THE JOURNAL OF PHYSICAL CHEMISTRY A

pubs.acs.org/JPCA Article

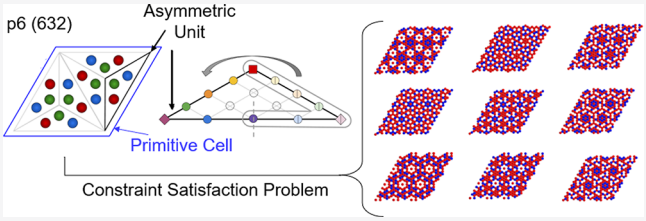
Symmetry-Based Crystal Structure Enumeration in Two Dimensions

Evan Pretti, Vincent K. Shen, Jeetain Mittal, and Nathan A. Mahynski*

Cite This: J. Phys. Chem. A 2020, 124, 3276–3285 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The accurate prediction of stable crystalline phases is a long-standing problem encountered in the study of conventional atomic and molecular solids as well as soft materials. One possible solution involves enumerating a reasonable set of candidate structures and then screening them to identify the one(s) with the lowest (free) energy. Candidate structures in this set can also serve as starting points for other routines, such as genetic algorithms, which search *via* optimization. Here, we present a framework for crystal structure enumeration of two-dimensional systems that utilizes a combination of symmetry- and stoichiometry-imposed constraints to compute valid configurations of particles that tile Euclidean space. With mild assumptions, this produces a computationally tractable total number of proposed candidates, enabling multicomponent systems to be screened by direct enumeration of possible crystalline ground states. The python code that enables these calculations is available at https://github.com/usnistgov/PACCS.



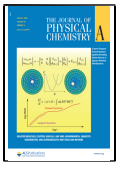
1. INTRODUCTION

Crystal structure prediction is a critically important but particularly challenging problem for understanding and designing ordered, periodic materials.[1−4] The problem is relatively simple to state: given the composition of a mixture and set of interaction potentials for its constituents, what is the thermodynamically stable periodic arrangement? Such a global optimization problem is difficult to solve due to the high dimensionality of the search (configurational) space and the potential for a large number of local minima, corresponding to metastable states, that exist in the free-energy landscape.[5,6] Different challenges can arise depending on the nature of the system in question: for nanoscale atomic and molecular systems where density functional theory (DFT) is often employed, evaluating the energy for a given configuration can incur a significant computational expense. For mesoscale colloidal systems, on the other hand, this is relatively inexpensive, but such systems may have features such as large size asymmetries between components,[7,8] short-range and discontinuous (hard-wall) interactions,[9,10] and anisotropy due to asphericity and directionality.[11−14] These can ultimately lead to complex energy landscapes with large regions of high energy and sharp local minima, making optimization particularly challenging.

A variety of methods have been developed to overcome this problem. Common techniques include random structure searching,[15] various Monte Carlo methods,[5,10,13,16,17] as well as evolutionary algorithms that explore configurational space by making mutations and combining structures from a population to yield new candidates.[9,18−27] Recent advances in computer hardware have also enabled a plethora of machine learning approaches.[28−33] Due to the typically nonconvex nature of the optimization problem, good initial guesses are generally required; often, these are generated at random or based on chemical, directional, or symmetry-based constraints.[15,27] Symmetry serves as a particularly powerful and general method for creating candidate structures, due to its inherent presence in the large majority of known crystalline systems.[15,34]

Methods based on directly generating candidates constrained by geometry and symmetry have been employed for a variety of chemical systems.[15,35−40] Perhaps the simplest methodology is to use space groups to fill space by applying matrix operations to the interior of a fundamental domain (FD), or direct-space asymmetric unit, containing particles whose positions are randomly assigned. This domain is the smallest region of a space that fills it without defects after operations determined by the group.[41] However, another way of representing symmetry is by considering the nature of the boundaries of the fundamental domain and how, by virtue of the operations defined by a given group, different edges and vertices will be mapped to their symmetrically equivalent sites at other locations on the boundary of the same domain. In two dimensions, this information is compactly represented by its orbifold ("orbit-manifold"),[42−44] which is a surface formed by collapsing each symmetrically equivalent location in a crystal (an "orbit") to a single point.

Received: January 30, 2020
Revised: March 9, 2020
Published: March 16, 2020



https://dx.doi.org/10.1021/acs.jpca.0c00846
J. Phys. Chem. A 2020, 124, 3276–3285

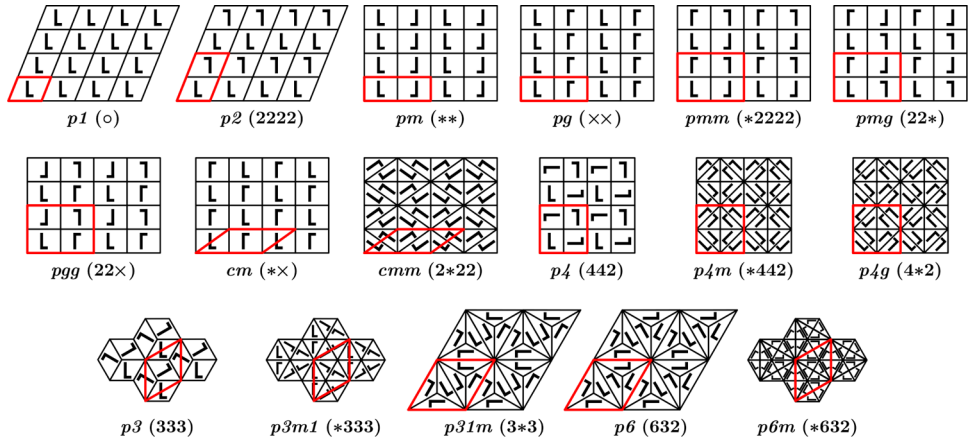


Figure 1. The seventeen wallpaper groups. Each group is listed with its Hermann−Mauguin symbol (*e.g.*, *p4m*) followed in parentheses by its orbifold signature (*e.g.*, *442). Any pattern can be placed in the interior of the fundamental domain (traced in black) and tiled throughout space by the groups; the "L" motif depicted is for illustrative purposes only. Example primitive cells, containing the minimum possible number of fundamental domains, are shown in red for each group.

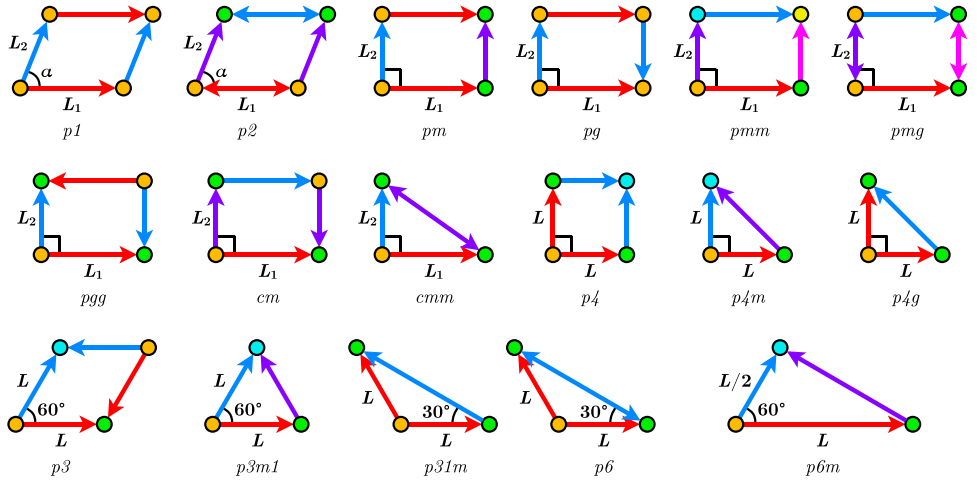


Figure 2. Symmetric edges and corners of the fundamental domains. Variable parameters α and L_1, L_2, or L are shown to illustrate the constraints on the geometries of the domains we have chosen. For a given domain, vertices of the same colors will overlap when the domain tiles space, and must therefore be identical. Likewise, edges of the same colors will overlap. Different colors imply independence. A single-headed arrow → indicates the direction along which distinct edges will be overlaid, while a double-headed arrow ↔ indicates that the given edge will overlay itself in reverse and must be symmetric about its midpoint (rotations of 180° occur there).

In essence, by focusing on determining equivalent sites along the boundaries of the fundamental domain, we develop an algorithm that enables us to combine this symmetry information with stoichiometry to create a constraint satisfaction problem (CSP) whose solutions may be easily enumerated. This is effectively a manifestation of the multiplicity of Wyckoff positions, which reflect the number of times a particle placed at such a position is repeated throughout the overall crystal.[41] Each CSP solution represents a crystal structure, enabling an extensive and diverse ensemble of periodic lattices to be enumerated. When a Hamiltonian is introduced so that (free) energies of candidates may be computed, this symmetry-based generation is a powerful technique for predicting stable crystal structures.[45] This method is particularly useful for systems whose interaction potentials are isotropic and moderate to short-ranged; this describes many colloidal systems, which are the focus of our efforts, but the method is not exclusive. We have demonstrated the utility of such an approach previously;[45] here we describe the inner workings of this method in detail and provide a code to enable these calculations.[46]

2. SYMMETRY-BASED CRYSTAL STRUCTURE ENUMERATION

2.1. Wallpaper Groups. Symmetry groups in two- or three-dimensional Euclidean space represent combinations of transformations, such as translations, rotations, and reflections, on some subset of the space, known as a fundamental domain (FD), which map the domain onto an image of itself.[41,47] In three dimensions these are the well-known crystallographic space groups, of which there are exactly 230,[43] and in two dimensions they are the analogous wallpaper groups (or plane symmetry groups), of which there are exactly 17.[41,44] The specific symmetries associated with each group dictate allowable shapes for its corresponding FD, and when symmetry operations are applied to its contents a translationally periodic structure will be produced. For periodic crystalline systems, a set of connected FDs known as a primitive (unit) cell can be identified, which will tile all of the space through translations

https://dx.doi.org/10.1021/acs.jpca.0c00846
J. Phys. Chem. A 2020, 124, 3276−3285

alone.[47] Figure 1 shows all wallpaper groups, demonstrating the actions of their symmetries on fundamental domains (black) and highlighting the primitive cells (red) of the resulting periodic tilings. There is not necessarily a unique shape to each group's FD,[41] and Figure 1 illustrates our choices. Throughout this work, we refer to the groups by their Hermann–Mauguin symbols listed in Figure 1. Also shown are the corresponding orbifold signatures of the groups.[44]

There is a substantial advantage of using symmetry groups for generating candidate crystal structures. For a two-dimensional lattice with $L \times L$ sites each with M possible states, there are a total of $O(M^{L^2})$ possibilities. This corresponds to the case of only restricting the allowable symmetries to those created by the group $p1$, for which the primitive cell is composed of a single fundamental domain and only translational symmetry is imposed. Note that this cell is typically employed in the computer simulation of fluids to approximate nonperiodic systems.[48] For the other wallpaper groups, there will in general be $1 < k \leq 12$ fundamental domains per primitive cell, where $p6m$ represents the maximum.[47] Since we need only consider the independent portion of the cell, sampling from such a group produces only $O(M^{L^2/k})$ configurations, corresponding to a reduction by a factor of $O(M^{L^2[1-1/k]})$.

Figure 2 shows the fundamental domains we have chosen for all of the groups, in which we have translated the constraints of the symmetry into requirements that must be met at the edges and corners of each domain. Here, similarly colored edges or points are symmetrically equivalent. This is the first step in defining the constraint satisfaction problem.

The International Tables for Crystallography provide standardized information on the shape of fundamental domains and generators to produce a unit cell. However, it is well known that the specifications in these tables are inexact precisely at the borders, requiring additional boundary-specific conditions to remove redundant coordinates.[49,50] In three dimensions, previous work has sought to define exact versions of the fundamental domains, consistent with the International Tables;[50] to circumvent this problem in two dimensions, we have elected to create our own fundamental domains, consistent with, but not identical to, those in the tables. Section 2.2 reviews our approach to creating domains in a consistent fashion across wallpaper groups that allows us to create a regular grid of points through the fundamental domain corresponding to different Wyckoff positions such that these redundancies are trivial to locate computationally. We note that screening for crystal candidates within select space groups with only certain desired Wyckoff positions has been successful at searching for optimal sphere packings and computing phase diagrams of purely repulsive systems;[40,51] since we are not focused on a specific class of systems, our algorithm searches all groups with all occupancies of all Wyckoff positions to remain as general as possible.

2.2. Discretizing Fundamental Domains. To use symmetry groups to generate candidate colloidal crystal structures, a method for placing individual particles over the FDs must be devised. Discretization of the FD into a lattice creates a finite and enumerable set of configurations. Although restrictive, generated candidate structures can later be optimized in continuum space, if desired. Furthermore, when done carefully, lattices have the advantage of easily providing nodes at special Wyckoff positions (always located at edges and vertices of FDs) and allowing symmetric equivalence between nodes to be determined. Only a certain fraction of each node will be contained within the FD depending on whether the node is located at a corner or edge, or is on the face of the domain. This enables us to introduce the constraints of stoichiometry, which represents the second half of the formulation of our method, discussed in more detail in Section 2.3. This is analogous to using the known "multiplicities" of Wyckoff positions, which are given as integers in the International Tables for Crystallography, though these positions are not discretized into nodes therein.[41]

Figure 3 shows examples of how we discretize fundamental domains using a "parallel-line construction" to exactly intersect

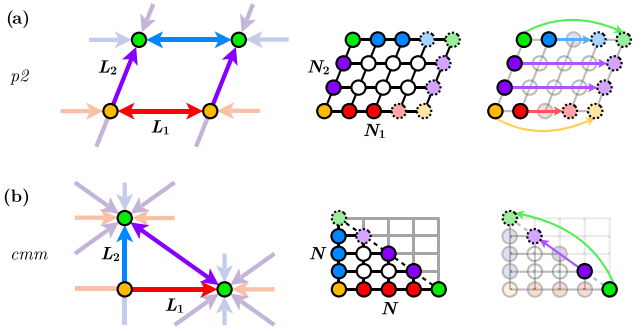


Figure 3. Examples demonstrating discretization for the wallpaper groups (a) $p2$ and (b) cmm. FDs are shown on the left with their boundary symmetry restrictions. On the right, lattice sites are shown with their symmetrically equivalent sites in dashed outlines, more explicitly demonstrated with colored arrows.

all boundaries of the FD. As previously illustrated in Figure 2, all domains produced by this method can be classified as either parallelograms or triangles.[47] For a parallelogram, a regular lattice can simply be laid down by creating sets of lines parallel to the sides of the domain and taking their intersections as nodes (*cf.* Figure 3a). Furthermore, any triangle may be represented as half of a parallelogram divided along its diagonal, provided the parallel-line spacing is such that nodes intersect the diagonal (*cf.* Figure 3b). Therefore, we may use a parallelogram motif to systematically generate a grid of lattice nodes for all groups. Note that groups $p3$, $p3m1$, $p31m$, $p6$, and $p6m$ exist on hexagonal lattices (based on equilateral triangles), while the remaining groups have lattices that are some subset of parallelograms not derived from triangles (including rectangle, rhombus, and square).[47] These correspond to the five Bravais lattices that exist in two dimensions. This parallel-line construction algorithm has the benefit that it works for all groups and can generate lattice points consistent with the underlying symmetry requirements when fundamental domains are chosen as in Figure 1.

Finally, for groups with domains having edges that must be symmetric about their own midpoints (*i.e.*, groups with double-headed arrows in Figure 2: $p2$, pmg, cmm, and $p6$), a central symmetrically unique site will appear on such an edge if and only if the number of sites along the edge is odd. These midpoints are, by construction, where a twofold center of rotation (180°) occurs, which is the source of that symmetry.

2.3. Solving the Constrained Symmetry and Stoichiometry Problem. Information regarding both symmetry and stoichiometry manifests on each group's FD *via* symmetrically equivalent positions and their multiplicity. Thus, we

https://dx.doi.org/10.1021/acs.jpca.0c00846
J. Phys. Chem. A 2020, 124, 3276−3285

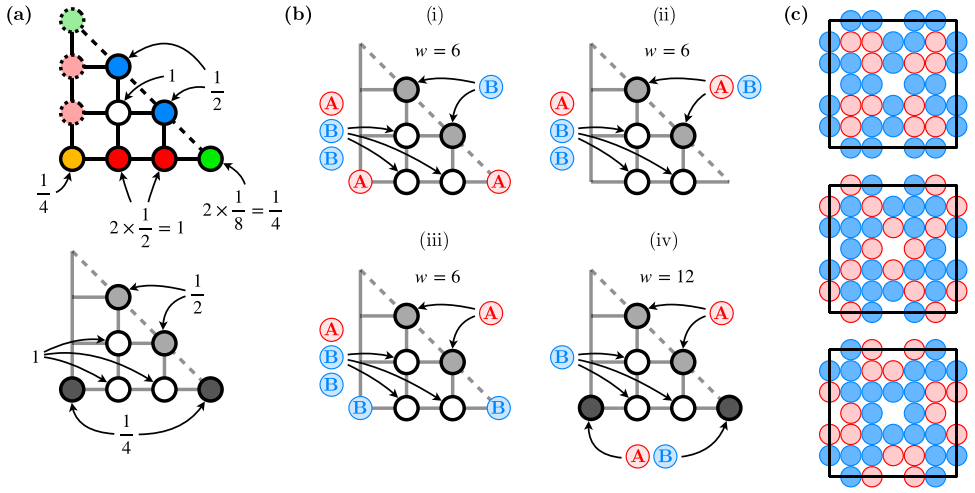


Figure 4. Generation of candidates from *p*4*g* with a 4 × 4 lattice. (a) Contribution of a particle on each site to the fundamental domain is determined by the number of domains that share each site and the number of symmetrically equivalent sites. Sites are classified based on their stoichiometric contributions. (b) All (four) solutions to the constraint satisfaction problem for an A_3B_3 stoichiometry, each with the number of realizations w for placing particles on sites. (c) Three unique structures obtainable from solution (ii); although $w = 6$, the other unit cells are superimposable mirror images (achiral) of those shown.

can combine symmetry and stoichiometry requirements to formulate a relatively simple constraint satisfaction problem (CSP).[52] That is, we may formulate the search for different candidate lattices as a solution that satisfies (1) the symmetry of the desired wallpaper group and (2) the ratio of the different components that we wish the lattice to contain.

To demonstrate this idea concretely, Figure 4 presents a complete example for the *p*4*g* group. Here, a right triangular FD possesses reflection symmetry across its hypotenuse and fourfold rotational symmetry about its right angle. As shown in Figure 4a, a lattice with $N = N_1 = N_2 = 4$ yields a total of 10 sites, 3 of which are redundant due to symmetry (2 red, 1 green). To understand the influence of stoichiometry on how particles can be placed on this lattice, we can consider each kind of site in turn. A particle placed on the internal (white, the general Wyckoff) site will contribute all of itself to the domain. However, a particle placed on a reflected edge (blue, a special Wyckoff) site will be shared between the domain and its reflected image, thus contributing only 1/2 of the particle to a single domain. Similarly, the orange corner site has a contribution of only 1/4. Analogously, in the International Tables, the multiplicities of these sites are reported as 8 (white), 4 (blue), and 2 (orange).[41] Although each node along one of the rotated edge (red) sites is split across two neighboring domains, a particle placed on such a site will be replicated on its symmetrically equivalent site along the adjacent edge. Thus, after accounting for this symmetry, a total contribution of 1 results. Similarly, the green corner site is shared by eight domains (see Figure 1) but has a symmetrically equivalent site, leading to a net contribution of 1/4.

Suppose n_{ij} represents the number of particles of type i to place on sites of type j, c_j is the stoichiometric contribution for each site of type j, m_j is the number of sites of type j, and d_i is the desired stoichiometric coefficient for a particle of type i in the final generated structures. Then we wish to find all sets of n_{ij} satisfying

$$0 \le \sum_i n_{ij} \le m_j \tag{1}$$

for all j, and

$$\frac{\sum_j c_j n_{ij}}{\sum_k \sum_j c_j n_{kj}} = \frac{d_i}{\sum_k d_k} \tag{2}$$

for all i. Additional constraints can be added as desired to, *e.g.*, limit the total number of particles placed on sites. Returning to our specific example, we will now look for solutions for a binary system of components A (type 1) and B (type 2) with stoichiometry given by A_3B_3. From eq 1, we obtain

$$\begin{cases} 0 \le n_{11} + n_{21} \le 3 \\ 0 \le n_{12} + n_{22} \le 2 \\ 0 \le n_{13} + n_{23} \le 2 \end{cases} \tag{3}$$

From eq 2, we can find

$$\begin{cases} 1n_{11} + \frac{1}{2}n_{12} + \frac{1}{4}n_{13} = \frac{3}{3+5}X = Y \\ 1n_{21} + \frac{1}{2}n_{22} + \frac{1}{4}n_{23} = \frac{5}{3+5}X = Z \end{cases} \tag{4}$$

where $X = \sum_k \sum_j c_j n_{kj}$, the denominator in eq 2. Thus, eq 4 is underspecified and admits multiple solutions when constrained by eq 3 and the fact that all n_{ij} must be integers.

There exist various computational approaches to solving such CSPs;[52] we have employed a backtracking approach.[53] In the end, there are four sets of integer solutions to these constraints, illustrated graphically in Figure 4b. A solution to the CSP simply specifies all n_{kj} values, but when $n_{kj} < m_j$ there are multiple ways to realize this solution. The number of possibilities for a site of type j is

$$w_j = \frac{m_j!}{\left(m_j - \sum_i n_{ij}\right)! \prod_i n_{ij}!} \tag{5}$$

This gives a total number of realizations $w = \prod_j w_j$, which can be generated from a given solution to the constraint satisfaction problem. Each realization corresponds to a configuration satisfying both symmetry and stoichiometric constraints.

https://dx.doi.org/10.1021/acs.jpca.0c00846
J. Phys. Chem. A 2020, 124, 3276−3285

Figure 4b shows all four solutions to the CSP for A_3B_3 in the *p*4*g* group example, along with the number of different combinatorial realizations of each solution (6, 6, 6, and 12 for a total of 30 solutions). However, it is important to note that the number of unique crystal structures that can be obtained from a given CSP solution may be less than its corresponding value of w. For example, solution (ii) has $w = 6$ but yields only the three distinct structures whose unit cells are depicted in Figure 4c. In this case, the others not shown have unit cells, which are superimposable mirror images of those displayed, *i.e.*, the unit cell has achiral solutions and both enantiomorphs represent solutions to the CSP though they do not represent different structures. This is addressed briefly in Section 4 and more thoroughly in the Supporting Information (SI).

3. CANDIDATE STRUCTURE GENERATION

We now have a method for systematically generating crystalline configurations by specifying only three things: the number of nodes to use on the edge(s) of a fundamental domain's lattice, N_g, the wallpaper group, and the desired stoichiometry. The CSP simply takes these three inputs and constructs a hierarchical tree, as depicted in Figure 5. The

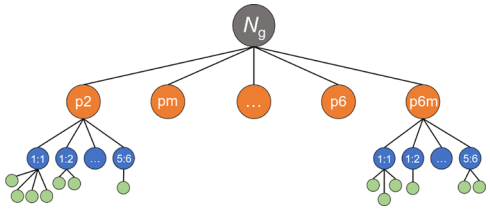


Figure 5. Schematic of the tree generated by the CSP. N_g is first chosen (gray) and then the wallpaper group (orange); these define the lattice. Then, the desired stoichiometry is specified (blue), which fully defines the CSP. The solutions (green) are configurations that can be sampled.

leaves of the tree correspond to realizations of solutions to each CSP defined by the branches upon which the leaves reside. Since there are a finite number of wallpaper groups, and one can often make a reasonable choice for the stoichiometries to consider, there are a tractable number of leaves (configurations) that can be generated. The result is a finite ensemble of crystalline candidates to consider. In principle, a Hamiltonian can then be chosen, and these candidates can be refined and screened as desired,[45] though this step is beyond the scope of this paper.

We highlight the fact that the grid spacing itself between nodes on the lattice has been left arbitrary until this point. In practice, we set this to unity; however, uniform scaling to any nearest-neighbor contact distance is possible as this isotropic scaling does not affect the symmetry. We generally scale a given CSP solution based on some assumed diameter (nominally unity) for different species so that nearest neighbors are in contact with each other. First, this implies that different structures can be generated from the same CSP solution if there are different characteristic length scales of interest. Second, this means that solutions that place particles on nodes spaced far apart may be scaled down and look identical to those configurations that packed them tightly to begin with. This can also be a source of the redundancy summarized in Section 4. Finally, regardless of scaling, "gaps" can develop if the right nodes are not used in the CSP solution, resulting in lattices of disconnected clusters (not in direct contact). For example, this can occur if all edge nodes are neglected, as in the orange curve of Figure 6. This means that this algorithm is capable of generating not only connected lattices but also disconnected cluster phase candidates. Figure 7 also contains examples of this.

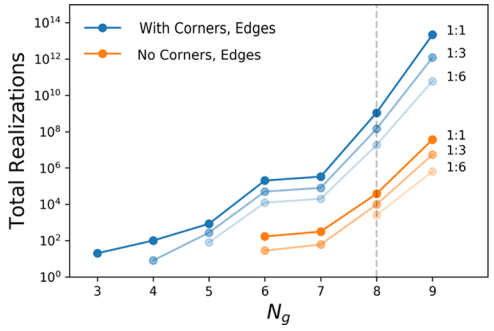


Figure 6. Total realizations, $\sum w$, for the CSP coming from the 16 wallpaper groups employed for various N_g. Various stoichiometries are shown for the cases when nodes located at the edges of the fundamental domains are allowed in the CSP (blue) and when any solution involving them is excluded (orange). The vertical line at $N_g = 8$ denotes the crossover where the number of edge and corner nodes is exceeded by the number of face nodes, averaged across the groups.[45]

3.1. Sampling from Enumerated Solutions.

First, we must define an approach to sample in an even-handed way between groups. Since the number of FDs per primitive cell varies from 1 (*p*1) to 12 (*p6m*), using the same number of lattice points on each group's FD would generate structures with drastically different unit cell sizes. Instead, we choose to query our algorithm, assuming we would like to consider all candidates that have a primitive (unit) cell no larger than some fixed size. Consider a *p*1 cell with N_g nodes along each side; this corresponds to minimal symmetry (typically employed in simulations of fluids[48]), where the fundamental domain is equal to its primitive cell. Taking this as a reference, we would like to find all crystals with more than this "trivial" symmetry that exist up to that primitive cell size; *i.e.*, the red parallelograms for each group (Figure 1) should not exceed this.

In general, we ignore *p*1 in favor of the other 16 groups when generating candidates due to the principle of maximum symmetry.[34,54] This heuristic states that structures with a high symmetry content tend to have either a very high or a very low energy and is a consequence of the structural correlations imposed by symmetry. Starting from this *p*1 reference cell, it is possible to generate structures by systematically placing particles within the cell. However, the number of arrangements undergoes a combinatorial explosion as N_g increases, quickly leading to an intractable number of possibilities.[45] Regardless, within these possibilities exist structures with symmetry that was not imposed *a priori*; arrangements that have, *e.g.*, a mirror plane, or are symmetric by some rotation will be found if the placement of particles is systematic and exhaustive. These structures would have been found directly if another wallpaper group, corresponding to its symmetry, was used as a generator instead. However, the set of *p*1-generated structures will also contain structures with minor variations on these higher symmetry ones that destroy their overall symmetry; *e.g.*, a

https://dx.doi.org/10.1021/acs.jpca.0c00846
J. Phys. Chem. A 2020, 124, 3276−3285

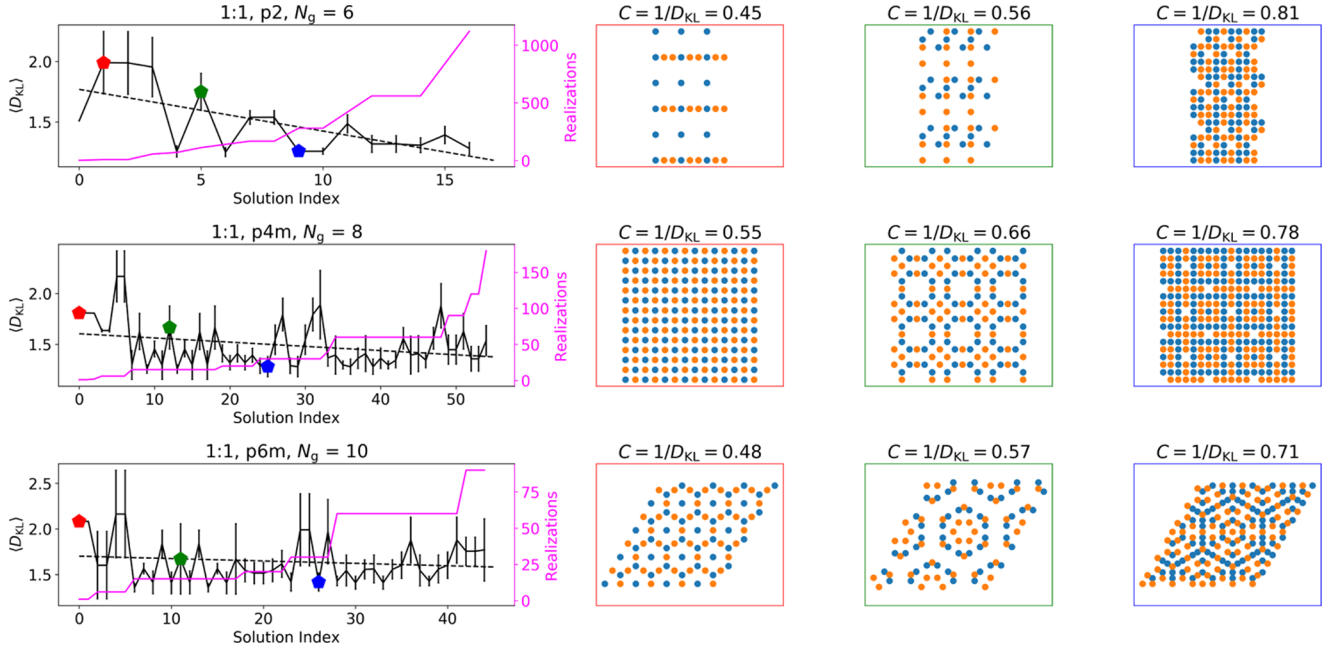


Figure 7. Mean complexity of different CSP solutions as a function of their number of realizations. For three groups, *p2*, *p4m*, and *p6m*, the solutions to their CSP for an equimolar binary system are shown for different selected N_g. Solutions are sorted from the smallest number of combinatorial realizations, *w*, to the largest (magenta), which is the solution index. The mean Kullback−Leibler divergence, $\langle D_{KL} \rangle$, from an ideal gas is also shown in black. Error bars correspond to 1 standard deviation. The dashed black line is a linear fit to $\langle D_{KL} \rangle$. Representative configurations from different solutions are shown on the right; the colored outlines correspond to the colored pentagons depicted on the graphs on the left.

defect in which a single particle is misplaced. As a result, the *p1*-generated structures effectively contain both the high symmetry structures and all of their defective variants. If these defective structures are ignored (by using only the other 16 groups), the reduction of possible arrangements can be on the order of $O(10^{13})$ or greater, making it possible to exhaustively search reasonably sized primitive cells.[45]

If we consider a *p1* cell with N_g^2 total nodes, we can compute the number of nodes along each edge, N_1 and N_2, that must be used on another group to reach the same node density, ρ

$$\rho = \frac{N_g^2}{A_g N_d} = \frac{N_1 N_2 \left(1 - \frac{1}{2}(N_s \bmod 2)\right)}{A_g} \tag{6}$$

where $N_s \in (3,4)$ is the number of sides a group's FD has, A_g is the area of the FD, and N_d is the number of FDs per primitive cell.[45,47] If symmetry constrains the ratio of the length of the sides to be $r = L_2/L_1 = N_2/N_1 \geq 1$, we arrive at (*cf.* SI for more details and caveats)

$$N_1 = \sqrt{\frac{N_g^2}{rN_d\left(1 - \frac{1}{2}(N_s \bmod 2)\right)}} \tag{7}$$

In practice, we generate a grid according to $\lfloor N_1 \rfloor$ and $N_2 = \lfloor rN_1 \rfloor$. The floor operation prevents the number of nodes from exceeding that of the reference *p1* cell. In general, we allow $r \in (1, \sqrt{2}, \sqrt{3}, 2)$, but this can be chosen as desired. Note that *p2* is the only group we sample from that has an unconstrained angle, and we typically sample $\alpha \in (\pi/2, \pi/3, \pi/4, \pi/6)$. A detailed table of allowable *r* and α values is given in the Supplementary Information of ref 45, so it is not reproduced here.

3.1.1. Exhaustive Sampling. Once lattice parameters have been selected, we can solve the CSP and obtain all realizations to each solution. For a given group, if the total number of lattice sites is small or the symmetry and stoichiometry restrictions are significant, the resulting structures can be explored exhaustively. Note that all nodes on the face of an FD belong to the general Wyckoff position and have a stoichiometric contribution factor of 1. The number of ways to place particles (only) there to achieve the desired stoichiometry grows combinatorially with the number of available nodes. Special Wyckoff positions may only occur at the edges and corners, though not all boundary nodes are special positions, and introduce specific factors that highly constrain the CSP. As previously shown, when averaged over all groups, $N_g \approx 8$ corresponds to the point where the total number of nodes on the face equals the total number of edge and corner nodes.[45] We expect that when $N_g < 8$ the CSP is highly constrained, leading to a relatively small number of realizations, which can be tractably enumerated for screening purposes, whereas when $N_g > 8$ the combinatorial explosion of possibilities leads to too many possibilities to reasonably screen [$>O(10^9)$]. Figure 6 shows the number of total realizations from the 16 wallpaper groups we consider for various stoichiometries in a binary mixture as a function of N_g. When edge and corner nodes are excluded (only face nodes allowed), there are three or more orders of magnitude fewer solutions, and, in fact, N_g must be substantially larger for any solutions to exist at all ($N_g \geq 6$) compared to the case when edge and corner nodes are allowed ($N_g \geq 3$ for a 1:1 stoichiometry).

3.1.2. Stochastic Sampling. If the number of total realizations to the CSP (structures) is too large to enumerate completely, we can instead sample stochastically. There are

https://dx.doi.org/10.1021/acs.jpca.0c00846
J. Phys. Chem. A 2020, 124, 3276−3285

different ways in which this could be done, perhaps the simplest being to choose leaves from the tree in Figure 5 at random. However, it is possible to miss important structures by chance; of course, this can be mitigated if these initial candidates are subsequently refined.[45] We propose a heuristic sampling method based on the number of realizations an individual CSP has.

We have empirically observed that, when an individual CSP solution has fewer realizations, it tends to correspond to the use of special Wyckoff positions, *e.g.*, a corner node in the FD, which has a very small, unique stoichiometric factor relative to the other available nodes. Moreover, these tend to correspond to relatively "simple" lattices that represent common, intuitive motifs. To quantify this, we compute a measure of structural "complexity," $C = 1/D_{KL}$, where D_{KL} is the Kullback−Leibler divergence between the radial distribution function (RDF), $g(r)$, of an ideal gas and that of the lattice in question. Since there are multiple components, we append individual pairwise RDFs into a single vector to produce an approximate fingerprint, *e.g.*, $\vec{h}_x = [g_{11}(r)][g_{12}(r)][g_{22}(r)]$. For an ideal gas, $\vec{h}_{ig} = 1$. After normalizing these fingerprints so they sum to unity

$$D_{KL} = \sum_i -h_x(i) \ln\left(\frac{h_{ig}(i)}{h_x(i)}\right) \tag{8}$$

D_{KL} is a measure of how one probability distribution differs from another;[55] we take this as a measure of how much order or correlation exists in a system since nonzero entries in $\vec{h}_x$ correspond to unique pairwise distances found in the crystal. We posit that the fewer of these that exist, the "simpler" the crystal; comparatively, a single instantaneous configuration of an ideal gas may be viewed as a crystalline configuration with an infinite unit cell size, allowing all pairwise distances. Consequently, we expect D_{KL} to be higher for "simpler" crystals, so its inverse, $C = 1/D_{KL}$, represents "complexity".

Typically, the larger the w is for a given CSP solution, the more complex the resulting structures will be on average. Figure 7 shows representative results for three different wallpaper groups in an equimolar binary mixture. The CSP solutions are sorted based on the number of realizations, w, they each have (magenta line). D_{KL} is computed for each structure described by a solution, and the mean is plotted in black. While clearly variable, when fitted to a line, the slope is generally negative (dashed black line). This indicates that the complexity increases from left to right, as the number of realizations increases. By weighting the probability to randomly select a solution by $p = (1 + \ln w)^{\gamma}$, we can favor higher-complexity structures for $\gamma > 0$ and lower-complexity structures for $\gamma < 0$.

While we have approached this problem from a colloidal perspective, we note that sampling with a bias toward "simpler" structures as we have defined them is consistent with Pauling's fifth principle, the "Rule of Parsimony," for ionic crystals.[56] The rule states that the number of essentially different particles and local environments in such crystals tends to be small. This is because ionic crystals tend to have a small number of optimal local arrangements (environments), which combine to fill space. While this is not inviolable, the Rule of Parsimony is a well-validated guiding principle for inorganic crystals, which can be analogous to those often obtained in soft matter systems.[57] Note that, by using different measures of complexity, Oganov and Valle[21] have also found the same trend in binary atomic crystals that simple structures tend to have lower energies, validating Pauling's fifth principle.

4. REDUNDANCY IN THE CSP

For various reasons, duplicate structures can appear as distinct solutions to the CSP we have formulated. This is often the case with very simple, achiral structures whose mirror is superimposable on itself, or when CSP solutions do not make use of lattice sites that critically distinguish the symmetry of the group from others. We performed a detailed analysis of the structures found by our algorithm in different instances, which is available in the SI and is summarized here.

Structural similarity was estimated by using a cosine similarity function based on a structure's radial distribution function (whose Fourier transform is its structure factor). First, we considered duplicates that may arise in a fixed group and N_g as obtained in our *p*4*g* example in Figure 4. For small N_g duplication is common, but it falls quickly and monotonically until less than 1% of pairs are duplicates for $N_g \gtrsim 9$; thus, this effect is minimal for cells of moderate size. However, it is important to also consider the effect of the lattice size itself, for a fixed group, as it is not necessarily guaranteed that CSP solutions on a larger lattice completely encompass those on a smaller one. For example, when the lattice has an odd number of nodes along an edge, a lattice site can be placed at twofold rotation centers occurring along that edge (*e.g.*, as in Figure 3 with *p*2 and *cmm*), whereas, when the number of nodes is even, the center of rotation occurs between lattice sites, and this special Wyckoff position does not contribute to the CSP. Our analysis suggests that this can be an important but not overwhelming effect, though the best practice is to enumerate structures for all possible lattices up to some $N_{g,max}$. Finally, we tested the overlap that can occur between structures generated by different groups and found this to be on the order of 10% between pairs of groups.

A pairwise comparison between structures can be computationally expensive, growing as $O(M^2)$ for M structures generated. Comparatively, the tree in Figure 5 can be generated in a matter of central processing unit (CPU) seconds to minutes depending on N_g. Consequently, it is generally more efficient to allow duplicates to be generated rather than to try to determine and remove them in advance.

5. AVAILABLE CODE

We have developed a software package in python,[58] "paccs", which implements this enumeration scheme and other operations, such as structure optimization given a Hamiltonian.[46] Although full documentation may be found therein, here we present a brief demonstration of the available functionality that can reproduce the results presented in this work. In the first snippet, we present the code to generate the 30 realizations of all four solutions to the CSP, as given in Figure 4b. The parameter "congruent" implies that N_g should be taken as the equivalent *p*1 cell's size, as used in the main text.

https://dx.doi.org/10.1021/acs.jpca.0c00846
J. Phys. Chem. A 2020, 124, 3276−3285

```
import paccs
Ng = 8
kwargs = {
    "stoichiometry": (3,5),
    "grid_count": Ng,
    "congruent": True,
    "sample_groups": [paccs.wallpaper.
    WallpaperGroup(name="p4g")]
}
generator = paccs.wallpaper.
    generate_wallpaper(**kwargs)
structures = list(generator)
```

Listing 1: Generate lattices from Fig. 4.

In the second example, rather than using an equivalent $p1$ cell, we can directly specify the number of nodes we want the FD to have. Setting "congruent" to False means the variable N_g directly assigns the number of nodes on the FD. For the $p4g$ lattice, symmetry constrains $r = 1$, so $N_1 = N_2 = N_g$. Clearly, a 4 × 4 grid used on the $p4g$ FD would produce an equivalent $p1$ cell with an 8 × 8 grid (*cf.* Figure 1; the red primitive cell of $p4g$ would thus be 8 × 8).

```
import paccs
Ng = 4
kwargs = {
    "stoichiometry": (3,5),
    "grid_count": Ng,
    "congruent": False,
    "sample_groups": [paccs.wallpaper.
    WallpaperGroup(name="p4g")],
    "chosen_solution_idx": 2
}
generator = paccs.wallpaper.
    generate_wallpaper(**kwargs)
structures = list(generator)
```

Listing 2: Generate all 6 realizations of solution (ii) in Fig. 4(b); only 3 of the 6 are unique.

Note that we have leveraged python generators, which only return configurations when requested rather than computing them all up-front. This is an efficient way to construct the tree in Figure 5 even when the number of true solutions to the CSP is too large to explicitly enumerate. The tree's topology and branches are simply defined by the keyword arguments, and its leaves (configurations) are traversed and sampled only when requested.

6. CONCLUSIONS

In conclusion, we have presented a method for systematically enumerating crystalline configurations in two-dimensional Euclidean space. Our formulation is based on defining a constraint satisfaction problem (CSP) using both symmetry and stoichiometry. In essence, this amounts to using the multiplicities of general and special Wyckoff positions when redundant lattice positions have been accounted for. This may be determined systematically using our parallel-line construction on the fundamental domains we defined. Our approach consists of two main steps. First, lattices are produced systematically for each domain; a particle placed at a node along the boundary will be divided into some fraction, which can be viewed as its stoichiometric contribution to the fundamental domain and, therefore, the primitive cell. Second, symmetrically equivalent nodes are collapsed, combining these stoichiometric factors from each equivalent node. In conjunction, a relatively simple CSP results, which defines all ways to achieve a desired stoichiometric ratio of an arbitrary number of components for a given wallpaper group. Solutions to the CSP detail how many of each type of particle (*e.g.*, A, B, *etc.*) to place at different types of nodes (*e.g.*, corners, faces, centers of edges, *etc.*), *i.e.*, different Wyckoff positions, with a given multiplicity.

We have described how to view solutions to the CSP as a tree, whose leaves are all realizations (configurations) of each branch. A combinatorial number of realizations to each solution exist, and, when enumerated, represent configurations that have the desired stoichiometric ratio of components and do not violate any of the symmetry conditions of the generating group. This implies that it is possible for the same structure to be generated from different groups if the pattern of the realization is simple enough. Nonetheless, we analyzed the degree of redundancy and found it is often insufficient to justify the additional computational effort to remove duplicates. We provided the python code that performs these calculations and can act as a structure generator, producing an ensemble of configurations that can be further optimized and screened to predict crystal structures. This algorithm is general and may be employed, in principle, to enumerate structures of any stoichiometry with any number of components. We anticipate a number of additional challenges in extending this method to three-dimensional systems, including developing an analogous node construction technique, and having to consider 230 space groups instead of just 17 wallpaper groups. However, we speculate that a stochastic, off-lattice approach to the placement of particles employing known Wyckoff positions[41] followed by more extensive refinement may be more computationally effective. This is the subject of future work.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.0c00846.

Wallpaper groups, discretizing fundamental domains, solving the constrained symmetry and stoichiometry problem, redundancy in the CSP, estimating structural similarity (Sections S1−S5); fraction of unique pairs of configurations (Figure S1); average fractional overlap for each group (Figure S2); similarity of structures in an equimolar, binary system for a fixed group as N_g increases (Table S1); and fraction of structures duplicated between groups at $N_{g,max}$ (Table S2) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Nathan A. Mahynski − *Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8320, United States;* orcid.org/0000-0002-0008-8749; Email: nathan.mahynski@nist.gov

Authors

Evan Pretti − *Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, Pennsylvania 18015-4791, United States*

https://dx.doi.org/10.1021/acs.jpca.0c00846
J. Phys. Chem. A 2020, 124, 3276−3285

Vincent K. Shen − *Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8320, United States*

Jeetain Mittal − *Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, Pennsylvania 18015-4791, United States;* orcid.org/0000-0002-9725-6402

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.0c00846

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

N.A.M. gratefully acknowledges helpful discussions with Prof. J. Dshemuchadse. This work was supported by the U.S. Department of Energy, Office of Basic Energy Science, Division of Material Sciences and Engineering under Award (DE-SC0013979). E.P. acknowledges support from the National Institute of Standards and Technology Summer Undergraduate Research Fellowship (NIST SURF) program with grant no. 70NANB16H. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported under Contract No. DE-AC02-05CH11231. Use of the high-performance computing capabilities of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation, project no. TG-MCB120014, is also gratefully acknowledged. Contribution of the National Institute of Standards and Technology, not subject to U.S. Copyright.

■ REFERENCES

(1) Maddox, J. Crystals from first principles. *Nature* **1988**, *335*, 201.
(2) Gavezzotti, A. Are crystal structures predictable? *Acc. Chem. Res.* **1994**, *27*, 309−314.
(3) Dunitz, J. D. Are crystal structures predictable? *Chem. Commun.* **2003**, 545−548.
(4) Woodley, S. M.; Catlow, R. Crystal structure prediction from first principles. *Nat. Mater.* **2008**, *7*, 937−946.
(5) Wales, D. J.; Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **1997**, *101*, 5111−5116.
(6) Stillinger, F. H. Exponential multiplicity of inherent structures. *Phys. Rev. E* **1999**, *59*, 48−51.
(7) Dijkstra, M.; van Roij, R.; Evans, R. Phase behavior and structure of binary hard-sphere mixtures. *Phys. Rev. Lett.* **1998**, *81*, 2268−2271.
(8) Dijkstra, M.; van Roij, R.; Evans, R. Phase diagram of highly asymmetric binary hard-sphere mixtures. *Phys. Rev. E* **1999**, *59*, 5744−5771.
(9) Filion, L.; Dijkstra, M. Prediction of binary hard-sphere crystal structures. *Phys. Rev. E* **2009**, *79*, No. 046714.
(10) Filion, L.; Marechal, M.; van Oorschot, B.; Pelt, D.; Smallenburg, F.; Dijkstra, M. Efficient method for predicting crystal structures at finite temperature: variable box shape simulations. *Phys. Rev. Lett.* **2009**, *103*, No. 188302.
(11) Zhang, Z.; Glotzer, S. C. Self-assembly of patchy particles. *Nano Lett.* **2004**, *4*, 1407−1413.
(12) Glotzer, S. C.; Solomon, M. J. Anisotropy of building blocks and their assembly into complex structures. *Nat. Mater.* **2007**, *6*, 557−562.
(13) 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.* **2012**, *137*, No. 214101.
(14) Bianchi, E.; Doppelbauer, G.; Filion, L.; Dijkstra, M.; Kahl, G. Predicting patchy particle crystals: variable box shape simulations and evolutionary algorithms. *J. Chem. Phys.* **2012**, *136*, No. 214102.
(15) Pickard, C. J.; Needs, R. J. Ab initio random structure searching. *J. Phys.: Condens. Matter* **2011**, *23*, No. 053201.
(16) Wales, D. J.; Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science* **1999**, *285*, 1368−1372.
(17) Goedecker, S. Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **2004**, *120*, 9911−9917.
(18) Goldberg, D. E.; Holland, J. H. Genetic algorithms and machine learning. *Mach. Learn.* **1988**, *3*, 95−99.
(19) Woodley, S. M.; Battle, P. D.; Gale, J. D.; Catlow, C. R. A. The prediction of inorganic crystal structures using a genetic algorithm and energy minimisation. *Phys. Chem. Chem. Phys.* **1999**, *1*, 2535−2542.
(20) Glass, C. W.; Oganov, A. R.; Hansen, N. USPEX—evolutionary crystal structure prediction. *Comput. Phys. Commun.* **2006**, *175*, 713−720.
(21) Oganov, A. R.; Valle, M. How to quantify energy landscapes of solids. *J. Chem. Phys.* **2009**, *130*, No. 104504.
(22) Fornleitner, J.; Lo Verso, F.; Kahl, G.; Likos, C. N. Genetic algorithms predict formation of exotic ordered configurations for two-component dipolar monolayers. *Soft Matter* **2008**, *4*, 480−484.
(23) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B.* **2010**, *82*, No. 094116.
(24) Oganov, A. R.; Lyakhov, A. O.; Valle, M. How evolutionary crystal structure prediction works—and why. *Acc. Chem. Res.* **2011**, *44*, 227−237.
(25) Lyakhov, A. O.; Oganov, A. R.; Stokes, H. T.; Zhu, Q. New developments in evolutionary structure prediction algorithm USPEX. *Comput. Phys. Commun.* **2013**, *184*, 1172−1182.
(26) Wu, S. Q.; Ji, M.; Wang, C. Z.; Nguyen, M. C.; Zhao, X.; Umemoto, K.; Wentzcovitch, R. M.; Ho, K. M. An adaptive genetic algorithm for crystal structure prediction. *J. Phys.: Condens. Matter* **2014**, *26*, No. 035402.
(27) Curtis, F.; Li, X.; Rose, T.; Vazquez-Mayagoitia, A.; Bhattacharya, S.; Ghiringhelli, L. M.; Marom, N. GAtor: A first-principles genetic algorithm for molecular crystal structure prediction. *J. Chem. Theory Comput.* **2018**, *14*, 2246−2264.
(28) Schütt, K.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K.; Gross, E. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, No. 205118.
(29) Lindquist, B. A.; Jadrich, R. B.; Truskett, T. M. Communication: Inverse design for self-assembly via on-the-fly optimization. *J. Chem. Phys.* **2016**, *145*, No. 111101.
(30) Huang, B.; Von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: the role of uniqueness and target similarity. *J. Chem. Phys.* **2016**, *145*, No. 161102.
(31) Dral, P. O.; Owens, A.; Yurchenko, S. N.; Thiel, W. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *J. Chem. Phys.* **2017**, *146*, No. 244108.
(32) Spellings, M.; Glotzer, S. C. Machine learning for crystal identification and discovery. *AIChE J.* **2018**, *64*, 2198−2206.
(33) Ryan, K.; Lengyel, J.; Shatruk, M. Crystal structure prediction via deep learning. *J. Am. Chem. Soc.* **2018**, *140*, 10158−10168.
(34) Wales, D. J. Symmetry, near-symmetry and energetics. *Chem. Phys. Lett.* **1998**, *285*, 330−336.
(35) Treacy, M. M. J.; Randall, K. H.; Rao, S.; Perry, J. A.; Chadi, D. J. Enumeration of periodic tetrahedral frameworks. *Z. Kristallogr.* **1997**, *212*, 768−791.
(36) Friedrichs, O. D.; Dress, A. W. M.; Huson, D. H.; Klinowski, J.; Mackay, A. L. Systematic enumeration of crystalline networks. *Nature* **1999**, *400*, 644−647.
(37) Foster, M. D.; Simperler, A.; Bell, R. G.; Delgado Friedrichs, O.; Almeida Paz, F. A.; Klinowski, J. Chemically feasible hypothetical crystalline networks. *Nat. Mater.* **2004**, *3*, 234−238.

(38) Foster, M. D.; Delgado Friedrichs, O.; Bell, R. G.; Almeida Paz, F. A.; Klinowski, J. Chemical evaluation of hypothetical uninodal zeolites. *J. Am. Chem. Soc.* **2004**, *126*, 9769−9775.
(39) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-scale screening of hypothetical metal-organic frameworks. *Nat. Chem.* **2012**, *4*, 83−89.
(40) LaCour, R. A.; Adorf, C. S.; Dshemuchadse, J.; Glotzer, S. C. Influence of softness on the stability of binary colloidal crystals. *ACS Nano* **2019**, *13*, 13829−13842.
(41) Hahn, T., Eds. *International Tables for Crystallography. Volume A: Space-Group Symmetry*; International Union of Crystallography, 1983; Vol. *A*.
(42) Thurston, W. P. *The Geometry and Topology of Three-Manifolds*; Princeton University: Princeton, NJ, 1979.
(43) Conway, J. H.; Delgado Friedrichs, O.; Huson, D. H.; Thurston, W. P. On three-dimensional space groups. *Contrib. Algebr. Geom.* **2001**, *42*, 475−507.
(44) Conway, J. H.; Huson, D. H. The orbifold notation for two-dimensional groups. *Struct. Chem.* **2002**, *13*, 247−257.
(45) Mahynski, N. A.; Pretti, E.; Shen, V. K.; Mittal, J. Using symmetry to elucidate the importance of stoichiometry in colloidal crystal assembly. *Nat. Commun.* **2019**, *10*, No. 2028.
(46) paccs: Python Analysis of Colloidal Crystal Structures. https://github.com/usnistgov/PACCS.
(47) Schattschneider, D. The plane symmetry groups: their recognition and notation. *Am. Math. Mon.* **1978**, *85*, 439−450.
(48) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Elsevier, 2001; Vol. *1*.
(49) Koch, E.; Fischer, W. Zur bestimmung asymmetrischer einheiten kubischer raumgruppen mit hilfe von wirkungsbereichen. *Acta Crystallogr., Sect. A* **1974**, *30*, 490−496.
(50) Grosse-Kunstleve, R. W.; Wong, B.; Mustyakimov, M.; Adams, P. D. Exact direct-space asymmetric units for the 230 crystallographic space groups. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2011**, *67*, 269−275.
(51) Hudson, T. S.; Harrowell, P. Structural searches using isopointal sets as generators: densest packings for binary hard sphere mixtures. *J. Phys.: Condens. Matter* **2011**, *23*, No. 194103.
(52) Apt, K. *Principles of Constraint Programming*; Cambridge University Press, 2003.
(53) Niemeyer, G. *Python-Constraint*, version 1.4.0; Python Software Foundation, 2018. https://labix.org/python-constraint.
(54) Calvo, F.; Schebarchov, D.; Wales, D. Grand and semigrand canonical basin-hopping. *J. Chem. Theory Comput.* **2016**, *12*, 902−909.
(55) Kullback, S.; Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79−86.
(56) Pauling, L. The principles determining the structure of complex ionic crystals. *J. Am. Chem. Soc.* **1929**, *51*, 1010−1026.
(57) Leunissen, M. E.; Christova, C. G.; Hynninen, A.-P.; Royall, C. P.; Campbell, A. I.; Imhof, A.; Dijkstra, M.; van Roij, R.; van Blaaderen, A. Ionic colloidal crystals of oppositely charged particles. *Nature* **2005**, *437*, 235−240.
(58) *Python Language Reference, version 3.6*, Python Software Foundation, 2018. http://www.python.org.

https://dx.doi.org/10.1021/acs.jpca.0c00846
J. Phys. Chem. A 2020, 124, 3276−3285