REVIEW ARTICLE | INSIGHT
https://doi.org/10.1038/s41563-021-01014-2

nature materials

From predictive modelling to machine learning and reverse engineering of colloidal self-assembly

Marjolein Dijkstra [1] ✉ and Erik Luijten [2] ✉

An overwhelming diversity of colloidal building blocks with distinct sizes, materials and tunable interaction potentials are now available for colloidal self-assembly. The application space for materials composed of these building blocks is vast. To make progress in the rational design of new self-assembled materials, it is desirable to guide the experimental synthesis efforts by computational modelling. Here, we discuss computer simulation methods and strategies used for the design of soft materials created through bottom-up self-assembly of colloids and nanoparticles. We describe simulation techniques for investigating the self-assembly behaviour of colloidal suspensions, including crystal structure prediction methods, phase diagram calculations and enhanced sampling techniques, as well as their limitations. We also discuss the recent surge of interest in machine learning and reverse-engineering methods. Although their implementation in the colloidal realm is still in its infancy, we anticipate that these data-science tools offer new paradigms in understanding, predicting and (inverse) design of novel colloidal materials.

The Stone Age, Bronze Age and Iron Age denote periods in history during which humans began to master these naturally occurring materials. But what if we could fabricate our own new materials with targeted properties by simply arranging the atoms as desired? As early as 1960, Richard Feynman challenged us to think 'from the bottom up' and create new materials by directing and manipulating the arrangements of individual atoms[1]. He invited us to enter a completely new field in physics, where we have unprecedented control over the properties and functionalities of new materials. Although this was only a distant dream at the time, advances in modern experimental synthesis techniques and revolutions in nanotechnology have brought us tantalizingly close to realizing this dream.

A promising approach towards achieving this goal is hierarchical self-assembly, the spontaneous organization of individual particles into ordered structures and the most important strategy used by nature for the formation of complex biofunctional structures. In this process, materials are fabricated by first assembling atoms into molecules, then combining molecules into larger units with sizes from a few nanometres to several micrometres, and finally letting these so-called colloidal building blocks, suspended in a liquid, self-organize into three-dimensional ordered structures. These self-assembled materials have well defined structure on the scale of tens to hundreds of nanometres and an incredibly large surface-to-volume ratio—properties that make them perfectly suited not only for optoelectronic, plasmonic and photonic applications, but also for catalysis and energy storage.

Successful implementation of this strategy hinges on the ability to synthesize and fabricate novel nanoparticles and colloidal particles. Although recent advances have resulted in a dazzling variety of new building blocks with interaction potentials that can be tailored from hard to soft repulsive, attractive, dipolar, shape anisotropic, patchy or even self-propelling, only a tiny fraction of the infinite number of possible colloidal building blocks, rivalling the molecular 'toolkit' of chemists, has been fabricated, despite extensive efforts devoted to developing new synthesis routes. Reviews that provide more details on colloidal interactions include refs. [2–7].

To accelerate advances in materials science, it is desirable to guide experimental efforts by theoretical predictions that permit the rational design of new materials. Owing to the high-dimensional parameter space that arises in multicomponent systems and the complexity that emerges when both short- and long-range interactions are present or when interactions become anisotropic or even many body in nature, it is often not possible to infer structures directly from the building blocks, as is sometimes possible for atoms by considering their valency. Computational modelling offers the potential to tackle such problems. During the past few decades, computer simulations have become a powerful tool for predicting the structure and phase behaviour of systems of particles with various shapes and interactions, including systems that have otherwise intractable complexity.

In this Review, we discuss computer simulation methods and strategies for the design of soft materials using self-assembly of colloids and nanoparticles. A comprehensive and up-to-date coverage of these topics is timely, given the rapid technical evolution of the field over the past decade, the burgeoning research efforts in colloid and nanoparticle self-assembly and the increasing prominence of simulation strategies in studies of soft materials. Our treatise is structured as follows. First, we briefly describe the simulation techniques that are commonly employed to study the phase behaviour and structure of colloidal systems, followed by a description of the effective interactions between colloidal particles in solution. These effective interactions can be used to predict self-assembled structures, for example crystal structures at infinite pressure or zero temperature, and subsequently the equilibrium phase diagram can be determined. Enhanced sampling techniques can accelerate these predictions. Finally, we discuss the classification of thermodynamic phases using order parameters, followed by a discussion of machine learning (ML) techniques and methods to reverse-engineer building blocks for targeted self-assembly. We conclude with an outlook.

Simulation strategies

Colloidal suspensions consist of mesoscopic particles dispersed in a solvent. Due to their size, these particles exhibit Brownian motion caused by the thermal agitation of solvent molecules that collide with their surfaces. In addition, the viscosity of the fluid results in viscous drag forces on the colloids. As a consequence, these systems have a genuine thermodynamic character and the entire machinery

[1]Soft Condensed Matter, Debye Institute for Nanomaterial Science, Department of Physics, Utrecht University, Utrecht, The Netherlands. [2]Departments of Materials Science and Engineering, Engineering Sciences & Applied Mathematics, Chemistry and Physics & Astronomy, Northwestern University, Evanston, IL, USA. ✉e-mail: M.Dijkstra@uu.nl; luijten@northwestern.edu

of equilibrium statistical mechanics can be used to describe the collective behaviour of these suspensions. Since the huge difference in time and length scales between the colloids and the solvent molecules prevents naïve implementation of a molecular dynamics (MD) scheme for both species, colloidal particles are typically treated as macroscopic bodies immersed in a structureless solvent. In this representation, the solvent is viewed as a homogeneous dielectric and viscous continuum and the colloidal particles interact via effective interactions. These colloid-only systems can be studied using standard Monte Carlo (MC) and MD simulations, thereby neglecting not only the Brownian motion of the colloids but also the long-range hydrodynamic interactions between them originating from the flow fields generated by their motion. Although hydrodynamic interactions are difficult to model due to their many-body character, various simulation techniques have been developed to take them into account, including Stokesian dynamics, the lattice Boltzmann method, fluid particle dynamics, multi-particle collision dynamics/stochastic rotation dynamics and dissipative particle dynamics[8]. These simulation techniques are computationally expensive and often the effect of hydrodynamic interactions may be neglected, for example in studies that are only concerned with the equilibrium behaviour of the suspension. Brownian dynamics simulations can be employed to mimic Brownian motion of a colloidal particle by friction and stochastic forces[9,10]. Finally, we note that solvents can be structured and thereby induce additional interactions. Examples include phase-separated solvents with particles adsorbed to the interface[11], binary solvent mixtures close to the critical point, resulting in critical Casimir forces between the colloids[12], and liquid-crystalline solvents with long-range forces between the particles induced by topological defects[13].

Effective interactions

A central ingredient in all simulations is the effective interactions between the colloids. Fluctuating dipoles give rise to an attractive dispersion or van der Waals force between each pair of atoms. When summed over all atom pairs in two colloidal particles, this results in a strong van der Waals attraction between the colloids, exceeding the thermal energy k_BT by orders of magnitude. To suppress irreversible aggregation, we can use either steric stabilization, realized by grafting a layer of polymers or ligands on the particle surface, or charge stabilization[14]. The resulting colloids then have effective interactions that can be mediated through a variety of mechanisms, summarized in Fig. 1 and discussed below.

Hard-particle interactions. The simplest effective interactions are those that merely represent the shape or excluded volume of the particles. In this case self-assembly is purely driven by entropy. One of the most famous examples is a hard-sphere fluid, comprised of equisized spheres that experience no interactions except that they cannot overlap. As demonstrated by computer simulations in 1957, this athermal system—counterintuitively—crystallizes at sufficiently high packing fractions[15,16]. Once particles are permitted to be anisotropic, a wide variety of crystalline, liquid-crystalline and quasicrystalline structures can arise[17–20]. The contact interactions make MC and event-driven (rather than force-based) MD the computational methods of choice. The primary challenge then consists of efficiently determining overlaps between particles. For convex objects, the Gilbert–Johnson–Keerthi algorithm[21] is an efficient choice, whereas non-convex polygonal objects can be dealt with through the detection of polygon interference[22].

Depletion interactions. When multiple types of particle are present, entropic effects become even more complicated. An important example of this is so-called depletion interactions[23], often illustrated in a binary suspension of colloids and a smaller species (for example flexible polymers or smaller colloids). If the surface separation between the larger colloids is smaller than the size of the smaller species, the concentration of the latter is reduced (hence the name 'depletant') in the region between the large colloids, resulting in an imbalanced force (osmotic pressure) that forces them together. The attraction induced by the smaller species depends on the size asymmetry between the two components as well as the concentration of the smaller component, and is often represented as an effective interaction. This is particularly advantageous in situations where the small particles are numerous compared with the colloidal component. Yet, important caveats are in order as well. The original Asakura–Oosawa–Vrij model assumes that the depletants behave ideally with respect to one another[23,24], an assumption that clearly does not hold when these particles occupy a non-negligible volume fraction or interact with one another, for example via electrostatic repulsions. The situation becomes even more complicated when the depletants interact with the larger species through more than a pure 'excluded-volume' repulsion. For example, charged depletants can be attracted by ionic surface groups on the colloid, and also van der Waals attractions between the depletants and the colloid must be considered. In these situations, the entropic effects of the depletion interaction compete with energetic (enthalpic) effects.

Another aspect that is sometimes overlooked is that effective interactions between the larger species are typically calculated on a pairwise basis. However, this approximation is not guaranteed to yield accurate effective potentials for situations where effective three-body and many-body interactions cannot be neglected, as in the case of large depletants. One way to tackle this problem is to derive a formal expression for the effective Hamiltonian of the colloids by integrating out the degrees of freedom of the depletants.

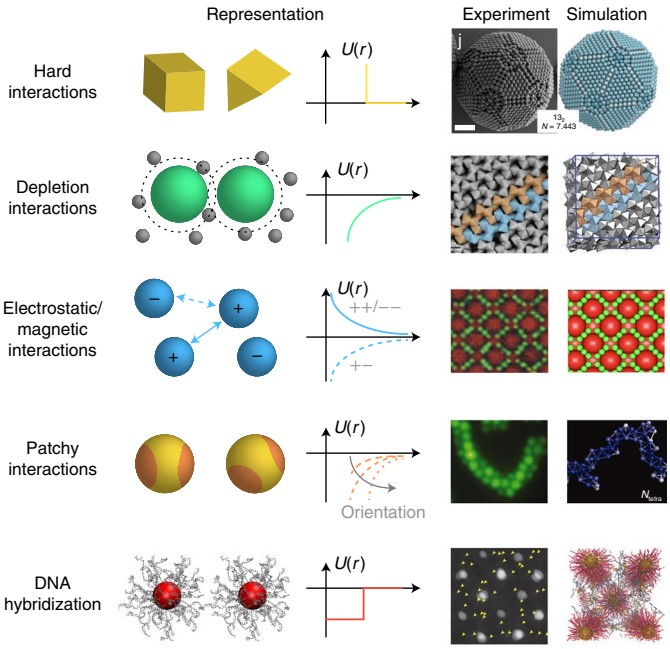


Fig. 1 | Effective interactions between colloids and nanoparticles, with typical examples of the resulting self-assembly behaviour observed in experiments and predicted in simulations. From top to bottom, the examples are excluded-volume interactions[133], depletion interactions[134], electrostatic interactions[70], hydrophobic patchy interactions[135] and interactions based on DNA hybridization[69]. Credit: figure reproduced with permission from ref. [133], under a Creative Commons License CC BY 4.0 (hard interactions); and ref. [134], Springer Nature Ltd (depletion interactions) and adapted with permission from ref. [70], APS (electrostatic/magnetic interactions); ref. [135] (patchy interactions) and ref. [69] (DNA hybridization), AAAS.

The effective Hamiltonian consisting of effective polymer-mediated two- and higher-body terms can subsequently be evaluated 'on the fly' in MC simulations[25].

An alternative approach to address complications that arise from colloid–depletant and depletant–depletant interactions is the explicit inclusion of the smaller species. Apart from the sheer number of particles that need to be dealt with, the primary bottleneck that arises in an MD simulation is the disparity in timescales. The motion of smaller particles takes place on far shorter timescales than that of the larger particles, and this dictates the magnitude of the time step in the MD integration scheme. As a result, we risk the simulation becoming non-ergodic. A standard Metropolis Markov-chain MC simulation encounters comparable problems, as the distribution of the smaller species sets the maximum displacement of the larger species that can be permitted without letting the acceptance rate become vanishingly small. Under certain conditions, a highly efficient approach is the geometric cluster algorithm[26] described below.

Coulombic interactions. Most colloids carry electrostatic charge, especially when immersed in an aqueous solution, as this promotes the ionization of chemical groups at the colloidal surface owing to the high permittivity of the medium. Even though this high permittivity also leads to a strong weakening of the electrostatic interactions between charge carriers, the long-range nature of the Coulomb potential often makes it one of the dominant interactions in colloidal suspensions[27]. Computationally, the long-range character is dealt with in systems with periodic boundary conditions via the Ewald summation technique[9,10]. In its traditional formulation, this method has an efficiency that scales with the number of discrete charge carriers N as $N^{3/2}$. Efficient electrostatic solvers, such as the particle–particle particle–mesh[28] and particle–mesh Ewald[29] methods, improve this to $\mathcal{O}(N\log N)$, whereas the fast multipole method even achieves $\mathcal{O}(N)$ scaling[30].

In salt-free conditions, each colloid still has its own cloud of neutralizing counterions. This condition is often presented as the equivalent of deionized water as a solvent, but self-ionization of water in fact imposes a lower bound on the salt concentration. The counterion shell along with the colloidal surface charges is referred to as the electric double layer, and the ion distribution is described by the Poisson–Boltzmann equation. At higher salt concentrations, the electrostatic interactions are screened by co- and counterions and hence for efficiency reasons in computer simulations the colloidal particles are often assumed to interact via screened Coulomb (Debye–Hückel) or Yukawa potentials in lieu of explicit salt ions[31]. A shortcoming of this mean-field approach is that it ignores fluctuation, correlation and many-body effects, causing its predictions to break down under so-called strong-coupling conditions. This occurs for example for multivalent ions in aqueous solutions at room temperature or for highly deionized colloidal suspensions, where a state-dependent volume term may drive a van der Waals instability[32]. Interesting colloidal phenomena that can be induced under these conditions are 'like-charge attraction'[33] and 'overcharging', also referred to as 'charge reversal'[34]. Moreover, electrostatic interactions can act as the driving force for colloidal crystallization[35] and stabilization, either through direct electrostatic repulsion or indirectly via surface binding of charged nanoparticles[36,37].

Coulombic interactions become far more complicated when polarizability is taken into account. Spatially non-uniform permittivity gives rise to many-body effects that are colloquially referred to as 'image effects'. In recent years, various approaches have been designed to account for non-uniform permittivity in particle-based simulations (Box 1).

Finally, we note that magnetic interactions allow particularly convenient control over colloidal assembly; the corresponding literature is vast. Here, we merely note that they can be treated with the same mathematical techniques as Coulombic interactions. The interactions between ferromagnetic particles are often represented by effective dipole moments, but in the case of paramagnetic interactions these must be solved self-consistently. The treatment of magnetic polarizability also follows the approach outlined above for electrostatic interactions and many phenomena are analogous.

Box 1 | Methods for modelling dielectric effects

For simple geometries (isolated planar, cylindrical and spherical interfaces), the induced surface charge can be easily calculated to very high precision through image charges. However, for systems with other dielectric geometries, no image-based solutions exist. Likewise, for systems comprised of for example multiple dielectric spheres, the situation rapidly becomes computationally intractable, since image charges iteratively induce additional image charges. This can be managed through truncation of the process, which in turns limits the numerical accuracy[140–142]. A highly efficient simulation technique to simulate the electrostatics in a medium with dielectric inhomogeneities was proposed by Maggs and Rossetto, in which an auxiliary electrostatic field is evolved using local MC simulations[143]. The advantage of this method is that it circumvents the time-consuming computation of the long-ranged Coulombic interactions and that it can be straightforwardly applied to systems with a spatially varying dielectric permittivity. An alternative that can be applied to arbitrary geometries with piecewise uniform permittivities relies on the boundary-element method and solves the Poisson equation on grid points at the dielectric interfaces. This approach has been proposed several times with varying degrees of efficiency[144–151], depending on the approach adopted to solve this problem self-consistently and the electrostatic solver employed. The refinements proposed by Barros et al.[151] improved the efficiency to such a degree that application to mobile dielectric interfaces became possible. This was subsequently exploited to demonstrate that polarization effects can radically alter self-assembled structures[152], precisely due to the many-body nature of these interactions. A noteworthy development is a hybrid method that combines the superior accuracy afforded by image charges at short separations with the efficiency of the method of moments at large distances. This approach—which is restricted to systems of dielectric spheres—is two orders of magnitude faster than the boundary-element method while simultaneously achieving an accuracy that is three orders of magnitude improved for colloidal aggregates[153].

Patchy interactions. Whereas most computer simulations traditionally represent colloids as spherical building blocks with isotropic interactions, the introduction of directional or 'patchy' interactions opens numerous possibilities for steering self-assembly and other collective phenomena. Going beyond the hard building blocks with anisotropic shapes discussed above, the most common realizations are so-called one-patch and Janus particles, which feature two sides with distinct surface chemistries[38–42]. Multicompartment particles are also possible[43,44]. The surface functionalization makes it possible to create particles with hemispheres that are charged, hydrophobic, paramagnetic, grafted with polymer brushes, different in roughness and so on. This in turn permits remarkable control over colloidal assembly as well as particle motion.

The surface anisotropy of (spherical) Janus colloids makes their effective interaction considerably more complicated. A widely used potential is due to Kern and Frenkel[45] and allows directional short-range attractions. In subsequent work on Janus particles with charged hemispheres, Cacciuto et al. developed a more complicated

potential that permits repulsions and attractions[46]. The study of Janus particles has grown exponentially over the past two decades, from fundamental investigations of their phase behaviour[47] and kinetics[48] to applications, for example in the pharmaceutical industry. Interestingly, variation of the number of patches permits control not only over the structure of the liquid, but also over the phase diagram[49,50], whereas careful choice of patterning symmetry and patch shape even makes it possible to achieve specific crystal structures[51].

Interactions based on DNA hybridization. A particularly successful technique for creating predefined crystal structures relies on the selective bonding of DNA oligonucleotides[52,53]. As demonstrated simultaneously by teams led by Mirkin[54] and Gang[55], mixtures of nanoparticles grafted with complementary single-stranded DNA can be made to assemble into face-centred-cubic and body-centred-cubic crystal structures. This seminal work sparked extensive exploration of the degree of programmability afforded by DNA hybridization, where the use of DNA linkers with different lengths and 'melting' temperatures and the composition of the nanoparticle or colloid mixture as well as particle sizes serve as control parameters[56]. Moreover, shape anisotropy of the particles can be used to impact directionality of the DNA bonding[57], thus permitting the formulation of design rules for nanoparticle-based materials[58–60]. Generalization to micrometre-sized colloids can lead to the formation of glassy random aggregates, which can be avoided through adjustment of grafting density and annealing protocol[61] or by DNA linkers that are fully mobile on the colloid surface[62]. These experimental efforts have been followed by theoretical and computational developments that have provided insights into the role of cooperative bonding and stability, and have started to yield predictive capabilities[63]. Strong DNA bonding makes direct MC or MD simulations very challenging, such that a substantial degree of coarse-graining is unavoidable[64–67], yet combined experimental–computational work continues to yield new conceptual insights[68,69]. Going beyond supramolecular self-assembly and crystallization, DNA-coated colloids also find important applications in nanomedicine thanks to their ability of selective and reversible binding, where the same computational strategies can be applied.

Crystal structure prediction

A central computational challenge in self-assembly is the prediction of the structure and phase behaviour of systems of colloidal particles and nanoparticles with various shapes and interactions. Given the interactions of the constituents, which crystal structures are stable? Conventional methods are often based on a preselection of structures for which the free energies are calculated to determine the thermodynamically most stable phase. This selection of structures relies on intuition and trial and error, and has the serious drawback that it rules out all non-selected structures, which might include stable equilibrium structures, at the outset.

Various simulated annealing techniques have been developed to predict the crystal structures for atomic and molecular systems (Box 2). These optimization techniques can also be employed to predict the energetically stabilized crystal structures of colloidal and nanoparticle systems. For instance, simulated annealing has been used to determine the minimum-energy crystal structures for oppositely charged colloids[35,70], whereas genetic algorithms were used to predict the ground-state candidate crystal structures of monolayers of dipolar particles[71], patchy colloids[72] and particles that interact through square-shoulder potentials[73].

An important factor that makes colloidal systems both more complicated and more interesting is that entropy may play an important role in the self-assembly, so the entropic contribution to the free energy cannot be neglected in the search for candidate crystal structures. For example, at finite temperatures new structures may appear in the phase diagram with symmetries that differ from the ground-state crystal structures.

In the extreme case of hard-particle systems, the self-assembled structures are stabilized by entropy alone. As a result, the minimization techniques for atomic and molecular systems cannot be straightforwardly applied to such systems, since it is difficult to construct an appropriate cost function—the potential energy of the system is always zero as only overlap-free configurations contribute to the partition function. One way to circumvent this problem is to replace the hard-core potential with a fictitious continuous potential and use the corresponding potential energy as a cost function[74,75]. However, when applied in a genetic algorithm, this approach is not very efficient for sampling configuration space, since new configurations generated by crossover and mutation often result in particle overlaps. Techniques to avoid or remove such overlaps markedly slow down the sampling. Alternatively, packing arguments are often employed to rationalize the self-assembled structures of hard-particle systems. The underlying idea is that structures with a higher close-packed density have a larger free volume per particle at lower densities, resulting in a higher entropy and hence a lower free energy. To predict the densest packings of hard-particle systems the adaptive shrinking-cell method[17], shape-fluctuating boxes[76] and the floppy-box MC simulation method[75] have been employed. These methods are based on simple MC simulations in the isobaric–isothermal ensemble. In the floppy-box MC method, the number of particles is chosen to be very small so that as little as one particle can be simulated in a unit cell. Since the shape of the simulation box can vary, all crystal symmetries can be sampled. If MC simulations are combined with simulated annealing, the densest packing can be obtained by gradually increasing the pressure and the ground-state structures can be found by progressively decreasing the temperature. This approach has been applied to spheres with various interactions, including hard interactions, attractive interactions, anisotropic interactions, medium-range soft interactions, truly long-range interactions[75] and patchy interactions[72]. For non-spherical particles, the floppy-box MC method has been combined with a triangular tessellation method to describe the surface of arbitrarily shaped particles, and a fast detection method for finding overlaps between triangles to predict infinite-pressure crystal

Box 2 | Crystal prediction of molecular systems

Simulated annealing techniques are commonly used for predicting crystal structures of atomic and molecular systems. These methods are based on (1) a description of the system or crystal structure in terms of a set of parameters such as the size and shape of the unit cell, (2) a procedure to explore the configuration space of the system and (3) a cost (energy) function, which is minimized using simulated annealing to establish the lowest-energy structure. These approaches can be applied to any atomic or molecular system for which a suitable cost function can be constructed. Alternative minimization techniques, typically aimed at locating the global potential-energy minimum of a system and thereby limited to probing the zero-temperature phase behaviour, include evolutionary algorithms[154,155] and particle swarm optimization[156]. These methods work well for a relatively smooth (free-) energy landscape with a well defined global minimum. In the case of 'rugged' (free-) energy landscapes the system can become trapped in local minima, and basin-hopping methods[157,158] or metadynamics techniques[159] are more suitable as they provide strategies to escape from such minima. We note that many of these optimization techniques are based on simulations with a variable box shape to efficiently sample crystal structures with different symmetries.

structures of shape-anisotropic particles[17,77], whereas shape-fluctuating boxes were employed to predict the self-assembly of hard polyhedra from the fluid phase[19].

Phase diagram calculations

Beyond the mere prediction of crystal structures lies the determination of the full phase behaviour of specific systems. Numerous simulation methods have been developed over the past decade to achieve this, the simplest of which amounts to performing simulations at different state points of the system and identifying the resulting phases and structures as a function of temperature, pressure or density. The thermodynamic phases can be identified either by visual inspection or by measuring thermodynamic variables, suitable order parameters or ML; see below. For colloidal particles with enthalpic interactions, we can perform cooling and heating runs, in which we either start from a dilute gas and lower the temperature until the system undergoes a phase transition, or initialize the system in a stable (ground-state) crystal structure and increase the temperature until it melts. Alternatively, compression runs may be carried out by increasing the pressure (starting from the fluid phase) or expansion runs can be employed by lowering the pressure from the (most dense-packed) crystal structure. The latter approach is commonly used for hard-particle systems.

In the case of a first-order phase transition, this approach will be highly inaccurate due to hysteresis, as a free-energy barrier has to be crossed to transform the system between phases. The height of this barrier is determined by the chemical potential difference between the stable and metastable phases and the free-energy cost of creating an interface between the two coexisting phases. In the common case of finite-sized systems with periodic boundary conditions, the barrier height is typically very high, since it scales with the interfacial area. To transform a system to the stable phase, high supersaturations or supercoolings are required, where it may become kinetically arrested or possible intermediate phases may be missed. For this reason, methods to study phase coexistence (Box 3) typically aim to eliminate the interface.

Alternatively, an older method that has recently received renewed attention is the direct-coexistence method[78]. We start with two well equilibrated bulk phases, for example a solid and a liquid phase, combined in a single simulation box. Subsequently, simulations are performed at varying pressures or temperatures to determine when the two-phase coexistence is stable. When the pressure (temperature) is much lower than the bulk pressure (temperature), the system will immediately melt (freeze), whereas the system will rapidly freeze (melt) if the pressure (temperature) is much higher.

Once a coexistence point has been located, the full coexistence curve can be traced out using numerical integration of the Clausius–Clapeyron equation[79]. This approach is not limited to the pressure–temperature plane, but has also been used to determine for example the freezing and melting curves of hard spheres as a function of size polydispersity[80] and the fluid–solid coexistence of charged colloids as a function of the screening parameter[31].

Enhanced sampling techniques

All approaches discussed so far focus on thermodynamic equilibrium. However, the formation of a thermodynamically stable crystal phase also depends on kinetic effects. Crystallization may be suppressed by glassy dynamics or gelation and may suffer from defects, such as stacking faults and vacancies. Elucidating nucleation rates and kinetic pathways for the transformation from a metastable phase to a thermodynamically stable phase requires efficient sampling of rare events, since free-energy barriers must be crossed. As noted above, high supersaturation or undercooling often are not viable solutions. Therefore, we discuss several enhanced sampling techniques to accelerate barrier crossing and sample configurational space in an effective manner. These methods can be categorized into (1) collective variable biasing techniques, such as umbrella sampling, hyperdynamics, metadynamics and adiabatic free-energy dynamics, (2) path-based techniques, such as transition path sampling, string methods and forward-flux sampling, (3) parallel tempering techniques, for example replica-exchange, Hamiltonian-exchange and multicanonical algorithms, and (4) cluster algorithms. We discuss the techniques most relevant for colloidal systems.

Umbrella sampling, first proposed by Torrie and Valleau[81], overcomes free-energy barriers by modulating the acceptance criterion in an MC simulation through application of a bias that depends on a suitably chosen reaction coordinate. Often the most appropriate choice for this coordinate is the order parameter (for example the density for a gas–liquid transition or the six-fold bond order parameter for a fluid–solid transition), but in general finding the optimal order parameter can be challenging. ML may help to find the relevant set of order parameters.

Alternatively, nucleation can be studied via forward-flux sampling[82], which introduces a reaction coordinate or order parameter similar to the umbrella sampling method. The nucleation rate between states A and B is then expressed as a product of the flux of trajectories crossing the A state boundary and the probability that a trajectory that has crossed this boundary will reach state B before

Box 3 | Methods for determining phase coexistence

In the Gibbs ensemble MC method[160], two phases are simulated in separate simulation boxes, which exchange volume and particles at a given temperature until the numbers of particles and the volumes of the two boxes reach their equilibrium values. The resulting densities of these boxes correspond to the densities of the coexisting phases for a given pressure and chemical potential. This method becomes inaccurate close to a critical point, where the densities of the two coexisting phases become identical and the boxes can change identity during a simulation run. Furthermore, this approach breaks down when one of the coexisting phases becomes so dense that exchange of particles and volume is prohibited, for example in a high-density fluid or solid phase.

A more accurate method to determine phase coexistence relies on measuring the probability $P(N)$ of observing N particles in a volume V at chemical potential μ using grand-canonical MC simulations. To increase the sampling efficiency of $P(N)$ we can employ (successive) umbrella sampling in combination with histogram reweighting[161]. At phase coexistence, the probability distribution function becomes bimodal with two separate peaks of equal area for the two coexisting phases. The coexisting densities can be determined from the average number of particles in the two peaks, whereas the interfacial tension of the two coexisting phases can be obtained from the height of the two peaks and the minimum separating them[162].

The most accurate but also most expensive way to determine phase coexistence is by calculating the free energies of all possible phases, for example the candidate structures as predicted by the (crystal) structure prediction methods discussed above. Direct computation of the free energy is intractable as it involves a high-dimensional phase-space integral. However, we can calculate free-energy differences between two bulk systems using the thermodynamic integration technique. In this method, we construct a reversible path that links the system of interest to a reference system for which the free energy is known, for example the ideal gas or the Einstein crystal[10]. Once the Helmholtz free energies of the candidate phases have been determined, we can apply common-tangent constructions to the free-energy curves to establish the thermodynamically stable phases and the two-phase coexistence regions.

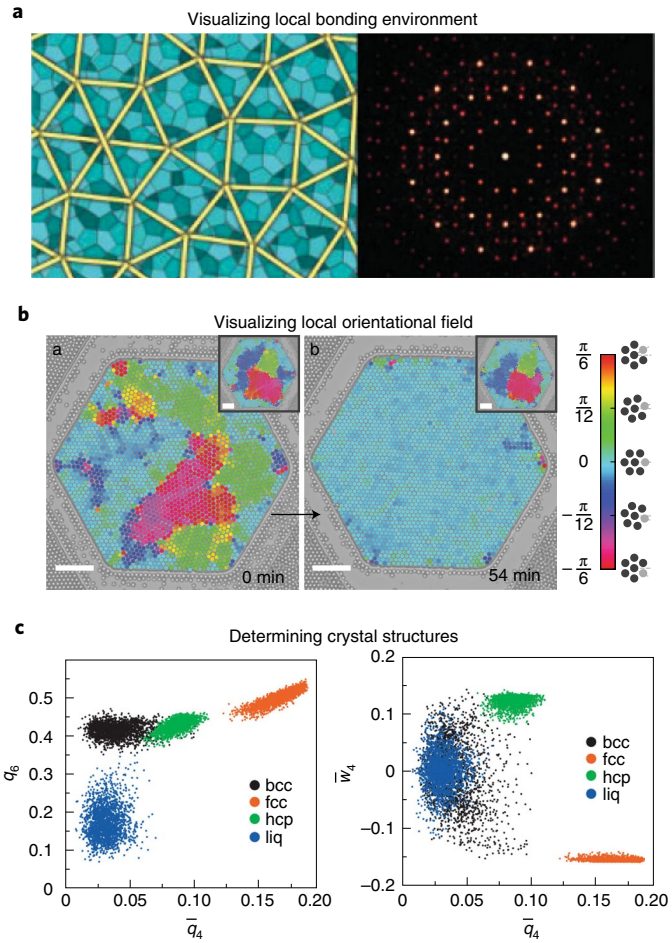


Fig. 2 | Example applications of correlation functions and bond order parameters to classify thermodynamic phases and pathways. a, Left: connected tetrahedra in a slice through a quasicrystal, with the tiling structure highlighted in yellow. Right: corresponding diffraction pattern indicating 12-fold symmetry[136]. **b**, Annealing of colloidal monolayers, with time evolution of the bond orientational correlation function shown in the inset[137]. Scale bars, 50 μm. **c**, Distinction of various fluid and solid phases via scatter plots of bond order parameters[94]. bcc, body-centred cubic; fcc, face-centred cubic; hcp, hexagonal close-packed; liq, liquid. Credit: figure adapted with permission from ref. [136], Springer Nature Ltd (**a**) and reproduced with permission from ref. [137], under a Creative Commons License CC BY 4.0 (**b**) and ref. [94], AIP (**c**).

returning to state A. Forward-flux sampling facilitates the calculation of this probability by breaking it up into a set of probabilities between sequential values of the reaction coordinate or order parameter. Consequently, forward-flux sampling requires no previous detailed knowledge of the nucleation process, and the choice of order parameter is less important than in umbrella sampling.

Finally, nucleation events can be studied in computer simulation via the transition path sampling method[83]. Here, a collection of trajectories connecting one state to another is generated via standard simulation tools such as MD or MC. By assigning a probability to each of the many pathways, we can construct an MC random walk in the space of the transition trajectories, using for example the 'shooting method', and generate an ensemble of transition paths. All relevant information, such as the nucleation mechanisms (classical or non-classical), the transition states and the nucleation rate, can subsequently be extracted from the ensemble of these paths. Because no a priori reaction coordinate is required, the resulting path ensemble gives an unbiased insight into the nucleation mechanism and its kinetics.

Going beyond nucleation events, parallel tempering[84] allows efficient and ergodic sampling by exploiting the increased propensity of a system at elevated temperatures to cross free-energy barriers. Multiple independent MC or MD simulations are started in parallel, where each of these 'replicas' has a different temperature. At fixed intervals, the configuration of a run is swapped with a configuration at a neighbouring temperature. The acceptance criterion of this swap is chosen such that detailed balance is obeyed. This approach permits configurations in deep free-energy minima (that is at low temperatures) to transition, via a 'ladder' of successive swaps, to other free-energy minima. An advantage of parallel tempering is that it requires minimal modification of a standard MC or MD simulation, allowing massively parallel simulations with very limited communication between the parallel processing units. On the other hand, the overall computational cost can be very high. For systems with considerable free-energy barriers a large number of parallel runs are needed, with only the run at the actual lowest temperature yielding relevant information.

MC simulations offer an alternative route to enhanced sampling, as they merely have to reproduce the Boltzmann distribution without the need to emulate the physical dynamics of a system. Liu and Luijten[26], building on earlier work[85,86], devised an MC algorithm for colloidal systems that bypasses physical dynamics in favour of collective ('cluster') moves that are rejection free regardless of the pairwise interactions between any of the components in a system. Remarkably, the process by which the cluster moves are constructed guarantees that, despite the rejection-free nature of the algorithm, the particle configurations are generated according to the Boltzmann distribution. As the method generates uncorrelated configurations at a rate that is exponentially faster than conventional MC or MD simulations, it outperforms standard approaches by orders of magnitude for highly size-asymmetric mixtures, provided that the density of the system is not too high. This property makes it particularly suitable for determining generalized depletion potentials. A derivative of this method, in which the requirement for rejection-free updates is dropped, has been proposed for studying the self-assembly of strongly attractive particles[87], whereas the rejection-free variant was extended to the determination of phase equilibria[88] and generalized to anisotropic particles[89]. An alternative rejection-free MC eliminates the detailed-balance condition but employs event-driven moves in a manner that generates the Boltzmann distribution[90,91].

Classification of thermodynamic phases

The techniques described to determine crystal structures and phase diagrams rely on the ability to define appropriate order parameters that measure the broken symmetries in the respective phases. Crystalline solids are characterized by a broken translational and bond-orientational symmetry. To distinguish crystalline and fluid phases, Steinhardt, Nelson and Ronchetti introduced the concept of bond-orientational order parameters[92]. The analysis starts by identifying the local environment of each particle in the system, for example all neighbouring particles within a certain radial distance, a Voronoi construction or a solid-angle nearest-neighbour criterion[93]. Subsequently, the set of distance vectors ('bonds') between a particle and its neighbours is expanded in a set of spherical harmonics. To obtain order parameters that are rotationally invariant, we can construct combinations of the spherical harmonics. Local bond order parameters or averaged and reweighted variations[94,95] can be used to study structural and dynamic processes such as nucleation and growth, melting, grain boundary dynamics and defect dynamics (Fig. 2). The six-fold bond order parameter is commonly used to bias the sampling towards crystal nucleation and to study crystallization in a wide variety of systems[96], whereas the six-fold

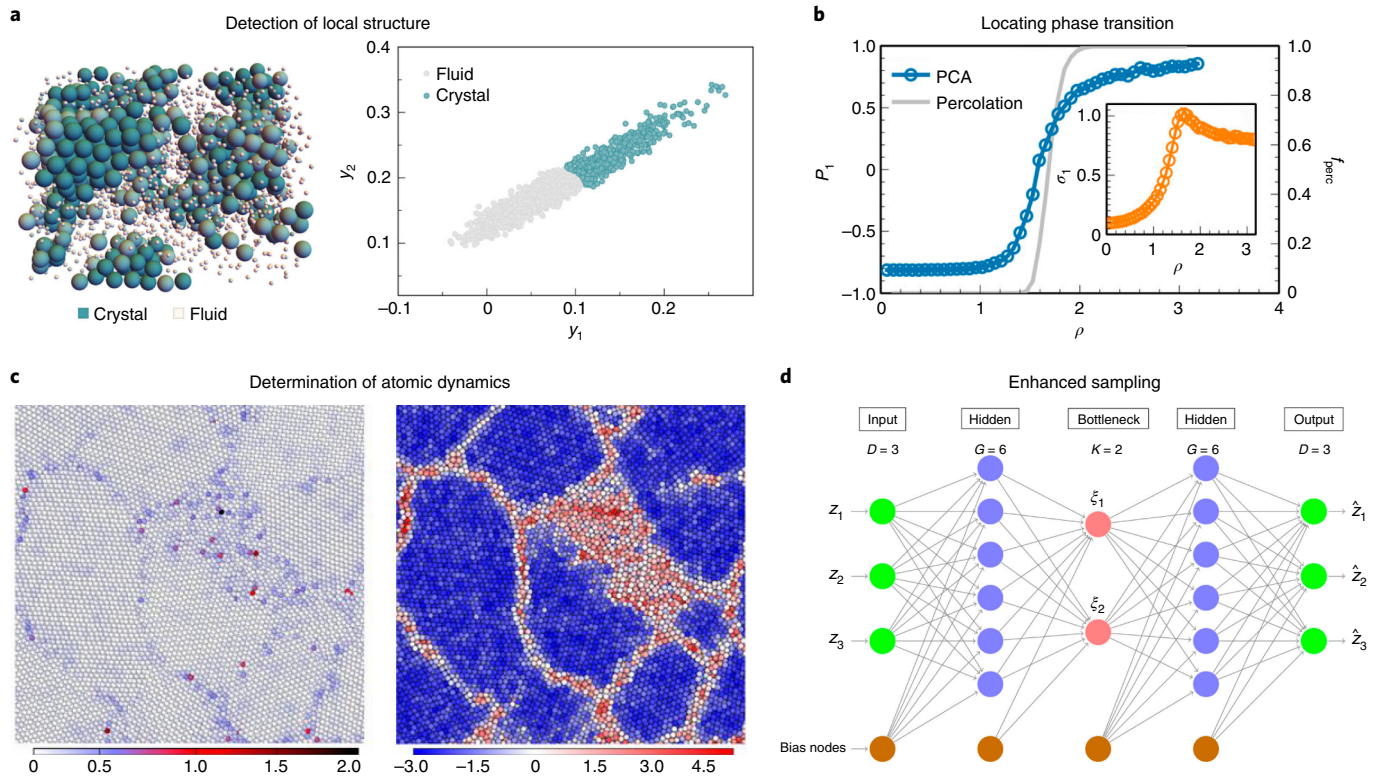


Fig. 3 | Examples of ML techniques applied to atomic and colloidal systems. a, Detection of local structure (left) via unsupervised learning, quantified in a two-dimensional parameter space identified by an autoencoder (right)[115]. **b**, Locating the phase transition in the Widom–Rowlinson model using principal-component analysis (PCA)[113]. **c**, Dynamics of atoms at grain boundaries in simulated polycrystalline aluminium characterized by their 'hopping' probability (left) and their 'softness' (right)[138]. **d**, Autoassociative neural network comprised of an encoding (leftmost three layers) and a decoding (rightmost three layers) part[139]. Credit: figure reproduced with permission from ref. [115], AIP (**a**); ref. [113], AIP (**b**); ref. [138], National Academy of Sciences (**c**); and ref. [139], AIP (**d**).

and four-fold bond order parameters are often used to distinguish body-centred-cubic, face-centred-cubic, hexagonal close-packed and fluid states. Other ways to describe the symmetry of the local environment of a particle are the common neighbourhood analysis, bond-orientational order diagrams (histograms of bond orientations), diffraction patterns and topological cluster classification algorithm[97]. To classify and identify distinct crystal structures, we can also use the angles and lengths of the lattice vectors of the respective unit cells[98] or evaluate the closeness of the lattice vectors to reference crystals[99].

Machine learning

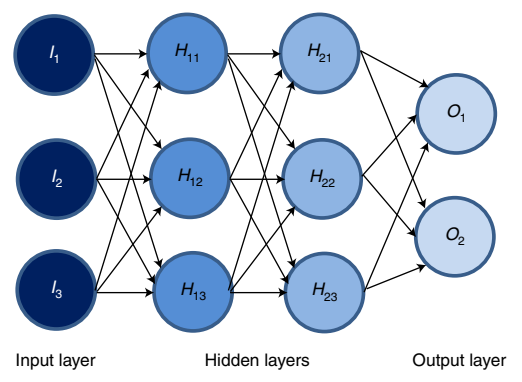
Recently, ML has become a viable tool for analysing computer simulations and/or experiments. ML can be used to characterize and classify complex data sets, to find non-obvious patterns in big data and to reduce the number of random variables in multivariate data. The focus of ML is on the development of computer programs that can (1) learn from data that are given without being programmed explicitly, (2) make predictions and decisions by finding patterns in data and (3) adapt (or learn) when exposed to new data. Although ML has been adopted successfully in many research fields, the soft-matter field has lagged behind in embracing this approach. There are several factors that make the application of ML in colloid science challenging and have hampered successful implementations. The main difficulty stems from the importance of entropy in soft-matter systems, which results not only in complex phase behaviour, but also in thermodynamic phases that are highly susceptible to thermal fluctuations. The abundance of 'exotic' phases with (partial) orientational and positional order, such as liquid crystals, quasicrystals and plastic crystals, along with the omnipresent thermal noise, makes classification of these states of matter using ML tools non-trivial. Despite these inherent challenges, the promise of materials design through ML is too great to ignore, calling for a thorough investigation of its applicability in soft matter (Fig. 3).

ML algorithms can roughly be divided into 'supervised' and 'unsupervised' learning. In supervised learning (Box 4), a computer program (or artificial neural network) is trained with a set of example inputs along with the desired outputs. The objective is to develop general rules between output and input and to subsequently provide output for previously unseen input data. This form of ML can be used for classification when the outputs are discrete and for regression in the case of continuous outputs. Neural networks have been trained to recognize various types of local crystalline order in systems of Lennard-Jones particles, Yukawa particles, water molecules and binary hard spheres[100–102]. In these cases, the different outputs (crystalline phases) are known and the input consists of the local environment of each particle, described with either symmetry functions[100], bond order parameters[101,102], particle positions[103] or image recognition[104]. The trained neural network can subsequently be used as an 'order parameter' to classify local structures, for example in nucleation studies. Furthermore, ML methods have been used to find correlations between local structure and dynamical rearrangements and to make predictions on the long-term dynamics of glassy systems[105,106].

Another exciting application of neural networks is to develop coarse-grained models of the system by integrating out the microscopic degrees of freedom of certain species. Using simulations of the full system, ML can be used to accurately fit the effective

Box 4 | Supervised learning

An artificial neural network is an ML algorithm that mimics the way the human brain recognizes patterns. It consists of interconnected layers of artificial 'neurons' and is trained with 'labelled' data consisting of a set of example inputs (represented by the vector $\mathbf{I}=I_1, I_2, \ldots$) along with the desired outputs ($\mathbf{O}=O_1, O_2, \ldots$). The remaining layers are so-called hidden layers. Each neuron (denoted by a circle) receives inputs from the neurons in the previous layer, processes these inputs by calculating a weighted sum of all inputs together with a bias and passes this weighted sum through an activation function to the neurons in the next layer. By adjusting the weights and biases we can tune the network's output to the training set. This process of adjusting the weights and biases is called 'learning'.



Neural network consisting of an input layer (dark blue), a number of hidden layers and an output layer.

Box 5 | Unsupervised learning

Unsupervised ML techniques, such as principal-component analysis and the diffusion map method, project high-dimensional data onto a low-dimensional manifold. The central idea of principal-component analysis is to reduce the dimension of the input data in a manner that retains most of the information of the original data set by projecting it onto a new set of orthogonal basis vectors. This transformation is defined such that the first principal component $\mathbf{p}_1$ accounts for the largest possible variance in the data, the second component $\mathbf{p}_2$ for the second largest variance and so on. A set of M simulation configurations described by L variables, for example Cartesian coordinates and particle orientations, is represented by an $M\times L$ matrix X. The principal components of this data set are obtained from the eigenvalues of the covariance matrix X^TX. The corresponding eigenvectors constitute the new basis of principal components, $\mathbf{p}_1, \ldots, \mathbf{p}_L$. Each configuration i, described by an L-dimensional vector $\mathbf{x}_i$, is projected onto the jth principal component $\mathbf{p}_j$ using $t_{j,i}=\mathbf{x}_i\cdot\mathbf{p}_j$ ($j=1,\ldots,L$; $i=1,\ldots,M$), such that the $t_{j,1},\ldots,t_{j,M}$ show successively the largest possible variance with increasing principal component j. As the jth eigenvalue of the covariance matrix is equal to the variance of the data along the jth principal component, the set of principal components that exhibit substantial variance can easily be identified. The data in the high-dimensional phase space can thus be projected onto the lower-dimensional manifold spanned by these dominant principal components, neglecting the remaining principal components. If the data exhibit nonlinear features, principal-component analysis will fail and a nonlinear dimensionality reduction method such as the diffusion map approach should be employed.

(many-body) interactions as a function of the particle coordinates in the same spirit as the Behler–Parrinello ML potentials derived from ab initio simulations[107]. The resulting potentials, which still capture the essential physics, can subsequently be employed in coarse-grained simulations[108]. Finally, we also mention the use of an active learning approach to efficiently sample a phase diagram, thereby accelerating materials design and discovery[109].

In the case of unsupervised learning (Box 5), unlabelled data are used as input without corresponding output variables. The algorithm aims to find structures, patterns and correlations in big data sets, thus reducing the number of variables by mapping high-dimensional data onto a low-dimensional space. By identifying groups of local environments and detecting correlations between various input data, different thermodynamic phases can be identified and new order parameters can be designed that identify the corresponding phases. Unsupervised learning methods, such as diffusion map dimensionality reduction, principal-component analysis, Gaussian mixture models and autoassociative neural networks, have been applied to identify different crystal structures for a wide variety of systems[110–115]. More recently, unsupervised learning has also been employed to detect phase transitions[113] and to study self-assembly pathways[116]. We refer the interested reader to several recent reviews that discuss recent ML developments in the soft-matter field[117–119].

Reverse-engineering methods

Conventional colloid science is based on forward design: which structures with what properties are formed by a given colloidal building block? By contrast, an inverse-design strategy starts from desired materials properties, translates them into a structure and then identifies the building blocks that self-assemble into this structure. This inverse problem is a tremendous challenge, owing to the number of possible structures, compositions, interactions and control parameters. A multitude of approaches has been exploited to devise interaction potentials for such targeted self-assembly. Most inverse-design methods start with an variational ansatz for the pair potential, parametrized by a set of parameters. Torquato determined the optimal values of these parameters by evaluating the difference in ground-state energy between the target lattice and its competing structures and minimizing the Lindemann parameter of the target structure near melting[120]. This approach was employed for the inverse design of interaction potentials for low-coordinated crystal lattices such as the honeycomb and square lattices, and disordered ground states, but also targeted bulk properties[120]. Truskett et al. introduced an inverse-design method[121] using maximum-likelihood ML or the relative-entropy coarse-graining approach[122]. Here, the optimal parameter values for the pair potentials are found by maximizing the likelihood of reproducing the targeted configurations. This approach has been used to rationally design pair potentials for porous structures, a variety of open lattices such as honeycomb, kagome, (truncated) square, rectangular and truncated hexagonal lattices, binary crystals[123] and Frank–Kasper and quasicrystalline phases[124]. Disordered structures with large photonic band gaps were designed with a constraint optimization method[125] and particle shapes for targeted colloidal crystals via entropy maximization[126]. We can also use a statistical physics approach[127] to design the interaction potentials and the system parameters for a target crystal structure[128]. Finally, we note that an inverse-design approach should take into account not only the thermodynamics but also the kinetics of the targeted self-assembly. A first step in this direction was demonstrated recently through the use of diffusion maps to project the stable states and assembly pathways of patchy

colloids onto a low-dimensional assembly landscape and to rationally design the interaction strength that optimizes the yield of the target self-assembly[129,130]. We refer the interested reader to 'Inverse design of soft materials with desired function' in this issue as well as refs. [118] and [131] for a more elaborate review.

Future challenges and paths towards their resolution

The computational prediction of colloidal self-assembly, from aggregates to crystals, has made spectacular strides in the past two decades. Nevertheless, resolution of this type of problem continues to be enormously demanding, so it is satisfying to note that the development of new techniques is still an active area of research. As a concrete example, ML techniques have played an important role in the field, with examples ranging from crystal structure classification to the identification of suitable order parameters.

Beyond the inevitable extension to systems with more accurate potentials and larger numbers of components (and hence correspondingly larger numbers of phases), we anticipate rapid growth in several specific directions. Active, that is self-propelled, particles represent a burgeoning field in themselves, with numerous exciting aspects as well as noteworthy connections to systems that represent life. A primary challenge for predictive capabilities for such active matter is the degree to which its behaviour is controlled by an underlying minimum-free-energy principle and a corresponding probability distribution, comparable to the way in which equilibrium phenomena are governed by the Boltzmann distribution. Another direction that is gaining traction is the extension to smaller length scales. In this case, assumptions that are valid for micrometre-sized colloids must be relinquished and aspects such as polydispersity, surface roughness, non-uniform surface chemistry including molecular details of the capping layer, and structuring of the solvent molecules must be taken into account. Moreover, the typical interaction range is no longer small compared with the size of the particle[132].

Finally, with increasingly powerful computational capabilities, the use of colloidal design rules to create particles suitable for specific applications is coming within reach. Given the vast swaths of parameter space that at best have received only a cursory glance, we propose that a more systematic framework correlating colloidal properties with their collective behaviour, perhaps in a high-throughput screening and data-driven design method, would greatly facilitate the adoption of these building blocks for the creation of new materials with targeted properties. Another route that could be undertaken is the 'inverse-design' strategy, which involves optimization of the shape and interactions of the building blocks as well as the system parameters for a targeted structure with specific properties.

Received: 20 May 2020; Accepted: 19 April 2021;
Published online: 27 May 2021

References

1. Feynman, R. P. There's plenty of room at the bottom. *Eng. Sci.* **23**, 22–36 (1960).
2. Glotzer, S. C. & Solomon, M. J. Anisotropy of building blocks and their assembly into complex structures. *Nat. Mater.* **6**, 557–562 (2007).
3. 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).
4. Likos, C. N. Soft matter with soft particles. *Soft Matter* **2**, 478–498 (2006).
5. Sacanna, S. & Pine, D. J. Shape-anisotropic colloids: building blocks for complex assemblies. *Curr. Opin. Colloid Interface Sci.* **16**, 96–105 (2011).
6. Cademartiri, L. & Bishop, K. J. Programmable self-assembly. *Nat. Mater.* **14**, 2–9 (2015).
7. Rovigatti, L., Gnan, N., Tavagnacco, L., Moreno, A. J. & Zaccarelli, E. Numerical modelling of non-ionic microgels: an overview. *Soft Matter* **15**, 1108–1119 (2019).
8. Bolintineanu, D. S. et al. Particle dynamics modeling methods for colloid suspensions. *Comput. Part. Mech.* **1**, 321–356 (2014).
9. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids* (Clarendon, 1987).
10. Frenkel, D. & Smit, B. *Understanding Molecular Simulation* 2nd edn (Academic, 2002).
11. Binks, B. P. & Horozov, T. S. *Colloidal Particles at Liquid Interfaces* (Cambridge University Press, 2006).
12. Maciołek, A. & Dietrich, S. Collective behavior of colloids due to critical Casimir interactions. *Rev. Mod. Phys.* **90**, 045001 (2018).
13. Muševič, I. Nematic liquid-crystal colloids. *Materials* **11**, 24 (2018).
14. Dijkstra, M. Computer simulations of charge and steric stabilised colloidal suspensions. *Curr. Opin. Colloid Interface Sci.* **6**, 372–382 (2001).
15. Alder, B. J. & Wainwright, T. E. Phase transition for a hard sphere system. *J. Chem. Phys.* **27**, 1208–1209 (1957).
16. Wood, W. W. & Jacobson, J. Preliminary results from a recalculation of the Monte Carlo equation of state of hard spheres. *J. Chem. Phys.* **27**, 1207–1208 (1957).
17. Torquato, S. & Jiao, Y. Dense packings of the Platonic and Archimedean solids. *Nature* **460**, 876–879 (2009).
18. Agarwal, U. & Escobedo, F. A. Mesophase behaviour of polyhedral particles. *Nat. Mater.* **10**, 230 (2011).
19. Damasceno, P. F., Engel, M. & Glotzer, S. C. Predictive self-assembly of polyhedra into complex structures. *Science* **337**, 453–457 (2012).
20. Dijkstra, M. Entropy-driven phase transitions in colloids: from spheres to anisotropic particles. *Adv. Chem. Phys.* **156**, 35 (2015).
21. 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).
22. GAMMA Research Group at the University of North Carolina *RAPID—Robust and Accurate Polygon Interference Detection* http://gamma.cs.unc.edu/OBB/ (1997).
23. Asakura, S. & Oosawa, F. On interaction between two bodies immersed in a solution of macromolecules. *J. Chem. Phys.* **22**, 1255–1256 (1954).
24. Vrij, A. Polymers at interfaces and the interactions in colloidal dispersions. *Pure Appl. Chem.* **48**, 471–483 (1976).
25. Dijkstra, M., van Roij, R., Roth, R. & Fortini, A. Effect of many-body interactions on the bulk and interfacial phase behavior of a model colloid–polymer mixture. *Phys. Rev. E* **73**, 041404 (2006).
26. Liu, J. & Luijten, E. Rejection-free geometric cluster algorithm for complex fluids. *Phys. Rev. Lett.* **92**, 035504 (2004).
27. Linse, P. Structure, phase stability, and thermodynamics in charged colloidal solutions. *J. Chem. Phys.* **113**, 4359–4373 (2000).
28. Hockney, R. W. & Eastwood, J. W. *Computer Simulation Using Particles* (McGraw-Hill, 1981).
29. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an $N\cdot\log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
30. Greengard, L. & Moura, M. On the numerical evaluation of electrostatic fields in composite materials. *Acta Numer.* **3**, 379–410 (1994).
31. Hynninen, A.-P. & Dijkstra, M. Phase diagrams of hard-core repulsive Yukawa particles. *Phys. Rev. E* **68**, 021407 (2003).
32. van Roij, R., Dijkstra, M. & Hansen, J.-P. Phase diagram of charge-stabilized colloidal suspensions: van der Waals instability without attractive forces. *Phys. Rev. E* **59**, 2010 (1999).
33. Linse, P. & Lobaskin, V. Electrostatic attraction and phase separation in solutions of like-charged colloidal particles. *Phys. Rev. Lett.* **83**, 4208–4211 (1999).
34. Levin, Y. Strange electrostatics in physics, chemistry, and biology. *Physica A* **352**, 43–52 (2005).
35. Leunissen, M. E. et al. Ionic colloidal crystals of oppositely charged particles. *Nature* **437**, 235–240 (2005).
36. Tohver, V., Smay, J. E., Braem, A., Braun, P. V. & Lewis, J. A. Nanoparticle halos: a new colloid stabilization mechanism. *Proc. Natl Acad. Sci. USA* **98**, 8950–8954 (2001).
37. Liu, J. & Luijten, E. Stabilization of colloidal suspensions by means of highly charged nanoparticles. *Phys. Rev. Lett.* **93**, 247802 (2004).
38. Sciortino, F., Giacometti, A. & Pastore, G. Phase diagram of Janus particles. *Phys. Rev. Lett.* **103**, 237801 (2009).
39. Jiang, S. et al. Janus particle synthesis and assembly. *Adv. Mater.* **22**, 1060–1071 (2010).
40. Walther, A. & Müller, A. H. E. Janus particles: synthesis, self-assembly, physical properties, and applications. *Chem. Rev.* **113**, 5194–5261 (2013).
41. Smallenburg, F. & Sciortino, F. Liquids more stable than crystals in particles with limited valence and flexible bonds. *Nat. Phys.* **9**, 554–558 (2013).
42. Zhang, J., Luijten, E. & Granick, S. Toward design rules of directional Janus colloidal assembly. *Annu. Rev. Phys. Chem.* **66**, 581–600 (2015).
43. Du, J. & O'Reilly, R. K. Anisotropic particles with patchy, multicompartment and Janus architectures: preparation and application. *Chem. Soc. Rev.* **40**, 24020–2416 (2011).

44. Chen, Q. et al. Triblock colloids for directed self-assembly. *J. Am. Chem. Soc.* **133**, 7725–7727 (2011).
45. Kern, N. & Frenkel, D. Fluid–fluid coexistence in colloidal systems with short-ranged strongly directional attraction. *J. Chem. Phys.* **118**, 9882–9889 (2003).
46. Hong, L., Cacciuto, A., Luijten, E. & Granick, S. Clusters of charged Janus spheres. *Nano Lett.* **6**, 2510–2514 (2006).
47. Sciortino, F., Giacometti, A. & Pastore, G. Phase diagram of Janus particles. *Phys. Rev. Lett.* **103**, 237801 (2009).
48. Zhang, J., Luijten, E., Grzybowski, B. A. & Granick, S. Active colloids with collective mobility: status and research opportunities. *Chem. Soc. Rev.* **46**, 5551–5569 (2017).
49. Bianchi, E., Largo, J., Tartaglia, P., Zaccarelli, E. & Sciortino, F. Phase diagram of patchy colloids: towards empty liquids. *Phys. Rev. Lett.* **97**, 168301 (2006).
50. Ruzicka, B. et al. Observation of empty liquids and equilibrium gels in a colloidal clay. *Nat. Mater.* **10**, 56–60 (2011).
51. Romano, F. & Sciortino, F. Patterning symmetry in the rational design of colloidal crystals. *Nat. Commun.* **3**, 975 (2012).
52. Mirkin, C. A., Letsinger, R. L., Mucic, R. C. & Storhoff, J. J. A DNA-based method for rationally assembling nanoparticles into macroscopic materials. *Nature* **382**, 607–609 (1996).
53. Alivisatos, A. P. et al. Organization of 'nanocrystal molecules' using DNA. *Nature* **382**, 609–611 (1996).
54. Park, S. Y. et al. DNA-programmable nanoparticle crystallization. *Nature* **451**, 553–556 (2008).
55. Nykypanchuk, D., Maye, M. M., van der Lelie, D. & Gang, O. DNA-guided crystallization of colloidal nanoparticles. *Nature* **451**, 549–552 (2008).
56. Jones, M. R., Seeman, N. C. & Mirkin, C. A. Programmable materials and the nature of the DNA bond. *Science* **347**, 1260901 (2015).
57. Jones, M. R., Macfarlane, R. J., Prigodich, A. E., Patel, P. C. & Mirkin, C. A. Nanoparticle shape anisotropy dictates the collective behavior of surface-bound ligands. *J. Am. Chem. Soc.* **133**, 18865–18869 (2011).
58. Martinez-Veracoechea, F. J., Mladek, B. M., Tkachenko, A. V. & Frenkel, D. Design rule for colloidal crystals of DNA-functionalized particles. *Phys. Rev. Lett.* **107**, 045902 (2011).
59. Macfarlane, R. J., O'Brien, M. N., Petrosko, S. H. & Mirkin, C. A. Nucleic acid-modified nanostructures as programmable atom equivalents: forging a new 'table of elements'. *Angew. Chem. Int. Ed.* **52**, 5688–5698 (2013).
60. McGinley, J. T., Wang, Y., Jenkins, I. C., Sinno, T. & Crocker, J. C. Crystal-templated colloidal clusters exhibit directional DNA interactions. *ACS Nano* **9**, 10817–10825 (2015).
61. Wang, Y. et al. Crystallization of DNA-coated colloids. *Nat. Commun.* **6**, 7253 (2015).
62. van der Meulen, S. A. J. & Leunissen, M. E. Solid colloids with surface-mobile DNA linkers. *J. Am. Chem. Soc.* **135**, 15129–15134 (2013).
63. Angioletti-Uberti, S., Mognetti, B. M. & Frenkel, D. Theory and simulation of DNA-coated colloids: a guide for rational design. *Phys. Chem. Chem. Phys.* **18**, 6373–6393 (2016).
64. Ouldridge, T. E., Louis, A. A. & Doye, J. P. K. Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model. *J. Chem. Phys.* **134**, 085101 (2011).
65. Li, T. I. N. G., Sknepnek, R., Macfarlane, R. J., Mirkin, C. A. & Olvera de la Cruz, M. Modeling the crystallization of spherical nucleic acid nanoparticle conjugates with molecular dynamics simulations. *Nano Lett.* **12**, 2509–2514 (2012).
66. Hinckley, D. M., Freeman, G. S., Whitmer, J. K. & de Pablo, J. J. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: structure. *J. Chem. Phys.* **139**, 144903 (2013).
67. Markegard, C. B., Gallivan, C. P., Cheng, D. D. & Nguyen, H. D. Effects of concentration and temperature on DNA hybridization by two closely related sequences via large-scale coarse-grained simulations. *J. Phys. Chem. B* **120**, 7795–7806 (2016).
68. Fong, L.-K., Wang, Z., Schatz, G. C., Luijten, E. & Mirkin, C. A. The role of structural enthalpy in spherical nucleic acid hybridization. *J. Am. Chem. Soc.* **140**, 6226–6230 (2018).
69. Girard, M. et al. Particle analogs of electrons in colloidal crystals. *Science* **364**, 1174–1178 (2019).
70. Hynninen, A.-P., Christova, C., van Roij, R., van Blaaderen, A. & Dijkstra, M. Prediction and observation of crystal structures of oppositely charged colloids. *Phys. Rev. Lett.* **96**, 138308 (2006).
71. Fornleitner, J., LoVerso, F., Kahl, G. & Likos, C. N. Genetic algorithms predict formation of exotic ordered configurations for two-component dipolar monolayers. *Soft Matter* **4**, 480–484 (2008).
72. Bianchi, E., Doppelbauer, G., Filion, L., Dijkstra, M. & Kahl, G. Predicting patchy particle crystals: variable box shape simulations and evolutionary algorithms. *J. Chem. Phys.* **136**, 214102 (2012).
73. Fornleitner, J. & Kahl, G. Lane formation vs. cluster formation in two-dimensional square-shoulder systems—a genetic algorithm approach. *Europhys. Lett.* **82**, 18001 (2008).
74. Stucke, D. P. & Crespi, V. H. Predictions of new crystalline states for assemblies of nanoparticles: perovskite analogues and 3-D arrays of self-assembled nanowires. *Nano Lett.* **3**, 1183–1186 (2003).
75. Filion, L. et al. Efficient method for predicting crystal structures at finite temperature: variable box shape simulations. *Phys. Rev. Lett.* **103**, 188302 (2009).
76. Haji-Akbari, A. et al. Disordered, quasicrystalline and crystalline phases of densely packed tetrahedra. *Nature* **462**, 773–777 (2009).
77. de Graaf, J., Filion, L., Marechal, M., van Roij, R. & Dijkstra, M. Crystal-structure prediction via the floppy-box Monte Carlo algorithm: method and application to hard (non)convex particles. *J. Chem. Phys.* **137**, 214101 (2012).
78. Ladd, A. & Woodcock, L. Interfacial and co-existence properties of the Lennard-Jones system at the triple point. *Mol. Phys.* **36**, 611–619 (1978).
79. Kofke, D. A. Gibbs–Duhem integration: a new method for direct evaluation of phase coexistence by molecular simulation. *Mol. Phys.* **78**, 1331–1336 (1993).
80. Bolhuis, P. G. & Kofke, D. A. Monte Carlo study of freezing of polydisperse hard spheres. *Phys. Rev. E* **54**, 634 (1996).
81. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J. Comp. Phys.* **23**, 187–199 (1977).
82. Allen, R. J., Frenkel, D. & ten Wolde, P. R. Simulating rare events in equilibrium or nonequilibrium stochastic systems. *J. Chem. Phys.* **124**, 024102 (2006).
83. Dellago, C., Bolhuis, P. G., Csajka, F. S. & Chandler, D. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **108**, 1964–1977 (1998).
84. Earl, D. J. & Deem, M. W. Parallel tempering: theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **7**, 3910–3916 (2005).
85. Dress, C. & Krauth, W. Cluster algorithm for hard spheres and related systems. *J. Phys. A* **28**, L597–L601 (1995).
86. Heringa, J. R. & Blöte, H. W. J. Geometric cluster Monte Carlo simulation. *Phys. Rev. E* **57**, 4976–4978 (1998).
87. Whitelam, S. & Geissler, P. L. Avoiding unphysical kinetic traps in Monte Carlo simulations of strongly attractive particles. *J. Chem. Phys.* **127**, 154101 (2007).
88. Liu, J., Wilding, N. B. & Luijten, E. Simulation of phase transitions in highly asymmetric fluid mixtures. *Phys. Rev. Lett.* **97**, 115705 (2006).
89. Sinkovits, D. W., Barr, S. A. & Luijten, E. Rejection-free Monte Carlo scheme for anisotropic particles. *J. Chem. Phys.* **136**, 144111 (2012).
90. Bernard, E. P., Krauth, W. & Wilson, D. B. Event-chain Monte Carlo algorithms for hard-sphere systems. *Phys. Rev. E* **80**, 056704 (2009).
91. Michel, M., Kapfer, S. C. & Krauth, W. Generalized event-chain Monte Carlo: constructing rejection-free global-balance algorithms from infinitesimal steps. *J. Chem. Phys.* **140**, 054116 (2014).
92. Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B* **28**, 784 (1983).
93. van Meel, J. A., Filion, L., Valeriani, C. & Frenkel, D. A parameter-free, solid-angle based, nearest-neighbor algorithm. *J. Chem. Phys.* **136**, 234107 (2012).
94. Lechner, W. & Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *J. Chem. Phys.* **129**, 114707 (2008).
95. Mickel, W., Kapfer, S. C., Schröder-Turk, G. E. & Mecke, K. Shortcomings of the bond orientational order parameters for the analysis of disordered particulate matter. *J. Chem. Phys.* **138**, 044501 (2013).
96. Auer, S. & Frenkel, D. Prediction of absolute crystal-nucleation rate in hard-sphere colloids. *Nature* **409**, 1020–1023 (2001).
97. Malins, A., Williams, S. R., Eggers, J. & Royall, C. P. Identification of structure in condensed matter with the topological cluster classification. *J. Chem. Phys.* **139**, 234506 (2013).
98. Gantapara, A. P., de Graaf, J., van Roij, R. & Dijkstra, M. Phase diagram and structural diversity of a family of truncated cubes: degenerate close-packed structures and vacancy-rich states. *Phys. Rev. Lett.* **111**, 015501 (2013).
99. Klotsa, D., Chen, E. R., Engel, M. & Glotzer, S. C. Intermediate crystalline structures of colloids in shape space. *Soft Matter* **14**, 8692–8697 (2018).
100. Geiger, P. & Dellago, C. Neural networks for local structure detection in polymorphic systems. *J. Chem. Phys.* **139**, 164105 (2013).
101. Dietz, C., Kretz, T. & Thoma, M. Machine-learning approach for local classification of crystalline structures in multiphase systems. *Phys. Rev. E* **96**, 011301 (2017).
102. Boattini, E., Ram, M., Smallenburg, F. & Filion, L. Neural-network-based order parameters for classification of binary hard-sphere crystal structures. *Mol. Phys.* **116**, 3066–3075 (2018).

103. DeFever, R. S., Targonski, C., Hall, S. W., Smith, M. C. & Sarupria, S. A generalized deep learning approach for local structure identification in molecular simulations. *Chem. Sci.* **10**, 7503–7515 (2019).
104. Terao, T. A machine learning approach to analyze the structural formation of soft matter via image recognition. *Soft Mater.* **18**, 215–227 (2020).
105. Schoenholz, S. S., Cubuk, E. D., Sussman, D. M., Kaxiras, E. & Liu, A. J. A structural approach to relaxation in glassy liquids. *Nat. Phys.* **12**, 469–471 (2016).
106. Bapst, V. et al. Unveiling the predictive power of static structure in glassy systems. *Nat. Phys.* **16**, 448–454 (2020).
107. Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
108. 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).
109. Dai, C. & Glotzer, S. C. Efficient phase diagram sampling by active learning. *J. Phys. Chem. B* **124**, 1275–1284 (2020).
110. Reinhart, W. F., Long, A. W., Howard, M. P., Ferguson, A. L. & Panagiotopoulos, A. Z. Machine learning for autonomous crystal structure identification. *Soft Matter* **13**, 4733–4745 (2017).
111. Reinhart, W. F. & Panagiotopoulos, A. Z. Automated crystal characterization with a fast neighborhood graph analysis method. *Soft Matter* **14**, 6083–6089 (2018).
112. Jadrich, R., Lindquist, B. & Truskett, T. Unsupervised machine learning for detection of phase transitions in off-lattice systems. I. Foundations. *J. Chem. Phys.* **149**, 194109 (2018).
113. Jadrich, R., Lindquist, B., Piñeros, W., Banerjee, D. & Truskett, T. Unsupervised machine learning for detection of phase transitions in off-lattice systems. II. Applications. *J. Chem. Phys.* **149**, 194110 (2018).
114. Spellings, M. & Glotzer, S. C. Machine learning for crystal identification and discovery. *AIChE J.* **64**, 2198–2206 (2018).
115. Boattini, E., Dijkstra, M. & Filion, L. Unsupervised learning for local structure detection in colloidal systems. *J. Chem. Phys.* **151**, 154901 (2019).
116. Adorf, C. S., Moore, T. C., Melle, Y. J. & Glotzer, S. C. Analysis of self-assembly pathways with unsupervised machine learning algorithms. *J. Phys. Chem. B* **124**, 69–78 (2019).
117. Bereau, T., Andrienko, D. & Kremer, K. Research update: Computational materials discovery in soft matter. *APL Mater.* **4**, 053101 (2016).
118. Ferguson, A. L. Machine learning and data science in soft materials engineering. *J. Phys.: Condens. Matter* **30**, 043002 (2017).
119. Wang, J. & Ferguson, A. Nonlinear machine learning in simulations of soft and biological materials. *Mol. Simul.* **44**, 1090–1107 (2018).
120. Torquato, S. Inverse optimization techniques for targeted self-assembly. *Soft Matter* **5**, 1157–1173 (2009).
121. Lindquist, B. A., Jadrich, R. B. & Truskett, T. M. Communication: Inverse design for self-assembly via on-the-fly optimization. *J. Chem. Phys.* **145**, 11110 (2016).
122. Shell, M. S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **129**, 144108 (2008).
123. Piñeros, W. D., Lindquist, B. A., Jadrich, R. B. & Truskett, T. M. Inverse design of multicomponent assemblies. *J. Chem. Phys.* **148**, 104509 (2018).
124. Lindquist, B. A., Jadrich, R. B., Piñeros, W. D. & Truskett, T. M. Inverse design of self-assembling Frank–Kasper phases and insights into emergent quasicrystals. *J. Phys. Chem. B* **122**, 5547–5556 (2018).
125. Florescu, M., Torquato, S. & Steinhardt, P. J. Designer disordered materials with large, complete photonic band gaps. *Proc. Natl Acad. Sci. USA* **106**, 20658–20663 (2009).
126. Geng, Y., van Anders, G., Dodd, P. M., Dshemuchadse, J. & Glotzer, S. C. Engineering entropy for the inverse design of colloidal crystals from hard shapes. *Sci. Adv.* **5**, eaaw0514 (2019).
127. Miskin, M. Z., Khaira, G., de Pablo, J. J. & Jaeger, H. M. Turning statistical physics models into materials design engines. *Proc. Natl Acad. Sci. USA* **113**, 34–39 (2016).
128. Kumar, R., Coli, G. M., Dijkstra, M. & Sastry, S. Inverse design of charged colloidal particle interactions for self assembly into specified crystal structures. *J. Chem. Phys.* **151**, 084109 (2019).
129. Long, A. W. & Ferguson, A. L. Rational design of patchy colloids via landscape engineering. *Mol. Syst. Des. Eng.* **3**, 49–65 (2018).
130. Ma, Y. & Ferguson, A. L. Inverse design of self-assembling colloidal crystals with omnidirectional photonic bandgaps. *Soft Matter* **15**, 8808–8826 (2019).
131. Sherman, Z. M., Howard, M. P., Lindquist, B. A., Jadrich, R. B. & Truskett, T. M. Inverse methods for design of soft materials. *J. Chem. Phys.* **152**, 140902 (2020).
132. Ou, Z., Wang, Z., Luo, B., Luijten, E. & Chen, Q. Kinetic pathways of crystallization at the nanoscale. *Nat. Mater.* **19**, 450–455 (2020).
133. Wang, J. et al. Magic number colloidal clusters as minimum free energy structures. *Nat. Commun.* **9**, 5259 (2018).
134. 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).
135. Chen, Q. et al. Supracolloidal reaction kinetics of Janus spheres. *Science* **331**, 199–202 (2011).
136. Haji-Akbari, A. et al. Disordered, quasicrystalline and crystalline phases of densely packed tetrahedra. *Nature* **462**, 773–777 (2009).
137. Ramananarivo, S., Ducrot, E. & Palacci, J. Activity-controlled annealing of colloidal monolayers. *Nat. Commun.* **10**, 3380 (2019).
138. Sharp, T. A. et al. Machine learning determination of atomic dynamics at grain boundaries. *Proc. Natl Acad. Sci. USA* **115**, 10943–10947 (2018).
139. Chen, W., Tan, A. R. & Ferguson, A. L. Collective variable discovery and enhanced sampling using autoencoders: innovations in network architecture and error function design. *J. Chem. Phys.* **149**, 072312 (2018).
140. Gan, Z. & Xu, Z. Multiple-image treatment of induced charges in Monte Carlo simulations of electrolytes near a spherical dielectric interface. *Phys. Rev. E* **84**, 016705 (2011).
141. Freed, K. F. Perturbative many-body expansion for electrostatic energy and field for system of polarizable charged spherical ions in a dielectric medium. *J. Chem. Phys.* **141**, 034115 (2014).
142. Qin, J., de Pablo, J. J. & Freed, K. F. Image method for induced surface charge from many-body system of dielectric spheres. *J. Chem. Phys.* **145**, 124903 (2016).
143. Maggs, A. & Rossetto, V. Local simulation algorithms for Coulomb interactions. *Phys. Rev. Lett.* **88**, 196402 (2002).
144. Levitt, D. G. Electrostatic calculations for an ion channel. I. Energy and potential profiles and interactions between ions. *Biophys. J.* **22**, 209–219 (1978).
145. Hoshi, H., Sakurai, M., Inoue, Y. & Chûjô, R. Medium effects on the molecular electronic structure. I. The formulation of a theory for the estimation of a molecular electronic structure surrounded by an anisotropic medium. *J. Chem. Phys.* **87**, 1107–1115 (1987).
146. Bharadwaj, R., Windemuth, A., Sridharan, S., Honig, B. & Nicholls, A. The fast multipole boundary element method for molecular electrostatics: an optimal approach for large systems. *J. Comput. Chem.* **16**, 898–913 (1995).
147. Allen, R., Hansen, J.-P. & Melchionna, S. Electrostatic potential inside ionic solutions confined by dielectrics: a variational approach. *Phys. Chem. Chem. Phys.* **3**, 4177–4186 (2001).
148. Boda, D., Gillespie, D., Eisenberg, B., Nonner W., & Henderson, D. in *Ionic Soft Matter: Modern Trends in Theory and Applications* (eds Henderson, D. et al.) 19–43 (NATO Science Series II: Mathematics, Physics and Chemistry Vol. 206, Springer, 2005).
149. Tyagi, S. et al. An iterative, fast, linear-scaling method for computing induced charges on arbitrary dielectric boundaries. *Phys. Chem. Chem. Phys.* **3**, 4177–4186 (2001).
150. Jadhao, V., Solis, F. J. & Olvera de la Cruz, M. Simulation of charged systems in heterogeneous dielectric media via a true energy functional. *Phys. Rev. Lett.* **109**, 223905 (2012).
151. Barros, K., Sinkovits, D. & Luijten, E. Efficient and accurate simulation of dynamic dielectric objects. *J. Chem. Phys.* **140**, 064903 (2014).
152. Barros, K. & Luijten, E. Dielectric effects in the self-assembly of binary colloidal aggregates. *Phys. Rev. Lett.* **113**, 017801 (2014).
153. Gan, Z., Wang, Z., Jiang, S., Xu, Z. & Luijten, E. Efficient dynamic simulations of charged dielectric colloids through a novel hybrid method. *J. Chem. Phys.* **151**, 024112 (2019).
154. Holland, J. H. *Adaptation in Natural and Artificial Systems* (MIT Press, 1992).
155. Oganov, A. R. & Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *J. Chem. Phys.* **124**, 244704 (2006).
156. Wang, Y., Lv, J., Zhu, L. & Ma, Y. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B* **82**, 094116 (2010).
157. Amsler, M. & Goedecker, S. Crystal structure prediction using the minima hopping method. *J. Chem. Phys.* **133**, 224104 (2010).
158. Wales, D. J. & Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science* **285**, 1368–1372 (1999).
159. Martoňák, R. et al. Simulation of structural phase transitions by metadynamics. *Z. Kristallogr. Cryst. Mater.* **220**, 489–498 (2009).
160. Panagiotopoulos, A. Z. Direct determination of phase coexistence properties of fluids by Monte Carlo simulation in a new ensemble. *Mol. Phys.* **61**, 813–826 (1987).
161. Ferrenberg, A. M. & Swendsen, R. H. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* **63**, 1195–1198 (1989).
162. Potoff, J. J. & Panagiotopoulos, A. Z. Surface tension of the three-dimensional Lennard-Jones fluid from histogram-reweighting Monte Carlo simulations. *J. Chem. Phys.* **112**, 6411–6415 (2000).

Acknowledgements

We thank Z. Wang for designing the figures. E.L. is supported by the Center for Bio-Inspired Energy Science (CBES), an Energy Frontier Research Center funded by the US Department of Energy, Office of Science, Basic Energy Sciences BES, under award no. DE-SC0000989.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to M.D. or E.L.

Peer review information *Nature Materials* thanks Emanuela Zaccarelli and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.