Machine-learning effective many-body potentials for anisotropic particles using orientation-dependent symmetry functions

Gerardo Campos-Villalobos, Giuliana Giunta, Susana Marín-Aguilar, et al.

View Online   Export Citation   CrossMark

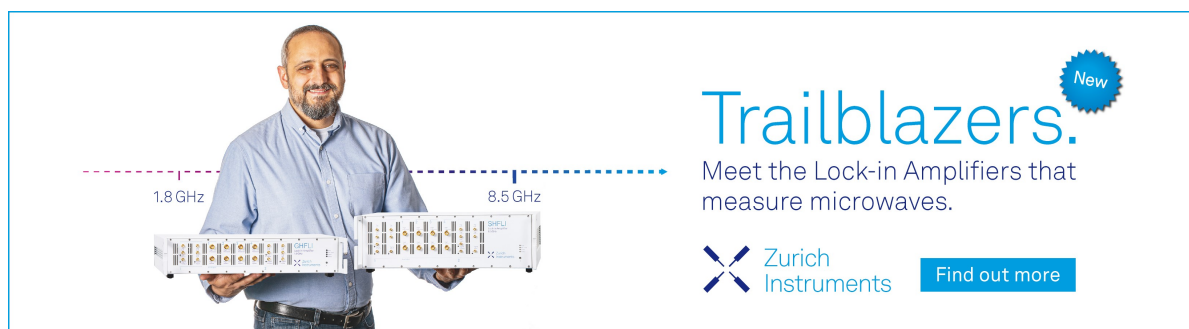## ARTICLES YOU MAY BE INTERESTED IN

Machine-learning effective many-body potentials for anisotropic particles using orientation-dependent symmetry functions

Gerardo Campos-Villalobos,[a] (iD) Giuliana Giunta, (iD) Susana Marín-Aguilar, (iD) and Marjolein Dijkstra[b] (iD)

AFFILIATIONS

Soft Condensed Matter, Debye Institute for Nanomaterials Science, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands

[a]Author to whom correspondence should be addressed: g.d.j.camposvillalobos@uu.nl
[b]Electronic mail: m.dijkstra@uu.nl

ABSTRACT

Spherically symmetric atom-centered descriptors of atomic environments have been widely used for constructing potential or free energy surfaces of atomistic and colloidal systems and to characterize local structures using machine learning techniques. However, when particle shapes are non-spherical, as in the case of rods and ellipsoids, standard spherically symmetric structure functions alone produce imprecise descriptions of local environments. In order to account for the effects of orientation, we introduce two- and three-body orientation-dependent particle-centered descriptors for systems composed of rod-like particles. To demonstrate the suitability of the proposed functions, we use an efficient feature selection scheme and simple linear regression to construct coarse-grained many-body interaction potentials for computationally efficient simulations of model systems consisting of colloidal particles with an anisotropic shape: mixtures of colloidal rods and non-adsorbing polymer coils, hard rods enclosed by an elastic microgel shell, and ligand-stabilized nanorods. We validate the machine-learning (ML) effective many-body potentials based on orientation-dependent symmetry functions by using them in direct coexistence simulations to map out the phase behavior of colloidal rods and non-adsorbing polymer coils. We find good agreement with the results obtained from simulations of the true binary mixture, demonstrating that the effective interactions are well described by the orientation-dependent ML potentials.

Published under an exclusive license by AIP Publishing. https://doi.org/10.1063/5.0091319

I. INTRODUCTION

Anisotropic molecules and colloids are able to self-assemble into complex structures with competing orientational and translational order, ranging from liquid crystals[1,2] to empty liquids.[3] Among colloidal systems, non-spherical particles have long been known and modern synthetic routes allow for a great variety of particles with different shapes.[4] For example, in addition to rod-like particles of biological origin, such as viruses,[5–7] colloidal rods can also be synthesized from a wide range of materials, including boehmite,[8] β-ferric oxyhydroxide (FeOOH),[9] gold,[10] silica,[11] cellulose,[12] and titanium dioxide (TiO$_2$) nanocrystals.[13] The assembly of such elongated particles has become increasingly popular because it allows for the formation of ordered superstructures exhibiting collective physical properties that depend on the shape and size of the constituent particles.[14] By precisely controlling physicochemical parameters and boundary conditions of a self-assembly process, particle-based simulations can be especially effective in providing a clear insight into the link between the detailed interactions among particles/molecules and the resulting equilibrium properties.

Historically, there have been basically two approaches to model non-spherical particles: atomistic and coarse grained.[15] In the former approach, chemistry-level details are well represented at the expense of a larger computational cost. In the latter, many atoms are grouped together into single sites (beads and superatoms) that generally do not show a spherical symmetry. In such cases, generic single-site ellipsoidal pair potentials and simple hard-particle models have been widely used, thereby providing an understanding of the physics behind their mesoscopic and macroscopic behaviors.[2]

For an accurate representation of specific systems with complex interactions, developing predictive and computationally tractable coarse-grained (CG) models is necessary. However, this is not a trivial task. Difficulties arise, in part, because at such a resolution level, the formal integration of a set of degrees of freedom leads to effective interaction potentials that typically require a description beyond the pairwise approximation.[16,17] The many-body terms in these CG potentials are, in general, very difficult to take into account[17] as it has been recognized in developing reduced-order models for star polymers,[18,19] colloid–polymer mixtures,[20,21] and ligand-stabilized nanoparticles.[22,23] In addition, even though the evaluation of these many-body potentials is, in general, computationally cheaper than simulating the full or fine-grained system, the computational cost may still be high and limit the range of accessible time- and length-scales.

In the past years, data-driven or machine learning (ML) approaches have been successfully applied to build accurate and computationally efficient potentials for molecular and atomistic systems and more generally to establish the relationship between a specific atomic configuration and the properties that can be computed by *ab initio* methods.[24] In the field of soft matter and colloidal systems, a ML approach based on simple linear regression has been recently employed to efficiently represent many-body interactions in spherical microgel particles in 2D[25] and to build an effective one-component interaction Hamiltonian for a mixture of colloidal hard spheres and non-adsorbing polymer coils.[26] In these approaches, the use of ML techniques was shown to serve as a powerful tool for speeding up by several orders of magnitude simulations that incorporate effective many-body interactions.

The construction of ML potentials follows two main steps: the transformation of the atomic positions into suitable descriptors and the subsequent association of a (free) energy with this structure using a functional form provided by a ML method.[27] In general terms, the descriptors correspond to quantities that are easier to evaluate than the properties one ultimately aims at predicting but strongly correlate with the to-be-predicted properties.[24] In practice, these descriptors are determined by applying geometric and algebraic operations on the Cartesian coordinates of the system, ultimately transforming them into mathematical objects that satisfy the conditions of smoothness and symmetry with respect to isometries.[24] Typical representations based on descriptors of atomic/particle environments involve bond-order parameters and Fourier series of structural invariants.[28] However, the most commonly used representations include atom-centered symmetry functions (SFs)[27,29] and the smooth overlap of atomic positions (SOAP).[30] For the purpose of constructing not only accurate but also computationally efficient potentials for anisotropic colloids or molecules, ML approaches, such as those from Refs. 25 and 26, are appealing. Nevertheless, available descriptors or structure functions, such as the original atom-centered SFs by Behler and Parrinello,[29] are, by construction, spherically symmetric and do not take into account orientation and alignment effects of non-spherical particles. Thus, in order to represent the many-body CG potentials in such systems using ML methods based on local structure characterization, suitable descriptors that capture this crucial aspect are needed.

Within the recent efforts in correlating the structure and dynamics in disordered systems, there have been some attempts

in describing the local structure of systems composed of elongated particles by using spherically symmetric SFs.[28] However, these approaches are simply based on describing continuous elongated bodies, such as ellipsoids by two distinct monomers. Hence, under such considerations, the individual structure functions for a reference elongated particle depend on either the position of the two individual monomers and the centroid of surrounding neighbors or a centroid and the two monomers composing a dimer of neighboring particles. Therefore, descriptors that are able to explicitly and simultaneously incorporate information on the orientation of a reference non-spherical particle and its neighbors are still missing.

Here, we propose two- and three-body orientation-dependent particle-centered descriptors suitable for describing the local structure and constructing many-body potentials for systems composed of rod-like particles, including spherocylinders, rigid linear chains, and ellipsoids. Using a recently proposed feature selection scheme and linear regression,[25,26] we construct effective many-body potentials for model systems of colloidal hard rods and non-adsorbing polymer coils and for core–shell microgel rods. Furthermore, the same approach is used to represent the effective orientation-dependent two-body potential of mean force (PMF) of ligand-stabilized nanorods.

The remainder of this article is organized as follows: in Sec. II, we discuss the orientation-dependent descriptors for rod-like particles. The feature selection scheme and regression method used to construct the ML potentials for the three systems discussed above are briefly described in Sec. III. In this section, we also demonstrate the suitability of the particle-centered orientation-dependent descriptors for encoding information on the local structure of systems composed of prolate ellipsoidal particles. The accuracy of the coarse-grained ML potentials is further tested by directly comparing the results obtained from Monte Carlo simulations of the "true" full systems with those coming from simulations using the reduced-order ML potentials. We conclude with a final discussion and remarks in Sec. IV.

II. SYMMETRY FUNCTIONS FOR ROD-LIKE PARTICLES

To introduce the orientation-dependent symmetry functions (ODSFs), we consider model systems of uniaxial particles with a spherocylindrical shape (cylinders of length L and diameter D capped with two hemispheres at both ends) and ellipsoids of revolution for which two of the perpendicular semi-axes are equal but different, in general, from the third: $\sigma_\parallel \neq \sigma_\perp = \sigma_{\perp'}$. In these cases, each particle i is characterized by its center-of-mass position vector, $\boldsymbol{R}_i$, and the orientation vector of its long axis, $\hat{\boldsymbol{u}}_i$. The descriptors discussed next are a function of scalars that depend on these quantities. We note that, due to their general functional form, the ODSFs are also valid for rigid linear chains of arbitrary dimensions (fused- and tangent-sphere models) and for oblate ellipsoids with an infinite-fold rotational symmetry around the $\hat{\boldsymbol{u}}_i$ axis.

We start the discussion of the functional form of the ODSFs by introducing a cutoff function $f_c(R_{ij})$: a monotonically decreasing function that smoothly goes to 0 in both value and slope at a cutoff distance r_c. Here, we consider a cutoff function of the form

$$f_c(R_{ij}) = \begin{cases} \tanh^3(1 - R_{ij}/r_c) & \text{for } R_{ij} \leq r_c, \\ 0 & \text{for } R_{ij} > r_c, \end{cases} \quad (1)$$

where $R_{ij} = |R_i - R_j|$ is the center-of-mass distance between particles i and j at positions R_i and R_j, respectively. Each symmetry function discussed below is multiplied by one or more cutoff functions to ensure that the total symmetry function decays to zero in value and slope at the cutoff radius. Consequently, particles beyond the cutoff radius do not enter the reference particle contributions. We note that Eq. (1) was introduced in Ref. 27 and also used in Refs. 26 and 31.

To describe the local environment of a rod-like particle, we start from the spherically symmetric two-body radial SFs $G^{(2),R}$ introduced by Behler and Parrinello for constructing high-dimensional neural network potentials,[29]

$$G^{(2),R}(i; \eta, R_s) = \sum_j e^{-\eta(R_{ij} - R_s)^2} \cdot f_c(R_{ij}), \quad (2)$$

which is a sum of Gaussians multiplied by a cutoff function. The width of the Gaussians is defined by parameter η, and the center of the Gaussian distributions can be shifted to a certain radial distance by parameter R_s. In Fig. 1, we plot several example spherically symmetric two-body radial SFs, $G^{(2),R}(i; \eta, R_s)$, for different sets of parameters. While these functions are able to capture two-body correlations between spherically symmetric particles, they are not suitable for encoding information on the relative orientation of rod-like particles. To correctly describe the local environment of such anisotropic bodies, it is necessary to account for the orientational degrees of freedom. To this end, we introduce two orientation-dependent two-body SFs, $G^{(2),OD_1}$ and $G^{(2),OD_2}$, and one orientation-dependent three-body SF, $G^{(3),OD}$. The first ODSF is physically motivated and is based on the assumption that for a reference particle i, the correlation with a neighboring particle j occurs within an ellipsoid of revolution centered around particle i, with a variable spatial extent of the ellipsoid $2\sigma_\parallel$ along the principal axis $\hat{u}_i$ and $2\sigma_\perp$ perpendicular to it. By associating an anisotropic (trivariate) Gaussian with every ellipsoid, we thus compute a two-body function proportional to the mathematical overlap of the Gaussian of the reference particle and that of a neighbor. Such a function is

efficiently represented by the so-called "overlap model" potential as introduced by Berne and Pechukas.[32] The resulting ODSF for particle i is therefore a sum of cut-off functions multiplied by the overlap of its anisotropic Gaussian with those of neighboring particles j,

$$G^{(2),OD_1}(i; \sigma_\parallel, \sigma_\perp) = \sum_j \varepsilon(\hat{u}_i, \hat{u}_j) \cdot e^{-(R_{ij}/\sigma(\hat{u}_i, \hat{u}_j, \hat{R}_{ij}))^2} \cdot f_c(R_{ij}), \quad (3)$$

where $\varepsilon(\hat{u}_i, \hat{u}_j)$ and $\sigma(\hat{u}_i, \hat{u}_j, \hat{R}_{ij})$ are the angle-dependent strength and range parameters, which, for pairs of identical particles, read

$$\varepsilon(\hat{u}_i, \hat{u}_j) = \left[1 - \chi^2(\hat{u}_i \cdot \hat{u}_j)^2\right]^{-1/2}, \quad (4)$$

$$\sigma(\hat{u}_i, \hat{u}_j, \hat{R}_{ij}) = \sigma\left(1 - \frac{1}{2}\chi\left\{\frac{(\hat{R}_{ij} \cdot \hat{u}_i + \hat{R}_{ij} \cdot \hat{u}_j)^2}{[1 + \chi(\hat{u}_i \cdot \hat{u}_j)]} + \frac{(\hat{R}_{ij} \cdot \hat{u}_i - \hat{R}_{ij} \cdot \hat{u}_j)^2}{[1 - \chi(\hat{u}_i \cdot \hat{u}_j)]}\right\}\right)^{-1/2}, \quad (5)$$

where $\hat{R}_{ij} = R_{ij}/R_{ij}$ is the unit vector along the center-of-mass distance vector between particles i and j, $R_{ij} = R_i - R_j$, and $\sigma = \sqrt{2}\sigma_\perp$ and $\chi = (\sigma_\parallel^2 - \sigma_\perp^2)/(\sigma_\parallel^2 + \sigma_\perp^2)$ are the range and anisotropy parameters, respectively. The two parameters that control the shape of this function are the spatial extents of the ellipsoid of revolution along the principal axis and perpendicular to it, i.e., $\sigma_\parallel$ and $\sigma_\perp$. This function carries a direct dependence on the relative orientation of the two rods and is sensitive enough to simultaneously capture information on the inter-particle distances as can be appreciated from Fig. 2.

The second ODSF considered here corresponds to a generalization of the $G^{(2),R}(i; \eta, R_s)$ function to non-spherical particles, where R_{ij} is replaced by the minimum distance between the long axes of the two particles, $d_{m,ij}(R_{ij}, \hat{u}_i, \hat{u}_j)$.[33,54] As in the case of the $G^{(2),R}(i; \eta, R_s)$-type functions, the width of the Gaussians is defined
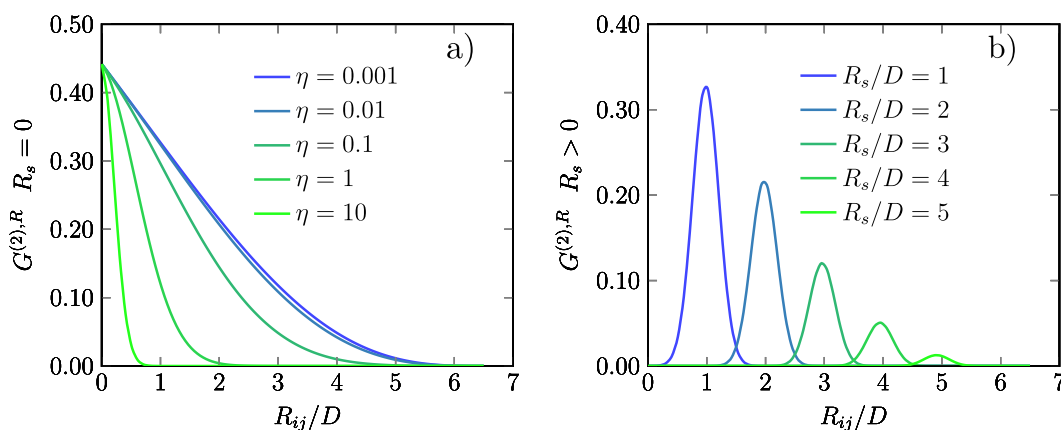
FIG. 1. Two-body radial (R) symmetry function $G^{(2),R}$ as a function of the center-of-mass distance R_{ij}/D between two (rod-like) particles with diameter D. The cut-off value is set to $r_c/D = 6.5$. (a) $G^{(2),R}$ for $R_s = 0$ and different values of η as labeled. (b) $G^{(2),R}$ at fixed $\eta = 10$ and for different shifting distances R_s as labeled.

by parameter α, and the center of the Gaussian distributions can be shifted by parameter R_m,

$$G^{(2),OD_2}(i;\alpha,R_m) = \sum_j e^{-\alpha\left(d_{m,ij}(R_{ij},\hat{u}_i,\hat{u}_j)-R_m\right)^2} \cdot f_c(R_{ij}). \quad (6)$$

In Fig. 3, we exemplify the plot of $G^{(2),OD_2}(i;\alpha,R_m)$ for varying parameters.

Finally, we also introduce an angular three-body ODSF $G^{(3),OD}(i;\sigma_\parallel,\sigma_\perp,\lambda,\xi)$ that depends on the angle $\theta_{ijk} = \arccos(R_{ij} \cdot R_{ik}/(R_{ij}R_{ik}))$ centered at reference particle i,

$$G^{(3),OD}(i;\sigma_\parallel,\sigma_\perp,\lambda,\xi) = 2^{1-\xi}\sum_{j,k\neq i}\left(1+\lambda\cos\theta_{ijk}\right)^\xi \varepsilon_{ij}(\hat{u}_i,\hat{u}_j)\varepsilon_{ik}(\hat{u}_i,\hat{u}_k)\varepsilon_{jk}(\hat{u}_j,\hat{u}_k)$$

$$\times e^{-\left(R_{ij}^2/\sigma_{ij}^2(\hat{u}_i,\hat{u}_j,\hat{R}_{ij})+R_{ik}^2/\sigma_{ik}^2(\hat{u}_i,\hat{u}_k,\hat{R}_{ik})+R_{jk}^2/\sigma_{jk}^2(\hat{u}_j,\hat{u}_k,\hat{R}_{jk})\right)}f_c(R_{ij})f_c(R_{ik})f_c(R_{jk}), \quad (7)$$

where the indices j and k run over all the neighbors of particle i and ξ and λ are two parameters that determine the shape of the function. The parameter λ can have the values $+1$ or -1 and determines the angle θ_{ijk} at which the angular part of the function has its maximum. The angular resolution is provided by parameter ξ, while the terms $\sigma_{ab}(\hat{u}_a,\hat{u}_b,\hat{R}_{ab})$ and $\varepsilon_{ab}(\hat{u}_a,\hat{u}_b)$ control the radial resolution via the anisotropy parameters as defined above. The angular part of this function for $\lambda = 1$ and -1 and different ξ values is shown in Fig. 4.

The ODSFs introduced here provide a rotationally and translationally invariant description of the environment because they depend on the internal coordinates R_{ij}, $d_{m,ij}$, and θ_{ijk} and scalar products of pairs of vectors. Because of the sum over all neighbors within r_c, they are invariant with respect to any permutation of equivalent particles in the environment. As it will become apparent in Sec. III, when describing the local environment of elongated particles, several ODSFs that carry different structural information are combined together. Therefore, the unique significance of the individual functions is generally lost, but when they are used as a group, a more complete and unbiased description of the local structure is achieved.

III. MACHINE-LEARNING POTENTIALS FOR ROD-LIKE PARTICLES

In this section, we demonstrate the suitability of the ODSFs for rod-like particles as descriptors for constructing ML coarse-grained potentials for varying systems with different levels of complexity. More specifically, in order to show the generality of the approach, we chose model systems of prolate Gay–Berne particles, mixtures of spherocylindrical colloids and non-adsorbing polymers, core–shell microgel rods, and ligand-stabilized nanorods. By selecting these models, we cover anisotropic particles represented as prolate ellipsoids, spherocylinders, and rigid linear chains of fused and tangent
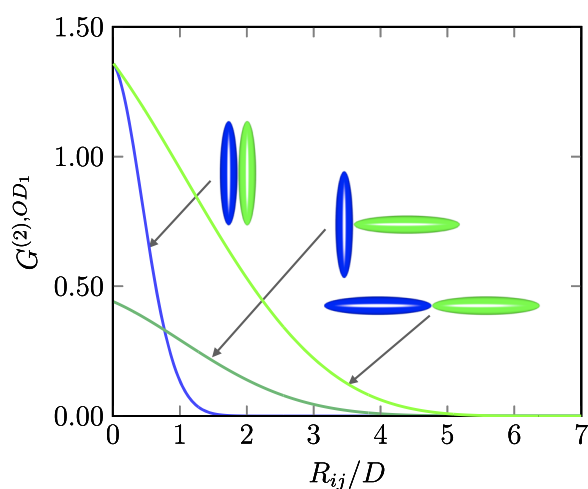
FIG. 2. Two-body orientation-dependent symmetry function $G^{(2),OD_1}$ as a function of the center-of-mass distance R_{ij}/D between two rods with diameter D. Three different relative orientations are considered. The function is shown for values $\sigma_\parallel/D = 3$ and $\sigma_\perp/D = 0.5$. The cut-off value is set to $r_c/D = 6.5$.
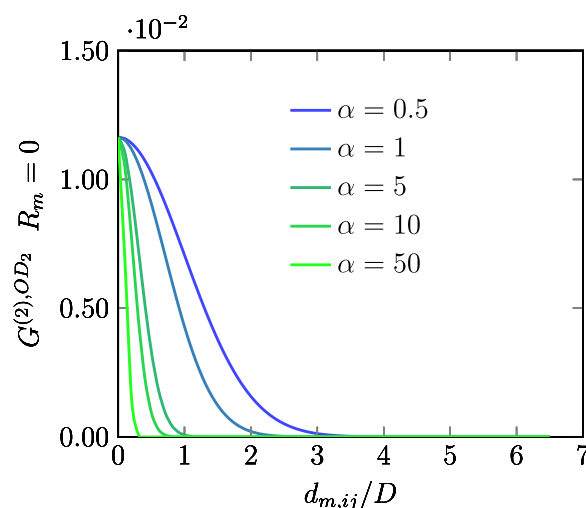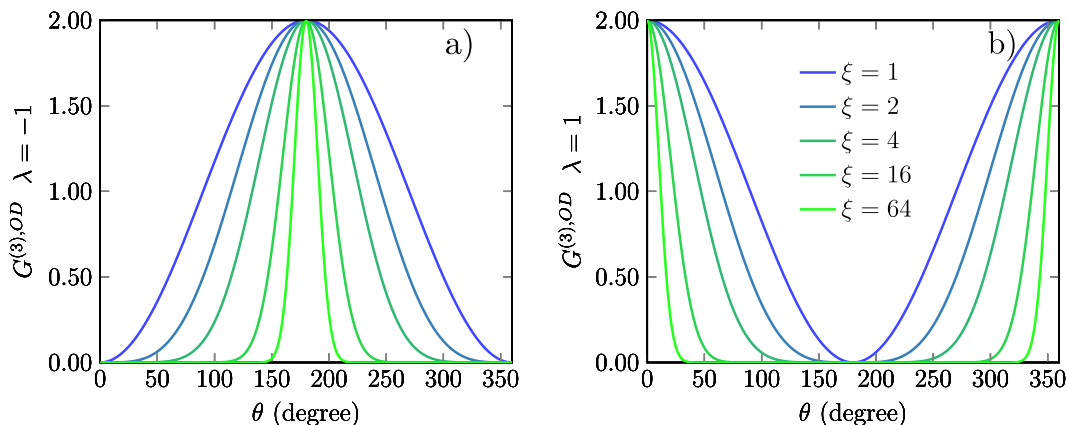
FIG. 3. Two-body orientation-dependent symmetry function $G^{(2),OD_2}$ as a function of the minimum distance $d_{m,ij}/D$ between the central axes of two rod-like particles with diameter D for different values of α as labeled. The cut-off value is set to $r_c/D = 6.5$, and the center-of-mass distance is fixed at $R_{ij}/D = 5$.

FIG. 4. Angular contribution of the three-body orientation-dependent symmetry function $G^{(3),OD}$ for a particle with only two neighbors. (a) $G^{(3),OD}$ for $\lambda = 1$ and (b) $G^{(3),OD}$ for $\lambda = -1$. The function is plotted for different values of ξ as labeled.

sites. In all cases, we construct ML potentials as linear combinations of a number of symmetry functions N_{SF}. Here, the goal is to express the total effective potential of a system of N non-spherical particles with positions $\{\boldsymbol{R}_i\}$ and orientations $\{\hat{\boldsymbol{u}}_i\}$ as

$$\Phi(\{\boldsymbol{R}_i, \hat{\boldsymbol{u}}_i\}) = \sum_i^N \sum_k^{N_{SF}} \omega_k G_k(i), \qquad (8)$$

where $G_k(i)$ is the kth descriptor describing the local environment of particle i and ω_k is the coefficient (weight) of the corresponding SF, which is fixed by the fitting procedure. Note that since $\Phi(\{\boldsymbol{R}_i, \hat{\boldsymbol{u}}_i\})$ is expressed as a sum of per-particle contributions, the fitting of an N-particle system can be extended to simulations with a different number of particles. We select the optimal subset of descriptors using the feature selection scheme of Ref. 25, which we briefly summarize below.

For a given dataset at hand, which typically consists of a collection of particle configurations and the corresponding values of the to-be-predicted quantities (e.g., energy), a training/validation split of the whole data is used. We note that unless stated otherwise, an 80/20 training/validation split of the whole datasets is adopted here. The first step of the method involves the creation of a large but manageable pool of candidate SFs. This is done by calculating, for every particle in the different configurations in the training dataset, M SFs with different sets of parameters. Then, an optimal subset of $N_{SF} < M$ SFs is selected from the pool in a step-wise fashion. The first SF that is selected corresponds to the one with the largest correlation with the target function as quantified by the square of the Pearson correlation coefficient, defined as

$$c_k = \frac{\sum_j \left(\sum_i G_k(i)|_j - \overline{\sum_i G_k(i)} \right) \left(\phi|_j - \overline{\phi} \right)}{\sigma_{SD} \left(\sum_i G_k(i) \right) \sigma_{SD}(\phi)}, \qquad (9)$$

where $\sum_i G_k(i)|_j$ represents the sum of the kth SF over all particles i in configuration j and $\phi|_j$ denotes the target variable evaluated for this configuration. $\overline{\sum_i G_k(i)}$ and $\overline{\phi}$ correspond to arithmetic means over all the configurations in the dataset and $\sigma_{SD}\left(\sum_i G_k(i)\right)$ and $\sigma_{SD}(\phi)$

to their standard deviations. The next SF is then selected based on the highest increase in the linear correlation between the currently selected set and the target data as determined using the coefficient of multiple correlation as follows:

$$R^2 = \boldsymbol{c}^T \boldsymbol{R}^{-1} \boldsymbol{c}, \qquad (10)$$

where $\boldsymbol{c}^T = (c_1, c_2, \ldots)$ is the vector whose ith component is given by the Pearson correlation coefficient, c_i, between the ith SF and the target data, and $\boldsymbol{R}$ is the correlation matrix of the current set of SFs with elements $\mathscr{R}_{ij}$ representing the Pearson correlation function between the ith and jth SFs. In the case of only one SF, R^2 reduces to c_i^2. We note that R^2 can also be computed as the fraction of variance that is explained by a linear fit (including an intercept) of the target function in terms of the SFs in the set. The latter way of computing R^2 turns out to be slightly more expensive but has the advantage of being numerically more stable.[25]

Maximizing the increase in the linear correlation with the target variable guarantees that only SFs that add relevant information are selected.[25] This process is repeated iteratively, and new SFs are selected until the correlation stops increasing appreciably. This indicates that the remaining SFs in the pool add negligible (irrelevant) information to the model. In turn, this constitutes a simple rule to optimize the number of selected SFs as their inclusion would simply imply an unnecessary numerical overhead. All the parameters employed to generate the pool of SFs that are used to build the CG effective potentials for the different systems discussed next are reported in the supplementary material. We note that (if sufficient) using linear regression instead of other more complex schemes, such as nonlinear neural networks, might have some important advantages, namely, the deterministic against the stochastic optimization of model parameters, the control on the number of features, and the final reduced computational cost.[25] The latter aspect, as it can be inferred, is, in part, determined using the optimal value of N_{SF} and the type of descriptor used in the ML potential. In the supplementary material, we briefly discuss the computational cost of the potentials constructed for the systems introduced next.

A. Application to ellipsoids: The Gay–Berne model

To test the performance of the orientation-dependent descriptors, we start by constructing interaction potentials of particles with ellipsoidal symmetry. In particular, for instructive purposes, we use the well-known Gay–Berne (GB) model, in which the anisotropic pair potential reads

$$\phi_{GB}(\boldsymbol{R}_{ij}, \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j) = 4\varepsilon'(\hat{\boldsymbol{R}}_{ij}, \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j)\left[\rho_{ij}^{-12} - \rho_{ij}^{-6}\right], \tag{11}$$

where

$$\rho_{ij} = \frac{R_{ij} - \sigma(\hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j, \hat{\boldsymbol{R}}_{ij}) + \sigma_0}{\sigma_0}, \tag{12}$$

R_{ij} is the distance between the centers of mass of particles i and j, and $\hat{\boldsymbol{R}}_{ij} = \boldsymbol{R}_{ij}/R_{ij}$ is the unit vector along the separation vector $\boldsymbol{R}_{ij} = \boldsymbol{R}_i - \boldsymbol{R}_j$. The anisotropic contact distance $\sigma(\hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j, \hat{\boldsymbol{R}}_{ij})$ and the depth of the interaction energy $\varepsilon'(\hat{\boldsymbol{R}}_{ij}, \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j)$ depend on the orientational unit vector, length-to-breadth ratio ($\kappa = \sigma_\parallel/\sigma_\perp$), and the energy depth anisotropy ($\kappa' = \varepsilon_\perp/\varepsilon_\parallel$), which correspond to the ratio of the size and energy parameters in the end-to-end ($\parallel$) and side-by-side ($\perp$) configurations. The contact distance function is given by Eq. (5), while the depth interaction energy reads

$$\varepsilon'(\hat{\boldsymbol{R}}_{ij}, \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j) = \varepsilon \times \left[\varepsilon(\hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j)\right]^\nu \times \left[\varepsilon_1(\hat{\boldsymbol{R}}_{ij}, \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j)\right]^\mu, \tag{13}$$

where $\varepsilon(\hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j)$ is defined by Eq. (4),

$$\varepsilon_1(\hat{\boldsymbol{R}}_{ij}, \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j) = 1 - \frac{\chi'}{2}\left[\frac{(\hat{\boldsymbol{R}}_{ij}\cdot\hat{\boldsymbol{u}}_i + \hat{\boldsymbol{R}}_{ij}\cdot\hat{\boldsymbol{u}}_j)^2}{[1 + \chi'(\hat{\boldsymbol{u}}_i\cdot\hat{\boldsymbol{u}}_j)]} + \frac{(\hat{\boldsymbol{R}}_{ij}\cdot\hat{\boldsymbol{u}}_i - \hat{\boldsymbol{R}}_{ij}\cdot\hat{\boldsymbol{u}}_j)^2}{[1 - \chi'(\hat{\boldsymbol{u}}_i\cdot\hat{\boldsymbol{u}}_j)]}\right], \tag{14}$$

and $\chi' = \left[(\kappa')^{1/\mu} - 1\right]/\left[(\kappa')^{1/\mu} + 1\right]$. In the above equations, ε and σ_0 represent the energy and length scales of the interaction, respectively. Here, we use the well-known GB model, with characteristic parameters $(\kappa, \kappa', \mu, \nu) = (3.0, 5.0, 2.0, 1.0)$, which were originally used by Gay and Berne.[34]

In particular, we focus on fitting the potential energies ϕ of both dimer and trimer configurations (confined to equilateral triangles only, i.e., $R_{ij} = R_{ik} = R_{jk}$). To generate a training dataset, we perform Monte Carlo (MC) simulations of dimers and trimers at reduced temperature $T^* \equiv k_B T/\varepsilon = 1.0$, where we fix the particle positions, but allow them to freely rotate in space and measure their total potential energy in 250 different samples collected from runs of 2.5×10^6 MC cycles. For each system, 100 separation distances in the range $0.85 \leq R_{ij}/\sigma_0 \leq 4.95$ are considered. The selected range of separation distances contains both high-energy (repulsive) and low-energy (attractive) configurations. With this choice, each set contains a total of 25 000 samples, from which 80% are used for training and 20% are reserved for validation. We perform two different fits: (i) using two-body radial $G^{(2),R}$ SFs, which only depend on the distance between centroids of particles, and (ii) using two-body orientation-dependent $G^{(2),OD_1}$ ODSFs. In both cases, we restrict the number of descriptors to $N_{SF} = 70$ and set the cut-off value at $r_c/\sigma_0 = 6.5$. Parity plots, comparing the original ground truth (training and validation) and predicted potential energies using the two

ML models for the dimers and trimers, are reported in Fig. 5. Since our datasets contain configurations at fixed R_{ij} and varying particle orientations (for which the potential energy differs slightly), small clouds of points are identified. The models constructed with only $G^{(2),R}$ SFs show a correlation coefficient of $R^2 \approx 0.9$ and root mean square errors (RMSE), which are defined by the simple relation $R^2 = 1 - \text{RMSE}^2/\sigma_{SD}^2(\phi)$, of ~0.30$\varepsilon$ and 0.55ε on both the training and validation sets of trimers and dimers, respectively. However, as it can be expected, they present a critical defect as shown in Fig. 5. In particular, since the descriptors are unable to encode information on the orientation-dependence of the interactions, the models assign the same energy to configurations with identical R_{ij} but with a different relative orientation of the particles. In contrast, when $G^{(2),OD_1}$ SFs are employed, the resulting models not only present a correlation coefficient of $R^2 \approx 0.99$ and RMSE values of 0.11ε and 0.46ε for trimers and dimers, respectively, but are also able to capture accurately the orientation-dependence of the interaction energy. If the dimer and trimer datasets are combined and fitted simultaneously, the quality of the resulting model with the same number of features remains virtually unaffected.

B. Effective one-component Hamiltonian for colloidal hard rods and non-adsorbing polymer coils

We continue our discussion by applying the ODSFs for constructing a direct relationship between the structure and effective many-body interactions in systems of rod-like particles with a spherocylindrical shape. Such a particle shape has been widely used to represent generic colloidal nanorods and to investigate their phase behavior and self-assembling behavior.[35–37] Here, we consider a system of sterically stabilized colloidal rods and non-adsorbing polymer coils. By departing from the thermodynamic potential of the full binary mixture, we formally reduce the problem to a colloid-only effective Hamiltonian, which incorporates many-body effects. A ML model is then constructed to represent such an effective Hamiltonian as a function of all colloid coordinates and orientations.

A simple model for the mixture is the so-called Asakura–Oosawa (AO) model, where the colloids are treated as hard particles, while the non-interacting polymer coils are regarded as point particles, which are excluded from the surface of the colloids by a distance equal to the radius of gyration of the polymer R_g.[20,21,38,39] The diameter of the coils is $\sigma_p = 2R_g$. The colloids are represented by hard spherocylinders, which consist of cylinders of diameter σ_c and length L with semi-spherical caps at both ends with diameter σ_c. The pair potentials of this model are given by

$$\phi_{cc}(\boldsymbol{R}_{ij}, \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j) = \begin{cases} \infty & \text{for } d_{m,ij}(\boldsymbol{R}_{ij}, \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j) < \sigma_c, \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

$$\phi_{cp}(\boldsymbol{R}_i - \boldsymbol{r}_j, \hat{\boldsymbol{u}}_i) = \begin{cases} \infty & \text{for } d_{m,ij}(\boldsymbol{R}_i - \boldsymbol{r}_j, \hat{\boldsymbol{u}}_i) < \sigma_{cp}, \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

$$\phi_{pp}(r_{ij}) = 0, \tag{17}$$
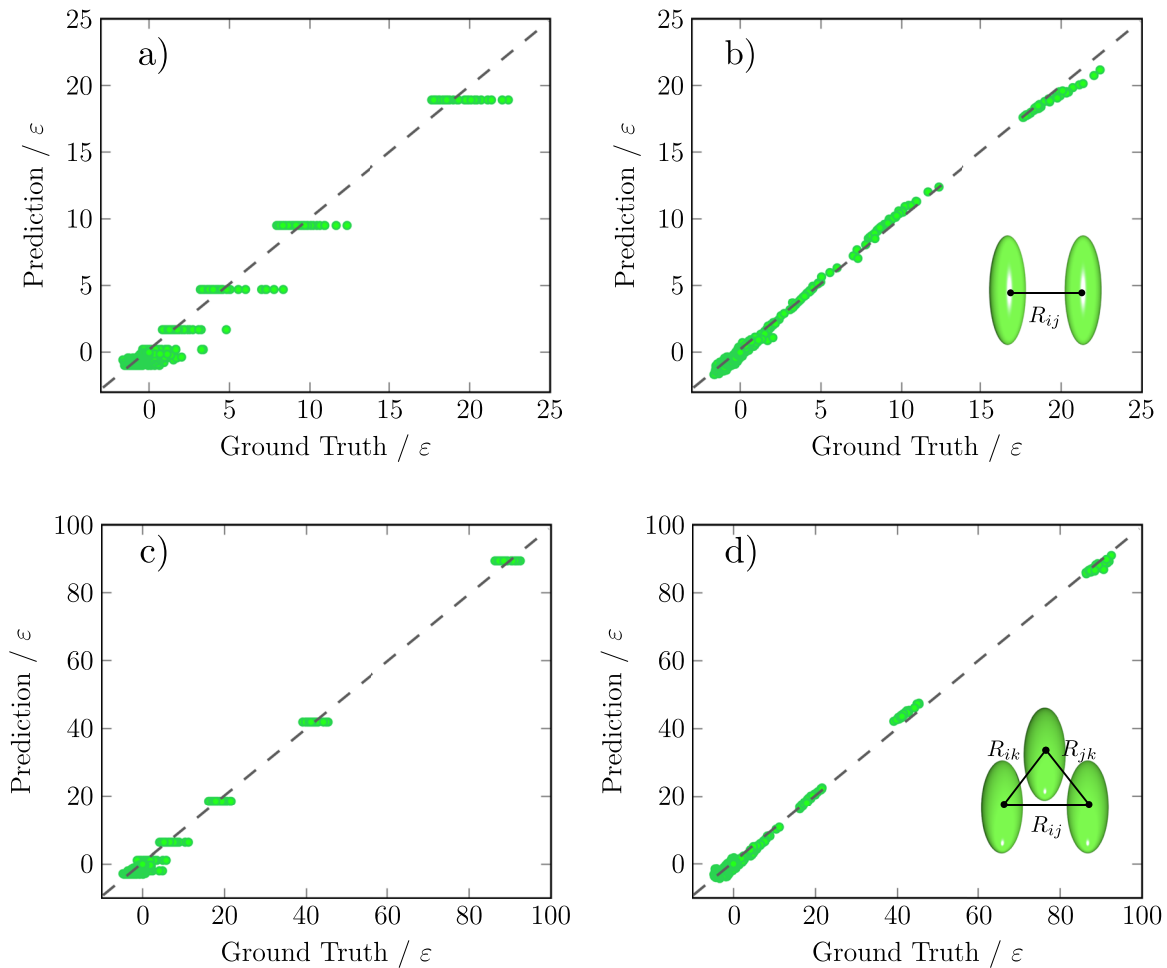
FIG. 5. Parity plots comparing total interaction energies of the true Gay–Berne model with characteristic parameters $(\kappa, \kappa', \mu, \nu) = (3.0, 5.0, 2.0, 1.0)$ and the ML model constructed using only two-body radial $G^{(2),R}$ SFs (left) and with two-body orientation-dependent $G^{(2),OD_1}$ SFs (right). In both models, the number of descriptors is $N_{SF} = 70$. (a) and (b) Results for the dimer configurations. (c) and (d) Comparison for the trimer configurations.

where $\sigma_{cp} = (\sigma_c + \sigma_p)/2$; $\boldsymbol{R}_{ij} = \boldsymbol{R}_i - \boldsymbol{R}_j$ with $\boldsymbol{R}_i$ and $\boldsymbol{R}_j$ the center-of-masses of spherocylinders i and j, respectively; $d_{m,ij}(\boldsymbol{R}_{ij}, \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j)$ is the minimum distance between the central axes of the two spherocylinders with orientations $\hat{\boldsymbol{u}}_i$ and $\hat{\boldsymbol{u}}_j$; $d_{m,ij}(\boldsymbol{R}_i - \boldsymbol{r}_j, \hat{\boldsymbol{u}}_i)$ is the minimum distance between the spherocylinder axis and the polymer center-of-mass at $\boldsymbol{r}_j$ in the case of colloid–polymer interactions; and $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$ is the distance between the two polymer center-of-masses.

The total interaction Hamiltonian of the mixture of N_c colloidal rods and N_p polymer coils in a volume V at absolute temperature T reads $H = H_{cc} + H_{cp} + H_{pp}$, where $H_{cc} = \sum_{i<j}^{N_c} \phi_{cc}(\boldsymbol{R}_{ij}, \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j)$, $H_{cp} = \sum_{i=1}^{N_c} \sum_{j=1}^{N_p} \phi_{cp}(\boldsymbol{R}_i - \boldsymbol{r}_j, \hat{\boldsymbol{u}}_i)$, and $H_{pp} = \sum_{i<j}^{N_p} \phi_{pp}(r_{ij}) = 0$. By keeping constant the number of colloids N_c and treating the polymer coils grand-canonically, in which the polymer fugacity z_p is fixed, the thermodynamic potential of the binary mixture, $F(N_c, z_p, V, T)$, can be written as a function of an effective one-component (only colloids) Hamiltonian where the polymer coils are formally integrated out,

$$\exp[-\beta F] = \frac{1}{N_c! \Lambda_c^{3N_c}} \int_V d\boldsymbol{R}^{N_c} \int d\hat{\boldsymbol{u}}^{N_c} \exp[-\beta H_{\text{eff}}], \qquad (18)$$

where $\beta = (k_{\text{B}} T)^{-1}$ with k_{B} the Boltzmann constant and $H_{\text{eff}} = H_{cc} + \Omega$, with Ω the grand potential of a "sea" of the ideal polymer at fugacity z_p in the external field of a fixed configuration of N_c colloidal rods.[20,21,40] For the AO model, the grand potential Ω reads

$$\Omega = -z_p V_f(\{\boldsymbol{R}_i, \hat{\boldsymbol{u}}_i\}), \qquad (19)$$

where $V_f(\{\boldsymbol{R}_i, \hat{\boldsymbol{u}}_i\})$ is the free volume available for the polymer in the static configuration of N_c colloidal rods with positions $\{\boldsymbol{R}_i\}$ and orientations $\{\hat{\boldsymbol{u}}_i\}$ and with $i = 1, \ldots, N_c$. V_f can be seen as the volume outside the N_c depletion zones and can be decomposed into zero-, one-, two-, three-, and higher-body contributions, $V_f = V_f^{(0)} + \sum_{i=1}^{N_c} V_f^{(1)}(\boldsymbol{R}_i, \hat{\boldsymbol{u}}_i) + \sum_{i<j}^{N_c} V_f^{(2)}(\boldsymbol{R}_i, \boldsymbol{R}_j, \hat{\boldsymbol{u}}_i, \hat{\boldsymbol{u}}_j) + V_f^{(3+)}$. In turn, for a fixed colloid configuration, the volume outside the

N_c depletion zones will be determined by the size ratio $q = \sigma_p/\sigma_c$ between the polymer coils and colloids. For $q < 0.1547$, a mapping onto an effective one-component system with an effective Hamiltonian based on pairwise additive depletion potentials is exact.[21,41,42] However, for larger q, the many-body contributions in $V_f^{(3+)}$ must be considered.

For spherical colloids, the effective pair potential is given by the AO depletion potential that is known analytically.[20] However, no analogous analytic expression exists for the overlap volume of two depletion zones of finite spherocylinders with arbitrary orientations and positions, and hence, $V_f^{(2)}$ can only be approximated as in Ref. 42 or calculated numerically. It turns out that the numerical evaluation of $V_f^{(2)}$ is almost as expensive as that of the whole free volume V_f, which prevents us from using such an approach in long simulations of large systems. Therefore, in order to construct for the first time a full many-body effective one-component interaction Hamiltonian for such a mixture, which is computationally efficient, we fit V_f as a function of all colloid coordinates using a set of ODSFs as discussed above.

To build the training dataset, we perform MC simulations on $N_c = 768$ colloids with $L/\sigma_c = 5$, size ratio $q = 1.0$, polymer fugacity $z_p = 0$, and colloid packing fraction $\eta_c = \pi\sigma_c^2(2\sigma_c/3 + L)N_c/(4V)$ $\in [0.13, 0.66]$, with a packing fraction spacing of $\delta\eta_c = 0.005$. From each simulation of 1×10^7 MC cycles, consisting of N_c attempts of rotating or translating particles, we collect 300 equilibrated, well-spaced configurations and measure V_f using a numerical integration.[20,26] The resulting dataset contains a total of 27 900 representative particle configurations at different colloid densities, from which 80% for training and 20% for testing are used. Among the different thermodynamic states used for training, isotropic (I), nematic (N), smectic (Sm), and crystal (X) phases,[35] which are characterized by different degrees of orientational and positional order, are considered. To quantify the importance of the many-body contributions to the effective potential in each of the stable phases, we calculate $P(n)$, the probability that we find $n = n(\mathbf{r})$ overlapping depletion layers at spatial coordinate $\mathbf{r}$.[20,26] In Fig. 6, we show $P(n)$ for varying packing fractions η_c, including examples of I, N, Sm, and X phases. In the high-density X phase at $\eta_c = 0.66$, we find that even four-body contributions to V_f become non-negligible.

To fit V_f, we create a manageable pool of $M = 365$ candidate SFs, setting the cut-off value at $r_c/\sigma_c = 7.0$. The results of the fitting are reported in Fig. 7. In particular, we show the RMSE of the linear fits with the actual V_f as a function of the number of selected SFs for both the training and test sets. For $N_{SF} > 104$, the correlation coefficient is $R^2 \approx 0.99$ and the normalized RMSE, defined as NRMSE $\equiv$ RMSE$/|V_f^{\max} - V_f^{\min}|$, is on the order of 1×10^{-3}, which clearly indicates the quality of the linear fit and ultimately the ability of the ODSFs as fingerprints to encode information about the structure of the system of non-spherical particles. As described above, the feature selection method used here sequentially picks up the descriptors from the pool of candidate SFs based on the increase in linear correlation. Figure 7(b) shows the type of descriptor that is sequentially selected from the pool of candidate SFs. We clearly observe that the $G^{(2),OD_2}(i; \alpha, R_s)$ descriptors describe best the variance of the data.
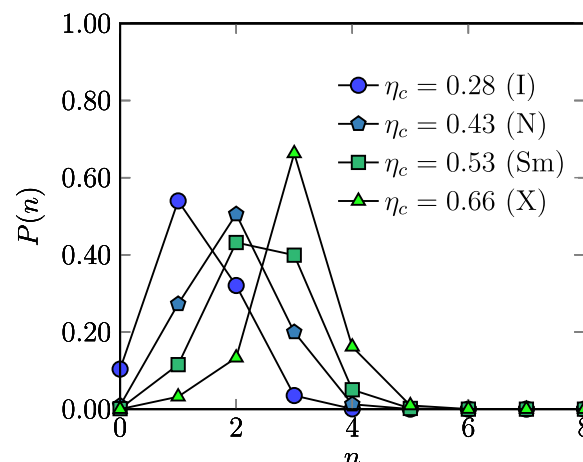

FIG. 6. Probability of n overlapping depletion layers for a mixture of colloidal rods with a length-to-diameter ratio $L/\sigma_c = 5$ and non-adsorbing polymer coils of diameter σ_p with size ratio $q = \sigma_p/\sigma_c = 1$ at polymer fugacity $z_p = 0$ and colloid packing fraction η_c as labeled.

To validate the ML model and in order to assess its transferability, we calculate the (partial) phase diagram of the effective one-component system using canonical MC simulations. In particular, we focus on the binodals for the isotropic fluid phases.[42] To compute such binodals, we perform direct coexistence simulations of the effective one-component system of $N_c = 896$ colloids in elongated simulation boxes of volume V with edges $L_x = L_y \geq 3L$ and $L_z = 3L_x$ at varying polymer fugacities z_p. We fit the effective many-body Hamiltonian as evaluated in simulations on the full binary system with $N_{SF} = 120$, which provides a NRMSE of 8×10^{-4}. In order to confirm the accuracy of the fitted model, we also compute the coexistence curves from grand canonical Monte Carlo (GCMC) simulations of the full rod-sphere binary mixture, where N_c, z_p, V, and T are kept constant. In Fig. 8, we report the phase diagrams in the (η_c, z_p) plane as obtained from simulations of the two models. At sufficiently high polymer fugacity, we find a coexistence between a low-density isotropic gas I_G phase and a high-density isotropic liquid I_L phase. We find a good correspondence between the two models and also agreement with previous results from GCMC simulations by Jungblut *et al.*[43] For $z_p \geq 1.38$, a broad coexistence between the isotropic gas I_G and nematic N phase is observed. Again, we find good agreement between the ML model and the "true" binary mixture. We highlight that since we do not include any two-phase systems with interfaces in our training set nor configurations at non-zero z_p, reproducing the coexisting curves with MC simulations of the ML model proves the transferability to finite z_p and also that the ML potential captures the dependence on density that a realistic description must have to be able to calculate phase diagrams accurately.

C. Many-body interactions of core–shell microgel rods

The third model we consider is a coarse-grained representation of core–shell microgel (CSM) rods based on the models proposed in Refs. 25 and 44. These models capture the many-body
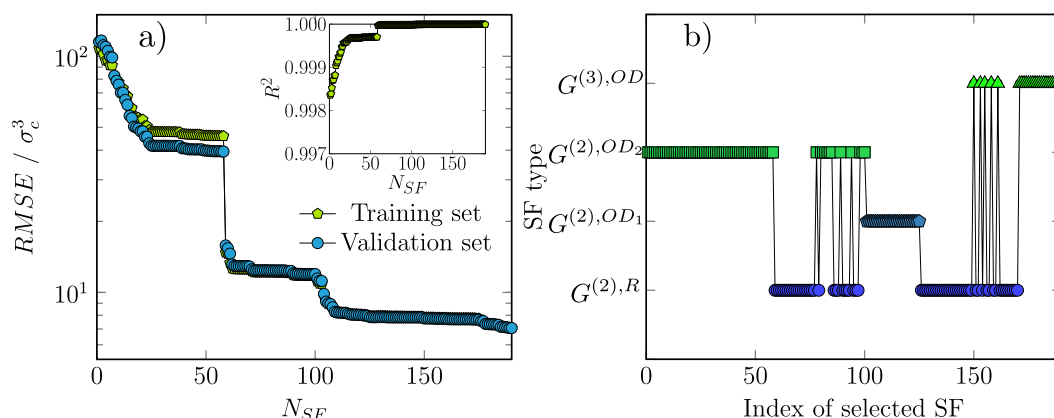
FIG. 7. (a) Root mean square error (RMSE) as a function of the number of SFs in the subset N_{SF} for the effective one-component many-body Hamiltonian of a mixture of sterically stabilized colloidal rods and non-adsorbing polymer coils. The RMSE values (in σ_c^3 units) are shown for the training and validation sets. The correlation coefficient R^2 as a function of N_{SF} is shown in the inset. (b) Type of SF as a function of the index chosen in the feature selection method.

interactions arising when elastic spheres are strongly deformed at high densities and high pressures.

Here, we represent the core–shell microgel rod as a hard spherocylindrical core surrounded by a deformable shell. To represent the shape anisotropy of the particle, we describe the rod as an assembly of deformable spheres with rigid hard cores. The particle is constructed as follows: the rigid hard cores of the particles are represented in a "linear tangent" fashion,[45] where n spherical hard cores of diameter σ_A are linearly aligned and tangent to their neighbors

as depicted in Fig. 9(a). Subsequently, each rigid core is surrounded by a deformable shell. The diameter of the rod is $\sigma_B = \sigma_A + \lambda$, where $\lambda/2$ is the effective length of the deformable shell. In order to capture the elastic deformation due to an interaction with another CSM particle, we discretize the surface of the deformable particle by a large number of N_p points, which are evenly distributed on the surface. When a microgel particle is deformed due to overlap with another particle, each point on the surface of this deformable shell will be pushed radially to the point of intersection of the two spheres. The
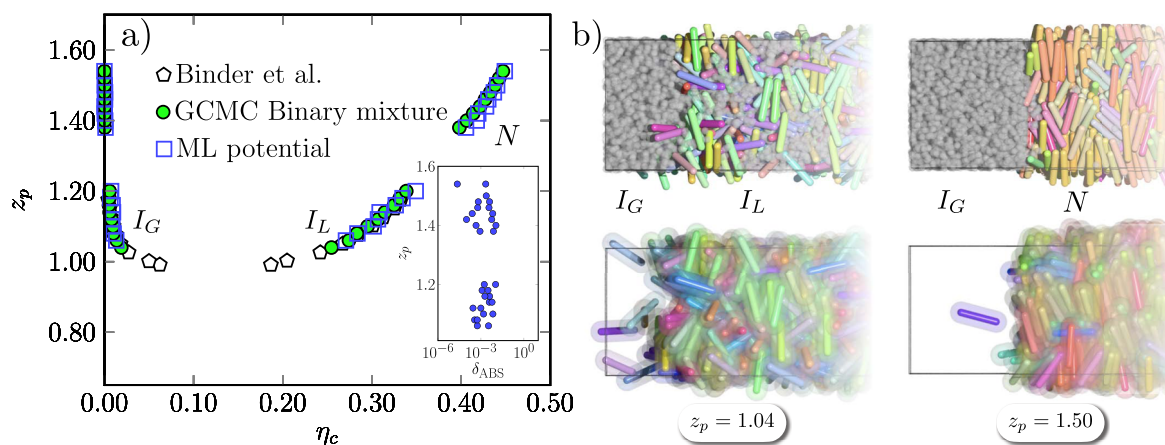
FIG. 8. (a) Phase diagram of a mixture of sterically stabilized colloidal rods of length-to-diameter ratio $L/\sigma_c = 5$ and non-adsorbing polymer coils of diameter $\sigma_p = 1.0$ in the colloid packing fraction η_c–polymer fugacity z_p representation. The size ratio of the mixture is $q = \sigma_p/\sigma_c = 1$. Filled circles correspond to results obtained from grand-canonical Monte Carlo (GCMC) simulations of the true binary mixture, whereas empty squares represent those obtained from direct coexistence Monte Carlo simulations using the fitted many-body ML potential. Empty pentagons represent previous results reported by Jungblut et al.[43] A broad coexistence region between an isotropic gas phase I_G and an isotropic liquid phase I_L is identified. At high polymer fugacity z_p, the I_G phase is in equilibrium with a nematic N phase. The absolute errors ($\delta_{ABS} \equiv |\eta_{c,ML} - \eta_{c,GCMC}|$) of the coexisting colloid packing fraction predictions are shown in the inset as a function on the polymer fugacity z_p. (b) Typical configurations of the fluid phase equilibria of the colloid–polymer mixture as obtained from direct coexistence simulations using both the full binary mixture (top) and the CG ML potential (bottom). The gray spheres correspond to the polymer coils, and in the CG representation, the depletion layers of every spherocylinder are displayed. The color of the rod-like particles is assigned according to their orientation in space.
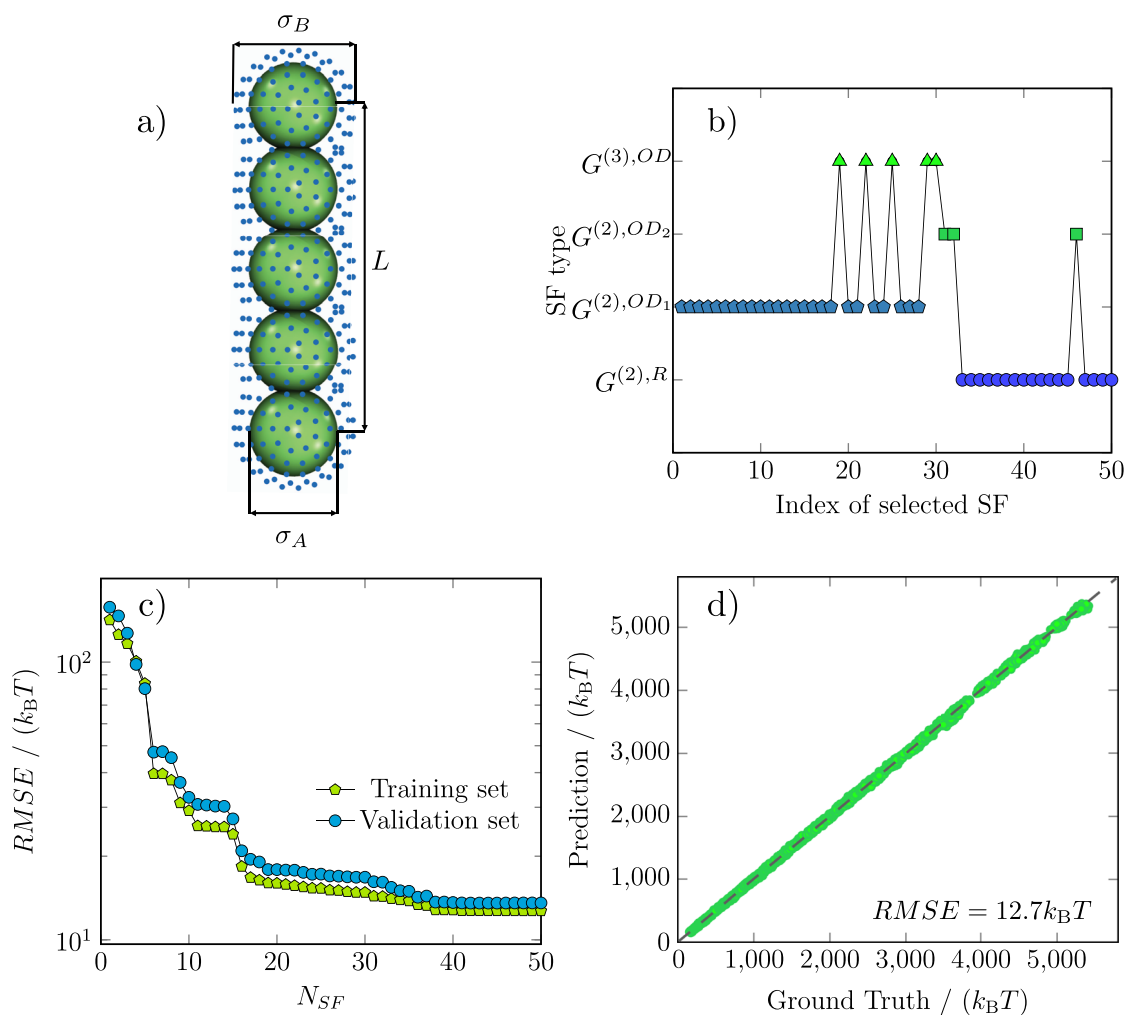
FIG. 9. (a) Schematic picture of a core–shell microgel (CSM) rod-like particle. The core consisting of tangent hard spheres is depicted in green. The surface of the elastic microgel shell is shown in blue, where only the points on the surface are depicted for clarity. The visual representation of the deformation of the shell due to an overlap between two different rods (top view) is shown below. (b) Type of SF as a function of the index chosen in the feature selection method. (c) Root mean square error (RMSE; in k_BT units) as a function of the number N_{SF} of SFs, where both the training and validation sets are plotted. The inset of (c) shows the correlation coefficient R^2 as a function of N_{SF}. (d) Parity plot comparing the total elastic energy predictions of the ML model and the true model in configurations consisting of $N_c = 768$ CSM rod-like particles.

elastic energy corresponding to the deformation of particle i is then approximated by the weighted sum of the elastic energies associated with each point on its surface as[25]

$$\beta\phi_i = \frac{K}{2}\sum_{l=1}^{N_p}\frac{w_l}{\sigma_B^2}\left(\frac{\delta r_l}{\sigma_B}\right)^2, \tag{20}$$

where K is an elastic constant with the dimension of energy, l runs over all the N_p points on the surface of the deformable particle i, w_l is the surface area of the associated Voronoi cell of point l, and δr_l is the length of the radial deformation from the surface of particle i to the plane of intersection of the two spheres, as shown in Fig. 9(a). When particle i overlaps with more than one particle at the

same time, we evaluate for each point l on the surface of particle i δr_l^j for each interacting particle j and use the maximum deformation to evaluate the elastic energy, i.e., the length of the radial deformation from the surface of particle i to the intersection of all the spheres is $\delta r_l = \max \delta r_l^j$.

We consider a core–shell microgel rod with a rigid core consisting of $n = 5$ hard spheres. Following our notation of the hard-rod model, the length of the hard rod that forms the CSM particle is $L = 4\sigma_A$ with semi-spherical caps at both ends with diameter σ_A. The elastic particle has a diameter of $\sigma_B = 1.35\sigma_A$, corresponding to $\lambda/2 = 0.175\sigma_A$. We discretize the surface of each sphere composing the rod with $N_p = 200$ points, which we find a good compromise between accuracy and efficiency. Finally, we set the elastic constant to $K = 500k_BT$.

The computational cost of evaluating the elastic energy of deformable spheres is very high, as all the N_p points on the surface of each sphere have to be taken into account when a Monte Carlo move is performed. The use of ML potentials to incorporate the many-body interactions of a 2D system of elastic spheres speeded up the energy evaluations considerably.[25] The need of a ML technique for the present model is even more crucial as there is a dramatic increase of points $N_p = 200 \times 5$ on the surface of each microgel rod to be evaluated. Hence, this model represents a perfect example where the orientation-dependent symmetry functions can be used to reduce the computational cost in simulating these systems.

In order to apply the ODSFs to the CSM rod-like particles, we first build a dataset based on the total elastic energy associated with configurations in different conditions. Due to the high computational cost of the model, we use the same equilibrated configurations as in Sec. III B that correspond to $N_c = 768$ rod-like particles. For each configuration, we calculate the elastic energy of each particle using Eq. (20) and subsequently the total elastic energy of the system. The training and validation configurations correspond to various thermodynamic states ranging from isotropic, nematic, smectic to crystalline phases. We proceed as in Sec. III B by generating a pool of ODSFs and selecting a subset of them that capture best the variance of the target energy. Here, we use a cut-off of $r_c/\sigma_A = 6.5$ and take 50 ODSFs, which give $R^2 \approx 0.99$ and minimize the root mean square error on both training and validation sets, providing a NRMSE of 2×10^{-3}. In contrast with the model shown in Sec. III B, the variance of the data is best captured using the $G^{(2),OD_1}(i; \sigma_\parallel, \sigma_\perp)$ descriptors as shown in Fig. 9(b). Note that with just a small number of ODSFs, the data are already well described as the correlation coefficient approaches quickly to $R^2 \approx 0.99$ and the RMSE is quickly minimized. Finally, we test the fitting by predicting the energies of the validation set. In Fig. 9(d), we show the comparison of the predicted energy and the one calculated from Eq. (20) and observe a perfect agreement between the two of them. We thus show that the ODSFs work well to capture the many-body elastic energy of CSM rod-like particles. This opens the door to explore the phase behavior and structural properties of complex systems, such as CSM rods.

D. Two-body potential of mean force of ligand-stabilized nanorods

Finally, we consider a chemistry-specific CG model of ligand-stabilized nanorods, for which we use the ODSFs to fit the effective two-body potential of mean force (PMF). In this model, the ligands are represented as chains of five CG beads, approximately corresponding to alkyl ligands of 18 carbon atoms and a headgroup (e.g., thiol or amine), as shown in Fig. 10(a). Nanocrystal cores (NCs) are modeled as rigid bodies composed of a cylinder of length $L = 20$ nm and diameter $D = 4.2$ nm, with semi-spherical caps at both ends. The orientation of a particle is thus determined using the central axis of its cylindrical core. The surface coverage, calculated as the number of ligand molecules per surface area, is 5.5 nm^{-2}.[46] The CG ligands are covalently bonded to the nanocrystal cores, and the interactions between the constituent CG beads are described by the MARTINI force field. For simplicity, we consider only "C1"-type MARTINI beads.[47] To account for the solvent implicitly, we use the approach reported by Fan and Grünwald,[48] where pair interactions between

non-bonded beads are described through a modified Lennard-Jones (LJ) potential, which reads

$$\phi(r; s) = \begin{cases} \phi^{LJ}(r) + (1-s)\varepsilon & \text{for } r \leq 2^{1/6}\sigma, \\ s\phi^{LJ}(r) & \text{for } 2^{1/6}\sigma < r \leq r_c, \end{cases} \quad (21)$$

where the quality of the solvent is controlled by the parameter $0 \leq s \leq 1$ with $s = 0$ corresponding to a good solvent and $s = 1$ to a bad solvent or vacuum, and

$$\phi^{LJ}(r) = 4\varepsilon\left[\left(\frac{\sigma}{\varepsilon}\right)^{12} - \left(\frac{\sigma}{\varepsilon}\right)^6\right] \quad (22)$$

is the standard Lennard-Jones (LJ) potential, where $\sigma = 0.47$ nm and $\varepsilon = 0.8365$ kcal mol^{-1} are the length- and energy-scale parameters of the pair interaction, respectively; r is the separation distance between pairs of coarse-grained sites; and $r_c = 1.2$ nm is the cut-off radius of the interactions. We set the solvent parameter $s = 0.3$. Within the chains, intramolecular interactions acting on the centers of bonded CG sites are described using a harmonic bond-stretching potential as follows:

$$\phi^{bond}(b) = \frac{1}{2}K_b(b - b_0)^2, \quad (23)$$

with the bond force constant $K_b = 149.3787$ kcal mol^{-1} nm^{-2} and b and $b_0 = 0.47$ nm the instantaneous and equilibrium bond distances, respectively. Similarly, the angle-bending between triplets of connected beads is modeled via a harmonic potential as follows:

$$\phi^{angle}(\theta) = \frac{1}{2}K_\theta(\cos\theta - \cos\theta_0)^2, \quad (24)$$

where $K_\theta = 2.9876$ kcal mol^{-1} denotes the angle force constant and θ and $\theta_0 = 180°$ the instantaneous and equilibrium angle-bending values, respectively. Interactions between NC cores are neglected as these forces, for small NCs, are typically much weaker than ligand–ligand interactions.[22,49,50] We perform molecular dynamics (MD) simulations on systems of 21132 CG beads using the software package Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS).[51] To avoid finite size effects, we employ a cubic simulation box of side length $10L$. We remove overlaps of ligand beads in the initial configurations by energy minimization. Simulations are performed in the canonical ensemble (NVT) at temperature $T = 300$ K, which is maintained constant with a Nosé–Hoover thermostat. The equations of motion are integrated with a time step of 20 fs for up to 10^7 steps, and statistics are collected over the last 10^6 steps.

To compute the effective pair interactions between the ligand-stabilized nanorods, we calculate the PMF for three different relative particle orientations, as shown in Figs. 10(b)–10(d), using constraint MD simulations.[52] For each PMF calculation, we perform simulations with the nanorod cores frozen at 120 different distances R_{ij}. For each of these simulations, the mean force F_m is calculated as the average force between the centers of mass of the two nanorods along the centre-of-mass distance vector $\mathbf{R}_{ij}$,[22]

$$F_m(R_{ij}) = \frac{1}{2}\left\langle(\mathbf{F}_i - \mathbf{F}_j) \cdot \hat{\mathbf{R}}_{ij}\right\rangle_{NVT; R_{ij}, \hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j}, \quad (25)$$
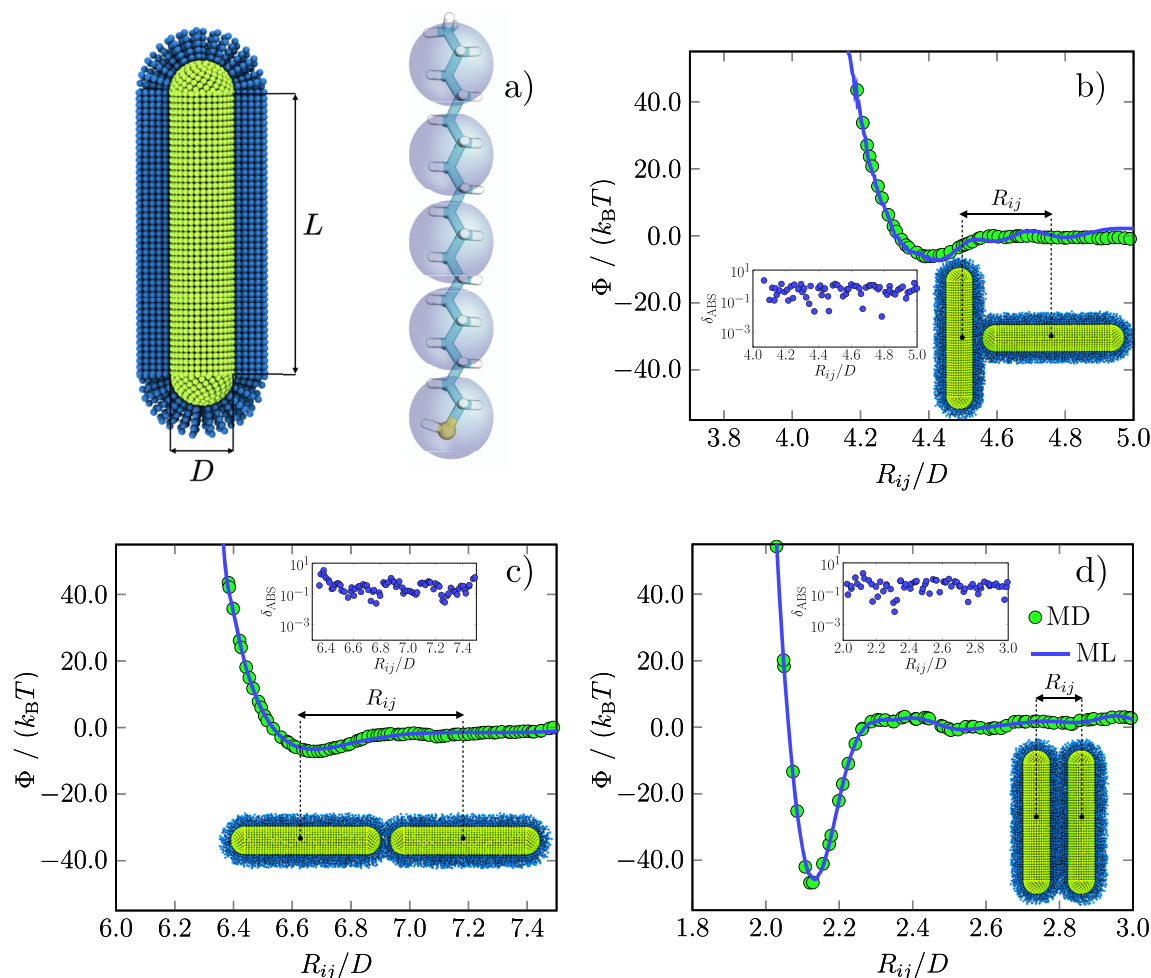
FIG. 10. (a) Schematic representation of a ligand-stabilized nanorod of length L and diameter D. Only a portion of the ligands is shown to better appreciate the morphology of the particles. The blue beads correspond to CG ligand atoms, and the green beads represent the nanocrystal core atoms. In the CG representation, a single ligand (octadecanethiol) contains five CG beads that represent 18 carbon atoms (cyan) and a thiol group (yellow). (b)–(d) Potential of mean force Φ as a function of center of mass distance R_{ij} for three different relative nanorod orientations, as illustrated in the snapshot of each plot. Green filled circles correspond to the values obtained from constraint MD simulations, whereas blue lines correspond to the values predicted by the ML model. The absolute errors ($\delta_{ABS} \equiv |\Phi_{ML} - \Phi_{MD}|$) of the potential of mean force predictions are shown in the insets as a function of the center of mass distance R_{ij}.

where $\mathbf{F}_i$ and $\mathbf{F}_j$ are the total forces acting on the center of mass of each core and $\hat{\mathbf{R}}_{ij}$ is the unit vector connecting the two rods along the reaction coordinate R_{ij}. Angular brackets denote ensemble averages in the canonical ensemble with the constraint separation R_{ij} and orientations $\hat{\boldsymbol{u}}_i$ and $\hat{\boldsymbol{u}}_j$ of the NCs. The PMF $\Phi(R_{ij})$ is then computed using

$$\Phi(R_{ij}) = \int_{R_{ij}}^{\infty} F_m(R_{ij}) dR'_{ij}. \tag{26}$$

In Figs. 10(b)–10(d), the computed PMF curves clearly show that the effective interactions between the nanorods are repulsive at short distances and attractive at larger distances. In particular, we find that when the particles are parallel aligned, the strength of the effective

attractive interaction is the strongest, with a deep local minimum of $-46.7k_BT$ at $R_{ij}/D = 2.1$. The value of the well depth for the side-by-side configuration is around seven times larger than that of the other two configurations.

To employ the ML approach explained above, we combine the data points of the three measured PMFs into one training dataset that includes particle–particle distances with corresponding orientations and free energies associated with them. We perform the fits using two-body ODSFs with a cut-off of $r_c/D = 8.0$ and restrict the number of descriptors to $N_{SF} = 100$. In the supplementary material, we show the RMSE as a function of the number of SFs and the type of SF as a function of the index chosen in the feature selection scheme. The ML models constructed with the two ODSFs introduced in Sec. II present a correlation coefficient $R^2 \approx 0.99$ and a

NRMSE of 7.8×10^{-3}. To further test the accuracy of the ML model, we evaluate Φ using a single interaction site representation for the particles. In particular, for each of the three relative orientations, we generate up to 500 configurations at fixed distance R_{ij} and evaluate the PMF using the fitted model. Considering that the generated particle configurations are all different from those included in the original training dataset, the agreement with the values obtained from the MD simulations [see Figs. 10(b)–10(d)] highlights the ability of the ML model to accurately interpolate between structures and smoothly recover (predict) Φ. This coarse-graining strategy, which is based on a mapping that projects fine-grained configurations of the complex nanorods onto a lower-dimensional representation (a single site model), can be used to efficiently explore the phase behavior and structural properties by long simulations of large system sizes. The quality of the fitted CG potential will obviously reflect that of the underlying fine-grained model. Thus, its predictive power and accuracy would strongly depend on the quality of the training dataset (e.g., the number of training samples and the numerical precision of the measured PMF). Furthermore, as discussed above, for a given dataset, the quality of the fitted CG potential can be controlled in our approach by the number of descriptors N_{SF}.

IV. CONCLUSIONS

In summary, we have proposed new orientation-dependent particle-centered descriptors that effectively map a static configuration of anisotropic rod-like particles into a suitable representation that can be employed to construct a machine learning model to regress a structure–property relation. To demonstrate the ability of the functions in describing orientation and alignment effects, we have used simple linear regression to construct an effective single-component Hamiltonian for hard colloidal rods and non-adsorbing polymer coils by formally integrating out the polymers and fitting the grand potential (free volume) that incorporates many-body effects. The resulting ML potential was used in direct coexistence simulations to calculate the phase diagram of the mixture. We found good agreement with the results obtained from simulations of the true binary mixture. Additionally, an accurate and computationally efficient many-body interaction potential of anisotropic core–shell microgel particles has been fitted on the basis of the proposed descriptors. The same approach has also been used to represent the effective orientation-dependent two-body potential of mean force of a chemistry-specific nearly-atomistic model of ligand-stabilized nanorods.

The methodology presented here can be seen as a bottom-up coarse-graining strategy valuable for speeding up simulations of complex anisotropic particles. The speedup achieved with the ML potentials depends on the details of the precise underlying fine-grained model and the type and number of descriptors that are used for the construction of the ML potentials. Roughly, we find that the ML potentials are 20 times slower for the case of Gay–Berne ellipsoids, three orders of magnitude faster for the colloid–polymer mixtures, and three times faster for the microgel particles (see the supplementary material). It would be interesting to extend this approach to construct CG models with anisotropic building blocks (superatoms and beads) that are able to capture the large-scale properties of specific molecular systems, in the same spirit as the ellipsoid-based models developed for semiflexible polymers.[53] This strategy may also allow one to move from idealized to predictive models of mesogenic molecules, relevant in the field of liquid crystals. Furthermore, the proposed functions could be exploited in a ML approach to address how the orientational structure and shape anisotropy determine the dynamics of elongated particles. Finally, we note that the functions discussed here are perfectly suited for describing the orientational and translational degrees of freedom of anisotropic uniaxial particles with an elongated or oblate shape. Therefore, it would be of interest to develop new descriptors suited for mapping the configurations of non-spherical particles with more complex symmetry (e.g., biaxial particles).

SUPPLEMENTARY MATERIAL

See the supplementary material for a complete list of parameters used to construct the initial pool of candidate SFs for building the CG potentials for the different models, learning curves, and a brief discussion on the computational efficiency of the ML potentials. Additionally, simple example FORTRAN functions to evaluate the SFs are also provided.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Gerardo Campos-Villalobos: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Software (lead); Writing – original draft (lead); Writing – review & editing (lead). **Giuliana Giunta:** Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Writing – original draft (equal); Writing – review & editing (equal). **Susana Marín-Aguilar:** Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Writing – original draft (equal); Writing – review & editing (equal). **Marjolein Dijkstra:** Conceptualization (lead); Funding acquisition (lead); Investigation (lead); Project administration (lead); Resources (lead); Supervision (lead); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] C. M. Care and D. J. Cleaver, "Computer simulation of liquid crystals," Rep. Prog. Phys. **68**, 2665 (2005).

[2] M. P. Allen, "Molecular simulation of liquid crystals," Mol. Phys. **117**, 2391–2417 (2019).

[3] B. Ruzicka and E. Zaccarelli, "A fresh look at the Laponite phase diagram," Soft Matter **7**, 1268–1286 (2011).

[4] S. C. Glotzer and M. J. Solomon, "Anisotropy of building blocks and their assembly into complex structures," Nat. Mater. **6**, 557–562 (2007).

[5] C. Wetter, "Die flüssigkristalle des tabakmosaikvirus," Biol. Unserer Zeit **15**, 81–89 (1985).

[6] Z. Dogic and S. Fraden, "Smectic phase in a colloidal suspension of semiflexible virus particles," Phys. Rev. Lett. **78**, 2417 (1997).

[7] E. Grelet, "Hard-rod behavior in dense mesophases of semiflexible and rigid charged viruses," Phys. Rev. X **4**, 021053 (2014).

[8] P. A. Buining, A. P. Philipse, and H. N. W. Lekkerkerker, "Phase behavior of aqueous dispersions of colloidal boehmite rods," Langmuir **10**, 2106–2114 (1994).

[9] H. Maeda and Y. Maeda, "Liquid crystal formation in suspensions of hard rodlike colloidal particles: Direct observation of particle arrangement and self-ordering behavior," Phys. Rev. Lett. **90**, 018303 (2003).

[10] B. D. Busbee, S. O. Obare, and C. J. Murphy, "An improved synthesis of high-aspect-ratio gold nanorods," Adv. Mater. **15**, 414–416 (2003).

[11] A. Kuijk, D. V. Byelov, A. V. Petukhov, A. Van Blaaderen, and A. Imhof, "Phase behavior of colloidal silica rods," Faraday Discuss. **159**, 181–199 (2012).

[12] C. Schütz, M. Agthe, A. B. Fall, K. Gordeyeva, V. Guccini, M. Salajková, T. S. Plivelic, J. P. F. Lagerwall, G. Salazar-Alvarez, and L. Bergström, "Rod packing in chiral nematic cellulose nanocrystal dispersions studied by small-angle X-ray scattering and laser diffraction," Langmuir **31**, 6507–6513 (2015).

[13] S. N. Hosseini, A. Grau-Carbonell, A. G. Nikolaenkova, X. Xie, X. Chen, A. Imhof, A. Blaaderen, and P. J. Baesjou, "Smectic liquid crystalline titanium dioxide nanorods: Reducing attractions by optimizing ligand density," Adv. Funct. Mater. **30**, 2005491 (2020).

[14] Y. Xie, S. Guo, Y. Ji, C. Guo, X. Liu, Z. Chen, X. Wu, and Q. Liu, "Self-assembly of gold nanorods into symmetric superlattices directed by OH-terminated hexa(ethylene glycol) alkanethiol," Langmuir **27**, 11394–11400 (2011).

[15] G. Odriozola and F. d. J. Guevara-Rodríguez, "Communication: Equation of state of hard oblate ellipsoids by replica exchange Monte Carlo," J. Chem. Phys. **134**, 201103 (2011).

[16] P. G. Bolhuis, A. A. Louis, and J. P. Hansen, "Many-body interactions and correlations in coarse-grained descriptions of polymer solutions," Phys. Rev. E **64**, 021801 (2001).

[17] C. N. Likos, "Effective interactions in soft condensed matter physics," Phys. Rep. **348**, 267–439 (2001).

[18] M. Watzlawek, C. N. Likos, and H. Löwen, "Phase diagram of star polymer solutions," Phys. Rev. Lett. **82**, 5289 (1999).

[19] C. Von Ferber, A. Jusufi, C. N. Likos, H. Löwen, and M. Watzlawek, "Triplet interactions in star polymer solutions," Eur. Phys. J. E **2**, 311–318 (2000).

[20] M. Dijkstra, R. van Roij, R. Roth, and A. Fortini, "Effect of many-body interactions on the bulk and interfacial phase behavior of a model colloid-polymer mixture," Phys. Rev. E **73**, 041404 (2006).

[21] M. Dijkstra, J. M. Brader, and R. Evans, "Phase behaviour and structure of model colloid-polymer mixtures," J. Phys.: Condens. Matter **11**, 10079 (1999).

[22] P. Schapotschnikow and T. J. H. Vlugt, "Understanding interactions between capped nanocrystals: Three-body and chain packing effects," J. Chem. Phys. **131**, 124705 (2009).

[23] G. Bauer, N. Gribova, A. Lange, C. Holm, and J. Gross, "Three-body effects in triplets of capped gold nanocrystals," Mol. Phys. **115**, 1031–1040 (2017).

[24] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Physics-inspired structural representations for molecules and materials," Chem. Rev. **121**, 9759–9815 (2021).

[25] E. Boattini, N. Bezem, S. N. Punnathanam, F. Smallenburg, and L. Filion, "Modeling of many-body interactions between elastic spheres through symmetry functions," J. Chem. Phys. **153**, 064902 (2020).

[26] G. Campos-Villalobos, E. Boattini, L. Filion, and M. Dijkstra, "Machine learning many-body potentials for colloidal systems," J. Chem. Phys. **155**, 174902 (2021).

[27] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys. **134**, 074106 (2011).

[28] M. Harrington, A. J. Liu, and D. J. Durian, "Machine learning characterization of structural defects in amorphous packings of dimers and ellipses," Phys. Rev. E **99**, 022903 (2019).

[29] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).

[30] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).

[31] A. Singraber, J. Behler, and C. Dellago, "Library-based *LAMMPS* implementation of high-dimensional neural network potentials," J. Chem. Theory Comput. **15**, 1827–1840 (2019).

[32] B. J. Berne and P. Pechukas, "Gaussian model potentials for molecular interactions," J. Chem. Phys. **56**, 4213–4216 (1972).

[33] A fast algorithm to calculate $d_{m,ij}$ for two bodies of spherocylindrical symmetry has been described by Vega and Lago.[54]

[34] J. G. Gay and B. J. Berne, "Modification of the overlap potential to mimic a linear site–site potential," J. Chem. Phys. **74**, 3316–3319 (1981).

[35] P. Bolhuis and D. Frenkel, "Tracing the phase boundaries of hard spherocylinders," J. Chem. Phys. **106**, 666–687 (1997).

[36] B. Martínez-haya, L. F. Rull, A. Cuetos, and S. Lago, "Gibbs ensemble simulation of the vapour-liquid equilibrium of square well spherocylinders," Mol. Phys. **99**, 509–516 (2001).

[37] G. Campos-Villalobos, M. Dijkstra, and A. Patti, "Nonconventional phases of colloidal nanorods with a soft corona," Phys. Rev. Lett. **126**, 158001 (2021).

[38] S. Asakura and F. Oosawa, "Interaction between particles suspended in solutions of macromolecules," J. Polym. Sci. **33**, 183–192 (1958).

[39] A. Vrij, "Polymers at interfaces and the interactions in colloidal dispersions," Pure Appl. Chem. **48**, 471–483 (1976).

[40] M. Dijkstra and R. van Roij, "Entropic wetting and many-body induced layering in a model colloid-polymer mixture," Phys. Rev. Lett. **89**, 208303 (2002).

[41] A. P. Gast, C. K. Hall, and W. B. Russel, "Polymer-induced phase separations in nonaqueous colloidal suspensions," J. Colloid Interface Sci. **96**, 251–267 (1983).

[42] S. V. Savenko and M. Dijkstra, "Phase behavior of a suspension of colloidal hard rods and nonadsorbing polymer," J. Chem. Phys. **124**, 234902 (2006).

[43] S. Jungblut, R. Tuinier, K. Binder, and T. Schilling, "Depletion induced isotropic-isotropic phase separation in suspensions of rod-like colloids," J. Chem. Phys. **127**, 244909 (2007).

[44] B. Pansu and J.-F. Sadoc, "Metallurgy of soft spheres with hard core: From BCC to Frank-Kasper phases," Eur. Phys. J. E **40**, 102 (2017).

[45] C. Vega, C. McBride, and L. G. MacDowell, "Liquid crystal phase formation for the linear tangent hard sphere model from Monte Carlo simulations," J. Chem. Phys. **115**, 4203–4211 (2001).

[46] D. Monego, T. Kister, N. Kirkwood, P. Mulvaney, A. Widmer-Cooper, and T. Kraus, "Colloidal stability of apolar nanoparticles: Role of ligand length," Langmuir **34**, 12982–12989 (2018).

[47] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. De Vries, "The MARTINI force field: Coarse grained model for biomolecular simulations," J. Phys. Chem. B **111**, 7812–7824 (2007).

[48] Z. Fan and M. Grünwald, "Orientational order in self-assembled nanocrystal superlattices," J. Am. Chem. Soc. **141**, 1980–1988 (2019).

[49] T. Kister, D. Monego, P. Mulvaney, A. Widmer-Cooper, and T. Kraus, "Colloidal stability of apolar nanoparticles: The role of particle size and ligand shell structure," ACS Nano **12**, 5969–5977 (2018).

[50] A. Widmer-Cooper and P. Geissler, "Orientational ordering of passivating ligands on CdS nanorods in solution generates strong rod–rod interactions," Nano Lett. **14**, 57–65 (2014).

[51] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen *et al.*, "LAMMPS—A flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," Comput. Phys. Commun. **271**, 108171 (2022).

[52] G. Ciccotti, M. Ferrario, J. T. Hynes, and R. Kapral, "Constrained molecular dynamics and the mean potential for an ion pair in a polar solvent," Chem. Phys. **129**, 241–251 (1989).

[53] C. K. Lee, C. C. Hua, and S. A. Chen, "An ellipsoid-chain model for conjugated polymer solutions," J. Chem. Phys. **136**, 084901 (2012).

[54] C. Vega and S. Lago, "A fast algorithm to evaluate the shortest distance between rods," Computers & Chemistry **18(1)**, 55–59 (1994).