# Methods for "Design rules for the self-assembly of a protein crystal"

Thomas K. Haxton and Stephen Whitelam
*The Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720*

## DYNAMIC SIMULATIONS

We conducted virtual-move Monte Carlo (VMMC) simulations [1] using the 'symmetrized' version of the algorithm described in Refs. [2, 3]. This algorithm approximates overdamped dynamics for short-range interacting particles in solution by self-consistently attempting and accepting cluster moves according to gradients of potential energy. The following description assumes familiarity with this algorithm.

At each Monte Carlo (MC) step, we attempt a rotation move with probability $p_{\rm r}$ and a translation move with probability $p_{\rm t} = 1 - p_{\rm r}$. A translation move shifts a monomer's $x$ and $y$ coordinates by random displacements in the interval $(-\Delta_{\rm t}/2, \Delta_{\rm t}/2)$. A rotation move changes the seed monomer's orientation vector from $\hat{\boldsymbol{u}}$ to $\hat{\boldsymbol{u}}' = (\hat{\boldsymbol{u}} + r\hat{\boldsymbol{u}}_\perp)/|\hat{\boldsymbol{u}} + r\hat{\boldsymbol{u}}_\perp|$, where $r$ is a random number in the interval $(-\Delta_{\rm r}/2, \Delta_{\rm r}/2)$ and $\hat{\boldsymbol{u}}_\perp$ is a unit vector perpendicular to $\hat{\boldsymbol{u}}$ and in the plane of the substrate. For an isolated monomer, these moves result in translational and rotational diffusion coefficients

$$\begin{aligned} D_{\rm t} &= \frac{p_{\rm t}}{24 t_{\rm cycle}} \left(\Delta_{\rm t}\right)^2, \\ D_{\rm r} &= \frac{p_{\rm r}}{24 t_{\rm cycle}} \left(\Delta_{\rm r}\right)^2, \end{aligned} \tag{1}$$

where $t_{\rm cycle}$ is the time interval assigned to each MC cycle.

We assume that the dominant source of drag is from the three-dimensional fluid surrounding the protein, rather than from the interaction with the two-dimensional substrate. We take the kinematic and dynamic viscosities of the aqueous solvent to be $\nu = 1.00 \times 10^{-6} \mathrm{m}^2/\mathrm{s}$ and $\eta = 1.00$ Pa s, respectively. Inertial effects are controlled by the Reynolds number $\mathrm{Re} = av/\nu$, where $a = 3.9$ nm is the characteristic length scale of the protein monomers and $v$ is a characteristic velocity. Using the thermal velocity $v = \sqrt{k_{\rm B}T/m}$, where $m = 132$ kDa is the protein mass, results in $\mathrm{Re} = 0.017$. Alternately, balancing a characteristic drag force $6\pi\eta a v$ with a characteristic inter-protein force $F = 100 k_{\rm B}T/a$, taking characteristic interaction strengths and separations to be on the order of $10 k_{\rm B}T$ and $0.1a$, respectively, results in a characteristic velocity $v = 100 k_{\rm B}T/6\pi\eta a^2$ and Re =. In either case, Re is small, so we neglect inertia.

In order to have a reasonably efficient simulation, we do not calculate the fluid flow; instead, we let the drag acting on a cluster be equivalent to the drag acting on an isolated sphere with the same hydrodynamic radius. We define the hydrodynamic radius of a cluster $\mathcal{C}$ as a generalization of the radius of gyration [1]. For translations,

$${R_{\rm t}}^2 \equiv \langle |(\boldsymbol{r} - \boldsymbol{r}_{\rm com}) \times \hat{\boldsymbol{n}}|^2 \rangle_{\boldsymbol{r} \in \mathcal{C}}, \tag{2}$$

where $\boldsymbol{r}_{\rm com}$ is the center of mass and $\hat{\boldsymbol{n}}$ is the direction of the translation. For rotations,

$${R_{\rm r}}^2 \equiv \langle |(\boldsymbol{r} - \boldsymbol{r}_{\rm axis}) \times \hat{\boldsymbol{z}}|^2 \rangle_{\boldsymbol{r} \in \mathcal{C}}, \tag{3}$$

where $\boldsymbol{r}_{\rm axis}$ is the center of rotation and $\hat{\boldsymbol{z}}$ is the axis of rotation, perpendicular to the substrate. We take $\boldsymbol{r} \in \mathcal{C}$ to include all points within the hard cores of the monomers. The radius of gyration of a (real) monomer depends not only on its two-dimensional footprint on the substrate, but also on its height. We take the height of a monomer to be equal to its width. Since the hydrodynamic radii of a sphere are ${R_{\rm t}}^2 = {R_{\rm r}}^2 = 2R^2/5$, the Stokes solutions for the drag on a sphere are

$$\begin{aligned} D_{\rm t}^\star(R_{\rm t}) &= \frac{k_{\rm B}T}{6\pi\eta\sqrt{5/2}R_{\rm t}}, \\ D_{\rm r}^\star(R_{\rm r}) &= \frac{k_{\rm B}T}{8\pi\eta(5/2)^{3/2}{R_{\rm r}}^3}. \end{aligned} \tag{4}$$

We parameterize the algorithm to yield Eq. (4) for isolated, strongly-bound clusters with hydrodynamic radii $R_{\rm t}$ and $R_{\rm r}$. We can enforce Eq. (4) *a priori* if we assume that the trial step sizes $\Delta_{\rm t}$ and $\Delta_{\rm r}$ are large compared to the size of the bonds, so that individual moves are always suppressed and only whole-cluster moves are accepted. We will consider realistic cases where single-monomer relaxations are allowed shortly. First, we ensure that clusters with the smallest possible radii of gyration satisfy Eq. (4). For rotations, the smallest possible radius of gyration is that of a monomer, $R_0^{\rm r} = 0.698a$. For translations, the radius of gyration of a monomer depends on the direction of translation. Drag is least for translation along the long axis, for which $R_0^{\rm t} = 0.408a$. Plugging these minimal radii of gyration into Eqs. 1 and 4 constrains the relative frequency of translation and rotation moves according to

$$\frac{p_{\rm r}}{p_{\rm t}} = 0.361 \left(\frac{\Delta_{\rm t}}{a\Delta_{\rm r}}\right)^2. \tag{5}$$

We will choose appropriate trial step sizes shortly.

Next, we ensure that larger clusters also obey Eq. (4). Since we have already constrained the diffusion coefficients of the smallest possible clusters, it is sufficient to require

$$\begin{aligned} \frac{D_{\rm t}(R_{\rm t})}{D_{\rm t}(R_0^{\rm t})} &= \frac{R_0^{\rm t}}{R_{\rm t}}, \\ \frac{D_{\rm r}(R_{\rm r})}{D_{\rm r}(R_0^{\rm r})} &= \left(\frac{R_0^{\rm r}}{R_{\rm r}}\right)^3. \end{aligned} \tag{6}$$

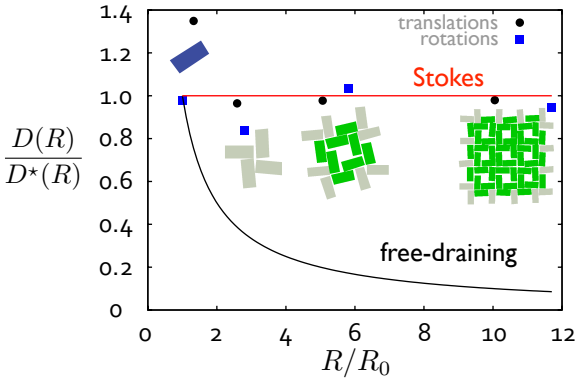


FIG. 1: Ratio of diffusion coefficients $D(R)$ to $D^\star(R)$, the solution of Eq. (4), for rotations and translations of clusters of 1, 4, 16, and 64 monomers, where $R$ is the average measured radius of gyration.

We enforce Eq. (6) by employing cutoffs that depend on the hydrodynamic radii, in addition to the usual particle number-dependent cutoff of the VMMC algorithm [1]. Before each attempted translation or rotation, we draw two random numbers, $q$ and $x$, from the interval $(0, 1)$. As we build a cluster according the the VMMC algorithm, we reject the move *in situ* if the hydrodynamic radius $R$ exceeds $R_0 q^{-\nu}$ or the number of monomers in the cluster $N_c$ exceeds $1/x$. According to this scheme, the diffusion coefficient per MC cycle is

$$D(N_c, R) = N_c D(1, R_0) \left(\frac{R}{R_0}\right)^{-1/\nu} \frac{1}{N_c}, \qquad (7)$$

where $D(1, R_0)$ is the diffusion coefficient of a single monomer with radius of gyration $R_0$. The prefactor $N_c$ is the average number of times the cluster is selected for a trial move per MC cycle, once per monomer in the cluster. Choosing $\nu = 1$ for translations and $\nu = 1/3$ for rotations reduces Eq. (7) to the desired form of Eq. (6).

In practice, we aim both to produce the correct diffusion coefficients of Eq. (4) and to allow local relaxation within clusters. This requires optimizing the trial step sizes $\Delta_\mathrm{t}$ and $\Delta_\mathrm{r}$. To do so, we performed VMMC simulations to measure the diffusion coefficients of isolated, compact clusters of size 1 to 256 bound by permanent specific bonds. For $\Delta_\mathrm{t}$ and $\Delta_\mathrm{r}$ much smaller than the specific bond range $s = 0.1a$, single-monomer trial moves rarely change the potential energy and are almost always accepted, leading to the incorrect free-draining diffusion scalings of single-particle MC or Brownian dynamics, $D_\mathrm{t} \propto {R_\mathrm{t}}^{-2}$ and $D_\mathrm{r} \propto {R_\mathrm{t}}^{-4}$. For $\Delta_\mathrm{t}$ and $\Delta_\mathrm{r}$ much larger than $s$, the algorithm efficiently rejects trial moves of incomplete clusters and proposes cluster moves of the entire cluster with a probability determined by the cutoffs for the radii of gyration. While this yields the correct diffusion scaling, it suppresses internal relaxation. In order to produce the correct diffusion scaling while allowing local rearrangements, we choose the smallest step sizes that produce the correct diffusion coefficients up to an accuracy of about 10%. We find these to be $\Delta_\mathrm{t} = 0.8a$ and $\Delta_\mathrm{r} = 0.5$. With these choices, single-monomer moves account for the majority of accepted moves, but whole-cluster moves dominate the long-time diffusive motion.

Fig. 1 shows that the choices $\Delta_\mathrm{t} = 0.8a$ and $\Delta_\mathrm{r} = 0.5$ yield the correct diffusion coefficients of Eq. (4). Notice in Fig. 1 that single monomers have a translational diffusion coefficient greater than predicted by Eq. (4) because of their anisotropic shape: diffusion occurs preferentially along the long axis, for which the hydrodynamic radius is smaller, but the average hydrodynamic radius appearing in Eq. (4) samples all directions equally. The translational diffusion coefficients of the larger, more compact clusters, as well as all of the rotational diffusion coefficients, lie near the Stokes solution, Eq. 4. The free-draining solutions for both translational and rotational diffusion fall off as $1/R$ relative to the Stokes solution. Fig. 1 demonstrates that this parameterization of the virtual-move algorithm can much more closely approximate Stokes flow than can free-draining motion. The latter is generated by simple implementations of Brownian dynamics, and by single-particle MC algorithms in the limit of vanishing step size. This difference is potentially significant: Fig. 1 reveals that for clusters of even modest size (e.g. $\sim 60$ particles), the free-draining diffusion constant is an *order of magnitude* smaller than the Stokes one.

Applying Eq. (5), our choices of $\Delta_\mathrm{t}$ and $\Delta_\mathrm{r}$ yield attempt frequencies $p_\mathrm{t} = 0.520$ and $p_\mathrm{r} = 0.480$. Combining Eq. (1) and (4) for a monomer with $R_0 = 0.408a$ yields a time per MC cycle of

$$t_\mathrm{cycle} = \sqrt{\frac{5}{2}} \frac{\pi \eta p_\mathrm{t} {\Delta_\mathrm{t}}^2 R_0}{4 k_\mathrm{B} T} = 2.42 \text{ ns}. \qquad (8)$$

We used dynamic simulations to calculate the scaled yield and to generate pathway diagrams shown in Fig. 4 of the main text. Letting $f_2$ and $f_3$ be the fraction of monomers with two and three satisfied specific bonds, respectively, we define the scaled yield as in Ref. [4] by $\hat{f}_3 \equiv f_3(f_3/(f_3 + f_2))^2$, rewarding crystalline clusters with large bulk-to-surface ratios. In the color maps of yield in the main text, we report the scaled yield after an arbitrary choice of $10^7$ MC cycles, or 24.2 ms. We find that the qualitative behavior is insensitive to this choice. We constructed pathway diagrams by recording the maximum fraction of monomers in various configurations over the course of assembly. Fig. 2 shows time traces of these fractions for two examples corresponding to points in Figs. 3 and 4 of the main text. Panel (a) shows a pathway that proceeds via misbound configurations, while panel (b) shows a pathway that proceeds directly.

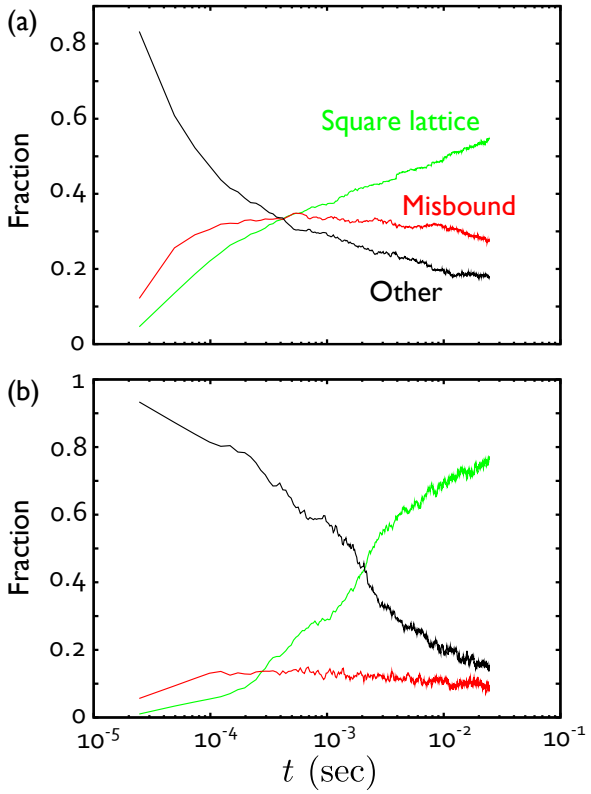


FIG. 2: Fraction of monomers in various configurations vs time for (a) $\phi = 0.1, \epsilon_{\rm n} = 1, \epsilon_{\rm s} = 10$ and (b) $\phi = 0.1, \epsilon_{\rm n} = 1, \epsilon_{\rm int} = 5$, demonstrating the difference between a 'misbound' pathway and a direct pathway. 'Square lattice' denotes monomers with all three specific bonds, and 'misbound' denotes monomers with their external specific bond satisfied but only one of their two internal specific bonds satisfied.

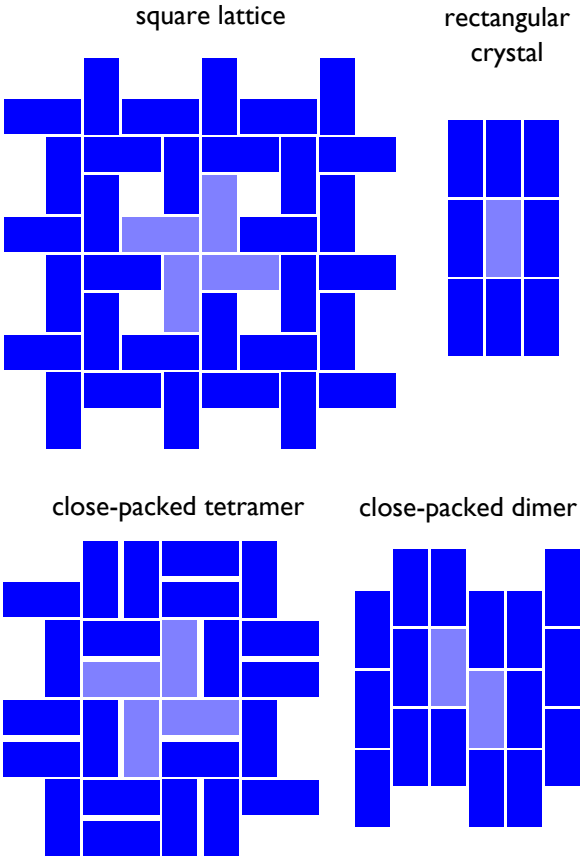


FIG. 3: Sketches of the four crystal phases.

## EQUILIBRIUM SIMULATIONS

We numerically calculated fluid-solid coexistence packing fractions $\phi_{\rm fluid}$ and $\phi_{\rm solid}$ by conducting direct coexistence Monte Carlo simulations at fixed $T$, $N$, and total area $A$. We initialized each periodic simulation box with an aspect ratio 4 : 1 and a crystal slab spanning the short axis of the box. We allowed the aspect ratio to fluctuate to let the system adopt its equilibrium lattice spacing. We used a mixture of moves, each obeying detailed balance, to facilitate efficient equilibration. We used non-local aggregation-volume bias [5] and teleportation [4] moves to facilitate exchange between solid and fluid phases. We used 180 degree rotations of single monomers to facilitate sampling of specific bonds and rigid 90 turn moves of two adjacent, parallel monomers to facilitate conversion among close-packed crystal phases. We also used rigid rotations and translations of bound dimers and tetramers. We chose appropriate probabilities for each move type depending on the interaction strengths (e.g. more nonlocal moves for systems with strong bonds), with the remaining probability split between single-particle translations and rotations.

We consider the four crystal phases sketched in Fig. 3. Depending on the values of $\epsilon_{\rm int}$, $\epsilon_{\rm ext}$, and $\epsilon_{\rm n}$, we performed one or more simulations starting with slabs of square lattice, rectangular, close-packed tetramer, or close-packed dimer crystals. While we did not attempt to accurately measure solid-solid equilibria with simulations, we only found narrow regions of interaction space over which two crystal phases are stable, and these regions are consistent with the solid-solid phase boundaries calculated analytically with mean-field theory (see next section).

We numerically calculated liquid-vapor coexistence packing fractions $\phi_{\rm liquid}$ and $\phi_{\rm vapor}$ by performing Gibbs ensemble simulations [6], using combinations of the same local and nonlocal moves. At intermediate values of $\epsilon_n$ and low values of $\epsilon_{\rm int}$ and $\epsilon_{\rm ext}$, we find coexisting liquid-vapor mixtures in which each phase consists mostly of unbound monomers. These packing fractions define a stable liquid-vapor binodal, as shown e.g. for protein 1 in Fig. 1(b) of the main text. We determine the liquid-vapor critical point $\{\phi_c, T_c\}$ by conducting least-squares fits to the functions

$$(\phi_{\rm liquid} - \phi_{\rm gas})^8 = c_1(T_c - T) \qquad (9)$$

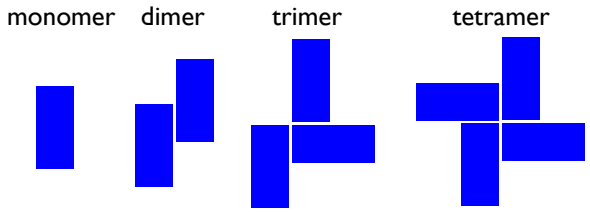


FIG. 4: Sketches of the four oligomers.

and

$$\frac{1}{2}\left(\phi_{\rm liquid}+\phi_{\rm gas}\right)=\phi_c+c_2(T_c-T), \tag{10}$$

where $c_1$ and $c_2$ are constants. Eq. 9 is the expected form for systems in the two-dimensional Ising universality class, and Eq. 10 is the empirical law of rectilinear diameter. As $\epsilon_{\rm int}$ and $\epsilon_{\rm ext}$ increase, the binodal is subsumed by the solubility curve and becomes metastable. While crystallization sets in too rapidly for us to measure the metastable binodal directly, we estimate the location of the metastable liquid-vapor critical point for proteins 2 ($\epsilon_{\rm int}/\epsilon_{\rm n}=1.5$) and 3 ($\epsilon_{\rm int}/\epsilon_{\rm n}=2$) in Fig. 1 of the main text by linear extrapolation along $\epsilon_{\rm ext}/\epsilon_{\rm int}=2$ from their locations at $\epsilon_{\rm int}/\epsilon_{\rm n}=0$ and 1. We find a weak dependence on the specific interaction strength,

$$\begin{aligned} T_c &= 0.431+0.013\epsilon_{\rm int}/\epsilon_{\rm n},\\ \phi_c &= 0.284+0.002\epsilon_{\rm int}/\epsilon_{\rm n}. \end{aligned} \tag{11}$$

In addition to the monomer liquid, at intermediate values of $\epsilon_{\rm n}$, low values of $\epsilon_{\rm int}$, and high values of $\epsilon_{\rm ext}$, we find coexisting liquid-vapor mixtures in which each phase consists mostly of bound dimers. For example, in Fig. 2 of the main text, the upper-left corner of the phase diagram corresponds to a dimer gas coexisting with a dimer liquid. We find no parameters for which a liquid of tetramers is stable, due to the smaller relative interaction range for tetramers.

## MEAN-FIELD THEORY

We used analytic mean-field theory to calculate stable and metastable solid-fluid and solid-solid phase boundaries, thermodynamic driving forces for assembly, and second virial coefficients. For the solid-vapor phase boundaries, we find excellent agreement between the mean field theory and the numerically calculated solubility curves, as shown in Fig. 1 of the main text. The solid-solid phase boundaries run through the narrow regions of parameter space in which both solids are stable on the timescale of the direct coexistence simulations. We do not attempt to account for liquid phases in the mean-field theory; we identify regions of liquid stability using Gibbs ensemble simulations.

We calculated the canonical partition function $Z\equiv Z_{\rm ideal}Q$ and the associated dimensionless Helmholtz free energy density $\mathcal{F}\equiv-\ln(Q)/N$ for homogeneous gas phases composed of each of the four oligomers sketched in Fig. 4 and coexisting combinations of the four gas phases and four crystal phases sketched in Fig. 3. For the homogeneous monomer gas,

$$Q_{\rm monomer}=\left(\frac{1}{2\pi A}\right)^N\int(d\boldsymbol{r})^N(d\theta)^N\prod_{i<j}(1+f_{ij}), \tag{12}$$

where $f_{ij}\equiv\exp(-U_{ij}/k_{\rm B}T)-1$ is the Mayer f-function and $U_{ij}$ is the potential energy between monomers $i$ and $j$. Employing a cluster expansion [7, 8] we find

$$\mathcal{F}_{\rm monomer}=\frac{B_2^{\rm monomer}}{la^2}\phi+O(\phi^2), \tag{13}$$

where

$$B_2^{\rm monomer}=-\frac{1}{4\pi}\int d\boldsymbol{r}_{12}d\theta_{12}f_{12}^{\rm monomer} \tag{14}$$

is the second virial coefficent. The full solution for the second virial coefficient is

$$B_2^{\rm full}=\frac{1}{4\pi}\left(\nu_{\rm h}-(e^{\epsilon_{\rm n}}-1)(\nu_{\rm n}-2\nu_{\rm int}-\nu_{\rm ext})-2(e^{\epsilon_{\rm n}+\epsilon_{\rm int}}-1)\nu_{\rm int}-(e^{\epsilon_{\rm n}+\epsilon_{\rm ext}}-1)\nu_{\rm ext}\right), \tag{15}$$

where $\nu_{\rm h}$ is the configurational volume excluded by the hard cores, $\nu_{\rm n}$ is the configurational volume within the nonspecific interaction range, and $\nu_{\rm int}$ and $\nu_{\rm ext}$ are the configurational volumes within the specific interaction ranges. We define the reduced second virial coefficient

(reported on plots in the main text) as

$$B_2^\star = B_2^{\text{full}}/B_2^{\text{hard core}}, \tag{16}$$

where

$$B_2^{\text{hard core}} = \frac{\nu_{\text{h}}}{4\pi} \tag{17}$$

is the hard-core part. However, since we define the monomer gas phase as the phase containing negligible specific bonds, we neglect the specific interactions in calculating the monomer gas free energy. Thus, we use a restricted version of the second virial coefficient,

$$B_2^{\text{monomer}} = \frac{1}{4\pi}\left(\nu_{\text{h}} - (e^{\epsilon_{\text{n}}} - 1)\nu_{\text{n}}\right). \tag{18}$$

In computing the solubility curves and the thermodynamic driving forces for Fig. 1 of the main text, we cut off the solution to Eq. 14 at $2B_2^{\text{monomer}}\phi/la^2 = -1$. For $2B_2^{\text{monomer}}\phi/la^2 < -1$ we expect higher-order terms in the cluster expansion to be significant.

Defining an oligomer gas phase as a gas satisfying all specific bonds commensurate with the oligomer, we separate the inter- and intra-oligomer degrees of freedom for each $n$-mer gas to obtain

$$Q_{\text{oligomer}} = \frac{N!}{M!n!}\left(\frac{\nu_{\text{oligomer}}\phi}{2\pi N la^2}\right)^{M(n-1)} e^{N\epsilon_{\text{oligomer}}} Q_{\text{com}}, \tag{19}$$

where $M = N/n$, $(\nu_{\text{oligomer}})^{n-1}$ is the configurational volume per oligomer given a fixed center of mass and global orientation, $-\epsilon_{\text{oligomer}}k_{\text{B}}T$ is the energy per monomer, and

$$Q_{\text{com}} = \left(\frac{1}{2\pi A}\right)^M \int (d\boldsymbol{r}_{\text{com}})^M (d\theta)^M \prod_{i<j}(1 + f_{ij}^{\text{com}}) \tag{20}$$

is the configurational integral for the center-of-mass degrees of freedom. Performing a cluster expansion, we obtain

$$\mathcal{F}_{\text{oligomer}} = 1 - \frac{\ln(n)+1}{n} + \frac{n-1}{n}\ln\left(\frac{2\pi la^2}{\nu_{\text{oligomer}}\phi}\right) - \epsilon_{\text{oligomer}} + \frac{B_2^{\text{com}}}{la^2}\phi + O(\phi^2), \tag{21}$$

where

$$B_2^{\text{com}} \equiv -\frac{1}{4n\pi}\int d\boldsymbol{r}_{12} d\theta_{12} f_{12}^{\text{com}}. \tag{22}$$

To calculate $B_2^{\text{com}}$, we neglect specific bonds external to the oligomers and fix the internal degrees of freedom in their mean-field coordinates. The solutions are

$$\begin{aligned}
B_2^{\text{dimer}} &= \frac{1}{8\pi}\left(\nu_{\text{h,dimer}} - (e^{\epsilon_{\text{n}}} - 1)\nu_{\text{n,dimer}}\right), \\
B_2^{\text{tetramer}} &= \frac{1}{16\pi}\left(\nu_{\text{h,tetramer}} - (e^{\epsilon_{\text{n}}} - 1)\nu_{\text{n,tetramer}}\right), \\
B_2^{\text{trimer}} &= \frac{1}{12\pi}\left(\nu_{\text{h,trimer}} - (e^{\epsilon_{\text{n}}} - 1)\nu_{\text{n,trimer}}\right),
\end{aligned} \tag{23}$$

where $\nu_{\text{h,dimer}}$, $\nu_{\text{h,tetramer}}$, and $\nu_{\text{h,trimer}}$ are the configurational volumes excluded by the hard cores and $\nu_{\text{n,dimer}}$, $\nu_{\text{n,tetramer}}$, and $\nu_{\text{n,trimer}}$ are the configurational volumes of overlapping nonspecific interaction ranges.

For crystal phases coexisting with gas phases, we separate the crystal and gas degrees of freedom in the canonical partition function to write

$$Z_{\text{co}}(N,\phi) = Z_{\text{crystal}}(N_{\text{crystal}},\phi_{\text{crystal}})Z_{\text{gas}}(N_{\text{gas}},\phi_{\text{gas}}). \tag{24}$$

We calculate $Z(N_{\text{crystal}})$ using the cell method [9] that approximates

$$Z_{\text{crystal}}(N_{\text{crystal}},\phi_{\text{crystal}}) = Z_1(\phi_{\text{crystal}})^{N_{\text{crystal}}}, \tag{25}$$

where $Z_1$ is the partition function of a single monomer with neighboring monomers fixed at their mean-field coordinates. With some algebraic simplifications we obtain

$$\mathcal{F}_{\text{co}}(N_{\text{gas}},\phi_{\text{gas}}) = \left(1 - \frac{N_{\text{gas}}}{N}\right)\left(1 + \ln\left(\frac{2\pi la^2}{\nu_{\text{crystal}}\phi}\right) - \epsilon_{\text{crystal}}\right) + \frac{N_{\text{gas}}}{N}\left(\mathcal{F}_{\text{oligomer}}(\phi_{\text{gas}}) - \ln\left(\frac{\phi}{\phi_{\text{gas}}}\right)\right), \tag{26}$$

where $\nu_{\text{crystal}}$ is the configurational volume available to a monomer given the fixed coordinates of its neighbors. For $\phi_{\text{crystal}} >> \phi$ and $\phi_{\text{crystal}} >> \phi_{\text{gas}}$, we use $N_{\text{gas}}/N \approx \phi_{\text{gas}}/\phi$. Using this approximation and com-

bining Eq. (21) and (26) yields

$$\mathcal{F}_{\rm co}(\phi_{\rm gas}) = \left(1 - \frac{\phi_{\rm gas}}{\phi}\right)\mathcal{F}_{\rm crystal} + \frac{\phi_{\rm gas}}{\phi}\left(\mathcal{F}_{\rm olig,0} + \frac{1}{la^2}B_2^{\rm olig}\phi_{\rm gas} + \frac{1}{n}\ln(\phi_{\rm gas})\right), \tag{27}$$

where

$$\mathcal{F}_{\rm crystal} \equiv 1 + \ln\left(\frac{2\pi la^2}{\nu_{\rm crystal}\phi}\right) - \epsilon_{\rm crystal} \tag{28}$$

and

$$\mathcal{F}_{\rm olig,0} = 1 - \frac{\ln(n)+1}{n} + \frac{n-1}{n}\ln\left(\frac{2\pi la^2}{\nu_{\rm olig}}\right) - \ln(\phi) - \epsilon_{\rm olig}. \tag{29}$$

Equation 27 must be minimized with respect to $\phi_{\rm gas}$ to determine $\phi_{\rm gas}$, the packing fraction (solubility concentration) of the gas phase. The solution for $\phi_{\rm gas}$ is

$$\phi_{\rm gas}^{\rm min} = \frac{la^2}{2nB_2^{\rm olig}}W(x), \tag{30}$$

where $W$ is the Lambert W-function and

$$x \equiv \frac{2nB_2^{\rm olig}}{la^2}e^{n(\mathcal{F}_{\rm crystal} - \mathcal{F}_{\rm olig,0}) - 1}. \tag{31}$$

Notice that the dependence on $\phi$ cancels in $\mathcal{F}_{\rm crystal} - \mathcal{F}_{\rm olig,0}$ so that $\phi_{\rm gas}^{\rm min}$ is independent of $\phi$. At large interaction strength $x$ is small, and using $W(x) = x + O(x^2)$ we obtain an algebraic expression for the solubility packing fraction:

$$\phi_{\rm gas}^{\rm min} = \exp\left(n\left(\ln\left(\frac{2\pi la^2}{\nu_{\rm crystal}}\right) - \epsilon_{\rm crystal} + \frac{\ln(n)+1}{n} - \frac{n-1}{n}\ln\left(\frac{2\pi la^2}{\nu_{\rm olig}}\right) + \epsilon_{\rm olig}\right) - 1\right) \tag{32}$$

For a monomer gas, this simplifies to

$$\phi_{\rm gas}^{\rm min} = \left(\frac{2\pi la^2}{\nu_{\rm crystal}}\right)e^{-\epsilon_{\rm crystal}}. \tag{33}$$

In addition to finding phase coexistence boundaries, we use the mean-field theory to determine free energy differences. In particular, we calculate $\mathcal{F}$, the thermodynamic driving force for assembly. Since the formation of small oligomers is a much faster process than crystallization, we define $\mathcal{F}$ as the difference in free energy between the most stable stable homogeneous fluid phase (from Eq. 21) and the most stable square lattice phase (from Eq. 27).

Finally, we calculate the supersaturation at a given thermodynamic driving force by setting

$$\mathcal{F} = \mathcal{F}_{\rm oligomer} - \mathcal{F}_{\rm co}(\phi_{\rm gas}), \tag{34}$$

using $\mathcal{F}_{\rm oligomer}$ from Eq. 21 and $\mathcal{F}_{\rm co}$ from Eq. 27. If both the homogeneous and coexisting systems contain a monomer gas, rather than an oligomer gas, this results in the particular simple expression

$$\mathcal{F} = \frac{B_2^{\rm monomer}\phi_{\rm gas}}{la^2}\left(\frac{\phi}{\phi_{\rm gas}} - \frac{\phi_{\rm gas}}{\phi}\right) - 1 + \frac{\phi_{\rm gas}}{\phi} - \ln\left(\frac{\phi_{\rm gas}}{\phi}\right) \tag{35}$$

Using the low-temperature expression Eq. 33, we obtain

$$\mathcal{F} = \frac{2\pi B_2^{\text{monomer}}}{\nu_{\text{crystal}}} e^{-\epsilon_{\text{crystal}}} \left( \frac{\phi}{\phi_{\text{gas}}} - \frac{\phi_{\text{gas}}}{\phi} \right) - 1 + \frac{\phi_{\text{gas}}}{\phi} - \ln\left( \frac{\phi_{\text{gas}}}{\phi} \right) \tag{36}$$

For large values of $\epsilon_{\text{crystal}}$ (low temperatures), we may neglect the first term, resulting in a concentration- and temperature-independent expression for the supersaturation $S$,

$$S \equiv \frac{\phi}{\phi_{\text{gas}}} = \frac{-1}{W(-\exp(-1-\mathcal{F}))}, \tag{37}$$

where, again, $W$ is the Lambert W-function. For large values of $\mathcal{F}$, this approximates to

$$S = e^{1+\mathcal{F}} - 1 + O\left( \left( e^{-1-\mathcal{F}} \right)^2 \right). \tag{38}$$

Thermodynamic driving forces of 1, 2, and 3 correspond to supersaturations of 6.3, 19.1, and 53.6. We can therefore transform our rule of thumb for good assembly, $\mathcal{F} = 1 - 2\,k_{\text{B}}T$, into a rule of thumb for the supersaturation, $S = 5 - 20$. We stress that the conversion between thermodynamic driving force and supersaturation is independent of concentration and temperature, as long as the temperature is low enough so that the first term in Eq. 37 may be neglected. We therefore expect that a similar window of optimal supersaturation may exist for real protein systems, though the precise value of the window will depend on the optimal values of $\mathcal{F}$.

To determine the numerical values of the phase boundaries, reduced second virial coefficients, and free energy differences, we must calculate the configurational volumes in the theory. We calculate the configurational volumes $\nu_{\text{h}}$, $\nu_{\text{n}}$, $\nu_{\text{h,dimer}}$, $\nu_{\text{n,dimer}}$, $\nu_{\text{h,tetramer}}$, and $\nu_{\text{n,tetramer}}$ by determining the excluded area at fixed relative orientation $\theta$, as depicted in Fig. 5, and then integrating over $\theta$ using Mathematica [10]. We assume that bound monomers are oriented either perpendicular or parallel and are separated by a distance $0.1a$ equal to one-half the specific interaction range. We obtain

$$\begin{aligned}
\nu_{\text{hard}} &= 73.886a^2, \\
\nu_{\text{n}} &= 45.296a^2, \\
\nu_{\text{hard,dimer}} &= 162.093a^2, \\
\nu_{\text{n,dimer}} &= 66.386a^2, \\
\nu_{\text{hard,tetramer}} &= 343.416a^2, \\
\nu_{\text{n,tetramer}} &= 99.745a^2.
\end{aligned} \tag{39}$$

Because of the low symmetry of the trimer, the configurational integrals are much more complicated, so we do not solve them. Instead, we estimate $\nu_{\text{h,trimer}} = (\nu_{\text{h,tetramer}} + \nu_{\text{h,dimer}})/2$ and $\nu_{\text{n,trimer}} = (\nu_{\text{n,tetramer}} - \nu_{\text{n,dimer}})/2$. This introduces only a small error to the free energy, because the configurational volumes only contribute logarithmically to the free energy.

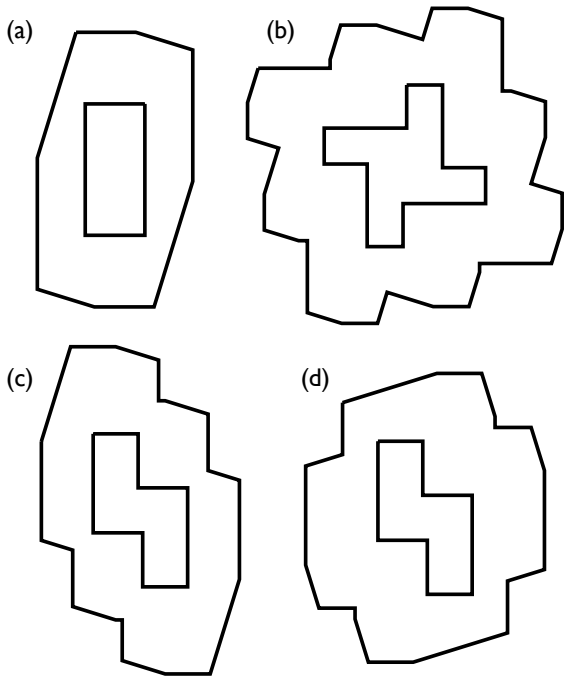


FIG. 5: Examples of areas excluded by the hard cores at arbitrary relative orientation $\theta$: (a) monomers, (b) tetramers, (c) dimers at small $\theta$, and (d) dimers at $\theta$ near $\pi/2$.

We determine the configurational volumes $\nu_{\text{int}}$, $\nu_{\text{ext}}$, $\nu_{\text{dimer}}$, $(\nu_{\text{tetramer}})^3$, and $(\nu_{\text{trimer}})^2$ by integrated over configurations satisfying both the hard-core and the specific interaction constraints. The integrals for $\nu_{\text{int}}$, $\nu_{\text{ext}}$, $\nu_{\text{dimer}}$, and $\nu_{\text{trimer}}$, are identical, so $\nu_{\text{int}} = \nu_{\text{ext}} = \nu_{\text{dimer}} = \nu_{\text{trimer}}$. We solve the remaining integrals by numerical integration on a regular grid. We obtain

$$\begin{aligned}
\nu_{\text{dimer}} &= 0.02157a^2, \\
\nu_{\text{tetramer}} &= 0.0118a^2.
\end{aligned} \tag{40}$$

To solve the crystal configurational volumes $\nu_{\text{rectangular}}$, $\nu_{\text{square}}$, $\nu_{\text{cpd}}$, and $\nu_{\text{cpt}}$, we fix the mean-field orientations of neighboring monomers at right angles and constrain the bound specific interaction patches to line up. Then, we calculate the configurational volume available to the one freely moving monomer as a function of one or more inter-monomer spacings, and we maximize the volume with respect to the spacing(s). We calculate the volumes subject to the constraint that the energy is the maximum energy characteristic of the crystal, and we used a combination of analytic and numeric techniques in Mathematica [10] to perform the calculations.

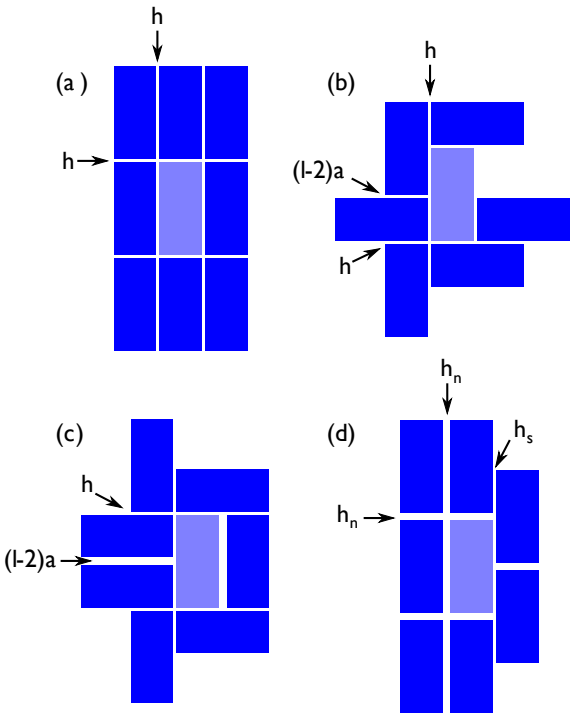


FIG. 6: Sketches of mean-field configurations and spacings used to calculate the crystal configurational volumes: (a) rectangular crystal, (b) square lattice, (c) close-packed tetramer crystal, and (d) close-packed dimer crystal.

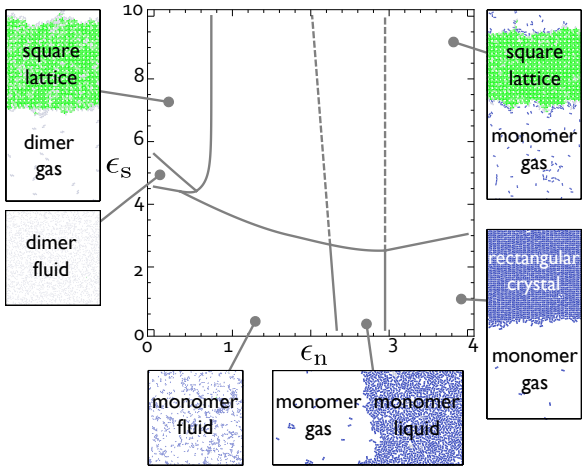


FIG. 7: Phase diagram for $\phi = 0.1$ as a function of nonspecific interaction $\epsilon_n$ and specific interaction $\epsilon_s$ for $\epsilon_s \equiv \epsilon_{int} = 2\epsilon_{ext}$. Solid (dashed) grey curves denote the stable (metastable) boundaries for the labeled simulated coexistence combinations. All boundaries were calculated using analytic theory, except for the boundary between homogeneous and phase-separated monomer fluids; this was determined using Gibbs ensemble simulations. The surrounding simulation snapshots label the equilibrium phase or coexisting phases within each region of the phase diagram.

We sketch the mean-field configurations in Fig. 6. For the rectangular crystal (panel (a)), we assume that the mean-field spacing $h$ between monomers is the same along the long and short directions. We obtain a mean-field spacing $h = 0.1932a$ and configurational volume $\nu_{\text{rectangular}} = 0.04085a^2$. For the square lattice (panel (b)), lining up the specific patches constrains the spacing between nonspecifically bound, perpendicular monomers to be $(l-2)a = 0.2a$. Symmetry and the previously described constraints dictate that the spacings between internally bound and externally bound monomers are the same. We obtain a mean-field spacing $h = 0.10a$ and a configurational volume $\nu_{\text{square}} = 0.00500$. For the close-packed tetramer crystal (panel (c)), symmetry dictates that the mean-field spacing between monomers that share a long edge should be $(l-2)a = 0.2a$. We obtain a spacing between perpendicularly oriented monomers $h = 0.10a$ and a configurational volume $\nu_{\text{cpt}} = 0.00892$. For the close-packed dimer crystal (panel (d)), we assume that the mean-field separations constrained by the nonspecific interaction–that is, the separation along the short edge and the separation along the long edge with no patchy interactions–are the same, but that the mean-field separation constrained by the patchy interaction is different. We obtain a nonspecifically bound spacing $h_n = 0.19$, a specifically-bound spacing $h_s = 0.18$, and a configurational volume $\nu_{\text{cpd}} = 0.0138$.

In the main text, we present results for a single choice of the ratio between external and internal interaction strength, $\epsilon_{ext}/\epsilon_{int} = 2$. However, if we allow both specific interaction strengths $\epsilon_{int}$ and $\epsilon_{ext}$ to vary separately, we find a total of 11 stable phase combinations: homogenous fluid phases of monomers, dimer, and tetramers; the square lattice coexisting with the monomer, dimer, and tetramer gases; the rectangular crystal coexisting with the monomer gas; the close-packed tetramer crystal coexisting with the monomer and tetramer gases; and the close-packed dimer crystal coexisting with the monomer and dimer gases. In addition, as mentioned in the previous section, Gibbs ensemble simulations reveal two additional phase combinations, a liquid of monomers coexisting with the monomer gas and a liquid of dimers coexisting with the dimer gas. We find that phases involving a gas of trimers are never stable.

Along the slice of parameter space discussed in the main text, $\epsilon_{ext}/\epsilon_{int} = 2$, we find the 6 stable phase combinations labeled in Fig. 7 and appearing in the analogous phase diagram in Fig. 3 of the main text. Notice that in addition to the monomer gas, the monomer liquid, the square lattice, and the rectangular crystal discussed in the main text, we also find phase combinations involving a dimer gas at low $\epsilon_n$. If we decrease $\epsilon_{ext}/\epsilon_{int}$, the dimer gas disappears from the phase diagram, but the dependence of the yield and pathway on $\epsilon_n$ and $\epsilon_{int}$ does not qualitatively change. We find that our design rules persist as we vary $\epsilon_{ext}/\epsilon_{int}$. As we will discuss in a subsequent publication, the detrimental effects of nonspecific

aggregation are exacerbated when $\epsilon_{\rm ext}/\epsilon_{\rm int}$ departs substantially from 2, and the window of moderate thermodynamic driving force remains a necessary condition for efficient crystallization.


[1] S. Whitelam and P. L. Geissler, J. Chem. Phys. **127**, 154101 (2007).

[2] S. Whitelam, E. Feng, M. Hagan, and P. Geissler, Soft Matter **5**, 1251 (2009).

[3] S. Whitelam, Molecular Simulation **37**, 606 (2011).

[4] S. Whitelam, Phys. Rev. Lett. **105**, 088102 (2010).

[5] B. Chen and J. I. Siepmann, J. Phys. Chem. B **104**, 8725 (2000).

[6] A. Z. Panagiotopoulos and M. R. Stapleton, Fluid Phase Equilibr. **53**, 133 (1989).

[7] N. G. V. Kampen, Physica **27**, 783 (1961).

[8] W. J. Mullin, Am. J. Physics **40**, 1473 (1972).

[9] J. E. Lennard-Jones and A. F. Devonshire, Proc. Royal Soc. London A **163**, 53 (1937).

[10] *Mathematica, Version 7.0* (Wolfram Research, Inc, Champaign, IL, 2008).