Design rules for the self-assembly of a protein crystal

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

Theoretical guidance for protein crystallization based on spheres that form close-packed structures predicts that optimal assembly occurs within a 'slot' of second virial coefficients near the metastable liquid-vapor critical point. However, most protein crystals are open structures stabilized by anisotropic interactions. Using theory and simulation, we show that the crystallization slot is not a sufficient condition for good assembly of a model protein crystal stabilized by anisotropic interactions, and large density fluctuations associated with the critical point generally *conflict* with symmetry fluctuations required for crystallization. Instead, good assembly requires a thermodynamic driving force of about $k_{\mathrm{B}}T$ and an optimization of interaction specificity. Our findings suggest that self-assembly of open crystals is more akin to viral capsid self-assembly than to the crystallization of spherical colloids.

The need to crystallize proteins for X-ray studies has spurred the development of theories of protein crystallization. These theories are largely based on the behavior of spheres with short-range isotropic attractions, a representation motivated by two observations: phase diagrams for typical proteins and spherical colloids with short range attractions are structurally similar, possessing a metastable liquid-vapor binodal [1–4]; and both proteins and spherical colloids tend to crystallize when solution osmotic second virial coefficient lies in a defined 'crystallization slot' [5–7]. On the computer, short-range isotropic spheres crystallize poorly above the metastable liquid-vapor binodal and show enhanced nucleation rates near or below it [3, 8–12]. Such enhancement is indeed seen in some protein solutions [13–15]. However, other experiments show disparities with this picture. Proteins can crystallize readily above the binodal [16, 17] and experience kinetically-impaired crystallization below it [18]. They can also lie in the crystallization slot and not crystallize [19]. In addition, although the *structure* of protein- and colloid phase diagrams is similar, the microscopic nature of the stable solid is not: most proteins do not form close-packed crystals [20].

These disparities motivate a theoretical approach to protein crystallization that acknowledges additional features of proteins' interactions, particularly their anisotropy [21–26]. This approach suggests that rules for optimal assembly of open structures are different from the rules for optimal assembly of close-packed structures. Here we explicitly demonstrate this difference. We have used extensive equilibrium and nonequilibrium numerical simulations and quantitatively accurate mean-field theory to exhaustively determine the design rules for optimal assembly of a model patterned after the SbpA surface-layer protein. The latter forms a porous square lattice with a tetrameric repeat unit on the surface of the bacterium *Lysinibacillus sphaericus*, and *in vitro* on surfaces or in solution [27–30]. We impose a simple set of model protein interactions that stabilize the two condensed phases observed in experiments: *specific* interactions to stabilize the open crystal structure [31] and *nonspecific* interactions to stabilize unstructured aggregates observed in recent experiments on lipid bilayers [30]. A similar distinction between orientationally specific and nonspecific interactions has been considered in models of polymer crystallization [32]. Such a model is crucially different from isotropic models in that the same microscopic interaction does not stabilize *both* crystal and liquid phases. Instead, specific and nonspecific interactions independently drive distinct critical behaviors [32–34]. Consequently, we find that design rules for assembly differ from those of spheres. While large density fluctuations promote crystallization of close-packed spheres [3], they tend to *inhibit* the symmetry fluctuations required to achieve assembly of the open S-layer lattice. Further, the second virial coefficient B_2, an orientationally-averaged measure of protein-protein attraction, bounds good assembly but does not predict it. It is intuitively reasonable that it should bound assembly: too strong an attraction results in kinetic trapping, while too weak an attraction suppresses crystal nucleation. However, its orientational averaging renders it blind to the microscopic origin of the attraction: model proteins with identical values of B_2 but different combinations of specific and nonspecific interactions can assemble well, poorly, or not at all.

Instead, we find that good assembly *can* be predicted by a combination of two design rules: the thermodynamic driving force for crystallization (defined as the free energy difference between the gas and the crystal) must be $1-2\,k_{\mathrm{B}}T$, and interactions should be made as nonspecific as possible without promoting liquid-vapor phase separation. In experimental terms, our results suggest adjusting solution conditions in order to impose a defined supersaturation at the liquid-vapor binodal. While crystallization *can* happen at or below the binodal, we find that such large-scale nonspecific association usually leads to slow dynamics and poor yield. Taken in the context of recent simulation work [35–37], our findings suggest that the rules governing the assembly of protein crystals are in important ways more like those governing viral capsid self-assembly than those underpinning the crystallization of spherical colloids.

Model and methods. We consider a generalization of the model SbpA surface-layer protein introduced in Ref. [24]. Hard rectangular monomers of width a ($\equiv 3.9$

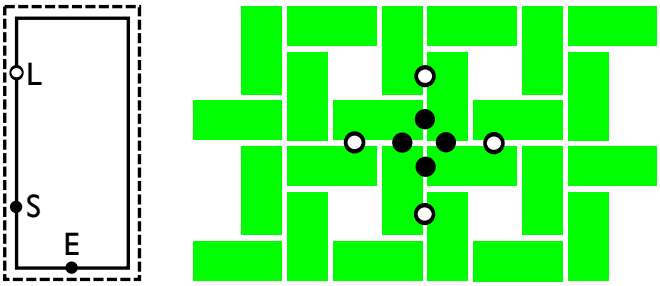

FIG. 1: (a) Monomer geometry. (b) Square lattice stabilized by the two chemically specific interactions: internal bonds (filled circles) and external bonds (open circles).

nm) and length la ($l = 2.2$) live on a smooth, two-dimensional substrate. Monomer interactions acknowledge the tendency of SbpA proteins to form both unstructured aggregates [30] and an open square lattice of tetramers [31]. To allow formation of the square lattice, monomers are decorated by three patches labeled E (edge), S (short arm) and L (long arm), each located on the hard-core boundary a distance $a/2$ from the nearest vertex, as shown in Fig. 1. Patches mediate a chemically specific internal bond of energy $-\epsilon_{\rm int}\, k_{\rm B}T$ if the E and S patches of neighboring monomers approach closer than $\Delta = a/5$ and an external bond of energy $-\epsilon_{\rm ext}\, k_{\rm B}T$ if two L patches of neighboring monomers approach closer than Δ. To permit unstructured aggregation, monomers also experience a nonspecific attraction of energy $-\epsilon_{\rm n}\, k_{\rm B}T$ if their surrounding rectangular forcefields (of width $a+2\Delta$ and length $la + 2\Delta$) overlap.

To determine design rules for assembly, we extensively varied all three energetic parameters and the packing fraction ϕ. We will discuss the effects of separately varying $\epsilon_{\rm int}$ and $\epsilon_{\rm ext}$ elsewhere. Here, we present results for $\epsilon_{\rm ext}/\epsilon_{\rm int} = 2$ in terms of a single specific interaction parameter $\epsilon_{\rm s} \equiv \epsilon_{\rm int} = 2\epsilon_{\rm ext}$. We find that the results presented here are largely insensitive to the choice of the ratio $\epsilon_{\rm ext}/\epsilon_{\rm int}$.

We solved model thermodynamics in two ways, as detailed in the Supplementary Methods. we used analytic mean-field theory to determine the thermodynamic driving force for assembly, phase boundaries for stable and metastable phases, and reduced second virial coefficients $B_2^\star \equiv B_2/B_2^{\rm hard\,core}$. $B_2 = (4\pi)^{-1}\int d\boldsymbol{r}_{12} d\theta_{12}\,(1-{\rm e}^{-\beta U_{12}})$ is calculated in the conventional way, integrating over the phase space of two model proteins interacting via the energy U_{12}. The hard-core normalization $B_2^{\rm hard\,core}$ is obtained similarly, but for a system with no attractive interactions. We calculated phase diagrams numerically using direct coexistence and Gibbs ensemble simulations [38]. We find that phase diagrams calculated by mean-field theory and simulation agree, except that simulation reveals a narrow region of thermodynamically stable liquid that the mean-field theory does not attempt to account for.

We determined self-assembly dynamics using virtual-move Monte Carlo simulations [39–41] of 1024 monomers at constant packing fraction, starting from well-mixed conditions. Although a truly physical dynamics cannot be effected by simulations that do not explicitly represent solvent, some important aspects of real overdamped motion are retained by this algorithm: bodies move locally according to potential energy gradients, and collective diffusion constants can be scaled according to cluster size and shape. Here, we parameterized the algorithm to ensure that tightly-bound protein clusters of hydrodynamic radius R diffuse according to the Stokes solution for the overdamped motion of a sphere of radius R, resulting in diffusive behavior for moderate to large clusters that is considerably more realistic than that effected by basic Brownian dynamics integrators. Taking $a = 3.9$ nm, $T = 300$ K, and solution viscosity $\eta = 1.00 \times 10^{-3}$ Pa s, each Monte Carlo (MC) cycle corresponds to 2.42 ns.

Results. We carried out two numerical protocols, each designed to mimic a particular experiment. First, for three selected 'proteins,' each with a different balance of specific and nonspecific interactions, we determined where on the conventional temperature-concentration phase diagram yield is best. Second, we determined the microscopic mechanisms for optimal assembly by independently varying specific and nonspecific interaction strength. As we will show, such a protocol mimics studying a large ensemble of related proteins *or* varying solvent chemistry to optimize assembly for a single protein.

In Fig. 2 (a) we show temperature-concentration phase diagrams for three model proteins: *specifin*, with $\epsilon_{\rm s}/\epsilon_{\rm n} = 2$; *intermedin*, with $\epsilon_{\rm s}/\epsilon_{\rm n} = 1.5$; and *nonspecifin*, with $\epsilon_{\rm s}/\epsilon_{\rm n} = 1$. From protein to protein, the solubility curve shifts with interaction specificity more than the liquid-vapor binodal, leading to a change of phase diagram structure [32] similar to that effected by changing the range of attraction of an isotropic sphere [3, 42–44]. Specifin and intermedin display a metastable liquid-vapor coexistence, while nonspecifin displays a stable liquid-vapor coexistence. Intermedin and nonspecifin also display a transition from a square lattice ($\phi \approx 0.70$) to a close-packed crystal ($\phi \approx 0.76$) at high temperature.

To reveal how well these proteins crystallize, we overlay phase diagrams with color maps quantifying the crystal yield obtained after long dynamic simulations. Green indicates high yield; red, low yield. Specifin self-assembles best above the liquid-vapor critical point, intermedin assembles best near or just below it, and nonspecifin crystallizes poorly throughout its phase diagram. Below the binodal, monomers generally form kinetically sluggish gel-like or microcrystalline clusters that lead to poor yield. We illustrate dynamic trajectories leading to these outcomes in Fig. 2 (b) by showing snapshots at early (10^5 MC cycles) and late (5×10^6 MC cycles) times for near-optimal conditions for each protein. (See also the corresponding Movies S1, S2, and S3.)

For this set of proteins, optimal assembly does not track the liquid-vapor binodal. Moreover, assembly is not

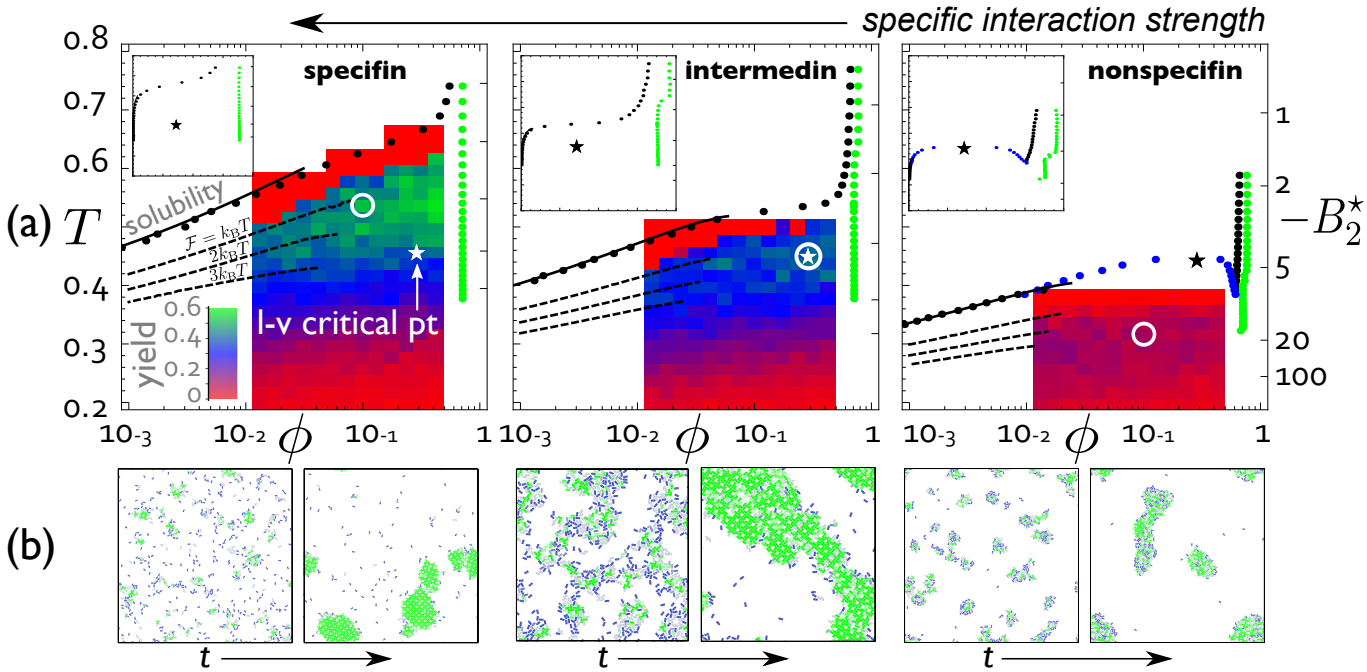

FIG. 2: *Optimal yield is not predicted by $B_2^\star$ or the position of the liquid-vapor binodal.* **(a)** Phase diagrams (temperature $T \equiv 1/\epsilon_n$ vs packing fraction ϕ) for three model proteins whose interaction specificities decrease from left to right. We overlay the analytic solubility curve (solid), which agrees with the numeric data with no adjustable parameters, lines of fixed driving force $\mathcal{F}$ (dashed), and a color map of square lattice yield (obtained after dynamic simulations of 10^7 MC cycles). The position of best yield does not track the liquid-vapor binodal (critical points shown as stars) or a fixed slot of $B_2^\star$ (right ticks). Insets show phase diagrams on a more conventional linear horizontal scale. **(b)** Snapshots show example dynamic simulations for each protein after 10^5 (left) and 5×10^6 (right) MC cycles for the points on the phase diagrams labeled by open circles.

predicted by the crystallization slot [5], which provides a necessary but not a sufficient condition for crystallization: good assembly indeed generally occurs in a slot $-100 \lesssim B_2^\star \lesssim -2$, but peak yield within this slot can be highly localized (specifin), or uniformly poor (nonspecifin). From our analytic theory we calculated lines of constant $\mathcal{F}$, the thermodynamic driving force for crystallization. The proteins that assemble well do so in the window $\mathcal{F} = 1 - 2\,k_\mathrm{B}T$. Our analytic theory demonstrates that this window is equivalent to a supersaturation $S \equiv \phi/\phi_\mathrm{gas}(T) = 5 - 20$, where $\phi_\mathrm{gas}(T)$ is the solubility packing fraction.

The temperature-concentration diagrams of Fig. 2 reveal where yield is good and where it is bad. To identify the molecular dynamical origins of these behaviors it is necessary to survey an ensemble of related proteins. We do this in Figs. 3 and 4 by independently varying the specific and nonspecific interactions at fixed packing fraction $\phi = 0.1$. Fig. 3 shows a color map of crystal yield overlaid on the phase diagram spanned by the two interactions. The surrounding simulation snapshots label the equilibrium phase or coexisting phases within each region of the phase diagram. The three proteins of Fig. 2 lie on the dashed yellow lines. Fig. 4 shows 'pathway diagrams' [34] which identify, along self-assembly trajectories, the maximum fractions of 'misbound' proteins (those with their external specific bond satisfied but only one of their two internal specific bonds satisfied) and nonspecifically aggregated proteins (those having no specific bonds and two or more nonspecific bonds).

Optimal yield occurs in the part of the phase diagram identified by two conditions: 1) the thermodynamic driving force $\mathcal{F}$ for crystallization must be $1 - 2\,k_\mathrm{B}T$, and 2) the nonspecific attraction must be as large as possible without inducing liquid-vapor phase separation. The window of optimal $\mathcal{F}$ lies between the weakly supersaturated region near the solubility curve, where nucleation barriers are too high for crystallization to happen in the alloted simulation time, and the strongly supersaturated region at large ϵ_s, where misbinding predominates (see Fig. 4). The thermodynamic driving force is substantially more predictive than the second virial coefficient; while most good assembly occurs with the displayed slot $-100 \lesssim B_2^\star \lesssim -5$, this slot includes large parts of the phase diagram where assembly is poor or nonexistent, including regions where the target crystal is not stable.

Within the window $\mathcal{F} = 1 - 2\,k_\mathrm{B}T$, yield initially increases as specific interaction strength is traded for nonspecific interaction strength. We argue that this is due to a combination of two effects. First, the nonspecific attraction provides a larger interaction 'cross-section' than do the sticky patches alone, enabling diffusing particles

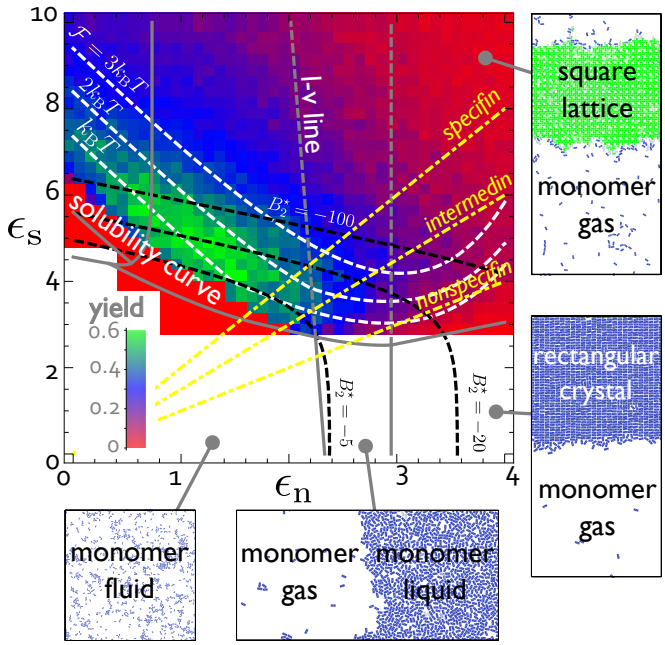

FIG. 3: *Weak nonspecific association is optimal.* Phase diagram and dynamic yield color map for $\phi = 0.1$ and a range of protein specific and nonspecific interactions. Solid (dashed) grey curves denote the stable (metastable) boundaries for the labeled simulated coexistence combinations. All boundaries were calculated using analytic theory, except for the boundary between homogeneous and phase-separated monomer fluids; this was determined using Gibbs ensemble simulations. The dashed black (white) curves denote lines of constant B_2^* (driving force $\mathcal{F}$). Dash-dotted yellow lines represent the proteins from Fig. 2; for a fixed protein, temperature increases to the left along these lines.

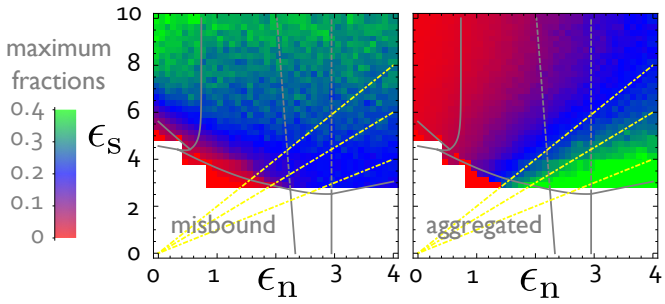

FIG. 4: *Large-scale nonspecific aggregation hinders crystal assembly.* 'Pathway diagrams' show color maps of the maximum fractions of misbound and nonspecifically aggregated proteins for $\phi = 0.1$ and a range of protein specific and nonspecific interactions (see Fig. 3). Comparison of the yield (Fig. 3) and pathway color maps reveals a rule of thumb for optimizing assembly: impose the strongest nonspecific interaction that does not induce liquid-vapor phase separation.

to associate more rapidly and thereby enhancing crystallization kinetics. Second, weak nonspecific interactions help to stabilize the target structure, allowing specific interactions to be made weaker at fixed thermodynamic driving force. Multiple weak interactions are easier to break than few strong interactions and allow defective binding motifs to be corrected more easily; this is the idea of microscopic 'reversibility' [37, 45, 46].

However, the trend of increasing yield with decreasing specificity terminates at the metastable liquid phase boundary. As the 'aggregated' pathway diagram of Fig. 4 indicates, this phase boundary signals the onset of large-scale nonspecific aggregation. Density fluctuations associated with phase separation therefore conflict with the symmetry fluctuations required to stabilize the open crystal lattice. This behavior is strikingly distinct from that of isotropic models, which crystallize best at the metastable critical point. However, assembly of model capsomer proteins into icosahedral viral capsids [35–37] shares the behavior of the present model; there, assembly is also impaired by liquid-like aggregation at low interaction specificity [35]. We conjecture that for *open* structures in two and three dimensions stabilized by *anisotropic* attractions, weak nonspecific association should aid assembly, but large density fluctuations associated with the liquid-vapor critical point should generally impair it.

Note that nonspecifin (lowest dash-dotted yellow line in Fig. 3) could be transformed into specifin (highest dash-dotted yellow line in Fig. 3) if its crystal-forming contacts were strengthened. Extrapolating our results to real protein systems suggests that solvent chemistry might be adjusted (even at constant B_2^*) to optimize the ability of a defined protein to crystallize. For instance, recent experiments suggest that multivalent salt concentration can be adjusted to tune specific protein-protein interactions [47].

Conclusions. Typical protein phase diagrams resemble those of isotropic colloids bearing short-range attractions, but, crucially, they describe different solid structures. Here we have shown that the self-assembly of an open crystal formed by a surface-layer protein is different in significant ways from the assembly of a close-packed crystal. However, it can be rationalized by a set of relatively simple design rules. First, the thermodynamic driving force for crystallization must be of order k_BT. Second, interactions should be adjusted to trade specific interaction strength for *weak* nonspecific association; substantial nonspecific aggregation is deleterious. Comparison of our results with those for model viral capsid self-assembly [35–37] suggest that our design rules are common to both extended *and* closed structures and hold in both two and three dimensions.

In experimental terms, our results suggest quantitative guidelines for optimizing crystal yield in real protein systems. Our window of optimal thermodynamic driving force corresponds to a supersaturation of 5 to 20. Combined with the observation that metastable liquid-vapor phase separation hinders crystal assembly, this observation suggests that there exists an optimal 'metastability

gap' [48] between the solubility curve and the binodal. Based on the present model, we advocate adjusting solution conditions to impose a supersaturation at the binodal of about 20.

Acknowledgements. We thank Caroline Ajo-Franklin, Robert Jack, Behzad Rad, and Jeremy Schmit for discussions, and we acknowledge NERSC for computing facilities. This work was performed at the Molecular Foundry, Lawrence Berkeley National Laboratory, and was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02–05CH11231.

[1] D. Rosenbaum, P. C. Zamora, and C. F. Zukoski, Phys. Rev. Lett. **76**, 150 (1996).
[2] D. F. Rosenbaum and C. F. Zukoski, J. Crystal Growth **169**, 752 (1996).
[3] P. R. ten Wolde and D. Frenkel, Science **277**, 1975 (1997).
[4] D. F. Rosenbaum, A. Kulkarni, S. Ramakrishnan, and C. F. Zukoski, J. Chem. Phys. **111**, 9882 (1999).
[5] A. George and W. W. Wilson, Acta Cryst. **D50**, 361 (1994).
[6] A. M. Kulkarni, N. M. Dixit, and C. F. Zukoski, Faraday Discuss. **123**, 37 (2003).
[7] T. Gibaud, F. Cardinaux, J. Bergenholtz, A. Stradner, and P. Schurtenberger, Soft Matter **7**, 857 (2011).
[8] V. Talanquer and D. W. Oxtoby, J. Chem. Phys. **109**, 223 (1998).
[9] K. G. Soga, J. R. Melrose, and R. C. Ball, J. Chem. Phys. **110**, 2280 (1999).
[10] D. Costa, P. Ballone, and C. Caccamo, J. Chem. Phys. **116**, 3327 (2002).
[11] A. Fortini, E. Sanz, and M. Dijkstra, Phys. Rev. E **78**, 041402 (2008).
[12] S. Babu, J.-C. Gimel, and T. Nicolai, J. Chem. Phys. **130**, 064504 (2009).
[13] O. Galkin and P. G. Vekilov, Proc. Natl. Acad. Sci. U.S.A. **97**, 6277 (2000).
[14] W. Pan, A. B. Kolomeisky, and P. G. Vekilov, J. Chem. Phys. **122**, 174905 (2005).
[15] P. G. Vekilov, Nanoscale **2**, 2346 (2010).
[16] M. Muschol and F. Rosenberger, J. Chem. Phys. **107**, 1953 (1997).
[17] Y. Liu, X. Wang, and C. B. Ching, Cryst. Growth Des. **10**, 548 (2010).
[18] S. Gorti, J. Konnert, E. Forsythe, and M. Pusey, Cryst. Growth Des. **5**, 535 (2005).
[19] W. W. Wilson, J. Struct. Biol. **142**, 56 (2003).
[20] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. H. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[21] N. Kern and D. Frenkel, J. Chem. Phys. **118**, 9882 (2003).
[22] J. P. K. Doye, A. A. Louis, I.-C. Lin, L. R. Allen, E. G. Noya, A. W. Wilber, H. C. Kok, and R. Lyus, Phys. Chem. Chem. Phys. **9**, 2197 (2007).
[23] H. Liu, S. K. Kumar, and J. F. Douglas, Phys. Rev. Lett. **103** (2009).
[24] S. Whitelam, Phys. Rev. Lett. **105**, 088102 (2010).
[25] F. Romano, E. Sanz, and F. Sciortino, J. Chem. Phys. **134**, 174502 (2011).
[26] F. Romano and F. Sciortino, arxiv:1101.3877.
[27] U. B. Sleytr, M. Sára, D. Pum, and B. Schuster, Prog. Surf. Sci. **68**, 231 (2001).
[28] U. B. Sleytr, E. Gy orvary, and D. Pum, Prog. Org. Coat. **47**, 279 (2003).
[29] U. B. Sleytr, C. Huber, N. Ilk, D. Pum, B. Schuster, and E. M. Egelseer, FEMS Microbiol. Lett. **267**, 151 (2007).
[30] S. Chung, S.-H. Shin, C. R. Bertozzi, and J. J. D. Yoreo, Proc. Natl. Acad. Sci. U.S.A. **107**, 16536 (2010).
[31] J. E. Norville, D. F. Kelly, T. F. K. Jr., A. M. Belcher, and T. Walz, J. Struct. Biol. **160**, 313 (2007).
[32] W. Hu and D. Frenkel, Adv. Polym. Sci. **191**, 1 (2005).
[33] S. Whitelam, J. Chem. Phys. **132**, 194901 (2010).
[34] L. Hedges and S. Whitelam, arXiv:1103.0334 (2011).
[35] A. W. Wilber, J. P. K. Doye, A. A. Louis, E. G. Noya, M. A. Miller, and P. Wong, J. Chem. Phys. **127**, 085106 (2007).
[36] M. F. Hagan and D. Chandler, Biophys. J. **91**, 42 (2006).
[37] D. Rapaport, Phys. Rev. Lett. **101**, 186101 (2008).
[38] A. Z. Panagiotopoulos and M. R. Stapleton, Fluid Phase Equilibr. **53**, 133 (1989).
[39] S. Whitelam and P. L. Geissler, J. Chem. Phys. **127**, 154101 (2007).
[40] S. Whitelam, E. Feng, M. Hagan, and P. Geissler, Soft Matter **5**, 1251 (2009).
[41] S. Whitelam, Molecular Simulation **37**, 606 (2011).
[42] A. P. Gast, W. B. Russell, and C. K. Hall, J. Colloid Interface Sci. **96**, 251 (1983).
[43] H. Liu, S. Garde, and S. Kumar, J. Chem. Phys. **123**, 174505 (2005).
[44] D. L. Pagan and J. D. Gunton, J. Chem. Phys. **122**, 184515 (2005).
[45] R. Jack, M. Hagan, and D. Chandler, Phys. Rev. E **76**, 021119 (2007).
[46] J. Grant, R. L. Jack, and S. Whitelam, ArXiv e-prints (2011), 1108.4542.
[47] F. Zhang, G. Zocher, A. Sauter, T. Stehle, and F. Schreiber, J. Appl. Cryst. **44**, 755 (2011).
[48] N. Asherie, A. Lomakin, and G. B. Benedek, Phys. Rev. Lett. **77**, 4832 (1996).