Soft Matter

Crystallizing protein assemblies via free and grafted linkers

Journal:	*Soft Matter*
Manuscript ID	SM-ART-04-2019-000693.R1
Article Type:	Paper
Date Submitted by the Author:	24-Apr-2019
Complete List of Authors:	Dahal, Yuba; Northwestern University, Material Science and Engineering Olvera de la Cruz, Monica; Northwestern University, Materials Science and Engineering

Journal Name

ARTICLE TYPE

Cite this: DOI: 10.1039/xxxxxxxxxx

Crystallizing protein assemblies via free and grafted linkers

Yuba Raj Dahal[a] and Monica Olvera de la Cruz[a,b,c,*]

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

Porous protein superlattices have plausible catalytic applications in biotechnology and nanotechnology. They are solid yet open structures with the potential of preserving the activity of enzymes. However, there is still a lack of understanding of the design parameters that are required to arrange proteins in a periodic porous fashion. Here, we introduce a coarse–grained molecular dynamics (MD) simulation approach to study the effects of length and geometry of linkers on the stability of $3D$ crystalline assemblies of metal ions anchored ferritin protein. By simulating a system of proteins (eight metal ions anchored sites per protein) and linkers (two free ends per linker), we find that there is a range of optimal linker lengths for crystalline order. The optimal linker length is found to depend on the linker to protein concentration ratio and binding energy. We also examine the case of grafted flexible linkers on the protein surface as an alternative route for constructing highly porous crystalline structures. Our study demonstrates that the length of grafted linkers is a better tunable parameter than the length of free linkers to achieve high porosity protein superlattices. The computational study developed here provides guidelines to assemble biomolecules into crystals with high porosity.

Introduction

Proteins are chemically and physically diverse biomolecules with important catalytic functions. The possibility of synthesizing protein–based functional materials has incited multiple studies to direct their assembly including the use of controlled protein functionalization [1–4] and the design of random copolymers sequences that can protect the enzymatic activity of the protein [5–7]. Periodic arrays of proteins are particularly attractive for applications as separation materials, and as heterogeneous catalysts to name a few due to their uniform pore size distribution [8,9]. However, there are challenges to arranging proteins in open periodic structures. Among the many complexities, the heterogeneous chemical composition and asymmetric shape of the proteins are the basic hurdles in the study of protein–protein interactions, which depend on several external parameters such as solution pH, salt types, and concentration [10–12], etc. Approaches such as surface functionalizations have guided protein assemblies into particular shapes and/or crystalline structures [1,14–18]. The metal organic frameworks ($MOFs$) design strategy has also been widely exploited to build porous crystalline structures [19–21]. In recent years, the Tezcan group has utilized $MOFs$ design approaches to assemble ferritin proteins in crystalline structures. First, they chemically anchored metal ions from transition II group at the $c3$ symmetric interfaces of ferritin protein and then by using the organic linkers, they built bcc and/or bct types of $3D$ crystalline structures of metal–protein hybrid system [2,4].

The linker directed crystalline assemblies of metal–protein integrated systems (protein–$MOFs$) are rich porosity materials like $MOFs$. In addition, they have periodic chemical diversity around the pores which, along with pore size, is sought as a tunable parameter [8] . In protein–$MOFs$, the major driving force in constructing $3D$ multi–component assembly is the specific interactions between two ends of organic linkers and metal ions anchored sites on the protein surface. The linker length and geometry are potential parameters for controlling the interparticle separation, crystalline structures, and the porosity of lattices. For example, by using a shorter linker length, one can shorten the interparticle distance and/or pore size and increase the volume fraction of protein. However, care is required to design assemblies with short linkers because if the linkers are too short then multiple sites of proteins may be involved in protein–protein interactions which could lead to disordered structures and to compact structures destroying the functionality of the enzymes. Here we address questions regarding the physical properties required for the linkers with the purpose of assembling porous $3D$ crystalline structures with various pore sizes using a few functional groups on the pro-

[a] Department of Material Science and Engineering, Northwestern University, Evanston, Illinois 60208, United States
[b] Department of Chemistry, Northwestern University, Evanston, USA
[c] Department of Physics and Astronomy, Northwestern University, Evanston, USA
* m-olvera@northwestern.edu

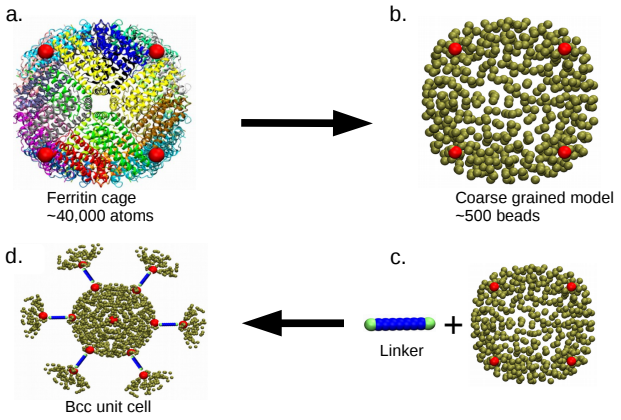

Fig. 1 Detailed structure of a ferritin protein, coarse–grained models of a protein and a linker, and an assembly of proteins in a bcc unit cell. a) Front view of a ferritin cage (PDB code: 5CMQ) with anchored metal ions. There are 8 metal ions anchored sites per protein which are represented by small red spheres separated by 72Å. In experiments[2,4], metal ions are anchored on protein surfaces after replacing Threonine amino acid at position 122 with Histidine. b) A coarse grained model of a ferritin cage which is composed of 500 beads. c) A system of coarse–grained linker and protein. The linker has identical beads at two ends (colored in green). The linear density of beads is approximated by $\frac{L}{5}$Å^{-1}. d) Proteins in a bcc unit cell held by linkers through the metallic ions installed sites.

tein surface to preserve their enzymatic activity.

We introduce a coarse–grained molecular dynamic (MD) simulation approach to study the effects of length and geometry of the linkers on the $3D$ crystalline assemblies of the ferritin protein, and the effect of linker flexibility and linker to protein concentration ratio on the symmetry and porosity of the lattices. We employ the Lennard–Jones (LJ) potential to account for the specific interactions between two ends of linkers and metal ion anchored sites on the protein surface and the Weeks-Chandler-Andersen (WCA) potential for excluded volume interactions. By simulating a system of proteins and linkers, we find a range of optimal linker lengths for crystalline arrays. The optimal linker length depends on the linker to protein concentration ratio and binding energy. Our analysis suggests that the emergence of an extreme optimal length is rooted at the expense of rotational degrees of freedom of the free linkers. We also investigate the effect of the length of grafted linkers on the formation of the protein arrays. Contrary to the free linkers case, we do not find an extreme optimal length in the linker grafted case suggesting that the grafting of linkers on the protein surface is a better route to yield rich porosity crystalline structures. The computationally inexpensive method that we introduce in this study is potentially useful to design open crystalline structures of proteins and serve as a guide for constructing higher order assemblies of complex molecules.

Model

We introduce a coarse–grained MD simulation approach to understand the underlying mechanisms behind the linker–directed–protein–self–assembly ($LDPSA$) process. We select ferritin as the model protein because this protein has been widely studied[2,4,18,22,23] and the experimental data are readily available for comparison with simulation results[4]. In experiments, Bailey et al. replaced the Threonine amino acid 122 located at the $c3$ symmetric interface of ferritin with Histidine to load essentially 8 metal ions from the transition II group. Then, by using the organic linkers having two functional head groups, they construct bcc and/or bct crystals. Ferritin is a 24–mers protein, nearly a spherical cage possessing octahedral (432) symmetry (shown in Fig.(1a). Its external diameter is approximately 120Å. The ferritin cage has a hollow interior of size 80Å and this cavity has been utilized as a container for nanoparticles and enzymes[18,22]. There are around $40,000$ atoms per ferritin cage which hints that the use of full atom simulation approaches to study the protein superlattices could be computationally expensive. To reduce the computational expense, we coarse grain the ferritin cage into a few hundred beads. In this coarse–grained model, the position of the metal ion anchored sites are kept same as in the detailed structure of protein. In doing so, the distribution of metal ions on the protein surface is accurate and the reduced number of beads is sufficient enough to mimic the shape, protein surface roughness and average internal and external diameters of the protein (shown in Fig.(1b)). At this level of coarse–graining, our simulation box consists of $50,000$ beads on average.

The ferritin model used in this study is built based on pdb codes 5CMQ, 5UP7 and 5VTD in the protein data bank. There could be multiple metal binding pockets in a ferritin. However, we consider only 8 linker binding sites per protein that represent the anchored metal ions at the $c3$ symmetric interfaces of ferritin (equivalent to the eight vertices of a cube). These 8 metal ions are located at the centers of RESID 122 in the ferritin structure's files. We consider only 8 metal sites per protein and the resulting lattices to be bcc/bct types for the following reasons. In references ([2,4]), Bailey et al. report only bcc/bct types of linker mediated protein lattices. Therefore, we assume that the metal ions other than the anchored ones are either insufficiently stable to hold proteins in the ordered structures via linkers or they are inaccessible to linker binding. Secondly, we are not aware of any published experimental works showing a range of crystalline structures of ferritin via free linkers apart from references ([2,4]). It has been shown that an appropriately engineered ferritin can be crystallized to form fcc lattice via metal coordination[13]. However, in this study we are interested in the linker mediated metal-protein hybrid assembly (open crystalline structures). We expect that if metal ions are anchored at other symmetric interfaces (c2 and c4) of ferritin protein then that may open paths to construct a range of open lattices via linkers. We have not explored this possibility here.

The experimentally determined ferritin–MOFs assemblies show that there are no inter–protein contacts (surface residues of proteins are not closer than 5Å) in the crystalline structures. In the assembly product, proteins are solely held together by the linkers via the metal ions that are anchored on the protein surface fully replacing protein-protein binding. In our model, we ignore the electrostatic interactions.

In this implicit solvent coarse grained model, we represent each metal ions anchored protein and linker by two types of beads. In the protein, one type of bead represents the metal ions anchored

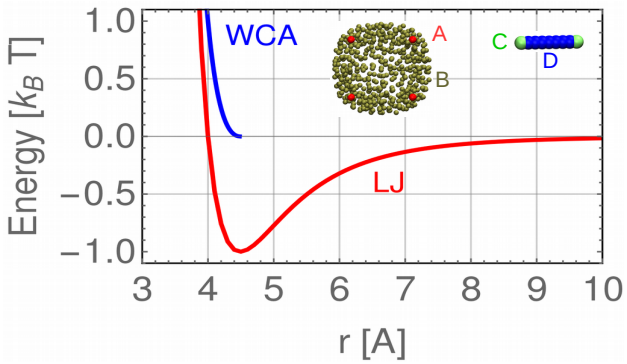

Fig. 2 Interaction potentials (Eqs. 1 (red curve) and 2 (blue curve)) between beads are plotted as a function of the distance. The value of the 6-12 LJ potential becomes smaller with the increase in separation between particles so we truncate the potential at 2.5σ ($r_{\text{cutoff}} = 2.5\sigma$). The repulsive interactions between particles are modeled using the WCA potential by shifting the LJ interaction energy by ε and shortening the cutoff distance to $2^{1/6}\sigma$ ($r_{\text{cutoff}} = 2^{1/6}\sigma$). In this plot, we chose $\varepsilon = 1k_BT$ and $\sigma = 4$Å. Except the interaction between A and C types of beads, all pair interactions are modeled by the WCA potential. The interaction between A and C types of beads is modeled by the 6-12 LJ potential.

sites (type A in Fig.2–red spheres) and the second type of bead represents the amino acids (type B in Fig.2). In the linker, the two terminals are represented by one type of bead (type C in Fig.2) and the remaining parts of the linker are denoted by another type of bead (type D in Fig.2). We employ a Lennard–Jones (12–6)[24] and/or a WCA potentials[25] to account for the pairwise interactions between beads, which are given by equations (1) and (2), respectively.

$$U_{LJ}(r) = \begin{cases} 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right], & \text{if } r \le 2.5\sigma. \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

$$U_{WCA}(r) = \begin{cases} 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] + \varepsilon, & \text{if } r \le 2^{1/6}\sigma. \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

In equations (1) and (2), ε is the interaction strength, σ is a distance between two particles when the potential between them is 0 and r is the center to center distance between particles. The interactions between metal ions on the protein surface (type A in Fig. (2)) and terminals of linkers (type C in Fig.(2)) are modeled by using the LJ potential and all other pairwise interactions (metal–metal, metal–amino acid, amino acid –amino acid, linker–amino acid, linker–linker) are accounted for by using the WCA potential.

Simulation details

We performed MD simulations using a HOOMD–blue[26] simulation toolkit. To begin the simulation, we randomly distributed the metal anchored proteins and linkers in the 3D cubic box. We introduced periodic boundary conditions in the x, y and z dimensions. Particles (proteins and linkers) were allowed to freely diffuse in the box obeying the forces mentioned in the model section. The particles have translational and rotational degrees of freedoms. However, in our model, we ignore the conformational changes of both proteins and linkers (unless linkers are grafted to proteins) by treating them as rigid bodies. Although we observed the formation of clusters of proteins resembling unit cells of bcc and/or bct crystals for simulations starting from disordered, our simulations failed to form long-ranged crystalline structures in simulation time. This is because the simulation times are extremely short in comparison to the experiment. In experiments, this may take anywhere from hours to days to grow micro/millimeter sized crystals given that the parameters are right.

Thus, to reduce sampling time, instead of distributing proteins and linkers randomly in the box, we place proteins in the crystalline lattice and leave only the linkers in the protein–free random positions. To construct a protein lattice, the lattice parameters corresponding to a linker are picked from the experimental results[4]. After placing the proteins in the crystalline structure, we measure the closest distance between the interprotein binding site pairs using the visualization tool "VMD"[27] and we find that this distance is equivalent to the length of the linkers that has yielded the crystal structure. The relationship between the lattice parameter of the bcc lattice and the linker length is $a = \frac{2}{\sqrt{3}}(2R_p + L)$, where R_p and L are the protein radius and length of the linker, respectively. If the experimentally measured lattice parameters are unavailable, then we place proteins in a lattice setting the distance between the centers of the body-centered and the corner proteins to be equal to the sum of the protein diameter $2R_p$ and the end to end distance of linker L. In doing so, the shortest distance between the interprotein binding sites is equivalent to the length of the linker which means the length of a linear linker will be sufficiently long enough to be able to link the proteins via metal anchored nodes.

After the simulation is set up, we integrate the system by utilizing the NVE integrator in HOOMD to avoid possible overlaps between particles. Then, by keeping proteins at their initial positions, we allow only linkers to diffuse freely in the box obeying Lennard–Jones and WCA force fields using the Langevin dynamics. In this step, linkers compete to find binding spots on the protein surface and some of them may link proteins via the closest interprotein metal ion anchored site pairs. We then allow both linkers and proteins to diffuse freely for the rest of the simulation time. If the system of proteins and linkers that link proteins move coherently in the box then the initial crystalline structure of the proteins remains stable and this happens only if the right parameters are implemented in the simulations. In the HOOMD simulation toolkit, there are three fundamental units: distance (D), mass (M) and energy (ε). In our simulation, we set $D = 1$ Å, $M = 1$ *amu* and $\varepsilon = 1\ k_BT$. We use 0.1 as a simulation time step which corresponds to $0.1\tau_0$ in real units, where $\tau_0 = \sqrt{\frac{MD^2}{\varepsilon}}$ is the time unit which is 50 fs. Therefore, the real time step is 5 fs, which is usually small in a coarse-grained model. We made this choice to make the simulations stable. The real time unit is expected to be larger than 5 fs due to the use of the coarse-grained model.

Results and discussion

Effects of linker to protein concentration ratio

The specific interactions between ditopic linkers and the metal ion anchored sites on the protein surfaces drive the assembly in *LDPSA*. Therefore, the linker to protein ratio (f) is a key parameter to determine the types of protein aggregates. Concentrations of linker, both too high and too low, are unfavorable for driving the ordered assembly of proteins. For example, if both ends of a linker find the binding sites (one end of the linker binds to a site of one protein and other end binds to a site of another protein) then ordered assemblies are favorable, whereas if only one end of the ditopic linker finds the binding site (unproductive binding) then disordered structures may result. Generally, if the ratio (f) is less than the number of binding sites per protein (metal ions anchored sites on the protein surface) then it is highly probable that ditopic linkers find binding spots for both of their ends. On the other hand, when f is greater than the number of binding sites, the probability of finding the binding spots for both of its ends is low because of the saturation of the binding spots.

We explore the binding modes of the linkers and the resulting protein aggregates by varying the ratio f from 2 to 20. Since there are only 8 linker binding sites per protein in this model (equivalent to the number of anchored metal ions per protein), the range of the ratio f studied here covers both low and high linker concentration regimes in comparison to the number of binding sites per protein. To connect $2x^3$ number of proteins in a bcc and/or bct lattice by linkers, $(2x^3 \times 4)$ number of linkers are required, where $x = 1, 2, 3...$ denotes the unit cell multiplicity along 3 mutually perpendicular directions. The numerical factor 4 represents the number of linkers required per protein to form a bcc and/or bct crystal via 8 coordination sites. In our study, we chose $x = 3$, which means there are 54 proteins and, when they are arranged in a bcc crystal with a lattice parameter $(a = \frac{2}{\sqrt{3}}(2R_p + L))$, then there exists 216 binding site pairs separated by a distance L equivalent to the length of a linker.

In principle, when $f < 4$, then the number of linkers are insufficient in number to link proteins through 8 coordination sites and this leaves some binding sites on the protein surfaces empty. In this study, when the ratio is $f = 2$ we find about 50% of the binding sites pairs empty (110 out of 216). Though the value of f is reasonably lower than the number of binding sites per protein, we observe unproductive bindings at two binding site pairs. As expected, at this low linker concentration ($f = 2$), we obtain a disordered assembly which comes from two sources– one is the finite probability of the linkers to occupy the binding sites pairs in an unproductive way and the second source to produce a disordered aggregate is the presence of some empty sites (shortage of linkers). At low linker concentrations, the presence of many empty binding sites is a major factor to produce disordered assemblies. Then we increase the value of ratio (f) to 4, which is the case that the number of linkers should ideally be sufficient to link proteins in bcc and/or bct crystals. In this case, we observe about 10% non–linked binding sites pairs. However, there are still nearly 90% linked binding site pairs which seems adequate to stabilize a protein superlattice.

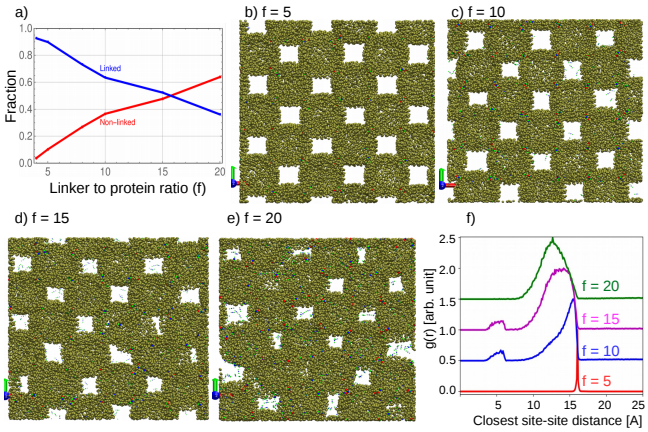

Fig. 3 The effects of linker to protein ratio (f) on linked and non–linked fractions and on the quality of the protein lattices. a) Fractions of linked and non-linked binding sites pairs as a function of linker to protein ratio f. As the value of f is increased, the fraction of non–linked binding sites pairs overtakes the fraction of linked pairs which is an adverse condition for yielding the stable protein superlattices. The front views of bcc crystals at different linker to protein ratios are shown : b) f =5 c) f = 10, d) f = 15, and e) f = 20. The snapshots show that the crystalline structures of proteins deteriorate as the ratio of the linker to protein increases. f) The normalized time averaged closest interprotein metal anchored sites distance is plotted by varying f. For the comparison purposes, distributions corresponding to $f = 10$, 15 and 20 are shifted vertically. The binding energy and the length of linkers are fixed at ($E_{bind} = 34$) and $L = 10$Å, respectively.

When we further increase the ratio (f) we observe more non–linked binding sites pairs which directly affects the stability of protein lattices. Contrary to the low linker concentrations, the dominant cause of more non–linked binding sites pairs is the binding of linkers in an unproductive way. The effects of ratio f on protein assemblies, linked and non–linked fractions are shown in Fig. (3). The binding probability of linkers, the linked and/or non–linked fractions of binding sites pair also depend on the energy difference between the bound and unbound states of the linkers. In general, if the binding energy between the linker and the specific sites on the protein surface is very strong, it may lead to disordered protein assemblies because of the irreversible linker–protein unproductive binding. On the other hand, weak linker–protein binding energy is insufficient to hold proteins in the crystalline structures. Therefore, the tuning of binding energy is required to find a value for which crystalline assemblies are stable. The non–linked fraction of binding sites varies non–monotonically with the binding energy of linkers. At the lower binding energy regime, the non–linked fraction (due to weak energy, there are more empty sites) decreases faster with the increase in energy. Then, after attaining a minimum value it slowly increases (unproductive binding) with the energy. The energy value corresponding to the minima of the non–linked fraction depends on the linker to protein ratio and the length of the linkers. In this study, the binding energy and the length of linkers are fixed at $34k_BT$ and 10Å respectively.

To further explore the ratio (f) effects on the stability of protein lattices, instead of distributing linkers randomly in the simulation box, we place $(2x^3 \times 4)$ number of linkers at the ideal positions

and the remaining linkers ($2x^3(f-4)$) are distributed randomly in the box. Here, the ideal positions of linkers mean that they are initially linking proteins via metal ions anchored sites. When the ratio is set to $f = 4$, there are no linkers in the solution that compete for the binding sites. At this value of f, linkers do not unbind from the binding sites and result in the stable protein lattices. However, the competition to get the binding sites increases if the number of linkers in the box is increased. When we increase the value of f from 4 to larger values, we observe that some initially bound linkers sacrifice one of their binding spots to another linker. The population of linkers losing one of their binding sites increases with the value of f which means the population of the unproductively bound linkers increases at the expense of the productively bound linkers. Note that in the productive binding case, a binding site pair is connected by a single linker whereas in the unproductive binding, the same binding site pair is occupied by two linkers. For either type of binding, the total energy gain per binding site pair is equal, however, the unproductive binding is entropically beneficial for the free linkers in the solution. As a result of this, more unproductive binding of linkers are observed as the value of f is increased.

The observation of the concentration dependent unbinding of linkers in this study is similar to the concentration dependent off rate study of DNA–binding protein by Erbas et al.[28,29]. In our study, however, we do not observe any empty binding sites when $f > 4$ at the end of the simulations. This may be due to the exchanges between the linkers in the binding site pairs and the lifetime of the empty sites may be shorter than the time taken by other linkers to bind. Moreover, we find that the initial linker distribution does not influence the type of output aggregate. Instead the linker concentration affects the types of protein assembly. To avoid the disordered outputs based on linker concentration, we select the value of f in such a way that the number of linkers are neither too low to leave the binding sites empty nor too high to occupy the sites unproductively. We select the linker to protein ratio f to be comparable to the number of binding sites per protein ($f = 8$) and study the stability of crystal structures predicted experimentally by Bailey et al. for different geometries and lengths of linkers. The comparisons of lattice constants obtained in simulation and experimental studies are shown in Fig.(4).

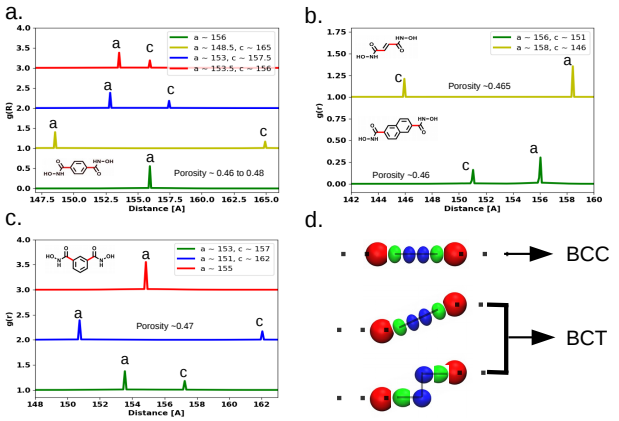

Fig. 4 Comparison of lattice constants obtained in experiments and simulations. In simulations, the time averaged distributions of interprotein centers with respect to the distance are measured. In the plots, only peaks corresponding to the lattice constants are shown. Lattice constants (a and c in legends) of bcc and/or bct unit cells yielded by different ditopic linkers: a) benzene–1,4–dihydroxamic acid. Ditopic ends of this linker are collinear. The linkers of this geometry yield either a bcc lattice or a bct lattice depending on how they are bound to the sites on the protein surfaces. The lattice constants of bct lattices are shifted vertically for the comparison purpose. b) E–ethylenedihydroxamic acid (upper part), and nepthalene–2,6–dihydroxamic acid (lower part). The hydroxamic acid head groups of both of these linkers are offset by some factors and both of them result in bct lattice. c) benzene–1,3–dihydroxamic acid. The Linkers used for assembling proteins are shown inside the frame of figures. The porosities of the crystalline structures corresponding to the linker are also shown inside the frame. Physically, linkers differ from each other in terms of the interhydroxamate spacing and their orientations. The fluctuations of the lattice constants are less than 2Å and these are caused by the lack of the directional interactions in the model. d) The binding mode of linkers to the metal ions and the resulting crystal structures. Red spheres represent a closest interprotein binding site pair. If a line joining a pair of binding site aligns with a line connecting the centers of a body centered protein and a corner protein then a linear linker yields bcc crystal. Otherwise, both linear and non–linear types of linkers yield a bct lattice.

Effects of linker length

In the linker directed protein superlattices, the interparticle distances, the pore sizes, and the porosity can be controlled treating the length of linker as a tuning parameter. However, the longer linker could be a hindrance for the construction of highly porous protein superlattices. For an example, the linkers with higher aspect ratio show Onsager transition[30] preferring to align towards the same direction. In the following paragraphs, we discuss how the length of the ditopic linker affects the stability of protein superlattices. To study the linker length effects on the porosity of the crystalline protein assemblies, we assume that there are 8 specific linker binding sites per protein (metal ions anchored sites) and 2 head groups per linker as before. The linker length effects on the porosity of the lattices could be a very interesting future

work in case additional metal ions are made stable and accessible to linker binding on the protein surface.

To explore the effects of linker length on linked and/or non–linked fractions of binding sites pairs, we perform simulations at various lengths of linker (from 10Å to 40Å). In these simulations, we keep the linker to protein ratio (f) and linker–protein binding energy E_b fixed at 5 and $34k_BT$, respectively. The number of beads in a linker is $(\frac{L}{5}+1)$, where L is the length of a linker. For example, linkers of length 10Å, 20Å, and 30Å are composed of 3, 5, and 7 beads, respectively, with an interbead separation of 5Å. Note that the number of binding sites per protein and their separation are independent of the length of the linkers, while the distance between interprotein binding sites in a lattice is linearly dependent on the linker length. At the end of the simulations, out of 216 binding sites pairs, we observe 22, 35, and 43 non–linked pairs for linkers of length 10Å, 20Å, and 30Å, respectively. The general trend here is that the fraction of non–linked binding site pairs increases with the length of the linker. To determine if the fraction of linked and/or non–linked binding site pairs is dependent on the number of beads or not, simulations at the above mentioned parameters are run keeping the number of beads constant at 3. We increase the interbead spacing to increase the linker length. For the linker lengths 10Å, 20Å, and 30Å, the number of non–linked pairs are 22, 49, and 60, respectively.

The non–linked fraction of binding sites pairs with respect to the length of the linker and the number of beads are given in Table(1) which shows that the non-linked fraction increases with the length of linker no matter how the length is increased either by adding the beads or by increasing the interbead distances. In the table, there are also non–linked fraction mismatches between the same length of linker but made with different number of beads. The magnitude of the non-linked fraction is higher when fewer beads are used to construct a linker in comparison to using more beads. The mismatches in non–linked fractions come from the dependence of linker binding probability on the number of beads. The probability of a bead finding a binding site decreases with the length of a linker and it further declines if a linker is made with more beads. The decline in the binding probability provides more chances for already bound linkers to find binding spots for their other end which, in return, decreases the non–linked fraction.

Table 1 Variation of non–linked fraction of binding site pairs with respect to the length of the linkers. The non–linked fraction values shown in parentheses correspond to the linkers made by 3 beads.

Linker length (Å)	$N_{beads} = (L/5+1)$ or 3	Non–linked fraction
10	3	0.10
20	5	0.16 (0.23)
30	7	0.20 (0.28)

The consequences of the linker length increments on the crystalline structures of proteins are shown in Fig.(5). Fig.(5a) shows distributions of the closest interprotein binding sites pairs with distances for different lengths of linker. The sharp peak values correspond to the initial separation between the closest interprotein binding site pairs. For comparison purpose, the initial peak positions for different length of linkers are horizontally shifted to a common point and distributions corresponding to the linkers of length 20Å, 30Å, and 40Å are shifted vertically. When the length of the linker is 10Å, the distribution is symmetric around an initial peak. However, as the length of the linker is increased, the distributions become asymmetric (skewed left) and largely deviate from the initial positions. Transition from crystalline order of proteins to global disorder is enhanced with the increase in linker length as shown in Fig.(5b,c,d) for linker lengths 20Å, 30Å, and 40Å, respectively. A stable protein crystalline structure resulting from linkers of 10Å length is shown in Fig.(5f). We set the binding energy between the ditopic ends of linkers and binding sites on the protein surface to $34k_BT$ in figures (a, b, c, d, f) in Fig.(5).

Since the lattice constant (a) increases with the linker–length (L), the available volume ($V_{\rm box} - n_{\rm p}V_{\rm p}$) in a simulation box increases with L. Where $V_{\rm p}$ and $n_{\rm p}$ are the volume of a protein and number of proteins in a box, respectively. Increasing the value of the binding energy between linker and protein sites from $E_b = 34$ to 42 and 48 for the linkers of lengths 20Å and 30Å, respectively, we get symmetric distributions of the binding sites pairs distances (Fig.5e). Stable crystals, as shown in Fig.(5g,h), are observed for linker lengths less than 40Å. The variation of energy required to obtain stable crystals with respect to the length of linker is shown in Fig.(6c).

When linker length is smaller than 40Å, stable crystals are obtained by adjusting the binding energy between linker and metal site on the protein. For linkers longer than 40Å, output assemblies are always disordered. We test a wide range of binding energies (from $E_b = 10$ to 100) to analyze the binding modes of linkers and the output assemblies. At energy values $E_b = 10$ and lower, we observe many empty binding sites pairs, which means that the selected energies are smaller than the threshold binding energy. Above the threshold energy, we increase the binding energy in the interval of $\Delta E_b = 3$, but none of them result in a stable protein lattice. By analyzing disordered outputs, we find that many proteins within the aggregate are bridged by linkers via two binding sites pairs, which are unfavorable for bcc and/or bct crystals.

To explain the inability of longer linkers to stabilize protein lattices, we evaluate the microstates corresponding to rotational degrees of freedom of linkers. For the rotational freedom, a linker of length L requires $(2\times\frac{4}{3}\pi L^3 - \frac{5}{12}\pi L^3)$ amount of volume. The first term, $\frac{4}{3}\pi L^3$ is the volume of a sphere having a radius equivalent to the length of a linker L. The volume of a sphere is multiplied by 2 to account for rotation around either ends of a linker. When two spheres are drawn, there is a $3D$ lens shaped region shared by both spheres. To correct this, we subtract the second term, $(\frac{5}{12}\pi L^3)$. The volume of a $3D$ lens formed by the intersection of two spheres of equal radii R located their centers d distance apart is given by equation (3) [31].

$$V_{\rm lens} = \frac{1}{12}\pi(4R+d)(2R-d)^2. \qquad (3)$$

In our case, $R = L$ and $d = L$, which means $V_{\rm lens} = \frac{5}{12}\pi L^3$. The net

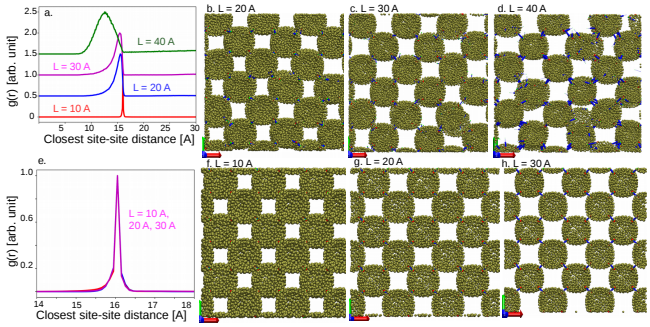

Fig. 5 Distributions of interprotein metal anchored sites distances and the front views of protein superlattices at different linker lengths. a) Normalized pair distributions of the closest interprotein metal anchored sites distances for linker lengths $L = 10$Å, 20Å, 30Å, and 40Å using the same binding energy ($E_b = 34$). The distributions become more asymmetric as the linker length increases. For comparison purposes, initial lattice positions are shifted to a same point and distributions corresponding to lengths 20Å, 30Å and 40Å are shifted vertically. b,c,d) Protein assemblies corresponding to linkers of lengths $L = 20$Å, 30Å and 40Å, respectively. The quality of structure declines with the linker length. Furthermore, in the each figure, the pore size is not uniform which means that the proteins are not in perfect lattice positions. e) Normalized pair distributions of the closest interprotein metal anchored sites distances for linker lengths $L = 10$Å, 20Å, and 30Å using different binding energies. The distributions are symmetric. To get the periodic structures of proteins for the linkers of lengths 20Å and 30Å, we use binding energies $E_b = 42$ and 48, respectively. f,g,h) Periodic arrays of protein yielded by linkers of lengths $L = 10$Å, 20Å, and 30Å, respectively. Pores are identical in each case.

volume required for a linker for its rotational freedom is,

$$\begin{aligned} V_{\rm net} &= 2\times\frac{4}{3}\pi L^3 - \frac{5}{12}\pi L^3, \\ &= \frac{9}{4}\pi L^3. \end{aligned} \quad (4)$$

For shorter linkers, the volume required for the rotational freedom ($V_{\rm net}$) does not have a noticeable effect. However, as the length of linker increases, $V_{\rm net}$ becomes large. The number of microstates associated with rotational degrees freedom of linkers are evaluated by dividing the free volume of a box ($V_{\rm box} - n_{\rm p}V_{\rm p}$) by $V_{\rm net}$. This ratio decreases faster with the linker length, as shown in Fig.(6b). This means that longer linkers lose their rotational freedom more than shorter linkers. To reduce the rotational freedom loss in the solution, longer linkers prefer linking proteins via two binding sites pairs forming a non–crystalline structure. We approximate an extreme optimal linker length using,

$$\begin{aligned} V_{\rm void} &= 0, \\ V_{\rm box} - n_{\rm p}V_{\rm p} - n_{\rm l}V_{\rm net} &= 0, \\ \left(\frac{2}{\sqrt{3}}(2R_p + L)\right)^3 - n_{\rm p}(\frac{4}{3}\pi R_p^3) - f n_{\rm p}(\frac{9}{4}\pi L^3) &= 0. \end{aligned} \quad (5)$$

Equation (5) is obtained after replacing $V_{\rm box}$, $V_{\rm p}$, and total number of linker ($n_{\rm l}$) by $\left(\frac{2}{\sqrt{3}}(2R_p + L)\right)^3$, $\frac{4}{3}\pi R_p^3$, and $f n_{\rm p}$, respectively. For a ferritin protein, $R_p \sim 60$Å. In a bcc unit cell, $n_{\rm p} = 2$, if we choose $f = 5$, then an extreme optimal linker length is ~ 40Å, which is approximately 65% of R_p.

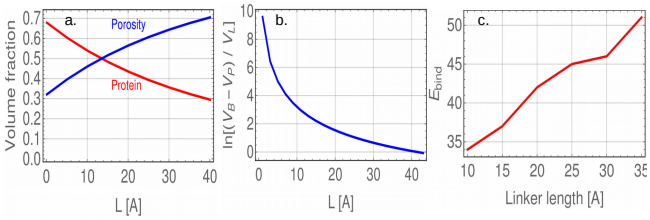

Fig. 6 a) Volume fraction of protein and/or porosity of $3D$ bcc protein superlattices with respect to the linker length. Since the interprotein distance increases with the linker length in bcc crystals, the volume fraction of porosity (and/or protein) increases (decreases). Volume fraction of protein is ($v_f = \frac{2V_{\rm protein}}{a^3}$), where $a = \frac{2}{\sqrt{3}}(2R_p + L)$ is a lattice parameter. b) Microstates of linkers corresponding to the rotational degree of freedom vs linker length. This declines with the linker length, which means that longer linkers lose their rotational freedom more than the shorter linkers do. c) The binding energy required for maintaining crystalline structures for different lengths of linkers.

Using free linkers, we find that the pore size and porosity of protein superlattices could not be increased once an optimal linker length reaches an extreme value. To find out whether or not the grafted linkers yield protein superlattices of wider pores, we grafted eight linkers per protein. To be consistent with the free linker case, the locations of the grafting points are chosen to be the same spots as the metal ion anchored sites in a ferritin protein. Using this way of grafting, we expect the resulting lattice to be bcc and/or bct. Each grafted linker is composed of 20 beads, 19 bonds, and 18 angles. The interbead equilibrium distance is set to be $r_0 = 3.0$Å. If the free ends of grafted linkers are $L/2$ away from their corresponding grafting points at equilibrium then, we expect, these could produce the same sized pores and equal porosity as produced by free linkers of length L.

In this type of linker grafted case, since we have no available lattice parameters to place the proteins into a lattice, we perform simulations choosing various initial lattice parameters. At the end of the simulations, we analyze fluctuations in each of the initial lattice constants. The initial lattice constant which fluctuates minimally is considered a real lattice constant of the crystal. Free ends of the interprotein grafted linkers interact via the Lennard–Jones potential. The other possible non–bonded interactions are accounted for using the WCA potential. Harmonic potentials are used to account for the bond stretching and bending interactions in grafted linkers.

The fluctuations of initially chosen lattice constants are shown in Fig.(7). In the figure, there is a minimal lattice constant fluctuation when a parameter L is 40Å. The lattice constant corresponding to $L = 40$Å is found to be ~ 190Å. We also calculate the volume fraction of the protein and the porosity of lattice, which are ~ 0.30 and ~ 0.70, respectively. The value of porosity is already higher than the maximum porosity that can be achieved using the free linkers. Furthermore, in the grafted linkers case, we do not find an extreme optimal length as in the free linker's case. These findings demonstrate that the extreme optimal linker length is emerged from the loss of the rotational freedom of the free linkers.

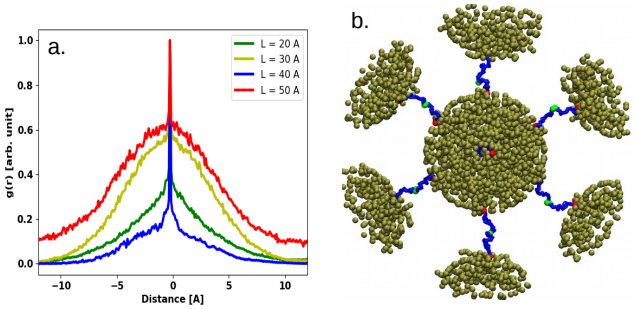

Fig. 7 Plots showing the fluctuations of initially chosen lattice constants and proteins in a bcc unit cell. a) Lattice constants fluctuations. We use an expression $(a = (\frac{2}{\sqrt{3}}(2R_p + L)))$ to determine initial lattice constants by varying the value of L to place proteins in a bcc lattice. Where $R_p \sim 60$Å is radius of ferritin. For comparison purposes, initial lattice constants are shifted to the same point. In the plot, there is a minimal fluctuation when $L = 40$Å. The lattice constant corresponding to $L = 40$Å is $a \sim 190$Å. Here, each grafted linkers are composed of 20 beads and the interbead equilibrium distance in the harmonic interaction is $r_0 = 3.0$Å. b) A bcc unit cell of proteins when lattice constant is $a \sim 190$Å. There are 8 grafted linkers (shown in blue colors) per protein and the grafting locations are shown in red colors. Free ends of interprotein grafted linkers interact via complementary interactions and they are represented by green color.

Conclusions

In this study we introduce a coarse–grained molecular dynamic simulation approach for studying the effects of linker length, geometry, and concentration on the stability of metal anchored ferritin protein crystallizing assembly. This model assumes that the anchored metal ions are sufficiently stable on the protein surface for holding proteins in the crystalline structures via linkers. Here, only anchored metal ions at the $c3$ symmetric interfaces of ferritin are considered. The linked and/or non–linked fraction of binding sites pairs as a function of the linker to protein concentration ratio is studied keeping linker length and interaction strength of pair potentials constant. A high (low) value of linked (non–linked) fraction is favorable for yielding crystalline protein assembly. In order to obtain stable crystallizing protein assembly, the linker to protein concentration ratio should be comparable to the total number of binding sites per protein. At low linker concentration, the protein assembly is disordered because many interprotein binding sites pairs remain empty due to the shortage of linkers. Similarly, if the ratio is too high in comparison to the number of binding sites per protein, the protein assembly is disordered because many binding sites pairs are non–linked by linkers due to the over population of linkers. By setting the linker to protein ratio constant at $f = 8$, equivalent to the number of metal anchored sites per protein, we find that bcc and/or bct crystals are obtained in agreement with the experiments[4] for linkers of various geometries (linear or non–linear) and lengths (from 9Å to 13Å).

The linker length effects on $3D$ protein superlattices are analyzed. The quality of bcc/bct protein superlattices deteriorates as the length of the linker is increased. The quality of structures is found to be regained if the binding energy between linker and metal ion site on the protein surface is adjusted. However, the adjustment of binding energy works only up to an extreme optimal linker length. Linkers having length longer than an extreme optimal length fail to hold proteins in crystalline structures. We have evaluated the number of microstates based on the rotational freedom of linkers that suggests that linkers lose their rotational freedom once their length exceeds an extreme optimal linker length. As a result, more longer linkers occupy binding sites in an unproductive way, which enforce disordered aggregates. Our work shows that the optimal linker length is dependent on the linker to protein concentration ratio and binding energy.

Here, we predict that the length of free linkers can be exploited to build crystalline protein assemblies up to an extreme optimal length. An option for yielding highly porous crystalline protein assemblies is analyzed by studying the effect of linker length when they are grafted on the protein surfaces. In this case, an extreme linker length is not found, which suggests that the grafting of linkers could be a better route to obtain crystalline protein assemblies having wider sized pores and/or high porosity. The computational method developed to study porous crystals of proteins provide the guidelines to design crystalline open assemblies.

Conflicts of interest

"There are no conflicts to declare".

Acknowledgements

Y.R.D. acknowledges Jaime Millan and Martin Girard for useful discussions. This work was supported by the Department of Energy, Basic Energy Science grant number DE-FG02-08ER46539, and by the Sherman Fairchild Foundation.

References

1 J. D. Brodin, E. Auyeung and C. A. Mirkin, *Proc. Natl. Acad. Sci. U.S.A.*, 2015, **112**, 4564–4569.
2 P. A. Sontz, J. B. Bailey, S. Ahn and F. A. Tezcan, *J. Am. Chem. Soc.*, 2015, **137**, 11598–11601.
3 J. R. Mcmillan, J. D. Brodin, J. A. Millan, B. Lee, M. O. D. L. Cruz and C. A. Mirkin, *J. Am. Chem. Soc.*, 2017, **139**, 1754–1757.
4 J. B. Bailey, L. Zhang, J. A. Chiong, S. Ahn and F. A. Tezcan, *J. Am. Chem. Soc.*, 2017, **139**, 8160–8166.
5 B. Panganiban, B. Qiao, T. Jiang, C. Delre, M. M. Obadia, T. D. Nguyen, A. A. A. Smith, A. Hall, I. Sit, M. G. Crosby and et al., *Science*, 2018, **359**, 1239–1243.
6 T. D. Nguyen, B. Qiao and M. O. D. L. Cruz, *Proc. Natl. Acad. Sci. U.S.A.*, 2018, **115**, 6578–6583.
7 A. Huang and B. D. Olsen, *Macromol. Rapid Commun.*, 2016, **37**, 1268–1274.
8 A. L. Margolin and M. A. Navia, *Angew. Chem. Int. Ed.*, 2001, **40**, 2204–2222.
9 T. Ueno, *Chem. Eur. J.*, 2013, **19**, 9096–9102.
10 Hofmeister, F., *Arch. Exp. Pathol. Pharmakol.*, 1888, **24**, 247–260.
11 A. A. Green, *J. Biol. Chem.*, 1931, **93**, 517–542.
12 Tanford, C., *Physical Chemistry of Macromolecules [Hardcover].*, 1966 *John Wiley & Sons, Inc.*.

13 D. M. Lawson, P. J. Artymiuk, S. J. Yewdall, J. M. A. Smith, J. C. Livingstone, A. Treffry, A. Luzzago, S. Levi, P. Arosio, G. Cesareni and et al., *Nature*, 1991, **349**, 541–544.

14 C. J. Lanci, C. M. MacDermaid, S.-g. Kang, R. Acharya, B. North, X. Yang, X. J. Qiu, W. F. DeGrado and J. G. Saven, *Proc. Natl. Acad. Sci. U.S.A.*, 2012, **109**, 7304–7309.

15 N. P. King, J. B. Bale, W. Sheffler, D. E. Mcnamara, S. Gonen, T. Gonen, T. O. Yeates and D. Baker, *Nature*, 2014, **510**, 103–108.

16 J. E. Padilla, C. Colovos and T. O. Yeates, *Proc. Natl. Acad. Sci. U.S.A.*, 2001, **98**, 2217–2221.

17 Y.-T. Lai, E. Reading, G. L. Hura, K.-L. Tsai, A. Laganowsky, F. J. Asturias, J. A. Tainer, C. V. Robinson and T. O. Yeates, *Nat. Chem.*, 2014, **6**, 1065–1071.

18 M. Uchida, K. Mccoy, M. Fukuto, L. Yang, H. Yoshimura, H. M. Miettinen, B. Lafrance, D. P. Patterson, B. Schwarz, J. A. Karty and et al., *ACS Nano*, 2017, **12**, 942–953.

19 H. Furukawa, K. E. Cordova, M. O'Keeffe and O. M. Yaghi, *Science*, 2013, **341**, year.

20 J. R. Holst, A. Trewin and A. I. Cooper, *Nat. Chem.*, 2010, **2**, 915–920.

21 G. Jutz, P. van Rijn, B. Santos Miranda and A. BÃűker, *Chem. Rev.*, 2015, **115**, 1653–1701.

22 M. Hestericová, T. Heinisch, M. Lenz and T. R. Ward, *Dalton Trans.*, 2018, **47**, 10837–10841.

23 W. M. Aumiller, M. Uchida and T. Douglas, *Chem. Soc. Rev.*, 2018, **47**, 3433–3469.

24 J. E. Jones and S. Chapman, *Proc. Royal Soc. Lond. Series A, Containing Papers of a Mathematical and Physical Character*, 1924, **106**, 463–477.

25 J. D. Weeks, D. Chandler and H. C. Andersen, *J. Chem. Phys.*, 1971, **54**, 5237–5247.

26 J. A. Anderson, C. D. Lorenz and A. Travesset, *J. Comput. Phys.*, 2008, **227**, 5342–5359.

27 W. Humphrey, A. Dalke and K. Schulten, *Journal of Molecular Graphics*, 1996, **14**, 33–38.

28 C. Sing, M. Olvera de la Cruz and J. Marko, *Nucleic Acids Res.*, 2014, **42**, 3783–3791.

29 A. Erbaş, M. O. de la Cruz and J. F. Marko, *Phys. Rev. E*, 2018, **97**, 022405.

30 L. Onsager, *Ann. N. Y. Acad. Sci.*, 1949, **51**, 627–659.

31 *Sphere-Sphere Intersection*, `http://mathworld.wolfram.com/Sphere-SphereIntersection.html`.