Assembly of Capsids from Hepatitis B Virus Core Protein Progresses through Highly Populated Intermediates in the Presence and Absence of RNA

Oliver, Ryan C.; Potrzebowski, Wojciech; Najibi, Seyed Morteza; Pedersen, Martin Nors; Arleth, Lise; Mahmoudi, Najet; Andre, Ingemar

Published in:
ACS NANO

DOI:
10.1021/acsnano.0c03569

Publication date:
2020

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Oliver, R. C., Potrzebowski, W., Najibi, S. M., Pedersen, M. N., Arleth, L., Mahmoudi, N., & Andre, I. (2020). Assembly of Capsids from Hepatitis B Virus Core Protein Progresses through Highly Populated Intermediates in the Presence and Absence of RNA. *ACS NANO*, *14*(8), 10226-10238. https://doi.org/10.1021/acsnano.0c03569

ACS NANO

www.acsnano.org

ARTICLE

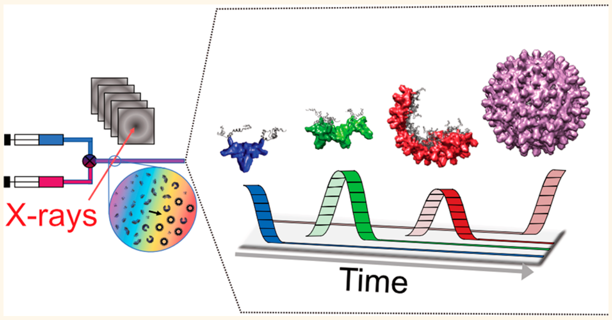
Assembly of Capsids from Hepatitis B Virus Core Protein Progresses through Highly Populated Intermediates in the Presence and Absence of RNA

Ryan C. Oliver,[#] Wojciech Potrzebowski,[#] Seyed Morteza Najibi, Martin Nors Pedersen, Lise Arleth, Najet Mahmoudi, and Ingemar André*

Cite This: *ACS Nano* 2020, 14, 10226−10238

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: **The genetic material of viruses is protected by protein shells that are assembled from a large number of subunits in a process that is efficient and robust. Many of the mechanistic details underpinning efficient assembly of virus capsids are still unknown. The assembly mechanism of hepatitis B capsids has been intensively researched using a truncated core protein lacking the C-terminal domain responsible for binding genomic RNA. To resolve the assembly intermediates of hepatitis B virus (HBV), we studied the formation of nucleocapsids and empty capsids from full-length hepatitis B core proteins, using time-resolved small-angle X-ray scattering. We developed a detailed structural model of the HBV capsid assembly process using a combination of analysis with multivariate curve resolution, structural modeling, and Bayesian ensemble inference. The detailed structural analysis supports an assembly pathway that proceeds through the formation of two highly populated intermediates, a trimer of dimers and a partially closed shell consisting of around 40 dimers. These intermediates are on-path, transient and efficiently convert into fully formed capsids. In the presence of an RNA oligo that binds specifically to the C-terminal domain the assembly proceeds *via* a similar mechanism to that in the absence of nucleic acids. Comparisons between truncated and full-length HBV capsid proteins reveal that the unstructured C-terminal domain has a significant impact on the assembly process and is required to obtain a more complete mechanistic understanding of HBV capsid formation. These results also illustrate how combining scattering information from different time-points during time-resolved experiments can be utilized to derive a structural model of protein self-assembly pathways.**

KEYWORDS: *hepatitis B virus, capsid assembly, time-resolved, small-angle scattering, electron microscopy, multivariate curve resolution, Bayesian statistics*

One essential step during viral−host infection is the encapsulation of the nucleic acids necessary for viral reproduction by a protein shell. Formation of this shell is one of the most complex processes in biology and typically involves the spontaneous self-assembly of hundreds of individual and identical subunits. The final assembled protein shell, or capsid, must also selectively transport and deliver its cargo to infect new host cells, where this cycle can be repeated. As the protein capsid serves a vital role not only in viral maturation but also for cellular recognition, a better understanding of these mechanisms is desirable for both the development of antiviral drugs[1] and other applications such as vaccine development,[2] biotechnology,[3] and protein design/engineering.

Our understanding of the mechanism for capsid formation has primarily been advanced through assembly experiments *in vitro*.[4] In successful examples, capsid proteins were purified under conditions where disassembly was favored and the

Received: April 29, 2020
Accepted: July 16, 2020
Published: July 16, 2020

https://dx.doi.org/10.1021/acsnano.0c03569
ACS Nano 2020, 14, 10226−10238

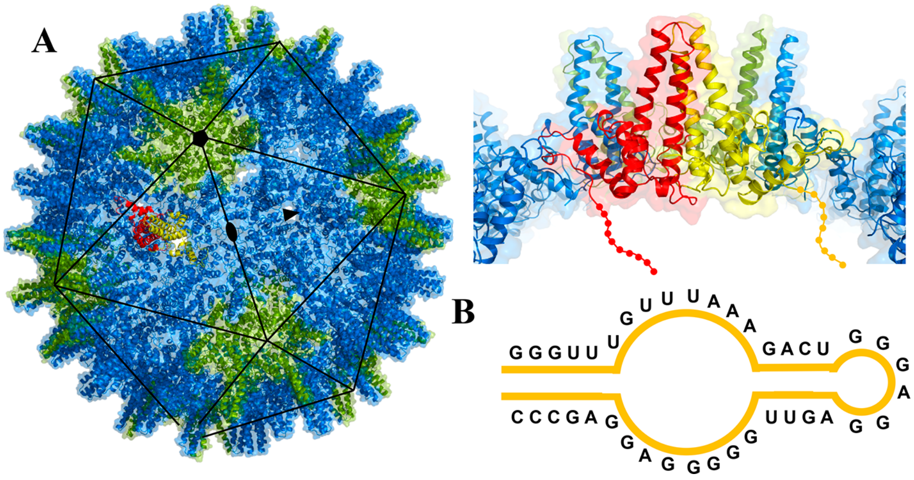


Figure 1. Structural arrangement and subunit organization of the HBV T = 4 viral capsid. (A) A Cp dimer in red and yellow is shown in the context of the T = 4 Cp149 capsid (from PDB code 1QGT[24]), with dashed lines indicating the C-terminal domain. Subunits at the 5-fold symmetry axis are shown in green with the icosahedral symmetry shown as black lines and 2-, 3-, and 5-fold symmetry centers indicated with black symbols. (B) The RNA oligo (PS; see Results) used in the TR-SAXS.

resulting stable subunits were used in kinetic experiments where self-assembly was triggered by changes in solution conditions (pH, temperature, ionic strength, *etc.*) that promote associations between protein subunits. The hepatitis B virus (HBV) mature capsid is composed of 120 dimers of core protein (Cp) forming icosahedral structures *in vivo* with T = 4 quasi-symmetry. A small fraction (∼5%) of T = 3 particles, consisting of 90 dimers, has also been observed *in vivo.*[5] The native Cp monomer contains 183 amino acids, with a large disordered C-terminal arginine rich domain (ARD) that binds to pregenomic RNA (pgRNA). A truncated variant containing the first 149 residues (Cp149) lacking the RNA binding domain was capable of assembling into T = 4 capsids (Figure 1).[6] *In vitro* self-assembly of Cp149 has been demonstrated by increasing the ionic strength, subunit concentration, or temperature.[6,7] Cp149 capsid formation was described by a classical nucleation−elongation mechanism,[7,8] occurring in three steps. Capsid formation is initiated through the formation of a nucleus.[8,9] This nucleus is then extended through an "elongation" phase in which capsid subunits are added to the growing assembly.[6] Finally, the incorporation of the last subunit completes the capsid.[10] Cp149 has been suggested to form a trimer of dimers as a nucleus,[6] while the elongation rapidly proceeds through an ensemble of lowly populated intermediates. Kinetic data from light scattering,[6] mass spectrometry,[10−14] and small-angle X-ray scattering (SAXS)[15] have provided a multifaceted picture of the formation of the Cp149 capsid. Much less is known about the assembly pathway of the full-length protein. Capsids have been reassembled from disassembled containers,[16] but the limited solubility of Cp at physiological conditions[16] has made kinetic experiments more complex to carry out. Cp assembly was monitored indirectly *in vitro* using single-molecule fluorescence correlation spectroscopy and following the rotational correlation time of fluorescently labeled pgRNA during encapsulation.[17]

Time-resolved small-angle X-ray scattering (TR-SAXS) is a powerful method to study the mechanism of virus self-assembly.[18−23] Macromolecular complexes can be studied over broad length scales (tens to hundreds of angstroms) and at relatively fast time scales (approaching milliseconds) to provide three-dimensional structural information (size and shape) about the molecular assemblies in solution. Downsides of TR-SAXS include orientational averaging of the scattering signal resulting in limited structural resolution and the fact that scattering from the sample is an ensemble average over all species found in solution. Hence, to develop a kinetic and structural model of the self-assembly process from TR-SAXS, experimental data must be deconvoluted into components and combined with additional information from three-dimensional structures.

In this study we developed a statistical methodology based on multivariate curve resolution to identify basis spectra from TR-SAXS measurements and an approach to predict the structural composition along the assembly trajectory of the full-length HBV core protein, using Bayesian ensemble inference. To characterize the capsid assembly mechanism, we first established experimental conditions that allowed for the controlled self-assembly of virus-like particles (VLPs) at high Cp concentrations, enabling kinetic measurements of assembly reactions. We then carried out kinetic experiments using stopped flow coupled to TR-SAXS. TR-SAXS was performed in both the presence and absence of RNA to determine the effect of RNA binding to capsid subunits on assembly kinetics. The selected RNA had previously been identified as forming stem loops that have sequence-specific interactions with the hepatitis B core protein dimers.[17] Interpretation of the data suggested that the formation of empty (RNA-removed) capsids proceeds along a pathway involving two highly populated intermediate states: an early buildup of trimers of dimers followed by the formation of a late intermediate consistent with large, partial shells. The binding of RNA to the capsid proteins did not significantly alter the self-assembly pathway. Our results demonstrate that full-length core proteins assemble with a significantly different mechanism compared to Cp149, which currently serves as the *de facto* model system for HBV capsid assembly.

RESULTS AND DISCUSSION

Assembly competent dimers of full-length Cp (Cp185, Supplementary Figures 1 and 2) were produced by disassembly of recombinantly expressed VLPs with guanidine hydrochloride (GuHCl), followed by purification using size-exclusion chromatography (SEC). SAXS data recorded for purified Cp185 under disassembly solution conditions were consistent with a folded protein having an extended, or flexible, portion and molecular mass of a dimer (Supplementary Figure 3 and Supplementary Table 1). To further characterize the structure of Cp185 in solution, we compared SAXS spectra predicted from atomistic models to the experimental data. A dimer model extracted from the crystal structure of a Cp149 T = 4 capsid[24] did not fit well to the SAXS spectrum (data not shown), as expected given that the C-terminal domain is not part of the Cp149 structure. To fit a model of Cp185 to the SAXS data, a set of structures was modeled with different conformations of the C-terminal domain. This set of structures was then fit to the SAXS data using a Bayesian inference algorithm[25] that balanced the fit to data with the complexity of the ensemble model. The resulting ensemble, consisting of two dimers each with different conformation of their C-terminal domains, was in excellent agreement with the experimental data (Figure 2). Collectively, the SAXS data was well described by a model in which Cp185, under disassembly conditions, is folded into a dimer adopting the conformation found in the Cp149 capsid with flexible C-terminal domains. However, the presence of very low concentrations of higher order species cannot be fully ruled out.

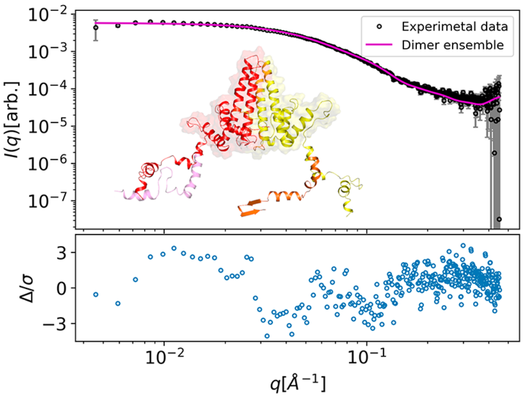


Figure 2. Structural characterization of disassembled Cp185 from the SAXS profile under disassembly buffer. These data are consistent with a monodisperse solution of folded protein having an extended and/or flexible domain and are well-fit by a simulated SAXS profile (magenta line, χ^2 = 2.0) corresponding to the Cp185 dimer ensemble. Fitting residuals are shown in the lower panel. The Cp185 ensemble of structures selected *via* Bayesian inference from a library of models having reconstructed C-termini is shown, with monomers in red and yellow. The two alternative conformers are shown in red/yellow and magenta/orange and contribute to the SAXS profile with nearly equal populations: 0.46 (red/orange C-termini) and 0.54 (yellow/magenta C-termini). The experimental curve was analyzed to extract the radius of gyration, R_g, and the molecular weight, M_w. The R_g was estimated to be 37.2 ± 1.6 Å. The molecular weights estimated using absolute scale-independent methods (45.7 ± 6.9 kDa from SAXSMoW[26] and 43.7 ± 6.6 kDa using the Porod-invariant method[27]) suggested that Cp185 exists as a dimer under these disassembly conditions.

Characterization of Assembled Cp185. Cp is not stable in its dimeric form at physiological conditions in the absence of high ionic strength provided by GuHCl or other ions.[16,28] An ionic strength corresponding to greater than 0.25 M NaCl is also required to solubilize the assembled capsids in the absence of RNA.[16] This requirement is attributed to screening of repulsive forces between strongly positively charged C-terminal domains inside the shell of the empty capsid. The binding of negatively charged RNA similarly leads to reduced repulsion and promotes intersubunit associations. *In vivo* a kinase is believed to solubilize dimers by binding to the ARD domain. The kinase then acts as a chaperone for the dimers prior to RNA encapsulation.[28] Nonphysiological conditions were therefore required to study the assembly pathway of full-length Cp. To develop an efficient self-assembly protocol suitable for stopped-flow TR-SAXS, dilution experiments were performed in which Cp185 under disassembly conditions (1.5 M GuHCl and 0.5 M LiCl to reduce nonspecific interaction with nucleic acids) was manually mixed 1:1 with assembly buffer (no GuHCl and 0.25 M NaCl to maintain solubility) and incubated for >1 h. SEC showed that capsid formation was essentially complete under these conditions, with ∼88% mass found in the capsid peak and no significant fraction of stable aggregates (Supplementary Figure 4). The mixture was further analyzed by SAXS, and scattering results were consistent with a spherical particle, having a maximum diameter (D_{max}) of 340 ± 10 Å, which is consistent with the 350 Å maximum diameter of Cp149 particles (Supplementary Figure 5).[24]

How does the presence of pgRNA affect the assembly pathway of HBV particles? It has been demonstrated that pgRNA contains preferred sites (PSs) that function as packaging signals.[17] The PSs are regions of pgRNA that bind preferentially to the Cp185 with high affinity and specificity. Patel *et al.* identified three preferred sites in pgRNA constituting around 30 nucleotides which were predicted to form stem-loops with an RGAG recognition motif.[17] Isolated PS sequences were shown to induce capsid formation, and the resulting capsid structures determined in the presence of a PS oligo by cryo-electron microscopy (cryo-EM) demonstrated that the RNA was bound below the 5-fold vertices of the T = 4 capsid. In this study, a synthetic 40-nucleotide-long RNA oligo encompassing one of the identified PS sequences was used in the capsid assembly experiments (Figure 1). Addition of this RNA oligo to Cp185 dimers in disassembly buffer immediately triggered capsid formation. For controlled assembly experiments the RNA was therefore added to the assembly buffer, resulting in a molar ratio between Cp185 and the PS RNA oligo of approximately 6:1 after mixing. SEC results were similar to the results in the absence of RNA, with an almost (∼89%) complete conversion to capsids. SAXS data from the reassembled nucleocapsid showed similar features to the empty capsids, with an estimated D_{max} of 330 Å (Supplementary Table 7).

The assembled capsids were also studied by single-particle cryo-EM after purification with SEC. Analysis of particle class averages showed that the core particles assembled into primarily T = 4 capsids (87.5%) with a smaller fraction of the T = 3 state (12.5%). We determined the structure of the empty and RNA-filled T = 4 capsids to a resolution of 5.1 and 7.5 Å, respectively. The capsid structures were consistent with prior models of full-length core protein capsids (Figure 3 and Supplementary Table 5).[29] In the nucleocapsid the RNA was readily identified as an additional shell of density in the interior of the capsid.

Stopped-Flow Rapid-Mixing Assembly of Cp185. To investigate the detailed assembly mechanism of Cp185 in the presence and absence of RNA (PS RNA oligo) we carried out TR-SAXS measurements using a stopped-flow mixing cell. One

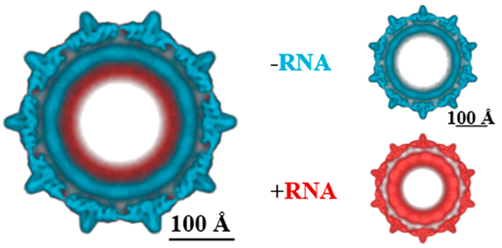


Figure 3. Comparison of cryo-EM particle reconstruction of Cp185 reassembled in the presence (red) and absence (blue) of complementary RNA. A central cross-section of each capsid is depicted on the right side and shown overlaid on the left.

mixing chamber contained Cp185 in disassembly buffer, while the second chamber contained assembly buffer with or without PS RNA oligo. After rapid mixing the sample was analyzed by SAXS at various time intervals to interrogate the oligomer distribution in the mixture. In the RNA assembly reaction, the final mixture contained a ratio of 1 PS RNA oligo per ~6.5 Cp185 monomers.

Twenty frames were collected for the RNA-free assembly reaction, which covered a total observation time of 58 s (Figure 4 and Supplementary Tables 2–4). The SAXS spectrum from the earliest time frame in the measurement series at 60 ms was similar to that of the Cp185 dimer in the disassembly buffer (Figure 6). After 10 s the spectra remained essentially unchanged, suggesting that the assembly reaction was complete at this point. Comparison of the spectrum at the final time frame of 58 s with the steady-state SAXS spectrum collected on capsids assembled by manual mixing (without further purification) showed a high degree of similarity (Figure 6), indicating stopped-flow mixing produced the same capsids as under manual mixing conditions. No signs of aggregation were observed during the assembly. In Figure 5 progression of the assembly reaction is illustrated by a plot of the apparent (z-averaged[30]) R_g and the apparent D_{max} from the SAXS spectrum of each frame. Both metrics showed similar sigmoidal time dependence with a half-time of around 1s and a complete assembly after 10 s. The apparent D_{max} plateaued at a value of 330 Å, which was consistent with the expected diameter of the capsid (350 Å). The maximum apparent D_{max} value was reached slightly before the maximum apparent R_g. The

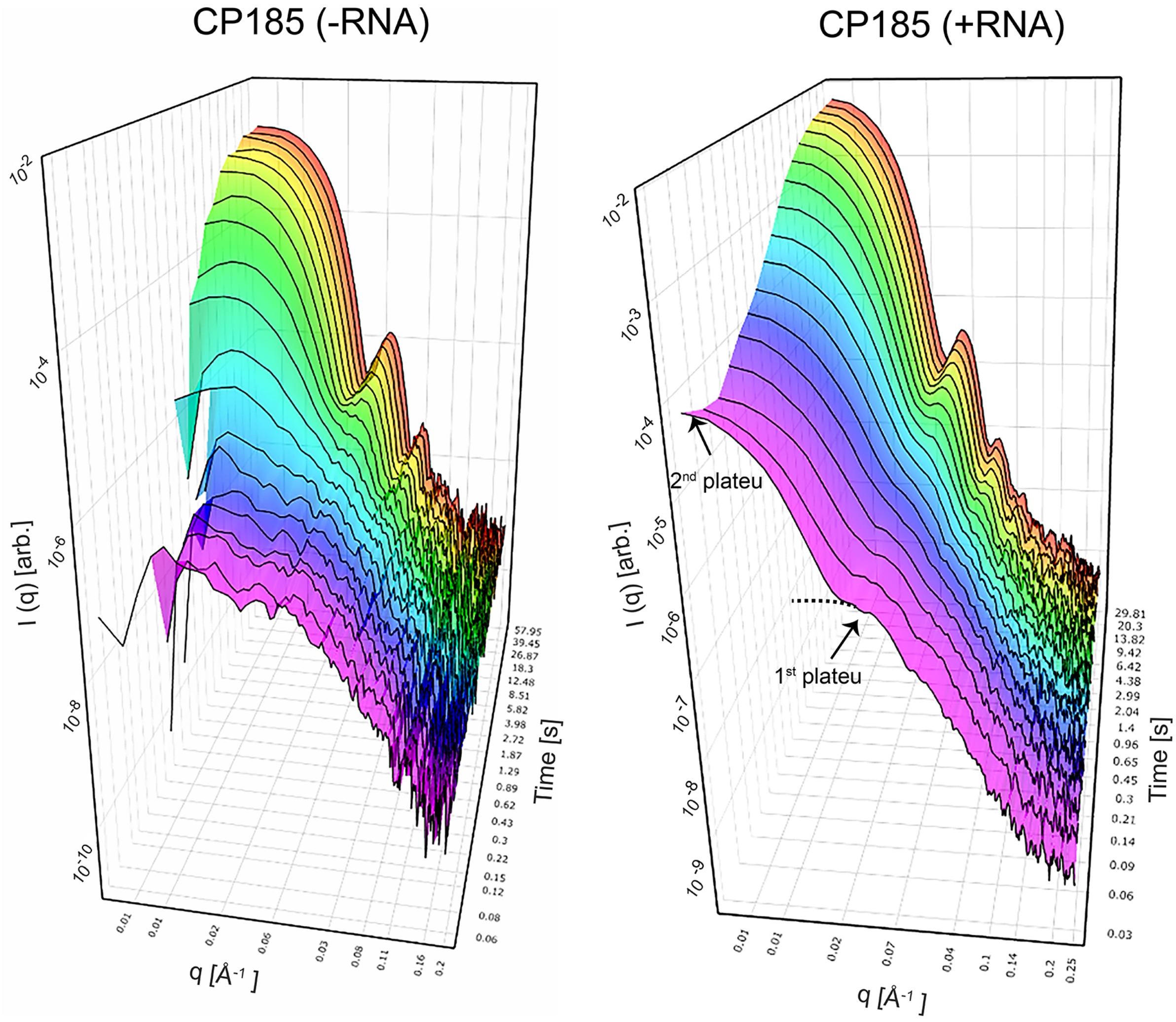


Figure 4. Time-dependent scattering of Cp185 in the presence and absence of RNA. Left, SAXS spectra collected at various time points after stopped-flow mixing of Cp185 in disassembly buffer with assembly buffer. Intensity $I(q)$ as a function of scattering vector q and time. Right, TR-SAXS of Cp185 in the presence of PS RNA oligo. The presence of two plateaus in the scattering profile in the presence of RNA is indicated in the figure. The dotted line is for illustrative purposes only.

https://dx.doi.org/10.1021/acsnano.0c03569
ACS Nano 2020, 14, 10226−10238

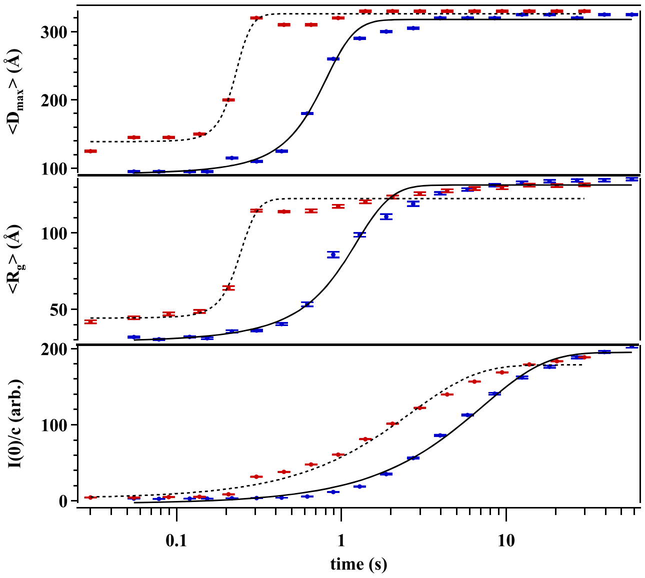


Figure 5. Observed rate changes derived from TR-SAXS measurements after triggered self-assembly of HBV Cp185 without RNA present (blue) and in the presence of RNA (red). Inflection points from sigmoidal fits, without RNA: D_{max}, 698 ms; R_g, 956 ms; $I(0)/c$, 5.07 s; and with RNA: D_{max}, 225 ms; R_g, 238 ms; $I(0)/c$, 3.45 s.

intensity extrapolated to zero angle, $I(0)$, is proportional to the mass of the protein sample and is mainly sensitive to the formation of high-molecular-weight species such as the T = 4 capsid. The midpoint of the transition observed for $I(0)$ occurred considerably later than the midpoints for R_g and D_{max}. If the assembly reaction was described by a transition between only two states, dimer−capsid as attributed to Cp149, a common midpoint between minimal and maximal values for R_g, D_{max}, and $I(0)$ would have been observed. This was not the case, suggesting that the assembly proceeds through some highly populated intermediate states between the initial-state dimers and final-state T = 4/T = 3 capsids. A similar picture emerged when the intensities were plotted as a function of time at other q-values, which report on the evolution of different length scales during the assembly (Supplementary Figure 6a). As described by Wallimann *et al.*,[31] intensities measured at two wavelengths should obey a linear relation for a two-state system, which was not observed here (Supplementary Figure 6b).

Assembly reactions were also carried out under similar conditions but with PS RNA oligo present in the assembly buffer. Steady-state SAXS data collected on PS RNA in assembly buffer demonstrated that the initial state of the RNA was a higher order assembly (Supplementary Figure 7). Eighteen time frames were collected for the RNA series extending up to 30 s. Up to the third time point, a decrease in intensity at low angles was observed. One possible interpretation of this feature was the presence of higher order RNA structures in the assembly buffer that rapidly disassembled after the initial mixing. The early spectra were highly similar to those recorded in the absence of RNA and for the steady-state dimer for q greater than ∼0.02 $Å^{-1}$ (Figures 4 and 6). A second, plateauing feature was pronounced at lower scattering angles with a second valid Guinier fit region suggesting a small and discrete population of larger ($R_g \approx$ 100 Å) particles (Figure 4). While this feature may be attributed to preformed capsids, we suggest that it rather corresponded to scattering from RNA based on the aforementioned observation of large RNA assemblies in the assembly buffer and decrease in forward scattering measurements over the first three frames. The small R_g values were similar to the initial assembly frames in the absence of RNA, as shown in Figure 5. Plateaus in D_{max}, R_g, and $I(0)/c$ before the sigmoidal transition indicated that the binding of RNA to capsid proteins occurred very early on in the trajectory, possibly during the delay time of the stopped-flow apparatus. Overall, the TR-SAXS spectra in the presence of RNA appeared very similar to the ones collected in the absence of RNA (Figure 4). The rate of assembly was higher compared to the RNA-free experiment (Figure 5 and Supplementary Tables 6 and 7), with midpoints between minimum and maximum values of R_g, D_{max}, and $I(0)/c$ occurring earlier. This can be consistently rationalized by the slightly higher protein concentration used in the RNA experiment, which increases the assembly rate and shortens assembly halftimes.[6,9]

Identification of Intermediates States with Multivariate Curve Resolution. SAXS spectra collected during the assembly process can be described as a linear combination of signals emanating from the species present in solution. To identify intermediate states, the collection of spectra must be deconvoluted into basis spectra that can be assigned to a single species, or collection of species, emerging at the same time points. A general and flexible method for deconvoluting time-series spectra is multivariate curve resolution. In multivariate curve resolution (MCR) a set of spectra at different time points are described by a matrix **D** (with intensities as a function of time stored as row vectors), which can be deconvoluted as

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where **C** is a matrix of concentration profiles, **S** is a matrix of basis spectra, and **E** is a matrix of residuals corresponding to uncertainties of the model. In contrast to singular-value decomposition analysis the basis components are not required to be orthogonal, which is rarely the case. MCR has been used previously to analyze time-resolved small- and wide-angle scattering data.[32,33] The decomposition in eq 1 does not produce a unique solution because of rotational ambiguities (any rotation applied to both C and S produces a new valid solution). Additional constraints, such as assumptions about non-negativity, unimodality, closeness, *etc.*, must therefore be applied. Furthermore, many alternative strategies exist for iterative optimization of the matrices **C** and **S** that can strongly influence the results. In order to develop an optimal analysis strategy for this set of TR-SAXS data, scattering profiles were simulated with noise and evaluated for the ability of MCR to recover the correct concentration and spectral matrices. The best performing analysis strategy involved the use of weighted alternating least-squares (MCR-WALS) with inverse standard errors from the experimental data as weights during the ALS.[34] A fundamental assumption in any MCR analysis is that errors for individual data points are independent. However, the highly covarying errors at low q-values were found to have a detrimental effect, which was resolved by removing such points in the analysis. Constraints of non-negativity on concentrations and spectra were also applied to reduce rotational ambiguity. The iterative process of finding **C** and **S** was initialized with a guess for the basis spectra through an orthogonal projection approach (OPA), modified to recover the correct intensity scale. This approach was shown to resolve

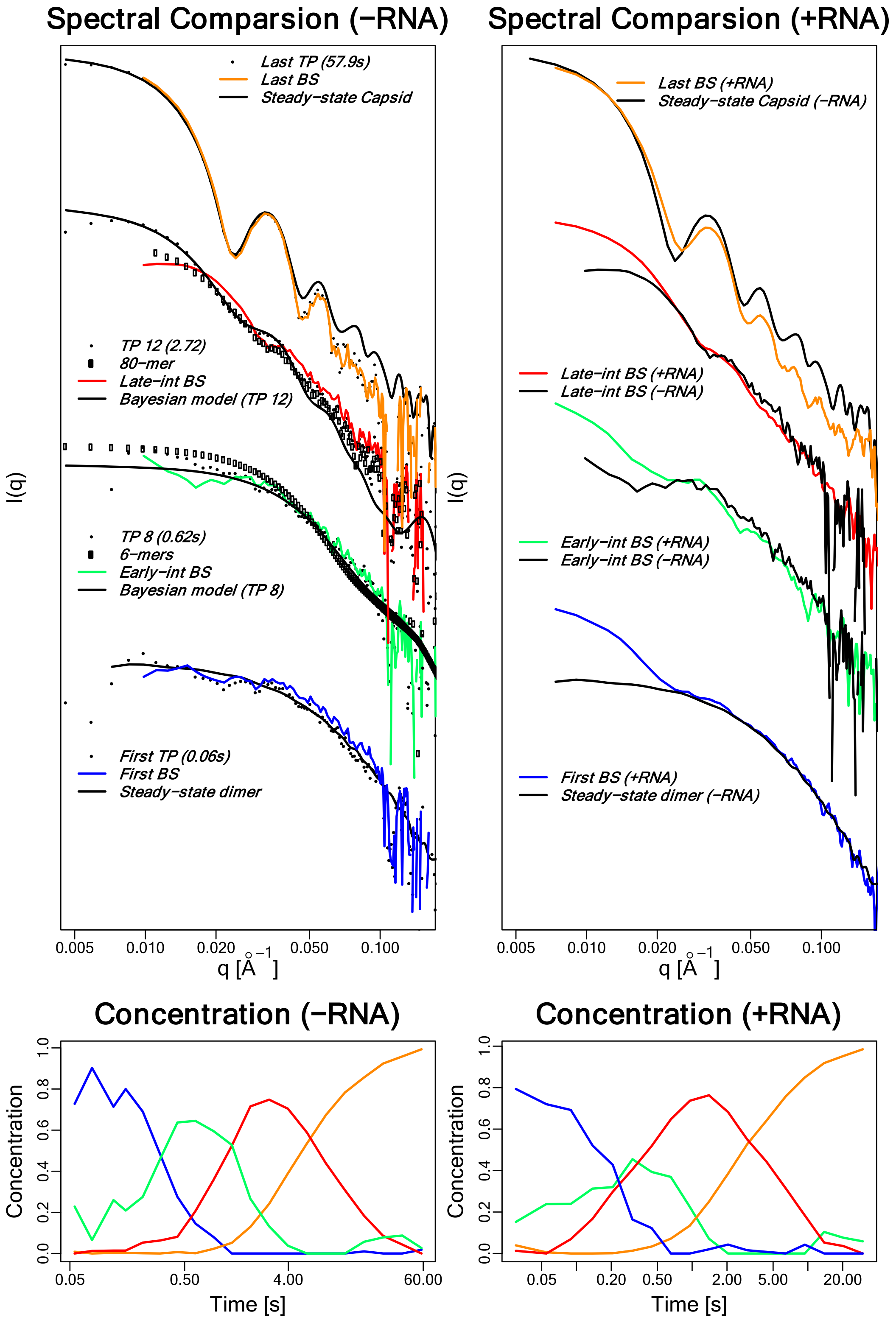


Figure 6. Reconstruction of species basis spectra using a four-component MCR-WALS model without (left) and with RNA (right). Lower panels, predicted concentration profiles for initial (blue), early intermediate (green), late intermediate (red), and final state (orange) in log of the time scale. Upper left panel, comparison of first basis spectrum (First BS) with first time point (First TP), early intermediate basis spectrum (Early int BS) with TP8, late intermediate basis spectrum (Late-int BS) with TP 12, and last basis spectrum (Last BS) with last time point (Last TP). TP8 and TP12 are the time frames at the predicted concentration maximum of the intermediates. For the right RNA panel, the basis spectra are compared to the RNA-free steady-state spectra (First and Last BS) and intermediate basis spectra. SAXS profiles derived from structural modeling are also shown in the RNA-free panel (results described in the structural modeling section). A Bayesian assembly model fit to TP8 (Bayesian model) and a weighted ensemble of two dimers of trimers extracted from $T = 3$ and $T = 4$ capsids is shown (6-mers). For the late intermediate the Bayesian assembly model fit to TP12 (Bayesian model) and a calculated spectrum for an 80-mer identified in the fitting are shown.

mixtures of three species in the synthetic data set with very high accuracy, in terms of both spectra and concentration profiles (Supplementary Figure 8). When the number of components was increased to four and the signal-to-noise increased, the prediction of basis spectra remained highly accurate, but small biases in the concentration profiles could occur (Supplementary Figure 9).

Identification of Intermediate State in the Cp185 Assembly. The MCR-WALS pipeline was first applied to the RNA-free TR-SAXS data. In the experimental data set, the number of components was unknown and had to be estimated from the data. A Scree plot typically used in principle component analysis (PCA) and the Cattell−Nelson−Gorsuch (CNG) index[35] both suggested that the data could be best described by a model of three components. This was consistent with the quality analysis of the MCR model fit to the experimental data as a function of number of components. A two-component model inaccurately reconstructs the experimental signal with significant deviations at intermediate time frames (Supplementary Figure 10). With three or more components the fit to the data was excellent, with low reduced χ^2 values (average 0.9) relative to all SAXS frames (Figure 6 and Supplementary Figure 10). However, PCA and the CNG index can often result in underestimation of the number of components. For example, analysis of our simulated four-component time series (Supplementary Figure 9) would suggest three rather than four components. When the number of components was increased to four, an additional early intermediate state was resolved, while the three other states remained highly similar to those identified in the three-component model. The profile of the early intermediate was also consistent with our prior expectations; the concentration was low at the beginning and the concentration profile was unimodal without making those assumptions in the analysis. Taken together, we suggest that the early intermediate represented a true intermediate, and given that it appeared early on in the assembly it could represent an assembly nucleus.

As expected, the components that appeared first and last during the assembly were very similar to the SAXS spectra collected before and after assembly and thus correspond to the dimer and capsid states (Figure 6). The concentration profiles indicate that the initial dimer population was converted into an early intermediate ensemble that had a maximum around 0.4 s. The absolute-scaling-independent M_w at this time frame was around 2.8 dimers. A late intermediate ensemble appeared next with a maximum at around 3 s. This late intermediate disappeared as the population of the capsid state increased. At the predicted concentration maximum R_g was ∼120 Å and D_{max} was estimated to be 305 Å. Scale-independent M_w determinations suggested the presence of oligomers with an average around 30 dimers (Supplementary Table 6).

Identification of Intermediate State in Cp185 Assembly in the Presence of RNA. Next, the MCR-WALS approach was used to assess the assembly trajectory of Cp185 in the presence of PS RNA oligo. With three or more components the fit to the data was excellent, with low reduced χ^2 values (average 1.0) relative to all SAXS frames (Figure 6 and Supplementary Figure 11). In the four-component model an early intermediate was observed with a maximum around 0.3 s, while a late intermediate emerged with a maximum around 1.4 s. Basis spectra with and without RNA appeared highly similar at q-values above 0.020 Å^{-1}, but differed at lower q-values (Figure 6). This difference was attributed to RNA binding to Cp185. One additional difference was that the concentration profile of the late intermediate had increased overlap with the first basis component in the presence of RNA. Consequently, the late intermediate appeared earlier on in the assembly reaction. Nonetheless, the overall similarities in concentration profiles and basis spectra indicated that the assembly passed through the same intermediate states in the presence and absence of RNA.

Structure-Based Modeling of Self-Assembly Pathway. One aim of this study was to attain a detailed structural understanding of the Cp185 assembly. This remains a challenge for a number of reasons. A TR-SAXS profile represents rotationally averaged mixture of states, encodes a low number of free parameters,[36] and has a lower signal-to-noise ratio than steady-state SAXS data. Care must therefore be taken to avoid overfitting during data analysis, and additional information must be provided to obtain a structural model of the assembly pathway. Bayesian statistical methods are perfectly suited to balance model complexity and fit in these scenarios. To simplify our analysis, capsid assembly was assumed to progress on the pathway that exclusively sampled the icosahedral lattice as observed in the crystal structures of Cp149 T = 3[37] and T = 4 HBV.[24] Models of intermediates were constructed by consecutive removal of dimers or trimers from the full capsid models using a graph-based algorithm. Furthermore, each monomer was modeled with a C-terminal extension identified in the ensemble inference of the steady-state dimer SAXS spectrum (Figure 2). To initially characterize the RNA-free four-state model from MCR-WALS analysis, structural ensembles were inferred using experimental data from the SAXS frames where the concentration of the intermediate reached a plateau (time points 7 and 8 for the early intermediate and time points 11, 12, and 13 for the late intermediate). Structural ensembles were determined from a library of 360 states extracted from the T = 3 and T = 4 capsid crystal structures containing the modeled C-terminal extensions. Each state was deemed equally probable during the initial Bayesian inference, and Bayesian model selection was used to derive the ensemble from experimental SAXS data and the library of structural models.[25] Ensembles fit to time points 7 and 8 corresponding to the early intermediates agreed well with the data (χ^2 1.9 and 1.6) and consisted of tetramers for TP7 and a mixture of trimers and hexamers for TP8. The ensembles for the late intermediate (time points 11−13) fitted well to the experimental data (χ^2 ranges from 2.1 to 2.5) and consisted of three models, which differed considerably at the structural level. For TP11 the ensemble consisted of models with 39, 48, and 50 monomers, for TP12 of 72, 74, and 76 monomers, and for TP13 of 84, 96, and 120 monomers. These results suggested that the intermediate forms an ensemble of large oligomers, but the lack of common components highlights the difficulty of extracting unique ensembles from individual noisy TR-SAXS frames. The same challenges were found in the analysis of the early intermediate. By fitting structural ensemble models to individual SAXS data frames independently, valuable information was lost. The appearance of a species at one time point is deemed more likely to occur if it was found at nearby time points (for example, the ensembles at time points 11, 12, and 13 should have contained similar species). Consequently, each species was expected to form and be replaced in a smooth and continuous manner. Additionally, when all SAXS frames were independently analyzed, the concentration of individual species fluctuated in an unrealistic manner (see Supplementary Figure 12), which was particularly apparent for early time points with lower signal-to-noise ratio.

To address these issues, we extended a Bayesian ensemble inference[25] method developed for steady-state data to describe how the ensembles evolved over time. In a Bayesian framework, we could bias the prior probability of a species appearing at a given time point by the probability that it was present during the last time point. Starting from the first time point the prior propagates from one frame to another until reaching the last frame. The result was that all time points were coupled together without the need for any underlying mechanistic model of the assembly process. This naive version has one major drawback: the end result of the ensemble predictions was biased by significant noise in the early time points. To reduce this bias, independent predictions for the population of individual species were collected at all time points. Thus, a smoothed line representing the time evolution of each population weight was fit along the trajectory using a Gaussian process model (Supplementary Figure 13). The benefit of using a Gaussian process for this purpose was threefold. First, it defined the extent of influence one time point has on another by choosing a covariance function; second, it acknowledged uncertainties of the inferred population weights; and last it provided confidence intervals to be assigned. The population weights from these smoothed lines were then used as priors in the ensemble prediction. This process enabled the prediction of a smooth and continuous concentration profile for each species, where population weight priors were influenced by predictions at all time points rather than a single experimental data frame. Ensemble prediction followed by population weight smoothing was then repeated until reaching convergence. The resulting time-dependent ensemble model reflects the number of species that can be supported by the data and identifies major species. The true species distribution at any given time point during the trajectory is likely more complex, but more data (experimental or additional modeling assumptions) would be required to support such a model.

Structural Model of the Cp185 Assembly Pathway. We applied this coupled Bayesian ensemble fitting model to the empty capsid assembly data using the conformational ensemble of 360 states extracted from the T = 3 and T = 4 capsid crystal structures and with the modeled C-terminal extensions. Because analysis of the entire landscape of population weights of 360 structures over the entire trajectory was computationally intractable, the structural library was limited to structures identified when the time frames were independently analyzed. This restriction was supported by observations that intermediates appeared to be on path with no aggregation. Using the resulting structural library of 35 models (out of 360), a structural ensemble was determined for each time frame with concentration profiles of species modeled by Gaussian process regression where the parameters of the covariance function and uncertainty of inferred population weights controlled the smoothness of the line. The estimate from the smoothed line as prior for the next iteration was used to repeat this process until reaching convergence. The coupling process did not significantly influence the fit to the data as judged by χ^2 values for individual time points. The model showed a good agreement with experimental data with χ^2 of individual data frames ranging from 2.0 to 4.8. The increase in χ^2 at later time points most likely reflects limitations in the modeling of the C-terminal extension. The C-terminal extensions should form an inner shell of the Cp149 capsid. This inner shell could be partly modeled by the smaller T = 3 capsid, which is likely why the portion of T = 3 particles is overestimated at the final time frame compared to the fraction observed in the cryo-EM data.

The inferred pathway (Figure 7) suggests that a mixture of dimers and hexamers existed at the early time points (up to 2 s), an intermediate state with 80 subunits (Sp_{80}) between 2 and 40 s, and capsid-like structures at the end (mixture of T = 3, T = 4). The mixture of dimers is at its maximum at the first time point (TP1) with a population weight of 0.65 ± 0.13 (combined from two conformations). However, the early intermediate was dominated by a mixture of trimer of dimers extracted from T = 3 and T = 4 capsids peaking at TP8 with a population weight of 0.74 ± 0.14. This mixture fits well to the

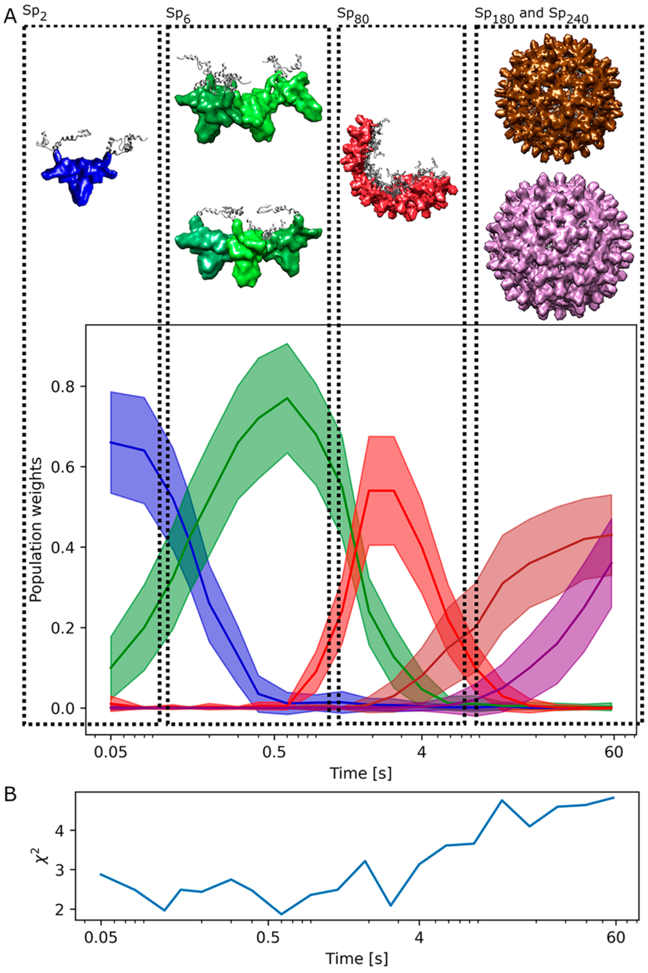


Figure 7. Predicted self-assembly pathway of Cp185 in the absence of RNA. (A) Predicted concentration profiles as a function of time of modeled species with a population above 20% at some time point during the trajectory. Shaded areas illustrate uncertainty in predicting population weights using Bayesian inference. Structural models are shown with colors matching those in the concentration profile. Dimer (Sp_2) and intermediate with 80 subunits (Sp_{80}) are shown with two alternative C-terminal extensions and are superimposed on residues 1−149, while two presented hexamer (Sp_6) models were extracted from T = 3 (top) and T = 4 (bottom) capsids. (B) χ^2 values between predicted spectra from the Bayesian assembly model and the experimental SAXS data.

early intermediate basis spectrum identified in the four-state MCR-WALS analysis (Figure 6). The 80-mer intermediate adopts a partially closed shell, with two almost equally contributing C-terminal extensions. As the dominant species for time points 11−13 (maximum population weight 0.55 ± 0.14 at TP12), it might well correspond to the late intermediate predicted by the MCR-WALS four-state model. Thus, a comparison was made between the obtained basis spectrum and the scattering profile calculated from the model of the 80-mer. Even though deviations occurred in the high q-region in the calculated curve, the overall profile of 80-mer agreed quite well with the MCR-WALS basis spectrum (Figure 6). This level of similarity was unexpected because the basis spectrum likely reflected an ensemble of species, rather than a single structure. The mixture of capsid-like structures reaches a maximum at the last frame (TP20) with population weights of 0.44 ± 0.09 and 0.36 ± 0.11 for Sp_{180} and Sp_{240}, respectively.

Intermediates were rapidly converted into higher order species without any indication of aggregation, which suggests that they were on pathway. Formation of intermediates during capsid assembly is commonly attributed to overnucleation, a phenomenon that can be observed in both kinetic simulations and assembly experiments. Overnucleation occurs in conditions where nucleation is favored relative to elongation, resulting in kinetically trapped intermediates due to a shortage of free capsomers required for elongation.[6,38,39] Induction of a Cp149 assembly with high concentrations of salt[11] or zinc[40] or in the presence of small-molecule assembly modifiers[38] results in overnucleation and formation of trapped intermediates believed to be on pathway but kinetically stable. These intermediates will eventually convert into fully formed capsids during a slower equilibration process (sometimes over days). This phenomenon is very different from our results on Cp185, suggesting that overnucleation was not the mechanism behind the formation of intermediates in this system. The results, however, are consistent with observations from TR-SAXS measurements on cowpea chlorotic mottle virus (CCMV), where an intermediate with the size of a half-capsid was inferred from the data.[20]

CONCLUSIONS

Our kinetic X-ray scattering data coupled to model-independent multivariate curve resolution analysis showed that full-length hepatitis B core protein Cp185 assembled through a mechanism that involved two highly populated intermediate states. Intermediates were transient, on pathway and rapidly converted into higher order species without formation of kinetic traps. By extending a Bayesian ensemble modeling method to time-domain data, the evolution of a structural ensemble was inferred without assumptions of a specific kinetic model. The structural model suggests that Cp185 assembles through an early intermediate largely composed of trimers of dimers, followed by a late intermediate ensemble dominated by a semiclosed capsid shape consisting of 40 dimers. This structure-based assembly mechanism was in strong agreement with the model-independent MCR-WALS analysis. By including RNA oligos in the assembly reaction, the influence of RNA binding on the assembly mechanism was investigated. In the presence of an RNA oligo binding specifically to the ARD domain of Cp the overall assembly mechanism was not significantly altered. However, the presence of RNA induced assembly in conditions where the RNA-free Cp was disassembled, suggesting that RNA binding can function as a trigger for capsid assembly.

The assembly mechanism of Cp185 shared common features with that of Cp149: in particular the formation of trimer of dimers that have been suggested to serve as a nucleus for the assembly of Cp149 capsids[6] and observations of a highly populated intermediate state during Cp185 capsid formation. Nonetheless, the overall assembly mechanism of Cp185 capsids was considerably different. Cp149 is well described by a two-state process, proceeding from dimers to capsid through a series of lowly populated intermediate states. Although aggressive assembly conditions involving high ionic strength have been shown to result in accumulation of intermediates, they had a broad size distribution and were slowly converted into capsids during the assembly process.[23] The presence of the flexible ARD thus has a significant impact on the capsid assembly process, highlighting the importance of further investigations as to how ARD modulates HBV capsid assembly.

The *in vivo* capsid assembly process occurs under conditions that are distinctly different from the ones used for *in vitro* experiments. Yet the *in vitro* self-assembly processes studied here described rapid kinetic trajectories that very efficiently converted capsomers into fully formed capsids with on-path intermediates. The *in vitro* assembly therefore represents a thermodynamically and kinetically accessible pathway that should be of relevance in understanding the more complex assembly scenario *in vivo*.

METHODS

Expression and Purification of HBV Cp185. The sequence of full-length Cp (Cp185) was based on a Cp variant with 185 residues described by Patel *et al.*[17] in their assessment of preferential binding to capsid core protein using different RNA structural motifs found within the viral pgRNA (Supplementary Figure 1). A modified plasmid (Genscript) containing an expression vector for the Cp185 sequence was prepared in *E. coli* Tuner (DE3) cells (Novagen) using standard transformation protocols. Expression conditions (time, temperature, and IPTG concentration) and the addition of detergent during cell lysis were investigated to optimize recovery of soluble Cp185. SDS-PAGE gels confirmed the purity of the samples and absence of insoluble aggregates (Supplementary Figure 2). Optimal yields were obtained when cells were incubated in Luria broth (LB) media supplemented with kanamycin antibiotic selection and induced with a final concentration of 1 mM IPTG overnight at 18 °C. This recombinantly expressed HBV Cp185 product was extracted and isolated following purification methods described by Patel *et al.*[17] with a few modifications.

Following the overnight induction period, cells were pelleted *via* centrifugation and the pellet was frozen at −20 °C after harvesting. Cell pellets were thawed before lysis and resuspended with lysis buffer consisting of 250 mM NaCl, 5 mM DTT, and 50 mM HEPES, pH 7.5, and freshly supplemented with protease inhibitor cocktail (Roche). All purification steps were performed on ice or at 4 °C. Cells were lysed *via* French press, and cellular debris removed *via* centrifugation. The supernatant was then centrifuged at ultrahigh speed overnight (120000g for ∼16 h) to pellet the soluble VLPs. After resuspending and homogenizing the pelleted VLPs in fresh lysis buffer, $(NH_4)_2SO_4$ was slowly added to a final concentration of 20% w/v to precipitate the HBV Cp185 proteins. Resuspension and dialysis of the resulting precipitate was performed using disassembly buffer consisting of 2 M GuHCl, 500 mM LiCl, 250 mM NaCl, 50 mM HEPES, pH 7.5, and 2 mM DTT and employing a two-stage dialysis over 48 h. After removal of the final dialysis precipitates *via* centrifugation, the supernatant was syringe-filtered at 0.2 μm and separated using a Superdex 200 Increase 10/300 GL column (GE

Healthcare), equilibrated with disassembled-state storage buffer containing 1.5 M GuHCl, 0.5 M LiCl, 50 mM HEPES, pH 7.5, and 2 mM DTT. Fractions with a column elution volume between 15.5 and 16.5 mL corresponding to disassembled-state Cp185 were collected for characterization and assembly reactions (Supplementary Figure 4). The appearance of a single dominant band in SDS-PAGE gels (Supplementary Figure 2) indicated high protein purity, and 260/280 nm absorbance ratios (∼0.6) after SEC purification implied negligible contamination by nucleic acids. The recovered volume corresponds to a molecular weight (M_w) range of 20−45 kDa, consistent with the expected M_w of ∼43 kDa for the Cp185 dimer. Purified Cp185 can be stored in this disassembled state for up to 4 weeks at 4 °C and remains assembly competent with negligible preformed capsids.

Preparation of HBV Cp185 Complementary-Binding RNA. A synthetic RNA oligo was obtained from Integrated DNA Technologies, Inc. with a sequence corresponding to the HBV Cp binding motif of design PS1 described by Patel *et al.*[17] The 40 nt sequence employed here was 5′-ggguuuguuuaaagacugggaggaguugggggaggagccc-3′. The capsid assembly buffer contained 250 mM NaCl, 50 mM HEPES, pH 7.5, and 2 mM DTT. The final concentration of RNA used in TR-SAXS measurements was 0.075 mg mL^{-1}, representing a theoretical maximum binding of 1 RNA per 6.5 Cp185 monomer.

Time-Resolved Small-Angle X-ray Scattering Measurements. TR-SAXS experiments were performed at the ID02 beamline of the European Synchrotron Radiation Facility (Grenoble, France).[41,42] A Rayonix MX-170HS CCD detector with an active area of 170 × 170 mm (pixel size 44.2 μm) and 8 × 8 binning was used to record the 2D SAXS images. The X-ray energy and sample-to-detector distance were 12.4 keV and 1.7 m, respectively, providing a usable q-range from 0.008 to 0.45 Å^{-1}. Data acquisition was triggered by the stopped-flow device,[17] with exposure times of 5 ms each. Corrected 2D images were radially averaged to obtain the 1D scattering profiles, and scattering from matched buffer measurements subtracted to yield the net macromolecular scattering intensities. All intensities are plotted as a function of the angular momentum transfer (scattering) vector q, defined as $4\pi\ \sin(\theta)/\lambda$, where 2θ is the scattering angle.

For each set of exposures, equal volumes (200 μL) of HBV Cp185 (conditions with RNA: 1.57 mg mL^{-1}, without RNA: 0.92 mg mL^{-1}) in dimer storage buffer and capsid assembly buffer (±0.15 mg mL^{-1} RNA PS oligo) were mixed to initiate capsid assembly by a stopped-flow, rapid-mixing apparatus (Bio-Logic). The total mixing time was 60 ms. Initial and final conditions were also recorded using a standard flow-through capillary with 10 frames of 0.1 s exposure times. Steady-state (nonflowing) exposure of the sample using 20 frames of 5 ms exposures was similarly used to assess potential radiation damage during the stopped-flow measurements. No radiation damage was observed over these cumulative exposures by comparing subsequent frames.

Measurements of steady-state conditions—1.14 mg mL^{-1} disassembled protein in disassembly buffer, disassembled protein mixed 1:1 with assembly buffer >1 h after mixing, and RNA in assembly buffer—were also collected using a standard flow-cell setup. Multiple frames for each sample were averaged; then buffer subtracted. Outlier frames, such as bubble flowing through cell, were identified and excluded.

Analysis of Small-Angle Scattering Data. A Guinier approximation of the low-q data was performed to determine reciprocal-space values for the forward scattering, $I(0)$, and radius of gyration, R_g, using methods for standard compact globular particles having a qR_g upper limit defined by 1.3.[43] When appropriate based on measurement quality, AUTORG (ATSAS) was employed.[44] Calculations of equivalent real-space parameters for $I(0)$ and R_g, as well as a maximum particle dimension $D_{\max}$ and the pairwise distance distribution function $P(r)$, were obtained from an indirect Fourier transform over a broad angular range of the scattering data (GNOM).[44] Molecular weight determinations were carried out using two absolute-scaling-independent methods, using the Porod-invariant methods described by Rambo and Tainer[45] and Piiadov *et al*[26] from the scattering profiles (Scatter[46] and SAXSMoW[26]). Scatter was used to approximate average molecular weights (number of dimer subunits) in Supplementary Table 6. Kratky plots were prepared as $q^2I(q)$ *vs* q and served as a qualitative assessment for the particle folding. The relevant X-ray contrast values and partial specific volume for HBV Cp185, complementary RNA, and buffer solutions were determined from the primary structure using MuLCH (modules for the analysis of contrast variation data).[47] These physical properties are summarized in Supplementary Table 1a.

Capsid Particle Volume Reconstruction with Cryo-Electron Microscopy. Electron micrographs were collected using the Talos Arctica microscope with an FEI Falcon II detector at the SciLifeLab Swedish National Facility (Stockholm). Sample grids were treated with a glow discharge (20 mA for 60 s) before sample application, blotting, and plunge freezing using a Vitribot. Quantifoil 1.2/1.3 and 2.2 grid types were screened and used. Approximately 3 μL of sample was applied before blotting with zero force under high humidity. Data for the empty capsid condition were acquired with 92000× magnification, 1.61 Å/pixel, a defocus range of −3.0 to −1.5, and a total dose of 31 e$^-$ per Å^2. Approximately 10k particles were extracted from 900 micrographs. After 2D classification, 9k "capsid-like" particles were selected, and the structures were solved using Scipion[48] and Relion;[49] details are found in Supplementary Table 5.

Intermediate Basis Spectra Deconvolution with Multivariate Curve Resolution. One-thousand synthetic TR-SAXS data sets of assembly kinetics were simulated to identify an optimal method to preprocess data and to find suitable parameters for the MCR-WALS algorithm used in analysis of experimental data. In each data set three or four high-resolution SAXS profiles with q-values in the range [0, 0.5] Å^{-1} were simulated using Pepsi-SAXS[9] (default parameters except for -ns 300) and mixed with their correspondence concentration profiles over a 20 s time window. The concentration profiles were simulated using beta distributions with suitable shape and scale parameters. To simulate noise with a similar pattern and similar scale to experimental data, synthetic noise was added to each mixed component data using the method of Sedlak and Lipfert.[50] The final synthetic data set is generated by adding a stochastic term to the matrix of mixed components. To do so, first the intensity at each time point is normalized (divided by $I(0)$ and multiplied by a factor of 100) and then simulated from a normal distribution with mean $I(q)$ and variance $\sigma_I^2(q)$ such that

$$\sigma_I^2(q) = \frac{\tau}{kq}\left[I(q) + \frac{2c \times I(q_{\mathrm{arb}})}{(1-c)}\right]$$

where $\tau = 0.2$, $k = 4500$, $c = 0.9$, and $q_{\mathrm{arb}} = 0.2$ Å^{-1}.

The concentration matrix and basis spectra were identified by weighted non-negative least-squares for the bilinear equation of MCR-WALS by minimization of the difference between computed and experimental (or simulated) spectra. The minimization is iteratively run until both criteria, convergence threshold of root-mean-square error (10^{-3}) and minimum iterations (20), passed. Initial guesses for the basis spectra were derived using the orthogonal projection approach. OPA identifies the most dissimilar basis spectra in the data set. This initial guess was provided without rescaling the intensities. Constraints on the non-negativity of concentrations and intensities were further employed. MCR-WALS was able to decompose the true basis spectra when the first q-values with covariance greater than a certain threshold were removed from the data set. As a result, the true concentration profiles are reconstructed with little bias. The same strategy was also employed in the analysis of experimental data sets. Therefore, in TR-SAXS with (without) RNA, 1(4) q-values were not considered in the deconvolution of the intermediates. The MCR-WALS methodology was implemented in a shiny R web application, from which readers can reproduce simulation and experimental analysis interactively at http://shiny.andrelab.org/TR-SAXS/.

Ensemble Modeling with Rosetta and Bayesian Statistics. One-thousand different conformations of C-terminal domain extending from the Cp149 crystal structure (PDB code: 1QGT[24]) were generated with an all-atom Monte Carlo simulation in

Rosetta.[25,51] C_5 symmetry from the Cp149 crystal structure was applied during simulation. Monomer subunits were subsequently extracted from resulting pentamers, and models of dimers and tetramers were created by superimposing them on the crystal structure of Cp149. In order not to introduce symmetry bias that may occur if monomer subunits are identical, we combined different monomers. Therefore, the resulting dimers and tetramers are asymmetric. Finally, we removed clashing conformations, and as a result we obtained a structural library of 1396 models. This library was used to fit an ensemble to the SAXS data using a Bayesian model fitting algorithm.[25] As a result, we obtained the ensemble consisting of two dimers. These two conformations were subsequently used to generate models of intermediates based on T = 3 (PDB code: 6BVN[37]) and T = 4 (PDB code: 1QGT[24]) capsid scaffolds. In each iteration either a dimer (T = 4) or trimer (T = 3) was removed at the time. Connectivity of the remaining subunits was ensured after each iteration using a graph-based algorithm. The corresponding graphs were created for T = 3 and T = 4 capsids using centers of masses of building blocks (trimers or dimers) and connections to nearest neighbors. As a result, we obtained a structural library of 360 intermediates (120 models of T = 3 and 240 models for T = 4). We subsequently used an adapted version of a Bayesian model fitting algorithm[25] to fit ensembles to all time points. At the first step we used Variational Bayesian Inference, which allows for the selection of the smaller number of models representing SAXS data at each time point. At this stage we assumed that each intermediate is equally probable. For each time point we inferred more than one state, and the resulting 35 models from all time points were taken to the next step. At this stage we performed Bayesian inference using no-u-turn Hamiltonian Monte Carlo sampling as implemented in the stan probabilistic modeling library.[52] We ran the method iteratively, and after each iteration population weights were smoothed using Gaussian process regression as implemented in the scikit-learn python package.[53] We used standard error of the mean from inferred population weights to define uncertainty of input parameters for regression. We defined the covariance function using a combination of RBF (radial basis function) and White Kernel to provide a smooth but not too coarse solution. In the first iteration we biased prior to the first five time frames with weights obtained from the modeling of steady-state dimer, and otherwise we assumed states are equally probable. We used smoothed weights from Gaussian process regression to bias subsequent simulations. The code for the Bayesian inference method can be found at github.com/Andre-lab/bioce.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsnano.0c03569.

> Supporting figures showing protein sequence, protein purification results, static SAXS analysis, MCR-WALS simulations, MCR-WALS analysis, and reconstruction accuracy as a function of number of components, analysis of TR-SAXS data with the two-state model, result of Bayesian inference analysis of TR-SAXS data without time-point coupling and weight smoothing with Gaussian process regression; supporting tables for SAXS data collection and analysis, R_g and D_{max} at each time frame of SAXS, cryo-EM data collection and analysis (PDF)

AUTHOR INFORMATION

Corresponding Author

Ingemar André − *Department of Biochemistry and Structural Biology, Lund University, Lund, Sweden 22100;* orcid.org/0000-0002-4753-8233; Email: ingemar.andre@biochemistry.lu.se

Authors

Ryan C. Oliver − *Department of Biochemistry and Structural Biology, Lund University, Lund, Sweden 22100;* orcid.org/0000-0003-4489-0861

Wojciech Potrzebowski − *Data Management and Software Centre, European Spallation Source ERIC, 2200 Copenhagen, Denmark; Department of Biochemistry and Structural Biology, Lund University, Lund, Sweden 22100;* orcid.org/0000-0002-7789-6779

Seyed Morteza Najibi − *Department of Biochemistry and Structural Biology, Lund University, Lund, Sweden 22100*

Martin Nors Pedersen − *Niels Bohr Institute, Faculty of Science, University of Copenhagen, 2100 Copenhagen, Denmark*

Lise Arleth − *Niels Bohr Institute, Faculty of Science, University of Copenhagen, 2100 Copenhagen, Denmark*

Najet Mahmoudi − *ISIS Neutron and Muon Source, STFC Rutherford Appleton Laboratory, Chilton, Didcot OX11 0QX, U. K.*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsnano.0c03569

Author Contributions

[#]R. C. Oliver and W. Potrzebowski contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Proteins were prepared at the Lund Protein Production Platform (LP3) facility. The authors are grateful for the access and technical assistance, specifically to W.K., C.S., A.R., E.K., Z.F., and H.C. The small-angle X-ray scattering experiments were performed on beamline ID02 at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. We thank Thomas Zinn at the ESRF for providing assistance in using beamline ID02. We thank Enrico Ravera for suggesting the use of MCR for data analysis. Cryo-EM micrographs were collected at the SciLifeLab facility (Stockholm). The structure determination was enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at HPC2N partially funded by the Swedish Research Council through grant agreement no. 2019/3-140. R.O., W.P., N.M., and I.A. were supported by Swedish Research Council (https://www.vr.se/english.html) grant (2016-06947). R.O., W.P., and I.A. were funded by interreg Öresund-Kattegat-Skagerrak (LU-063). M.N. and I.A. were also supported by a grant from eSSENCE@LU.

REFERENCES

(1) Kondylis, P.; Schlicksup, C. J.; Katen, S. P.; Lee, L. S.; Zlotnick, A.; Jacobson, S. C. Evolution of Intermediates during Capsid Assembly of Hepatitis B Virus with Phenylpropenamide-Based Antivirals. *ACS Infect. Dis.* **2019**, *5*, 769−777.

(2) Hume, H. K. C.; Vidigal, J.; Carrondo, M. J. T.; Middelberg, A. P. J.; Roldão, A.; Lua, L. H. L. Synthetic Biology for Bioengineering Virus-Like Particle Vaccines. *Biotechnol. Bioeng.* **2019**, *116*, 919−935.

(3) Ding, X.; Liu, D.; Booth, G.; Gao, W.; Lu, Y. Virus-Like Particle Engineering: From Rational Design to Versatile Applications. *Biotechnol. J.* **2018**, *13*, 1700324.

(4) Perlmutter, J. D.; Hagan, M. F. Mechanisms of Virus Assembly. *Annu. Rev. Phys. Chem.* **2015**, *66*, 217−239.

(5) Stannard, L. M.; Hodgkiss, M. Morphological Irregularities in Dane Particle Cores. *J. Gen. Virol.* **1979**, *45*, 509−514.

(6) Zlotnick, A.; Johnson, J. M.; Wingfield, P. W.; Stahl, S. J.; Endres, D. A Theoretical Model Successfully Identifies Features of Hepatitis B Virus Capsid Assembly. *Biochemistry* **1999**, *38*, 14644−14652.

(7) Morozov, A. Y.; Bruinsma, R. F.; Rudnick, J. Assembly of Viruses and the Pseudo-Law of Mass Action. *J. Chem. Phys.* **2009**, *131*, 155101.

(8) Endres, D.; Zlotnick, A. Model-Based Analysis of Assembly Kinetics for Virus Capsids or Other Spherical Polymers. *Biophys. J.* **2002**, *83*, 1217−1230.

(9) Hagan, M. F.; Elrad, O. M. Understanding the Concentration Dependence of Viral Capsid Assembly Kinetics-The Origin of the Lag Time and Identifying the Critical Nucleus Size. *Biophys. J.* **2010**, *98*, 1065−1074.

(10) Lutomski, C. A.; Lyktey, N. A.; Zhao, Z.; Pierson, E. E.; Zlotnick, A.; Jarrold, M. F. Hepatitis B Virus Capsid Completion Occurs through Error Correction. *J. Am. Chem. Soc.* **2017**, *139*, 16932−16938.

(11) Lutomski, C. A.; Lyktey, N. A.; Pierson, E. E.; Zhao, Z.; Zlotnick, A.; Jarrold, M. F. Multiple Pathways in Capsid Assembly. *J. Am. Chem. Soc.* **2018**, *140*, 5784−5790.

(12) Pierson, E. E.; Keifer, D. Z.; Selzer, L.; Lee, L. S.; Contino, N. C.; Wang, J. C. Y.; Zlotnick, A.; Jarrold, M. F. Detection of Late Intermediates in Virus Capsid Assembly by Charge Detection Mass Spectrometry. *J. Am. Chem. Soc.* **2014**, *136*, 3536−3541.

(13) Uetrecht, C.; Barbu, I. M.; Shoemaker, G. K.; Duijn, E. v.; Heck, A. J. R. Interrogating Viral Capsid Assembly with Ion Mobility-Mass Spectrometry. *Nat. Chem.* **2011**, *3*, 126−132.

(14) Uetrecht, C.; Versluis, C.; Watts, N. R.; Roos, W. H.; Wuite, G. J.; Wingfield, P. T.; Steven, A. C.; Heck, A. J. High-Resolution Mass Spectrometry of Viral Assemblies: Molecular Composition and Stability of Dimorphic Hepatitis B Virus Capsids. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 9216−9220.

(15) Asor, R.; Selzer, L.; Schlicksup, C. J.; Zhao, Z.; Zlotnick, A.; Raviv, U. Assembly Reactions of Hepatitis B Capsid Protein into Capsid Nanoparticles Follow a Narrow Path through a Complex Reaction Landscape. *ACS Nano* **2019**, *13*, 7610−7626.

(16) Porterfield, J. Z.; Dhason, M. S.; Loeb, D. D.; Nassal, M.; Stray, S. J.; Zlotnick, A. Full-Length Hepatitis B Virus Core Protein Packages Viral and Heterologous RNA with Similarly High Levels of Cooperativity. *J. Virol.* **2010**, *84*, 7174−7184.

(17) Patel, N.; White, S. J.; Thompson, R. F.; Bingham, R.; Weiss, E. U.; Maskell, D. P.; Zlotnick, A.; Dykeman, E.; Tuma, R.; Twarock, R.; Ranson, N. A.; Stockley, P. G. HBV RNA Pre-Genome Encodes Specific Motifs That Mediate Interactions with the Viral Core Protein That Promote Nucleocapsid Assembly. *Nat. Microbiol* **2017**, *2*, 17098.

(18) Chevreuil, M.; Law-Hine, D.; Chen, J.; Bressanelli, S.; Combet, S.; Constantin, D.; Degrouard, J.; Moller, J.; Zeghal, M.; Tresset, G. Nonequilibrium Self-Assembly Dynamics of Icosahedral Viral Capsids Packaging Genome or Polyelectrolyte. *Nat. Commun.* **2018**, *9*, 3071.

(19) Kler, S.; Asor, R.; Li, C.; Ginsburg, A.; Harries, D.; Oppenheim, A.; Zlotnick, A.; Raviv, U. RNA Encapsidation by Sv40-Derived Nanoparticles Follows a Rapid Two-State Mechanism. *J. Am. Chem. Soc.* **2012**, *134*, 8823−8830.

(20) Law-Hine, D.; Zeghal, M.; Bressanelli, S.; Constantin, D.; Tresset, G. Identification of a Major Intermediate along the Self-Assembly Pathway of an Icosahedral Viral Capsid by Using an Analytical Model of a Spherical Patch. *Soft Matter* **2016**, *12*, 6728−6736.

(21) Tresset, G.; Le Coeur, C.; Bryche, J. F.; Tatou, M.; Zeghal, M.; Charpilienne, A.; Poncet, D.; Constantin, D.; Bressanelli, S. Norovirus Capsid Proteins Self-Assemble through Biphasic Kinetics *via* Long-Lived Stave-Like Intermediates. *J. Am. Chem. Soc.* **2013**, *135*, 15373−15381.

(22) Tuma, R.; Tsuruta, H.; French, K. H.; Prevelige, P. E. Detection of Intermediates and Kinetic Control during Assembly of Bacteriophage P22 Procapsid. *J. Mol. Biol.* **2008**, *381*, 1395−1406.

(23) Asor, R.; Schlicksup, C. J.; Zhao, Z.; Zlotnick, A.; Raviv, U. Rapidly Forming Early Intermediate Structures Dictate the Pathway of Capsid Assembly. *J. Am. Chem. Soc.* **2020**, *142*, 7868.

(24) Wynne, S. A.; Crowther, R. A.; Leslie, A. G. W. The Crystal Structure of the Human Hepatitis B Virus Capsid. *Mol. Cell* **1999**, *3*, 771−780.

(25) Potrzebowski, W.; Trewhella, J.; Andre, I. Bayesian Inference of Protein Conformational Ensembles from Limited Structural Data. *PLoS Comput. Biol.* **2018**, *14*, No. e1006641.

(26) Piiadov, V.; Araújo, E. A. d.; Neto, M. O.; Craievich, A. F.; Polikarpov, I. SAXSMOW 2.0: Online Calculator of the Molecular Weight of Proteins in Dilute Solution from Experimental SAXS Data Measured on a Relative Scale. *Protein Sci.* **2019**, *28*, 454−463.

(27) Rambo, R. P.; Tainer, J. A. Accurate Assessment of Mass, Models and Resolution by Small-Angle Scattering. *Nature* **2013**, *496*, 477−481.

(28) Chen, C.; Wang, J. C.-Y.; Zlotnick, A. A Kinase Chaperones Hepatitis B Virus Capsid Assembly and Captures Capsid Dynamics *In Vitro*. *PLoS Pathog.* **2011**, *7*, No. e1002388.

(29) Yu, X.; Jin, L.; Jih, J.; Shih, C.; Zhou, Z. H. 3.5Å Cryoem Structure of Hepatitis B Virus Core Assembled from Full-Length Core Protein. *PLoS One* **2013**, *8*, No. e69729.

(30) Feigin, L. A.; Svergun, D. I. *Structure Analysis by Small-Angle X-Ray and Neutron Scattering*; Plenum Press: New York, 1987.

(31) Wallimann, P.; Kennedy, R. J.; Miller, J. S.; Shalongo, W.; Kemp, D. S. Dual Wavelength Parametric Test of Two-State Models for Circular Dichroism Spectra of Helical Polypeptides: Anomalous Dichroic Properties of Alanine-Rich Peptides. *J. Am. Chem. Soc.* **2003**, *125*, 1203−1220.

(32) Herranz-Trillo, F.; Groenning, M.; Maarschalkerweerd, A. v.; Tauler, R.; Vestergaard, B.; Bernadó, P. Structural Analysis of Multi-Component Amyloid Systems by Chemometric SAXS Data Decomposition. *Structure* **2017**, *25*, 5−15.

(33) Minh, D. D. L.; Makowski, L. Wide-Angle X-Ray Solution Scattering for Protein-Ligand Binding: Multivariate Curve Resolution with Bayesian Confidence Intervals. *Biophys. J.* **2013**, *104*, 873−883.

(34) Wentzell, P. D.; Karakach, T. K.; Roy, S.; Martinez, M. J.; Allen, C. P.; Werner-Washburne, M. Multivariate Curve Resolution of Time Course Microarray Data. *BMC Bioinf.* **2006**, *7*, 343.

(35) Raiche, G.; Walls, T. A.; Magis, D.; Riopel, M.; Blais, J. G. Non-Graphical Solutions for Cattell's Scree Test. *Methodology-Eur.* **2013**, *9*, 23−29.

(36) Vestergaard, B.; Hansen, S. Application of Bayesian Analysis to Indirect Fourier Transformation in Small-Angle Scattering. *J. Appl. Crystallogr.* **2006**, *39*, 797−804.

(37) Schlicksup, C. J.; Wang, J. C.-Y.; Francis, S.; Venkatakrishnan, B.; Turner, W. W.; VanNieuwenhze, M.; Zlotnick, A. Hepatitis B Virus Core Protein Allosteric Modulators Can Distort and Disrupt Intact Capsids. *eLife* **2018**, *7*, No. e31473.

(38) Katen, S. P.; Chirapu, S. R.; Finn, M. G.; Zlotnick, A. Trapping of Hepatitis B Virus Capsid Assembly Intermediates by Phenylpropenamide Assembly Accelerators. *ACS Chem. Biol.* **2010**, *5*, 1125−1136.

(39) Zhang, T.; Schwartz, R. Simulation Study of the Contribution of Oligomer/Oligomer Binding to Capsid Assembly Kinetics. *Biophys. J.* **2006**, *90*, 57−64.

(40) Stray, S. J.; Ceres, P.; Zlotnick, A. Zinc Ions Trigger Conformational Change and Oligomerization of Hepatitis B Virus Capsid Protein. *Biochemistry* **2004**, *43*, 9989−9998.

(41) Narayanan, T.; Sztucki, M.; Van Vaerenbergh, P.; Leonardon, J.; Gorini, J.; Claustre, L.; Sever, F.; Morse, J.; Boesecke, P. A Multipurpose Instrument for Time-Resolved Ultra-Small-Angle and Coherent X-Ray Scattering. *J. Appl. Crystallogr.* **2018**, *51*, 1511−1524.

(42) Narayanan, T.; Gummel, J.; Gradzielski, M. Probing the Self-Assembly of Unilamellar Vesicles using Time-Resolved Saxs. *Adv. Planar Lipid Bilayers Liposomes* **2014**, *20*, 171−196.

(43) Glatter, O. Scattering Methods and Their Application in Colloid and Interface Science. *Scatt Methods Appl. Colloid Interface Sci.* **2018**, 19−32.

(44) Petoukhov, M. V.; Konarev, P. V.; Kikhney, A. G.; Svergun, D. I. ATSAS 2.1 − Towards Automated and Web-Supported Small-

Angle Scattering Data Analysis. *J. Appl. Crystallogr.* **2007**, *40*, s223−s228.

(45) Rambo, R. P.; Tainer, J. A. Characterizing Flexible and Intrinsically Unstructured Biological Macromolecules by SAS Using the Porod-Debye Law. *Biopolymers* **2011**, *95*, 559−571.

(46) Förster, S.; Apostol, L.; Bras, W. Scatter: Software for the Analysis of Nano- and Mesoscale Small-Angle Scattering. *J. Appl. Crystallogr.* **2010**, *43*, 639−646.

(47) Whitten, A. E.; Cai, S.; Trewhella, J. Mulch: Modules for the Analysis of Small-Angle Neutron Contrast Variation Data from Biomolecular Assemblies. *J. Appl. Crystallogr.* **2008**, *41*, 222−226.

(48) Rosa-Trevín, J. M. d. l.; Quintana, A.; Cano, L. D.; Zaldívar, A.; Foche, I.; Gutiérrez, J.; Gómez-Blanco, J.; Burguet-Castell, J.; Cuenca-Alba, J.; Abrishami, V.; Vargas, J.; Otón, J.; Sharov, G.; Vilas, J. L.; Navas, J.; Conesa, P.; Kazemi, M.; Marabini, R.; Sorzano, C. O. S.; Carazo, J. M. Scipion: A Software Framework toward Integration, Reproducibility and Validation in 3D Electron Microscopy. *J. Struct. Biol.* **2016**, *195*, 93−99.

(49) Scheres, S. H. W. Relion: Implementation of a Bayesian Approach to Cryo-Em Structure Determination. *J. Struct. Biol.* **2012**, *180*, 519−530.

(50) Sedlak, S. M.; Bruetzel, L. K.; Lipfert, J. Quantitative Evaluation of Statistical Errors in Small-Angle X-Ray Scattering Measurements. *J. Appl. Crystallogr.* **2017**, *50*, 621−630.

(51) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K. W.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y.-E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P., Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. In *Methods in Enzymology*; Johnson, M. L.; Brand, L., Eds.; Academic Press: New York, 2011; Vol. *487*, pp 545−574.

(52) Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; Riddell, A. Stan: A Probabilistic Programming Language. *J. Stat. Softw.* **2017**, *76*, DOI: 10.18637/jss.v076.i01.

(53) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, D. A.; Cournapeau; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Machine Learning Res.* **2011**, *12*, 2825−2830.