distance in the dimensionality reduction. Moreover, the probe particles that have closer first NNs have greater weights; we can interpret the role of using multiple probe particles as capturing an accurate measure of $r_1^{(1)}$ in a statistical sense, i.e., not all probe particles are required, but sampling is needed to make sure that sufficiently representative interparticle separations are included in each feature vector.

While the importance of the first NN distance is intuitive given that the RandOrg phase transition is defined by the presence or absence of particle overlaps, we did not incorporate knowledge of the transition in constructing the features. In other words, our results suggest that modifying the feature vector can be used to infer characteristics of a transition, even if its nature is unknown at the outset. Specifically, for the RandOrg model, the importance of the first NN distance revealed by the PCA implies a transition that is local in character in real-space, and the necessity of multiple probe particles indicates that multiple distinct particle types or environments are an important characteristic.

B. Hard ellipses

The freezing transition for hard ellipses differs from that of hard disks because the former features an intervening nematic phase between the disordered fluid and the positionally ordered solid. The nematic phase manifests when the ellipses display disordered center-of-mass positions but quasi-long range orientational order.[32–35] A conventional OP that reports on the fluid-nematic transition, P_2^{max}, as well as simulation configurations at densities below and above the phase transition, is shown in Fig. 4 for ellipses with an aspect ratio $\kappa = 4$. The continuous, second-order nature of the fluid-nematic transition is apparent from the behavior of P_2^{max}, from which the precise density for the underlying phase transition is not readily apparent. Therefore, one typically monitors the long-range power-law decay of a pairwise angular correlation function versus interparticle separation to identify the transition. The nematic phase transition point is identified when the power law decay exponent (ϕ) falls below an approximate value of 1/4.[33,34] As shown in the inset of Fig. 4(a), $\phi = 1/4$ at $\eta \approx 0.696$, in close agreement with prior identification of the transition.[33,34]

To detect the fluid-nematic phase transition via PCA, we use a feature vector constructed from the relative orientation of pairs of ellipses that are sorted in ascending order by the distance between the probe particle and its neighbor, as described in Sec. II B 2. In Fig. 5, we present the results of PCA using this orientational feature vector for ellipses with an aspect ratio $\kappa = 4$. As seen by the explained variance λ_i in Fig. 5(b), the first PC captures approximately 40% of the data variance, indicating effective dimensionality reduction. From the component weights $[\mathbf{q}_1]_k$ in Fig. 5(a), it is clear that long-range orientations (larger values of k) are much more important than the closer neighbor orientations which tend toward zero.
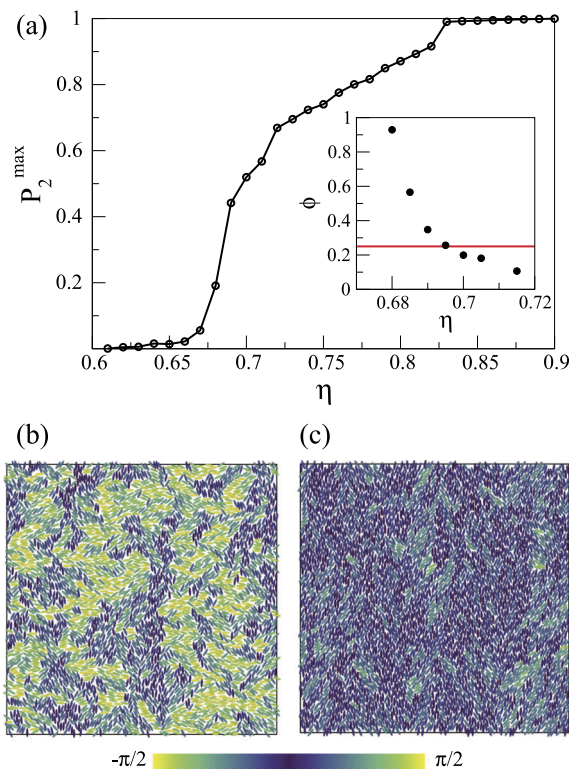

FIG. 4. Density-driven isotropic fluid to nematic phase transition in a system of hard ellipses with aspect ratio $\kappa = 4$. (a) The packing fraction η dependence of the conventional order parameter for this transition, $P_2^{max} = [\langle 1/N\Sigma_i^N \cos(2\theta_i)\rangle^2 + \langle 1/N\Sigma_i^N \sin(2\theta_i)\rangle^2]^{1/2}$, where θ_i is the angle between the semi-minor axis of the ith ellipse and the x-axis and N is the number of ellipses, as per Ref. 35. (Inset) Power law exponent (ϕ) for the angular correlations as a function of η around the fluid-nematic transition. The horizontal line indicates that the phase transition occurs when $\phi = 1/4$. Simulation configurations of (b) the isotropic fluid at $\eta = 0.65$ and (c) the nematic phase at $\eta = 0.75$. Ellipses are color coded according to θ_i as defined above with the angular range limited to $[-\pi/2, \pi/2]$ due to orientation symmetry of the ellipse.

The above weights reflect the underlying structural motifs present in hard ellipses at various values of η. Orientationally aligned clusters of ellipses are present in both fluid and nematic phases [compare, for instance, the snapshots in Figs. 4(b) and 4(c)]. Therefore, orientations between nearby ellipses [smaller values of k in Fig. 5(a)] are not useful indicators of nascent orientational long-range order, and their contributions to the OP are suppressed by the PCA. On the other hand, long distance components are approximately equal-weighed as they correlate proportionally to the presence of an emerging nematic director but average out for random configurations in the fluid state.

Regarding the OP itself, we show P_1 and its standard deviation σ_1 in Fig. 5(c); the latter quantity can be interpreted as a type of susceptibility of the OP. Note that P_1 resembles the traditional OP P_2^{max} of Fig. 4 and, while responsive to the nematic phase change, does not provide a unique transition point. As such, we use σ_1 to correlate the PCA results to the fluid-nematic phase transition: the maximum of σ_1 indicates the region most consistent with large-scale configuration fluctuations near the critical point of a continuous phase transition. The maximum occurs at $\eta \approx 0.695$, in close agreement with $\eta \approx 0.696$ identified by the point at which the power law
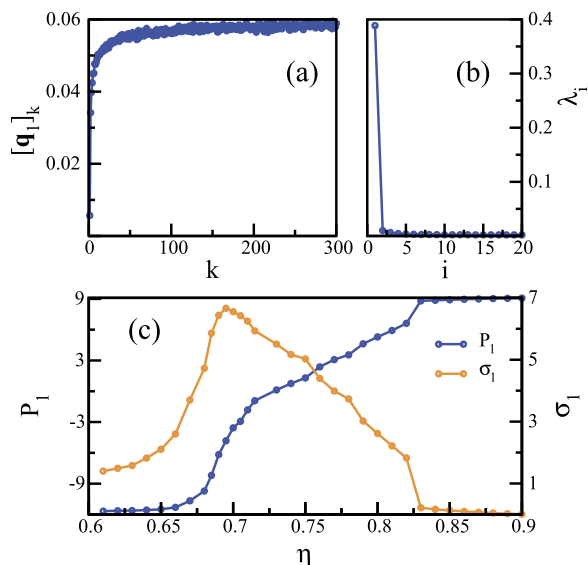
FIG. 5. Based on PCA of the 2D system of hard ellipses, we show (a) component weights $[\mathbf{q}_1]_k$, (b) the explained variance λ_i, and (c) the OP (P_1) and standard deviation (σ_1).

exponent ϕ decays to 1/4 in Fig. 4(a). Indeed, for all values of κ investigated here, we find that the density η associated with the maximum value of σ_1 is in excellent agreement with fitted fluid-nematic boundary reported by Xu *et al.*[34] (see Fig. 6), without requiring a tedious analysis of the long range scaling behavior in the angular correlations employed by the latter study.

While the use of orientational features as input to PCA provides a means to detect the fluid-nematic phase transition, the relationship between the above PCA results and the nematic-solid transition is less obvious. In Figs. 7(a) and 7(b), we plot a normalized version of P_1 and its standard deviation σ_1 derived from the orientational features against the known nematic-solid coexistence region (the gray shaded area). There is a weak response to the nematic-solid region as $\tilde{\sigma}_1$ drops abruptly—perhaps an indicator of reduced orientational freedom of the ellipses upon solidification. The weak response is not surprising as angular degrees of freedom are an indirect measure of the nematic-solid transition. Thus, in this case, angular features do not capture the dominant physics
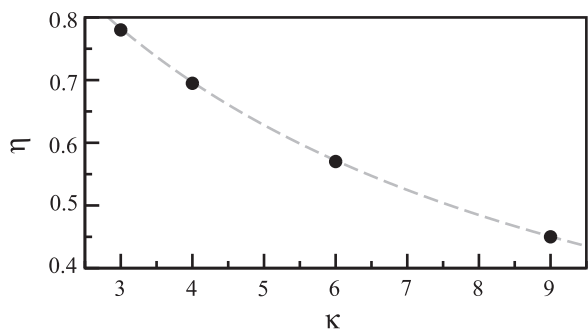

FIG. 6. Phase boundary of the isotropic fluid to nematic transition for a system of hard ellipses as a function of packing fraction η and aspect ratio κ. Solid black dots indicate the phase transition point identified from the position of maximum susceptibility, max(σ_1). The dashed gray line indicates the phase boundary fit reported by Xu *et al.*[34]
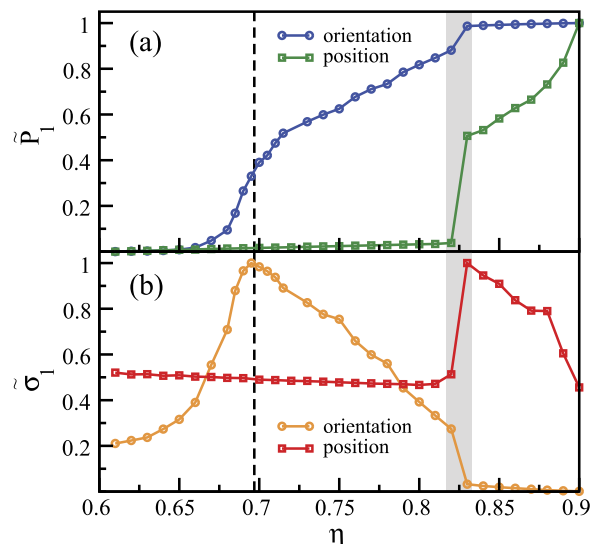

FIG. 7. For the first PC of hard ellipses, comparison of the (a) shifted and normalized PC scores $\tilde{P}_1 \equiv \frac{P_1 - \min(P_1)}{\max(P_1 - \min(P_1))}$ and (b) normalized standard deviations $\tilde{\sigma}_1 \equiv \frac{\sigma_1}{\max(\sigma_1)}$, as derived from either orientational or positional feature vectors. The dashed black vertical line indicates fluid-nematic boundary, and the shaded gray region indicates the nematic-solid phase coexistence region reported in Ref. 35.

governing the freezing transition and will require more suitable features to more clearly delineate the phase transition.

In order to detect the center-of-mass level ordering that occurs at the nematic-solid transition, we employ the positional NN features used for hard disks in Paper I.[52] That is, instead of including the relative angle between two ellipses in the feature vector, we employ the interparticle distances. In Figs. 7(a) and 7(b), we compare the normalized PC scores $\tilde{P}_1$ and the associated $\tilde{\sigma}_1$, respectively, when these positional features are used as input to the PCA. The resulting OP is insensitive to the fluid-nematic boundary but grows sharply across the known nematic-solid phase-coexistence region. Similar to the orientational features, the maximum in the position-based susceptibility σ_1 in Fig. 7(b) is an appropriate identifier for the underlying phase transition. The form of the OP as a function of η—flat over the fluid phase, rapid growth upon solidification, and more muted growth in the solid phase—is qualitatively similar to the OP reported in Paper I[52] for the densification of hard disks. Together, the above results attest to the ability of PCA to provide insights into the character of a given phase transition by varying the form of the feature vector.

C. Widom-Rowlinson model

As mentioned in Sec. II B 3, the WR model contains two particles types—A and B—where like particles are non-interacting and unlike particles interact via a hard-core repulsion of diameter σ. At low densities, the two species are mixed. However, upon densification, a phase transition occurs[37,38,48] where the WR mixture phase separates into A-rich and B-rich regions [Figs. 8(a) and 8(b)] as the excluded volume effects experienced by the unlike particles overcome the mixing entropy. The density at which the demixing transition occurs varies with composition; we denote x as the fraction
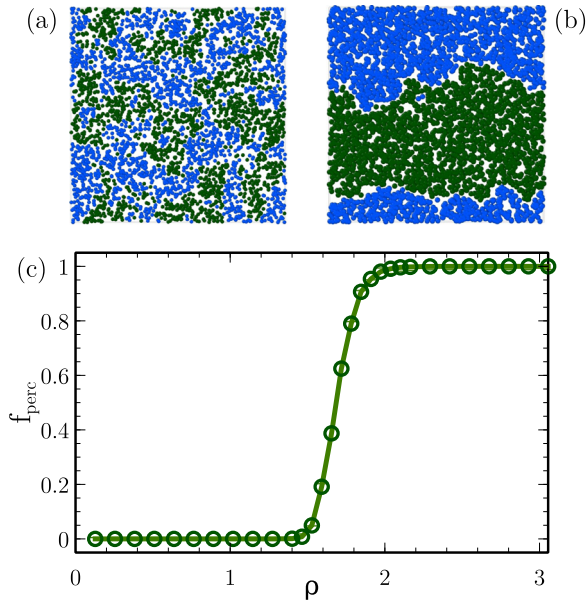
FIG. 8. For the WR model at $x = 0.5$, simulated configuration snapshots of (a) mixed WR particles at $\rho = 1.25$ (below the phase transition) and (b) demixed WR particles at $\rho = 2.5$ (above the phase transition). (c) Fraction of percolated configurations (f_{perc}) as a function of number density.

FIG. 9. For the first PC upon application of PCA to the WR model, (a) its component weights $[q_1]_k$, (b) the explained variance λ_i for the first 20 components, and (c) the PCA-deduced OP (P_1) and percolation-based OP (f_{perc}). (Inset) The standard deviation σ_1 of the PCA OP.

of A particles. In the present work, we study the mixture for which $x = 0.5$.

When $x = 0.5$, the density at which clusters of like particles become percolated can be used to determine the demixing transition.[38,49,50] For the WR model, a cluster is defined as a group of particles that are all either directly overlapping or connected via a contiguous pathway of overlapping particles when periodic boundary conditions are properly taken into account. For a finite-sized, periodically replicated simulation box, a percolated cluster is one that grows in size upon replication of the simulation cell. Therefore, for each species at $x = 0.5$, we computed the fraction of configurations possessing at least one percolated cluster of that particle type and averaged the results for the A and B particles to yield f_{perc}. Figure 8(c) shows f_{perc} as a function of density; percolated clusters were identified as described in Ref. 51. One choice for the percolation threshold—the point when at least 50% of the configurations are percolated—yields a de-mixing transition density of $\rho_t = 1.68$.

Positional features of the form defined by Eqs. (2)–(4) are unable to detect the above de-mixing transition if compositional degrees of freedom are not taken into account—a consequence of the absence of large scale fluctuations in the packings of the particles (agnostic to particle type) as the phase transition occurs. However, for any multicomponent mixture, features can be constructed using particle type data as well as spatial information. For the WR model, one such strategy is to design a feature vector that only includes interparticle distances if the corresponding pair of particles meets some criterion based on the particle type. One such choice (though others are possible) is to only include distances between two A particles in the feature vector defined by Eqs. (2)–(4)—akin to the liquid-gas formulation of the WR model. The outcome of PCA with the above feature vector is shown in Fig. 9. The component weights $[q_1]_k$ in Fig. 9(a) indicate that the
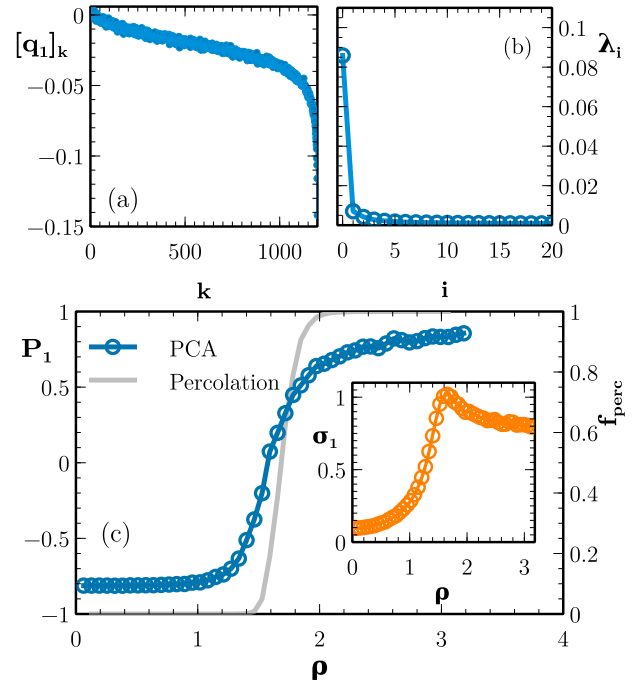
long-ranged positional correlations dominate, whereas the smallest interparticle separations with respect to a given probe are essentially meaningless to the PCA. Reminiscent of the fluid-nematic transition seen in ellipses and described in Sec. III B, some local clustering on the basis of the particle type occurs at lower densities than phase separation does; see Fig. 8(a), for example. Therefore, it is the long-range correlations that change sharply as the phase transition occurs. Figure 9(b) depicts the explained variance λ_i for the first 20 PCs, where the first PC accounts for ~10% of the data variance—an order of magnitude more than the succeeding components.

The PCA-deduced OP is shown in Fig. 9(c). Relative to f_{perc}, the PCA-based OP varies more slowly and has a wider transition window. To identify the unique transition point, the standard deviation σ_1, as outlined in Sec. III B, is computed for every density and shown in the inset of Fig. 9(c). The maximum value of σ_1, denoting the region with most variance in the feature vectors, is accepted as being associated with the transition density ρ_t. For the above one-component mixture, a peak in the standard deviation is observed at $\rho_t = 1.66$, which is in excellent agreement with the value obtained through percolation arguments ($\rho_t = 1.68$)—indicating successful identification of the de-mixing transition.

IV. CONCLUSIONS

In this article, we extended the PCA framework introduced in Paper I[52] to detect phase transitions in three new model systems, each characterized by very different physics. The success of the method in these cases highlights the importance of exploring various feature representations

when seeking to detect such phase transitions and understand their underlying physics.

Moving forward, two avenues seem fruitful for developing a routine analysis toolkit: (1) curate a sufficiently diverse library of features, each focused on different physical aspects that may be relevant to various phase transitions of interest, and (2) explore the ability of more sophisticated, nonlinear learners to autonomously extract the physical intuition underlying such transitions on-the-fly.

Finally, we comment on general trends that we observed to indicate that the PCA calculation was usefully reporting on the phase transition of interest. First we note that meaningful dimensionality reduction into the first PC is generally indicated by a large relative explained variance in comparison to the higher order PCs. Furthermore, we found that appropriate choices for the features resulted in an OP with strong convergence properties that required relatively small amounts of data to overcome sampling noise. We expect that these trends are relevant to other machine-learning approaches for the detection of phase transitions as well.

ACKNOWLEDGMENTS

[1] D. Barber, *Bayesian Reasoning and Machine Learning* (Cambridge University Press, 2012).

[2] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, Springer Series in Statistics (Springer-Verlag, New York, 2001), Vol. 1.

[3] J. Shlens, "A tutorial on principal component analysis," e-print arXiv:1404.1100 (2014).

[4] R. Ocampo, M. A. de Luna, R. Vega, G. Sanchez-Ante, L. E. Falcon-Morales, and H. Sossa, "Pattern analysis in DNA microarray data through PCA-based gene selection," in *Iberoamerican Congress on Pattern Recognition* (Springer, 2014), pp. 532–539.

[5] B. Hess, "Similarities between principal components of protein dynamics and random diffusion," Phys. Rev. E **62**, 8438–8448 (2000).

[6] M. Ernst, F. Sittel, and G. Stock, "Contact- and distance-based principal component analysis of protein dynamics," J. Chem. Phys. **143**, 244114 (2015).

[7] J. Carrasquilla and R. G. Melko, "Machine learning phases of matter," Nat. Phys. **13**, 431 (2017).

[8] E. P. van Nieuwenburg, Y.-H. Liu, and S. D. Huber, "Learning phase transitions by confusion," Nat. Phys. **13**, 435 (2017).

[9] W. Hu, R. R. P. Singh, and R. T. Scalettar, "Discovering phases, phase transitions, and crossovers through unsupervised machine learning: A critical examination," Phys. Rev. E **95**, 062122 (2017).

[10] L. Wang, "Discovering phase transitions with unsupervised learning," Phys. Rev. B **94**, 195105 (2016).

[11] S. J. Wetzel, "Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders," Phys. Rev. E **96**, 022140 (2017).

[12] P. Ponte and R. G. Melko, "Kernel methods for interpretable machine learning of order parameters," Phys. Rev. B **96**, 205146 (2017).

[13] C. Wang and H. Zhai, "Machine learning of frustrated classical spin models. I. Principal component analysis," Phys. Rev. B **96**, 144432 (2017).

[14] C. Wang and H. Zhai, "Machine learning of frustrated classical spin models (II): Kernel principal component analysis," Front. Phys. **13**, 130507 (2018).

[15] O. Lars, "The effects of shape on the interaction of colloidal particles," Ann. N. Y. Acad. Sci. **51**, 627–659 (1949).

[16] S. Ermakov, A. Beletskii, O. Eismont, and V. Nikolaev, "Brief review of liquid crystals," in *Liquid Crystals in Biotribology* (Springer, 2016), pp. 37–56.

[17] E. H. Chimowitz, *Introduction to Critical Phenomena in Fluids* (Oxford University Press, 2005).

[18] H. Hinrichsen, "Non-equilibrium phase transitions," Physica A **369**, 1–28 (2006), Fundamental Problems in Statistical Physics.

[19] M. Henkel, H. Hinrichsen, and S. Lubeck, *Non-Equilibrium Phase Transitions: Volume 1: Absorbing Phase Transitions* (Springer Science and Business Media BV, The Netherlands, 2008).

[20] M. Henkel and M. Pleimling, *Non-Equilibrium Phase Transitions: Volume 2: Ageing and Dynamical Scaling Far from Equilibrium* (Springer Science and Business Media BV, The Netherlands, 2011).

[21] D. Marenduzzo, "An introduction to the statistical physics of active matter: Motility-induced phase separation and the 'generic instability' of active gels," Eur. Phys. J.: Spec. Top. **225**, 2065–2077 (2016).

[22] N. Heptner, F. Chu, Y. Lu, P. Lindner, M. Ballauff, and J. Dzubiella, "Nonequilibrium structure of colloidal dumbbells under oscillatory shear," Phys. Rev. E **92**, 052311 (2015).

[23] T. H. Besseling, M. Hermes, A. Fortini, M. Dijkstra, A. Imhof, and A. van Blaaderen, "Oscillatory shear-induced 3D crystalline order in colloidal hard-sphere fluids," Soft Matter **8**, 6931–6939 (2012).

[24] C. Bechinger, R. Di Leonardo, H. Löwen, C. Reichhardt, G. Volpe, and G. Volpe, "Active particles in complex and crowded environments," Rev. Mod. Phys. **88**, 045006 (2016).

[25] J. E. Martin and A. Snezhko, "Driving self-assembly and emergent dynamics in colloidal suspensions by time-dependent magnetic fields," Rep. Prog. Phys. **76**, 126601 (2013).

[26] Y. Yang, D. McDermott, C. J. O. Reichhardt, and C. Reichhardt, "Dynamic phases, clustering, and chain formation for driven disk systems in the presence of quenched disorder," Phys. Rev. E **95**, 042902 (2017).

[27] C. Sándor, A. Libál, C. Reichhardt, and C. J. Olson Reichhardt, "Dynamic phases of active matter systems with quenched disorder," Phys. Rev. E **95**, 032606 (2017).

[28] D. Hexner, P. M. Chaikin, and D. Levine, "Enhanced hyperuniformity from random reorganization," Proc. Natl. Acad. Sci. U. S. A. **114**, 4294–4299 (2017).

[29] L. Corté, P. M. Chaikin, J. P. Gollub, and D. J. Pine, "Random organization in periodically driven systems," Nat. Phys. **4**, 420–424 (2008).

[30] E. Tjhung and L. Berthier, "Criticality and correlated dynamics at the irreversibility transition in periodically driven colloidal suspensions," J. Stat. Mech.: Theory Exp. **2016**, 033501.

[31] L. Milz and M. Schmiedeberg, "Connecting the random organization transition and jamming within a unifying model system," Phys. Rev. E **88**, 062308 (2013).

[32] J. Vieillard-Baron, "Phase transitions of the classical hard-ellipse system," J. Chem. Phys. **56**, 4729–4744 (1972).

[33] J. A. Cuesta and D. Frenkel, "Monte Carlo simulation of two-dimensional hard ellipses," Phys. Rev. A **42**, 2126–2136 (1990).

[34] W.-S. Xu, Y.-W. Li, Z.-Y. Sun, and L.-J. An, "Hard ellipses: Equation of state, structure, and self-diffusion," J. Chem. Phys. **139**, 024501 (2013).

[35] G. Bautista-Carbajal and G. Odriozola, "Phase diagram of two-dimensional hard ellipses," J. Chem. Phys. **140**, 204502 (2014).

[36] B. Widom and J. S. Rowlinson, "New model for the study of liquid–vapor phase transitions," J. Chem. Phys. **52**, 1670–1684 (1970).

[37] D. Ruelle, "Existence of a phase transition in a continuous classical system," Phys. Rev. Lett. **27**, 1040–1041 (1971).

[38] J. T. Chayes, L. Chayes, and R. Kotecký, "The analysis of the Widom-Rowlinson model by stochastic geometric methods," Commun. Math. Phys. **172**, 551–569 (1995).

[39] V. Botu and R. Ramprasad, "Learning scheme to predict atomic forces and accelerate materials simulations," Phys. Rev. B **92**, 094306 (2015).

[40] B. Venkatesh and R. Rampi, "Adaptive machine learning framework to accelerate *ab initio* molecular dynamics," Int. J. Quantum Chem. **115**, 1074–1083 (2014).

[41] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet—A deep learning architecture for molecules and materials," J. Chem. Phys. **148**, 241722 (2018).

[42] N. Lubbers, J. S. Smith, and K. Barros, "Hierarchical modeling of molecular energies using a deep neural network," J. Chem. Phys. **148**, 241715 (2018).

[43] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," Am. Stat. (published online).

[44] J. A. Anderson, C. D. Lorenz, and A. Travesset, "General purpose molecular dynamics simulations fully implemented on graphics processing units," J. Comput. Phys. **227**, 5342–5359 (2008).

[45] J. Glaser, T. D. Nguyen, J. A. Anderson, P. Lui, F. Spiga, J. A. Millan, D. C. Morse, and S. C. Glotzer, "Strong scaling of general-purpose molecular dynamics simulations on GPUs," Comput. Phys. Commun. **192**, 97–107 (2015).

[46] J. A. Anderson, M. E. Irrgang, and S. C. Glotzer, "Scalable metropolis Monte Carlo for simulation of hard shapes," Comput. Phys. Commun. **204**, 21–30 (2016).

[47] We do not exclude the possibility that the other PC scores may be useful or that a better order parameter could involve multiple PC scores.[9–14] For simplicity, we focus on p_1.

[48] H.-O. Georgii, J. Lorinczi, and J. Lukkarinen, "The continuum potts model at the disorder 'order transition' a study by cluster dynamics," J. Stat. Mech.: Theory Exp. **2005**, P06011.

[49] W. Klein, "Potts-model formulation of continuum percolation," Phys. Rev. B **26**, 2677–2678 (1982).

[50] R. Sun, H. Gould, J. Machta, and L. W. Chayes, "Cluster Monte Carlo study of multicomponent fluids of the Stillinger-Helfand and Widom-Rowlinson type," Phys. Rev. E **62**, 2226–2232 (2000).

[51] B. A. Lindquist, R. B. Jadrich, D. J. Milliron, and T. M. Truskett, "On the formation of equilibrium gels via a macroscopic bond limitation," J. Chem. Phys. **145**, 074906 (2016).

[52] R. B. Jadrich, B. A. Lindquist, and T. M. Truskett, "Unsupervised machine learning for detection of phase transitions in off-lattice systems I. Foundations." J. Chem. Phys. **149**, 194109 (2018).