Pre-Nucleation Clusters Predict Crystal Structures in Models of Chiral Molecules

John E. Carpenter and Michael Grünwald*

Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States

E-mail: michael.gruenwald@utah.edu

Abstract

Kinetics can play an important role in the crystallization of molecules and can give rise to polymorphism, the tendency of molecules to form more than one crystal structure. Current computational methods of crystal structure prediction, however, focus almost exclusively on identifying the thermodynamically stable polymorph. Kinetic factors of nucleation and growth are often neglected because the underlying microscopic processes can be complex and accurate rate calculations are numerically cumbersome. In this work, we use molecular dynamics computer simulations to study simple molecular models that reproduce the crystallization behavior of real chiral molecules, including the formation of enantiopure and racemic crystals, as well as polymorphism. A significant fraction of these molecules forms crystals that do not have the lowest free energy. We demonstrate that at high supersaturation crystal formation can be accurately predicted by considering the similarities between oligomeric species in solution and molecular motifs in the crystal structure. For the case of racemic mixtures, we even find that knowledge of crystal free energies is not necessary and kinetic considerations are sufficient to determine if the system will undergo spontaneous chiral separation. Our results suggest conceptually simple ways of improving current crystal structure prediction methods.

Introduction

Predicting which crystal structure a given molecule will form is a long-standing problem[1] with considerable practical significance for the industrial production of many chemical compounds, including medicinal drugs,[2–4] pesticides,[5–8] and explosives.[9–13] Finding the crystal structure with the lowest free energy is a daunting task, requiring efficient methods for sampling the space of periodic molecular packings as well as accurate methods for calculating crystal (free) energies.[14–19] Crystal structure prediction (CSP) is further complicated by the fact that the majority of molecules can form more than one polymorph, depending on crystallization conditions.[20] Predicting polymorphism requires not only knowledge of the thermodynamic stability of different polymorphs, but also insight into the mechanistic details of crystal formation as well as methods to estimate the rates of these processes.[21]

Most current computational frameworks of CSP focus entirely on the thermodynamic aspects of crystallization, and much progress has been made in recent years. The most accurate CSP methods now frequently identify all experimentally known polymorphs and their ranking in terms of free energies.[14,18,22] Nevertheless, CSP has still not replaced time-consuming experimental polymorph screening procedures. Many of the computationally predicted structures never materialize in experiments, in some cases even those with free energies lower than known polymorphs.[23] In order to determine which of the predicted low-energy polymorphs can likely be realized in experiments and which cannot, kinetic effects need to be incorporated in CSP.[24,25]

Why are kinetic factors not considered more routinely in CSP? Rates of crystal nucleation and growth depend sensitively on the experimental details (including solvent, molecular concentrations, and temperature) and are determined by a series of microscopic rare events including the desolvation, attachment, and perhaps rearrangement of molecular species on the surface of a growing crystallite. Determination of the timescales of these events requires numerically expensive molecular dynamics methods and highly accurate force fields.[26] Routine calculation of crystal formation rates of many different polymorphs is therefore currently

not feasible.

Traditional theories of crystal nucleation and growth assume that the building blocks attaching to a growing crystallite are monomers, or at least a unique species. There is growing evidence, however, that molecules can associate in solution to produce substantial concentrations of oligomers.[27–36] These oligomers can act as important building blocks in the nucleation and growth of specific polymorphs.[37–45] Concentrations of amino acid oligomers in solution, for instance, have been studied extensively using different techniques.[27–36] Electrospray ionization experiments have shown strong evidence of oligomerization of amino acids at low concentrations[35] and suggest the existence of 'magic number' oligomers with particularly large concentrations (e.g., tetramers of arginine).[36] A recent sedimentation study has shown that although monomers are the dominant species in undersaturated aqueous environments, large oligomers are present even at very low monomer concentrations and relative oligomer concentrations increase as supersaturation is approached.[28] Oligomeric species that serve as precursors for specific crystal structures are often referred to as pre-nucleation clusters (PNCs).[46] Substantial populations of PNCs in solution prior to crystallization have been observed in a range of systems.[39,40,42,43,47,48] The related concept of *synthons* describes energetically important binding motifs within a crystal structure.[49] In a recent review, Davey and coworkers reported that synthons present in solution also appear in the final crystal structure in 11 out of 14 cases.[21] Although oligomers, PNCs, or synthons have been frequently shown to correlate with the formation of specific polymorphs in experiments, the role of these species in molecular crystallization has not been systematically studied and no computational framework exists that incorporates oligomer concentration in CSP.

In this paper, we demonstrate with computer simulations that oligomers can play a decisive role in determining crystallization outcomes. Our study is based on a family of simple models of chiral molecules[50] capable of replicating the rich crystallization behavior found in real molecules. We show that the crystallization of these models in molecular dynamics (MD) simulations can be accurately predicted based on classical nucleation theory if available

oligomer building blocks are accounted for. Our theoretical framework successfully balances the kinetic and thermodynamic factors leading to polymorph formation. Even though our study is based on models that lack chemical detail, it suggests computationally tractable ways of augmenting existing CSP frameworks with kinetic information.

Results and Discussion

Molecular model and crystallization simulations

We simulated the crystallization of racemic mixtures of simple chiral molecules in two dimensions. Model molecules are rigid and consist of 5 beads that represent different functional groups and interact via short-ranged attractive pair potentials, as specified in the Methods section. We chose the strengths of interactions between functional groups in a way that mimics the heterogeneity of interactions found in real molecules: While most functional groups interact only weakly, a small number of pairs of functional groups have interactions that are several times stronger. By varying the spatial arrangement of functional groups and the distribution of weak and strong interactions, a large family of different molecules can be constructed. In recent work, we studied a subset of 159 of these molecules and showed that their simulated crystallization yields the same types of outcomes found in experiments: racemic crystals, enantiopure crystals (via spontaneous chiral separation) and several types of partially ordered and disordered solids.[50] In addition, the model shows good agreement with several other experimental facts: It produces racemic and enantiopure crystals with a relative frequency that is consistent with molecular crystallization experiments on surfaces, and produces crystalline (rather than disordered) outcomes at a rate consistent with a recent experimental study.[51] In this work, we focus on molecules that robustly form large crystalline clusters with few defects in our simulations. In addition to the 29 molecules from our previous work that meet this criterion (called set A in the following), we added 34 new molecules that we specifically selected because of their tendency to form kinetically rather than ther-

modynamically preferred polymorphs (set B, see Methods). While we expect molecules in set A to be representative of a typical set of real organic molecules, set B presents a worst-case scenario for CSP methods that rank polymorphs according to their free energies. (Shapes and interactions of all molecules studied in this work are specified in Fig. S1, Table S2, and Table S3.)

For each of these molecules, we performed crystallization simulations by placing racemic mixtures of 5184 molecules in square simulation boxes at a packing fraction of $0.04\,\sigma^{-2}$ and solving the Langevin equations of motion for several hundred millions of time steps. Solvent molecules were not represented explicitly. To facilitate crystallization, we used the following temperature protocol: Starting from an initial temperature well above crystallization conditions, we lowered the temperature until the first cluster of 50 molecules was observed. We then automatically adjusted the temperature to grow the largest molecular cluster at a fixed rate. In the final step of the procedure, temperature was linearly increased to facilitate defect annealing. For the vast majority of molecules, this protocol yields the same polymorphs as constant-temperature simulations but with fewer defects and without the need for manual optimization of simulation conditions. Independent simulation runs of the same molecules consistently yield the same polymorphs. A detailed description of simulation methods can be found in the Methods section and Ref. 50. Images of some of the molecules we studied and the crystals they form are shown in Figure 1. (All other molecules and their crystals are depicted in Figs. S2 and S3.)

Thermodynamic Polymorph Landscapes

We use a recently developed algorithm (POLYNUM) to identify millions of polymorphs for each of the 65 molecules and evaluate their thermodynamic stability in terms of their lattice energy. Because molecules are rigid and the range of intermolecular interactions is short, polymorph *free* energies typically deviate from lattice energies by less than 1%.[50] We will therefore use the two terms interchangeably. To quantify the thermodynamic role of the

polymorphs that form in our simulations, we calculate the relative energy difference

$$\Delta E_{\text{form}} = \frac{E_{\text{form}} - E_{\text{comp}}}{|\min(E_{\text{form}}, E_{\text{comp}})|},$$

where E_{form} is the energy of the observed polymorph and E_{comp} is the energy of the most stable competing polymorph, *i.e.*, the lowest energy of all polymorphs that did *not* form. Negative values of ΔE_{form} therefore indicate that the observed polymorph is the thermodynamic equilibrium structure, while cases of kinetically driven polymorph formation are indicated by positive values of ΔE_{form}.

Polymorph energy landscapes of our model molecules show that a substantial fraction of the crystals that form in our simulations are kinetic products ($\Delta E_{\text{form}} \geq 0$), as illustrated in Fig. 2. Of the 29 molecules in set A (which includes all good crystallizers from our previous work), 11 (41%) feature energy landscapes with polymorphs that have energies equal to or lower than the polymorph that is observed. This fraction of "kinetic" crystallizers is consistent with a recent estimate based on real organic molecules.[52] The fraction of molecules with $\Delta E_{\text{form}} \geq 0$ is even larger in set B since most of these molecules were selected because they form enantiopure crystals even though a racemic crystal is thermodynamically stable. To further characterize the polymorph landscape for each molecule, we calculate the number $N_{0.95}$ of polymorphs that have an energy smaller than $0.95E_0$, where E_0 is the minimum polymorph energy of a given molecule. $N_{0.95}$ varies substantially between molecules, but in all but a few cases we find at least several (and up to ≈ 100) competing polymorphs at low energies. Numerical values of E_{form} and $N_{0.95}$ for all molecules are given in Figure 1, Figure S2, and Figure S3.

Polymorphs are heralded by oligomer species

Visual inspection of crystallization trajectories suggests that oligomer species can be a decisive factor in the crystallization of our model molecules. As an example, let us consider

molecule s9-A1, which forms an enantiopure polymorph (called X in the following) that does not have the lowest energy ($\Delta E_{\mathrm{form}} = 0.03$, $N_{0.95} = 168$). The polymorph landscape of s9-A1 is illustrated in Figure 4a and features 55 polymorphs with lower energies than X. The unit cells of X and four racemic and enantiopure polymorphs with substantially lower energy (called V, W, Y, and Z)[1] are shown in Figure 3b.

Why does polymorph X prevail? Figure 3a shows a snapshot from a simulation of the crystallization of s9-A1. As evident from this image, the solution surrounding the growing nucleus of polymorph X contains various oligomeric species at substantial concentrations. Conspicuously, many of these oligomers closely resemble motifs found in polymorph X; only rarely do we observe oligomers that "belong" to any of the four other polymorphs. The relative concentrations of these oligomers are consistent with their zero-temperature energies. Figure 5 shows images of the lowest energy oligomers comprising between two and six molecules for each of the five polymorphs; Figure 4c shows a plot of these energies as a function of oligomer size. Even though polymorph X has a higher lattice energy than any of the other four polymorphs, its oligomeric motifs have the lowest energies for all oligomer sizes considered here. Energy differences between oligomers amount to several $k_{\mathrm{B}}T$ at the temperature of crystal formation ($T \approx 1.0\,\epsilon/k_{\mathrm{B}}$), consistent with the high concentrations of oligomers of polymorph X in our simulations. This observation suggests that polymorph X has a kinetic advantage over competing polymorphs: Oligomers resembling motifs of a given polymorph are likely to attach productively to the surface of a nucleus of that polymorph, while they will either only transiently attach to a different crystal surface or will need to undergo energetically activated rearrangements before they can be incorporated. Can such differences in attachment rates of oligomeric species be sufficient to overcome a substantial thermodynamic disadvantage? And can relative formation rates of different polymorphs be predicted from knowledge of oligomeric species in solution?

Our simulations suggest that polymorphs cannot be predicted based on speciation of

[1]We selected these four polymorphs because they illustrate the broad range of structural motifs that appear in low-energy polymorphs of this molecule.

oligomers alone. Polymorphs with substantial thermodynamic advantage can form despite a lack of suitable oligomers in solution. An example of such a system is molecule s7-A1, whose polymorph landscape is shown in Figure 4b. There is a large energy gap between the lowest energy polymorph (labeled Q) and its competitors. This decisive thermodynamic advantage of polymorph Q is, however, not reflected in the energies of small oligomeric motifs, as illustrated in Figure 7 and Figure 4d. Several other polymorphs (e.g., polymorphs R and S) have oligomeric motifs at substantially lower energies and larger concentrations in solution, as evident from the simulation snapshot in Figure 6. Nevertheless, polymorph Q forms, either through monomer addition or through a more complicated growth process. Clearly, kinetic factors associated with growth of crystalline clusters through oligomer addition must be balanced appropriately with polymorph energies in order to predict which polymorph will form in our simulations.

Estimating nucleation rates from crystal structures

Crystallization rates are typically analyzed assuming either nucleation or crystal growth as the rate limiting step. In the former, the polymorph with the largest nucleation rate, usually estimated using classical nucleation theory (CNT), is thought to prevail. In a growth-dominated scenario, one assumes that any kinetic advantages in the nucleation stage are irrelevant in comparison to differences in polymorph growth rates, which determine the final crystallization outcome. Convincing experimental and theoretical evidence exists for both scenarios and it is reasonable to assume that molecular crystallization can be determined by either, depending on conditions. In the majority of our simulations, clusters containing several unit cells of the eventually successful polymorph form already in the early stages of our simulations, albeit with high concentrations of defects. Concomitant polymorphism, *i.e.*, simultaneous formation of large clusters of different polymorphs, is observed in only a few cases (molecules s10-B2, s2-B5, and s5-A3). We therefore concluded that crystallization outcomes are determined primarily at the nucleation stage in our simulations and accordingly

chose to estimate crystallization rates based on classical nucleation theory. However, the kinetic factors associated with attachment of oligomeric species, which are the focus of this work, are relevant also for crystal growth rates. For one case (molecule s9-A1), we demonstrate further below that our nucleation rate estimates correctly identify the fastest growing polymorph, too.

According to classical nucleation theory, the rate at which super-critical nuclei are produced per unit area in two dimensions is given by[53]

$$J = A \exp\left(-\frac{\Delta G}{k_{\mathrm{B}} T}\right), \tag{1}$$

where A is the kinetic prefactor and ΔG is the free energy barrier associated with forming a nucleus of critical size. The simplicity of Eqn. 1 belies tremendous complexity in practical applications. The kinetic prefactor A encompasses various factors relating to the attachment of growth units to the nucleus. These factors include the concentration and diffusion rate of growth units in solution, free energies of de-solvating growth units and nucleus surface, as well as factors that determine the probability of growth units to attach correctly, including the symmetry of growth units and the structural complexity of the crystal surface. Differences in the kinetic prefactor are often neglected in the analysis of nucleation rates of different polymorphs under the assumption that the nucleation barrier is the most important term.[19,54] Other studies have shown that, on the contrary, A can be a decisive factor.[55,56] We show below that both prefactor and nucleation barrier need to be accounted for to predict molecular crystallization in our simulations.

Assuming that the polymorph with the largest nucleation rate will form, we wish to calculate nucleation rates J_p of all polymorphs p with sufficiently low lattice energy. But accurate estimates of J_p cannot be easily obtained, even for a simple coarse-grained model like the one analyzed here, because of the large number of energetically competitive polymorphs and the substantial numerical effort associated with determining the various thermodynamic

and kinetic factors entering into Eq. 1. Motivated by the crystallization dynamics observed in our simulations and with an eye towards applicability to real molecules, we therefore focus on those elements of Eq. 1 that describe attachment of various oligomer species to a growing nucleus and that can be straightforwardly estimated from a set of predicted crystal structures. Other factors are assumed to vary insignificantly between different polymorphs and are not considered here in any detail.

We rank polymorphs p of a given molecule according to a "nucleation score" χ_p, which is proportional to the CNT nucleation rate (under assumptions described below) and given by

$$\chi_p = \nu_p \exp\left(-\eta_p\right). \tag{2}$$

Here, ν_p is a dimensionless quantity proportional to the rate of attachment of oligomeric species to the nucleus, and η_p is a simple estimate of the nucleation barrier $\Delta G_p/k_\mathrm{B}T_\mathrm{c}$ at the temperature T_c of crystallization. The most important steps in the calculation of the nucleation score include the following:

1. We enumerate all oligomeric motifs containing up to six molecules that occur in the crystal structures of the 100 lowest-energy polymorphs of a given molecule.

2. We estimate the concentrations of these oligomers in solution based on their potential energies.

3. By appropriately summing up the concentrations of those oligomers that occur in a given polymorph p, we estimate the total rate of attachment of molecules to a nucleus of p to arrive at ν_p.

4. To calculate the nucleation barrier η_p of polymorph p, we estimate the chemical potential difference of molecules in p and in solution based on the lattice energy of p and a single global fitting parameter that is the same for all polymorphs and all molecules.

We discuss the functional forms of ν_p and η_p, as well as major approximations we make, in

detail in the Methods and Discussion sections. In the following, we describe the success of the nucleation score in predicting crystallization outcomes in our simulations.

Nucleation score predicts outcomes of crystallization simulations

To assess the predictive power of the nucleation score introduced in the previous section, we computed χ_p for each of the 100 lowest-energy enantiopure and racemic polymorphs of each of the 63 molecular models. If the polymorph with the largest value of χ_p matched the polymorph identified in MD simulations, the prediction was considered successful. In the few cases of molecules that formed two polymorphs simultaneously, the prediction was considered a success if either of the two polymorphs received the highest nucleation score. (For a handful of molecules, the predicted and observed polymorphs were super-cell variants of each other; consistent with other studies, we considered these to be identical.[57])

Despite the approximations underlying our model, the nucleation score correctly predicts almost all simulated crystallization outcomes. For molecules in set A, we predict the correct polymorph in 28 out of 29 cases (97%). This success rate constitutes a significant increase over a purely thermodynamic ranking based on lattice energies, which selects the correct polymorph in 24 out of 29 cases (83%) or 18 out of 29 cases (62%), depending on whether the correct polymorph is selected in cases where more than one polymorph is found at the lowest energy or not. For molecules in set B, which contains a much larger fraction of "kinetic" crystallizers with $\Delta E_{\text{form}} > 0$, the nucleation score correctly predicted 32 out of 34 cases (94%). For this set of molecules, lattice energies correctly predict only 7 out of 34 cases (21%). (Due to the way interactions between molecules in set B were selected—see Methods—all of these 7 models had non-degenerate ground states.) Remarkably, in both sets A and B the nucleation score predicts spontaneous chiral separation (*i.e.*, whether a racemic or enantiopure polymorph will form) with 100% accuracy. Overall, the nucleation score failed to predict the correct polymorph in only 3 out of 63 cases; for these molecules, the polymorph with the highest nucleation score shared many similarities with the polymorph

that formed in MD simulations (see Figures S2 and S3).

Dimers are not enough

Many studies have demonstrated correlations between polymorphs observed in experiments and a specific pre-nucleation cluster with low energy. These investigations usually concentrate on dimers.[38–43] We find that in most cases studied here it is not sufficient to consider a single oligomer species; larger oligomers need to be included to achieve the best results. As illustrated in Figure 8, prediction based on dimer species alone is successful only in 58% of all cases, a modest improvement over a ranking based on lattice energies (45%). (Note that because differences in surface energies and packing fractions of polymorphs are neglected in our model, rankings based on lattice energy are identical to rankings based on nucleation barriers η_p.) When oligomers of larger sizes are included in the analysis, the success rate increases approximately linearly and reaches 95% when all oligomer sizes up to hexamers are considered.

As an illustration of the importance of larger oligomers, consider the crystallization of molecule s9-A1/5. If only dimers and trimers of this molecule are included in the calculation of ν_p, polymorph Y receives the largest nucleation score, rather than the observed polymorph X. A look at Figs. 4c and 5 reveals why: The dimer with the lowest energy occurs in both X and Y, and the most important trimers present in the two polymorphs have the same energy. Since polymorph Y has the lower lattice energy, it receives the higher prediction score. The substantial kinetic advantage of polymorph X over Y only becomes apparent at oligomer sizes larger than three.

Thermodynamics vs. Kinetics

Figure 8 demonstrates that our model achieves the highest prediction accuracy when both kinetic and thermodynamic factors are included, as encoded in the attachment rate ν_p and the nucleation barrier η_p, respectively. To illustrate the competition between thermody-

namic and kinetic factors in the nucleation score and in our MD simulations, we revisit the crystallization of molecules s7-A1 and s9-A1. Table 1 lists the values of the nucleation score χ_p, oligomer attachment rate ν_p, and nucleation barrier η_p for the competing polymorphs discussed earlier (see Figs. 3–7). Molecule s7-A1 forms the thermodynamically preferred polymorph Q. As evident from the simulation snapshot in Figure 6, few of the oligomers present in solution can directly contribute to the growth of Q, as reflected in a small attachment rate ($\nu_{\rm Q} << \nu_{\rm S} < \nu_{\rm R}$). However, Q has a large energetic advantage over polymorphs R and S, resulting in a nucleation barrier that is much smaller than those of competing polymorphs ($\eta_{\rm Q} << \eta_{\rm R} < \eta_{\rm S}$). As a result of its low lattice energy and despite its low oligomer attachment rate, Q receives the largest nucleation score and indeed forms in simulations; nuclei of R or S are not observed.

Table 1: Kinetic and thermodynamic factors in the nucleation of molecules s9-A1 and s7-A1. Normalized nucleation score χ_p, barrier η_p, and oligomer attachment rate ν_p, lattice energy per molecule E_p, number of molecules in the asymmetric unit cell Z'_p, and composition (racemic/enantiopure) of several competing polymorphs of molecules s9-A1 and s7-A1. ($\chi_{\rm max}$ and $\nu_{\rm max}$ are the largest nucleation score and attachment rate, respectively, found for any polymorph of a given molecule, and $\eta_{\rm min}$ is the smallest nucleation barrier.) Polymorphs X and Q form spontaneously in MD simulations, as illustrated in Figs. 3 and 6, respectively.

mol.	**pol.**	$\frac{\chi_p}{\chi_{\rm max}}$	$\frac{\exp(-\eta_p)}{\exp(-\eta_{\rm min})}$	$\frac{\nu_p}{\nu_{\rm max}}$	$E_p(\epsilon)$	Z'_p	**comp.**
	Q	1.0	1.0	6.8E-3	-19.0	1	rac
s7-A1	R	0.18	1.3E-3	1.0	-16.5	2	rac
	S	3.8E-3	1.1E-4	0.23	-16.1	6	pure
	X	1.0	0.46	1.0	-18.5	1	pure
	Y	0.50	1.0	0.23	-19.0	1	rac
s9-A1	V	1.8E-3	1.0	8.3E-4	-19.0	1	rac
	W	3.6E-4	1.0	1.6E-4	-19.0	1	pure
	Z	2.4E-4	1.0	1.1E-4	-19.0	1	rac

By contrast, molecule s9-A1 crystallizes through a kinetically preferred route. The polymorph that forms in MD simulations, X, has a clear energetic disadvantage and therefore a larger nucleation barrier than the four other polymorphs (Y, V, W, Z) considered here ($\eta_{\rm X} > \eta_{\rm Y}, \eta_{\rm V}, \eta_{\rm W}, \eta_{\rm Z}$). However, V, W, and Z have structural motifs that are not reflected in

the prevalent oligomer species in solution, resulting in small oligomer attachment rates ν_p for these polymorphs. Polymorph Y shares some of the same dimer motifs with X, but larger oligomers of Y have lower concentrations than those contributing to the growth of X, leaving X with a modestly larger nucleation score than Y despite its higher lattice energy. These two examples illustrate that the proposed nucleation score is able to successfully balance lattice energies and kinetic factors of oligomer attachment to produce reliable polymorph predictions.

Depending on the relative values of attachment rate ν_{form} and nucleation barrier η_{form} of the polymorph that forms in our simulation, we identify four physically distinct crystallization scenarios:

I: The crystal that forms has both a lower energy and a larger oligomer attachment rate than any competing polymorph.

II: The crystal that forms has the lowest energy but there is at least one competing polymorph that has a larger oligomer attachment rate. Molecule s7-A1 is an example of this crystallization scenario.

III: The crystal that forms is thermodynamically metastable but has the largest oligomer attachment rate. Molecule s9-A1 is an example of this crystallization scenario.

IV: The crystal that forms has neither the lowest energy nor the largest oligomer attachment rate, but nevertheless has the largest nucleation rate.

Crystallization scenarios I and II result in thermodynamically stable crystals and can in principle be predicted with CPS methods, given that accurate free energies can be calculated. (Note, however, that these scenarios include molecules that have ground state energies shared by several polymorphs.) Crystallization scenarios III and IV result in metastable, "kinetic" polymorphs that cannot be easily predicted with current CSP methods. Table 2 shows that all four types are represented in molecule sets A and B, with a majority of cases of type I in set A, and a majority of type III in set B.

Table 2: Number of molecules in sets A and B that crystallize according to scenarios I–IV defined in the text.

Scenario	**I**	**II**	**III**	**IV**
Set A	17	7	3	2
Set B	3	4	22	5

Accounting for kinetic effects of oligomer attachment is particularly effective if one only wishes to know if a given molecule is likely to undergo spontaneous chiral separation, *i.e.*, if it will form an enantiopure or racemic crystal. In contrast to predictions of specific crystal lattices (Fig. 8), our ability to predict chiral separation does *not* improve when lattice energies (via the nucleation barrier η_p) are included; predictions based on oligomer attachment rates alone (ν_p) are just as successful, as illustrated in Figure S9. Note that also in this case all oligomer sizes up to hexamers need to be included in the calculation of ν_p to achieve the best results. We hypothesize that accounting for these larger oligomer sizes successfully captures the fact that racemic polymorphs typically do not contain enantiopure motifs consisting of more than a few monomers. While low-energy enantiopure dimer or trimer motifs frequently appear in both enantiopure and racemic crystals, the presence of larger enantiopure oligomers in solution clearly favors formation of enantiopure crystals.

Discussion

The nucleation score presented in this paper is designed to systematically capture contributions of oligomeric species (or pre-nucleation clusters) in solution to the nucleation rate of different polymorphs within a numerical framework that emphasizes computational simplicity. The nucleation score contains only a single fitting parameter and can be evaluated from a list of low-energy crystal structures, as furnished routinely by CSP methods—no dynamic information is needed. We have demonstrated that the nucleation score accurately predicts crystallization and chiral separation in simulations of a family of model molecules that display a range of crystallization outcomes similar to real molecules.

We discuss several limitations and caveats of our model. The hexagonal geometry and simple interactions of our molecules facilitate the formation of close-packed crystal structures. However, these features of our model are not the primary driver of the observed crystallization behavior. In our previous work,[50] we studied a related model that in addition to short-ranged attractions includes repulsive interactions between functional groups, providing a more realistic description of electrostatic interactions. Crystals formed by that model are typically more open and in many of these crystals the positions of functional groups within unit cells did not align with the sites of a hexagonal lattice. Nevertheless, we observed the same qualitative crystallization behavior for these "charged" models: Molecules form substantial numbers of oligomers in solution and these oligomeric motifs tend to appear in the crystal structure that forms in simulations. We are therefore confident that the kinetics of oligomer attachment observed in our work are not caused by the simple geometry of our model molecules.

Starting from classical nucleation theory, we made several approximations to render the nucleation score useful for practical application. The most severe of these approximations arguably include the neglect of variations in surface tension of different polymorphs and the assumption that oligomers of size 2–6 are present in solution at similar concentrations. As shown in Fig. S7, surface tensions of different polymorphs of a given molecule can in fact vary by up to $\approx 1\epsilon/\sigma$; such variations, if included in our model, would result in substantially different nucleation barriers. At the same time, the total concentrations of oligomers of different size at the temperature of crystallization can vary substantially between different molecules, as evident from simulation snapshots in Figs. 3 and 6. These variations are not captured in the nucleation score. In particular, our model likely overestimates the concentrations (and therefore also the attachment rates) of larger oligomers. Why are successful polymorph predictions possible despite these simplifications? We hypothesize that variations of polymorph surface tension are partly encoded in the energies of oligomers, particularly larger ones. Polymorphs containing low-energy oligomeric motifs can be dissected into sub-

units that have strong bonds within a given motif but much weaker interactions between different motifs, as illustrated in Figure S8. Such a pronounced separation of strong and weak interactions within a given crystal typically allows for cleavage of the polymorph along planes of weaker interactions, resulting in a small surface tension. In contrast, a polymorph in which all monomers are bound to their neighbors with similar strength will, on average, have a larger surface tension and few oligomers with low energy, resulting in a lower estimated attachment rate ν_p. Effects of varying surface tension are thus included effectively by overemphasizing larger oligomers in our nucleation score.

Another potential caveat of our approach is related to the short time and length scales accessible to our MD simulations. On much longer, experimental time scales, differences in crystal growth rates rather than nucleation rates might determine the fate of the crystallization process. However, oligomer attachment, as estimated in our model, can be an equally important factor in crystal growth. We have convinced ourselves for one case (molecule s9-A1) that the nucleation score also predicts the fastest-growing polymorph at temperatures at which spontaneous nucleation cannot be observed in our simulations. Figure 9 shows the time evolution of the number of molecules, averaged over three independent simulation runs, of large seed crystallites of different polymorphs in supersaturated solutions (see Methods). We find that the polymorph with the largest nucleation score (X) also grows fastest. Furthermore, we find that the ranking of polymorphs according to increasing growth rates (Z–W–V–Y–X) is identical to the ranking according to increasing oligomer attachment rates (see Table 1). This observation is consistent with reports showing that nucleation rates and growth rates can be highly correlated.[58,59] We are therefore optimistic that our model robustly captures several important factors in molecular crystallization.

We expect our model to be most predictive under conditions of large supersaturation. In this regime, oligomers will be present at substantial concentrations and nucleation barriers will be small. Closer to the saturation curve (*i.e.*, at higher temperatures or smaller concentrations), the importance of oligomer attachment in nucleation and growth is diminished and

differences in nucleation barrier heights, which our model only crudely estimates, can become decisive. On the other hand, recent work has shown that oligomers can be the dominant growth unit even when their concentration in solution is low compared to monomers.[60] We cannot straightforwardly simulate crystal nucleation at low supersaturation due to excessively long time scales required to cross nucleation barriers that exceed a few $k_{\mathrm{B}}T$. We have, however, simulated growth of crystalline seeds of molecule s9-A1 at different temperatures above T_{c}, as illustrated in Figure S5. Polymorph X grows fastest within $\approx 10\%$ of T_{c}, in good agreement with our model. At higher temperatures polymorph Y prevails, which has a substantially lower lattice energy than X and an estimated oligomer attachment rate that is smaller but comparable to that of X (Table 1). This result is consistent with the diminished role of larger oligomers in the growth of polymorphs at higher temperatures. Figure 5 shows that the larger oligomer attachment rate of X estimated by our model is primarily due to larger oligomers; this advantage vanishes if only dimer and trimers are considered. To more accurately predict nucleation (and growth) at smaller supersaturation, more accurate estimates of nucleation barriers and oligomer concentrations need to be employed in our model.

Conclusion

How applicable is our model to the crystallization of real molecules from solution? In its current form, our model neglects explicit solvent effects, which can markedly influence oligomer concentrations and polymorph surface energies. Solvent can also play an important kinetic role in the attachment of molecules and oligomers to the crystal surface, as these growth units need to be partially desolvated before they can be incorporated into the lattice.[55,61] In addition, while organic molecules can have substantial flexibility, our model molecules are rigid. As a result, configurations of oligomeric motifs appearing in crystal structures and in solution are essentially identical in our model. Oligomer motifs in real crystal structures,

however, might rearrange substantially in a solvent environment. This makes identification of low-energy oligomers from real crystal structure less straightforward than in our model, as oligomer transformations and associated energy changes need to be accounted for. Still, we believe that the nucleation score presented in this paper constitutes a significant step towards effective and numerically tractable incorporation of kinetic effects into existing methods of crystal structure prediction. While correlations between oligomer motifs and crystallization are well documented, our work is the first successful attempt to systematically connect energetically favorable oligomeric motifs and lattice energies with crystallization outcomes in a molecular model with realistic polymorph landscapes. We expect that appropriate extensions of our model will be useful for the prediction of chiral separation, crystallization, and co-crystallization of organic and inorganic molecules, as well as the self-assembly of larger particles including proteins and nanoparticles.

Methods

Molecular dynamics simulations.

All molecular dynamics simulations were performed with HOOMD.[62,63] All functional groups of molecules have the same mass m and diameter σ, which we use as our units of mass and length; the unit of energy is ϵ. Langevin equations of motion for rigid bodies are integrated with a time step of $0.004\sqrt{m\sigma^2/\epsilon}$ and a damping coefficient of $5.0\sqrt{m\epsilon/\sigma^2}$. All simulation snapshots were produced with OVITO.[64]

Molecular Interactions

Functional groups (beads) of molecules interact via the short-ranged pair potential

$$u(r) = u_{\mathrm{rep}}(r) + u_{\mathrm{att}}(r).$$

The repulsive part of the potential is of the WCA form,[65]

$$u_{\mathrm{rep}}(r) = \begin{cases} \epsilon_{\mathrm{rep}} \left[\left(\frac{\sigma}{r}\right)^{12} - 2\left(\frac{\sigma}{r}\right)^{6} \right] + \epsilon_{\mathrm{rep}} & \text{if } r < \sigma, \\ 0 & \text{else.} \end{cases}$$

The attractive part is given by

$$u_{\mathrm{att}}(r) = \begin{cases} -\epsilon_{\mathrm{att}}, & \text{if } r < \sigma, \\ -\frac{\epsilon_{\mathrm{att}}}{2} \left(\cos\left[\frac{(r-\sigma)\pi}{\omega} \right] + 1 \right) & \text{if } \sigma \leq r < \sigma + \omega, \\ 0 & \text{if } r \geq \sigma + \omega. \end{cases}$$

We set $\epsilon_{\mathrm{rep}} = 5.0\,\epsilon$ and $\omega = 0.2\,\sigma$. For molecules in set A, we use $\epsilon_{\mathrm{att}} = \epsilon$ for all weakly interacting functional groups and $\epsilon_{\mathrm{att}} = 5\,\epsilon$ for strongly interacting functional groups. The process we used to select attractive interactions between molecules in set B is described below. A comprehensive list of attractive interactions in sets A and B is given in the SI, tables S1 and S2, respectively.

Molecular interactions in set B

Set B contains many molecules that undergo spontaneous chiral separation even though the lowest-energy crystal is racemic. To select molecules that would produce the desired behavior, we only considered molecular shapes s2, s4, s5, s7, and s10, because these shapes tend to crystallize best in our MD simulations.[50] We then generated a set of random interaction vectors using the Bayesian Bootstrap method.[66] Here, the interaction vector $\vec{\epsilon}$ is defined as an ordered list of attractive interactions between all pairs of functional groups (numbered 1–5) of a given molecule,

$$\vec{\epsilon} = (\epsilon_{\mathrm{att},1:1}, \epsilon_{\mathrm{att},1:2}, \ldots, \epsilon_{\mathrm{att},5:5})$$

For each interaction vector, we determined the "heterogeneity" $\varphi_{\vec{\epsilon}}$ of interactions according to

$$\varphi_{\vec{\epsilon}} = \cos^{-1}\left(\frac{\vec{\epsilon_0}\cdot\vec{\epsilon}}{|\vec{\epsilon_0}||\vec{\epsilon}|}\right),$$

where $\vec{\epsilon_0}$ is the uniform interaction vector

$$\vec{\epsilon_0} = (\epsilon, \epsilon, \ldots, \epsilon).$$

We discarded interaction vectors with $\varphi_{\vec{\epsilon}} < 38^\circ$ since we have previously shown that molecules with such interactions have a small likelihood of producing good crystals in MD simulations.[50] For interaction vectors with $\varphi_{\vec{\epsilon}} \geq 38^\circ$, we determined the lowest-energy racemic ($E_0^{(\mathrm{R})}$) and enantiopure ($E_0^{(\mathrm{P})}$) polymorphs using the molecular shapes mentioned above. Approximately 1000 molecules with $0 \leq E_0^{(\mathrm{P})} - E_0^{(\mathrm{R})} \leq 0.4\,\epsilon$/molecule were selected as potential candidates for MD simulations. (These molecules have a racemic polymorph at the lowest energy but its energetic advantage over enantiopure polymorphs is small enough to be overcome by kinetic factors.)

For all candidate molecules, we performed MD simulations using the crystallization protocol described in our previous study.[50] All molecules that either formed an enantiopure crystal or produced a racemic polymorph that had not been observed in set A were selected for set B, resulting in a set of 34 molecules.

Polymorph enumeration (POLYNUM)

POLYNUM uses a numerically efficient exact-cover algorithm to tile periodic unit cells with molecular shapes, exploiting the simple hexagonal symmetry of our molecules. This method naturally generates polymorphs with all symmetries, with different numbers of molecules in the asymmetric unit, and different enantiomer ratios ranging from enantiopure to racemic. While POLYNUM is limited to unit cells containing less than ≈ 15 molecules, it has identified all but one polymorph found in MD simulations in this work (a polymorph with exceptionally

large unit cell), as well as the vast majority of polymorphs for hundreds of other molecules we have studied so far. We are therefore confident that POLYNUM identifies essentially all low-energy polymorphs of a given molecule. Details of POLYNUM are described in Ref. 50.

Nucleation Score

Before deriving the nucleation score $\chi_p = \nu_p \exp(-\eta_p)$, we briefly review common expressions for the nucleation rate $J_p = A_p \exp\left(-\frac{\Delta G_p}{k_B T}\right)$ of a given polymorph p in two dimensions.[53] According to CNT, $A_p \propto \omega_p D_p$, where ω_p is related to the curvature of the free energy barrier at its top and D_p is the effective diffusion coefficient associated with the number of molecules in the nucleus. ω_p can be expressed in terms of the surface tension γ_p (*i.e.*, the line tension in 2D) and $\Delta\mu_p = |\mu_p - \mu_{\text{sol}}|$, the chemical potential difference of molecules in the crystal and solution phase, respectively: $\omega_p \propto \frac{(\Delta\mu_p)^{3/2}}{\gamma_p}$. The diffusion coefficient D_p is related to the rates of attachment $r_{\text{att},p}$ and detachment $r_{\text{det},p}$ of molecules to and from the critical nucleus, respectively. At the barrier top, one can show that $r_{\text{att},p} \approx r_{\text{det},p}$ and therefore $D_p = r_{\text{att},p}$. The rate of attachment can be written as $r_{\text{att},p} \propto C R_p^*$, where C is the concentration of molecules in solution and $R_p^* \propto \frac{\gamma_p}{\Delta\mu_p}$ is the radius of the critical nucleus. The CNT expression for the kinetic prefactor is therefore $A_p \propto C(\Delta\mu_p)^{\frac{1}{2}}$.

In our nucleation score, we neglect the weak dependence of the kinetic prefactor on $\Delta\mu_p$ and focus on the concentration C of molecular building blocks in solution. We assume that the kinetic prefactor for a nucleus of polymorph p is proportional to the total rate of attachment of all oligomeric species in solution that match at least one oligomeric motif found in p. We denote the set of these special oligomers as O_p. We restrict our analysis to hexamers and smaller oligomers. The numerical procedure used to identify these oligomers is described in section . We ignore attachment of monomers (assuming they contribute equally to the growth of all polymorphs) and attachment of oligomeric species that are structurally incompatible with p. The rate of attachment of a specific oligomer i is assumed to be proportional to its concentration C_i in solution and to its likelihood to attach to the nucleus

in the "right" place. This "sticking probability", $\lambda_{i,p}$, depends in complicated ways on the details of the surface of the nucleus and on the structure of the oligomer. For simplicity, we assume that $\lambda_{i,p} \propto s_i/Z'_p$, where Z'_p is the number of molecules in the asymmetric unit cell of p and s_i is the rotational symmetry of the oligomer. Crystals with larger Z'_p have structurally more complex surfaces on average and display a smaller density of attachment points for a given oligomer, resulting in a smaller attachment rate, as illustrated schematically in Figure S4. Oligomers with higher rotational symmetry s_i have a larger probability to attach to the surface in the correct orientation. The contribution of a specific oligomer i comprised of n_i molecules to the total growth rate of the nucleus of polymorph p is therefore proportional to $n_i C_i s_i/Z'_p$. (Differences in the diffusion constants of oligomers are neglected.) The part of the nucleation score describing oligomer attachment, ν_p, is thus obtained by summing over all oligomers of size 2–6,

$$\nu_p = \frac{1}{Z'_p} \sum_{i \in O_p} n_i s_i C_i. \tag{3}$$

While oligomers consisting of more than six molecules can sometimes be observed in our simulations, we show below that satisfactory polymorph predictions can be achieved based on hexamers and smaller species.

We estimate oligomer concentrations C_i straightforwardly based on their energies and geometry. Accurate estimates of oligomer concentrations would require the numerically strenuous calculation of partition functions for all oligomeric species, including loosely bound oligomers and oligomers not found in any low-energy polymorph. To obtain a computationally more tractable estimate of oligomer concentrations, we first assume that the total concentrations of species of size n do not vary much, *i.e.*, that there are, on average, the same total numbers of dimers, trimers, 4-mers, 5-mers, and 6-mers in solution (see Discussion section). Furthermore, we assume that the relative concentrations of oligomers of the same size are well represented by Boltzmann factors and rotational partition functions of energy-optimized configurations. Based on these considerations, we estimate the concentration of

oligomer i of size n_i as

$$C_i \propto \frac{s_i^{-1}\sqrt{I_i}e^{-E_i/k_\mathrm{B}T_\mathrm{c}}}{\sum_{j\in O_{n_i}} s_j^{-1}\sqrt{I_j}e^{-E_j/k_\mathrm{B}T_\mathrm{c}}}. \tag{4}$$

Here, I_i and E_i are the moment of inertia and potential energy of the oligomer i in its optimized configuration, respectively, the sum extends over the set O_n of *all* n-mers that appear in at least one low-energy polymorph (see Methods), and T_c is the temperature at which crystallization is observed. While T_c could be obtained directly from our simulations, it is more convenient to estimate it based on crystal energies. In fact, we find that T_c is approximately proportional to the lowest polymorph energy, $k_\mathrm{B}T_\mathrm{c} \approx \alpha E_0$, where $\alpha = -0.055$, as shown in Fig. S6. (This relation is reminiscent of the well-known relation between crystal melting points and sublimation enthalpies.[67])

We now turn to estimating the polymer-specific nucleation barrier, $\eta_p = \Delta G_p/k_\mathrm{B}T_\mathrm{c}$. According to CNT,

$$\Delta G_p = \frac{\pi\gamma_p^2}{\rho_p \Delta\mu_p},$$

where ρ_p is the number of molecules per unit area in the crystal. We use a single value $\rho_p = \bar{\rho} = 2\sqrt{3}/15\,\sigma^{-2}$ (*i.e.*, the density of a close-packed crystal) for all polymorphs because polymorph packing fractions vary only little across our models. Since the surface tension γ_p cannot be straightforwardly determined from knowledge of the crystal structure alone, we similarly use the same value of $\gamma_p = \bar{\gamma} = 1.78\,\epsilon/\sigma$ for all polymorphs, which we determined as an average over several large crystalline clusters observed in our simulations (see SI).

To obtain an estimate of the polymorph supersaturation $\Delta\mu_p$, we assume that upon cooling a solution of a given molecule, crystallization is observed at a temperature (T_c) at which the nucleation barrier of the lowest-energy polymorph is small enough to be surmounted spontaneously on the simulation time scale. Specifically, we assume that $\Delta G_0/k_\mathrm{B}T_\mathrm{c} \equiv g$, where ΔG_0 is the nucleation barrier of the polymorph with the lowest energy and g is a constant that we treat as a fitting parameter. We find that our model is most predictive for $g = 7.62$. (This value implies $\Delta G_0 = 7.62\,k_\mathrm{B}T_\mathrm{c}$, consistent with the assumption of

spontaneous barrier crossing on the microsecond timescale of simulations.)

With $\Delta G_0/k_\mathrm{B}T_\mathrm{c}$ thus fixed, the nucleation barrier of a given polymorph p can be obtained by expressing $\Delta\mu_p$ in terms of $\Delta\mu_0$,

$$\Delta\mu_p = \Delta\mu_0 - \Delta E_p.$$

Here, $\Delta E_p = E_p - E_0 > 0$ is the difference in lattice energy between polymorph p and the lowest energy polymorph. (Entropic differences between polymorphs are neglected.) We therefore have

$$\frac{\Delta G_p}{k_\mathrm{B}T_\mathrm{c}} = \frac{\pi\bar{\gamma}^2}{k_\mathrm{B}T_\mathrm{c}\bar{\rho}\Delta\mu_p} = \frac{\pi\bar{\gamma}^2}{k_\mathrm{B}T_\mathrm{c}\bar{\rho}(\Delta\mu_0 - \Delta E_p)} = \left(g^{-1} - \frac{k_\mathrm{B}T_\mathrm{c}\bar{\rho}\Delta E_p}{\pi\bar{\gamma}^2}\right)^{-1}.$$

Substituting our estimate for T_c, the barrier part of the nucleation score is therefore given by

$$\eta_p = \left(g^{-1} - \frac{\alpha E_0\bar{\rho}\Delta E_p}{\pi\bar{\gamma}^2}\right)^{-1}. \tag{5}$$

Equations 2, 3, 4, and 5 completely describe the nucleation score χ_p used to rank polymorphs in this work.

Oligomer Enumeration

In order to enumerate all oligomeric motifs of a given polymorph that contain n = 2–6 molecules, we first replicate the unit cell of the polymorph in two dimensions until both edge lengths of the resulting supercell are longer than $16\,\sigma$, ensuring that oligomers do not make contact with their periodic images. We then construct a neighbor list of functional groups belonging to different molecules with a cutoff distance of $1.65\,\sigma$. (This cutoff selects only directly contacting molecules.) From the neighbor list, we generate a graph that represents molecules as nodes and contacts between molecules as edges. An efficient connected induced subgraph algorithm is used to enumerate all subgraphs with n nodes.[68] These subgraphs

represent oligomers of size n within the given polymorph. Resulting oligomers are then checked for uniqueness using *oligomer fingerprints*, which are described in the SI.

Simulations of seeded crystal growth

We create initial configurations for the simulated growth of a given polymorph from a seed as follows. First, we create a single compact crystallite of the polymorph containing at least 500 molecules by replicating its unit cell. To equilibrate the shape of the crystallite, it is then surrounded by a racemic mixture of 4684 molecules in a simulation box of dimension $361.25\,\sigma \times 361.25\,\sigma$. A molecular dynamics simulation is performed at a temperature of $1.5\,\epsilon/k_{\mathrm{B}}$, resulting in slow dissolution of the crystallite. When the crystallite has reached a size of 250 molecules, the simulation is terminated and the configuration is saved. This configuration is then used as the initial condition in growth simulations at different temperatures (Fig. 9 and S5).

Supporting Information

Additional methods, simulation results, and data analysis.

Acknowledgement

The authors thank Ryan Looper and Julio Facelli for useful discussions. The support and resources of the Center for High Performance Computing at the University of Utah are gratefully acknowledged. This work was supported by the National Science Foundation under Grant No. CHE-1900626.

References

(1) Maddox, J. Crystals from first principles. *Nature* **1988**, *335*, 201–201.

(2) von Raumer, M.; Hilfiker, R. *Polymorphism in the Pharmaceutical Industry*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2018; pp 1–30.

(3) Lee, A. Y.; Erdemir, D.; Myerson, A. S. Crystal Polymorphism in Chemical Process Development. *Annual Review of Chemical and Biomolecular Engineering* **2011**, *2*, 259–280.

(4) Price, S. L.; Price, L. S. In *Polymorphism in the Pharmaceutical Industry: Solid Form and Drug Development*; Brittain, H. G., Ed.; CRC Press, 2018; pp 133–157.

(5) Yang, J.; Zhu, X.; Hu, C. T.; Qiu, M.; Zhu, Q.; Ward, M. D.; Kahr, B. Inverse Correlation between Lethality and Thermodynamic Stability of Contact Insecticide Polymorphs. *Crystal Growth & Design* **2019**, *19*, 1839–1844.

(6) Zhu, X.; Hu, C. T.; Yang, J.; Joyce, L. A.; Qiu, M.; Ward, M. D.; Kahr, B. Manipulating Solid Forms of Contact Insecticides for Infectious Disease Prevention. *Journal of the American Chemical Society* **2019**, *141*, 16858–16864.

(7) Yang, J.; Erriah, B.; Hu, C. T.; Reiter, E.; Zhu, X.; López-Mejías, V.; Carmona-Sepúlveda, I. P.; Ward, M. D.; Kahr, B. A deltamethrin crystal polymorph for more effective malaria control. *Proceedings of the National Academy of Sciences* **2020**, *117*, 26633–26638.

(8) Olenik, B.; Thielking, G. *Modern Methods in Crop Protection Research*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2013; pp 249–272.

(9) Tang, J.; Cheng, G.; Zhao, Y.; Yang, P.; Ju, X.; Yang, H. Optimizing the molecular structure and packing style of a crystal by intramolecular cyclization from picrylhydrazone to indazole. *CrystEngComm* **2019**, *21*, 4701–4706.

(10) Fondren, N. S.; Fondren, Z. T.; Unruh, D. K.; Weeks, B. L. Effects of Solution Con-

ditions on Polymorph Development in 2,4,6-Trinitrotoluene. *Crystal Growth & Design* **2020**, *20*, 568–579.

(11) Ma, Q.; Lu, Z.; Liao, L.; Huang, J.; Liu, D.; Li, J.; Fan, G. 5,6-Di(2-fluoro-2,2-dinitroethoxy)furazano[3,4-b]pyrazine: a high performance melt-cast energetic material and its polycrystalline properties. *RSC Advances* **2017**, *7*, 38844–38852.

(12) Tang, J.; Cheng, G.; Feng, S.; Zhao, X.; Zhang, Z.; Ju, X.; Yang, H. Boosting Performance and Safety of Energetic Materials by Polymorphic Transition. *Crystal Growth & Design* **2019**, *19*, 4822–4828.

(13) Yan, T.; Cheng, G.; Yang, H. 1,2,4-Oxadiazole-Bridged Polynitropyrazole Energetic Materials with Enhanced Thermal Stability and Low Sensitivity. *ChemPlusChem* **2019**, *84*, 1567–1577.

(14) Hoja, J.; Ko, H.-Y.; Neumann, M. A.; Car, R.; DiStasio, R. A.; Tkatchenko, A. Reliable and practical computational description of molecular crystal polymorphs. *Science Advances* **2019**, *5*, eaau3338.

(15) Reilly, A. M. et al. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2016**, *72*, 439–459.

(16) Egorova, O.; Hafizi, R.; Woods, D. C.; Day, G. M. Multifidelity Statistical Machine Learning for Molecular Crystal Structure Prediction. *The Journal of Physical Chemistry A* **2020**, *124*, 8065–8078.

(17) McDonagh, D.; Skylaris, C.-K.; Day, G. M. Machine-Learned Fragment-Based Energies for Crystal Structure Prediction. *Journal of Chemical Theory and Computation* **2019**, *15*, 2743–2758.

(18) Yang, M. et al. Prediction of the Relative Free Energies of Drug Polymorphs above Zero Kelvin. *Crystal Growth and Design* **2020**, *20*, 5211–5224.

(19) Abraham, N. S.; Shirts, M. R. Statistical Mechanical Approximations to More Efficiently Determine Polymorph Free Energy Differences for Small Organic Molecules. *Journal of Chemical Theory and Computation* **2020**, *16*, 6503–6512.

(20) Stahly, G. P. Diversity in Single- and Multiple-Component Crystals. The Search for and Prevalence of Polymorphs and Cocrystals. *Crystal Growth & Design* **2007**, *7*, 1007–1026.

(21) Davey, R. J.; Schroeder, S. L.; Ter Horst, J. H. Nucleation of organic crystals - A molecular perspective. *Angewandte Chemie - International Edition* **2013**, *52*, 2167–2179.

(22) Greenwell, C.; McKinley, J. L.; Zhang, P.; Zeng, Q.; Sun, G.; Li, B.; Wen, S.; Beran, G. J. O. Overcoming the difficulties of predicting conformational polymorph energetics in molecular crystals via correlated wavefunction methods. *Chemical Science* **2020**, *11*, 2200–2214.

(23) Price, S. L. Why don't we find more polymorphs? *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2013**, *69*, 313–328.

(24) Cruz-Cabeza, A. J. Crystal structure prediction: Are we there yet? *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2016**, *72*, 437–438.

(25) Montis, R.; Davey, R. J.; Wright, S. E.; Woollam, G. R.; Cruz-Cabeza, A. J. Transforming Computed Energy Landscapes into Experimental Realities: The Role of Structural Rugosity. *Angewandte Chemie* **2020**, *132*, 20537–20540.

(26) Sosso, G. C.; Chen, J.; Cox, S. J.; Fitzner, M.; Pedevilla, P.; Zen, A.; Michaelides, A. Crystal Nucleation in Liquids: Open Questions and Future Challenges in Molecular Dynamics Simulations. *Chemical Reviews* **2016**, *116*, 7078–7116.

(27) Chang, Y. C.; Myerson, A. S. Diffusivity of glycine in concentrated saturated and supersaturated aqueous solutions. *AIChE Journal* **1986**, *32*, 1567–1569.

(28) Kellermeier, M.; Rosenberg, R.; Moise, A.; Anders, U.; Przybylski, M.; Cölfen, H. Amino acids form prenucleation clusters: ESI-MS as a fast detection method in comparison to analytical ultracentrifugation. *Faraday Discussions* **2012**, *159*, 23–45.

(29) Koch, K. J.; Gozzo, F. C.; Nanita, S. C.; Takats, Z.; Eberlin, M. N.; Cooks, R. G. Chiral Transmission between Amino Acids: Chirally Selective Amino Acid Substitution in the Serine Octamer as a Possible Step in Homochirogenesis. *Angewandte Chemie - International Edition* **2002**, *41*, 1721–1724.

(30) Takats, Z.; Nanita, S. C.; Cooks, R. G.; Schlosser, G.; Vekey, K. Amino acid clusters formed by sonic spray ionization. *Analytical Chemistry* **2003**, *75*, 1514–1523.

(31) Toyama, N.; Kohno, J. Y.; Mafuné, F.; Kondow, T. Solvation structure of arginine in aqueous solution studied by liquid beam technique. *Chemical Physics Letters* **2006**, *419*, 369–373.

(32) Yang, P.; Xu, R.; Nanita, S. C.; Cooks, R. G. Thermal formation of homochiral serine clusters and implications for the origin of homochirality. *Journal of the American Chemical Society* **2006**, *128*, 17074–17086.

(33) Hughes, C. E.; Hamad, S.; Harris, K. D. M.; Catlow, C. R. A.; Griffiths, P. C. A multitechnique approach for probing the evolution of structural properties during crystallization of organic materials from solution. *Faraday Discussions* **2007**, *136*, 71.

(34) Hamad, S.; Hughes, C. E.; Catlow, C. R. A.; Harris, K. D. Clustering of glycine molecules in aqueous solution studied by molecular dynamics simulation. *Journal of Physical Chemistry B* **2008**, *112*, 7280–7288.

(35) Nemes, P.; Schlosser, G.; Vékey, K. Amino acid cluster formation studied by electrospray ionization mass spectrometry. *Journal of Mass Spectrometry* **2005**, *40*, 43–49.

(36) Zhang, D.; Wu, L.; Koch, K. J.; Cooks, R. G. Arginine clusters generated by electrospray ionization and identified by tandem mass spectrometry. *European Journal of Mass Spectrometry* **1999**, *5*, 353–361.

(37) Hunter, C. A.; McCabe, J. F.; Spitaleri, A. Solvent effects of the structures of prenucleation aggregates of carbamazepine. *CrystEngComm* **2012**, *14*, 7115–7117.

(38) Spitaleri, A.; Hunter, C. A.; McCabe, J. F.; Packer, M. J.; Cockroft, S. L. A 1H NMR study of crystal nucleation in solution. *CrystEngComm* **2004**, *6*, 489.

(39) Chiarella, R. A.; Gillon, A. L.; Burton, R. C.; Davey, R. J.; Sadiq, G.; Auffret, A.; Cioffi, M.; Hunter, C. A. The nucleation of inosine: The impact of solution chemistry on the appearance of polymorphic and hydrated crystal forms. *Faraday Discussions* **2007**, *136*, 179–193.

(40) Parveen, S.; Davey, R. J.; Dent, G.; Pritchard, R. G. Linking solution chemistry to crystal nucleation: The case of tetrolic acid. *Chemical Communications* **2005**, 1531–1533.

(41) Jie, C.; Trout, B. L. Computational study of solvent effects on the molecular self-assembly of tetrolic acid in solution and implications for the polymorph formed from crystallization. *Journal of Physical Chemistry B* **2008**, *112*, 7794–7802.

(42) Kulkarni, S. A.; McGarrity, E. S.; Meekes, H.; Ter Horst, J. H. Isonicotinamide self-

association: The link between solvent and polymorph nucleation. *Chemical Communications* **2012**, *48*, 4983–4985.

(43) Chadwick, K.; Davey, R. J.; Dent, G.; Pritchard, R. G.; Hunter, C. A.; Musumeci, D. Cocrystallization: A solution chemistry perspective and the case of benzophenone and diphenylamine. *Crystal Growth and Design* **2009**, *9*, 1990–1999.

(44) Gebauer, D.; Kellermeier, M.; Gale, J. D.; Bergström, L.; Cölfen, H. Pre-nucleation clusters as solute precursors in crystallisation. *Chemical Society Reviews* **2014**, *43*, 2348–2371.

(45) Hamad, S.; Moon, C.; Richard, C.; Catlow, A.; Hulme, A. T.; Price, S. L. Kinetic insights into the role of the solvent in the polymorphism of 5-fluorouracil from molecular dynamics simulations. *Journal of Physical Chemistry B* **2006**, *110*, 3323–3329.

(46) Gebauer, D.; Cölfen, H. Prenucleation clusters and non-classical nucleation. *Nano Today* **2011**, *6*, 564–584.

(47) Peral, F.; Gallego, E. Self-association of imidazole and its methyl derivatives in aqueous solution. A study by ultraviolet spectroscopy. *Journal of Molecular Structure* **1997**, *415*, 187–196.

(48) Burton, R. C.; Ferrari, E. S.; Davey, R. J.; Finney, J. L.; Bowron, D. T. The Relationship between Solution Structure and Crystal Nucleation: A Neutron Scattering Study of Supersaturated Methanolic Solutions of Benzoic Acid. *The Journal of Physical Chemistry B* **2010**, *114*, 8807–8816.

(49) Desiraju, G. R. Crystal engineering: A holistic view. *Angewandte Chemie - International Edition* **2007**, *46*, 8342–8356.

(50) Carpenter, J. E.; Grünwald, M. Heterogeneous Interactions Promote Crystallization

and Spontaneous Resolution of Chiral Molecules. *Journal of the American Chemical Society* **2020**, *142*, 10755–10768.

(51) Pillong, M.; Marx, C.; Piechon, P.; Wicker, J. G.; Cooper, R. I.; Wagner, T. A publicly available crystallisation data set and its application in machine learning. *CrystEngComm* **2017**, *19*, 3737–3745.

(52) Neumann, M. A.; van de Streek, J. How many ritonavir cases are there still out there? *Faraday Discussions* **2018**, *211*, 441–458.

(53) Jungblut, S.; Dellago, C. Pathways to self-organization: Crystallization via nucleation and growth. *The European Physical Journal E* **2016**, *39*, 77.

(54) Deij, M. A.; Ter Horst, J. H.; Meekes, H.; Jansens, P.; Vlieg, E. Polymorph formation studied by 3D nucleation simulations. Application to a yellow isoxazolone dye, paracetamol, and L-glutamic acid. *Journal of Physical Chemistry B* **2007**, *111*, 1523–1530.

(55) Davey, R. J.; Back, K. R.; Sullivan, R. A. Crystal nucleation from solutions - Transition states, rate determining steps and complexity. *Faraday Discussions* **2015**, *179*, 9–26.

(56) Sullivan, R. A.; Davey, R. J.; Sadiq, G.; Dent, G.; Back, K. R.; Ter Horst, J. H.; Toroz, D.; Hammond, R. B. Revealing the roles of desolvation and molecular self-assembly in crystal nucleation from solution: Benzoic and p -aminobenzoic acids. *Crystal Growth and Design* **2014**, *14*, 2689–2696.

(57) Bernstein, J.; Dunitz, J. D.; Gavezzotti, A. Polymorphic perversity: Crystal structures with many symmetry-independent molecules in the unit cell. *Crystal Growth and Design* **2008**, *8*, 2011–2018.

(58) Cruz-Cabeza, A. J.; Davey, R. J.; Sachithananthan, S. S.; Smith, R.; Tang, S. K.; Vetter, T.; Xiao, Y. Aromatic stacking-a key step in nucleation. *Chemical Communications* **2017**, *53*, 7905–7908.

(59) Cruz-Cabeza, A. J.; Feeder, N.; Davey, R. J. Open questions in organic crystal polymorphism. *Communications Chemistry* **2020**, *3*, 142.

(60) Warzecha, M.; Verma, L.; Johnston, B. F.; Palmer, J. C.; Florence, A. J.; Vekilov, P. G. Olanzapine crystal symmetry originates in preformed centrosymmetric solute dimers. *Nature Chemistry* **2020**, *12*, 914–920.

(61) Dunning, W. J.; Shipman, A. J. Nucleation in Sucrose Solutions. *Proc. Agric. Industries 10th International Conference* **1954**, 1448–1456.

(62) Glaser, J.; Nguyen, T. D.; Anderson, J. A.; Lui, P.; Spiga, F.; Millan, J. A.; Morse, D. C.; Glotzer, S. C. Strong scaling of general-purpose molecular dynamics simulations on GPUs. *Computer Physics Communications* **2015**, *192*, 97–107.

(63) Anderson, J. A.; Lorenz, C. D.; Travesset, A. General purpose molecular dynamics simulations fully implemented on graphics processing units. *Journal of Computational Physics* **2008**, *227*, 5342–5359.

(64) Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool. *Modelling and Simulation in Materials Science and Engineering* **2010**, *18*, 015012.

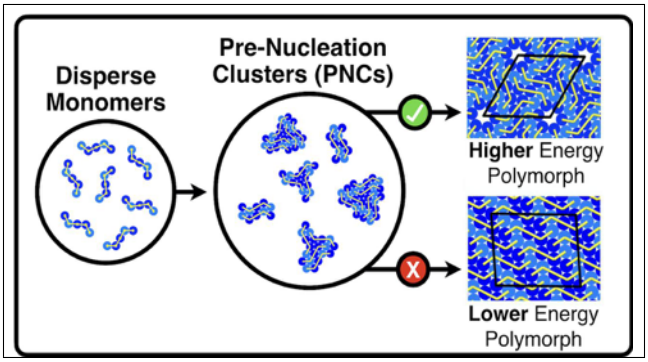
(65) Weeks, J. D.; Chandler, D.; Andersen, H. C. Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids. *The Journal of Chemical Physics* **1971**, *54*, 5237–5247.

(66) Rubin, D. B. The Bayesian Bootstrap. *The Annals of Statistics* **1981**, *9*, 130–134.

(67) Salahinejad, M.; Le, T. C.; Winkler, D. A. Capturing the Crystal: Prediction of Enthalpy of Sublimation, Crystal Lattice Energy, and Melting Points of Organic Compounds. *Journal of Chemical Information and Modeling* **2013**, *53*, 223–229.

(68) Komusiewicz, C.; Sommer, F. Enumerating Connected Induced Subgraphs: Improved Delay and Experimental Comparison. SOFSEM 2019: Theory and Practice of Computer Science. Cham, 2019; pp 272–284.

Graphical TOC Entry



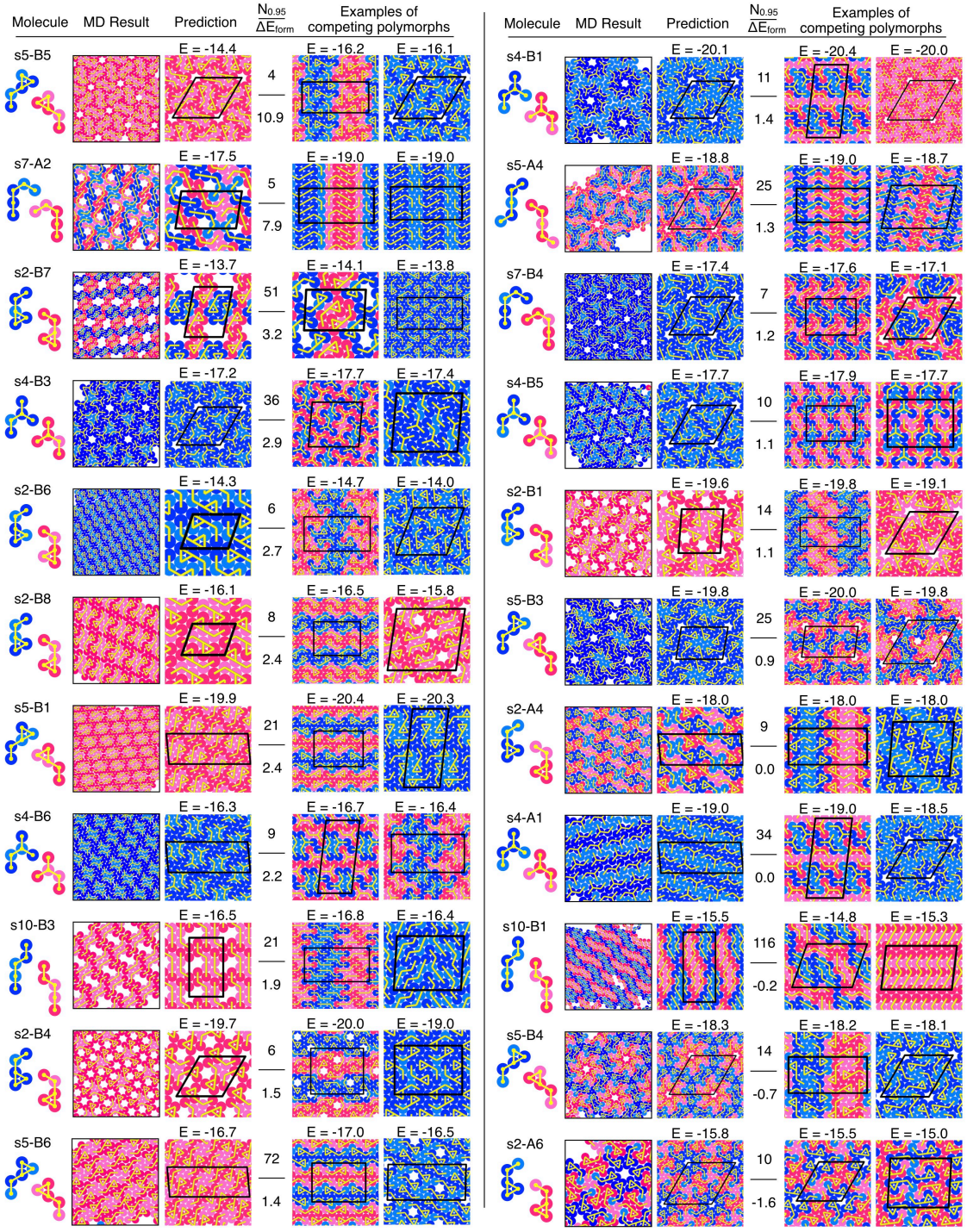


Figure 1: Examples of the 65 chiral molecules studied in this paper sorted by ΔE_{form}. For each molecule, we show from left to right: space-filling representation of the two enantiomers (light colors indicate functional groups with strong interactions); snapshot of the largest crystalline cluster observed in MD simulations and the bulk energy of the polymorph; number $N_{0.95}$ of energetically competing polymorphs; unit cells of examples of competing polymorphs and their energies. Crystal energies are given in units of ϵ per molecule. ΔE_{form} values are given in percent. Interactions between functional groups are specified in the SI.

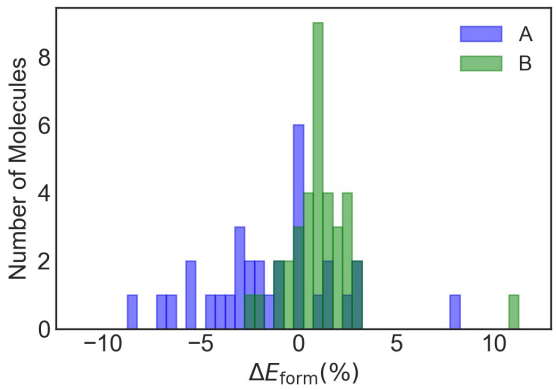


Figure 2: Histograms of ΔE_{form} for molecules in sets A (blue) and B (green).

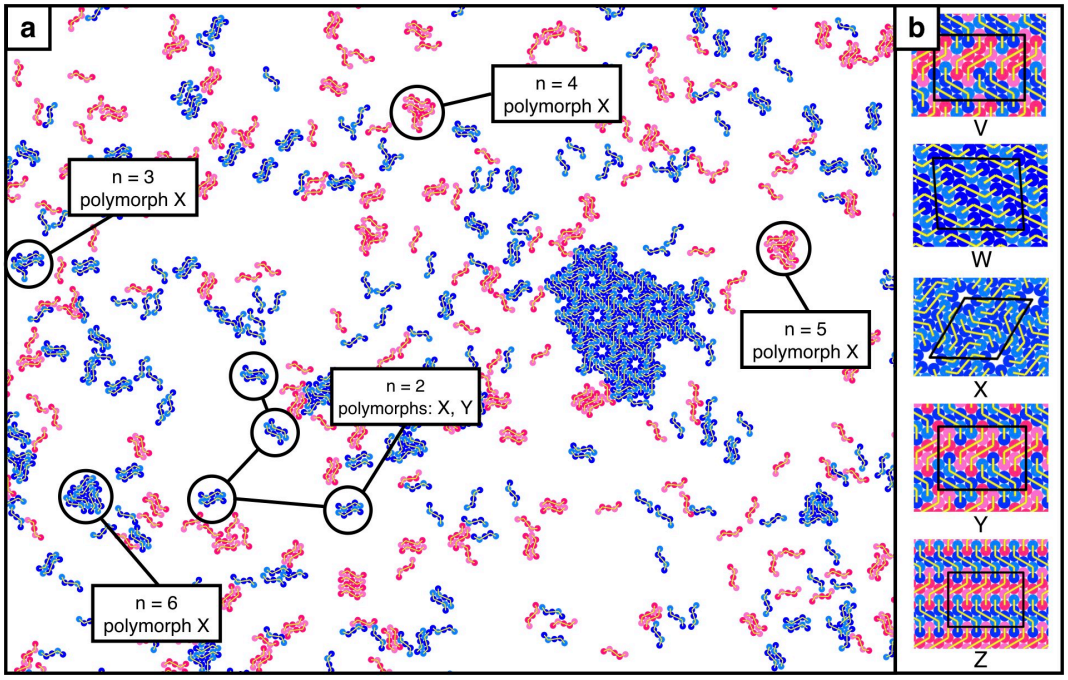


Figure 3: Crystallization of molecule s9-A1. (a) Snapshot from an MD simulation of a racemic mixture of s9-A1, showing formation of a cluster of polymorph X. Various oligomers of size n are highlighted and labeled according to the polymorph(s)they occur in. (b) Unit cells of select low energy polymorphs ($E_{\mathrm{X}} = -18.5\epsilon$, $E_{\mathrm{V}} = E_{\mathrm{W}} = E_{\mathrm{Y}} = E_{\mathrm{Z}} = -19.0\epsilon$).

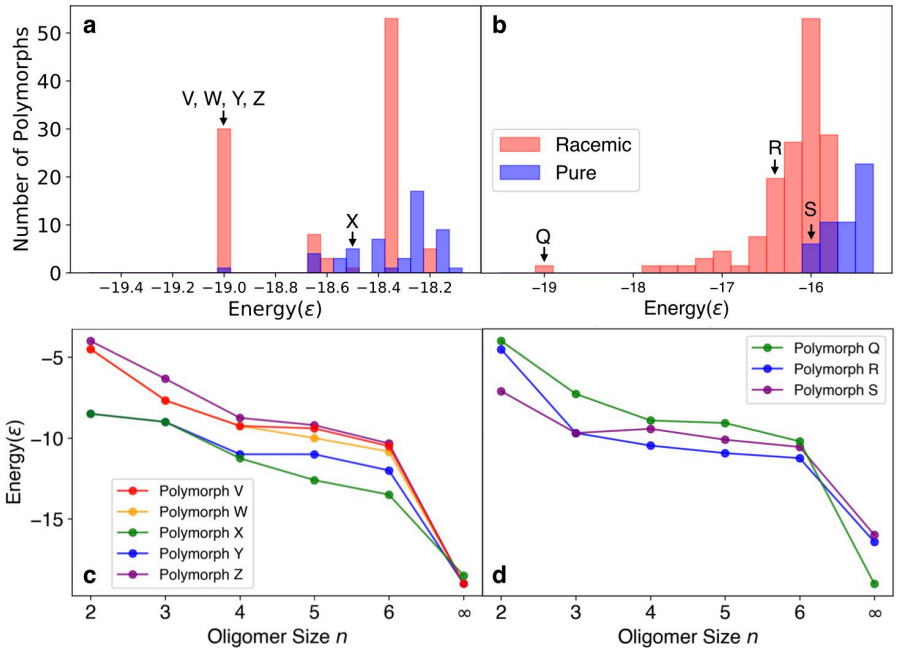


Figure 4: Thermodynamic landscapes of polymorphs and oligomers of molecules s9-A1 and s7-A1. (a-b) Histogram of polymorph energies of (a) s9-A1 and (b) s7-A1. (c-d) Energies per molecule of most stable oligomers of size n found in select polymorphs of molecules (c) s9-A1 and (d) s7-A1. $n = \infty$ indicates lattice energy per molecule.

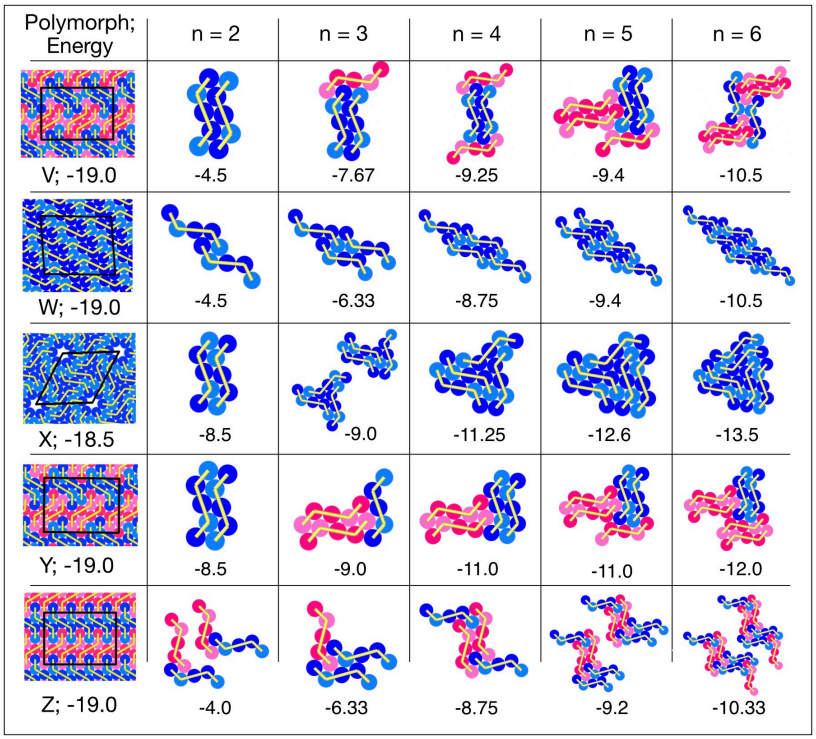


Figure 5: MD snapshots and energies (per molecule) of the most stable oligomers of size n found in select polymorphs of molecule s9-A1. Polymorph unit cells and lattice energies are shown in the left-most column.

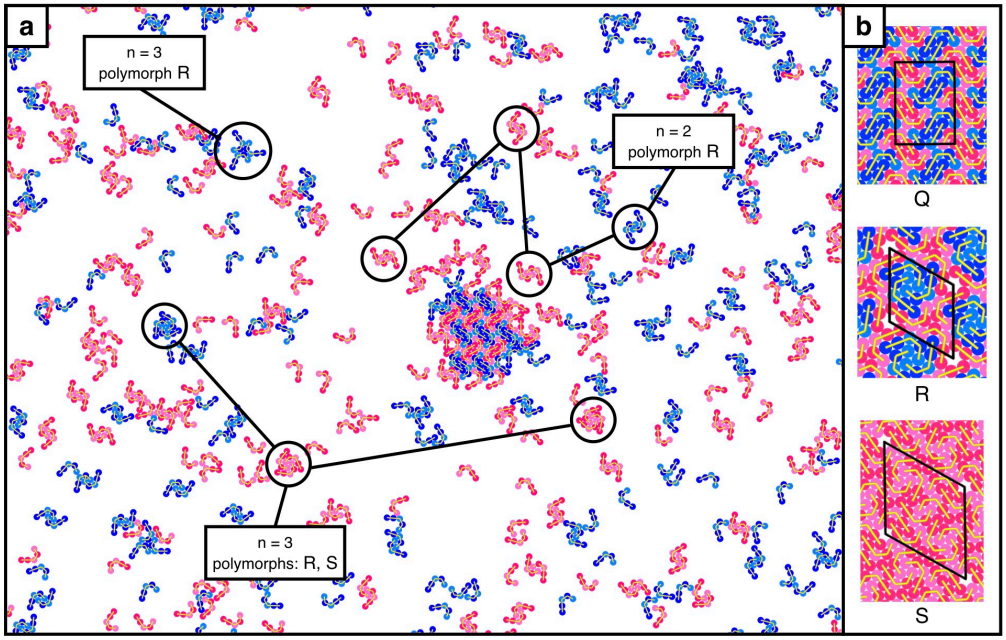


Figure 6: Crystallization of molecule s7-A1. (a) Snapshot from an MD simulation of a racemic mixture of s7-A1, showing formation of a cluster of polymorph Q. Various oligomers of size n are highlighted and labeled according to the polymorph(s) they occur in. (b) Unit cells of select low energy polymorphs ($E_{\mathrm{Q}} = -19.0\epsilon$, $E_{\mathrm{R}} - 16.5\epsilon$, and $E_{\mathrm{S}} - 16.1\epsilon$).

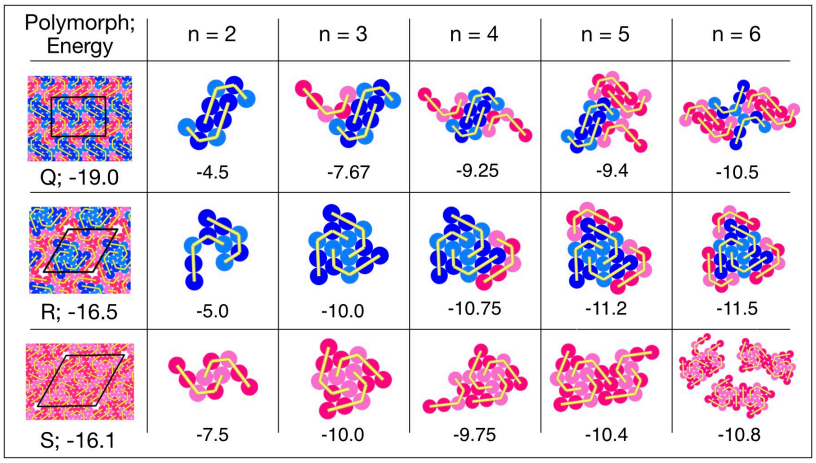


Figure 7: MD snapshots and energies (per molecule) of the most stable oligomers of size n found in select polymorphs of molecule s7-A1. Polymorph unit cells and lattice energies are shown in the left-most column.

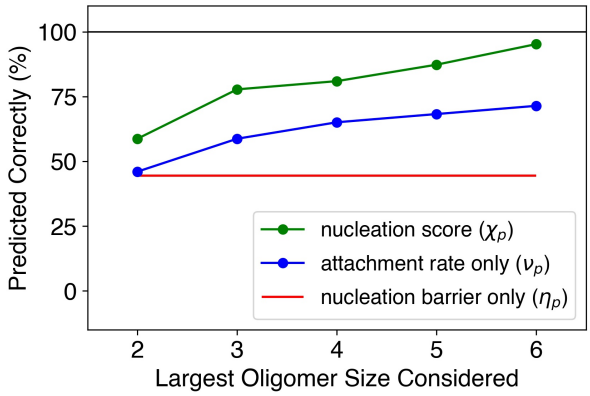


Figure 8: Fraction of correctly predicted polymorphs as a function of the largest oligomer size included in χ_p (green). The red line represents the success rate (44%) based on lattice energies alone (assuming that the correct polymorph is selected in half of all cases with multiple polymorphs at the lowest energy). The success rate using only estimated oligomer attachment rates ν_p (disregarding lattice energies) is shown in blue color.

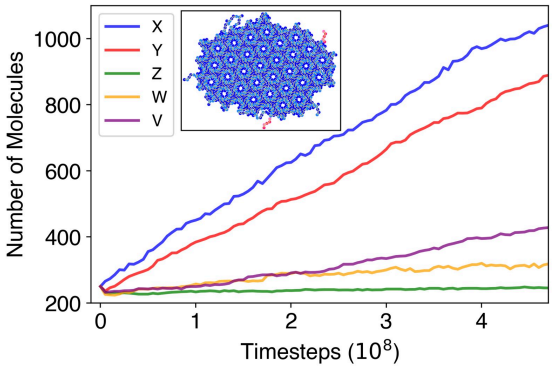


Figure 9: Average number of molecules in growing nuclei of polymorphs X, Y, Z, W, and V of molecule s9-A1 at a temperature of $T = 1.125\,\epsilon/k_{\mathrm{B}}$, slightly above the temperature at which spontaneous nucleation is observed. The inset shows a snapshot of a crystallite of polymorph X containing 254 molecules, at the beginning of the simulation.