Shoemark, D. K., Avila Ibarra, A., Ross, J. F., Beesley, J. L., Bray, H. E. V., Mosayebi, M., Linden, N., Liverpool, T. B., McIntosh-Smith, S. N., Woolfson, D. N., & Sessions, R. B. (2018). The dynamical interplay between a megadalton peptide nanocage and solutes probed by microsecond atomistic MD; Implications for design. *Physical Chemistry Chemical Physics*, 21(1), 137-147.
<https://doi.org/10.1039/c8cp06282j>

Peer reviewed version

Link to published version (if available):
[10.1039/c8cp06282j](https://doi.org/10.1039/c8cp06282j)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via RSC at <https://pubs.rsc.org/en/Content/ArticleLanding/2019/CP/C8CP06282J#!divAbstract>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

The dynamical interplay between a megadalton peptide nanocage and solutes probed by microsecond atomistic MD; implications for design.

Deborah K. Shoemark^{1,2}, Amaury Avila Ibarra¹, James F. Ross³, Joseph L. Beesley³, Harriet E.V. Bray³, Majid Mosayebi^{2,4}, Noah Linden⁴, Tanniemola B. Liverpool^{2,4}, Simon N. McIntosh-Smith⁵, Derek N. Woolfson^{2,3*}, Richard B. Sessions^{1,2*}

¹School of Biochemistry, Biomedical Sciences Building, University Walk, Bristol, BS8 1TD, UK., ²BrisSynBio, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ UK., ³School of Chemistry, Cantock's Close, Bristol, BS8 1TS, UK., ⁴School of Mathematics, University Walk, Bristol, BS8 1TW, UK., ⁵Computer Sciences, Merchant Venturers' Building, Woodland Road, Bristol, BS8 1UB, UK.

R.Sessions@bristol.ac.uk, D.N.Woolfson@bristol.ac.uk

KEYWORDS *Microsecond, atomistic, molecular dynamics simulation, synthetic, megadalton nanoparticle.*

Abstract. Understanding the assembly and dynamics of protein-based supramolecular capsids and cages is of fundamental importance and could lead to applications in synthetic biology and biotechnology. Here we present long and large atomistic molecular dynamics simulations of *de novo* designed self-assembling protein nanocages (SAGEs) in aqueous media. Microsecond simulations, comprised of ≈ 42 million atoms for three pre-formed SAGEs of different charges, in the presence of solutes and solvent have been completed. Here, the dynamics, stability and porosity of the peptide networks are explored along with their interactions with ions, small molecules and macromolecular solutes. All assemblies are stable over the μ s timescale; and the solutes show a mixture of transport behavior across or adherence to the fabric of the SAGE particles. Solute proteins largely retained native-like conformation on contact with SAGE. Certain residues of the SAGE peptides are identified as “repeat offenders” for contacting many different solutes, which suggest modifications to reduce non-specific binding. These studies highlight how molecular dynamics can aid the design process of SAGE and similar assemblies for potential applications as diverse as platforms for drug and vaccine delivery and nanoreactors to encapsulate enzyme pathways.

Introduction

The repurposing of natural protein-based capsids and cages,¹⁻⁴ involving protein engineering and directed evolution of natural protein assemblies is an extensive field.⁵⁻⁸ Several groups now focus on more *de novo* designs for protein-based higher-order cage-like assemblies and their work is perhaps more pertinent to this study. The majority of these, still employ protein engineering of natural proteins to guide the assembly of variants. This includes the pioneering work of Padilla and Yeates,⁹ and the equally important, more recent work from Baker, King and Yeates.¹⁰⁻¹⁴ These assemblies are impressive in terms of scope, accuracy of the designs achieved and reveal new areas for applications.¹⁵ In addition, Marsh and co-workers have employed hybrid/fusion approaches that use *de novo* peptides to assemble natural proteins.¹⁶⁻¹⁷ Taking this further still, Jerala and co-leagues have created cage-like objects through a ‘protein-origami’ approach, which employs largely *de novo* coiled-coil peptides that are concatenated into single polypeptide chains.¹⁸⁻¹⁹

With the aim of designing entirely *de novo* peptide-based cage-like assemblies, we have introduced the Self-Assembled peptide caGES (SAG-Es).²⁰ In these, fully characterized *de novo* peptides²¹ are used as modules to assemble a honey-comb peptide network (Figure 1A), which folds upon itself and closes to form spherical objects. Specifically, the design rationale is to create two self-complementary hubs, one acidic (Hub-A) and the other basic (Hub-B), that interact and co-assemble to form the network only when mixed. Each hub comprises a central trimeric module (CC-Tri),²¹ and each peptide of the trimer has a pendant acidic (CC-Di-A) or basic (CC-Di-B) peptide, which are the two halves of an obligate *de novo* design heterodimer.²² Experimentally, the hub peptides are discrete, water-soluble and partly folded trimers as designed, and when mixed they form ≈ 100 nm spherical objects as visualized by electron microscopy (EM) and atomic force microscopy (AFM). Detailed

AFM studies also show that these objects flatten consistent with lamellar structures, and they reveal hexagonal surfaces features. On this basis, we estimate that SAGE particles each comprise $\approx 10,000$ component peptides. These features have been confirmed through silicification and further AFM imaging,²³ and rationalized by coarse-grained modelling.²⁴ Moreover, and to move towards applications of the SAGEs, we have shown the component peptides can be fused to a variety of natural proteins via the expression of synthetic genes, and that the resulting fusion proteins can be incorporated into SAGE particles. We estimate that the particles can tolerate up to $\sim 20\%$ incorporation of functional proteins through this process of active decoration.²⁵ Here we describe the construction of a model of SAGE particles *in silico*, and subsequent long and large molecular dynamics simulations in aqueous media containing a variety of solutes. In this way, we probe the dynamics and stability of the peptide network, and how this fabric allows the transport of or interaction with the solutes. More specifically, we have compared three simulations for three SAGEs that differ in charge.

Encapsulation in biology -natural capsids and cages. Throughout nature the ability to separate, contain and preserve cargo within defined structures affords countless benefits to organisms from the smallest virus, through bacteria to the largest eukaryotes. The viral capsid packages genomic RNA or DNA, providing protection for transport into and around host.²⁶ The bacterial microcompartment provides a selective advantage to those capable of utilizing substrates that yield toxic intermediates.^{27,28} Enclosing enzyme pathways and containing their reactive species has the added advantage of reducing the diffusion path for the products of one enzyme to become substrate for the next.²⁵

In nature, organisms accomplish encapsulation using a surprisingly limited repertoire of protein building blocks that assemble to form beautiful and often symmetric structures. Exploring the interplay of component subunits and factors affecting stability, perfusion and deformability have long been the subject of investigation. In order to determine at a molecular level, the attributes responsible for “phenotypic” behavior, static images are invaluable. However, studying dynamical properties of large and/or complex systems, provides insights for aspects we would find hard to explore experimentally.²⁹⁻³¹

Advances in modelling and molecular dynamics (MD) studies of natural systems. In 2006, molecular dynamics (MD) simulations of the satellite tobacco mosaic virus capsid and RNA represented the largest systems tractable for simulation, comprising ~ 1 million atoms and simulated for a total of 50 ns³². Since then, advances in high performance computing and software has significantly increased this capability. Recently, the HIV empty viral capsid comprised of 64 million atoms was simulated for 1 microsecond.³³ Extending the study of the dynamical properties of systems to ever greater size, complexity and time-scales will no doubt enhance our understanding of the drivers for complex molecular machinery. Exploring the drivers for the assembly of these systems are, as yet, beyond the timescales we can usefully sample with atomistic methods and for this, methods such as coarse grain modeling is required.^{24,34}

Synthetic biology aims to build self-assembling structures from simpler components for bespoke applications. The benefits of developing a “suite” of simple building block components, with a range of properties, capable of self-assembly to form a nanostructure, provides the opportunity to mix and match components to tailor the properties of the assembled nanocage for a specific function. Individual components may be easier to produce in bulk and have known, reproducible physical properties. Engineered self-assembling systems are now emerging.^{11,15,35} The reduced complexity of these synthetic structures compared with e.g. viral capsids, affords the opportunity to understand in greater detail, at the molecular level, factors that influence their physical properties. Expanding our understanding informs the design process and the modularity of a synthetic systems affords greater control. Desired characteristics could be more readily “dialed-in”. This approach may in the future provide solutions for applications in which timely responses to rapidly changing need is vital, such as fast-track vaccine design and production.

We describe the relative motion of peptide components within each SAGE assembly and follow their interaction with, and effect on, a variety of globular proteins and small molecules added to the solvent box. Solute permeation through the hexagonal pores and adhesion to SAGE surface was monitored and analyzed throughout these simulations. We envisage that exploring the drivers that mediate interactions with various solutes will inform the design of SAGE vehicles tailored for different purposes. For example, a SAGE drug delivery system must retain its contents until it reaches its target cell type and once taken up, release the cargo appropriately.

Likewise, a SAGE nanoreactor incorporating an enzyme pathway for a synthetic task must allow starting materials to diffuse inward and end-products out.

Simulations (0.6-1.0 μ s) were carried out for three SAGE constructs; the original SAGE (parent) and those with four-residue extensions at the N-termini of half of the trimeric hubs, either lysines (K4) or glutamates (E4). These constructs represent a range of charged assemblies; SAGEs with net positive charge (parent), pronounced net and surface positive charge (K4) and the net neutral, but negatively surface charged, (E4). This SAGE set represents the basis for the constructs being explored experimentally towards applications including targeted drug delivery, vaccine development and nanoreactors. All structures were simulated in a box of explicit water containing 0.15 M sodium chloride ions. Typical simulation boxes contained $\sim$ 42 million atoms and a total of 2.6 microseconds of trajectory data were acquired. It is worth noting here that the size and repeating modularity of the SAGE assemblies allows 3720 individual SAGE peptide chains to be followed dynamically. Here we describe their behavior in the context of experimentally and biologically reasonable concentrations of a variety of proteins (40 μ M each) and small molecules (1 mM each) shown in tables 1 and 2 in SI-2. The length of the simulations coupled with the sheer number of repeating units within the SAGE assemblies provides extensive sampling. We anticipate our observations will prove useful for application-driven design.

Figure 1.

Figure 1. A) Honeycomb lattice assembly between acidic and basic hubs. Green, trimeric peptide (CC-Tri3); red, acidic half of heterodimer (CC-DiA); blue, basic half of heterodimer (CC-DiB); bars between circles represent disulfide bonds. B) is a space-filled representation of the parent-SAGE construct before molecular dynamics. C) Illustration of the vectors fitted to the homotrimers (green) and heterodimers (red and blue) labelled for defining the internal geometry parameters.

Results and Discussion

SAGE construction and simulations

A megadalton model of a SAGE can be built from X-ray structures of peptide modules. The SAGE model was built based on an array of 312 points equally spaced (with icosahedral symmetry) on a sphere corresponding to polygon centers. These points were converted into a hexagonal net and scaled such that the gaps between points matched the distance between trimers in our model of a planar hexagonal mesh. Alternating acidic (A) and basic (B) hubs were placed at each of the 620 vertices, aligned along the radial vector such that the N-termini of all peptides were oriented outside the SAGE in accordance with the curvature observed previously for simulations

of hexagonal patches⁸. The twelve pentagons required for closure (in accordance with Euler's rules) were accommodated by using mixed hubs ensuring that all peptide hydrophobic interfaces were satisfied, Figure 1. (see SI-1 for further details).

SAGE simulations (overview): A cubic periodic boundary box, larger than the SAGE by 5 nm in each direction was defined and filled with water, giving approximately equivalent volumes inside and outside the SAGE particle. Proteins and small molecules were added at random positions and orientations using BUDE³⁶ while maintaining the same concentration of solute inside the SAGE as outside. Simulations were performed under the amber99SB-ildn³⁷ forcefield using GROMACS 4.6.7³⁸ under conditions of constant pressure (1 bar) and temperature (300 K) with the PME method used for long range electrostatics. Virtual sites for chemical groups exhibiting fast motions (specifically, v-site hydrogens) allowed a 5 fs timestep to be used for the longer simulations without compromising their thermodynamic and structural behavior.³⁹ For further details of the setup, including sequence information, of the SAGE simulations please see Supplementary Information sections SI-2. Structures were saved every 100 ps for further analysis. The simulations acquired during this work are listed in SI-2 Table 1 and SI-2 Table 2 that also detail the proteins and small molecules added to three long simulations for parent, K4 and E4 SAGEs. The proteins and small molecule solutes were chosen to represent a wide variety of sizes, physical properties and potential utility. Details of the parallelization performance and data analysis can be found in SI-3-4. In this study, a single layer of peptides was built to form a hexagonal mesh representing the simplest model of SAGE assembly. This structure comprises 3720 peptides and a total of 93000-115320 amino acid residues (depending on the SAGE construct) giving a hollow object about 70 nm in diameter. We built the SAGEs based on a spherical array of 312 points as described in methods and SI-1. The parent-SAGE model is shown in Figure 1B This model maximizes contacts between adjacent trimers and dimers in the honeycomb lattice leading to all N-termini being located on one face of the lattice. MD simulation of a flat section of this lattice leads to spherical cap formation with all N-termini on the convex face, as previously described.²⁰ Each dimer is connected to the flanking trimers via disulfide bonds, enabling alternative orientations of the dimer, while the trimers are locked in an “N-terminus-out” orientation in the SAGE model courtesy of their 3 linkages to dimers. It is also appropriate to observe that the simulations described here are too short to observe switching of peptides between partners (as doubtless happens over longer timescales) rather we are exploring the conformational flexibility around local minima corresponding to instances of already self-assembled and idealized SAGE structures. Assembly characteristics must be studied by other theoretical approaches.²⁴

Figure 2.

Figure 2. RMSD of the C α atoms of the SAGE particles with time in the five simulations for: A) parent-SAGE-dim-o (cyan); parent-SAGE-dim-i (brown) and B) parent-SAGE-mols (green); K4-SAGE-mols (blue); E4-SAGE-mols (red). Radius of gyration of the SAGE particle in the five simulations for: C) parent-SAGE-dim-o (cyan); parent-SAGE-dim-i (brown). D) parent-SAGE-mols (green); K4-SAGE-mols (blue); E4-SAGE-mols (red) (an expanded view can be found in SI-5.2)

Five simulations were performed over a two-year period with a largely continuous scripted queue/run cycle using 128 nodes (3072 cores) of the UK national machine Archer generating 2.4 μ s of simulation time and some 6 TB of unformatted trajectory data. In all simulations but one, the N-termini of all the peptides in the initial structures are located on the outer surface of the SAGE. These constructs are named parent-SAGE-dim-o, parent-SAGE-mols, K4-SAGE-mols and E4-SAGE-mols, with the suffix mols corresponding to cases with added solutes. The exception has the N-termini of all the dimers located inside the SAGE and is named parent-SAGE-dim-i. Every peptide in the parent SAGE carries one excess positive charge giving an overall charge of +3720. The K4-SAGE has a 4-lysine extension on the N-terminus of all trimer peptides in the basic hub giving this construct an overall charge +7440. The E4-SAGE has these lysine extensions replaced by glutamates giving a neutral species overall, albeit with a negative surface charge. Further details of the simulations are given in SI-2 Table 1.

Global structure and dynamics

Firstly, we examined the typical time-dependent structural parameters that provide information regarding the structural stability of the assembly during the simulations via the root mean squared deviation (RMSD) to the initial structure and the radius of gyration (Rg).

Dimer peptide orientation affects SAGE contractility: The first 100 ns simulation, parent-SAGE-dim-o, has all the dimers oriented with N-termini out, and most dimers remain in this conformation (see SI: parent-SAGE-o_movie). The size of the SAGE sphere as reported by the evolution of the radius of gyration remains constant over the 100 ns. In contrast, the second 100 ns simulation, parent-SAGE-dim-i, was started from a conformation in which all the dimers were rotated by 180° (dimer N-in). This showed less robust behavior, with a gradual contraction of the radius during 100 ns, see Figure 2C. Also, the surface of the parent-SAGE-dim-i is more

“floppy” than its (dimer N-out) counterpart. In the last frame the standard deviation of the average radius for parent-SAGE-dim-i is twice that for the, parent-SAGE-dim-o (1.08 nm and 0.49 nm respectively).

The effect of appending charged tags to the trimers: Following an initial relaxation, the radius of gyration of all three constructs, parent-SAGE-mols, K4-SAGE-mols and E4-SAGE-mols showed a low but discernible contraction across the simulations (Figure 2D), unfortunately the longer simulations required to track this trend over longer (> millisecond) time periods are not possible on currently available hardware. Nevertheless, we have shown that these models of SAGE assemblies are stable and persistent over the microsecond timescale *in silico*. It is noteworthy that the neutral E4-SAGE exhibits a higher RMSD and lower Rg than either the parent or K4-SAGEs. We suggest that this a consequence of the overall positive charges of the parent and K4-SAGEs resisting compression of the SAGE particle.

Secondary structure persists over the simulations: Helical secondary structure is well-preserved over the simulations, providing confidence that this designed structure behaves as intended. This is shown graphically in SI-5.1. Here, each pixel (representing a single peptide) is white if it retains more than 50% helicity within helical Ramachandran space and is colored if it has lost more than 50% helicity. The spot density is poised such that if all peptides had lost more than 50% helical structure, over the trajectories, the field would be a single block of color. Hence the rather faint spot-density plot illustrates that most of the peptides retain most of their helical structure over time.

Tracking the mobility of SAGE peptide components guides strategy for “elaboration”: A variety of parameters related to hub geometry were extracted from the trajectories to reveal the evolution of the structure and conformational dynamics of the basic SAGE building blocks. The first set of data comprises distances and angles between peptide units, specifically between vectors fitted to the trimer and dimer helical components of the SAGE as illustrated in Figure 1 C.

The distances between neighboring hub trimers (Figure 1C: vector BE) showed no significant difference between all five simulations. These range between 3 and 4.5 nm with the histograms peaking just below 4 nm. (Figure 3, Panels 1-5A). There is however some broadening observed in the three longer simulations (Figure 3 panels 3-5A) compared with the two shorter ones (1A and 2A), consistent with increased conformational sampling in the former. This behavior is also evident in the distribution of angles between each trimeric coiled coil (e.g. Figure 1C: vectors BA and EF) and the corresponding radial vector (e.g. Figure 1C: where R=SAGE-center, vectors RB and RE). The angle histograms are shown in Figure 3 (panels 1-5B).

Figure 3.

Figure 3. Histograms showing the distribution of peptide helix orientation throughout the simulations, by distance ((panels A) illustrated in Figure S5 as B→E) and by angle (thus how perpendicular) to the center of the SAGE the three-fold axis of the trimers (panels B) and two-fold axis of the dimers (panels C) remain. Panel numbers (and colour-code) correspond to the simulations of: (1) parent SAGE dimer N-out (cyan); (2) parent SAGE dimer N-in (gold); (3) 1 microsecond of the parent SAGE with solutes (green); (4) 650 ns of the K4 SAGE with solutes (blue) (5) 650 ns of the E4 SAGE with solutes (red). The probability (black lines) in these plots are derived from the number of counts in a bin divided by the surface area of the spherical sector of the bin.

The orientation of the dimers can be represented by the distribution of dihedral angles (Figure 1C: defined by ABCD and FECD). Histograms for these data are shown in Figure 3_new (panels 1-5C). Analysis of the two shorter simulations show no interconversion between the dimer-N-out, panels (1C) orientation, and the dimer-N-in case (2C), but a wider range of angles are sampled in the former. During the longer simulations there are some dimers that reorient from dimer-N-out to dimer-N-in. However, longer simulations (by more than an order of magnitude) are required to determine whether conformational equilibration has occurred. In terms of

informing design, such dimer flexibility suggests that appending “functional decoration” to the dimer pair, with peptides or proteins, would be a riskier strategy for maintaining current SAGE properties than modifying the trimers. This is consistent with experimental findings that appending peptides to the dimers compromises SAGE formation (unpublished data).

87% of all possible dimer rotations was sampled, despite non-equilibration. Dimer orientation was also examined in the context of the sidechain (χ_1 , χ_2) and disulfide (S-S) dihedral angles of the two cystine bridges coupling a dimer to its neighboring trimers. Hence, changes in the 10 torsion angles of the two cystine groups govern the set of alternative orientations of the dimer. For example, assuming 3 possible values each for the χ_1 and χ_2 angles and 2 for the disulfide (i.e. corresponding to low energy conformations), an exploration of the 26244 possible conformations using rigid body geometry found 344 accessible conformations for the dimer in the context of the hexagonal lattice. By measuring these 10 dihedral angles for all dimers in the 2.4 μ s of all the trajectories and assigning these to the nearest minima showed that most of these 344 torsion angle permutations (87%) were sampled during the simulations. These results allow us to conclude that over this time frame the simulations have accessed most of the possible dimer conformations despite being a long way from equilibrium. Consequently, it appears that the only barrier to achieving an equilibrated simulation is sampling time.

Informative anomaly occurring in parent SAGE-mols run: The simulations were stable apart from an unexpected perturbation occurring around 100 ns in the parent-SAGE-mols calculation. An unrealistic movement of at least one SAGE atom towards the edge of the simulation box resulted in a rent in the SAGE, as four adjacent homodimer pairs were pulled apart. The simulation recovered from this insult and we chose to allow it to continue so we could monitor whether the tear would propagate around the SAGE or re-anneal. This would provide some insight into integrity requirements for stability. In the event, the defect remained stable throughout the subsequent 900 ns without any discernible effect on the SAGE structure in general, beyond a slight and expected increase in flexibility around this region. In reality, we might expect such defects to be self-healing, but this simulation is far too short to observe such behavior, because of the large entropic barrier to reforming the cognate coiled coil interactions. We repeated this portion of the run using checkpoint point files from just prior to the rupturing event and the second run over the same time-period failed to produce a tear (see figure SI-5.3). The fact that this repeat produced the expected, yet different outcome, leads us to suspect that the rupture was an artefact resulting from a runtime error.

SAGEs and solutes.

In the preceding section we addressed the dynamic behavior of the different SAGE constructs in terms of their geometry and stability. Next, we turn our attention to the interactions of the SAGE peptide network with solutes present in the simulations from salt through small molecules to proteins. The small molecules and proteins were chosen to represent a range of sizes and charge. These were added to the simulation box, with the same number of molecules inside and outside the cage structures. The solutes are listed in SI-1 Table 2 with their respective charges and molecular weights. Ten copies of each protein and 250 copies of each small molecule were distributed in random positions and orientations in the simulation boxes, corresponding to concentrations of 40 μ M and 1 mM respectively. The parent-SAGE was simulated for 1 μ s (parent-SAGE-mols) and the K4-SAGE (K4-SAGE-mols) and E4-SAGE (E4-SAGE-mols) for 0.6 μ s each (see SI: parent-SAGE-mols_movie, K4-SAGE-mols_movie, E4-SAGE-mols_movie). We performed analyses to measure solute adherence to, or passage through, the SAGE peptide network.

The passage of molecules through the SAGE pores was monitored by dividing the simulation box into 3 regions defined as I = inside, closest to the geometric center of the SAGE coordinates (C), S = surface-associated, within the radial zone occupied by the SAGE and O = outside, in the region furthest from the SAGE center. The radial zone is defined as the volume lying between two concentric spheres, the inner touching the SAGE atom closest to the center C and the outer touching the SAGE atom furthest from C. The radial zone was recalculated at each timepoint. Each molecule was labelled at each 100 ps timepoint of the simulation as either I (inside), S (in the SAGE zone) or O (outside) based on the distance between their center of coordinates from that of the center of the SAGE. This allowed transitions to be identified where molecules had moved through a pore in either an outward, I $\rightarrow$ S $\rightarrow$ O or inward, O $\rightarrow$ S $\rightarrow$ I direction. A molecule was deemed to be in contact with the SAGE if at least one atom from each are within 6 Å of each other.

Ions pass freely through the peptide network of the SAGEs. Each system comprises a SAGE particle, protein and small molecule solutes, water and NaCl at 150 mM and was adjusted to be charge and pH neutral overall.

Consequently, there were 38514, 36618 and 40338 sodium ions for the parent, K4 and E4 SAGE-mols simulations, respectively and 42774, 44598, and 40878 chloride ions for the parent, K4 and E4 SAGE-mols simulations respectively. We tracked individual sodium and chloride ions within each system to gauge how often they traversed the SAGE boundary by diffusion. Plots showing the number of times ions passed from outside to inside the SAGE assemblies are shown in Figure 4. Trajectory coordinates were saved every 0.1 ns. We have calculated the distance travelled for all single Na^+ Cl^- ions between these timepoints, to estimate the probability it has traversed more than once. Sodium and chloride ions have diffusion constants in water at 25 °C close to 1×10^{-5} and $2 \times 10^{-5} \text{ cm}^2 \text{ s}^{-1}$ respectively⁴⁰, corresponding to a sub-nanometer average linear displacement per 0.1 ns. The width of the SAGE S layer is $10.8 \pm 1.0 \text{ nm}$, hence the great majority of events measured will correspond to single passages of an ion across the SAGE network. The simulation data correspond reasonably well with these experimental values, having average displacements of $1.2 \pm 0.5 \text{ nm}$ (Na^+) and $1.3 \pm 0.6 \text{ nm}$ (Cl^-) over 0.1 ns. (SI-6.1.2).

Figure 4.

Figure 4. The frequency distribution of the passage of sodium ions (blue) and chloride ions (green) crossing the SAGE boundaries inwards during the time-period 0-600 ns for the E4-SAGE (panel A), parent-SAGE (panel B) and K4-SAGE (panel C). Corresponding graphs showing ion passage in the outwards direction are shown in Figure SI-6.1.1.

Small molecule permeability and surface adherence: The percentage of small molecule ligands in contact with the SAGE assemblies over the trajectories is shown in figure 5. Here charge seems to have had the largest influence on differential colocalization. For example, the highly positively charged molecule spermine adheres to E4-SAGE, shows intermediate behavior with parent-SAGE and is repelled from K4-SAGE.

Figure 5.

Figure 5. Graphs showing the association of small-molecule solutes with the SAGE constructs over time for A parent-SAGE; B K4-SAGE and C E4-SAGE. MGA methylgalactoside; MGL methylglucoside; SPM spermine; ATP adenosine triphosphate; MG magnesium; CFL carboxyfluorescein; BYP bipyridyl platinum; IMI imidazolium; IMD imidazole.

We envisage future applications where solute passage needs to be either permissive or prevented. The results, shown in Figure 5, indicate that the SAGEs are remarkably sticky. Most of the molecules spend most of their time adhering to the SAGE surface, the exceptions being small and very hydrophilic solutes, particularly imidazole and its protonated ion. The mixed behavior of the small molecules (some adhering and some passing freely through the SAGE peptide network) was illustrated by plotting the distance from the center of the SAGE of each molecule with time. By representing the molecules as points on the graph and coloring these differently depending on their initial location inside versus outside the SAGE, it is possible to visually follow adherence and permeation with time. See SI-6.2 for these plots.

Protein solute adherence: The percentage of protein solutes in contact with parent-SAGE, K4-SAGE and E4-SAGE over the course of their trajectories reveals again how sticky they are (figure 6). In all cases (parent, K4 and E4 SAGEs) roughly 80% of the individual proteins have stuck to the SAGEs by 200 – 300 ns. The exceptions are ubiquitin for parent and K4-SAGEs and GFP where only 40% stuck to the E4 SAGE. This latter behavior may reflect a surface charge influence. GFP has a net charge of -3, which may account for its reduced binding to the negative surface charge on the E4 SAGE.

FAB, although heavily positively charged seems to bind with almost equivalent avidity to all assemblies regardless of overall or surface charge, this may be related to the lack of glycosylation in the FAB model.

Figure 6.

Figure 6. Graphs showing the association of protein solutes with the SAGE constructs over time. Vertical axes show the percentage of molecules within 6 Å of the SAGE. A, parent-SAGE; B, K4-SAGE; C, E4-SAGE. Key for proteins: (FAB, anti-body fragment; GFP, green fluorescent protein; SPY, spycatcher; UBI, ubiquitin; LZIP, leucine zipper; CRB, crambin.

We have evidence for this adherent behavior *in vitro* as adding GFP to parent-SAGE is enough to stain in correlated light EM and sharpen the SAGE DLS signal.²⁵ In addition, we have experimentally quantified the amount of soluble GFP binding to the parent, K4 and E4 SAGEs according to the methods and figure shown in SI-7.1. These data reflect a similar tendency for solutes to adhere to the SAGE surface *in vitro* as *in silico*. The experiments performed here indicate that around 200 GFP proteins bind to one SAGE particle and there is little dependence on the charge of the SAGEs.

Heatmaps help identify “sticky” spots. We drilled down further into the contact data to identify which residues (both by type and position) are primarily responsible for adhesion. An example of the information that can be gleaned from heat maps is shown in Figure 7. Since the trajectories contain structures saved every 100 ps, a contact score of 10000 would mean a molecule was in contact with the parent-SAGE throughout a 1000 ns trajectory. In these maps, protein solutes are numbered 1-5 if they began the simulation inside the SAGE and 6-10 for those outside. Further inspection of panels A and B reveal that the second residue in the parent-SAGE trimer interacts most persistently with both GFP and FAB molecules (denoted by a red block). This information correlates well with the information in panels C and D showing that the parent-SAGE residue type that most frequently interacts with both GFP and FAB were glutamates and the second residue of the parent SAGE is indeed a glutamate. The E2 residue in the trimer also is a frequent interaction site with other proteins for the parent, K4 and E4-SAGE interactions, indicating it is a residue to avoid if non-specific binding is an issue. The most frequently interacting residues are the ones that started out nearest the external surface of the SAGE, suggesting that penetration of the peptide network by these protein solutes is limited, at least on this time-scale. By heat-mapping residue contacts in this way, we identified SAGE residues that are most likely to contribute to non-specific binding events, thus informing the design process for the next generation of *in vitro* SAGE

assemblies. Representations for all the protein solutes in the context of parent, K4 and E4 SAGEs can be found in supplemental information SI-6.3.

Figure 7.

Figure 7. Panels A and B are example heatmaps showing parent-SAGE trimer residue positions in contact with either GFP proteins (panel A) or FAB proteins (panel B). The left-hand y-axis numbers the proteins according to their starting position inside (1-5) the SAGE or outside (6-10). In all panels, the right-hand y-axis shows the number of contacts over the trajectory according to the color key shown (red being most contacts). In panels C and D the left-hand y-axis shows which parent SAGE residue type most often contacted which residue type in any of the GFP molecules (panel C) or FABs (panel D).

Effects on solute protein structure when bound to SAGE assemblies: Finally, we explored whether association between the protein solutes and the large SAGE object have any discernable effect on their structure. The inherent modularity of the SAGE system provides the potential to “dial in” protein attachments to confer properties related to targeting or function.²⁵ For many potential applications these will need to be appropriately presented in a native conformation for e.g. targeting to specific cell receptors or as antigens. Backbone RMSDs were calculated for each of the protein solute molecules and graphical representations for all solute protein RMSDs in the context of SAGE contact for all the simulations (Parent, E4 and K4) can be found in SI-6.4. There is little consistent correlation (SI-6.4.4) between fluctuations in RMSD and contact with SAGE assemblies. This suggests that for the most part, contact with any of the current SAGE constructs does not induce significant conformational changes of the solute proteins. We regard this as encouraging in the context of adding proteins or peptides to the SAGE assembly that need to retain as near native conformations as possible to function appropriately.

Conclusions

In anticipation of the SAGEs’ potential in synthetic biology, we have performed long and large dynamic simulations and monitored the interplay between SAGE assemblies and protein and small molecule solutes. SAGEs as nanoreactors²⁵ will require permeability to enable the flow of substrates in and products out. Our

studies have shown that though sodium and chloride ions diffuse freely through the peptide network, small molecule permeation is largely influenced by charge, an effect that is more pronounced for the protein solutes. This behavior will need to be factored in to nanoreactor design. Medical applications using SAGEs as drug or vaccine delivery platforms will require many specific attributes around targeting, rigidity, stability (shelf-life) and *in vivo* susceptibility to degradation. SAGE assemblies are stable under simulation conditions, even with an artefactually induced rent in the side. Component peptide orientation has revealed that cell-targeting or epitope presenting peptides and proteins are best attached to the less motile trimer peptides. Lastly, the negligible influence on solute protein conformation by SAGE contact, bodes well for presenting native-like larger peptide structures as conformational epitopes in a SAGE-based vaccine delivery system.

The ~42 million atom simulations reported here have pushed the boundaries of size, complexity and duration possible with current hardware. We have shown that with high performance computing such as Archer and software such as GROMACS, atomistic molecular dynamics simulations of this size are possible and yield informative results. Considering the inexorable advances in computer architecture, these types of simulation will become run-of-the-mill in the near-future.

Our analyses have helped us gain a better understanding of how to manipulate this inherently versatile system towards synthetic biological applications as diverse as nanoreactors for complex, eco-friendly chemistry and delivery systems for drugs and vaccines.

Acknowledgments

Extensive computer time on the UK HPC machine Archer was secured via an EPSRC LeaderShip Award. We also thank the UK HECBiosim Consortium and the University of Bristol Advanced Computing Research Center for further provision of computer time and storage. DKS, MM, RBS, DNW and TBL are supported by BrisSynBio, a BBSRC/EPSRC Synthetic Biology Research Center (BB/L01386X/1). RBS and DNW are funded by a BBSRC LoLa grant (BB/M002969/1). DNW is a Royal Society Wolfson Research Merit Award holder (WM140008).

Conflicts of interest. The authors declare no conflicts of interest.

ABBREVIATIONS

RMSD, root mean square deviation; MD molecular dynamics; PME, particle mesh Ewald; SAGE, self-assembling protein nanocage; E4-SAGE, self-assembling nanocage with 4 glutamate residues at the N-terminus of half of the trimer peptides; K4-SAGE, self-assembling nanocage with 4 lysine residues at the N-terminus of half of the trimer peptides; BUDE Bristol University Docking Engine; AFM atomic force microscopy; GROMACS GROMINGEN MACHINE for Chemical Simulations; EM, electron microscopy; DLS, dynamic light scattering.

References

1. Y. Zhang, M.S. Ardejani, B.P. Orner, *Chem. Asian J.*, 2016, **11**, 2814-2828.
2. J.G. Heddle, S. Chakraborti, K. Iwasaki, K. *Current Opinion in Structural Biology*, 2017, **43**, 148-155. <https://doi.org/10.1016/j.sbi.2017.03.007>
3. W.M. Aumiller, M. Uchida, T. Douglas, *Chem. Soc. Rev.*, 2018 **47**, 3433-3469. <http://dx.doi.org/10.1039/C7CS00818J>
4. F. Lapenta, J. Aupic, Z. Strmsek, R. Jerala, *Chem. Soc. Rev.* 2018, **47**, 3530-3542. <http://dx.doi.org/10.1039/C7CS00822H>
5. M.S. Ardejani, X.L. Chok, C.J. Foo, B.P. Orner, *Chem. Commun.*, 2013, **49**, 3528-3530. <http://dx.doi.org/10.1039/C3CC40886H>
6. T.A. Cornell, J. Fu, S.H. Newland, B.P. Orner, *J. Am. Chem. Soc.*, 2013, **135**, 16618-16624. <https://doi.org/10.1021/ja4085034>
7. J.G. Heddle, I. Fujiwara, H. Yamadaki, S. Yoshii, K. Nishio, C. Addy, I. Yamashita, J.R. Tame, *Small*, 2007 **3**, 1950-6.
8. T.A. Cornell, M.S. Ardejani, J. Fu, S.H. Newland, Y. Zhang, B.P. Orner, *Am. Chem. Soc.*, 2018, **57**, 604-613. doi: 10.1021/acs.biochem.7b01000
9. J.E. Padilla, C. Colovos, T.O. Yeates, *Proc. Natl. Acad. Sci. USA.*, 2001, **98**, 2217-2221. <http://www.pnas.org/content/98/5/2217.abstract>
10. Y.T. Lai, D. Cascio, T.O. Yeates, *Scienc*, 2012, **336**, 1129. DOI: 10.1126/science.1219351
11. N.P. King, W. Sheffler, M.R. Sawaya, B.S. Vollmar, J.P. Sumida, I. André, T. Gonen, T.O. Yeates, D. Baker, *Science*, 2012, **336**, 1171-1174. doi: 10.1126/science.1219364
12. N.P. King, J.B. Bale, W. Sheffler, D.E. McNamara, S. Gonen, T. Gonen, T.O. Yeates, D. Baker, *Nature*, 2014, **510**, 103. <http://dx.doi.org/10.1038/nature13404>.
13. Y.T. Lai, E. Reading, G.L. Hura, K.-L. Tsai, A. Laganowsky, F.J. Asturias, J.A. Tainer, C.V. Robinson, T.O. Yeates, *Nat. Chem.*, 2014, **6**, 1065. <http://dx.doi.org/10.1038/nchem.2107>
14. J.B. Bale, S. Gonen, Y. Liu, W. Sheffler, D. Ellis, C. Thomas, D. Cascio, T.O. Yeates, T. Gonen, N.P. King, D. Baker, 2016, *Science*, **353**, 389-394. <http://science.sciencemag.org/content/353/6297/389.abstract>
15. G.L. Butterfield, M.J. Lajoie, H.H. Gustafson, D.L. Sellers, U. Nattermann, D. Ellis, J.B. Bale, S. Ke, G.H. Lenz, A. Yehdego, R. Ravichandran, S.H. Pun, N.P. King, D. Baker, *Nature*, 2017, **552**, 415. <http://dx.doi.org/10.1038/nature25157>

16. A.Sciore, M.Su, P.Koldewey, J.D.Eschweiler, K.A.Diffley, B.M.Linhares, B.T.Ruotolo, J.C.A.Bardwell, G.Skiniotis, E.N.G.Marsh, *Proc. Natl. Acad. Sci. USA*, 2016, **113**, 8681-8686. <http://www.pnas.org/content/113/31/8681.abstract>
17. S. Badiyan, A. Sciore, J. D. Eschweiler, P. Koldewey, A. S. Cristie-David, B. T. Ruotolo, J. C. A. Bardwell, M. Su and E. N. G. Marsh, *ChemBioChem*, 2017, **18**, 1871, DOI: 10.1002/cbic.201700481.
18. H.Gradizar, S.Bozic, T.Doles, D.Vengust, I.Hafner-Bratkovic, A.Mertelj, B.Webb, A.Sali, S.Klavzar, R.Jerala, *Nat. Chem. Biol.*, 2013, **9**, 362-366. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3661711/>
19. A.Ljubetič, F.Lapenta, H.Gradišar, I.Drobnak, J.Aupič, Z.Strmšek, D.Lainšček, I.Hafner-Bratkovič, A.Majerle, N.Krivec, M.Benčina, T.Pisanski, T.C.Veličković, A.Round, J.M.Carazo, R.Melero, R.Jerala, *Nat. Biotechnol.*, 2017, **35**, 1094-1101. doi: 10.1038/nbt.3994.
20. J.M.Fletcher, R.L.Harniman, F.R.Barnes, A.L.Boyle, A.Collins, J.Mantell, T.H.Sharp, M.Antognozzi, P.J.Booth, N.Linden, M.J.Miles, R.B.Sessions, P.Verkaide, D.N.Woolfson, *Science*, 2013, **340**, (6132), 595-9.DOI: 10.1021/ja2082476
21. J.M.Fletcher, A.L.Boyle, M.Bruning, G.J.Bartlett, T.L.Vincent, N.R.Zaccai, C.T.Armstrong, E.H.C.Bromley, P.J.Booth, R.L.Brady, A.R.Thomson, D.N.Woolfson, *ACS Synthetic Biology*, 2012, **1**, 240-250. <https://doi.org/10.1021/sb300028q>
22. F.Thomas, A.I.Boyle, A.J.Burton, D.N.Woolfson, *J. Am. Chem. Soc.*, 2013, **135**, 5161-5166. <https://doi.org/10.1021/ja312310g>
23. J.M.Galloway, L.Senior, J.M.Fletcher, J.L.Beasley, L.R.Hodgson, R.L.Harniman, J.M.Mantell, J.Coombs, G.G.Rhys, W-F.Xue, M.Mosayebi, N.Linden, T.B.Liverpool, P.Curnow, P.Verkaide, D.N.Woolfson, *ACS Nano*, 2017, **12** (2) 1420-1432.
24. M.Mosayebi, D.K.Shoemark, J.M.Fletcher, R.B.Sessions, N.Linden, D.N.Woolfson, D.N., T.B.Liverpool, *Proc. Natl. Acad. Sci. USA*, 2017 **114**, 9014 - 9019 DOI:10.1073/pnas.1706825114
25. J.F.Ross, A.Bridges, J.M.Fletcher, D.K.Shoemark, D.Alibhai, H.E.V.Bray, J.L.Beasley, W.M.Dawson, L.R.Hodgson, J.M.Mantell, P.Verkaide, C.M.Edge, R.B.Sessions, D.Tew, D.N.Woolfson, *ACS Nano*, 2017, **11**, 7901–7914
26. N.Nandhagopal, A.A.Simpson, J.R.Gurnon, X.Yan, T.S.Baker, M.V.Graves, J.L.Van Etten, M.G.Rossmann, *Proc. Natl. Acad. Sci. U S A*, 2002, **99**, 14758-14763. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC137492/>
27. C.A.Kerfeld, C.Aussignargues, J.Zarzycki, F.Cai, M.Sutter, *Rev Microbiol.*, 2018, **16**(5):277-290. doi: 10.1038/nrmicro.201
28. M.J.Lee, J.M.Mantell, L.Hodgson, D.Alibhai, J.M.Fletcher, I.R.Brown, S.Frank, W-F.Xue, P.Verkaide, D.N.Woolfson, M.Warren, *Nature Chem. Biol.*, 2017, **14**, 142-147 <http://dx.doi.org/10.1038/nchembio.2535>
29. J.R.Perilla, B.C.Goh, C.K.Cassidy, B.Liu, R.C.Bernardi, T.Rudack, H.Yu, Z.Wu, K.Schulten, *Curr. Op. Struct. Biol.* 2015, **31**, 64-74
30. E.Tarasova, V.Farafonov, R.Khayat, N.Okimoto, T.S.Komatsu, M.Taiji, D.Nerukh, *J. Phys. Chem. Lett.*, 2017, **8**, 779-784.
31. J.R.Perilla, J.A.Hadden, B.C.Goh, C.G.Mayne, K.Schulten, *J. Phys. Chem. Lett.*, 2016, **7**, 1836–1844.
32. P.L.Freddolino, A.S.Arhipov, S.B.Larson, A.McPherson, K.Schulten, *Structure*, 2006, **14**, 437–449.
33. J.R.Perilla, K.Schulten, *Nature Comms.*, 2017, **8**, 15959. DOI: 10.1038/ncomms15959
34. D.C.Rapaport, *Journal of Biological Physics*, 2018, **44**, (2) 147–162
35. S.Kwon, H.S.Shin, J.Gong, J-H.Eom, A.Jeon, S.H.Yoo, I.S. Chung, S.J.Cho, H.S.Lee, *J. Am. Chem. Soc.*, 2011 **133** (44), 17618-17621.
36. S.N.McIntosh-Smith, J.R.Price, R.B.Sessions, A.Avila Ibarra, *International Journal of High Performance Computing Applications*, 2015, **29**, no. 2, 119-134. DOI: 10.1177/1094342014528252
37. K.Lindorff-Larsen, S.Piana, K.Palmo, P.Maragakis, J.L.Klepeis, R.O.Dror, D.E.Shaw, *Proteins* 2010, **78**, 1950–1958.
38. B.Hess, C.Kutzner, D.van der Spoel, E.Lindahl, *J. Chem. Theory Comp.*, 2008, **4** (3), 435–447.
39. H.J.C.Berendsen, W.F.van Gunsteren, W. F. Molecular dynamics simulations: Techniques and approaches. In *Molecular Liquids-Dynamics and Interactions*, Barnes, A. J., Orville-Thomas, W.J., Yarwood, J. Ed. Reidel Dordrecht: The Netherlands, 1984, 475-500
40. E. Hawlicka, *Z. Naturforsch.* 1987, **42a**, 1014-1016.