

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

John E. Carpenter and Michael Grünwald<sup>a)</sup>

*Department of Chemistry, University of Utah, Salt Lake City, 84112 Utah*

(Dated: 15 July 2021)

Kinetics can play an important role in the crystallization of molecules and give rise to polymorphism, the prevalent ability 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 are thought to be complex and accurate rate calculations are numerically cumbersome. In this work, we use molecular dynamics computer simulations to study a simple molecular model that reproduces the crystallization behavior of real chiral molecules, including the formation of enantiopure and racemic crystals, as well as polymorphism. We show that in many cases, the crystal that robustly forms in simulations is not the one with the lowest free energy. We demonstrate that at high supersaturation the prevailing polymorph can be accurately predicted by considering the similarities between prevalent 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.

## I. INTRODUCTION

Predicting which crystal structure a given molecule will form is a long-standing problem<sup>1</sup> with considerable practical significance for the industrial production of many chemical compounds, including medicinal drugs,<sup>2-4</sup> pesticides,<sup>5-8</sup> and explosives.<sup>9-13</sup> 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.<sup>14-19</sup> 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.<sup>20</sup> 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.<sup>21</sup>

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.<sup>18,22,23</sup> 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.<sup>24</sup> 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.

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.<sup>25</sup> 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.<sup>26-35</sup> These oligomers can act as important building blocks in the nucleation and growth of specific polymorphs.<sup>36-44</sup> Concentrations of amino acid oligomers in solution, for instance, have been studied extensively using different techniques.<sup>26-35</sup> Electrospray ionization experiments have shown strong evidence of oligomerization of amino acids at low concentrations<sup>34</sup> and suggest the existence of ‘magic number’ oligomers with particularly large concentrations (e.g., tetramers of arginine).<sup>35</sup> 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.<sup>27</sup> Oligomeric species that serve as precursors for specific crystal structures are often referred to as pre-nucleation clusters (PNCs).<sup>45</sup> Substantial populations of PNCs in solution prior to crystallization

---

<sup>a)</sup>Electronic mail: michael.gruenwald@utah.edu

have been observed in a range of systems with spectroscopic techniques<sup>38,39,41,42,46,47</sup> The idea that oligomers can foreshadow the crystal structure that will eventually form has given rise to the concept of the "synthon", which refers to the dominant binding motifs in oligomeric species.<sup>48</sup> 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.<sup>21</sup> 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<sup>49</sup> 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 estimated nucleation rates successfully balance 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.

## II. RESULTS AND DISCUSSION

### A. Molecular model and crystallization simulations

We simulated the crystallization of racemic mixtures of simple chiral molecules in two dimension. Model molecules are rigid and consist of 5 beads that represent different functional groups and interact via short-ranged attractive pair potentials (see Methods section). By varying the spatial arrangement of functional groups and their interaction strengths, 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 categories of products found in experiments: racemic crystals, enantiopure crystals (via spontaneous chiral separation) and several types of partially ordered and disordered solids. 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 thermodynamically preferred polymorphs (set B).

For each of these molecules, we performed crystallization simulations by placing racemic mixtures of 5184 molecules in square simulation boxes at a packing frac-

tion 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. 49. 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.)

### B. 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 small, polymorph *free* energies typically deviate from lattice energies by less than 1%.<sup>49</sup> 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  $E_{\text{form}}$  is the energy of the observed polymorph and  $E_{\text{comp}}$  is energy of the most stable competing polymorph, i.e., the smallest energy of all polymorphs that did *not* form. Negative values of  $\Delta E_{\text{form}}$  therefore indicate that the observed polymorph is the thermodynamic equilibrium structure, while cases of kinetically driven polymorph formation are indicated by positive values of  $\Delta E_{\text{form}}$ .

The polymorph energy landscapes 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.<sup>50</sup> 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

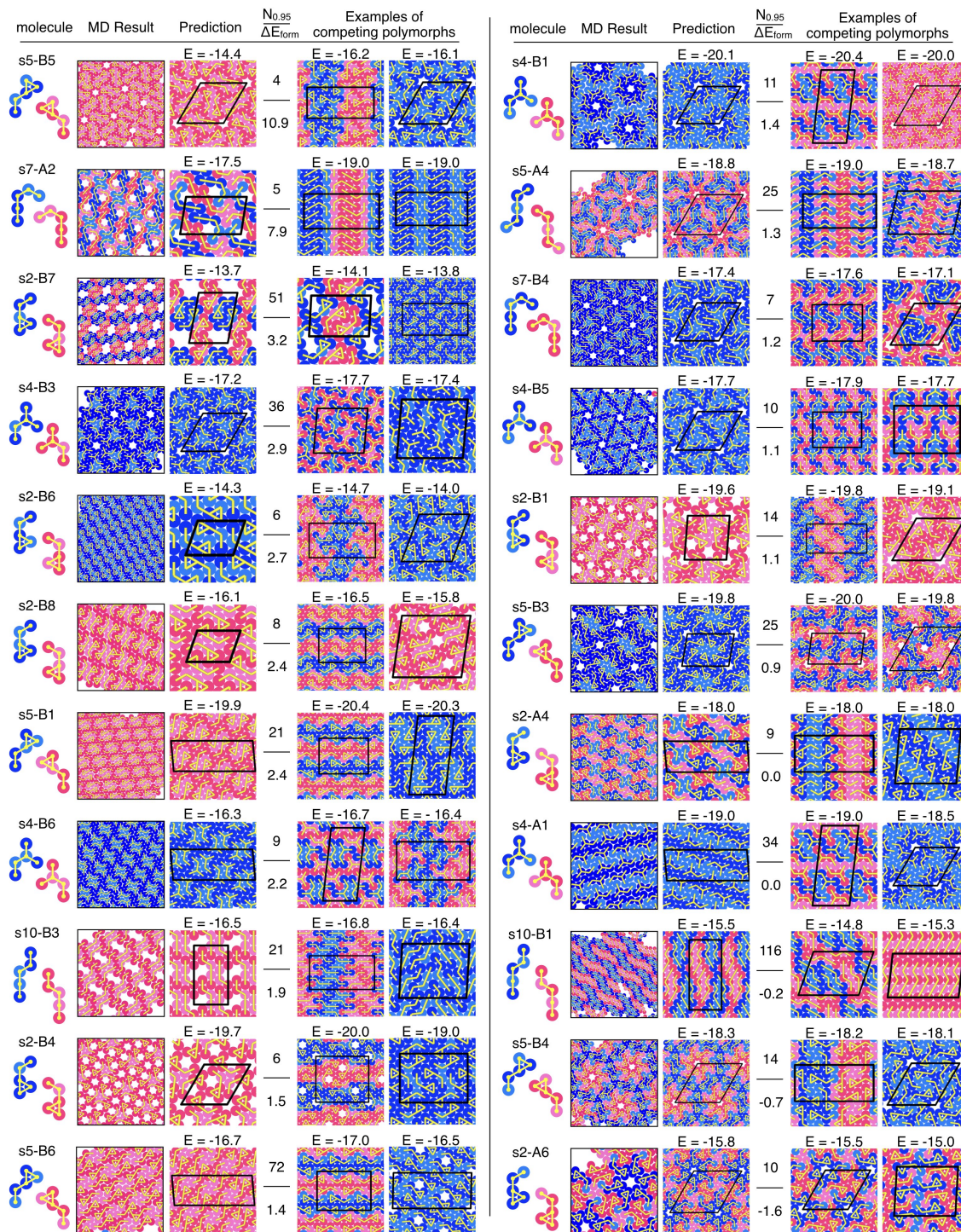


FIG. 1. Examples of the 65 chiral molecules studied in this paper sorted by  $\Delta E_{\text{form}}$ . For each molecule, we show from left to right: space-filling representation of the two enantiomers (light colors indicate functional groups with strong interactions); snapshot of the largest crystalline 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. Energies are given in units of  $\epsilon$  per molecule.  $\Delta E_{\text{form}}$  values are given in percent. Interactions between functional groups are specified in the SI.

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

within 95% of the lowest energy.  $N_{0.95}$  varies substantially between molecules, but in all but a few cases we find at least several (and up to  $\approx 100$ ) competing poly-



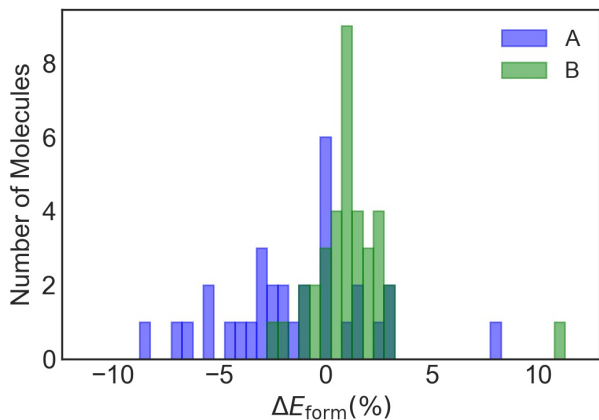


FIG. 2. Histograms of  $\Delta E_{\text{form}}$  for molecules in sets A (blue) and B (green).

morphs at low energies. Numerical values of  $E_{\text{form}}$  and  $N_{0.95}$  for all molecules are given in Figure 1, Figure S2, and Figure S3.

### C. 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 ( $E_{\text{form}} = 0.03$ ,  $N_{0.95} = 168$ ). 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) 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  $k_{\text{B}}T$  at the temperature of crystal formation ( $T \approx 1.0 \epsilon/k_{\text{B}}$ ), 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 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.

### D. 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 chose to estimate crystallization rates in our simulations based on classical nucleation theory. How-

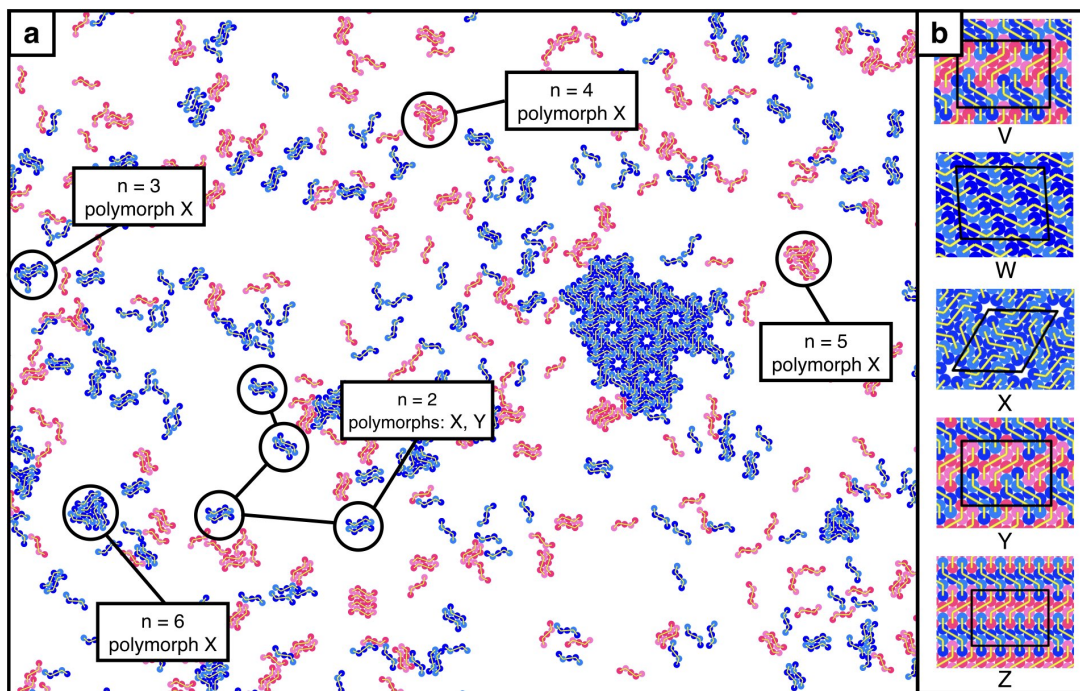


FIG. 3. Crystallization of molecule s9-A1. (a) Snapshot from an MD simulation of a racemic mixture of s9-A1, showing formation of a cluster of polymorph X. Various oligomers of size  $n$  are highlighted and labeled according to the polymorph(s) they occur in. (b) Unit cells of prominent low energy polymorphs ( $E_X = -18.5\epsilon$ ,  $E_V = E_W = E_Y = E_Z = -19.0\epsilon$ ).

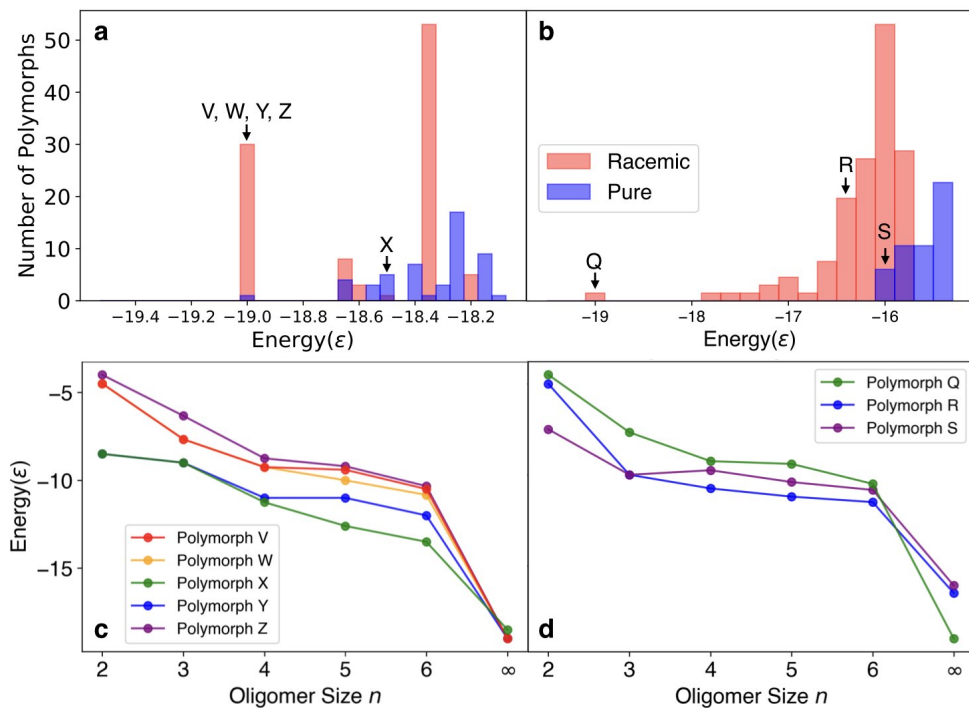


FIG. 4. Thermodynamic landscapes of polymorphs and oligomers of molecules s9-A1 and s7-A1. Histogram of polymorph energies of (a) s9-A1 and (b) s7-A1. (c) Energies per molecule of most stable oligomers of size  $n$  found in prominent polymorphs of molecules (c) s9-A1 and (d) s7-A1.  $n = \infty$  indicates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

ever, the kinetic factors associated with attachment of oligomeric species, which are the focus of this work, are

relevant also for crystal growth rates. We demonstrate below for one case (molecule s9-A1) that our nucleation

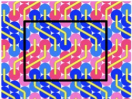



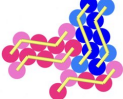
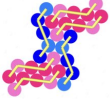
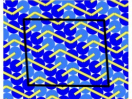
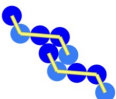
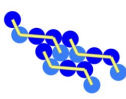
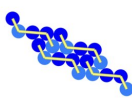
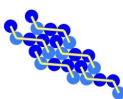
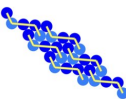
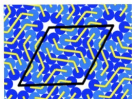

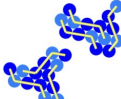
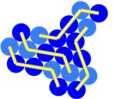
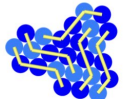
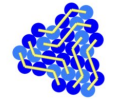
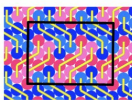
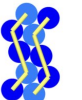
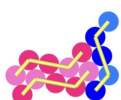
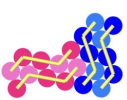
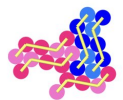
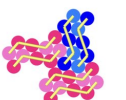
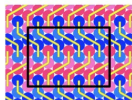
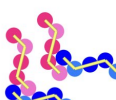
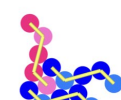
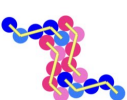
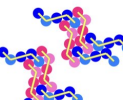
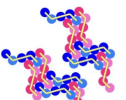
Polymorph; Energy	n = 2	n = 3	n = 4	n = 5	n = 6
 V; -19.0	 -4.5	 -7.67	 -9.25	 -9.4	 -10.5
 W; -19.0	 -4.5	 -6.33	 -8.75	 -9.4	 -10.5
 X; -18.5	 -8.5	 -9.0	 -11.25	 -12.6	 -13.5
 Y; -19.0	 -8.5	 -9.0	 -11.0	 -11.0	 -12.0
 Z; -19.0	 -4.0	 -6.33	 -8.75	 -9.2	 -10.33

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

rate estimates correctly identify also the fastest growing polymorph.

According to classical nucleation theory, the rate at which super-critical nuclei are produced per unit area in two dimensions is given by<sup>51</sup>

$$J = A \exp\left(-\frac{\Delta G}{k_B T}\right), \quad (1)$$

where  $A$  is the kinetic prefactor and  $\Delta G$  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. The kinetic prefactor is often neglected entirely in the analysis of nucleation rates of different polymorphs, under the assumption that the nucleation barrier is the most important term.<sup>19,52</sup> Other studies have shown that, on the contrary,  $A$  can be a decisive factor.<sup>53,54</sup> 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"  $\chi_p$ , which is proportional to the CNT nucleation rate (under assumptions described below) and given by

$$\chi_p = \nu_p \exp(-\eta_p). \quad (2)$$

Here,  $\nu_p$  is a dimensionless quantity proportional to the



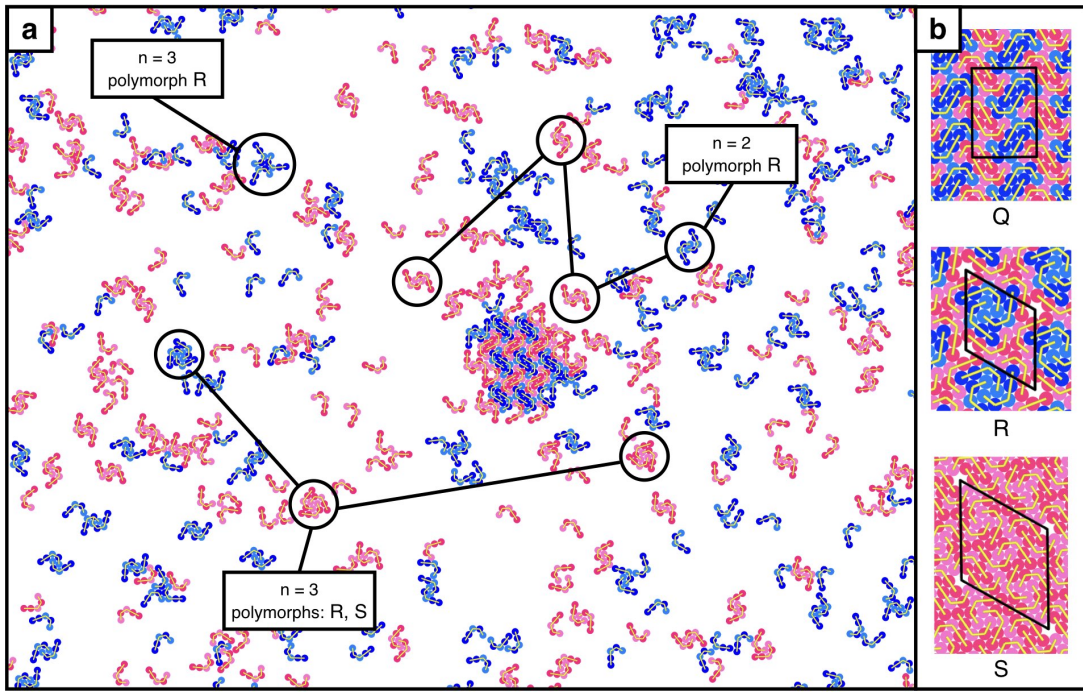


FIG. 6. Crystallization of molecule s7-A1. (a) Snapshot from an MD simulation of a racemic mixture of s7-A1, showing formation of a cluster of polymorph Q. Various oligomers of size  $n$  are highlighted and labeled according to the polymorph(s) they occur in. (b) Unit cells of prominent low energy polymorphs ( $E_Q = -19.0\epsilon$ ,  $E_R = -16.5\epsilon$ , and  $E_S = -16.1\epsilon$ ).

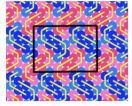


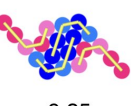

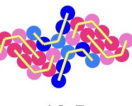
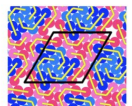


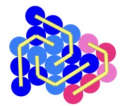


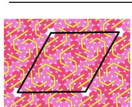

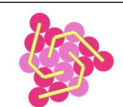
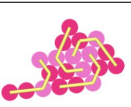
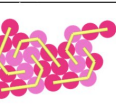
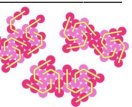
Polymorph; Energy	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
 Q; -19.0	 -4.5	 -7.67	 -9.25	 -9.4	 -10.5
 R; -16.5	 -5.0	 -10.0	 -10.75	 -11.2	 -11.5
 S; -16.1	 -7.5	 -10.0	 -9.75	 -10.4	 -10.8

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

rate of attachment of oligomeric species to the nucleus, and  $\eta_p$  is a simple estimate of the nucleation barrier  $\Delta G_p/k_B T_c$  at the temperature  $T_c$  of crystallization. In the following, we describe the functional forms of  $\nu_p$  and  $\eta_p$ .

We assume that the growth rate of 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  $O_p$ . As discussed below, we restrict our analysis to hexamers and smaller oligomers. The numerical procedure used to identify these oligomers is described in the Methods section. We ignore attachment of monomers (assuming they contribute equally to the growth of all polymorphs), attachment of oligomeric species that are structurally incompatible with  $p$ , and all

detachment events. 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  $n_i C_i s_i / Z'_p$ . The part of the nucleation score describing oligomer attachment,  $\nu_p$ , is thus obtained by summing over all oligomers of size 2–6,

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

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. (We return to this assumption in the 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

$$C_i \propto \frac{s_i^{-1} \sqrt{I_i} e^{-E_i/k_B T_c}}{\sum_{j \in O_{n_i}} s_j^{-1} \sqrt{I_j} e^{-E_j/k_B T_c}}. \quad (4)$$

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,  $k_B T_c \approx \alpha E_0$ , where  $\alpha = -0.055$ , as shown in SI Fig. S6. (This relation is reminiscent of the well-known relation between crystal melting points and sublimation enthalpies.<sup>55</sup>)

We now turn to estimating the polymer-specific nucleation barrier,  $\eta_p = \Delta G_p / k_B T_c$ . According to CNT,

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

Here,  $\gamma_p$  is the surface tension (*i.e.*, the line tension in 2D),  $\rho_p$  is the number of molecules per unit area in the crystal, and  $\Delta \mu_p = |\mu_p - \mu_{\text{sol}}|$  is the chemical potential difference of molecules in the crystal and solution phase, respectively.<sup>51</sup> Since  $\gamma_p$  cannot be straightforwardly determined from knowledge of the crystal structure alone, we use the same value of  $\bar{\gamma} = 1.78 \epsilon / \sigma$  for all polymorphs, which we determined as an average over several large crystalline clusters observed in our simulations (see SI). We similarly use a single value  $\bar{\rho} = 2\sqrt{3}/15 \sigma^{-2}$  (*i.e.*, the density of a close-packed crystal) for all polymorphs because polymorph packing fractions vary only little across our models.

In order to estimate  $\eta_p$ , we need an expression for the polymorph supersaturation  $\Delta \mu_p$ . To make progress, 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  $\Delta G_0 / k_B T_c \equiv g$ , where  $\Delta G_0$  is the nucleation barrier of the polymorph with the lowest energy and  $g$  is a constant that we treat as a fitting parameter. As shown below, we find that our model is most predictive for  $g = 7.62$ . (This value implies  $\Delta G_0 = 7.62 k_B T_c$ , consistent with the assumption of spontaneous barrier crossing on the microsecond timescale.)

With  $\Delta G_0 / k_B T_c$  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,  $\Delta E = E_p - E_0 > 0$  is the difference in lattice energy between polymorph  $p$  and the lowest energy polymorph. We therefore have

$$\begin{aligned} \frac{\Delta G_p}{k_B T_c} &= \frac{\pi \bar{\gamma}^2}{k_B T_c \bar{\rho} \Delta \mu_p} = \frac{\pi \bar{\gamma}^2}{k_B T_c \bar{\rho} (\Delta \mu_0 - \Delta E_p)} \\ &= \left( g^{-1} - \frac{k_B T_c \bar{\rho} \Delta E_p}{\pi \bar{\gamma}^2} \right)^{-1}. \end{aligned}$$

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

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



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

### E. Nucleation score predicts outcomes of crystallization simulations

To assess the predictive power of the nucleation score introduced in the previous section, we computed  $\chi_p$  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  $\chi_p$  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.<sup>56</sup>)

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 several polymorphs are 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 unique ground state.) 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).

#### 1. Dimers are not enough

Many studies have demonstrated correlations between polymorphs observed in experiments and a specific prenucleation cluster with low energy. These investigations usually concentrate on dimers.<sup>37–42</sup> 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  $\eta_p$ .) 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.

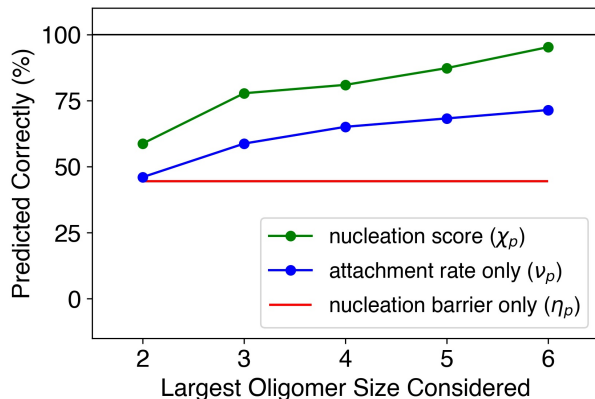


FIG. 8. Fraction of correctly predicted polymorphs as a function of the largest oligomer size included in  $\chi_p$  (green). The red line represents the success rate (44%) based on lattice energies alone (assuming that the correct polymorph is selected in half of all cases with multiple polymorphs at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

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  $\nu_p$ , 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.

#### 2. 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  $\nu_p$  and the nucleation barrier  $\eta_p$ , respectively.

To evaluate the relative importance of kinetic and thermodynamic factors in the crystallization of a given model molecule, we define the normalized quantities:  $X_p = \chi_p/\chi_{\text{max}}$ ,  $K_p = \nu_p/\nu_{\text{max}}$ , and  $B_p = \exp(-\eta_p)/\exp(-\eta_{\text{min}})$ . Here,  $\chi_{\text{max}}$  and  $\nu_{\text{max}}$  are the

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

TABLE I. Number of molecules of type I-IV (as defined in the text) found in sets A and B.

largest nucleation score and attachment rate, respectively, found for any polymorph, and  $\eta_{\min}$  is the nucleation barrier of the polymorph with the minimum lattice energy.

We identify 4 different types of molecules based on the values of  $K_{\text{form}}$  and  $B_{\text{form}}$  of the polymorph observed to form in simulations. (While  $X_{\text{form}} = 1$  for the vast majority of molecules,  $K_{\text{form}}$  and  $B_{\text{form}}$  may have values smaller than one.) Type I molecules are those that produce polymorphs with  $K_{\text{form}} = B_{\text{form}} = 1$ , *i.e.*, polymorphs that have both a larger attachment rate and a smaller lattice energy than any other polymorph of the same molecule. Molecules of type II are defined by  $K_{\text{form}} < 1$ ,  $B_{\text{form}} = 1$ ; these molecules form the polymorph with the smallest lattice energy even though oligomer concentrations in solution favor one or more other polymorphs. Type III molecules ( $K_{\text{form}} = 1$ ,  $B_{\text{form}} < 1$ ) form the polymorph most favored by oligomer species in solution, even though that polymorph does not have the lowest energy. Finally, molecules of type IV ( $K_{\text{form}} < 1$ ,  $B_{\text{form}} < 1$ ) produce polymorphs that neither have the lowest lattice energy nor the largest kinetic attachment rate, but still receive the largest nucleation score. Molecules of type III and IV form "kinetic" polymorphs that cannot be easily predicted with most current CSP methods. Table I shows that all four types are represented in molecule sets A and B, with a majority of clear-cut cases of type I in set A, and a majority of type III in set B.

To illustrate the competition between thermodynamic and kinetic factors in the nucleation score and in our MD simulations, we revisit the crystallization of molecules s9-A1 and s7-A1. Table II lists the values  $X_p$ ,  $K_p$ , and  $B_p$  for the competing polymorphs discussed earlier (see Figs. 3-7). Molecule s7-A1 is of type II: As evident from the simulation snapshot in Figure 6, few of the oligomers present in solution can directly contribute to the growth of Q ( $K_Q \ll K_S < K_R$ ) polymorphs. However, Q has a large energetic advantage over polymorphs R and S ( $B_Q \gg B_R > B_S$ ). As a result, Q receives the largest nucleation score and indeed forms in simulations; nuclei of R or S are not observed.

By contrast, molecule s9-A1 is of type III. The polymorph that forms in MD simulations, X, has a clear energetic disadvantage ( $B_X < B_Y, B_V, B_W, B_Z$ ) over the four other polymorphs (Y, V, W, Z) considered here. However, V, W, and Z have structural motifs that are not reflected in the prevalent oligomer species in solution, resulting in small values of  $K_p$  and thus  $X_p$  for these polymorphs. Polymorph Y shares some of the same dimer motifs with X, but larger oligomers of Y have lower con-

centrations than those contributing to the growth of X, leaving X with a modestly larger nucleation score than Y despite its higher lattice energy. These examples are characteristic for all molecules considered in this work and illustrate that the proposed nucleation score is able to successfully balance lattice energies and kinetic factors of oligomer attachment to produce reliable polymorph predictions.

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  $\eta_p$ ) are included; predictions based on oligomer attachment rates alone ( $\nu_p$ ) 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  $\nu_p$  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, larger enantiopure oligomers in solution clearly favor formation of enantiopure crystals.

## F. Discussion

The nucleation score presented in this paper is designed to systematically capture contributions of oligomeric species (or prenucleation 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. 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

molecule	polymorph	$X_p$	$B_p$	$K_p$	$E_p(\epsilon)$	$Z'_p$	Rac/Pure
s7-A1	Q	1.0	1.0	6.8E-3	-19.0	1	Rac
	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
s9-A1	X	1.0	0.46	1.0	-18.5	1	Pure
	Y	0.50	1.0	0.23	-19.0	1	Rac
	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

TABLE II. Kinetic and thermodynamic factors in the nucleation of molecules s9-A1 and s7-A1. Normalized nucleation scores  $X_p$ ,  $B_p$ , and  $K_p$ , 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. Polymorphs X and Q form spontaneously in MD simulations, as illustrated in Figs. 3 and 6, respectively.

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  $\nu_p$ . 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 ( $K_p$ , see Table II). We

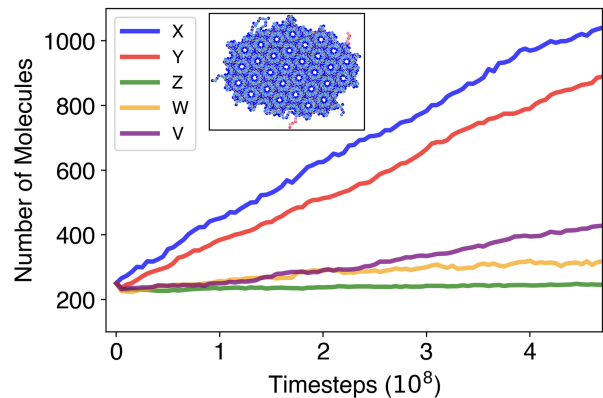


FIG. 9. Average number of molecules in growing nuclei of polymorphs X, Y, Z, W, and V of molecule s9-A1 at a temperature of  $T = 1.125 \epsilon/k_B$ , slightly above the temperature at which spontaneous nucleation is observed. The inset shows a snapshot of a crystallite of polymorph X containing 254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

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, the importance of oligomer attachment in nucleation and growth is diminished and differences in nucleation barrier heights, which our model only crudely estimates, will become decisive. We cannot straightforwardly simulate crystal nucleation in this regime due to excessively long time scales required to cross nucleation barriers that exceed a few  $k_B T$ . 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 II). This result is consistent with the di-



minished 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.

How applicable is our model to the crystallization of real molecules from solution? In its current form, our model neglects all 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.<sup>53,57</sup> 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 model 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.

### III. METHODS

#### A. Molecular dynamics simulations.

All molecular dynamics simulations were performed with HOOMD<sup>58,59</sup>. All functional groups of molecules have the same mass  $m$  and diameter  $\sigma$ , which we use as our units of mass and length; the unit of energy is  $\epsilon$ . 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.<sup>60</sup>

#### B. Molecular Interactions

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

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

The repulsive part of the potential is of the WCA form,<sup>61</sup>

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

The attractive part is given by

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

We set  $\epsilon_{\text{rep}} = 5.0\epsilon$  and  $\omega = 0.2\sigma$ . For molecules in set A, we use  $\epsilon_{\text{att}} = \epsilon$  for all weakly interacting functional groups. For strongly interacting functional groups of molecules in set A, we use  $\epsilon_{\text{att}} = 5\epsilon$ . 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.

#### C. Molecular interactions in set B

Set B was created with a bias towards 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.<sup>49</sup> We then generated a set of random interaction vectors using the Bayesian Bootstrap method.<sup>62</sup> 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,

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

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

$$\vec{\epsilon}_0 = (\epsilon, \epsilon, \dots, \epsilon).$$

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.<sup>49</sup> For interaction vectors with  $\varphi_{\vec{\epsilon}} \geq$

38°, we determined the lowest-energy racemic ( $E_0^{(R)}$ ) and enantiopure ( $E_0^{(P)}$ ) polymorphs using the molecular shapes mentioned above. Approximately 1000 molecules with  $0 \leq E_0^{(P)} - E_0^{(R)} \leq 0.4 \epsilon/\text{molecule}$  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.<sup>49</sup> 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.

#### D. 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  $\approx 15$  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. 49.

#### E. 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.<sup>63</sup> 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.

#### F. 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  $1.5 \epsilon/k_B$ , 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).

#### ACKNOWLEDGMENTS

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.

- 1 J. Maddox, "Crystals from first principles," *Nature* **335**, 201–201 (1988).
- 2 M. Hilfiker, Rolf; Raumer von, *Polymorphism in the Pharmaceutical Industry: Solid Form and Drug Development*, edited by R. Hilfiker and M. von Raumer (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2018).
- 3 A. Y. Lee, D. Erdemir, and A. S. Myerson, "Crystal Polymorphism in Chemical Process Development," *Annual Review of Chemical and Biomolecular Engineering* **2**, 259–280 (2011).
- 4 H. G. Brittain, *Polymorphism in Pharmaceutical Solids: Second Edition*, edited by H. G. Brittain (CRC Press, 2018).
- 5 J. Yang, X. Zhu, C. T. Hu, M. Qiu, Q. Zhu, M. D. Ward, and B. Kahr, "Inverse Correlation between Lethality and Thermodynamic Stability of Contact Insecticide Polymorphs," *Crystal Growth & Design* **19**, 1839–1844 (2019).
- 6 X. Zhu, C. T. Hu, J. Yang, L. A. Joyce, M. Qiu, M. D. Ward, and B. Kahr, "Manipulating Solid Forms of Contact Insecticides for Infectious Disease Prevention," *Journal of the American Chemical Society* **141**, 16858–16864 (2019).
- 7 J. Yang, B. Erriah, C. T. Hu, E. Reiter, X. Zhu, V. López-Mejías, I. P. Carmona-Sepúlveda, M. D. Ward, and B. Kahr, "A deltamethrin crystal polymorph for more effective malaria control," *Proceedings of the National Academy of Sciences* **117**, 26633–26638 (2020).
- 8 B. Olenik and G. Thielking, "Polymorphism and the Organic Solid State: Influence on the Optimization of Agrochemicals," in *Modern Methods in Crop Protection Research*, 2012 (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2013) pp. 249–272.
- 9 J. Tang, G. Cheng, Y. Zhao, P. Yang, X. Ju, and H. Yang, "Optimizing the molecular structure and packing style of a crystal by intramolecular cyclization from picrylhydrazone to indazole," *CrystEngComm* **21**, 4701–4706 (2019).
- 10 N. S. Fondren, Z. T. Fondren, D. K. Unruh, and B. L. Weeks, "Effects of Solution Conditions on Polymorph Development in 2,4,6-Trinitrotoluene," *Crystal Growth & Design* **20**, 568–579 (2020).
- 11 Q. Ma, Z. Lu, L. Liao, J. Huang, D. Liu, J. Li, and G. Fan, "5,6-Di(2-fluoro-2,2-dinitroethoxy)furanazo[3,4-b]pyrazine: a high

- performance melt-cast energetic material and its polycrystalline properties,” *RSC Advances* **7**, 38844–38852 (2017).
- <sup>12</sup>J. Tang, G. Cheng, S. Feng, X. Zhao, Z. Zhang, X. Ju, and H. Yang, “Boosting Performance and Safety of Energetic Materials by Polymorphic Transition,” *Crystal Growth & Design* **19**, 4822–4828 (2019).
  - <sup>13</sup>T. Yan, G. Cheng, and H. Yang, “1,2,4-Oxadiazole-Bridged Polynitropyrazole Energetic Materials with Enhanced Thermal Stability and Low Sensitivity,” *ChemPlusChem* **84**, 1567–1577 (2019).
  - <sup>14</sup>J. Hoja, H.-Y. Ko, M. A. Neumann, R. Car, R. A. DiStasio, and A. Tkatchenko, “Reliable and practical computational description of molecular crystal polymorphs,” *Science Advances* **5** (2019), 10.1126/sciadv.aau3338.
  - <sup>15</sup>A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiStasio, A. Dzyabchenko, B. P. Van Eijck, D. M. Elking, J. A. Van Den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C. A. Gatsiou, T. S. Gee, R. De Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. De Jong, J. Kendrick, N. J. De Klerk, H. Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. Leusen, A. M. Lund, J. Lv, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meeke, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monserrat, R. J. Needs, M. A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E. Schneider, C. Schober, G. P. Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. Wang, R. E. Watson, G. A. De Wijs, J. Yang, Q. Zhu, and C. R. Groom, “Report on the sixth blind test of organic crystal structure prediction methods,” *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **72**, 439–459 (2016).
  - <sup>16</sup>O. Egorova, R. Hafizi, D. C. Woods, and G. M. Day, “Multifidelity Statistical Machine Learning for Molecular Crystal Structure Prediction,” *The Journal of Physical Chemistry A* **124**, 8065–8078 (2020).
  - <sup>17</sup>D. McDonagh, C.-K. Skylaris, and G. M. Day, “Machine-Learned Fragment-Based Energies for Crystal Structure Prediction,” *Journal of Chemical Theory and Computation* **15**, 2743–2758 (2019).
  - <sup>18</sup>M. Yang, E. Dybeck, G. Sun, C. Peng, B. Samas, V. M. Burger, Q. Zeng, Y. Jin, M. A. Bellucci, Y. Liu, P. Zhang, J. Ma, Y. A. Jiang, B. C. Hancock, S. Wen, and G. P. Wood, “Prediction of the Relative Free Energies of Drug Polymorphs above Zero Kelvin,” *Crystal Growth and Design* **20**, 5211–5224 (2020).
  - <sup>19</sup>N. S. Abraham and M. R. Shirts, “Statistical Mechanical Approximations to More Efficiently Determine Polymorph Free Energy Differences for Small Organic Molecules,” *Journal of Chemical Theory and Computation* **16**, 6503–6512 (2020), arXiv:2006.03101.
  - <sup>20</sup>G. P. Stahly, “Diversity in single- and multiple-component crystals: the search for and prevalence of polymorphs and cocrystals,” *Crystal Growth and Design* **7**, 1007–1026 (2007).
  - <sup>21</sup>R. J. Davey, S. L. Schroeder, and J. H. Ter Horst, “Nucleation of organic crystals - A molecular perspective,” *Angewandte Chemie - International Edition* **52**, 2167–2179 (2013).
  - <sup>22</sup>J. Hoja, H.-Y. Ko, M. A. Neumann, R. Car, R. A. DiStasio, and A. Tkatchenko, “Reliable and practical computational description of molecular crystal polymorphs,” *Science Advances* **5** (2019), 10.1126/sciadv.aau3338, arXiv:1803.07503.
  - <sup>23</sup>C. Greenwell, J. L. McKinley, P. Zhang, Q. Zeng, G. Sun, B. Li, S. Wen, and G. J. Beran, “Overcoming the difficulties of predicting conformational polymorph energetics in molecular crystals: Via correlated wavefunction methods,” *Chemical Science* **11**, 2200–2214 (2020).
  - <sup>24</sup>S. L. Price, “Why don’t we find more polymorphs?” *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **69**, 313–328 (2013).
  - <sup>25</sup>G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen, and A. Michaelides, “Crystal Nucleation in Liquids: Open Questions and Future Challenges in Molecular Dynamics Simulations,” *Chemical Reviews* **116**, 7078–7116 (2016), arXiv:1611.06456.
  - <sup>26</sup>Y. C. Chang and A. S. Myerson, “Diffusivity of glycine in concentrated saturated and supersaturated aqueous solutions,” *AIChE Journal* **32**, 1567–1569 (1986).
  - <sup>27</sup>M. Kellermeier, R. Rosenberg, A. Moise, U. Anders, M. Przybylski, and H. Cölfen, “Amino acids form prenucleation clusters: ESI-MS as a fast detection method in comparison to analytical ultracentrifugation,” *Faraday Discussions* **159**, 23–45 (2012).
  - <sup>28</sup>K. J. Koch, F. C. Gozzo, S. C. Nanita, Z. Takats, M. N. Eberlin, and R. G. Cooks, “Chiral transmission between amino acids: Chirally selective amino acid substitution in the serine octamer as a possible step in homochirogenesis,” *Angewandte Chemie - International Edition* **41**, 1721–1724 (2002).
  - <sup>29</sup>Z. Takats, S. C. Nanita, R. G. Cooks, G. Schlosser, and K. Vekey, “Amino acid clusters formed by sonic spray ionization,” *Analytical Chemistry* **75**, 1514–1523 (2003).
  - <sup>30</sup>N. Toyama, J. Y. Kohno, F. Mafuné, and T. Kondow, “Solvation structure of arginine in aqueous solution studied by liquid beam technique,” *Chemical Physics Letters* **419**, 369–373 (2006).
  - <sup>31</sup>P. Yang, R. Xu, S. C. Nanita, and R. G. Cooks, “Thermal formation of homochiral serine clusters and implications for the origin of homochirality,” *Journal of the American Chemical Society* **128**, 17074–17086 (2006).
  - <sup>32</sup>C. E. Hughes, S. Hamad, K. D. M. Harris, C. R. A. Catlow, and P. C. Griffiths, “A multi-technique approach for probing the evolution of structural properties during crystallization of organic materials from solution,” *Faraday Discussions* **136**, 71 (2007).
  - <sup>33</sup>S. Hamad, C. E. Hughes, C. R. A. Catlow, and K. D. Harris, “Clustering of glycine molecules in aqueous solution studied by molecular dynamics simulation,” *Journal of Physical Chemistry B* **112**, 7280–7288 (2008).
  - <sup>34</sup>P. Nemes, G. Schlosser, and K. Vékey, “Amino acid cluster formation studied by electrospray ionization mass spectrometry,” *Journal of Mass Spectrometry* **40**, 43–49 (2005).
  - <sup>35</sup>D. Zhang, L. Wu, K. J. Koch, and R. G. Cooks, “Arginine clusters generated by electrospray ionization and identified by tandem mass spectrometry,” *European Journal of Mass Spectrometry* **5**, 353–361 (1999).
  - <sup>36</sup>C. A. Hunter, J. F. McCabe, and A. Spitaleri, “Solvent effects of the structures of prenucleation aggregates of carbamazepine,” *CrystEngComm* **14**, 7115–7117 (2012).
  - <sup>37</sup>A. Spitaleri, C. A. Hunter, J. F. McCabe, M. J. Packer, and S. L. Cockroft, “A 1H NMR study of crystal nucleation in solution,” *CrystEngComm* **6**, 489 (2004).
  - <sup>38</sup>R. A. Chiarella, A. L. Gillon, R. C. Burton, R. J. Davey, G. Sadiq, A. Auffret, M. Cioffi, and C. A. Hunter, “The nucleation of inosine: The impact of solution chemistry on the appearance of polymorphic and hydrated crystal forms,” *Faraday Discussions* **136**, 179–193 (2007).
  - <sup>39</sup>S. Parveen, R. J. Davey, G. Dent, and R. G. Pritchard, “Linking solution chemistry to crystal nucleation: The case of tetrolic acid,” *Chemical Communications* , 1531–1533 (2005).
  - <sup>40</sup>C. Jie and B. L. Trout, “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* **112**, 7794–7802 (2008).
  - <sup>41</sup>S. A. Kulkarni, E. S. McGarrity, H. Meeke, and J. H. Ter Horst, “Isonicotinamide self-association: The link between solvent and polymorph nucleation,” *Chemical Communications* **48**, 4983–4985 (2012).
  - <sup>42</sup>K. Chadwick, R. J. Davey, G. Dent, R. G. Pritchard, C. A. Hunter, and D. Musumeci, “Cocrystallization: A solution chem-



- istry perspective and the case of benzophenone and diphenylamine,” *Crystal Growth and Design* **9**, 1990–1999 (2009).
- <sup>43</sup>D. Gebauer, M. Kellermeier, J. D. Gale, L. Bergström, and H. Cölfen, “Pre-nucleation clusters as solute precursors in crystallisation,” *Chemical Society Reviews* **43**, 2348–2371 (2014).
- <sup>44</sup>S. Hamad, C. Moon, C. Richard, A. Catlow, A. T. Hulme, and S. L. Price, “Kinetic insights into the role of the solvent in the polymorphism of 5-fluorouracil from molecular dynamics simulations,” *Journal of Physical Chemistry B* **110**, 3323–3329 (2006).
- <sup>45</sup>D. Gebauer and H. Cölfen, “Prenucleation clusters and non-classical nucleation,” *Nano Today* **6**, 564–584 (2011).
- <sup>46</sup>F. Peraland E. Gallego, “Self-association of imidazole and its methyl derivatives in aqueous solution. A study by ultraviolet spectroscopy,” *Journal of Molecular Structure* **415**, 187–196 (1997).
- <sup>47</sup>R. C. Burton, E. S. Ferrari, R. J. Davey, J. L. Finney, and D. T. Bowron, “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* **114**, 8807–8816 (2010).
- <sup>48</sup>G. R. Desiraju, “Crystal engineering: A holistic view,” *Angewandte Chemie - International Edition* **46**, 8342–8356 (2007).
- <sup>49</sup>J. E. Carpenter and M. Grünwald, “Heterogeneous Interactions Promote Crystallization and Spontaneous Resolution of Chiral Molecules,” *Journal of the American Chemical Society* **142**, 10755–10768 (2020).
- <sup>50</sup>M. A. Neumann and J. van de Streek, “How many ritonavir cases are there still out there?” *Faraday Discussions* **211**, 441–458 (2018).
- <sup>51</sup>S. Jungblut and C. Dellago, “Pathways to self-organization: Crystallization via nucleation and growth,” *The European Physical Journal E* **39**, 77 (2016).
- <sup>52</sup>M. A. Deij, J. H. Ter Horst, H. Meekes, P. Jansens, and E. Vlieg, “Polymorph formation studied by 3D nucleation simulations. Application to a yellow isoxazolone dye, paracetamol, and L-glutamic acid,” *Journal of Physical Chemistry B* **111**, 1523–1530 (2007).
- <sup>53</sup>R. J. Davey, K. R. Back, and R. A. Sullivan, “Crystal nucleation from solutions - Transition states, rate determining steps and complexity,” *Faraday Discussions* **179**, 9–26 (2015).
- <sup>54</sup>R. A. Sullivan, R. J. Davey, G. Sadiq, G. Dent, K. R. Back, J. H. Ter Horst, D. Toroz, and R. B. Hammond, “Revealing the roles of desolvation and molecular self-assembly in crystal nucleation from solution: Benzoic and p-aminobenzoic acids,” *Crystal Growth and Design* **14**, 2689–2696 (2014).
- <sup>55</sup>M. Salahinejad, T. C. Le, and D. A. Winkler, “Capturing the Crystal: Prediction of Enthalpy of Sublimation, Crystal Lattice Energy, and Melting Points of Organic Compounds,” *Journal of Chemical Information and Modeling* **53**, 223–229 (2013).
- <sup>56</sup>J. Bernstein, J. D. Dunitz, and A. Gavezzotti, “Polymorphic perversity: Crystal structures with many symmetry-independent molecules in the unit cell,” *Crystal Growth and Design* **8**, 2011–2018 (2008).
- <sup>57</sup>W. J. Dunning and A. J. Shipman, “Nucleation in Sucrose Solutions,” *Proc. Agric. Industries 10th International Conference*, 1448–1456 (1954).
- <sup>58</sup>J. Glaser, T. D. Nguyen, J. A. Anderson, P. Lui, F. Spiga, J. A. Millan, D. C. Morse, and S. C. Glotzer, “Strong scaling of general-purpose molecular dynamics simulations on GPUs,” *Computer Physics Communications* **192**, 97–107 (2015), arXiv:1412.3387.
- <sup>59</sup>J. A. Anderson, C. D. Lorenz, and A. Travestet, “General purpose molecular dynamics simulations fully implemented on graphics processing units,” *Journal of Computational Physics* **227**, 5342–5359 (2008).
- <sup>60</sup>A. Stukowski, “Visualization and analysis of atomistic simulation data with OVITO—the Open Visualization Tool,” *Modelling and Simulation in Materials Science and Engineering* **18** (2010), 10.1088/0965-0393/18/1/015012.
- <sup>61</sup>J. D. Weeks, D. Chandler, and H. C. Andersen, “Role of repulsive forces in determining the,” *The Journal of Chemical Physics* **54**, 5237–5247 (1971).
- <sup>62</sup>D. B. Rubin, “The Bayesian Bootstrap,” *The Annals of Statistics* **9**, 130–134 (1981).
- <sup>63</sup>C. Komusiewicz and F. Sommer, “Enumerating connected induced subgraphs: Improved delay and experimental comparison,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 11376 LNCS (2019) pp. 272–284.