There are about $10^{30}$ possible intermediates on the assembly path from hepatitis B capsid protein dimers to the 120-dimer capsid. If every intermediate was tested, assembly would often get stuck in an entropic trap and essentially every capsid would follow a unique assembly path. Yet, capsids assemble rapidly with minimal trapped intermediates, a realization of the Levinthal paradox. To understand the fundamental mechanisms of capsid assembly it is critical to resolve the early stages of the reaction. We have used Time-Resolved Small Angle X-ray Scattering, which is sensitive to solute size and shape and has millisecond temporal resolution. Scattering curves were fit to a thermodynamically curated library of assembly intermediates, using the principle of maximum entropy. Maximum entropy also provides a physical rationale for the selection of species. We found that the capsid assembly pathway was exquisitely sensitive to initial assembly conditions. With the mildest conditions tested, the reaction appeared two-state from dimer to 120-dimer capsid with some dimers-of-dimers and trimers-of-dimers. In slightly more aggressive conditions, we observed transient accumulation of a decamer-of-dimers and appearance of 90-dimer capsids. In conditions where there is measurable kinetic trapping, we found that a highly diverse early intermediates accumulated within a fraction of a second and propagated into long-lived kinetically trapped states (>90-mer). In all cases, intermediates between 35 and 90 subunits did not accumulate. These results are consistent with the presence of low barrier paths that connect early and late intermediates and direct the ultimate assembly path to late intermediates where assembly can be paused.
Rapidly Forming Early Intermediate Structures Dictate the Pathway of Capsid Assembly

Roi Asor,† Christopher John Schlicksup,‡ Zhongchao Zhao,‡ Adam Zlotnick,‡ and Uri Raviv*,†

†Institute of Chemistry and the Center for Nanoscience and Nanotechnology, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram, Jerusalem, 9190401, Israel
‡Department of Molecular and Cellular Biochemistry, Indiana University, Bloomington, Indiana 47405, United States

E-mail: uri.raviv@mail.huji.ac.il
Phone: +972-2-6586030. Fax: +972-2-566-0425

Abstract

There are \( \sim 10^{30} \) possible intermediates on the assembly path from hepatitis B capsid protein dimers to the 120-dimer capsid. If every intermediate was tested, assembly would often get stuck in an entropic trap and essentially every capsid would follow a unique assembly path. Yet, capsids assemble rapidly with minimal trapped intermediates, a realization of the Levinthal paradox. To understand the fundamental mechanisms of capsid assembly it is critical to resolve the early stages of the reaction. We have used Time-Resolved Small Angle X-ray Scattering, which is sensitive to solute size and shape and has millisecond temporal resolution. Scattering curves were fit to a thermodynamically curated library of assembly intermediates, using the principle of maximum entropy. Maximum entropy also provides a physical rationale for the selection of species. We found that the capsid assembly pathway was exquisitely sensitive to
initial assembly conditions. With the mildest conditions tested, the reaction appeared
two-state from dimer to 120-dimer capsid with some dimers-of-dimers and trimers-of-
dimers. In slightly more aggressive conditions, we observed transient accumulation of
a decamer-of-dimers and appearance of 90-dimer capsids. In conditions where there is
measurable kinetic trapping, we found that a highly diverse early intermediates accu-
mulated within a fraction of a second and propagated into long-lived kinetically trapped
states (≥ 90-mer). In all cases, intermediates between 35 and 90 subunits did not accu-
mulate. These results are consistent with the presence of low barrier paths that connect
early and late intermediates and direct the ultimate assembly path to late intermediates
where assembly can be paused.

Introduction

A virus capsid is the protein shell that protects the genome of a virus. To minimize the
amount of the viral genome devoted to structural proteins, capsids are comprised of many
copies of a small number of proteins, often only one. Capsid assembly is a critical step
in the lifecycle of viruses. Despite years of research, the mechanism of icosahedral capsid
assembly (about half of the known viruses) has remained poorly understood because it
involves a large number of capsid protein subunits, a huge number of possible intermediates,
and many more potential assembly pathways. The assembly process includes nucleation,
elongation, and final closure, the timescale for which can cover many orders of magnitude:
from msec to, in vitro, days. The assembly pathways and their timescales are regulated
by the interactions between the viral components, usually with no additional chaperone.
The early reaction steps can be very fast hence tracking them and resolving the involved
structures can be challenging. Experimental kinetic data at high temporal resolution,
covering the early stages of assembly, are critical for resolving the underlying mechanism of
virus assembly.

For self-assembly of a hollow spherical polymer, capsid subunits must be at least triva-
Subunits are often themselves small oligomers. Assembly theory and simulation suggest that for assembly to alleviate errors, follow the most efficient path, and avoid kinetic traps, association energy must be relatively weak, nucleation must be relatively slow, and association must be reversible. In practice, these regulatory features have an overlapping basis and can be controlled by solution conditions, adjusting association energy and/or protein concentration. By reference to classical polymerization theory, adjusting solution conditions modifies the initial subunit supersaturation of the system.

For this study, we examined the assembly of the hepatitis B Virus (HBV) capsid because of the medical importance of the system, because its assembly in vitro is well-characterized, and because HBV capsid assembly has been identified as a promising target for direct-acting antiviral agents. HBV is an endemic pathogen that causes chronic infection in more than 250 million people and leads to about 880,000 deaths each year, by liver failure, cirrhosis, and liver cancer. HBV is an enveloped dsDNA virus that has an icosahedral capsid composed of homodimeric core protein (Cp). In-vivo, assembly can nucleate spontaneously to form empty particles, which comprise 90% of the particles present during infection. The remaining population of infectious particles assemble around a complex of viral RNA and reverse transcriptase, which may itself play a role in catalyzing assembly nucleation. Assembly-directed antiviral agents, which include molecules now in clinical trials, act by accelerating assembly, strengthening association energy, and inducing aberrant geometry.

HBV assembly can be recapitulated with purified protein. Recombinant capsid protein assembly domain, Cp149, the first 149 residues of Cp, lacking the C-terminal nucleic acid binding domain, assembles in vitro into empty capsids that are identical to the capsids isolated from virus-expressing cells. Thus, Cp is a tractable and important system for characterizing assembly. Data suggests that Cp dimers undergo a conformational change from dimer to assembly active state, and then associate predominantly by burial of hydrophobic surface. An ionic strength dependent interaction, at the level of allostery and/or screening a repulsive interaction, also affects assembly. Thus, assembly of Cp149
can be induced by increasing ionic strength and temperature.\textsuperscript{32,33} Ionic strength also alters the molar ratio between T = 3 and T = 4 capsids, consistent with an allosteric role.\textsuperscript{34,35}

Recently, we computed the grand canonical free energy landscape of HBV capsid at the onset of assembly and calibrated it based on experimental data. We found that the number of stable species is rather small (a few hundreds out of $\approx 10^{30}$) and the distribution of products depends on the Cp-Cp interaction strength.\textsuperscript{33} Kinetic pathways, however, cannot be directly predicted from free energy landscapes and remained largely unknown.

Light scattering measurements showed that under mild conditions the assembly reaction of $T = 4$ capsids is well described by the nucleation and growth model where a single subunit can be added at each step.\textsuperscript{36} Charge detection mass spectroscopy (CDMS), a single molecule technique, revealed HBV assembly products several minutes after the onset of assembly.\textsuperscript{33,37} CDMS showed that at high ionic strength (relatively strong dimer-dimer interaction) kinetically trapped complexes accumulate. The formation mechanism of these trapped states, however, was unclear owing to the limited time resolution of CDMS. Similar observations were reported using single molecule resistive-pulse sensing where very low Cp149 concentrations enabled visualization of early times, but not small species.\textsuperscript{38}

In this paper, we used Time-Resolved SAXS (TR-SAXS) with millisecond temporal resolution to track in real-time the assembly of empty HBV capsids. The supersaturation state at the onset of assembly, for Cp (a function of protein concentration, ionic strength, and temperature), has a dominant effect on the assembly path, biasing the competition between the formation of full capsids and malformed structures.\textsuperscript{8,33,36} From rigorous analyses of our data, using maximum information entropy and examination of the free energy landscape,\textsuperscript{33} we find that an increase of $1 \text{k}_B T$ in the interaction strength between subunits can dramatically affect the reaction rates, accumulation of intermediates, and assembly mechanism. Remarkably, under the conditions that we tested, the path of assembly was determined in less than a second.
Results and Discussion

Equilibrium Analysis of Assembly Products

HBV capsids were assembled at pH 7.5 and 25°C from different concentrations of Cp149 dimer and ammonium acetate (Figure 1). The use of the volatile salt simplifies direct comparison with CDMS data. The equilibrium reaction products were evaluated by SAXS. To quantitatively analyze the data, spectra were fit to a thermodynamically weighted sum of basis spectra. The weights were calculated according to the grand canonical ensemble (see Equation 1 in Materials and Methods and our earlier publication33) and provided a diversity of species that afforded an adequate fit. In an equilibrium situation, it is a reasonable assumption that labile, higher energy complexes will dissipate in favor of lower energy species. In kinetic experiments, this assumption must be treated with caution.

The scattering curve for 38 µM Cp149 dimer in 20 mM ammonium acetate at 5°C (Figure 1a) fit the modeled scattering intensity for Cp149 dimer, calculated by D+ software,39,40 on the basis of the atomic dimer model given in PDB entry 2G33,41 to which a hydration layer was added. The quality of this fit indicates that no detectable oligomers of dimers were present. This state served as the initial protein solution in all of our time-resolved experiment.

Figure 1b and Figure S1 in the supporting information (SI) show the measured scattering intensities and the best fitted thermodynamic model (Equations 1 and 3) of the equilibrated reactions between 5 and 45 µM Cp149 dimer in between 163 and 1010 mM ammonium acetate. The selection of species from the thermodynamic model (Equations 1 and 3), resulted in a good fit to the data at low to moderate ionic strength that slightly deteriorated at high ionic strength. The law of mass action predicts that most of the mass of Cp149 should be divided between complete capsids (T = 4 and T = 3) and free dimer.42 The amount of free dimer, a pseudo-critical concentration, is characteristic of the intersubunit association energy and decreased with increasing ionic strength from about 3.9 µM at 163 mM ammo-
nium acetate to 0.5 μM at 510 mM ammonium acetate (see Figure 2b). The mass fraction of dimer and capsids (T = 3 and T = 4) as a function of ammonium acetate concentration for 20 μM Cp149, ranged from >99% at 163 mM ammonium acetate to <99% above 513 mM ammonium acetate, indicating a small amounts of intermediates.

The mass fraction of T = 3 capsids (Figures 1 and S1–S3) increased with salt concentration from about 0.08 at 163 mM ammonium acetate to 0.22 at 513 mM ammonium acetate, in agreement with CDMS results. At high ammonium acetate concentrations (see red arrows in Figure 1b, d), the best fit of our model, based exclusively on on-path capsid-like intermediates, slightly deviated from the data. Similar deviations (smearing of the oscillations and increased intensity at low q) were previously reported at high temperature and NaCl concentrations. It has been suggested that at strong association energy, kinetic traps are likely to accumulate, including off-path species such as aggregated capsid fragments, capsid oligomers with aberrant geometry, or overgrown capsids. Section 5 in the SI and Figure S5 provide additional support for aberrant particle products with a smaller average radius than that of a T = 4 capsid. A phase diagram (Figure 1d) classifies the different assembly conditions according to the quality of the fit of Eq. 3 to our SAXS data (Figure S1).

**Dimer-Dimer Association Free Energy**

Figure 2 shows the salt dependence of the dimer-dimer association free energy in T = 4 capsids and the pseudo critical dimer concentration. The values were calculated from SAXS data that fit well to the thermodynamic model (blue symbols in Figure 1d). The association free energy in ammonium acetate was stronger than in equivalent NaCl concentrations. Association energies decreased from −8.2 to −9.2 k_BT when the ammonium acetate concentrations increased from 163 to 510 mM (the energy values were calculated based on the molar fraction scale and are 2 k_BT higher than would have been obtained on the molarity scale, used in some other literature). The pseudo-critical concentration from SAXS shows ex-
Figure 1: HBV core protein assembly reaction products at pH 7.5. (a) SAXS analysis shows that starting material is pure dimer. Azimuthally integrated background-subtracted SAXS absolute intensity as a functions of $q$, the magnitude of the momentum transfer vector, from $38\mu M$ (1.36 mg/mL) Cp149 in 20mM ammonium acetate at 5°C (blue symbols). The data closely fit a computed scattering curve (red) for an atomic model of Cp149 dimer (PDB ID 2G33) with a solvation layer (b) SAXS data (blue symbols) of Cp149 assembly reactions at $25^\circ C$, $\sim$24h after the addition of ammonium acetate salt. The computed scattering curves (red) used atomic models of reaction products and mass fractions from a thermodynamic model (Equation 1), in which only two free parameters (the standard dimer-dimer association free energies of $T=3$ and $T=4$) were fitted to the data. The total concentrations of Cp149 and ammonium acetate are indicated. (c) Mass fraction of the major assembly products of $20\mu M$ Cp149 at 25°C as a function of ammonium acetate concentration (data from panel b). (d) Phase diagram of the quality of fit of SAXS data to the thermodynamic model in the protein concentration-salt concentration plane. Blue, orange, and red symbols correspond to excellent agreement, slight deviations, and small deviations between the data and models, respectively. The red arrows in panel b indicate the characteristic small deviation between the model and the data at conditions that were classified into red symbols in the phase diagram (panel d, see also Figure S1)
cellent agreement with concentrations determined by size exclusion chromatography (SEC) (Figure 2b).

Figure 2: Association free energy and assembly pseudo-critical concentration obtained SAXS data (Figures 1 and S1). (a) Association free energy change per dimer-dimer contact in the $T = 4$ capsid symmetry as a function of ammonium acetate concentration calculated from SAXS data (using only data that fit well to Equation 1 - blue symbols in Figure 1b). The standard deviation was calculated from the scatter and by assuming 10% error in the measured total protein concentration. The free energies were calculated on the molar fraction scale which resulted in an offset ($2k_B T$) with respect to the association energies calculated on the molar scale. (b) Pseudo critical concentrations of Cp149 dimer as a function of ammonium acetate concentration obtained from SAXS data (blue circles) and by size exclusion chromatography (red diamond symbols).
Time-Resolved SAXS Measurements and Singular Value Decomposition

Because of the differences in the distributions of products as a function of protein concentration and ionic strength, we hypothesized that the initial supersaturation state of the reaction would affect the early stages of the assembly process and thus dictate the subsequent assembly path. We used TR-SAXS to examine assembly kinetics (Figure 3). At the onset of all the reactions (between 20 and 40 msec), dimer was the dominant state. The oscillation in the scattering curve, corresponding to the capsid structure, became stronger as time proceeded. The oscillations appeared earlier at higher salt concentrations.

The minimum number of independent states contributing to the time-resolved data was estimated by singular value decomposition (SVD). This analysis compares the information content of TR-SAXS data sets at the different assembly conditions (see Materials and Methods and Section 4 in the SI). At 163 mM ammonium acetate the scattering data set could be reconstructed, within the noise level, using only two orthonormal vectors corresponding to reactant and final product. At 313 mM and 513 mM ammonium acetate, at least 3 and 4 singular vectors were needed for the reconstruction, respectively. Figures S6 and S7 show that at higher salt concentration the effective rank of the time-resolved data matrix was higher.

Maximum Information Entropy Analysis of Time-Resolved SAXS Data

To identify the structure of the dominant species involved in capsid assembly, we fitted the TR-SAXS curves to a weighted sum of calculated scattering intensities for intermediates (Equation 7). To reduce the possibility of overfitting, arising from the large number of degrees of freedom in the configurational space and the limited information of TR-SAXS curves, we used maximum information entropy method with a representative library of on-path configurations from dimer to full \( T = 3 \) and \( T = 4 \) capsids. This method finds the
Figure 3: TR-SAXS data and maximum entropy analysis at selected time points during the assembly reactions. TR-SAXS intensity (blue symbols and gray error bars) fits well to scattering curves calculated from maximum entropy-based models (red curves) for reactions with 25 µM (panels a. and b.) and 41 µM (panel c.). Cp149 assembly was initiated by increasing the ammonium acetate concentration from 20 mM to (a.) 163 mM, (b.) 313 mM, or (c.) 513 mM, using a stopped-flow device. Reactions were performed at 25°C. Below each SAXS panel, a quantile-quantile (Q-Q) plot compares histograms of normalized residuals for the presented scattering curves (Equation S9, where different colored open symbols correspond to different time points), with an expected normal distribution (red lines). The small deviations (~0.1σ) of the mean value from 0 and of the standard deviations from 1, in σ units, may be attributed to small inaccuracy (~1 – 2%) in the measured absolute intensity and inaccuracy in estimating the experimental and modeling errors (see also Limits of TR-SAXS Detection in Materials and Methods). Figures S9–S16 in section 6 in the SI show the complete data set, fitting results, and residuals as a function of q.
probability distribution that has maximum information entropy subject to the SAXS data and prior knowledge. Maximum entropy assigns a positive weight to every component that is not excluded by the given information, hence ensures that no arbitrary assumptions are introduced. Our library of representative states is large to minimize bias.

Initial attempts to fit equilibrium SAXS data (that have a much better SNR than TR-SAXS data) using a maximum information entropy approach, assuming all intermediates in the library are equally accessible, failed because of the enormous number of intermediates. By filtering particles for stability (see Equation 14, Figure S8, section 5 in the SI), we eliminated many fragile and extended intermediates, as well as particles that would have arisen from those intermediates, resulting in a library of species that is tractable, thermodynamically realistic, kinetically accessible, and consistent with numerous experimental works and simulations. The scattering curves calculated from maximum informational entropy optimization (Figures 3 and Figures S9–S16) agree well with experimental measurements. The quality of the fits can be seen both in the $I$ vs. $q$ curves and by the distribution of the normalized residuals (Equation S9).

Mild Assembly Conditions: The Appearance of a Two-State Reaction

After 24h, assembly reactions with 163 mM ammonium acetate and 25 µM Cp149 (Figure 1), had more than 99% of the capsid protein distributed between free dimer and full capsids. Therefore, the assembly in 163 mM ammonium acetate provided a means to examine a reaction with low probability for off-path kinetic traps. Figure 4a shows that intermediates larger than three dimers did not accumulate to detectable amounts during the assembly process. The state of the system at 20 msec was dominated by free dimer ($D \sim 90\%$) with small amounts of dimers of dimers ($D_2, \sim 4\%$) and trimers of dimers ($D_3, \sim 6\%$). At 256 sec, the final reaction product was almost exclusively $T = 4$ capsid. These results are consistent with the SVD analysis, indicated an effective rank of 2 (Figure S7). An initial lag
Figure 4: Effect of ammonium acetate concentration on the kinetics of Cp149 dimer assembly. Assembly reactions show fundamental differences at ammonium acetate concentrations of (a) 163 mM, (b) 313 mM, and (c) 513 mM. (left panel) SAXS data (gray curves and error bars) at selected early times were fitted to a library of intermediates by maximum entropy (black curves). (Middle panel) The results of the fitting are displayed in a 3D plot in terms of mass fraction, size ($s$), and time. Note that time is on a log scale. Illustrations were added to indicate the major components that accumulated along the assembly path. (right panel) These results are rearranged to show the mass fractions of the major components as a function of time. Besides dimer (D), $T = 3$ capsid, and $T = 4$ capsid, dimers and trimers of dimers ($D_2$ and $D_3$) are major components. At 313 mM ammonium acetate, $D_{10}$ is also noted. At 513 mM ammonium acetate, the broad peak of intermediates contains between 7 and 35 dimers and in the right panel is considered a single state, $D_{7-35}$. 
phase of $\sim 10$ sec was followed by a depletion of free subunits and formation of capsids, this
lag is attribute to the time required to build up a steady state pipeline of intermediates.\textsuperscript{10} Though intermediates towards capsid completion were presumably synthesized during the
lag phase,\textsuperscript{42,43,45} their concentrations were too low to be detected.

The Reaction Can Be Approximated by a Series of Reversible Assembly Steps

To examine the assembly pathway, we calculated the grand canonical free energy change,
$\Delta \Omega_G$, for the formation of $T = 4$ capsid (Equation 5) at the onset ($t = 0$) of the reaction
(Figure 5a). This landscape maps the differences between the chemical potential of $s$ free
dimers, $s\mu_1$ and the chemical potential, $\mu^\circ_{4,s,c}$, of $s$ dimers incorporated into $T = 4$ interme-
diate, $T^4_{s,c}$, forming $c$ inter-dimer contacts. This difference is the driving force for assembly.

For this calculation we used the standard association energy between subunits , $\Delta F^\circ_4$, in
163 mM ammonium acetate ($8.15 k_B T$), obtained from the thermodynamic analysis of the
equilibrium SAXS data (Figure 2). The free energy is plotted as heat map over a plane of
intermediate size ($s$, in dimers) and the degree of connectivity, $D_c$ (the number of intersub-
unit contacts in an intermediate of a given size divided by the maximum number of contacts
for an intermediate of that size (Equation 4).

By plotting the minimum free energy path between the free dimer and the complete
$T = 4$ capsid (the $D_c = 1$ cut in Figure 5a), we describe the assembly path with the lowest
free energy barriers at the onset of the reaction (black curve in Figure 5d). We found a
relatively high ($\sim 17 k_B T$) and broad free energy barrier for assembly (a broad peak between
$s = 11$ and 23). Following this, the free energy decreases towards the full capsid state where
no local minima, deeper than $\sim 1 k_B T$, can be seen.

This type of free energy landscape is consistent with a nucleation and growth mecha-
nism,\textsuperscript{17,36} where at least the early stages of assembly (before the peak in the free energy bar-
rier) are controlled by weak and reversible binding, important for correct assembly.\textsuperscript{7,14,16,17,32}

In a reversible assembly step, nonoptimal contacts can be corrected, which favors the case
where the populated intermediates along the assembly path are the most compact and most stable. Additionally, kinetic traps that may result from sudden depletion of free subunits in the solution are avoided owing to the gradual decrease in the concentration of free subunits and also for intermediates to rearrange.

In a fully reversible process, the free energy landscape at different time points can be approximated by Equation 5, where the variation in time is given by the decrease in the molar fraction of free dimer, $X_1(t)$. In this case, the dimer and $T = 4$ capsid (the two stable states), are separated by a free energy barrier that increases with time. In the absence of deep local minima, the concentrations of other intermediates are expected to be low throughout the process (Figure 4a). In other words, intermediate will either pass the barrier and form capsid or completely disassemble. The broad energy barrier, shallow local minima, and weak binding are consistent with heterogeneous nucleation mechanism.

**Aggressive Assembly Conditions: Low Barrier Assembly Leads to Rapid Accumulation of Intermediates.**

In conditions where association energy is relatively strong (513 mM ammonium acetate), the scattering intensity increased much faster than in 163 mM salt (Figure 4a). The data indicated fast accumulation of a broad distribution of intermediates containing between 7 and 35 dimers. The final assembly products included appreciable concentrations of $T = 3$ capsids ($s = 90$) as well as $T = 4$ capsids. The mass fractions of the different species detected during assembly may be roughly clustered into four groups (in agreement with the SVD analysis), whose concentrations were correlated with time: the earliest distribution after initiating assembly (D, D₂, and D₃), a broad peak of mid-size intermediates (D₇₋₃₅), $T = 3$ capsid, and $T = 4$ capsid (see Limits of TR-SAXS Detection in Materials and Methods).

Figure 4c shows the mass fraction results of the dominant species as a function of time, extracted from three different mixing experiments. Within ~ 250 msec, the mass fraction of free dimer rapidly decreased (without a measurable lag-phase) to half of its initial value, the
mass fraction of mid-size intermediates grew to about 0.3, and $T = 3$ and $T = 4$ capsids, started to accumulate to detectable amounts (mass fraction of about 0.1 each). As the reaction proceeded, the mass fraction of the mid-size intermediates decayed to less than 0.05 (at 200 sec), and the dominant assemblies were $T = 3$ and $T = 4$ capsids.

At strong association energy (high salt concentration), the fast depletion of free subunit is consistent with a much lower barrier for assembly. Figures 5c and d show the expected initial grand canonical free energy landscape (c), representing the thermodynamic driving force for assembly, and the minimum free energy path (d, red curve). Under those conditions, the barrier for assembly was very low and limited to species with $s < 10$. The shape of the contour lines of the initial grand canonical free energy landscape (Figure 5c) suggests that there are low barriers for sampling less compact mid-size intermediates ($D_c$ as low as 0.6, representing structures with suboptimal binding).

The rapid (within 250 msec) accumulation of mid-size intermediates and the deviation from the expected equilibrium state (Figures 1 and S12) suggest that malformed particles could be formed by interactions between multi-dimer complexes. The relatively strong dimer-dimer association interactions are expected to slow rearrangement of malformed particles into the more stable $T = 4$ capsids.

Figure 4c shows that within 1 sec the mass fraction of $s = 90$ particles ($T = 3$ capsid) reached $\sim 0.22$, similar to the value measured after 40 h (Figure S4). This observation is consistent with single molecule observations that $T = 3$ capsids assemble only at the earliest times of the reaction.

After the first second of the reaction, the dominant processes were an increase in the concentration of $T = 4$ capsid and a decrease in dimer concentration (Figure 4c). Figure 5c shows the expected change in the grand canonical free energy along the minimum free energy path ($D_c = 1$) at different time points during the measured assembly process at 513 mM ammonium acetate. After 1 sec the free energy profile resembled the initial free energy profile of the 163 mM reaction, which predominantly gave $T = 4$ capsids without
detectable intermediates.

The increase in the concentration of $T = 4$ capsid and the exclusion of new $T = 3$ assembly could have been achieved by: (i) Creation of new 'capsid assembly lines' with relatively low dimer concentration, which are expected to act by the nucleation and elongation mechanism, as observed with the 163 mM salt. This observation is consistent with the hypothesis offered based on single molecule observations of assembly. (ii) Slow elongation of the medium-size intermediates that still possessed $\sim 10\%$ of the total mass. (iii) Annealing of malformed particles. As the mass fraction of $T = 3$ particles, which may include malformed $T = 3$-like particles, was constant at these time scale and equal to the steady state results, the third option is unlikely. Additionally, CDMS results suggest that the annealing of malformed particles can take days. We therefore expect that mechanisms (i) and (ii) dominated in our case.

Our analysis of the equilibrated reactions at the high salt concentration indicated that in addition to a higher fraction of $T = 3$ capsids (compared with low salt conditions), large aggregates formed, observed by the higher intensity at low $q$ (Figure 1b). Figure S12 shows that 50 sec after the onset of the reaction, the measured intensity at the low $q$ range deviated from the modeled intensity (the residuals at low $q$ were larger than $4\sigma$) in a similar manner to that of the equilibrated reaction (Figures 1 and S1). These observations suggest that the accumulation of large aggregates (larger than full capsids) took tens of seconds.

Kinetic traps via starvation for free subunit can be excluded because, after 1 sec, the dimer concentration was still 6.25 $\mu$M, which is an order of magnitude higher than the pseudocritical concentration (about 0.5 $\mu$M, Figure 2). Therefore, there were enough available free dimers to elongate incomplete intermediates.
In Moderate Assembly Conditions, Compact 10-Mer Intermediates Accumulated

To test our understanding of the assembly reaction, we examined assembly at a higher dimer concentration of 41 µM and intermediate dimer-dimer association free energy (313 mM ammonium acetate) (Figure 4b). Experiments at higher protein concentration provide better SNR but also higher molar activity to accelerate the reaction and trap reactants. However, the time scale for the assembly reaction, the variation of the scattering intensity at short timescale, the effective SVD rank of 3 (Figures S6and S7), and the average growth of the particles (⟨s⟩) as a function of time (Figure 5f) suggest that the assembly process proceeded by a mechanism that is distinct from the ones observed for the low (163 mM) and high (513 mM) salt conditions.

The reaction in Figure 4b, exhibited a lag phase of ∼ 0.4 sec, followed by a gradual decrease in the free dimer concentration and accumulation of T = 4 capsids. The accumulated mass fraction of T = 3 particles was only ∼ 0.03. The shorter lag phase, compared with the 163 mM salt, is attributed to the higher initial supersaturation state due to high initial protein concentration and a lower pseudo-critical concentration. In addition, the stronger association decreases reversibility of intermediate reactions contributing to a faster rate of elongation. The high protein concentration and stronger association energy also contributed to the lower calculated free energy barrier for assembly (7.5 k_B T; green curve in Figure 5d). The assembly in 513 mM Ammonium Acetate, where association energy is stronger, was much faster than in 313 mM because the minimum free energy barrier for assembly in 513 mM was lower by about 2 k_B T than the barrier at 313 mM (Figure 5d).

Maximum entropy fitting of scattering data suggests the accumulation of a specific intermediate of ten dimers. The most stable and compact 10-mer intermediate, which is consistent with scattering, has 15 inter-dimer contacts, arranged about a 5-fold symmetry axis (Figure 4b). At its highest concentration, at t = 2 sec, when free dimer had dropped to about 70 % of its initial concentration, 5 % of the dimers were in this 10-mer intermediate form. At
this free-dimer concentration, the compact 10-mer is a local free energy minimum, along the \( D_c = 1 \) line (see Figure S17), right after the assembly barrier. The contribution of the compact 10-mer intermediate to the overall scattering data is important for reconstructing the experimental data at the early points of assembly. We note that the compact 10-mer may be an intermediate that is on-path to capsid assembly or may be a trap.

Figure 5: Heat maps of energy (in \( k_B T \) units) plotted in the plane of degree of connectivity (\( D_c \)) versus size in dimers (\( s \)) from the change in the grand canonical free energy surface, \( \Delta \Omega_G \), for \( T = 4 \) capsid at the onset (\( t = 0 \)) of the assembly reactions (Equation 3). Degree of connectivity is the number of intersubunit contacts in an intermediate of a given size divided by the maximum number of contacts of intermediates of that size (Equation 4). Calculations of \( \Delta \Omega_G \) are for free energies per contact, \( \Delta F_p \), extracted from equilibrated reactions (Figures 2) at ammonium acetate concentrations of (a) 163 mM (8.15 \( k_B T \)) (b), 313 mM (8.6 \( k_B T \)) (b), or (c) 513 mM (9.2 \( k_B T \)). The energy surfaces are calculated for the reactions shown in Figures 3 and 4 where the initial Cp149 concentrations were 25 \( \mu \)M (a and c), and 41 \( \mu \)M (b). (d.) The initial grand canonical free energy change, \( \Delta \Omega_G (t = 0) \), along the minimum free energy path (\( D_c = 1 \)) for \( T = 4 \) intermediates varies at different ammonium acetate concentrations. The inset shows the free energy barrier on an expanded scale. (e.) The predicated grand canonical free energy change, \( \Delta \Omega_G (t) \), for \( T = 4 \) capsid assembly in 513 mM ammonium acetate at different time points after the onset of the reaction, along the \( D_c = 1 \) path (calculated by Equation 5) using the molar fraction of free dimer subunits as a function of time, \( X_1 (t) \), from Figure 4b. (f.) Number averaged intermediate size, \( \langle s \rangle \), as a function of time during the first minute of the three assembly reactions, calculated by Equation 15 using the TR-SAXS data analysis in Figure 4.
Conclusions

We deduce that the pathway of HBV capsid assembly is dictated by the intermediates formed within less than a second. At mild dimer-dimer association free energy, assembly appears to be a two-state reaction (dimer and $T = 4$), though intermediates were necessarily present (but did not accumulate to detectable amounts). This reaction had a 10 sec lag phase. The grand canonical free energy landscape had a relatively high and broad barrier (that broadened with time and prevented the accumulation of intermediates), following which the energy decreased towards the full capsid with no local minima. This landscape is consistent with multiple reversible steps, allowing the reaction to follow the minimum free-energy path, at which the most stable and compact intermediate structures are dominant, and is consistent with heterogeneous nucleation mechanism. At aggressive assembly conditions, the dimer-dimer association free energy was about $1 \text{k}_B \text{T}$ higher, the reaction was much faster, and the dimer concentration rapidly decreased without any measurable lag phase. A diverse array of intermediates, containing between 7 and 35 dimers accumulated within the first 250 msec after which $T = 4$ and $T = 3$ (or $T = 3$-like) particles were detected. After the first second, the dimer concentration was still above the pseudo critical concentration and supported assembly of $T = 4$ capsids by either slow elongation of the mid-size intermediates or establishing new ‘capsid assembly lines’. At moderate assembly conditions, the main intermediate was a compact 10-mer, a deep local free energy minimum after the assembly barrier. The high temporal resolution of our data and analyses showed that small changes in the dimer-dimer association free energy control the earliest steps of the reaction and dictate the subsequent assembly pathway. Our findings may provide strategies for understanding, regulating, and designing assembly of protein cages.
Materials and Methods

Sample Preparation

The N-terminal truncated dimer, Cp149, was expressed in E.coli using a pET 11-based vector. The dimer was then purified as described. To prepare oligomer free dimer for SAXS, solid urea was added to the purified Cp149 dimer solution to reach a final concentration of 3M urea. After one hour, buffer exchange with 20 mM ammonium acetate at pH 7.5 was performed at 4°C, using a pre-equilibrated PD-10 column. The fraction that contained the Cp149 dimer was collected and its concentration determined by UV-Vis absorption spectroscopy using an extinction coefficient of 60,900 M⁻¹ cm⁻¹. Before measurements were performed, the solution was incubated between 0.5 and 40 h at ambient room temperature.

Size Exclusion Chromatography

Similar to previous publications, pseudo-critical concentrations of Cp149 assembly were determined at pH 7.5 for three salt conditions, 163 mM, 313 mM and 513 mM ammonium acetate, by size exclusion chromatography. Various concentrations of Cp149 were assembled and incubated at 23°C for over 24 h. All assembly reactions were analyzed using a Superose 6 10/300 GL column (GE Healthcare) mounted on an HPLC system (Shimadzu Corp.). The column was pre-equilibrated with the corresponding salt solutions. Assembled capsids and dimers were quantified by their UV-Vis absorption to determine pseudo-critical concentrations.

SAXS Measurements.

Solution small X-ray scattering (SAXS) measurements of capsid assembly were performed at P12 EMBL BioSAXS Beamline (headed by D. Svergun) in PETRA III (DESY, Hamburg). Measurement were taken using an automated sample changer setup in which samples were stored on a temperature controlled plate and injected into a 2mm thick quartz capillary.
that was previously equilibrated at the same temperature. The wavelength of the incident X-ray beam was 0.124 nm$^{-1}$ and the scattering intensity was recorded on a single-photon PILATUS 2M pixel area detector (DECTRIS).

Background measurements before and after each sample were performed on the solvent of each sample, under identical measurement conditions to that of the sample. The intensity frames were normalized to the intensity of the transmitted beam, and azimuthally averaged to yield the scattering intensity as a function of the magnitude of the scattering vector, $q$. Background scattering curves were averaged and the averaged background signal was subtracted from the averaged sample, and gave the final background subtracted scattering intensity curve of the assembly reactions, as explained in our earlier papers. The products of all the assembly reactions were measured at 25°C. The sample to detector distance was 3.1 m, resulting in $q_{\text{min}} = 0.025$ nm$^{-1}$ and $q_{\text{max}} = 5$ nm$^{-1}$. 40 µL of sample were injected in each measurements and 30 frames were recorded by exposing the sample for 45 ms per frame. Additional measurements were performed at ID02 beamline (headed by T. Narayanan) in the European synchrotron radiation facility (ESRF, Grenoble). Static measurements were taken using the flow-cell setup which included a temperature controlled, 2 mm thick, quartz capillary. The wavelength of the incident beam was 0.995 nm and the scattered intensity was recorded on a FReLoN 16M Kodak CCD detector.

**Time-Resolved SAXS Measurements**

Time resolved SAXS (TR-SAXS) experiments were performed at ID02 beamline in ESRF. The assembly reactions in 163 and 513 mM ammonium acetate were initiated by mixing 1.4 mg/ml Cp149 in 20 mM ammonium acetate with either 0.45 or 1.5 M ammonium acetate, at volume ratio of 2:1. The assembly reaction in 313 mM ammonium acetate was initiated by mixing 2.2 mg/ml Cp149 and 0.9 M ammonium acetate using the same procedure. All the solutions were at pH 7.5. The reactions were observed over the initial 3 or 4 min, using a stopped-flow setup as explained in our earlier publication.
of ID02 was used to follow the slower phase of the reactions.

**Fitting the Thermodynamic Equilibrium Model**

In our earlier paper,\textsuperscript{33} we established the thermodynamic analysis of capsid virus assembly. We considered the following set of coupled assembly reactions, induce by increasing the salt concentration:

\[
\nu_D D_{(aq)} \xrightleftharpoons{[\text{Salt}]} \sum_n \sum s \sum_{c>0} \nu_{n,s,c} T_{n}^{s,c}_{(aq)}
\]

$s$ is the number of dimer molecules, $D$, which assemble into a $T_{n}^{s,c}$ icosahedral capsid intermediate, whose triangulation, $T$, number is $n$, and $c$ is the number of its dimer - dimer contacts (or inter-dimer interactions). $n$ is either 3 or 4 and $\nu_{n,s,c}$ and $\nu_D$ are the stoichiometric coefficients of $T_{n}^{s,c}$ and $D$, respectively (note that $\nu_D = \nu_{3,1,0} = \nu_{4,1,0}$). From mass conservation we get:

\[
\nu_D = \sum_n \sum s \sum_{c>0} s \cdot \nu_{n,s,c}
\]

We have used graph representation of a comprehensive library of $T = 4$ and $T = 3$ unique capsid assembly intermediates, generated by umbrella sampling of MC simulations, to get the degeneracy factors, $\Omega_{n,s,c}$, of all the $T_{n}^{s,c}$ icosahedral capsid intermediates. The details about the simulations were provided in our earlier paper. The degeneracy factors were used in a thermodynamic theory of macromolecular self-assembly, assuming a negative standard free energy for the association between capsid protein subunits, $\Delta F_n^c$ (for $n = 3$ or 4). By minimizing the total Helmholtz free energy of the grand canonical ensemble, we obtained the expected equilibrium distribution of dimer subunits molar fractions, $X_{n,s,c}$, in each of the possible $T_{n}^{s,c}$ intermediate structures in the $(n, s, c)$ configurational space:

\[
X_{n,s,c} = s \times \exp \left( -\frac{\mu_{n,s,c}^o - s\mu_1}{k_B T} \right).
\]
at a given temperature, $T$, and total protein molar fraction, $X_{\text{Total}}$. The chemical potential of the free dimer (in the solution) is

$$\mu_1 = \mu_1^0 + k_B T \ln X_1$$  \hspace{1cm} (2)

where $X_1$ is the molar fraction of free dimer ($= X_{n,1,0}$). The change in the standard chemical potential of $T_n^{s,c}$ relative to $s$ free dimers is $\mu_{n,s,c}^0 - s\mu_1^0 = c\Delta F_n^0 + k_B T \ln \Omega_{n,s,c}$. Eq. 1 was derived and well fitted to X-ray scattering data from HBV capsid in NaCl solutions, in our earlier paper.\textsuperscript{33}

We computed the solution scattering intensity curves, $I_n^{s,c}$, of each representative of the $T_n^{s,c}$ family of intermediates by docking the atomic model of the dimer (Cp149, PDB ID 2G33) into the symmetry of the intermediate (the set of all the translation vectors and the rotation matrices of the dimers in the intermediate complex). $q$ The computations took into account the contribution of the dimer solvation layer and the experimental resolution function as explained,\textsuperscript{33} using our home-developed state of the art scattering data analysis software, D+.\textsuperscript{39,40} Based on clustering algorithm analysis of the scattering curves,\textsuperscript{33} when $s$ was larger than 30, the variation of the scattering intensity curves between different members of the same family was very small. We therefore selected only one representative model for each $s$ value.

When $s$ was smaller than 30, the variation between scattering curves was not negligible. Hence, to better represent the families of small intermediates, while keeping the computation times of the optimization procedures for time resolved and equilibrium analysis feasible, we randomly selected five models from each family of type $T_n^{s,c}$. The total number of model was therefore 8477. Based on the thermodynamic analysis, the predicted total intensity at equilibrium

$$I_{\text{Model}} (q) = X_{\text{Total}}^{-1} \sum_{n,s,c} X_{n,s,c} I_n^{s,c} (q)$$  \hspace{1cm} (3)

was computed and then fitted, as explained,\textsuperscript{33} to the experimental X-ray scattering data,
where the only free parameters were the dimer-dimer standard association free energy in both $T = 4$ and $T = 3$ symmetries ($\Delta F_4^0$ and $\Delta F_3^0$). $q$ is the magnitude of the scattering vector and $X_{\text{Total}}^{-1}$ is the total molar fraction of the dimer protein. The best fit to the scattering data revealed the mass fractions, $X_{n,s,c}$, of all the intermediates at equilibrium at the relevant experimental conditions.

**Grand Canonical Free Energy Landscape**

Figure 5 shows heat maps of the grand canonical free energy landscape at the onset ($t = 0$) of the assembly reaction as a function of the entire configurational space of $T = 4$ symmetry in the $s - D_c$ plane, where

$$D_c = \frac{c(s) - c_{\text{max}}(s)}{c_{\text{max}}(s) - c_{\text{min}}(s)}$$

(4)

is the degree of connectivity of $T_{4}^{s,c}$ intermediate. $c_{\text{max}}$ and $c_{\text{min}}$ are the maximum and minimum number of contacts in intermediates containing $s$ dimers, respectively.

The grand canonical free energy change $\Delta \Omega_G$ for the formation of $T_{4}^{s,c}$ intermediates at time $t$ is:

$$\Delta \Omega_G (\Delta F_4^o, \mu_{1,t}) = \mu_{4,s,c}^o - s \mu_{1,t} + c \Delta F_4^o + k_B T \ln \Omega_{4,s,c} - s k_B T \ln X_1(t),$$

(5)

where $\mu_{4,s,c}^o$ is the standard chemical potential of $T_{4}^{s,c}$ intermediate, and $\mu_{1,t}$ is the free dimer chemical potential at time $t$, calculated according to Equation 2, using the molar fraction of free dimer subunits, $X_1(t)$, at time $t$.

**Singular Value Decomposition (SVD) Analysis**

For each of the three assembly reactions we defined an $n$ by $m$ data matrix, $D$, in which each column represented a one dimensional scattering intensity curve, $I(q,t)$, measured at time $t$ following the initiation of the reaction. The total number of rows, $n$, was set by the size of the $\vec{q}$ vector, whereas the total number of columns, $m$, was set by the total number
of measurement time points along the assembly reaction. The singular value deconvolution (SVD) of the matrix $D$, containing the time evolution of a measured spectra, is give by,

$$D = U \Sigma V^T$$  \hspace{1cm} (6)$$

where, $U$ and $V$ are unitary matrices and $\Sigma$ is a diagonal matrix with non-negative real values along its diagonal. The columns of the matrix $U$ and $V$ are the left and right orthonormal set of singular vectors of the matrix $D$. The singular values, $\sigma_i$, may be sorted (along with the corresponding columns of $U$ and $V$), from the largest ($\sigma_1$) to the smallest value ($\sigma_n$). With this ordering, the largest index $r$ with a positive singular value is the effective rank of $D$ and the first $r$ columns of $U$ comprise an orthonormal basis of the space spanned by the columns of $D$.

As previously described, the first $k \leq r$ columns of $U$, forming the matrix $U_k$, along with the corresponding first $k$ columns of $V$, forming the matrix $V_k$, and the first $k$ rows and and $k$ column of $\Sigma$, forming the matrix $\Sigma_k$, provide the best least squares approximation, $D_k = U_k \Sigma_k V_k^T$ with a rank of $k$, to the matrix $D$, where $\|D - D_k\|^2 = \sum_{i=k+1}^{n} \sigma_i^2$. By finding $r$ one can estimate the (minimal) number of species that are involved in the kinetic process described by $D$. The detailed protocol for finding $r$ using SVD analysis was previously described. However, since the basis spectra provided by \{$U_1, ..., U_r$\} has no physical meaning, the result of the SVD analysis can give only a rough approximation for the number of physical states along the measured process. SVD analysis cannot detect intermediates that accumulate at small amounts or that their appearance or disappearance as a function time is correlated with that of the reactants or products. In this work we used SVD analysis to get additional qualitative information regarding the differences in the kinetic process of different data sets. The complete detailed analysis and the results are provided in section 4 in the SI.
Using Maximum Information Entropy to Fit the Time-Resolved SAXS Data

In this method, information entropy is applied to determine the probability distribution, $p_{n}^{s,c}$, of $T_{n}^{s,c}$ intermediate structures, contributing to the scattering data (either at equilibrium or during kinetics). Information entropy is then computed from the probability distribution. The probability distribution, which maximizes the information entropy, subject to a set of constraints obtained from the experimental data, justifies the use of that distribution for inferring about the properties of the system, because it does not exclude any region of the phase space that is allowed by the available information.\textsuperscript{55,56} The method can be applied for many different physical problems. Here, we adopted the principle of maximum information entropy to interpret our solution X-ray scattering data under conditions, in which capsid protein solutions contained ensembles of $T_{n}^{s,c}$ intermediate structures. The computed scattering intensity curve,

$$I_{\text{model}}(q) = \sum_{s,c,n} p_{n}^{s,c} I_{n}^{s,c}(q)$$ \tag{7}

is compared with the measured scattering intensity signal, $I_{\text{exp}}(q)$. Our goal is to assign probabilities, $p_{n}^{s,c}$, to each of the possible intermediate structures, in a way that avoids uncontrolled bias, while agreeing with the experimental scattering data and whatever other information is given (for example, the probabilities are non-negative and satisfy the normalization condition, $\sum_{s,c,n} p_{n}^{s,c} = 1$ or known experimental evidence from current and past experience).

The probabilities, $p_{n}^{s,c}$, express our expectation to find each of the intermediate structures, on the basis of the available information. Information theory provides an unambiguous criterion for the uncertainty level of a given probability distribution. The criterion agrees with our intuition that a broad distribution represents more uncertainty than does a sharply peaked distribution (as long as it satisfies all the other conditions). Shannon proved that the
positive quantity, which increases with increasing uncertainty, and is additive for independent sources of uncertainty, is

\[ S = -K \sum_{s,c,n} p_{n}^{s,c} \ln p_{n}^{s,c} \]

where \( K \) is a positive constant that we shall set to unity.\(^{57}\) As this expression is identical to the expression of Gibbs entropy in statistical thermodynamics, it is called the entropy of the probability distribution \( p_{n}^{s,c} \). Hence, "entropy" measures the level of "uncertainty" in the probability values, \( p_{n}^{s,c} \).

To provide enough states we have used our library of intermediates.\(^{33}\) The degeneracy of each state, \( \Omega_{n,s,c} \), was then used to compute the prior probability distribution of HBV capsid intermediates. We have shown\(^{33}\) that the degeneracy factors, \( \Omega_{n,s,c} \) and the SAXS data are insufficient to reproduce the physical distribution of the assembly products at equilibrium, owing to the overwhelming number of possible intermediates (about \( 10^{30} \)) that act as an entropic barrier (given the information content of the SAXS curve).

To reduce the huge space of possible intermediates we have incorporated a stability bias to our prior distribution (\( \Delta F_{n}^{\circ} < 0 \) in Equation 1) and included the contribution of the free dimer chemical potential (see Equation 14 and Section 5 in the SI).\(^{33}\) The scattering curves from each intermediate in the library was then computed using atomic models. Finally, maximum entropy optimization was used to determine the probability distribution of intermediates at each of the TR-SAXS curves. The resulting mass fraction distribution could be then compared with CDMS data, when performed under similar conditions.\(^{33}\)

**Maximum Information Entropy Probability Distribution**

The following section describes the essential ideas and derivations that were used to perform the maximum informational entropy analysis. Full derivation of the presented equations can by found in section 5 in the SI. Given a set of \( M \) possible models, before any additional information is available, each of the possible states are expected to be equally probable. The information entropy of the distribution is then

\[ S = - \sum_{k}^{M} p_{k} \ln p_{k} \]

where, \( p_{k} \) is the probability.
to find state \( k \). When the distribution is uniform, \( S \) is maximal. Our prior knowledge may, however, dictates that the expected distribution is non-uniform and assign probability \( p_i \) to obtain the \( i \)th outcome. If, for example, there are \( g_i \) equally probable ways to obtain outcome \( i \), then \( p_k \) is given by \( p_k = \frac{p_i}{g_i} \). If \( N \) is the total number of different outcomes, \( M = \sum_i g_i \), the information entropy is

\[
S = -\sum_i g_i \left( \frac{p_i}{g_i} \ln \frac{p_i}{g_i} \right) = -\sum_i p_i \ln p_i + \sum_i p_i \ln g_i
\]

In addition, we define \( S^0 \equiv \ln (\sum_i g_i) \), and the prior probability to obtain outcome \( i \) as \( p_i^0 \equiv \frac{g_i}{\sum_i g_i} = g_i \cdot e^{-S^0} \), where \( g_i \) is the degeneracy factor of the \( i \)th outcome (likewise \( p_k^0 \equiv \frac{1}{\sum_i g_i} \)). We then get that

\[
S = S^0 - \sum_i p_i \ln \frac{p_i}{p_i^0} \quad (8)
\]

\( S^0 \) is the maximal value of \( S \), which is obtained when the actual distribution, \( p_i \), is equal to the prior distribution \( p_i^0 \). In other words, the prior distribution is the distribution of maximum entropy before taking into account the new constraints imposed by the data (beyond the degeneracy factors that always present and are inherent to the problem and were taken into account in the prior distribution). When the actual probability distribution is different than the prior distribution (owing to the additional constraints that became available from the experiments), the entropy is lower than \( S^0 \). The term \(-\sum_i p_i \ln \frac{p_i}{p_i^0} \) is the difference between the maximal and the actual value of the entropy and is therefore called the "entropy deficiency".

In this paper, we maximized the information entropy (Eq. 8), which takes into account the prior distribution, \( p_i^0 \), subject to the following three constrains.

1. All the probabilities are positive:

\[
p_i \geq 0 \ \forall \ i,
\]
2. The probability distribution is normalized:

\[ \sum_i p_i = 1, \]

3. The average signal should fit the experimental scattering data:

\[-\epsilon_q \leq \left[ \sum_i p_i \cdot I_i(q) \right] - I_{\text{exp}}(q) \leq \epsilon_q\]

where \( \epsilon_q \) defined the noise level at each scattering angle.

Note that the assumptions used to compute the prior distribution impose constraints (the constraints of the prior distribution will be discussed in the next section). It is convenient to solve the minimization problem for \(-S\). The inequality constrained minimization problem, can be solved by Lagrange multipliers method \(^{58,59}\) (the full procedure is described in section 5 in the SI). The resultant distribution that maximizes the informational entropy subject to the constraint imposed by the data and our prior assumptions is given by:

\[ p_i = \frac{e^{S_0} \cdot p_i^0}{Z} \exp \left\{ - \sum_q \lambda_q \cdot I_i(q) \right\}. \tag{9} \]

where,

\[ Z \equiv \sum_i g_i \exp \left\{ - \sum_q \lambda_q \cdot I_i(q) \right\}. \]

and \( \lambda \) is the vector of Lagrange multipliers (whose length is equal to the number of \( q \) points in the scattering curve), which sets the required probability distribution and was found by finding the solution to the Lagrange dual problem (as explain in section 5 in the SI) defined as:

\[ \min_{\lambda} \quad -L(\lambda; \epsilon) \tag{10} \]
where,

$$- L (\lambda; \epsilon) = S^0 + \ln \left[ \sum_i g_i \exp \left\{ - \sum_q \lambda_q \cdot I_i (q) \right\} \right] + \sum_q \lambda_q \cdot I_{exp} (q) + \sum_q |\lambda_q| \cdot \epsilon_q. \quad (11)$$

In this minimization problem, we can define a constraint to be active if $\lambda_q \neq 0$ or inactive if $\lambda_q = 0$. Note that the last term in Eq. (11) can be interpreted as a form of $L_1$ regularization term, which promotes the sparsity of $\lambda$. Therefore, the minimization will result with the minimum number of active constraints, needed to satisfy all of the introduced constraints. As the values of $\epsilon$ increases, the sparsity of $\lambda$ increases, hence the information content of the constraints decreases. The parameter vector $\epsilon$ is proportional to the noise level of the experimental signal, $\epsilon_q = \beta \cdot \sigma_q$, where $\sigma_q$ is the measured standard deviation of $I_{exp} (q)$ and $\beta$ is a global relaxation (or regularization) parameter. The Lagrange multiplier $\lambda_q$ is related to the contribution of the added information from $I_{exp} (q)$ and can therefore be used to estimate the information content of the signal by the number of active constraints in the entire measured $q$-range.

**Thermodynamic Constraints**

As we know additional chemical information on the problem, we can add additional constraints to the minimization problem. SAXS data alone contains limited information which may not necessarily overcome the overwhelming number of possible configurations, described by the degeneracies, $g_i$. Therefore, the additional chemical constraints confines the space of possibilities into a more realistic subspace that takes into consideration the stability of a given configuration. Common constraints for a self-assembly problem are given by the free energy gain in forming subunit-subunit interaction and a constraint regarding the expected mean number of dimers in an aggregate. Given these additional constraints, the new distribution is given by
\[ p_i = \frac{p_i^0}{Z} \exp \left\{ -\sum_q \lambda_q \cdot I_i(q) \right\} \]  \hspace{1cm} (12)

where

\[ Z = \sum_i p_i^0 \exp \left\{ -\sum_q \lambda_q \cdot I_i(q) \right\} . \]  \hspace{1cm} (13)

The prior distribution, \( p_i^0 \), is given by,

\[ p_i^0 = e^{S_0} \cdot p_i^0 \exp \left\{ -\Delta \Omega_G^i \right\} = e^{S_0} \cdot p_i^0 \cdot \exp \left\{ -\mu_E \cdot E_i - \mu_n \cdot n_i \right\} \]  \hspace{1cm} (14)

where, \( \Delta \Omega_G^i \) corresponds to the grand canonical free energy bias (for outcome \( i \)), which is a function of the free energy gain for creating inter-dimers bonds (see Equation [5], \( E_i \)), and the free energy cost of taking \( n_i \) free subunits from the solution. The multipliers \( \mu_E \) and \( \mu_n \) are associated with the parameters \( -\frac{1}{k_BT} \) and \( \frac{\mu}{k_BT} \), respectively. Complete derivation of adding the thermodynamic prior is shown in section 5 in the SI.

Performing the Maximum Information Entropy Optimization on a Set of Time Series Data

Each time-resolved experiment was initiated by mixing a cold dimer solution with a concentrated ammonium acetate solution at 25°C, resulting with a temperature and ionic strength jump. The time evolution of an assembly reaction was given by the set of scattering intensities, \( \{ I(q, t_i) \} \), where \( t_i \) corresponds to the time interval between the mixing time and the time of the \( i \)-th measurement and \( I(q, t_i) \) corresponds to the average scattering intensity during the 20 ms exposure of the measurement. To approximate the distribution of intermediates, \( p_n^{s,c}(t_i) \), we performed maximum entropy optimization (Equation [10]) on the entire series of signals, where \( i \in [1, .., m] \) and \( m \) is the number of measurement time points. If the assembly process is sampled at adequate frequency, the intermediate distributions, \( p_n^{s,c}(t) \), are likely to vary to a limited extent between consecutive measurements. The values of
should continuously vary with time. As was discussed in our earlier paper, \[p_n^{s,c}(t_i)\] when fitting a data series the continuity assumption may help to speed-up convergence.

To analyze our time-resolved data, we started the optimization from the first measured signal, \[I(q,t_1)\]. As a prior distribution, we used the closest known result, which was the state of the protein solution before mixing with the salt solution. In this state, the interaction was weak and the protein was in its pure dimeric form (Figure 1a). The thermodynamic state of the system could be well describe by Equation 14 with a weak association energy per contact of \(E_c = 5 \text{k}_B T\) and the total protein molar fraction \(X_{\text{total}}\), determined by UV-Vis adsorption measurement.

Following the optimization of the first time point, we have used the continuity of the probability distributions as a function of time. Hence, the prior distribution for the next signal, \[I(q,t_2)\], was the result of the optimization of the earlier time point \((t_1)\). This extrapolation was applied until the last measurement, \[I(q,t_m)\], was analyzed. Following the analysis of the last signal, the direction was reversed and the procedure continued in the same way from \(t_m\) backwards to \(t_1\). In this way we minimized the effect of the initial prior, which was based on the state of the protein solution (before mixing with the salt solution).

To minimize the effect of the value of \(E_c\) on the prior distribution and thereby the results of the maximum entropy optimization, we performed an additional optimization set. In the second procedure we started from the latest measured time point, \(t_m\), of the assembly reaction and used as a prior the thermodynamic probability distribution of the reaction products with an association free energy of \(E_c = 8.5 \text{k}_B T\). This value was a result of our equilibrium measurement calibration of \(E_c\), presented in Figure 2a). In this analysis, we assumed that the distribution in the latest time point, \(t_m\), was not far removed from the equilibrium distribution. The same procedure was perform in the reversed order (from \(t_1\) to \(t_m\)) and yielded similar results (\(p_n^{s,c}(t_i)\) values). Figures 3 and 4 present the average of the two procedures and the error bars correspond to the deviations between the two sets of prior distributions (obtained with the two \(E_c\) values). The same two prior distributions
were used for all the reactions conditions because even at high salt concentrations, the equilibrium SAXS data only slightly deviated from the thermodynamic model (Figure 1). Each assembly reaction was repeated between two and four times. The adequate fitting of the TR-SAXS data (Figure 3) confirms that our prior knowledge provided a good starting point for describing the assembly process.

**Clustering of SAXS Models**

Clustering of SAXS models, presented in Figure S18, was applied to the library of scattering models, used to analyze the static and time-resolved data. The full procedure was explained in our earlier work.\(^{33}\) Briefly, we defined a weighted matrix, \( M \), which included the set of scattering models to be classified into clusters:

\[
M = \\
\begin{pmatrix}
I_1(q_1) & \cdots & I_m(q_1) \\
\sigma_i(q_1) & \cdots & \sigma_m(q_1) \\
\vdots & \ddots & \vdots \\
I_1(q_N) & \cdots & I_m(q_N) \\
\sigma_i(q_N) & \cdots & \sigma_m(q_N)
\end{pmatrix}
\]

Each column in the matrix contained a computed scattering intensity, \( I_i(q) \), with \( i \in [1, m] \). The models were computed between \( q_1 = 0.1 \text{ nm}^{-1} \) and \( q_N = 1.1 \text{ nm}^{-1} \), the \( q \) range used for the time-resolved data analysis. The intensities were weighted according to the measured noise level, \( \sigma^i(q_j) \), of the time resolved measurements, where \( j \in [1, N] \). The dimensions of the matrices were therefore \( N \times m \) where \( N = 280 \) was the length of the model \( \vec{q} \)-vectors in the given \( q \) range. The dimensions of the matrix were reduced using the SVD procedure\(^{61}\) and a \( k \)-means clustering algorithm\(^{61}\) was applied to the reduced space. The number of clusters was defined as the minimal \( k \) that fulfilled the \( \chi^2 \) condition,

\[
\frac{1}{N} \sum_{j=1}^{N} \frac{(I_i(q_j) - I_c(q_j))^2}{\sigma^i(q_j)^2} < 1
\]

for all the models that were classified into the same cluster. Here, \( I_i(q_j) \) is a given model that
was classified into cluster, \( c \) and \( I_c(q_j) \) is the modeled scattering intensity of the centroid of this cluster.

**Limits of TR-SAXS Detection**

Figure 4 shows sharp mass fraction peaks for \( T = 3 \) and \( T = 4 \) capsids, attributed to the thermodynamic prior (Figure S8), favoring stable complexes. It is important to note that TR-SAXS data are insufficient to distinguish between complete capsids and capsids that are missing few subunits, observed by CDMS. To take into consideration the limited sensitivity of TR-SAXS, we applied a clustering algorithm (Figure S18 in section 8 in the SI) to divide the configurational space into clusters that are likely to be indistinguishable by TR-SAXS (owing to its lower SNR, compared with static SAXS data). With maximum broadening, particles missing six dimers may be included within the complete \( T = 4 \) peak. This broadening becomes wider with incomplete and degenerate particles (lower \( D_c \) values). Similar effects were observed for the \( T = 3 \) symmetry.

Intermediates containing 35 dimers or less could not be subclassified into \( T = 3 \) or \( T = 4 \) symmetries owing to the similarity in their scattering curves (Figure S19) and their low mass fraction. Within the SNR of our data at the high \( q \)-range, the distinction between \( T = 3 \) and \( T = 4 \) particles is mostly limited to their different diameter. Therefore, the mass fraction at \( s = 90 \) may represent both well formed \( T = 3 \) particles and incomplete \( T = 3 \)-like particles with \( s > 90 \) that deformed and assumed an average diameter, close to that of a \( T = 3 \), as suggested by CDMS experiments.

**Averaged Intermediate Size**

The number averaged size \( \langle s \rangle \) of intermediates as a function of time, \( t \), is given by:

\[
\langle s \rangle (t) = \frac{\sum_{n,s,c} X_{n,s,c}(t)}{\sum_{n,s,c} s^{-1} X_{n,s,c}(t)}
\]  

(15)
where \( s^{-1}X_{n,s,c}(t) \) is the molar fraction of \( T_{n}^{s,c} \) intermediate structures at time \( t \).

**Acknowledgement**

We thank Corinne A. Lutomski and Martin F. Jarrold for sharing their CDMS data, presented in Figure S5. We thank Daniel Harries for very helpful discussions. We acknowledge the European Synchrotron Radiation Facility (ESRF) beamline ID02 (T. Narayanan and his team), the Desy synchrotron at Hamburg, beamline P12 (D. Svergun and his team), and Soleil synchrotron, Swing beamline (J. Perez and his team), for provision of synchrotron radiation facilities and for assistance in using the beamlines. This project was supported by the NIH (award number 1R01AI118933 to A.Z.). R.A. acknowledges support from the Kaye-Einstein Fellowship Foundation. U.R. and R.A. acknowledge financial support from the Israel Science Foundation (grant 656/17).

**References**

(1) 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. *Journal of the American Chemical Society* 2012, *134*, 8823–8830.

(2) Chevreuil, M.; Law-Hine, D.; Chen, J.; Bressanelli, S.; Combet, S.; Constantin, D.; Degrouard, J.; Möller, J.; Zeghal, M.; Tresset, G. Nonequilibrium self-assembly dynamics of icosahedral viral capsids packaging genome or polyelectrolyte. *Nature Communications* 2018, *9*, 3071.

(3) 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.

(4) 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. *Journal of the American Chemical Society* 2014, *136*, 3536–3541, PMID: 24548133.

(5) Garmann, R. F.; Comas-Garcia, M.; Gopal, A.; Knobler, C. M.; Gelbart, W. M. The assembly pathway of an icosahedral single-stranded RNA virus depends on the strength of inter-subunit attractions. *Journal of molecular biology* 2014, *426*, 1050–1060.

(6) Perlmutter, J. D.; Hagan, M. F. Mechanisms of virus assembly. *Annual review of physical chemistry* 2015, *66*.

(7) Hagan, M. F.; Elrad, O. M.; Jack, R. L. Mechanisms of kinetic trapping in self-assembly and phase transformation. *The Journal of chemical physics* 2011, *135*, 104115.

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

(9) 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. *Journal of the American Chemical Society* 2013, *135*, 15373–15381.

(10) Garmann, R. F.; Goldfain, A. M.; Manoharan, V. N. Measurements of the self-assembly kinetics of individual viral capsids around their RNA genome. *Proceedings of the National Academy of Sciences* 2019, *116*, 22485–22490.

(11) Borodavka, A.; Tuma, R.; Stockley, P. G. Evidence that viral RNAs have evolved for efficient, two-stage packaging. *Proceedings of the National Academy of Sciences* 2012, *109*, 15769–15774.

(12) Marchetti, M.; Kamsma, D.; Cazares Vargas, E.; Hernandez GarcÄњa, A.; van der Schoot, P.; de Vries, R.; Wuite, G. J. L.; Roos, W. H. Real-Time Assembly of Viruslike
Nucleocapsids Elucidated at the Single-Particle Level. *Nano Letters* 2019, 19, 5746–5753, PMID: 31368710.

(13) Zlotnick, A. Are weak protein–protein interactions the general rule in capsid assembly? *Virology* 2003, 315, 269–274.

(14) Rapaport, D. Molecular dynamics study of T= 3 capsid assembly. *Journal of biological physics* 2018, 44, 147–162.

(15) Schwartz, R.; Shor, P. W.; Prevelige, P. E.; Berger, B. Local Rules Simulation of the Kinetics of Virus Capsid Self-Assembly. *Biophysical Journal* 1998, 75, 2626 – 2636.

(16) Rapaport, D. Role of reversibility in viral capsid growth: a paradigm for self-assembly. *Physical Review Letters* 2008, 101, 186101.

(17) Zandi, R.; van der Schoot, P.; Reguera, D.; Kegel, W.; Reiss, H. Classical nucleation theory of virus capsids. *Biophysical journal* 2006, 90, 1939–1948.

(18) Nguyen, H. D.; Reddy, V. S.; Brooks, C. L. Deciphering the Kinetic Mechanism of Spontaneous Self-Assembly of Icosahedral Capsids. *Nano Letters* 2007, 7, 338–344.

(19) Oosawa, F.; Asakura, S. *Thermodynamics of the Polymerization of Protein*; Academic Press London, 1975.

(20) Razavi-Shearer, D. et al. Global prevalence, treatment, and prevention of hepatitis B virus infection in 2016: a modelling study. *The Lancet Gastroenterology and Hepatology* 2018, 3, 383 – 403.

(21) Venkatakrishnan, B.; Zlotnick, A. The Structural Biology of Hepatitis B Virus: Form and Function. *Annual Review of Virology* 2016, 3, 429–451, PMID: 27482896.

(22) Ning, X.; Nguyen, D.; Mentzer, L.; Adams, C.; Lee, H.; Ashley, R.; Hafenstein, S.; Hu, J. Secretion of genome-free hepatitis B virus–single strand blocking model for virion morphogenesis of para-retrovirus. *PLoS pathogens* 2011, 7, e1002255.
(23) Seeger, C.; Mason, W. S. Molecular biology of hepatitis B virus infection. *Virology* **2015**, *479*, 672–686.

(24) Patel, N.; White, S. J.; Thompson, R. F.; Bingham, R.; Weiß, E. U.; Maskell, D. P.; Zlotnick, A.; Dykeman, E. C.; Tuna, 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. *Nature microbiology* **2017**, *2*, 17098.

(25) Stray, S. J.; Bourne, C. R.; Punna, S.; Lewis, W. G.; Finn, M. G.; Zlotnick, A. A heteroaryldihydropyrimidine activates and can misdirect hepatitis B virus capsid assembly. *Proceedings of the National Academy of Sciences* **2005**, *102*, 8138–8143.

(26) Kondylis, P.; Schlicksup, C. J.; Brunk, N. E.; Zhou, J.; Zlotnick, A.; Jacobson, S. C. Competition between Normative and Drug-Induced Virus Self-Assembly Observed with Single-Particle Methods. *Journal of the American Chemical Society* **2019**, *141*, 1251–1260.

(27) Yuen, M.-F.; Agarwal, K.; Gane, E. J.; Schwabe, C.; Ahn, S. H.; Kim, D. J.; Lim, Y.-S.; Cheng, W.; Sievert, W.; Visvanathan, K.; Ruby, E.; Liaw, S.; Yan, R.; Connelly, E.; Cai, D.; Huang, Q.; Colonno, R.; Lopatin, U. A. Final Results of a Phase 1b 28-Day Study of ABI-H0731, a Novel Core Inhibitor, in Non-Cirrhotic Viremic Subjects with Chronic Hepatitis B. *HEPATOLOGY* **2018**, *68*, 46A–47A, Annual Meeting of the American-Association-for-the-Study-of-Liver-Diseases (AASLD) / Liver Meeting, San Francisco, CA, NOV 09-13, 2018.

(28) Wang, J. C.-Y.; Nickens, D. G.; Lentz, T. B.; Loeb, D. D.; Zlotnick, A. Encapsidated hepatitis B virus reverse transcriptase is poised on an ordered RNA lattice. *Proceedings of the National Academy of Sciences* **2014**, *111*, 11329–11334.

(29) Packianathan, C.; Katen, S. P.; Dann, C. E.; Zlotnick, A. Conformational Changes in
the Hepatitis B Virus Core Protein Are Consistent with a Role for Allostery in Virus Assembly. *Journal of Virology* **2010**, *84*, 1607–1615.

(30) Zhao, Z.; Wang, J. C.-Y.; Gonzalez-Gutierrez, G.; Venkatakshnan, B.; Asor, R.; Khaykelson, D.; Raviv, U.; Zlotnick, A. Structural Differences between the Woodchuck Hepatitis Virus Core Protein in the Dimer and Capsid States Are Consistent with Entropic and Conformational Regulation of Assembly. *Journal of Virology* **2019**, *93*.

(31) Wynne, S.; Crowther, R.; Leslie, A. The Crystal Structure of the Human Hepatitis B Virus Capsid. *Molecular Cell* **1999**, *3*, 771 – 780.

(32) Ceres, P.; Zlotnick, A. Weak protein–protein interactions are sufficient to drive assembly of hepatitis B virus capsids. *Biochemistry* **2002**, *41*, 11525–11531.

(33) 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.

(34) Kegel, W. K.; van der Schoot, P. Competing Hydrophobic and Screened-Coulomb Interactions in Hepatitis B Virus Capsid Assembly. *Biophysical Journal* **2004**, *86*, 3905 – 3913.

(35) Harms, Z. D.; Selzer, L.; Zlotnick, A.; Jacobson, S. C. Monitoring Assembly of Virus Capsids with Nanofluidic Devices. *ACS Nano* **2015**, *9*, 9087–9096, PMID: 26266555.

(36) 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.

(37) 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, PMID: 29125756.

(38) Zhou, J.; Kondylis, P.; Haywood, D. G.; Harms, Z. D.; Lee, L. S.; Zlotnick, A.; Jacobson, S. C. Characterization of Virus Capsids and Their Assembly Intermediates by Multicycle Resistive-Pulse Sensing with Four Pores in Series. *Analytical Chemistry* **2018**, *90*, 7267–7274.

(39) Ginsburg, A.; Ben-Nun, T.; Asor, R.; Shemesh, A.; Ringel, I.; Raviv, U. Reciprocal Grids: A Hierarchical Algorithm for Computing Solution X-ray Scattering Curves from Supramolecular Complexes at High Resolution. *Journal of Chemical Information and Modeling* **2016**, *56*, PMID: 27410762.

(40) Ginsburg, A.; Ben-Nun, T.; Asor, R.; Shemesh, A.; Fink, L.; Tekoah, R.; Levartovskiy, Y.; Khaykelson, D.; Dharan, R.; Fellig, A.; Raviv, U. D+: Software for High-Resolution Hierarchical Modeling of Solution X-Ray Scattering from Complex Structures. *J. Appl. Crystallogr.* **2019**, *52*, 219–242.

(41) Bourne, C. R.; Finn, M.; Zlotnick, A. Global structural changes in hepatitis B virus capsids induced by the assembly effector HAP1. *Journal of virology* **2006**, *80*, 11055–11061.

(42) Zlotnick, A. To Build a Virus Capsid: An Equilibrium Model of the Self Assembly of Polyhedral Protein Complexes. *Journal of Molecular Biology* **1994**, *241*, 59 – 67.

(43) Endres, D.; Miyahara, M.; Moisant, P.; Zlotnick, A. A reaction landscape identifies the intermediates critical for self-assembly of virus capsids and other polyhedral structures. *Protein Sci.* **2005**, *14*, 1518–1525.

(44) Henry, E.; Hofrichter, J. [8] Singular value decomposition: Application to analysis of experimental data. In *Numerical Computer Methods*; Methods in Enzymology; Academic Press, 1992; Vol. 210; pp 129 – 192.
(45) 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. *Biophysical journal* **2010**, *98*, 1065–1074.

(46) Rapaport, D. C. Self-assembly of polyhedral shells: A molecular dynamics study. *Phys. Rev. E* **2004**, *70*, 051905.

(47) Zlotnick, A.; Ceres, P.; Singh, S.; Johnson, J. M. A small molecule inhibits and misdirects assembly of hepatitis B virus capsids. *Journal of virology* **2002**, *76*, 4848–4854.

(48) Blanchet, C. E.; Spilotros, A.; Schwemmer, F.; Graewert, M. A.; Kikhney, A.; Jeffries, C. M.; Franke, D.; Mark, D.; Zengerle, R.; Cipriani, F.; Fiedler, S.; Roessle, M.; Svergun, D. I. Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). *Journal of Applied Crystallography* **2015**, *48*, 431–443.

(49) Round, A.; Felisaz, F.; Fodinger, L.; Gobbo, A.; Huet, J.; Villard, C.; Blanchet, C. E.; Pernot, P.; McSweeney, S.; Roessle, M.; Svergun, D. I.; Cipriani, F. BioSAXS Sample Changer: a robotic sample changer for rapid and reliable high-throughput X-ray solution scattering experiments. *Acta Crystallographica Section D* **2015**, *71*, 67–75.

(50) Hammersley, A. FIT2D: a multi-purpose data reduction, analysis and visualization program. *Journal of Applied Crystallography* **2016**, *49*.

(51) Asor, R.; Ben-nun Shaul, O.; Oppenheim, A.; Raviv, U. Crystallization, Reentrant Melting, and Resolubilization of Virus Nanoparticles. *ACS Nano* **2017**, *11*, 9814–9824.

(52) Ginsburg, A.; Shemesh, A.; Millgram, A.; Dharan, R.; Levi-Kalisman, Y.; Ringel, I.; Raviv, U. Structure of Dynamic, Taxol-Stabilized, and GMPPCP-Stabilized Microtubule. *The Journal of Physical Chemistry B* **2017**, *121*, 8427–8436.
(53) Van Vaerenbergh, P.; Léonardon, J.; Sztucki, M.; Boesecke, P.; Gorini, J.; Claustre, L.; Sever, F.; Morse, J.; Narayanan, T. An upgrade beamline for combined wide, small and ultra small-angle x-ray scattering at the ESRF. Proceedings of the 12th International Conference on Synchrotron Radiation Instrumentation (SRI2015). 2016; p 030034.

(54) Narayanan, T.; Sztucki, M.; Van Vaerenbergh, P.; Léonardon, 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. *Journal of Applied Crystallography* **2018**, *51*, 1511–1524.

(55) Jaynes, E. T. Information theory and statistical mechanics. *Physical review* **1957**, *106*, 620.

(56) Levine, R. D. An information theoretical approach to inversion problems. *Journal of Physics A: Mathematical and General* **1980**, *13*, 91.

(57) Shannon, C. E. A mathematical theory of communication (parts I and II). *Bell System Tech. J.* **1948**, *27*, 379–423.

(58) Kazama, J.; Tsujii, J. Maximum entropy models with inequality constraints: A case study on text categorization. *Machine Learning* **2005**, *60*, 159–194.

(59) Boyd, S.; Vandenberghe, L. *Convex optimization*; Cambridge university press, 2004.

(60) Ben-Nun, T.; Barak, A.; Raviv, U. Spline-based parallel nonlinear optimization of function sequences. *Journal of Parallel and Distributed Computing* **2016**, *93*, 132–145.

(61) MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. Berkeley, Calif., 1967; pp 281–297.
Graphical TOC Entry
Supplementary information

Rapidly Forming Early Intermediate Structures Dictate the Pathway of Capsid Assembly

Roi Asor,† Christopher John Schlicksup,‡ Zhongchao Zhao,‡ Adam Zlotnick,‡ and Uri Raviv*,†

†Institute of Chemistry and the Center for Nanoscience and Nanotechnology, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram, Jerusalem, 9190401, Israel
‡Department of Molecular and Cellular Biochemistry, Indiana University, Bloomington, Indiana 47405, United States

E-mail: uri.raviv@mail.huji.ac.il
Phone: +972-2-6586030. Fax: +972-2-566-0425

1. Static SAXS Fitting Using the Thermodynamic Model
Figure S1: Azimuthally integrated background subtracted solution small-angle X-ray scattering (SAXS) data, measured at 25 °C (blue symbols), of equilibrated Cp149 dimer at different total dimer and ammonium acetate concentrations, as indicated. The gray bars correspond to the measured standard deviations. The red curves are the best fitted SAXS model, based on the thermodynamic distribution of intermediates, given by Equation 1. The data and fitted curves correspond to the different points in the phase diagram, presented in Figure 1a. The insets show the $\chi^2$ value of the fitted model as a function of ammonium acetate concentration, for the indicated Cp149 concentrations.
2. Slow Time Scale Kinetics

Figure S2: Concentrations of free dimer, $T = 3$ capsid, and $T = 4$ capsid as a function of time following the onset of capsid assembly reaction at pH 7.5 and 25°C. Assembly of 5µM Cp149 dimer was initiated by increasing the ammonium acetate concentration from 20mM to 210mM. The concentrations were obtained by fitting the measured SAXS curves to a linear combination of the computed scattering contributions from the three atomic models.
Figure S3: Concentrations of free dimer, $T = 3$ capsid, and $T = 4$ capsid as a function of time, as in Figure S2. Assembly of 20 µM Cp149 dimer was initiated by increasing the ammonium acetate concentration from 20 mM to 210 mM.
Figure S4: Concentrations of free dimer, $T = 3$ capsid, and $T = 4$ capsid as a function of time, as in Figure S2. Assembly of 20 $\mu$M Cp149 dimer was initiated by increasing the ammonium acetate concentration from 20 mM to 510 mM.
3. Comparing CDMS Results with SAXS Signals

At Aggressive Assembly Conditions the Radius of Incomplete Intermediates Was Smaller Than the Radius of a Complete $T = 4$ Capsid

At $20 \mu M$ Cp149 and $510 \text{mM}$ ammonium acetate, after $\sim 24$–$48\text{h}$ CDMS showed that the dominant products contained 120-dimer ($T = 4$ like) and 90-dimer ($T = 3$ like) particles. A relatively small population of particles containing between 90 and 120 dimers was also observed by CDMS. The CDMS results agree with our SAXS results (Figure 1c), showing that the mass fraction of $T = 3$ particles increased to about 0.22 as the ammonium acetate concentration was increased to $510 \text{mM}$. The SAXS data, however, cannot exclude the presence of intermediates containing between 90 and 120 dimers with a total mass fraction of $\sim 0.05$ or less. At shorter timescales (between 10 and 90 min), however, CDMS measurements showed that the population of intermediates containing between 90 and 120 dimers was much higher. As SAXS may add additional information regarding the shape of the particles we compared the CDMS distribution against our SAXS data, measured 1 h after initiating the assembly reaction as in the CDMS experiments (Figure S5 in the SI). Figure S5 suggest that the CDMS detected intermediates were probably not only incomplete $T = 4$ capsids. When $T = 3$ particles were included in the model the fit to the SAXS data was extremely good, suggesting that the observed intermediates could included deformed particles with average dimensions closer to $T = 3$ capsid.
Figure S5: Comparing SXAS fitting results with alternative mass fraction distributions measured by charge detection mass spectrometry (CDMS) under similar assembly conditions.\textsuperscript{2} (a.) Azimuthally averaged background subtraction SAXS data (blue symbols) and standard deviations (gray bars) of 20 µM Cp149 dimer assembled at 25 °C by increasing the ammonium acetate concentration from 20 mM to 210 mM at pH 7.5. The measurement was performed 1 h following the salt concentration jump. The red curves correspond to the calculated scattering intensity of three different mass fraction distributions, labeled as I, II and III. (b.) The best fitted mass fraction distribution corresponding to the red curve in a, labeled as I. The fit was done with the three most populated states in our thermodynamic analysis which included free dimer, complete $T = 3$, and complete $T = 4$ capsids. (c.) The mass fraction distribution as a function of the number of dimers, $s$, measured by CDMS under similar assembly conditions (labeled as II). The CDMS data were used to calculate the red intensity curve in a, labeled as II. We assumed that all the complexes of size 120 or larger (blue columns) could be well represented by the complete $T = 4$ capsid. Complexes around the $s = 90$ peak (green columns) were assumed to be of $T = 3$ symmetry; either the most stable incomplete capsids (for $s < 90$) or complete $T = 3$ capsid (for $90 \leq s \leq 93$). Complexes whose $s$ values were between 94 and 119, shown as red columns, were modeled as the most stable $T = 4$ intermediates of size $s$. (d.) Identical mass fraction distribution as in c, however in this case, to compute the red scattering curve, labeled as III in a, we assumed that all the red column complexes ($s < 120$) were of $T = 4$ symmetry. At each size, the most stable $T = 4$ intermediate was taken into account. The integrated mass fraction of the blue columns was assumed to be the mass fraction of the complete $T = 4$ capsid.
4. Singular Value Decomposition (SVD) Analysis for Time-Resolved Data

Figure S6: Singular value decomposition (SVD) analysis of time time-resolved SAXS (TR-SAXS) data matrices of assembly reactions at pH 7.5 and 25°C (see Materials and Methods). Assembly reactions were performed with 25 µM Cp149 dimer in 163 mM and 513 mM ammonium acetate, or with 41 µM Cp149 dimer in 313 mM ammonium acetate. The residuals are plotted as a function of the number of singular vector components that were chosen to reconstruct the data matrix. The red lines indicate the cutoff where the residual is equal to 1.
Figure S7: Histograms of the distribution of residuals, $R_{q_i,t_j}$, as a function of sigma units, normalized to the measured noise. The number of the orthonormal singular vectors, $k$, used to reconstruct the TR-SAXS data matrix, for the three assembly conditions, is indicated. The Figure shows that at 163 mM ammonium acetate at least two vectors were required to get a Gaussian residual distribution. At 313 mM at least three vectors were required and at 513 mM ammonium acetate at least four vectors were needed.
5. Maximum Information Entropy Method.

Given a set of $M$ possible models, before any additional information is available, each of the possible states are expected to be equally probable. The information entropy of the distribution is then $S = -\sum_k p_k \ln p_k$ where, $p_k$ is the probability to find state $k$. When the distribution is uniform, $S$ is maximal. Our prior knowledge may, however, dictates that the expected distribution is non-uniform and assign probability $p_i$ to obtain the $i$th outcome. If, for example, there are $g_i$ equally probable ways to obtain outcome $i$, then $p_k = p_i/g_i$. If $N$ is the total number of different outcomes, $M = \sum_i g_i$, the information entropy is

$$S = -\sum_i g_i \left( \frac{p_i}{g_i} \ln \frac{p_i}{g_i} \right) = -\sum_i p_i \ln p_i + \sum_i p_i \ln g_i$$

In addition, we define $S^0 \equiv \ln \left( \sum_i g_i \right)$, and the prior probability to obtain outcome $i$ as $p_i^0 \equiv g_i / \sum_i g_i = g_i \cdot e^{-S^0}$, where $g_i$ is the degeneracy factor of the $i$th outcome (likewise $p_k^0 \equiv 1 / \sum_i g_i$). We then get that

$$S = S^0 - \sum_i p_i \ln \frac{p_i}{p_i^0}$$  \hspace{1cm} (S1)

$S^0$ is the maximal value of $S$, which is obtained when the actual distribution, $p_i$, is equal to the prior distribution $p_i^0$. In other words, the prior distribution is the distribution of maximum entropy before taking into account the new constraints imposed by the data (beyond the degeneracy factors that always present and are inherent to the problem and were taken into account in the prior distribution). When the actual probability distribution is different than the prior distribution (owing to the additional constraints that became available from the experiments), the entropy is lower than $S^0$. The term $-\sum_i p_i \ln p_i/p_i^0$ is the difference between the maximal and the actual value of the entropy and is therefore called the "entropy deficiency".

In this paper, we maximized the information entropy (Eq. (S1)), which takes into account the prior distribution, $p_i^0$, subject to the following three constrains.
1. All the probabilities are positive:

\[ p_i \geq 0 \quad \forall \ i, \]

2. The probability distribution is normalized:

\[ \sum_i p_i = 1, \]

3. The average signal should fit the experimental scattering data:

\[ -B_q \leq \left[ \sum_i p_i \cdot I_i(q) \right] - I_{\exp}(q) \leq A_q \]

where \( A_q > 0, \ B_q > 0. \)

Note that the assumptions used to compute the prior distribution impose constraints. It is convenient to solve the minimization problem for \(-S\). The inequality constrained minimization problem, can be solved by Lagrange multipliers method. The Lagrangian is defined as:

\[
L(p, a, b, \mu, A, B) = -S^0 + \sum_i p_i \ln \frac{p_i}{p_0^i} + \sum_q a_q \left( \sum_i p_i \cdot I_i(q) - I_{\exp}(q) - A_q \right) \\
+ \sum_q b_q \left( I_{\exp}(q) - \sum_i p_i I_i(q) - B_q \right) + \mu \left( \sum_i p_i - 1 \right)
\]

where \( a_q \) and \( b_q \) are the Lagrange multipliers that are associated with the constraints imposed by the measured scattering intensity, \( I_{\exp} \) at each \( q \). \( \mu \) is the Lagrange multiplier that is associated with the normalization condition. The solution to the minimization problem is given by the maximum of the Lagrangian. At the maximum:

\[
\left( \frac{\partial L}{\partial p_i} \right)_{p_i \neq j} = 0
\]
and we get
\[ \ln \frac{p_i}{P_i^0} + 1 + \sum_q a_q \cdot I_i(q) - \sum_q b_q \cdot I_i(q) + \mu = 0 \]
hence the probabilities are
\[ p_i = p_i^0 \exp \left\{ - (\mu + 1) - \sum_q (a_q - b_q) \cdot I_i(q) \right\} \]

From the normalization condition, \( \sum_i p_i = 1 \), we get:
\[ \exp \{-(\mu + 1)\} \sum_i p_i^0 \exp \left\{ - \sum_q (a_q - b_q) \cdot I_i(q) \right\} = 1 \]
hence,
\[ \sum_i p_i^0 \exp \left\{ - \sum_q (a_q - b_q) \cdot I_i(q) \right\} = \exp(\mu + 1) = \frac{Z}{e^{S^0}}, \]
where \( Z \) is the canonical partition function:
\[ Z \equiv \sum_i g_i \exp \left\{ - \sum_q (a_q - b_q) \cdot I_i(q) \right\}. \]
The distribution is
\[ p_i = \frac{e^{S^0} \cdot p_i^0}{Z} \exp \left\{ - \sum_q (a_q - b_q) \cdot I_i(q) \right\}. \] (S2)

By substituting the probability distribution, \( p_i \), back to the Lagrangian we get:
\[ L(\mathbf{a}, \mathbf{b}; \mathbf{A}, \mathbf{B}) = -S^0 - \ln Z - \sum_q (a_q - b_q) \cdot I_{\exp}(q) - \sum_q a_q \cdot A_q - \sum_q b_q \cdot B_q \]
If \( \forall q, A_q = B_q = \epsilon_q \) then
\[ L(\mathbf{a}, \mathbf{b}; \epsilon) = -S^0 - \ln Z - \sum_q (a_q - b_q) \cdot I_{\exp}(q) - \sum_q (a_q + b_q) \cdot \epsilon_q \] (S3)

Using the new Lagrangian we can define the Lagrange dual problem as:
\[
\max_{a,b} L(a, b; \epsilon)
\]
subject to \(a_q, b_q \geq 0 \ \forall \ q.
\]

At the maximum:

\[
\frac{\partial L(a, b; \epsilon)}{\partial a_q} = \frac{e^0}{Z} \sum_i I_i(q) \cdot p_0 \cdot \exp \left\{ -\sum_q (a_q - b_q) \cdot I_i(q) \right\} - I_{\text{exp}}(q) - \epsilon_q = 0
\]

(S4)

Similarly,

\[
\frac{\partial L(a, b; \epsilon)}{\partial b_q} = -\sum_i p_i \cdot I_i(q) - I_{\text{exp}}(q) - \epsilon_q = 0.
\]

Following the Karush-Kuhn-Tucker (KKT) conditions for inequality Lagrange minimization, the values of each \((a_q, b_q)\) pair must satisfy, at the optimal minimization point, one of the three conditions:

1. \(a_q > 0\) and \(b_q = 0\)
2. \(a_q = 0\) and \(b_q > 0\)
3. \(a_q = 0\) and \(b_q = 0\)

where cases 1 and 2 are considered active constraints and case 3 corresponds to an inactive constraint. In our case, an inactive constraint means that the measured value \(I(q)\) does not contribute additional information. As was previously shown by using the above conditions one can formulate a simpler problem by defining \(\lambda_q \equiv a_q - b_q\), and then \(|\lambda_q| = a_q + b_q\), where \(\lambda_q \in \mathbb{R}\), and by solving

\[
\min_{\lambda} - L(\lambda; \epsilon)
\]

(S5)

by rewriting Eq. S3 as,

\[
-L(\lambda; \epsilon) = S^0 + \ln \left[ \sum_i g_i \exp \left\{ -\sum_q \lambda_q \cdot I_i(q) \right\} \right] + \sum_q \lambda_q \cdot I_{\text{exp}}(q) + \sum_q |\lambda_q| \cdot \epsilon_q.
\]

(S6)
In the new minimization problem, described in Eqs. S5 and S6, we can define a constraint to be active if \( \lambda_q \neq 0 \) or inactive if \( \lambda_q = 0 \). The result of the minimization is a \( \lambda \) vector of Lagrange multipliers, which sets the required probability distribution,

\[
p_i = \frac{e^{S_0} \cdot p^0_i}{Z} \exp \left\{ - \sum_q \lambda_q \cdot I_i(q) \right\}.
\] (S7)

Note that if the prior distribution represents only the degeneracy factor that was obtained from the MC simulations, which effectively assumed zero association energy between capsid protein dimers, the argument of the exponential term in Eq. S7 can be linked to the physical association energy where, \( \sum_q \lambda_q \cdot I_i(q) = E_i/k_BT \), and \( E_i \) is the internal (association) energy of the \( i \)th state. The model (or the average state) is

\[
I_{\text{model}}(q) = \sum_i p_i I_i(q) = -\frac{\partial \ln Z}{\partial \lambda_q},
\]

and its variance is

\[
\Delta^2 I_{\text{model}}(q) \equiv \left[ \sum_i p_i I_i(q) \right]^2 - \sum_i p_i I_i^2(q) = \frac{\partial^2 \ln Z}{\partial \lambda_q^2}.
\]

Substituting Eq. S2 into Eq. S1 we get

\[
S = \ln Z + \sum_i \left( \sum_q \lambda_q \cdot I_i(q) \right) \cdot p_i
\]

The logarithm of the ratio between the final and the prior probability distribution, sometimes called the surprisal, is then:

\[
\ln \frac{p_i}{p^0_i} = S^0 - \ln Z - \sum_q \lambda_q \cdot I_i(q)
\]

The entropy of the system is then obtained by the average surprisal and the degeneracy
factor,
\[ k_B \left( S^0 - \sum_i p_i \ln \frac{p_i}{p_0^i} \right) = k_B \left( \ln Z + \sum_i \sum_q \lambda_q \cdot I_i(q) \cdot p_i \right) = S, \]
where \( \sum_i \sum_q \lambda_q \cdot I_i(q) \cdot p_i = E/k_B T \), and \( E \) is the average internal energy.

Thermodynamic Constraints

Given the additional thermodynamic constraints regarding the free energy gain in forming subunit-subunit interaction and a constraint regarding the expected mean number of dimers in an aggregate, we can write the new Lagrangian as,

\[ L(p, a, b, \mu, A, B) = -S^0 + \sum_i p_i \ln \frac{p_i}{p_0^i} + \sum_q a_q \left( \sum_i p_i \cdot I_i(q) - I_{\text{exp}}(q) - A_q \right) \\
+ \sum_q b_q \left( I_{\text{exp}}(q) - \sum_i p_i I_i(q) - B_q \right) + \mu \left( \sum_i p_i - 1 \right) + \mu_E \left( \sum_i E_i p_i - \langle E \rangle \right) + \mu_n \left( \sum_i n_i p_i - \langle n \rangle \right). \]

At the maximum:
\[ \left( \frac{\partial L}{\partial p_i} \right)_{p_{i,j}} = 0 \]
and we get
\[ \ln \frac{p_i}{p_0^i} + 1 + \sum_q a_q \cdot I_i(q) - \sum_q b_q \cdot I_i(q) + \mu + \mu_E \cdot E_i + \mu_n \cdot n_i = 0 \]

and the probabilities are given by
\[ p_i = p_0^i \exp \left\{ -(\mu + 1) - \mu_E \cdot E_i - \mu_n \cdot n_i - \sum_q \lambda_q \cdot I_i(q) \right\}. \]

From the normalization condition, \( \sum_i p_i = 1 \), we get:
\[ \exp \left\{ -(\mu + 1) \right\} \sum_i p_0^i \exp \left\{ -\mu_E \cdot E_i - \mu_n \cdot n_i - \sum_q \lambda_q \cdot I_i(q) \right\} = 1 \]
hence,

\[
\sum_i p^0_i \exp \left\{ -\mu_E \cdot E_i - \mu_n \cdot n_i - \sum_q \lambda_q \cdot I_i(q) \right\} = \exp(\mu + 1) = \frac{Z}{e^{S_0}},
\]

where \(Z\) can be considered as a modified grand canonical partition function:

\[
Z \equiv \sum_i g_i \exp \left\{ -\mu_E \cdot E_i - \mu_n \cdot n_i - \sum_q \lambda_q \cdot I_i(q) \right\} = \sum_i \exp \left\{ -\Theta_i \right\} \exp \left\{ - \sum_q \lambda_q \cdot I_i(q) \right\}.
\]

where, \(\Theta_i = \mu_E \cdot E_i - \ln g_i + \mu_n \cdot n_i\). The distribution is

\[
p_i = \frac{p^0_i}{Z} \exp \left\{ - \sum_q \lambda_q \cdot I_i(q) \right\}.
\]

where, \(p^0_i = \exp \left\{ -\Theta_i \right\} = e^{S_0} \cdot p^0_1 \cdot \exp \left\{ -\mu_E \cdot E_i - \mu_n \cdot n_i \right\}\), and the multipliers \(\mu_E\) and \(\mu_n\) are associated with the parameters \(-\frac{1}{k_B T}\) and \(\frac{\mu_1}{k_B T}\) respectively.

6. Results of Maximum Information Entropy Optimization

Priors
Figure S8: The two prior distributions that were used for the maximum entropy analysis of our TR-SAXS data, projected onto the $s$ axis. The prior distribution assuming weak dimer-dimer association free energy ($5k_B T$) is plotted on a linear (a.) or logarithmic (b.) scale. In this case, the dominant state is free dimer because the association free energy is significantly lower than the minimal association free energy gain required for assembly. The prior distribution assuming strong dimer-dimer association free energy ($8.5k_B T$) in plotted on a linear (c.) and logarithmic (d.) scales. The plots correspond to the expected state of the system at (or near) equilibrium. When the association free energy is high, the predominant states are the free dimer and full capsids. For both prior distributions the value of $\Delta F_4$ and $\alpha \equiv \Delta F_3/\Delta F_4$ are indicated in the figure. The peaks at $s = 90$ and $s = 120$ correspond to the $T = 3$ and $T = 4$ full capsids.
Fitting Results

The following Figures show time resolved scattering results (blue symbols and gray error bars) and the results of the maximum informational entropy optimization (red curves) along the assembly process, at three assembly conditions. The quality of the fits can be seen both in the I vs. q curves or by the distribution of the normalized residuals, $r(q)$, given by

$$r(q) = \frac{(I_{\text{exp}}(q) - I_{\text{model}}(q))}{\sigma_m(q)}$$  \hspace{1cm} (S9)$$

where $\sigma_m$ is the measured standard deviation at $q$.

Assembly at 163 mM Ammonium Acetate
Figure S9: The entire azimuthally integrated background-subtracted 163 mM ammonium acetate TR-SAXS data (blue symbols), measurement error (gray bars), and best fitted models (red curves). Selected time points are shown in Figure 3a). The data were fitted by the maximum entropy optimization method, using the prior distributions presented in Figure S8. The residuals between the best fitted model and the data were normalized with respect to the measurement standard deviation, $\sigma(q_i)$, are plotted on the right side of each panel in $\sigma$ units.
Figure S10: Maximum entropy fitting results of TR-SAXS data obtained by repeating the assembly reaction in 163 mM ammonium acetate (shown in Figure S9).
Figures S11: Maximum entropy fitting results of TR-SAXS data obtained from the assembly reaction in 313 mM ammonium acetate (as in Figure S9). Selected time points are shown in Figure 3b.
Assembly at 513 mM Ammonium Acetate

Figure S12: Maximum entropy fitting results of TR-SAXS data obtained from the assembly reaction in 513 mM ammonium acetate (as in Figure S9). Selected time points are shown in Figure 3c.
Figure S13: Maximum entropy fitting results of TR-SAXS data obtained by repeating the assembly reaction in 513 mM ammonium acetate (shown in Figure [S12]), focusing on the first 4 sec of the reaction.
Figure S14: Maximum entropy fitting results of TR-SAXS data during the first 4 sec of the assembly reaction in 513 mM ammonium acetate. The data were acquired by repeating the assembly experiment shown in Figure S13.
Figure S15: Maximum entropy fitting results of TR-SAXS data during the first 4 sec of the assembly reaction in 513 mM ammonium acetate. The data were acquired by repeating the assembly experiment shown in Figure S14.
Figure S16: Maximum entropy fitting results of TR-SAXS data during the first 4 sec of the assembly reaction in 513 mM ammonium acetate. The data were acquired by repeating the assembly experiment shown in Figure S14.
7. At 313 mM Ammonium Acetate the Observed $D_{10}$ Intermediate is Located at a Local Free Energy Minima

Figure S17: The grand canonical free energy change along the minimum free energy path ($D_{C} = 1$) as a function of $s$, for $T = 4$ intermediates at 313 mM ammonium acetate concentrations and $t = 2$ sec. The chemical potential at this time point was calculated by $\mu_{1,t=2\text{sec}} = k_B T \ln (0.7X_{\text{total}})$, where 0.7 is the mass fraction of dimers at this time point and $X_{\text{total}}$ is the total molar fraction of Cp149 dimers.
8. Clustering of SAXS Signals

Figure S18: Classification of SAXS models, whose $s$ values were in the top 33% of the $T=3$ and $T=4$ capsid symmetries, into clusters. The clustering procedure was applied (as explained in Material and Method section) with respect to the measured noise level in the TR-SAXS data. Clusters are shown in the $D_c$ plane and are distinguished by different colors. Cartoons of the full $T=3$ capsid and the three structures that delimit the cluster of the full $T=4$ capsid are shown.
Figure S19: Comparison between mid-size intermediates having either $T = 3$ or $T = 4$ symmetry. We compare between intermediates containing 7 or 35 dimers ($D_7$ and $D_{35}$), corresponding to the lower and upper limit of the intermediates peak, observed during the assembly at 513 mM ammonium acetate. The selected intermediates are the most compacted and stable at each size and $T$ number (the most connected intermediates). The gray error bars indicate the noise level about the average intensity of the two structure in each panel. The noise level was based on the early time points of the reaction (Figure 4c), where the concentration of intermediates was maximal. The figure shows that the differences between the expected scattering intensities of $T = 3$ and $T = 4$ of mid-size intermediates are within the noise level.

References

(1) 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.

(2) 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.

(3) Shannon, C. E. A mathematical theory of communication (parts I and II). *Bell System Tech. J.* 1948, *27*, 379–423.
(4) Levine, R. D. An information theoretical approach to inversion problems. *Journal of Physics A: Mathematical and General* **1980**, *13*, 91.

(5) Kazama, J.; Tsujii, J. Maximum entropy models with inequality constraints: A case study on text categorization. *Machine Learning* **2005**, *60*, 159–194.

(6) Boyd, S.; Vandenberghe, L. *Convex optimization*; Cambridge university press, 2004.

(7) Dudik, M.; Phillips, S. J.; Schapire, R. E. Performance guarantees for regularized maximum entropy density estimation. International Conference on Computational Learning Theory. 2004; pp 472–486.
