npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01405-4

Machine-learned coarse-grained potentials for particles with anisotropic shapes and interactions

Check for updates

Gerardo Campos-Villalobos[1], Rodolfo Subert[1], Giuliana Giunta[2] & Marjolein Dijkstra[1,3]

Computational investigations of biological and soft-matter systems governed by strongly anisotropic interactions typically require resource-demanding methods such as atomistic simulations. However, these techniques frequently prove to be prohibitively expensive for accessing the long-time and large-length scales inherent to such systems. Conversely, coarse-grained models offer a computationally efficient alternative. Nonetheless, models of this type have seldom been developed to accurately represent anisotropic or directional interactions. In this work, we introduce a straightforward bottom-up, data-driven approach for constructing single-site coarse-grained potentials suitable for particles with arbitrary shapes and highly directional interactions. Our method for constructing these coarse-grained potentials relies on particle-centered descriptors of local structure that effectively encode dependencies on rotational degrees of freedom in the interactions. By using these descriptors as regressors in a linear model and employing a simple feature selection scheme, we construct single-site coarse-grained potentials for particles with anisotropic interactions, including surface-patterned particles and colloidal superballs in the presence of non-adsorbing polymers. We validate the efficacy of our models by accurately capturing the intricacies of the potential-energy surfaces from the underlying fine-grained models. Additionally, we demonstrate that this simple approach can accurately represent the contact function (shape) of non-spherical particles, which may be leveraged to construct continuous potentials suitable for large-scale simulations.

In molecular systems, the interactions that determine the equilibrium structural, thermodynamic and dynamic properties are generally a combination of non-covalent forces, including long-range electrostatic and short-range dispersion forces, which are typically well described by spherically-symmetric atom-atom potentials. At the meso- or microscopic scale, weaker interactions between particles such as colloids, nanoparticles, and macromolecules, which involve electromagnetic forces and entropic interactions, become relevant. Moreover, many of these systems are governed by effective interactions that are anisotropic in nature, i.e. they depend on both distance and orientation[1–4]. Non-spherical particle shapes represent an exemplary source of this type of interaction[1,5–7], which in turn may give rise to novel and fascinating behavior that distinguishes these systems from those composed of simple spherical particles with centrosymmetric isotropic interactions. For instance, whereas hard spheres can only form isotropic fluid and crystalline solid phases, numerous non-spherical particles can give rise to so-called mesophases, exhibiting translational and orientational symmetries that lie between those of isotropic fluids and crystalline materials[8]. In recent decades, significant progress has been made in the fields of chemistry and physics at the nanometer scale, leading to an astonishing level of maturity in colloid synthesis[9–13]. Consequently, a wide array of particles with different shapes, compositions, patterns, and functionalities is now readily available[14–24]. These systems are particularly gaining relevance in bottom-up self-assembly approaches for creating superstructures with tailored properties[6].

The emergence of anisotropic interparticle interactions is not exclusively due to the non-spherical shape of particles; rather, they can stem from various other sources such as induced or embedded dipoles[25] and chemically or physically patterned surfaces. Particles with the latter characteristics include the so-called Janus colloids and are generally referred to as patchy particles[26,27]. Additionally, there is a growing interest in the anisotropic, effective colloidal interactions arising from the orientational elastic energy of anisotropic host fluids[28–31], as well as those exhibited by topological solitons,

[1]Soft Condensed Matter & Biophysics, Debye Institute for Nanomaterials Science, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands. [2]BASF SE, Carl-Bosch-Strasse 38, 67056 Ludwigshafen am Rhein, Germany. [3]International Institute for Sustainability with Knotted Chiral Meta Matter (SKCM2), Hiroshima University, 2-313 Kagamiyama, Higashi-Hiroshima, Hiroshima, 739-8527, Japan. e-mail: g.d.j.camposvillalobos@uu.nl; m.dijkstra@uu.nl

which are often described as quasiparticles[32]. The latter type of interactions include the cohesion mediated by out-of-equilibrium dipoles in baby skyrmions[33], and the directional interactions between three-dimensional knots emerging in helical liquid-crystal fields[34].

While significant progress has been made in developing accurate atom-atom potentials (force-fields) over the last decades, as well as bottom-up coarse-graining strategies that represent small clusters of heavy atoms in large molecules like polymers as effective beads interacting via isotropic potentials or as small ellipsoids[35,36], only a few computational approaches have been developed for modeling whole molecules or larger-sized (colloidal) particles as single sites interacting with anisotropic interactions. An interesting approach, based on anisotropic force-matching, has been recently introduced by Nguyen and Huang[37] for parameterizing coarse-grained molecular models consisting of anisotropic building blocks. Developing such approaches is essential to facilitate the study of (chemically) specific systems while overcoming spatial and temporal constraints. Nevertheless, due to the inherent difficulties in this task, particularly in accurately representing the orientational dependence of interparticle interactions mathematically[38], much of the coarse-grained (CG) modeling of these systems at the single-site level has relied on the use of prototype anisotropic pair potentials, which implicitly assign a simple geometric shape to each particle. Notable examples of this class of potentials include the Gay-Berne pair potential[39] and its variants[40–43], as well as some generalizations of conventional spherically-symmetric attractive/repulsive potentials[44,45]. These potentials have enabled large-scale simulations of liquid crystals of oblate or prolate mesogens, but they have mainly been restricted to representing generic systems. When dealing with particles of arbitrary aspherical shapes, the situation is even less satisfactory, as one is typically forced to model solely their hard-core character by relying on algorithms based on collision detection methods[8], thereby limiting their applicability when exploring the role of attractive and repulsive forces in shape-dependent processes like self-assembly, packing, and transport. Recently, however, Lee and Arya introduced a model specifically for the Van der Waals interaction energy of diverse faceted particles, including nanocubes, triangular prisms, faceted rods, and square pyramids[46]. While this model represents significant progress, its applicability to faceted particles with more than six facets or rounded edges has not yet been tested. Within discrete element models (DEM), widely used for modeling granular materials[47], one can incorporate cohesive forces, but again, they are confined to describing generic short-range attractions between bare or uniformly coated nanoparticles or colloidal particles of aspherical shape[2]. Similarly, for generic spherical Janus and patchy particles, a common approach involves representing them as hard-core objects with a discontinuous potential that is sensitive to the orientation of the patches[45–47], as in the Kern-Frenkel model[48,49], or embedding additional attractive virtual sites within larger isotropic beads[50].

By aiming to capture only essential physical features while minimizing computational complexity, the aforementioned CG computational approaches have shed light on numerous aspects of many-body systems composed of model anisotropic particles. However, when the interest is to represent the intricacies and details of specific systems, one typically has to resort to slightly more fine-grained (FG) representations. One such approach is based on surface tessellations, where a large number of small spherical beads–already coarse-grained objects–are distributed on the surface of each particle. The high-frequency motion of these constituent beads is then effectively constrained by treating the composite as a rigid object, and the interaction between two large bodies is computed as a pairwise summation over the interactions between individual beads, one from each body. Naturally, this enables the simultaneous and explicit description of shape anisotropies and directional interactions stemming from diverse origins, such as the presence of heterogeneities on the surface of patchy colloidal particles. However, due to its inherent computational cost, this FG approach is only rarely adopted[48,49].

Following the success of bottom-up coarse-graining strategies applied in the study of soft-matter systems, there is a desire to devise a general platform for the development of efficient and accurate single-site CG potentials capable of capturing key information on anisotropy in shape and interactions found in the corresponding FG models. As mentioned earlier, the core challenge lies in finding a way to represent orientational dependence in a compact and physically interpretable manner[38]. With rapid advancements in machine learning (ML) and data-driven methods used for representing potential energy surfaces, new opportunities are emerging to accomplish this task. In the case of atomistic simulations, these techniques now enable fast simulations at a classical level with forces and energies that carry the accuracy of electronic structure calculations[50]. Furthermore, these approaches have been extended to construct CG two- and many-body potentials for colloidal particles with spherical shape and isotropic interactions[51–54], as well as for uniaxial particles with cylindrical symmetry, i.e. rod-like particles[55]. Very recently, deep learning has been leveraged to accelerate Molecular Dynamics simulations of rigid bodies made of smaller composite beads[56] and to coarse-grain small rigid molecules by force and torque matching[57]. In the majority of these data-driven approaches, the main idea is to represent the total CG (potential or free) energy Φ of an N-body system as a sum of per-particle contributions Φ_I. Each individual contribution to the potential is, in itself, a function of a set of N_s functions centered on the particle, i.e. $\Phi_I = \Phi_I(\{G_1(I), \ldots, G_{N_s}(I)\})$. These functions $\{G\}$ are rotational invariants that describe the local environments of the particles in relation to their neighbors within a cutoff sphere[58]. The current limitation lies in the fact that the available structural descriptors of local particle environments are primarily designed for either spherically symmetric sites[59,60], or for elongated sites with an infinite-fold rotational symmetry around their long axes and a head-tail symmetry, meaning that no arrow or directionality can be assigned to each particle[55].

In this work, we introduce a straightforward bottom-up approach that leverages a suitable set of particle-centered descriptors as regressors in a simple ML regression scheme. This approach enables us to construct accurate single-site CG interaction potentials for particles with arbitrary anisotropic shape and interactions, using FG reference data. We demonstrate the robustness of our method by constructing CG potentials for surface-patterned spheroids and for colloidal superballs in the presence of small depletants. Additionally, we present an application of our method for constructing continuous pair potentials that accurately represent the shape of aspherical particles.

Results

Coarse-grained models

We consider the simple case involving a pair of anisotropic (non-linear) particles I and J, as illustrated in Fig. 1a. Without loss of generality, we assume the existence of a high-resolution fine-grained (FG) representation for which we can accurately measure the total potential Φ for a static dimer configuration. In a CG description, where each particle is treated as a rigid object, the state of a dimer configuration is characterized by the separation distance $R_{IJ} = |\boldsymbol{R}_{IJ}| = |\boldsymbol{R}_J - \boldsymbol{R}_I|$ from the origin of particle I to the origin of particle J (The origin of a particle can be conveniently taken as its center of mass.), and three angular variables: $\Omega_I = (\alpha_I, \beta_I, \gamma_I)$, describing the orientation of particle I relative to an arbitrary space-fixed axis system (XYZ), $\Omega_J = (\alpha_J, \beta_J, \gamma_J)$, representing the orientation of particle J, and $\Omega_{IJ} \equiv (\phi, \theta, 0)$, defining the direction of the vector $\boldsymbol{R}_{IJ}$. Equivalently, one can describe the orientation of a particle using the coordinates of a reference orthonormal frame mounted on the particle itself $(\hat{\boldsymbol{x}}_i(\alpha_i, \beta_i, \gamma_i), \hat{\boldsymbol{y}}_i(\alpha_i, \beta_i, \gamma_i), \hat{\boldsymbol{z}}_i(\alpha_i, \beta_i, \gamma_i))$. Here, α, β and γ represent the classic Euler angles describing the orientation of a rigid body with respect to a fixed coordinate system (see Fig. 1b). Clearly, the CG interaction potential Φ depends on the aforementioned variables, i.e. $\Phi = \Phi(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ})$[61]. We remind that while the pair potential strictly depends on 6 degrees of freedom, amounting to 3 Euler angles describing the relative orientation of particle J in the reference frame of particle I, and 3 Cartesian coordinates describing its relative position, explicitly representing the orientation of both particles with respect to the laboratory frame of reference has multiple advantages. It simplifies the mathematics[61], requires information directly accessible in simulations[38], and provides a naturally invariant expression with respect to the system symmetries[62].

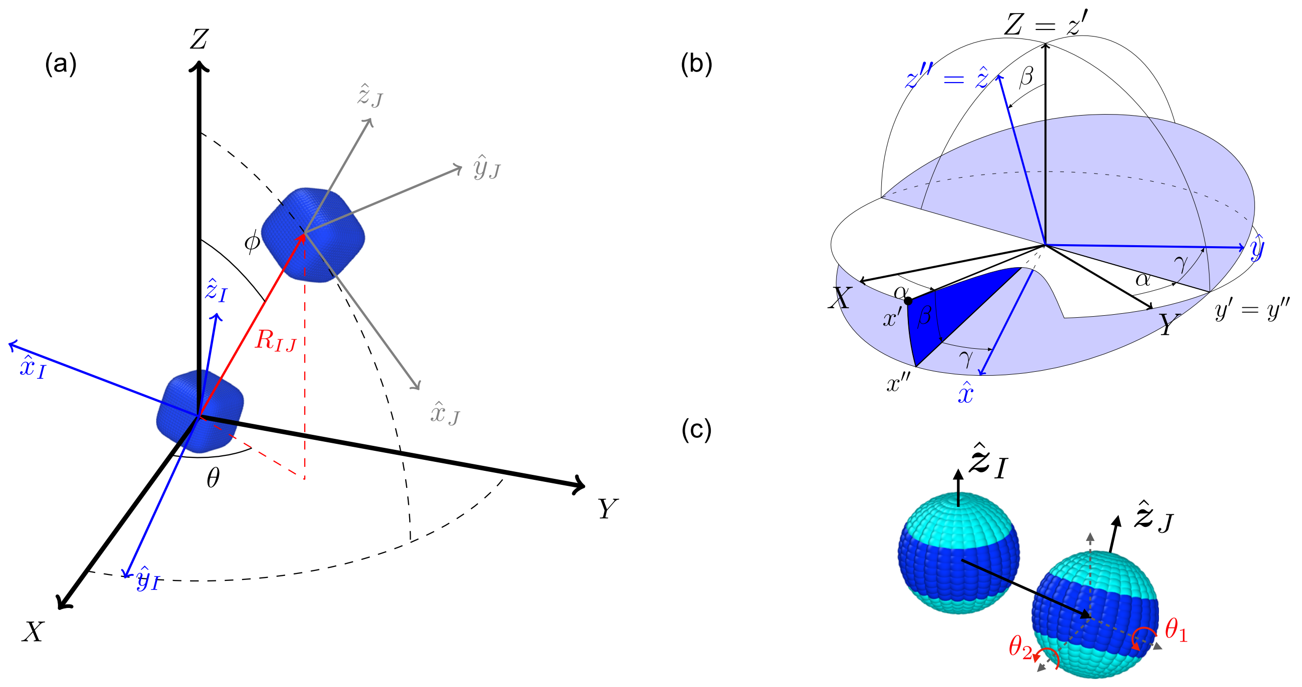

Fig. 1 | Variables used to describe pairs of particles with anisotropic shapes and interactions. a In the `rigid particle' approximation, the state of a pair of particles is defined in terms of $\Omega_{IJ} = (\theta, \phi, 0)$, which characterizes the direction of the vector $\boldsymbol{R}_{IJ} = \boldsymbol{R}_J - \boldsymbol{R}_I$ from the origin of particle I to the origin of particle J, and the orientation $\Omega_i = (\alpha_i, \beta_i, \gamma_i)$ of the body-fixed axes $(\hat{\boldsymbol{x}}_i(\alpha_i, \beta_i, \gamma_i), \hat{\boldsymbol{y}}_i(\alpha_i, \beta_i, \gamma_i), \hat{\boldsymbol{z}}_i(\alpha_i, \beta_i, \gamma_i))$ of particle $i = I, J$ relative to a space fixed set of axes (XYZ). **b** Geometrical definition of classic Euler angles (α, β, γ). **c** Schematic representation of a pair of spherical colloids with a patterned surface in the fine-grained (FG) representation. The angular variables $\theta_1, \theta_2 \in [0, 2\pi]$ are uniformly sampled to extract the potential energy landscapes shown as 3D plots in Figs. 2 and 4.

Building on the work of Blum and Torruella[62] and Stone[38] on the theory of intermolecular interactions, it is well established that at long range Φ can be expanded as

$$\Phi(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ}) = \sum_{j,l_I,l_J,k_I,k_J} f^{k_I,k_J}_{j,l_I,l_J}(R_{IJ}) S^{k_I,k_J}_{j,l_I,l_J}(\Omega_I, \Omega_J, \Omega_{IJ}), \quad (1)$$

where $f^{k_I,k_J}_{j,l_I,l_J}(R_{IJ})$ represent the associated expansion coefficients.

The scalar function $S^{k_I,k_J}_{j,l_I,l_J}(\Omega_I, \Omega_J, \Omega_{IJ})$, which depends on the orientations and relative orientation of a pair of arbitrarily shaped objects, can be expanded using a complete set of orthogonal functions known as 'S functions'[38]

$$S^{k_I,k_J}_{j,l_I,l_J}(\Omega_I, \Omega_J, \Omega_{IJ}) = i^{l_I - l_J - j} \sum_{m_I,m_J,m} \mathscr{D}^{l_I}_{m_I,k_I}(\Omega_I) \mathscr{D}^{l_J}_{m_J,k_J}(\Omega_J) \mathscr{D}^{j}_{m,0}(\Omega_{IJ}) \begin{pmatrix} l_I & l_J & j \\ m_I & m_J & m \end{pmatrix}, \quad (2)$$

with

$$\begin{pmatrix} l_I & l_J & j \\ m_I & m_J & m \end{pmatrix}, \quad (3)$$

a Wigner $3j$ symbol, $\mathscr{D}^{l}_{m,k}(\Omega)$ the Wigner D-matrix, l_I, l_J and j are non-negative integers, $-l_I \le k_I \le l_I$, $-l_J \le k_J \le l_J$, and where the summation runs from $m_I = -l_I$ to l_I, from $m_J = -l_J$ to l_J, and from $m = -j$ to j. In Eq. (2), the prefactor $i^{l_I - l_J - j}$ ensures that $S^{0,0}_{j,l_I,l_J}$ is real for all l_I, l_J and j. These functions are scalars that remain invariant under rigid rotations of the entire system. It also guarantees invariance against permutation of particle indices[38], i.e. $S^{k_I,k_J}_{j,l_I,l_J}(\Omega_I, \Omega_J, \Omega_{IJ}) = S^{k_J,k_I}_{j,l_J,l_I}(\Omega_J, \Omega_I, \Omega_{JI})$, where $\Omega_{JI} = (\pi + \phi, \pi - \theta, 0)$.

The origin and nature of the above expansion are rooted in the quantum mechanical structure of molecular orbitals as originally developed by Wigner[63,64], to which we refer the interested reader. For practical purposes, expansions like Eq. (1) for properties of particles with cylindrical symmetry have usually been truncated at the second rank level[65], where the rank of the expansion is defined as $r = j + l_I + l_J$. The expansion can be extended to arbitrary shapes, but this, of course, is accompanied by a large number of coefficients. Here, we will limit ourselves to showing how to rearrange the above expansion to be efficiently used in combination with a simple feature selection scheme in order to fit orientation-dependent CG interaction potentials.

The explicit form of the S functions in Eq. (2), with three Wigner rotation matrices, might initially seem cumbersome. However, it offers a crucial advantage by enabling significant simplification of the enumeration task without the need to manipulate nj symbols and Wigner matrices algebraically. As noted early on by Stone[38], the first step is to realize that S functions straightforwardly inherit the recursion relation of Wigner matrices

$$\mathscr{D}^{l'}_{m',k'}(\Omega) \mathscr{D}^{l''}_{m'',k''}(\Omega) = \sum_{l=|l'-l''|}^{l'+l''} (-1)^{-(m'+m'')-(k'+k'')} (2l+1) \mathscr{D}^{l}_{m'+m'',k'+k''} \begin{pmatrix} l' & l'' & l \\ m' & m'' & m \end{pmatrix} \begin{pmatrix} l' & l'' & l \\ k' & k'' & k \end{pmatrix}. \quad (4)$$

Combined with the fact that S functions form a complete basis set, this property allows us to express the product of any two functions as a linear combination

$$S^{k'_I,k'_J}_{j',l'_I,l'_J}(\Omega_I, \Omega_J, \Omega_{IJ}) S^{k''_I,k''_J}_{j'',l''_I,l''_J}(\Omega_I, \Omega_J, \Omega_{IJ}) = \sum_{l_I=|l'_I-l''_I|, l_J=|l'_J-l''_J|, j=|j'-j''|}^{l'_I+l''_I, l'_J+l''_J, j'+j''} i^{l'_I - l'_J - j' + l''_I - l''_J - j'' + l_I - l_J - j} (-1)^{k'_I + k''_I + k'_J + k''_J} (2l_I+1)(2l_J+1)(2j+1) S^{k_I,k_J}_{j,l_I,l_J}(\Omega_I, \Omega_J, \Omega_{IJ}) \begin{pmatrix} l'_I & l''_I & l_I \\ k'_I & k''_I & k_I \end{pmatrix} \begin{pmatrix} l'_J & l''_J & l_J \\ k'_J & k''_J & k_J \end{pmatrix} \begin{pmatrix} j' & j'' & j \\ 0 & 0 & 0 \end{pmatrix} \begin{Bmatrix} l'_I & l''_I & l_I \\ l'_J & l''_J & l_J \\ j' & j'' & j \end{Bmatrix}, \quad (5)$$

where the last term within braces represents a 9*j* symbol. To lighten the notation we represent a specific set of integer values as $\alpha = (j, l_I, l_J; k_I, k_J)$ or $\alpha\prime = (j', l_I', l_J'; k_I', k_J')$, respectively. So that given a set of lowest rank *S* functions we define a monomial as

$$M_S(\{a_\alpha\}) = \prod_{\alpha \in \text{ lowest rank}} S_\alpha(\Omega_I, \Omega_J, \Omega_{IJ})^{a_\alpha}, \tag{6}$$

where certain non-negative integer exponents a_α are involved. By inverting the recursive expression of Eq. (5), isolating the highest rank *S* function on the left-hand side, and recursively applying the same formula to all *S* functions on the right-hand side up to the lowest rank functions, we can express any *S* function as a polynomial of the lowest rank *S* functions

$$S_\alpha(\Omega_I, \Omega_J, \Omega_{IJ}) = \sum_{\alpha'} m_{\alpha'} M_S(\{a_{\alpha'}\}), \tag{7}$$

for certain monomials $M_S(\{a_{\alpha'}\})$ with coefficients $m_{\alpha'}$ as dictated by the recursive equation Eq. (5). Nonetheless, as it will become clear below, the explicit expressions of *S* functions in terms of polynomials will not be relevant for the procedure we propose, rather we will only be interested in generating all possible monomials of lowest rank *S* functions.

The set of lowest rank functions that allows to recursively generate any term in the expansion consists of 15 rank 2 *even* functions and 9 rank 3 *odd* functions. This implies that the number of *S* functions to account for scales exponentially with the rank *r* as 24^r. Accounting for particle point group symmetries can, in principle, reduce the number of terms in the expansion of Eq. (7). However, there is no general rule to achieve this reduction, and each term must be considered separately at every rank in the expansion, as originally shown by Steele, Blum and Torruella[62,66]. It can be constructively shown with a counterexample[67] that Eq. (5) does not necessarily preserve the symmetries of a selected group of the lowest rank functions. The only exceptions are cylindrically symmetric and mirror symmetric particles, as discussed below. For cylindrically symmetric particles, the interparticle potential depends only on the three second-rank *S* functions with $k_I = k_J = 0$

$$\begin{aligned} S^{0,0}_{0,1,1} &= \tfrac{1}{\sqrt{3}} \hat{\boldsymbol{z}}_J \cdot \hat{\boldsymbol{R}}_{IJ}, \\ S^{0,0}_{1,0,1} &= -\tfrac{1}{\sqrt{3}} \hat{\boldsymbol{z}}_I \cdot \hat{\boldsymbol{R}}_{IJ}, \\ S^{0,0}_{1,1,0} &= -\tfrac{1}{\sqrt{3}} \hat{\boldsymbol{z}}_I \cdot \hat{\boldsymbol{z}}_J, \end{aligned} \tag{8}$$

where the Wigner 3j symbols selection rule $k_I' + k_I'' + k_I = 0$ of Eq. (5) ensures that the recursion relation is closed with respect to cylindrical symmetry. This is physically justified, as the remaining 12 second-rank *S*-functions

$$\begin{aligned} S^{0,1}_{0,1,1} + S^{0,-1}_{0,1,1} &= \tfrac{2}{\sqrt{6}} \hat{\boldsymbol{y}}_J \cdot \hat{\boldsymbol{R}}_{IJ}, \\ S^{0,1}_{0,1,1} - S^{0,-1}_{0,1,1} &= \tfrac{2}{\sqrt{6}} \hat{\boldsymbol{x}}_J \cdot \hat{\boldsymbol{R}}_{IJ}, \\ S^{0,1}_{1,0,1} + S^{0,-1}_{1,0,1} &= \tfrac{2}{\sqrt{6}} \hat{\boldsymbol{y}}_I \cdot \hat{\boldsymbol{R}}_{IJ}, \\ S^{0,1}_{1,0,1} - S^{0,-1}_{1,0,1} &= \tfrac{2}{\sqrt{6}} \hat{\boldsymbol{x}}_I \cdot \hat{\boldsymbol{R}}_{IJ}, \\ S^{0,1}_{1,1,0} + S^{0,-1}_{1,1,0} &= \tfrac{2}{\sqrt{6}} \hat{\boldsymbol{y}}_I \cdot \hat{\boldsymbol{z}}_J, \\ S^{0,1}_{1,1,0} - S^{0,-1}_{1,1,0} &= \tfrac{2}{\sqrt{6}} \hat{\boldsymbol{x}}_I \cdot \hat{\boldsymbol{z}}_J, \\ S^{1,0}_{1,1,0} + S^{-1,0}_{1,1,0} &= \tfrac{2}{\sqrt{6}} \hat{\boldsymbol{z}}_I \cdot \hat{\boldsymbol{y}}_J, \\ S^{1,0}_{1,1,0} - S^{-1,0}_{1,1,0} &= \tfrac{2}{\sqrt{6}} \hat{\boldsymbol{z}}_I \cdot \hat{\boldsymbol{x}}_J, \\ S^{1,1}_{1,1,0} + S^{-1,-1}_{1,1,0} &= \tfrac{1}{\sqrt{6}} (\hat{\boldsymbol{x}}_I \cdot \hat{\boldsymbol{y}}_J + \hat{\boldsymbol{x}}_J \cdot \hat{\boldsymbol{y}}_I), \\ S^{1,1}_{1,1,0} - S^{-1,-1}_{1,1,0} &= \tfrac{1}{\sqrt{6}} (\hat{\boldsymbol{x}}_I \cdot \hat{\boldsymbol{x}}_J + \hat{\boldsymbol{y}}_I \cdot \hat{\boldsymbol{y}}_J), \\ S^{1,-1}_{1,1,0} + S^{-1,1}_{1,1,0} &= \tfrac{1}{\sqrt{6}} (\hat{\boldsymbol{x}}_I \cdot \hat{\boldsymbol{y}}_J - \hat{\boldsymbol{x}}_J \cdot \hat{\boldsymbol{y}}_I), \\ S^{-1,1}_{1,1,0} - S^{1,-1}_{1,1,0} &= \tfrac{1}{\sqrt{6}} (\hat{\boldsymbol{x}}_I \cdot \hat{\boldsymbol{x}}_J - \hat{\boldsymbol{y}}_I \cdot \hat{\boldsymbol{y}}_J). \end{aligned} \tag{9}$$

explicitly depend on either of the other axes and are sufficient when the potential (or any scalar property of interest) is mirror symmetric. This is understood because, as scalar products, they are inherently inversion symmetric by definition. The closure of the recursion relation with respect to the entire set of second-rank *S*-functions requires $l' + l_I'' + l_I$ to be even. In cases where space inversion symmetry is broken, we must also consider the rank 3 *S*-functions, which include the following 9 odd functions

$$\begin{aligned} S^{0,0}_{1,1,1} &= \tfrac{1}{\sqrt{6}} \hat{\boldsymbol{z}}_I \cdot (\hat{\boldsymbol{z}}_J \times \hat{\boldsymbol{R}}_{IJ}), \\ S^{1,0}_{1,1,1} - S^{-1,0}_{1,1,1} &= \tfrac{1}{2\sqrt{3}} \hat{\boldsymbol{x}}_I \cdot (\hat{\boldsymbol{z}}_J \times \hat{\boldsymbol{R}}_{IJ}), \\ S^{1,0}_{1,1,1} + S^{-1,0}_{1,1,1} &= \tfrac{1}{2\sqrt{3}} \hat{\boldsymbol{y}}_I \cdot (\hat{\boldsymbol{z}}_J \times \hat{\boldsymbol{R}}_{IJ}), \\ S^{0,1}_{1,1,1} - S^{0,-1}_{1,1,1} &= \tfrac{1}{2\sqrt{3}} \hat{\boldsymbol{z}}_I \cdot (\hat{\boldsymbol{x}}_J \times \hat{\boldsymbol{R}}_{IJ}), \\ S^{0,1}_{1,1,1} + S^{0,-1}_{1,1,1} &= \tfrac{1}{2\sqrt{3}} \hat{\boldsymbol{z}}_I \cdot (\hat{\boldsymbol{y}}_J \times \hat{\boldsymbol{R}}_{IJ}), \\ S^{1,1}_{1,1,1} + S^{-1,-1}_{1,1,1} &= \tfrac{1}{2\sqrt{3}} (\hat{\boldsymbol{x}}_I \cdot (\hat{\boldsymbol{y}}_J \times \hat{\boldsymbol{R}}_{IJ}) + \hat{\boldsymbol{y}}_I \cdot (\hat{\boldsymbol{x}}_J \times \hat{\boldsymbol{R}}_{IJ})), \\ S^{1,1}_{1,1,1} - S^{-1,-1}_{1,1,1} &= \tfrac{1}{2\sqrt{3}} (\hat{\boldsymbol{x}}_I \cdot (\hat{\boldsymbol{x}}_J \times \hat{\boldsymbol{R}}_{IJ}) + \hat{\boldsymbol{y}}_I \cdot (\hat{\boldsymbol{y}}_J \times \hat{\boldsymbol{R}}_{IJ})), \\ S^{1,-1}_{1,1,1} + S^{-1,1}_{1,1,1} &= \tfrac{1}{2\sqrt{3}} (\hat{\boldsymbol{x}}_I \cdot (\hat{\boldsymbol{y}}_J \times \hat{\boldsymbol{R}}_{IJ}) - \hat{\boldsymbol{y}}_I \cdot (\hat{\boldsymbol{x}}_J \times \hat{\boldsymbol{R}}_{IJ})), \\ S^{1-1}_{1,1,1} - S^{-1,1}_{1,1,1} &= \tfrac{1}{2\sqrt{3}} (\hat{\boldsymbol{x}}_I \cdot (\hat{\boldsymbol{x}}_J \times \hat{\boldsymbol{R}}_{IJ}) - \hat{\boldsymbol{y}}_I \cdot (\hat{\boldsymbol{y}}_J \times \hat{\boldsymbol{R}}_{IJ})). \end{aligned} \tag{10}$$

The cross products in these expressions explicitly demonstrate the broken chiral symmetry.

Introducing Eq. (7) into Eq. (1), grouping the double sum and adsorbing the monomial coefficients m_α in the coefficients of Eq. (1) leads to

$$\begin{aligned} \Phi(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ}) &= \textstyle\sum_\alpha f_\alpha(R_{IJ}) \left[\sum_\alpha m_\alpha M_S(\{a_\alpha\})\right] \\ &= \textstyle\sum_\alpha \tilde{f}_\alpha(R_{IJ}) M_S(\{a_\alpha\}), \end{aligned} \tag{11}$$

which is the expression that we will use in practice by using either of the three sets of functions specified above.

Yet, our aim is to construct accurate single-site CG interaction potentials for particles with arbitrary anisotropic shape and interactions using a ML approach. Given a large dataset of sample configurations and associated values of the total potential $\hat{\Phi}$ measured at some FG level, a CG potential Φ at the single-site level can be constructed in terms of descriptors $\{G\}$. Following the work of Behler and Parrinello[59] and for the sake of generality, we adopt the assumption that Φ results from the individual particle contributions, $\Phi = \Phi_I + \Phi_J$, which in the case of an isolated pair, are identical. While non-linear regression schemes are increasingly being used to learn from a reference database, here, we conveniently assume a simple linear relationship between the target function (individual particle contribution to the total energy) and its local environment: $\Phi_I = \sum_{k=1}^{N_s} \omega_k G_k(I)$, with ω_k representing the weight or linear coefficient of the *k*-th descriptor. From Eq. (11), we clearly observe that suitable particle-centered descriptors $\{G\}$ for a machine-learning approach to characterize the angular dependence of the CG interaction potential Φ are polynomials of *S* functions. Using these polynomials offers several benefits. Once the initial set of *S*-functions is fixed, they are computationally straightforward to enumerate. Moreover, they are easy to use in Monte Carlo (MC) simulations and are simple to differentiate[68,69]. The forces and torques expressed in terms of these polynomials, which are scalar and cross products of Cartesian vectors, remain as such, which, in combination with the chain rule, provides a significant advantage for efficient evaluation within Molecular Dynamics simulations. While we describe the radial dependence in terms of functions of the separation distance between the centers of two particles, $\Lambda(R_{IJ})$, (for example, a Gaussian function), and a cutoff function $f_c(R_{IJ})$ that gradually approaches zero, both in value and in slope, as it reaches the cutoff distance

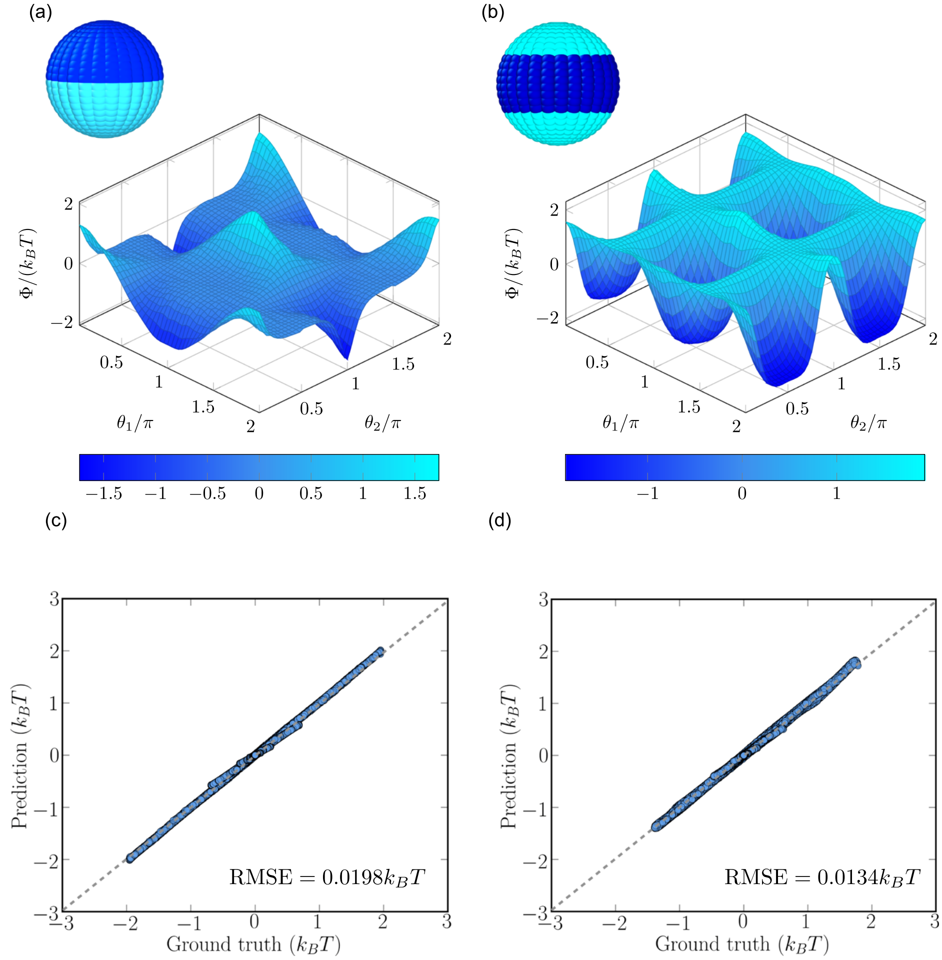

Fig. 2 | Potential energy landscapes of surface-patterned spherical particles and parity plots comparing FG and CG models. Orientational dependence of the pair interaction energy between charged Janus colloids (JCs) (**a**) and inverse patchy colloids (IPCs) (**b**) with $q(s_{i\in I}) = q(s_{j\in J}) = q = 20\sigma^{-2}$ and $\kappa\sigma = 10$ as measured in the FG models. The two-body potential is plotted as a function of the rotation angles θ_1 and θ_2 (see Fig. 1), with the separation distance between particles fixed at $R_{IJ}/\sigma = 1.015$. Parity plots comparing the energies in training and test configurations (see text for details) as obtained using the FG models (Ground truth) with those predicted by the ML-CG models (Prediction) for JCs (**c**) and IPCs (**d**).

R_c. The descriptor centered on particle I is then defined as

$$G_k(I) = \Lambda(R_{IJ}) f_c(R_{IJ}) M_S(\{a_\alpha\}), \tag{12}$$

where $k = \alpha$, a_α represents a specific integer value set. Here, we consider a cutoff function of the form

$$f_l(R_{IJ}) = \begin{cases} \tanh^3(1 - R_{IJ}/R_c) & \text{for } R_{IJ} \le R_c \\ 0 & \text{for } R_{IJ} > R_c. \end{cases} \tag{13}$$

We wish to make the following remarks. Firstly, we note that removing the angular term $M_S(\{a_\alpha\})$ in Eq. (12), and choosing specific $\Lambda(R_{IJ})$-functions lead to the well-known radial Behler and Parrinello Symmetry Functions (for a particle and a single neighbor)[58,59].

Secondly, although initially developed within the context of intermolecular potentials[38], the scalar S functions have been employed across various frameworks for constructing potential models for non-spherical particles[65,70]. Truncated expansions in rotational invariants akin to S functions have also been used to capture the orientational dependence in pair interactions of charged Janus spheres[48], highlighting their versatility in representing intricate functions within the $(\Omega_I, \Omega_J, \Omega_{IJ})$ space. Despite a few isolated attempts, a general systematic approach for constructing interaction potentials for particles with anisotropic interactions in a bottom-up fashion is currently lacking.

In our bottom-up coarse-graining approach, the inital step involves generating a sizable yet manageable pool of $\mathcal{N}$ candidate descriptors $\{G\}$. This is accomplished by computing various functions with different parameter sets for each sample within the reference training dataset. Each sample corresponds to a dimer configuration along with the associated potential values measured at a FG resolution ($\hat{\Phi}$). Subsequently, employing the feature selection scheme described in ref. 51, we identify an optimal subset comprising $N_s < \mathcal{N}$ functions that offer the most effective representation of the target function when combined linearly (see Methods for an extended description). Throughout the remainder of this article, we demonstrate the efficacy of this straightforward approach in constructing accurate CG models for particles with highly anisotropic interactions.

Potential energy surfaces

Surface-patterned particles. We start by examining the anisotropic interaction energy between heterogeneously charged spherical particles. In particular, we focus on the so-called inverse patchy colloids (IPCs), and on charged Janus colloids (JCs). Previous studies have developed single-site CG models based on Debye–Hückel theory for spherical IPCs[71], while for JCs, CG models based on expansions on rotational invariants have been proposed[48]. In our approach, we start with a rigid-body FG representation where the surface of a hard sphere with size σ is uniformly decorated with sites carrying either positive or negative charge density. These arrangements reflect the characteristic patterns of the corresponding particles (see Fig. 2). Consequently, a single particle at the FG level effectively comprises $n = 626$ sites with positions $\boldsymbol{r}^n = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n\}$, where $\boldsymbol{r}_i \in R^3$ with $i = 1, \cdots, n$. In contrast, the CG representation we aim to derive only includes the center of mass position of the particle $\boldsymbol{R} \in R^3$ and its associated orientation. We implicitly assume a decoupling of the hard-core and anisotropic interaction terms, allowing us to express the total

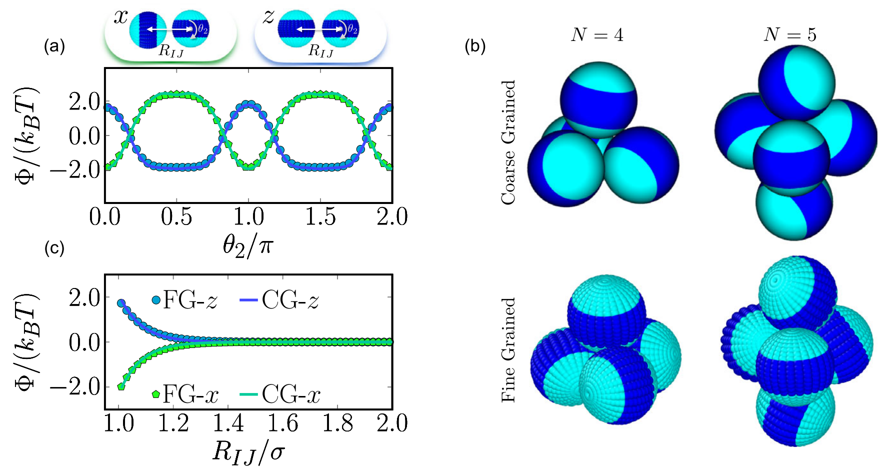

Fig. 3 | Validation of the CG model for the anisotropic interactions of spherical IPCs. Interaction energy between two inverse patchy colloids (IPCs) with $q(\boldsymbol{s}_{i\in I}) = q(\boldsymbol{s}_{j\in J}) = q = 20\sigma^{-2}$ and $\kappa\sigma = 10$. **a** The two-body potential $\Phi/(k_BT)$ as a function of the rotation angle θ_2 with the left particle fixed and its orientation vector aligned either in the x direction (green) or in the z direction (blue). The separation distance between particles is set at $R_{IJ}/\sigma = 1.015$. **b** Typical low-energy configurations of clusters of $N = 4$ and $N = 5$ IPCs as obtained from Monte Carlo simulations of the FG and ML-CG models. **c** The two-body potential $\Phi/(k_BT)$ as a function of the separation distance R_{IJ} between the two particles. Filled symbols correspond to the values obtained in the FG model, while lines represent the values predicted by the CG potential.

pair potential for particles I and J as follows

$$\Phi^{\text{total}}(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ}) = \Phi^{\text{hc}}(R_{IJ}) + \Phi(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ}). \tag{14}$$

Since $\Phi^{\text{hc}}(R_{IJ})$ represents a simple hard-sphere potential, our focus is solely on crafting a CG representation of the anisotropic term $\Phi(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ})$, which contains information about the orientation-dependent screened electrostatic interactions. Given their architecture, both IPCs and JCs can be treated as uniaxial objects with orientation vectors $\hat{\boldsymbol{z}}$ (in the patch-to-patch direction, see Fig. 1c). IPCs exhibit an additional symmetry of $\hat{\boldsymbol{z}} = -\hat{\boldsymbol{z}}$. To evaluate the interaction energy in the FG models, we follow the approach of ref. 48, where each pair of surface points on different particles interacts via a screened electrostatic potential

$$\phi_q(r_{ij}) = \frac{q_{i\in I}q_{j\in J}}{4\pi\varepsilon r_{ij}}\exp(-\kappa r_{ij}), \tag{15}$$

where ε is the dielectric constant, $\kappa\sigma = 10$ represents the screening (Debye) length for the interaction between charges $q_{i\in I}$ and $q_{j\in J}$ located at $\boldsymbol{s}_{i\in I}$ and $\boldsymbol{s}_{j\in J}$ on the surfaces of the two particles, respectively, with a separation of $r_{ij} = |\boldsymbol{R}_{IJ} + \boldsymbol{s}_{i\in I} - \boldsymbol{s}_{j\in J}|$. By representing the point charges by local densities $q(\boldsymbol{s}_{i\in I})$ and $q(\boldsymbol{s}_{j\in J})$, respectively, one can compute the total interaction energy for a given configuration $(\hat{\boldsymbol{z}}_I, \hat{\boldsymbol{z}}_J, \boldsymbol{R}_{IJ})$ by integrating over both particle surfaces, $\mathcal{S}_{I,J}$

$$\Phi(\hat{\boldsymbol{z}}_I, \hat{\boldsymbol{z}}_J, \boldsymbol{R}_{IJ}) = \int_{\mathcal{S}_I}\int_{\mathcal{S}_J}\frac{q(\boldsymbol{s}_{i\in I})q(\boldsymbol{s}_{j\in J})}{4\pi\varepsilon r_{ij}}\exp(-\kappa r_{ij})d\boldsymbol{s}_{i\in I}d\boldsymbol{s}_{j\in J}. \tag{16}$$

As shown in Fig. 2a, b, the pair interaction exhibits a strong dependence on the relative orientation and largely differs for the two types of particles. In these potential energy landscapes Φ, particles are maintained at a separation distance of $R_{IJ}/\sigma = 1.015$. The first particle I is kept fixed with an orientation $\hat{\boldsymbol{z}}_I = (0, 0, 1)$, while the second particle is rotated by angles θ_1 and θ_2 about the two orthogonal axes perpendicular to $\hat{\boldsymbol{z}}_I$, as pictorially illustrated in Fig. 1c.

For both IPCs and JCs, we generate reference datasets comprising a large number of dimer configurations, spanning the entire space of translational and rotational degrees of freedom, along with their corresponding energies evaluated via Eq. (16) (see Methods for details). Subsequently, we employ linear regression using the particle-centered descriptors given in Eq. (12). Due to the symmetry of the particles, we use only monomials of the three first-rank S functions, commonly employed for linear particles. These include $\sqrt{3}S^{0,0}_{1,1,0} = \hat{\boldsymbol{z}}_I \cdot \hat{\boldsymbol{R}}_{IJ}$, $\sqrt{3}S^{0,0}_{0,1,1} = \hat{\boldsymbol{z}}_J \cdot \hat{\boldsymbol{R}}_{IJ}$ and $-\sqrt{3}S^{0,0}_{1,1,0} = \hat{\boldsymbol{z}}_I \cdot \hat{\boldsymbol{z}}_J$. Furthermore, we use a cutoff value of $R_c/\sigma = 2.0$ and for the radial part of the candidate descriptors, we use an exponential function $\Lambda(R_{IJ}) = \exp(-\mu R_{IJ})$, where μ is a hyper-parameter optimized through a grid search. This choice is primarily made for convenience, as in this specific case, we anticipate that the radial dependence of the interaction should correspond with that of a screened electrostatic interaction. Indeed, in the optimal solution we find $\mu = 10\sigma^{-1}$. However, we emphasize that greater flexibility can be achieved using other functions (e.g. Gaussian distributions, as in the standard symmetry functions by Behler and Parrinello[59]).

We note that even with a small number of descriptors, $N_s = 106$ and 64 for JCs and IPCs, respectively, the CG models accurately capture the underlying energies measured in the FG models, as demonstrated by the parity plots reported in Fig. 2c, d, accompagnied by small Root Mean Square Error (RMSE) values. In these plots, the 'Ground truth' values correspond to the energies measured in all the FG samples (training + test sets), while the 'Predicted' values are obtained by evaluating the CG model constructed using simple linear regression and feature selection schemes. To better appreciate the accuracy of the CG models, we illustrate the orientational and translational dependence of the pair interaction of IPCs in Fig. 3a, c. Additionally, we perform replica exchange Monte Carlo (MC) simulations on small systems of IPCs using the total potential (Eq. (14)), which includes the anisotropic ML potential contribution Φ. We find that the low-energy self-assembled clusters of $N = 4$ and $N = 5$ particles closely match those obtained by simulations of the FG model. In particular, we find not only a clear structural resemblance (see Fig. 3b and SI for a quantitative comparison) but also very similar average energies per particle, with absolute differences of about 2–5%.

Furthermore, we explore systems characterized by anisotropic interactions stemming from a combination of non-spherical shape and a heterogeneous distribution of sites on the particle surface. In particular, we investigate prolate uniaxial ellipsoidal particles with axes $\sigma_{||} = 2\sigma_{\perp}$ and surfaces patterned similarly to the JCs and IPCs cases. These particles are referred to as ellipsoidal Janus colloids (EJCs) and ellipsoidal inverse patchy colloids (EIPCs). The architecture and shape of the particles can be appreciated from Fig. 4a, b, where examples of orientation-dependent potential energy landscapes are also presented, clearly differing from those of their spherical counterparts. These landscapes are measured similarly to the spherical colloids, but the distance is now fixed at $R_{IJ}/\sigma_{\perp} = 2.015$. For these cases, ML-CG potentials are constructed based on the three first-rank S functions, a cutoff value of $R_c/\sigma_{\perp} = 4.0$ and a radial term $\Lambda(R_{IJ}) = \exp(-\mu R_{IJ})$ (with $\mu = 10\sigma_{\perp}^{-1}$), by performing linear regression. The resulting models accurately capture the orientational and translational dependence of the interaction measured in the FG models. Parity plots demonstrating the nice agreement between both levels of representations are shown in Fig. 4c, d.

Depletion potential for colloidal superballs. We now consider the effective depletion pair interaction between non-spherical colloids

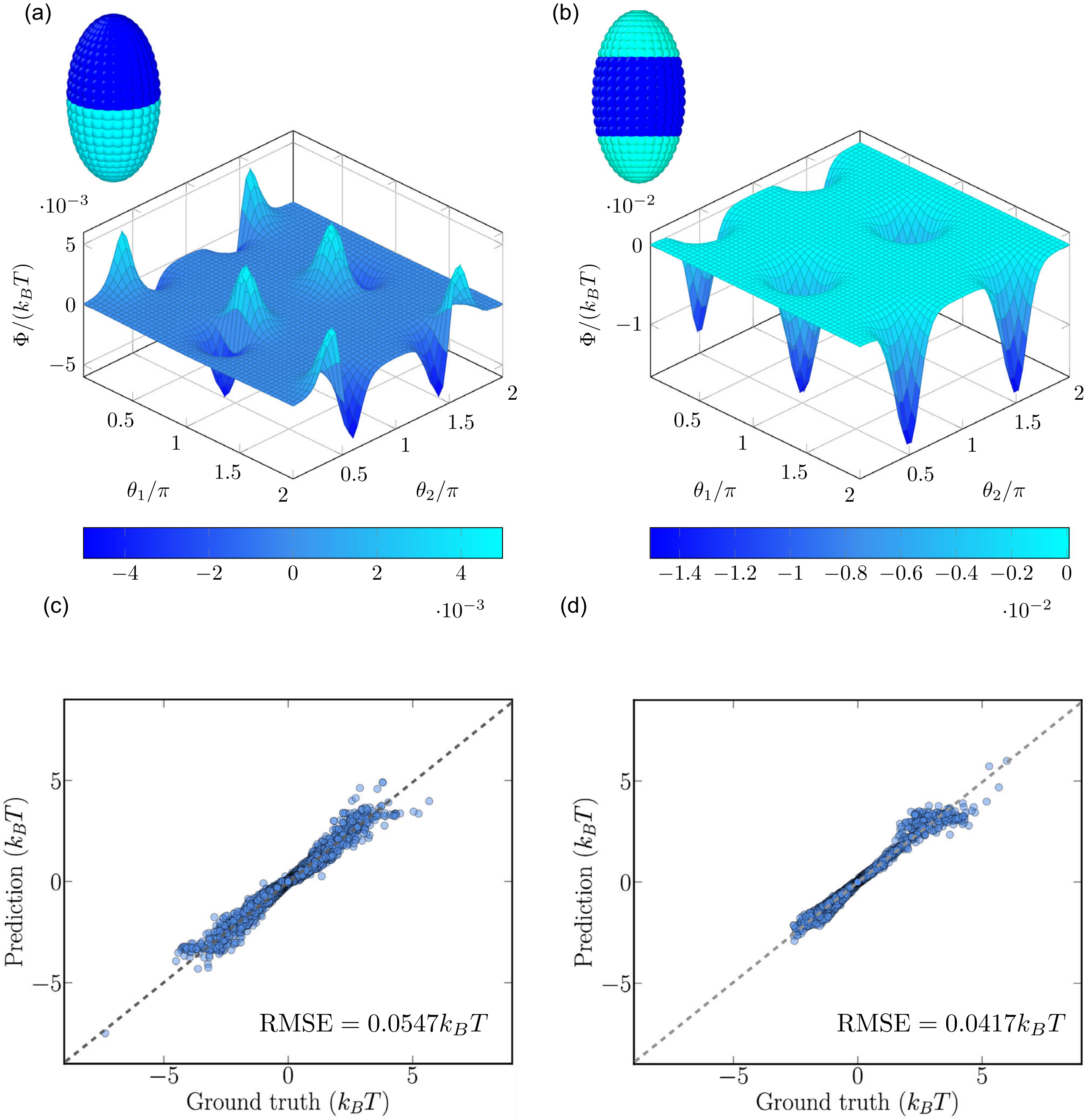

Fig. 4 | Potential energy landscapes of surface-patterned ellipsoidal particles and parity plots comparing FG and CG models. Orientational dependence of the pair interaction energy $\Phi/(k_BT)$ between ellipsoidal Janus colloids (EJCs) (**a**) and ellipsoidal inverse patchy colloids (EIPCs) (**b**) with $q(\mathbf{s}_{i\in I}) = q(\mathbf{s}_{j\in J}) = q = 20\sigma^{-2}$ and $\kappa\sigma = 10$. The FG two-body potential is plotted as a function of the rotation angles θ_1 and θ_2 (see Fig. 1), with the separation distance between particles fixed at $R_{IJ}/\sigma_\perp = 2.015$. Parity plots comparing the energies of training and test configurations (see text for details) obtained using the FG models with those predicted by the ML-CG models for EJCs (**c**) and EIPCs (**d**).

immersed in a suspension of small depletants. More specifically, the FG system consists of spherical depletants of diameter σ_d and anisotropic hard colloidal superballs with a surface defined by the equality $|x|^M + |y|^M + |z|^M = (\sigma/2)^M$, where σ represents the superball diameter (at it narrowest point) and M controls the particle shape (see Fig. 5a). Previous attempts at developing analytical CG two-body depletion potentials for non-spherical particles have been made[72,73], albeit only approximations to the *true* interaction potential have been achieved. Recently, we demonstrated that an ML-based model can effectively represent such a CG potential for uniaxial spherocylinders, even accurately capturing many-body effects[55]. Here, we demonstrate how the present approach can be used to construct depletion potentials for anisotropic superballs, whose orientation in space is described by the three orthonormal orientation vectors $(\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})$. Similar to the standard Asakura-Oosawa (AO) pair potential for spherical particles, the effective two-body (colloids only) depletion interaction is proportional to the orientation-dependent overlap volume of two depletion zones (gray shaded layers enclosing the colloids in Fig. 5a), denoted as $V_f^{(2)} = V_f^{(2)}(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ})$, with the interaction scale or *strength* determined by the depletant fugacity z_d[74,75]. Hence, the range of this

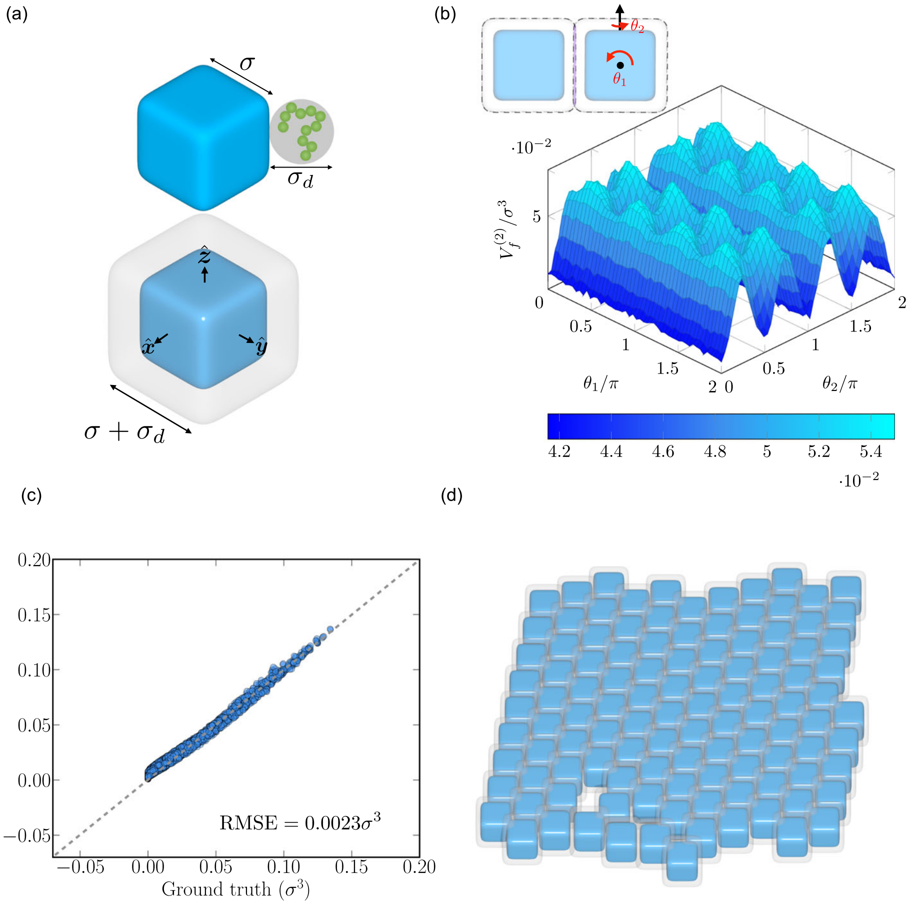

Fig. 5 | CG model for the depletion interaction of colloidal superballs. a Schematic representation of a colloidal superball with a diameter σ (blue superball) and a depletant of size σ_d (gray sphere, roughly representing a non-adsorbing polymer chain). The corresponding colloidal superball, along with its depletion zone (gray shaded superball), and orthonormal orientation vectors are shown. **b** Example of the overlap volume of two depletion zones as a function of two angular variables, as depicted in the sketch. **c** A comparison between the overlap volume sampled through MC integration (Ground truth) and the volume predicted by the ML model (Prediction). **d** A self-assembled Λ_1 cluster obtained from MC simulations of $N = 110$ hard colloidal superballs interacting via the ML-CG $\Phi(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ})$ potential. The parameters considered here are: $M = 3.5$, $Q = 0.3$ and $\eta_d^r = 5.6$.

attractive CG potential $\Phi \equiv -z_d V_f^{(2)}(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ})$ is controlled by the ratio of depletant and colloid diameters $Q = \sigma_d/\sigma$.

We select a FG model consisting of colloidal superballs with $M = 3.5$, resembling *rounded cubes* (see Fig. 5a), in the presence of depletants of size $\sigma_d = 0.3\sigma$. To efficiently gather information on the pair interaction at the FG level and given the potential's form, we simply focus on sampling $V_f^{(2)}(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ})$. We generate a large number of configurations of superball pairs with different relative orientations and separation distances, computing the overlap volume between their depletion layers through MC integration (see Methods for details). An example landscape of $V_f^{(2)}(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ})$ as a function of angular variables is shown in Fig. 5b. In this example, one particle remains fixed while the other is rotated about two orthogonal axes (see inset of Fig. 5b). To construct the ML-CG potential, we employ the first 15 S functions (non-chiral) and an exponential function for the radial dependence, akin to the case of surface patterned particles (setting $R_c/\sigma = 1.5$ and with $\mu = 4\sigma^{-1}$). As discussed previously, the number of descriptors to consider scales rapidly with the rank r. Within our implemented feature selection scheme, handling vast numbers of descriptors becomes computationally expensive. Thus, for these anisotropic particles, we limit the rank up to $r = 3$. Alternatively, we choose angular terms comprising individual 'S' functions raised to integer exponents. Interestingly, we find that with either approach, a relatively small number of descriptors suffices to accurately represent $V_f^{(2)}(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ})$. Particularly within the latter method, a model with $N_s = 30$ yields a RMSE $= 0.0023\sigma^3$. The quality of the model is clearly evident from the parity plot in Fig. 5c.

Earlier experiments and grand-canonical MC simulations involving the full binary mixture have demonstrated that for relatively large size ratios Q, sufficiently high values of M, and high depletant fugacity z_d, superballs in a monolayer spontaneously self-assemble into a unique structure termed the Λ_1 lattice[15]. This decanted lattice structure is characterized by interparticle bond angles distinct from 60° or 90° and had previously been predicted as the densest packing of superdisks[76]. For the parameters of our model, the rounded cubes should assemble into this Λ_1 lattice, provided that the pair interaction mediated by z_d is strong enough to overcome thermal fluctuations. To investigate this, we perform MC simulations of $N = 110$ hard rounded cubes interacting through the ML-CG $\Phi(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ})$ potential in the canonical ensemble. In our quasi-2D MC simulations, the rounded cubes are allowed to move and rotate in a plane. We maintain a low colloid number density $\rho_{2D} \equiv N\sigma^2/A = 0.05$, where A is the area of the simulation box.

Systems characterized by strong and short-ranged interactions can pose challenges for spontaneous self-assembly, often requiring significantly long time scales to reach equilibrium. Because of the computational efficiency of our models, we explore a range of depletant reservoir packing fractions $\eta_d^r \equiv \pi\sigma_d^3 z_d/6$ to determine the optimal assembly conditions. Starting from either isotropic configurations or perfect square lattices, we observe rapid nucleation and growth of a Λ_1 cluster. A representative configuration from a simulation illustrating an assembled Λ_1 lattice at $\eta_d^r = 5.6$ is shown in Fig. 5d. The resulting structure exhibits an average interparticle bond angle of ~53°, closely matching the values reported in ref. 15. This provides compelling evidence of the accuracy of the effective potential in capturing the depletion interaction between superballs.

Anisotropic particle shape

In the previous section, we decomposed the effective interactions into hard-core and anisotropic short-range interaction terms. Our primary focus has been on representing the latter contributions, which reflect attractive or repulsive interactions that are effective from the point of contact between pairs of particles up to a specific separation distance.

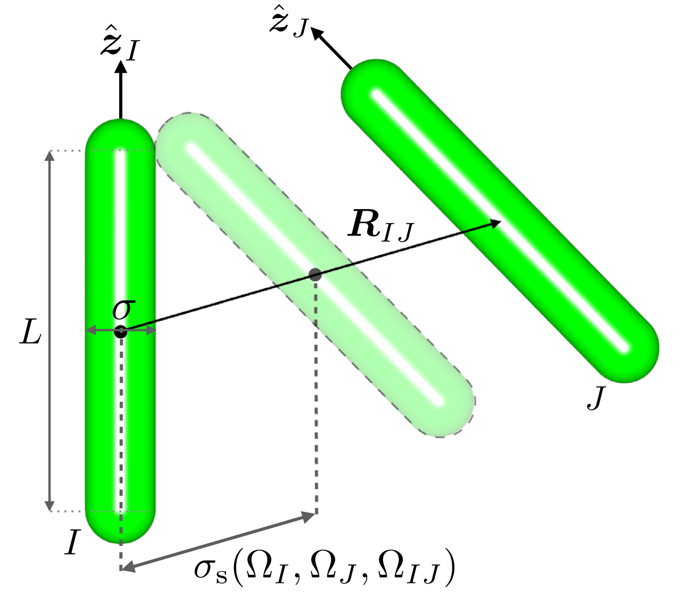

Fig. 6 | Schematic representation of the contact distance between two hard spherocylinders. The contact distance σ_s for a specific configuration of two spherocylinders with length L and diameter σ, is determined by the orientations of particle I and J (Ω_I and Ω_J, respectively), along with the orientation of the vector connecting the centroids of these particles ($\hat{\boldsymbol{R}}_{IJ}$ or Ω_{IJ}).

This strategy proves particularly advantageous when the core contribution to the interactions can be accurately modeled via simple potentials or fast numerical algorithms. In numerous colloidal systems, especially those involving non-spherical particles, the dominant excluded volume or hard-core interactions themselves can significantly impact the equilibrium properties. Typical examples include hard colloidal rod-like particles, where it is well-established that the degree of anisotropy or elongation determines the stability of 'mesophases' characterized by competing orientational and translational symmetries, even in the absence of attractive forces[77]. As mentioned earlier, when dealing with anisotropic hard-core particles, it is common practice to employ numerical algorithms that approximate the minimum distance between two non-spherical objects to determine if overlaps occur for a given configuration. An alternative, albeit less frequently employed approach, involves determining the contact distance $\sigma_s = \sigma_s(\Omega_I, \Omega_J, \Omega_{IJ})$ for a pair of particles. For a given dimer configuration characterized by Ω_I, Ω_J and Ω_{IJ}, such a quantity can be interpreted as the minimum distance between the centers of two particles where no overlap occurs. In other words, it represents the distance between particle centers at the point of contact between their surfaces. The contact distance for a pair of uniaxial spherocylinders is pictorially represented in Fig. 6. If this contact distance is known, it is theoretically feasible to construct a CG single-site, purely-repulsive continuous interaction potential[40,65,70,78].

Since the contact distance for a pair of non-spherical particles I and J depends solely on the orientations Ω_I, Ω_J and Ω_{IJ}, we can employ a special case of our method. In this situation, we set the radial terms $\Lambda(R_{IJ})$ and $f_c(R_{IJ})$ in Eq. (12) to unity and utilize an expansion $\sigma_s(\Omega_I, \Omega_J, \Omega_{IJ}) = \Phi_I = \sum_k^{N_s} \omega_k G_k(I)$, where the coefficients are determined by matching the model to numerically determined data.

For simplicity, we focus on representing the contact distance of hard spherocylinders (HSCs) with a length-to-diameter ratio of L/σ = 4. We systematically generate configurations spanning the entire (Ω_I, Ω_J, Ω_{IJ}) space for pairs of HSCs, and numerically compute the corresponding σ_s. We employ the Vega-Lago algorithm[79] to accurately determine the contact distance for each configuration. These numerically determined values constitute our dataset, which we will refer to as the FG values. Subsequently, we apply the fitting procedure to this dataset as described above.

We find that a model with as few as N_s = 64 descriptors provides an accurate representation of the contact function, resulting in an RMSE = 0.0129σ, which is less than 1% of the range of values observed in the numerically determined data. In Fig. 7a, we present a parity plot illustrating the agreement between numerically determined and predicted values of σ_s for this system. Furthermore, as shown in the top panel of Fig. 7b, apart from some minor surface 'imperfections' (e.g. small oscillations towards the HSC center and apparently sharper tips), the contour of the particle encoded in σ_s closely matches that of a true HSC with L/σ = 4. Leveraging the constructed contact function, we proceed to define an arbitrary prototype continuous and purely-repulsive potential for HSCs in the following form

$$\Phi = \epsilon \left(\frac{\sigma}{R_{IJ} - \sigma_{\mathrm{s}}(\Omega_I, \Omega_J, \Omega_{IJ}) + \sigma} \right)^{\lambda_r}, \tag{17}$$

where R_{IJ} represents the separation distance between the centers of the particles, ϵ denotes the energy scale of the interaction, and λ_r serves as an exponent controlling the steepness of the potential. To mimic a hard-core-like interaction between HSCs, we set λ_r = 50. The pair potential is displayed in the lower panel of Fig. 7b as a function of the distance between HSCs for three different relative orientations. The continuous potential curves for these base configurations effectively represent the anisotropy and dimensions of the reference HSCs with L/σ = 4.

To assess the performance of this model potential, we perform *NPT*-MC simulations of N = 896 particles to map out the bulk phase behavior of the system at k_BT/ϵ = 1.0. For comparison, we also investigate a system of true HSCs with L/σ = 4. We use $P^* \equiv P\sigma^3/(k_BT)$ as the reduced pressure and $\rho^* \equiv \rho\sigma^3$ as the reduced density. However, we note that in the CG model, the effective hard-core diameter may differ. Although we observe numerical differences in the equation of state (EOS) curves (see SI) due to the inherent softness of the CG potential, we confirm that the latter model exhibits the same phases as the system of HSCs. More specifically, we observe that both models give rise to the formation of isotropic (I) phases at low density (or pressure); approximately for $\rho^* < 0.128$ in the HSC (FG) system and for $\rho^* < 0.119$ in the CG model. A slight increase in density (or equivalently, in pressure) leads to a transition to a smectic (SM) phase with broken unidimensional translational symmetry. In the FG model, the first instance of an SM phase occurs at $\rho^* \approx 0.143$, while in the CG model, we find it at $\rho^* \approx 0.141$. For larger densities–$\rho^* > 0.163$ in the CG model and $\rho^* > 0.159$ in the FG model–a solid phase appears. The softness of the CG potential becomes evident in the ability of particles to be packed more tightly, thereby allowing access to higher density values than those achievable in the FG system. However, it is important to note that these densities should be considered approximate, as they were simply obtained from the respective EOS derived from simulations on small system sizes, where we used an arbitrary step in pressure of ΔP^* = 0.198. This may explain why we do not observe a clear uniaxial nematic state between the I and SM phases, which has been reported to be stable in the HSC system with L/σ = 4 within a rather narrow density regime[77].

Finally, to assess the robustness of our approach, we also apply it to hard superballs (HSBs) with varying exponent M determining the *roundness* of the particles and to hard pears of revolution (HPs). The adopted HSB and HP models, along with their corresponding results, are discussed in the SI. Overall, we find that the method remains valid at the expenses of requiring larger numbers of descriptors. In the SI, we also discuss how a non-linear regression scheme may be alternatively used.

Discussion

In summary, we have introduced particle-centered descriptors that effectively map static configurations of particles with arbitrary shapes and anisotropic interactions into a suitable representation, facilitating the construction of a ML model for regressing structure-property relationships. This approach is versatile and offers a solution to the challenging problem of developing accurate single-site CG potentials for particles with anisotropic shapes and interactions in a bottom-up fashion. In this work, we focused on

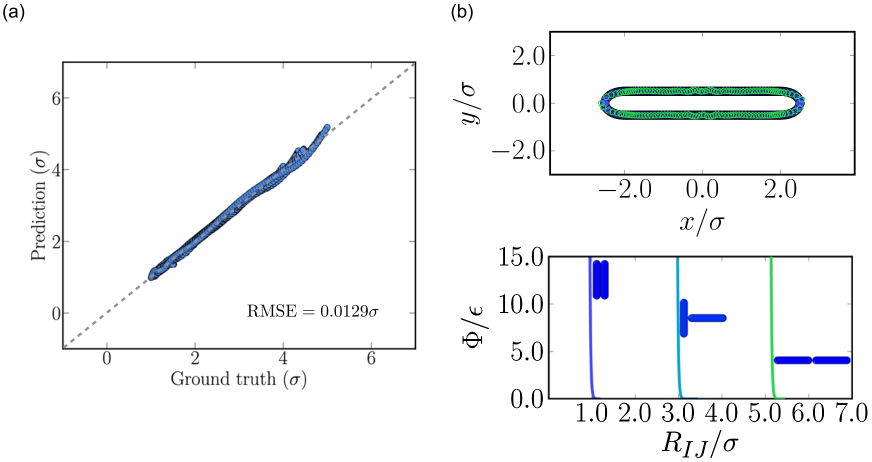

Fig. 7 | CG model for the contact function of spherocylinders and prototype continuous potential. a Comparison between the numerically-determined contact function σ_s of hard spherocylinders with a length-to-diameter ratio $L/\sigma = 4$ (Ground truth) and the corresponding prediction by the ML-CG model (Prediction). **b** Top: Contour of the particle shape described by σ_s for actual hard spherocylinders (blue circles) and the model constructed using ML (empty pentagons). Bottom: Prototype pair potential for these anisotropic particles (as defined in Eq. (17)) as a function of their separation distance R_{IJ} for three distinct relative orientations as shown in the accompanying sketches.

applying our data-driven method to precisely represent CG two-body interactions between surface-patterned spheroids and between anisotropic colloidal superballs in the presence of small depletants, as well as encoding information on the contact distance between aspherical particles (also a two-body function). Hence, the proposed representation naturally relies on so-called S functions that depend on the rotational degrees of freedom of up to two particles. If there is a need for functions involving three or more particles (similar to the angular symmetry functions of Behler and Parrinello[59]), our approach can be readily extended, as there are viable approaches for combining three or more sets of spherical tensor quantities[61].

In our selected examples, the use of simple exponentially decaying distance functions $\Lambda(R_{IJ})$ sufficed to accurately represent the separation dependence of their anisotropic pair interaction. This interaction was generally decoupled from the hard-core term (see Eq. (14)) as FG models of colloidal particles involved already some level of coarse-graining. As discussed in the SI, our application of the method to represent the orientation-dependent potential of mean force between ligand-stabilized nanocubes demonstrates that the approach remains effective for coarse-graining (quasi) atomistic models. However, in such cases the $\Lambda(R_{IJ})$ functions could be replaced by the more general radial functions such as the Behler and Parrinello symmetry functions[59] to capture more intricate radial distance dependencies, which could result in larger numbers of descriptors, both in the initial pool of candidates and in the final model. We note that a potential drawback of the present method may precisely lie in the management of large numbers of candidate descriptors $\{G\}$; particularly in the case of non-linear particles. One possible solution to mitigate this limitation could be leveraging group-theoretical arguments to reduce the number of independent parameters within the models and determining which parameters are non-zero in specific point groups[80]. However, it is important to highlight that the feature selection algorithm employed in our study effectively identifies the pertinent non-vanishing regressors that collectively best correlate with the target function.

Finally, as demonstrated throughout the article, the number of descriptors in our CG models remain relatively low, making their evaluation in MC simulations feasible. The computational cost of the developed CG potentials will not only scale with the number of descriptors, it will also depend on parameters like the cut-off radius as well as on the shape of the $\Lambda(R_{IJ})$ functions. Ultimately, as we show in the SI, the speedup achieved with the data-driven single-site potentials, relative to the parental FG model, will be determined by the complexity of the functions describing the interactions between sites in the FG model and how a single body is represented at such a resolution level (e.g., the number of sites tessellating its surface). It is important to mention that while model potentials like the one defined in Eq. (17) could serve various purposes, their computational cost may still remain relatively high due to the evaluation of the contact function. Nevertheless, considering that the function $\sigma_s(\Omega_I, \Omega_J, \Omega_{IJ})$ effectively represents the contour surface of the particles, it would be interesting to investigate whether such models could be conveniently exploited in reverse-engineering schemes, where particle shape can be optimized to obtain a target assembled structure. We anticipate that our straightforward yet accurate coarse-graining framework will facilitate the characterization, comprehension, and prediction of the structure and phase behavior of relevant anisotropic colloidal and molecular systems through direct simulations.

Methods

Construction of coarse-grained models

Single-site CG models for the interaction potential or contact function are constructed as linear combinations of an optimal number of particle-centered descriptors. In order to identify the optimal subset of descriptors, we implement the feature selection scheme of ref. 51, which we briefly summarize below.

For a given dataset, consisting of a collection of two-particle configurations and the corresponding values of the to-be-predicted quantities (e.g. energy or shape function), a training/validation split of the whole data is first adopted (typically 80/20). The first step of the method involves the creation of a large but manageable pool of $\mathcal{N}$ candidate descriptors $G_{k=1,\ldots,\mathcal{N}}$. This is accomplished by computing various functions with different parameter sets for each sample within the reference training dataset. Then, an optimal subset comprising $N_s < \mathcal{N}$ functions is selected from the pool in a step-wise fashion. The first function 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

$$c_k = \frac{\sum_j \left(\sum_I^N G_k(I)|_j - \overline{\sum_I^N G_k(I)}\right)\left(\Phi|_j - \overline{\Phi}\right)}{\sigma_{\mathrm{SD}}\left(\sum_I^N G_k(I)\right)\sigma_{\mathrm{SD}}(\Phi)}, \tag{18}$$

where $\sum_I^N G_k(I)|_j$ represents the sum of the k-th function over the N particles in configuration j and $\Phi|_j$ denotes the target variable evaluated for this configuration. Note that in the case of energy representation, $N = 2$ and $\Phi|_j = \Phi_I|_j + \Phi_J|_j$, while in the case of a contact function, $N = 1$ and $\Phi|_j = \Phi_I|_j$. In Eq. (18), $\overline{\sum_i^N G_k(I)}$ and $\overline{\Phi}$ correspond to arithmetic means over all the configurations in the data set, and $\sigma_{\mathrm{SD}}(\sum_I^N G_k(I))$ and $\sigma_{\mathrm{SD}}(\Phi)$ to their standard deviations. The next function is then selected based on the highest increase in the linear correlation between the currently selected set and the target data as determined by the coefficient of multiple correlation

$$R^2 = \mathbf{c}^T\mathbf{R}^{-1}\mathbf{c}, \tag{19}$$

where $\mathbf{c}^T = (c_1, c_2, \cdots)$ is the vector whose i-th component is given by the Pearson correlation coefficient, c_i, between the i-th function and the target data, and $\mathbf{R}$ is the correlation matrix of the current set of functions with elements $\mathcal{R}_{ij}$ representing the Pearson correlation function between the i-th and j-th functions. 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 of the target function in terms of the functions 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[51]. Maximizing the increase in the linear correlation with the target variable guarantees that only functions that add relevant information are selected[51]. This process is repeated in an iterative fashion and new functions are selected until the correlation stops increasing appreciably. This, indicates that the remaining descriptors in the pool do not add *relevant* information to the model. In turn, this constitutes a simple rule to optimize the number of selected functions as their inclusion would simply imply an unnecessary numerical overhead. In Supplementary Fig. 1, we show, as an example, the evolution of R^2 and the root mean square error over training and test sets in the process of constructing a CG model for the interaction of inverse patchy colloids.

Fine-grained data

Surface-patterned particles. The potential energy landscapes of spherical and ellipsoidal charged Janus and Inverse Patchy particles were sampled by evaluating the integrals of the site-site potentials over the particles' surfaces (Eq. (16)). The diverse configurations spanning the (R_{IJ}, $\boldsymbol{\Omega}_I$, $\boldsymbol{\Omega}_J$, $\boldsymbol{\Omega}_{IJ}$) space were generated by placing the particles at different mutual distances and then rotating the particles about two space-fixed axes by equally spaced angles in the range 0 to 2π. The separation distances were sampled uniformly from $R_{IJ}/\sigma = 1.04050$ to $R_{IJ}/\sigma = 2.0$ in the case of spherical particles, and from $R_{IJ}/\sigma_\perp = 1.010$ to $R_{IJ}/\sigma_\perp = 4.0$ in the case of ellipsoidal particles. For the latter, a simple collision detection algorithm is employed to discard configurations with overlapping particles. In building the CG models, we generated a total of 4,296,875 and 4,303,750 samples of spherical Janus and IPC particles, respectively, while 3,750,676 and 3,776,912 configurations were generated for Janus and IPC ellipsoids, respectively. For each case, 10^5 randomly selected configurations are considered for training and testing.

Overlap volume of depletion layers of colloidal superballs. The FG model of colloidal superballs consists of a hard particle surface defined by the equality $|x|^M + |y|^M + |z|^M = (\sigma/2)^M$, where σ represents the superball diameter and M controls the particle shape. The superball is enclosed by a shell of thickness equal to the size of spherical depletants of diameter σ_d; formally the Minkowski sum of a superball and a sphere and that we approximate to a superball with a different exponent M' and diameter σ'. We select a model consisting of colloidal superballs with $M = 3.5$ and $\sigma = 1$, in the presence of depletants of size $\sigma_d = 0.3\sigma$, that are well captured by a more rounded superball with $M' = 2.5$ and $\sigma' = \sigma + \sigma_d$. In order to construct a database suitable for fitting with the feature selection algorithm and efficiently gather information on the pair interaction at the FG level, we simply focus on sampling $V_f^{(2)} = V_f^{(2)}(R_{IJ}, \boldsymbol{\Omega}_I, \boldsymbol{\Omega}_J, \boldsymbol{\Omega}_{IJ})$, i.e. the volume of the overlapping depletion zones, for particle configurations with overlapping shells only. First we fix the first particle in the center and generate the orientations $\boldsymbol{\Omega}_I = (\theta_I, \phi_I)$ and $\boldsymbol{\Omega}_J = (\theta_J, \phi_J)$ on a regular grid with 8 equally spaced orientations from 0 to 2π. Next, we take 50 equally spaced positions for the second superball within the overlapping region only. This is done by a preliminary sampling to find the boundaries of the overlapping volume. Per each pair of particles the overlapping volume is computed by performing a MC integration of the volume using 10^7 samples, where the boundaries of the MC region are taken as the smallest rectangular prism encapsulating the volume itself.

Contact function of hard spherocylinders. In order to numerically evaluate the contact distance between two HSCs, we first place them at the origin of the fixed coordinate axes frame (XYZ). Then, particles I and J are rotated about X and Y axes; subsequently, by keeping particle I fixed, particle J is first shifted in the Z direction by a fixed amount (still overlapping with particle I), and finally iteratively displaced in the X direction with a fine step until the minimum distance d_m[79] between the two cylinders' central axes equals D, which corresponds to having the rods with their surfaces 'touching'. From a total of 431,316 samples, 10^5 randomly selected were used as the reference FG data to build the model. As usual, we perform a 20/80 split of the data into test/training sets.

Monte Carlo simulations

We perform MC simulations of systems containing $N = 4$ and 5 particles at low temperature to determine the equilibrium clusters and to compare the FG and CG models of spherical surface patterned particles. Due to the *rugged* potential energy landscapes, we use a parallel tempering method, also known as Replica Exchange Monte Carlo (REMC), which has previously been used to simulate Janus spheres[48]. For an efficient sampling, we perform each simulation using a total of 8 replicas at temperatures $T_i = T_0 \exp(\alpha_T i)$, with a reference (target) temperature of $k_B T_0/\epsilon_{\max} = 0.01$ and maximum temperature of 0.1, where $\epsilon_{\max}$ represents the (absolute) maximum value of the two-body interaction energy between surface-patterned particles. In the MC simulations of individual replicas, rotation and translation moves are attempted with equal probability on a randomly selected particle. Prior to the replica exchange moves, the maximum displacements are adjusted to ensure an acceptance rate of approximately 30%. During the parallel tempering runs, a MC cycle (corresponding to N trial moves) is performed on each replica in parallel, followed by an attempted exchange of two adjacent replicas. The exchange moves are attempted by alternating between pairs of replicas on each cycle. More precisely, during odd MC cycles, exchange moves are attempted only between replicas 1–2, 3–4, 5–6 and 7–8, while during even MC cycles, such moves are attempted between replicas 2–3, 4–5 and 6–7. Simulations were typically performed for 1.5×10^6 steps. Examples of low-energy clusters comprising $N = 4$ and 5 particles of spherical charged Janus particles and IPCs as obtained from REMC simulations using the FG and CG models are shown in Fig. 3b and in Supplementary Fig. 3.

In the case of colloidal rounded cubes interacting via an effective two-body depletion potential, we performed canonical MC simulations of $N = 110$ particles interacting through the ML-CG $\Phi(R_{IJ}, \Omega_I, \Omega_J, \Omega_{IJ})$ potential on top of their hard-core interaction. To model the hard core character of the particles, we use the Gilbert–Johnson–Keerthi (GJK)[81] collision-detection algorithm, as in ref. 82. In the quasi-2D simulations, equally-probable rotation and translation moves on randomly selected particles were performed only in two dimensions, say the XY-plane, and the maximum displacements were tuned to achieve a 30% acceptance rate. Simulations starting from either low-density states or square lattices, were fun for a total of 10^7 MC steps.

Data availability

The authors declare that the data supporting the findings of this study are available within supplementary information files. Other datasets are available from the corresponding author upon reasonable request.

Code availability

Example codes supporting the findings of this study are available within the supplementary information files. Other programs and scripts are available from the corresponding author upon reasonable request.

Received: 29 February 2024; Accepted: 19 August 2024;
Published online: 28 September 2024

References

1. Glotzer, S. C. & Solomon, M. J. Anisotropy of building blocks and their assembly into complex structures. *Nat. Mater.* **6**, 557–562 (2007).
2. Nguyen, T. D. & Plimpton, S. J. Aspherical particle models for molecular dynamics simulation. *Computer Phys. Commun.* **243**, 12–24 (2019).
3. Roberts, C. J. & Blanco, M. A. Role of anisotropic interactions for proteins and patchy nanoparticles. *J. Phys. Chem. B* **118**, 12599–12611 (2014).

4. Dijkstra, M. & Luijten, E. From predictive modelling to machine learning and reverse engineering of colloidal self-assembly. *Nat. Mater.* **20**, 762–773 (2021).
5. Dijkstra, M. Entropy-driven phase transitions in colloids: From spheres to anisotropic particles. *Adv. Chem. Phys.* **156**, 35–71 (2014).
6. Boles, M. A., Engel, M. & Talapin, D. V. Self-assembly of colloidal nanocrystals: From intricate structures to functional materials. *Chem. Rev.* **116**, 11220–11289 (2016).
7. Fejer, S. N., Chakrabarti, D. & Wales, D. J. Self-assembly of anisotropic particles. *Soft Matter* **7**, 3553–3564 (2011).
8. Allen, M., Evans, G., Frenkel, D. & Mulder, B. Hard convex body fluids. *Adv. Chem. Phys.* **86**, 1–166 (1993).
9. Bassani, C. L. et al. Nanocrystal assemblies: Current advances and open problems. *ACS Nano* **18**, 14791–14840 (2024).
10. Grzelczak, M., Pérez-Juste, J., Mulvaney, P. & Liz-Marzán, L. M. Shape control in gold nanoparticle synthesis. *Chem. Soc. Rev*. **37**, 1783–1791 (2008).
11. Rogach, A. L. et al. Organization of matter on different size scales: monodisperse nanocrystals and their superstructures. *Adv. Funct. Mater.* **12**, 653–664 (2002).
12. Murphy, C. J. et al. Anisotropic metal nanoparticles: synthesis, assembly, and optical applications. *J. Phys. Chem. B* **109**, 13857–13870 (2005).
13. Sacanna, S. & Pine, D. J. Shape-anisotropic colloids: Building blocks for complex assemblies. *Curr. Opin. Colloid Interface Sci.* **16**, 96–105 (2011).
14. Gou, L. & Murphy, C. J. Solution-phase synthesis of cu2o nanocubes. *Nano Lett.* **3**, 231–234 (2003).
15. Rossi, L. et al. Shape-sensitive crystallization in colloidal superball fluids. *Proc. Natl Acad. Sci. USA* **112**, 5286–5290 (2015).
16. Ahmadi, T. S., Wang, Z. L., Green, T. C., Henglein, A. & El-Sayed, M. A. Shape-controlled synthesis of colloidal platinum nanoparticles. *Science* **272**, 1924–1925 (1996).
17. Malikova, N., Pastoriza-Santos, I., Schierhorn, M., Kotov, N. A. & Liz-Marzán, L. M. Layer-by-layer assembled mixed spherical and planar gold nanoparticles: control of interparticle interactions. *Langmuir* **18**, 3694–3697 (2002).
18. Greyson, E. C., Barton, J. E. & Odom, T. W. Tetrahedral zinc blende tin sulfide nano-and microcrystals. *small* **2**, 368–371 (2006).
19. Hong, L., Cacciuto, A., Luijten, E. & Granick, S. Clusters of charged janus spheres. *Nano Lett.* **6**, 2510–2514 (2006).
20. Walther, A. & Muller, A. H. Janus particles: synthesis, self-assembly, physical properties, and applications. *Chem. Rev.* **113**, 5194–5261 (2013).
21. Yi, G.-R., Pine, D. J. & Sacanna, S. Recent progress on patchy colloids and their self-assembly. *J. Phys. Condens. Matter* **25**, 193101 (2013).
22. van Blaaderen, A. Chemistry: Colloidal molecules and beyond. *Science* **301**, 470–471 (2003).
23. Kuijk, A., Byelov, D. V., Petukhov, A. V., Van Blaaderen, A. & Imhof, A. Phase behavior of colloidal silica rods. *Faraday Discuss.* **159**, 181–199 (2012).
24. Fernández-Rico, C. et al. Shaping colloidal bananas to reveal biaxial, splay-bend nematic, and smectic phases. *Science* **369**, 950–955 (2020).
25. Tlusty, T. & Safran, S. Defect-induced phase separation in dipolar fluids. *Science* **290**, 1328–1331 (2000).
26. Pawar, A. B. & Kretzschmar, I. Fabrication, assembly, and application of patchy particles. *Macromol. rapid Commun.* **31**, 150–168 (2010).
27. Bianchi, E., van Oostrum, P. D., Likos, C. N. & Kahl, G. Inverse patchy colloids: Synthesis, modeling and self-organization. *Curr. Opin. Colloid Interface Sci.* **30**, 8–15 (2017).
28. Poulin, P., Stark, H., Lubensky, T. & Weitz, D. Novel colloidal interactions in anisotropic fluids. *Science* **275**, 1770–1773 (1997).
29. Senyuk, B., Puls, O., Tovkach, O. M., Chernyshuk, S. B. & Smalyukh, I. I. Hexadecapolar colloids. *Nat. Commun.* **7**, 10659 (2016).
30. Yuan, Y., Tasinkevych, M. & Smalyukh, I. I. Colloidal interactions and unusual crystallization versus de-mixing of elastic multipoles formed by gold mesoflowers. *Nat. Commun.* **11**, 188 (2020).
31. Chernyshuk, S. High-order elastic terms, boojums and general paradigm of the elastic interaction between colloidal particles in the nematic liquid crystals. *Eur. Phys. J. E* **37**, 1–9 (2014).
32. Ge, Y. et al. Constructing coarse-grained skyrmion potentials from experimental data with iterative boltzmann inversion. *Commun. Phys.* **6**, 30 (2023).
33. Sohn, H. R., Liu, C. D. & Smalyukh, I. I. Schools of skyrmions with electrically tunable elastic interactions. *Nat. Commun.* **10**, 4744 (2019).
34. Tai, J.-S. B. & Smalyukh, I. I. Three-dimensional crystals of adaptive knots. *Science* **365**, 1449–1453 (2019).
35. Goujon, F. et al. Backbone oriented anisotropic coarse grains for efficient simulations of polymers. *J. Chem. Phys.* **153**, 214901 (2020).
36. Cohen, A. E., Jackson, N. E. & De Pablo, J. J. Anisotropic coarse-grained model for conjugated polymers: Investigations into solution morphologies. *Macromolecules* **54**, 3780–3789 (2021).
37. Nguyen, H. T. & Huang, D. M., Systematic bottom-up molecular coarse-graining via force and torque matching using anisotropic particles. J. Chem. Phys. **156**, 184118 (2022).
38. Stone, A. The description of bimolecular potentials, forces and torques: the s and v function expansions. *Mol. Phys.* **36**, 241–256 (1978).
39. Gay, J. & Berne, B. Modification of the overlap potential to mimic a linear site–site potential. *J. Chem. Phys.* **74**, 3316–3319 (1981).
40. Berardi, R., Fava, C. & Zannoni, C. A generalized gay-berne intermolecular potential for biaxial particles. *Chem. Phys. Lett.* **236**, 462–468 (1995).
41. Cleaver, D. J., Care, C. M., Allen, M. P. & Neal, M. P. Extension and generalization of the gay-berne potential. *Phys. Rev. E* **54**, 559 (1996).
42. Everaers, R. & Ejtehadi, M. Interaction potentials for soft and hard ellipsoids. *Phys. Rev. E* **67**, 041710 (2003).
43. Memmer, R., Kuball, H.-G. & Schönhofer, A. Computer simulation of chiral liquid crystal phases. i. the polymorphism of the chiral gay-berne fluid. *Liq. Cryst.* **15**, 345–360 (1993).
44. Kihara, T. The second virial coefficient of non-spherical molecules. *J. Phys. Soc. Jpn.* **6**, 289–296 (1951).
45. Campos-Villalobos, G., Dijkstra, M. & Patti, A. Nonconventional phases of colloidal nanorods with a soft corona. *Phys. Rev. Lett.* **126**, 158001 (2021).
46. Lee, B. H.-j & Arya, G. Analytical van der waals interaction potential for faceted nanoparticles. *Nanoscale Horiz.* **5**, 1628–1642 (2020).
47. Wang, J., Yu, H., Langston, P. & Fraige, F. Particle shape effects in discrete element modelling of cohesive angular particles. *Granul. Matter* **13**, 1–12 (2011).
48. Hieronimus, R., Raschke, S. & Heuer, A. How to model the interaction of charged janus particles. *J. Chem. Phys.* **145**, 064303 (2016).
49. Camerin, F., Aguilar, S. M., & Dijkstra, M., Depletion-induced crystallization of anisotropic triblock colloids, *Nanoscale* **16**, 4724–4736 (2024).
50. Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 9759–9815 (2021).
51. Boattini, E., Bezem, N., Punnathanam, S. N., Smallenburg, F. & Filion, L. Modeling of many-body interactions between elastic spheres through symmetry functions. *J. Chem. Phys.* **153**, 064902 (2020).
52. Campos-Villalobos, G., Boattini, E., Filion, L. & Dijkstra, M. Machine learning many-body potentials for colloidal systems. *J. Chem. Phys.* **155**, 174902 (2021).
53. Giunta, G., Campos-Villalobos, G., & Dijkstra, M., Coarse-grained many-body potentials of ligand-stabilized nanoparticles from machine-learned mean forces. *ACS Nano* **17**, 23391–23404 (2023).
54. Zhou, Y., Bore, S. L., Tao, A. R., Paesani, F. & Arya, G. Many-body potential for simulating the self-assembly of polymer-grafted nanoparticles in a polymer matrix. *npj Computational Mater.* **9**, 224 (2023).

55. Campos-Villalobos, G., Giunta, G., Marín-Aguilar, S. & Dijkstra, M. Machine-learning effective many-body potentials for anisotropic particles using orientation-dependent symmetry functions. *J. Chem. Phys.* **157**, 024902 (2022).
56. Argun, B. R., Fu, Y., & Statt, A., Molecular dynamics simulations of anisotropic particles accelerated by neural-net predicted interactions. *J. Chem. Phys.* **160**, 244901 (2024).
57. Wilson, M. O. & Huang, D. M. Anisotropic molecular coarse-graining by force and torque matching with neural networks. *J. Chem. Phys.* **159**, 024110 (2023).
58. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
59. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
60. Guidarelli Mattioli, F., Sciortino, F., & Russo, J., A neural network potential with self-trained atomic fingerprints: a test with the mw water potential. *J. Chem. Phys.* **158**, 104501 (2023).
61. Stone, A., *The theory of intermolecular forces* (oUP oxford, 2013).
62. Blum, L. & Torruella, A. Invariant expansion for two-body correlations: Thermodynamic functions, scattering, and the ornstein-zernike equation. *J. Chem. Phys.* **56**, 303–310 (1972).
63. Martin, D. E. p. wigner, group theory and its application to the quantum mechanics of atomic spectra, (academic press inc., new york, 1959), j. j. griffin, ix + 372 pp.,80s. *Proc. Edinb. Math. Soc.* **12**, 67–67 (1960).
64. Brink, D. & Satchler, G. *Angular Momentum*. Oxford library of the physical sciences (Clarendon Press, 1962) https://books.google.nl/books?id=i1TNswEACAAJ.
65. Zewdie, H. Computer simulation studies of liquid crystals: A new corner potential for cylindrically symmetric particles. *J. Chem. Phys.* **108**, 2117–2133 (1998).
66. Steele, W. A. & Pecora, R. Scattering from fluids of nonspherical molecules. i. x rays and neutrons. *J. Chem. Phys.* **42**, 1863–1871 (1965).
67. Fiałkowski, M., Kapanowski, A. & Sokalski, K. Microscopic approach to theory of biaxial nematic liquid crystals. *Mol. Cryst. Liq. Cryst. Sci. Technol. Sect. A. Mol. Cryst. Liq. Cryst.* **265**, 371–385 (1995).
68. Berardi, R., Muccioli, L., & Zannoni, C., Field response and switching times in biaxial nematics. *J. Chem. Phys.* **128**, 024905 (2008).
69. Querciagrossa, L., Orlandi, S., Ricci, M., Arcioni, A. & Berardi, R. Chiral gay–berne model for molecular dynamics computer simulations. *Mol. Cryst. Liq. Cryst.* **684**, 66–81 (2019).
70. Berardi, R., Ricci, M. & Zannoni, C. Ferroelectric nematic and smectic liquid crystals from tapered molecules. *ChemPhysChem* **2**, 443–447 (2001).
71. Bianchi, E., Kahl, G. & Likos, C. N. Inverse patchy colloids: from microscopic description to mesoscopic coarse-graining. *Soft Matter* **7**, 8313–8323 (2011).
72. Savenko, S. & Dijkstra, M. Phase behavior of a suspension of colloidal hard rods and nonadsorbing polymer. *J. Chem. Phys.* **124**, 234902 (2006).
73. Henzie, J., Grünwald, M., Widmer-Cooper, A., Geissler, P. L. & Yang, P. Self-assembly of uniform polyhedral silver nanocrystals into densest packings and exotic superlattices. *Nat. Mater.* **11**, 131–137 (2012).
74. Asakura, S. & Oosawa, F. Interaction between particles suspended in solutions of macromolecules. *J. Polym. Sci.* **33**, 183–192 (1958).
75. Dijkstra, M., Brader, J. M. & Evans, R. Phase behaviour and structure of model colloid-polymer mixtures. *J. Phys. Condens. Matter* **11**, 10079 (1999).
76. Jiao, Y., Stillinger, F. & Torquato, S. Optimal packings of superdisks and the role of symmetry. *Phys. Rev. Lett.* **100**, 245504 (2008).
77. Bolhuis, P. & Frenkel, D. Tracing the phase boundaries of hard spherocylinders. *J. Chem. Phys.* **106**, 666–687 (1997).
78. Corner, J. The second virial coefficient of a gas of non-spherical molecules. *Proc. R. Soc. Lond. Ser. A* **192**, 275–292 (1948).
79. Vega, C. & Lago, S. A fast algorithm to evaluate the shortest distance between rods. *Comput. Chem.* **18**, 55–59 (1994).
80. Leavitt, R. P. An irreducible tensor method of deriving the long-range anisotropic interactions between molecules of arbitrary symmetry. *J. Chem. Phys.* **72**, 3472–3482 (1980).
81. Gilbert, E. G., Johnson, D. W. & Keerthi, S. S. A fast procedure for computing the distance between complex objects in three-dimensional space. *IEEE J. Robot. Autom.* **4**, 193–203 (1988).
82. Wang, D. et al. Interplay between spherical confinement and particle shape on the self-assembly of rounded cubes. *Nat. Commun.* **9**, 2228 (2018).

Acknowledgements

G.C.-V. thanks Susana Marín-Aguilar and Fabrizio Camerin for stimulating and useful discussions. G. C.-V. and R.S. acknowledge funding from The Netherlands Organization for Scientific Research (NWO) for the ENW PPS Fund 2018 - Technology Area Soft Advanced Materials ENPPS.TA. 018.002 and for the OCENW.KLEIN.432, respectively. M.D. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. ERC-2019-ADG 884902= SoftML).

Author contributions

G.C.-V. and R.S. contributed equally. G.C.-V., G.G. and M.D. conceived and designed the project. G.C.-V. and R.S. developed and implemented the approach. G.C.-V., R.S. and M.D. wrote the manuscript with advice and comments from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01405-4.

Correspondence and requests for materials should be addressed to Gerardo Campos-Villalobos or Marjolein Dijkstra.

Reprints and permissions information is available at http://www.nature.com/reprints