



Illuminating an invisible state of the HIV-1 capsid protein CTD dimer using ^{19}F NMR and weighted ensemble simulations

Darian T. Yang^{a,b,c} , Lillian T. Chong^{c,1} , and Angela M. Gronenborn^{b,c,1}

Affiliations are included on p. 10.

Edited by Michael Summers, Department of Chemistry and Biochemistry, University of Maryland Baltimore County, Baltimore, MD; received October 4, 2024; accepted January 6, 2025

The HIV-1 capsid protein (CA) assembles into a conical shell during viral maturation, encasing and protecting the viral RNA genome. The C-terminal domain (CTD) of the two-domain capsid protein dimerizes, and this dimer connects individual chains in the mature capsid lattice. Previous NMR studies have shown that different dimer arrangements can be formed by isolated capsid protein chains and in assembled capsid lattices; however, the dynamics and functional relevance of these alternate dimers are unknown. To explore the conformational landscape of the CA-CTD dimer, we carried out atomistic molecular dynamics simulations using the weighted ensemble path sampling strategy, generating an ensemble of conformations. Focusing on the two dimer forms previously observed via solution NMR, we refined the conformational ensemble to highlight two metastable states using a Markov state model. Experimentally, we measured the interconversion rates between the two alternate dimers using ^{19}F NMR, and these rates showed good agreement with the interconversion rates derived from the simulations. After identifying the key interactions that distinguish the dimer states, the alternate dimer was further experimentally verified through disulfide crosslinking. Our results demonstrate the advantages of pairing weighted ensemble path sampling with ^{19}F NMR to gain atomistic insights into the hidden dimer state of the HIV-1 capsid protein.

HIV-1 | capsid | protein conformational dynamics | weighted ensemble path sampling | ^{19}F NMR

The mature HIV-1 virion contains a conical, ovoid structure known as the capsid, which protects the viral genome and shields it from detection by the host cell. The capsid is essential for successful infection and plays pivotal roles throughout the HIV-1 lifecycle, from cytoplasmic trafficking and nuclear entry to genome integration (1). The assembled capsid is composed of 1,200 to 1,500 individual capsid protein (CA) chains and is arranged in a hexameric lattice into which 12 pentameric units (Fig. 1) are integrated to allow for closure of the ovoid (2). The individual CA chains in the lattice are held together via the C-terminal domain (CTD) dimer, located on the interior side of the shell, forming the flexible core. Dimerization of the CA-CTD is pivotal for maintaining the overall shape and structure of the assembled capsid, knitting together each subunit. CA-CTD dimers also hold together the immature Gag lattice, and during viral maturation, stochastic cleavage between CA and spacer peptide 1 initiates transformation of the immature Gag lattice into the mature capsid lattice (3, 4). For both the immature and mature lattice, mechanistic insights into the dynamics of the assembly process are still unknown.

The CA polypeptide chain consists of two domains, a 145 amino acid N-terminal domain (NTD) and an 80 amino acid CTD, connected by a five-residue flexible linker. In solution, CA exists in a monomer-dimer equilibrium, dimerizing via the CTD (5, 6). The isolated CA-CTD can adopt different quaternary structures (5–11) depending on the environment, and in solution, several different dimers coexist (5). While the structure of the major CA-CTD dimer (D1) in solution has been determined by NMR (6), structural and dynamic details of the minor population (D2) have been elusive since this alternate state is only populated to 10% at best. The equilibrium population of the alternate state can be modulated by shortening the amino acid sequence or by lowering solution pH levels (5). However, even in conditions with the highest equilibrium population of D2, the ensemble averaged experimental NMR data are not sufficiently resolved to determine the second dimer conformation with confidence, since extensive resonance overlap precludes complete resonance assignments in traditional ^1H , ^{15}N , and ^{13}C heteronuclear spectra.

Significance

Protein conformational dynamics are critical for understanding biological function. To derive atomic-resolution details of an alternate, “invisible” dimeric state of the HIV-1 capsid protein C-terminal domain (CTD), we used ^{19}F NMR and extended-timescale molecular dynamics simulations. We obtained experimentally validated structures and pathways of interconversion between multiple protein conformations. Our methodology leverages the efficiency of weighted ensemble path sampling to access long timescales, aligning with the experimental timescales probed by high-sensitivity ^{19}F NMR. We identified key differences between the two alternate dimer states, and we expect our integrated methodology to be applicable to other systems with alternate states that cannot be characterized by experiment or simulation alone.

Author contributions: D.T.Y., L.T.C., and A.M.G. designed research; D.T.Y. performed research; D.T.Y., L.T.C., and A.M.G. analyzed data; and D.T.Y., L.T.C., and A.M.G. wrote the paper.

Competing interest statement: L.T.C. serves on the scientific advisory board of OpenEye Scientific Software.

This article is a PNAS Direct Submission.

Copyright © 2025 the Author(s). Published by PNAS. This article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](#).

¹To whom correspondence may be addressed. Email: ltchong@pitt.edu or amg100@pitt.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2420371122/-/DCSupplemental>.

Published February 18, 2025.

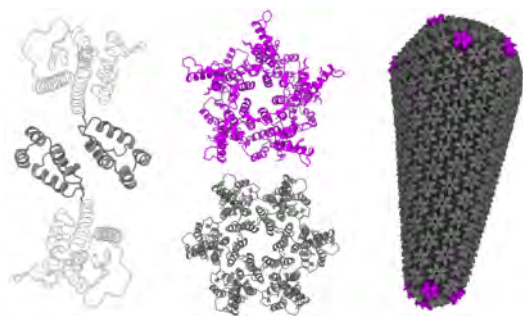


Fig. 1. The HIV-1 capsid protein C-terminal domain (CA-CTD) dimer. The capsid is made up of repeating units of the self-assembling CA. Dimerization of CA through the CTD holds the capsid together and connects adjacent hexameric (gray) and pentameric (magenta) subunits. The CA-CTD dimer is shown in gray ribbon representation and the CA-NTD is in white ribbon representation. The fully assembled capsid (15) consists of a hexameric lattice, with exactly 12 pentameric subunits which allow for closure of the ovoid.

Here, we report on the structure and dynamics of the previously identified, but not further characterized, alternate CA-CTD D2 conformational state. We used weighted ensemble path sampling (12, 13) to extensively explore the conformational landscape of the CA-CTD, ultimately identifying two major metastable states using a Markov state model (14). We experimentally validated our computational findings using ^{19}F NMR and determined the interconversion rates between the two dimer states. Structurally, we identified a key interaction network that differentiates the two dimer interfaces in our simulations, and guided by the simulation data, we introduced cysteines into positions of potential close contact in the CA-CTD that permitted us to experimentally trap a D2-like conformer of CA-CTD. This alternate D2 dimer state possesses a twisted subunit interface that may allow the lattice to rearrange during the assembly, maturation, or disassembly stages of the viral life cycle.

Results and Discussion

Weighted Ensemble Simulations Capture D1/D2 Exchange. We initially attempted to sample the D1 to D2 interconversion process using all-atom molecular dynamics (MD) simulations, starting from the previously solved NMR structure of D1 and from one of the CA-CTD [PDB ID: 1A43 (8)] crystal structures. The 1A43 crystal structure conformation, compared to the NMR structure, is slightly different in the relative orientation of the monomeric units of the dimer from the NMR structure (*SI Appendix, Fig. S1A*). We initially thought that this dimer quaternary state may resemble D2, and we used it as the initial candidate in our simulations. Despite multiple 1- μs conventional MD simulations starting from either the NMR or crystal structure, the conformational ensembles remained similar to the respective starting structure (*SI Appendix, Fig. S1*), with one exception: one simulation which started from the crystal structure began transitioning to the NMR structure-like state (*SI Appendix, Fig. S1B*).

Overall, this result was not unexpected, given that the timescales of the D1 to D2 interconversion processes are on the order of milliseconds (5), beyond the reach of conventional MD simulations. Such longer timescales, however, can be accessed using the weighted ensemble (WE) path sampling strategy (12, 13). Therefore, we carried out WE simulations using the open-source and highly scalable Weighted Ensemble Simulation Toolkit with Parallelization and Analysis (WESTPA) software package (16). While many enhanced sampling methods are

available for exploring the conformational dynamics of proteins (17–19), WE is advantageous as an efficient method for sampling continuous pathways with rigorous kinetics for the process of interest. The WE strategy involves running multiple weighted trajectories in parallel and iteratively applying a resampling procedure at fixed time intervals τ (WE iteration). Typically, configurational space is first divided into a set of bins along a chosen progress coordinate. The resampling procedure then replicates promising trajectories that make transitions to newly visited bins and prunes trajectories that occupy previously visited bins. Overall, this process aims to evenly distribute sampling across the progress coordinate space by maintaining a constant number of trajectories in each bin. Trajectory weights are rigorously tracked during the resampling process and the sum of the weights is always maintained as one, ensuring that no statistical bias is introduced (20). Therefore, rate constants between any pair of arbitrary states can be directly calculated, and ideally, the result is a complete ensemble of pathways with corresponding probabilities (21).

To explore the conformational landscape of the CA-CTD dimer, in search of the alternate D2 state, we performed WE simulations, starting from the final coordinates of our conventional MD simulations. An initial challenge in this endeavor was to identify an effective progress coordinate for the D1 to D2 conformational transition, given that the structure of the target D2 conformation was completely unknown. We explored several trial coordinates such as the RMSD of the backbone heavy atoms with respect to the starting structure [2KOD (6)], the solvent-accessible-surface area, and the radius-of-gyration. Unfortunately, WE simulations using these progress coordinates reported predominantly on unbinding or disruption of the dimer interface, making these trial coordinates poor choices for assessing the conformational transition.

To focus WE simulations on sampling conformational interconversions, we exploited the differences between the NMR and crystal structures of the CA-CTD dimer. A key difference between these structures pertains to the crossing angle of helix 9 of each monomer relative to the dimer interface (*SI Appendix, Fig. S1A*). The two helices form a V-shape with crossing angles of 56° and 38° in the NMR and 1A43 crystal structures, respectively. The vertex of this angle is positioned

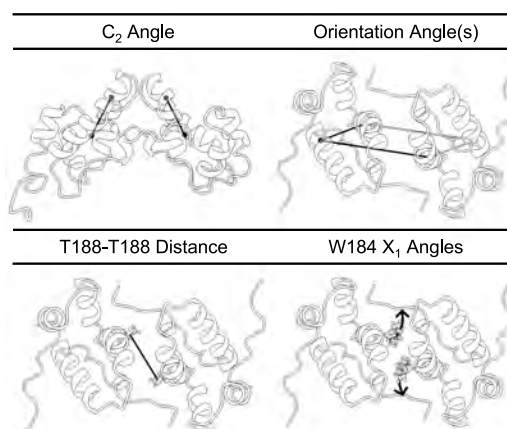


Fig. 2. Various progress coordinates used in WE simulations for evaluating the interconversion between the CA-CTD D1 and D2 states. The C_2 angle, orientation angle(s), T188–T188 distance, and W184 χ_1 angles are each indicated by black arrows or lines in the context of the NMR structure of the CA-CTD dimer [PDB ID: 2KOD (6)], which is shown in a transparent representation.

on the C_2 symmetry axis (6, 8), and we refer to this angle herein as the " C_2 angle" (Fig. 2). Using this coordinate for our WE simulations, we were able to sample the interconversion between the crystal structure and the NMR structural ensembles. However, the interconversion process was too fast compared to the experimental interconversion rates (5), effectively eliminating the crystal structure as a candidate for the alternate D2 state.

Since the C_2 angle performed well as a progress coordinate without promoting disruption of the dimer interface, we explored another angle that was designed to effectively track a wide range of orientations between the two individual chains in the CA-CTD, but in a direction perpendicular to the C_2 angle. We refer to this angle as the "orientation angle" (Fig. 2). This orientation angle describes changes of each monomer with respect to the other monomer and effectively results in a quadrilateral arrangement, permitting exploration of symmetric and asymmetric changes around the dimer interface. Using this angle as a progress coordinate, we were able to sample a wide range of alternate CA-CTD dimer conformations beyond the crystal structure conformation (8), the NMR structure conformation (6), as well as the dimer arrangement observed in the assembled capsid (15) (Fig. 3A). Importantly, WE simulations sampled conformations diverging from the initial NMR structure in two directions: one toward smaller orientation angles, passing the crystal structure, and the other toward larger angles, leading to an asymmetric state (*SI Appendix, Fig. S2*). To sample smaller orientation angles, it was most efficient to use a one-dimensional progress coordinate that considered only the larger value of the two orientation angles during each WE iteration; this approach prevented the WE simulations from narrowly focusing on highly asymmetric states, enabling the identification of a more diverse and viable CA-CTD dimer ensemble with potential conformations for the D2 state.

Identification of Two Main Routes of D1/D2 Exchange. Based on our previous NMR results, only two dimer populations were seen (5). However, additional states were identified by our WE simulations, and we assessed whether they were energetically

stable enough to remain occupied on the μ s timescale, or if the states were low-probability extrema that would only be transiently occupied. We used a diverse set of end states from our WE simulations to seed 68 conventional MD simulations run for 1.33 μ s each, totaling 90 μ s of aggregate simulation time after the initial 25 μ s of aggregate WE simulation time. Although we started with four to five dimer states, the simulations eventually settled into two major state populations (Fig. 3B), which aligned with the NMR results. Therefore, these final states are our most likely candidates for D1 and D2 (Fig. 3C).

To validate the stability of the two states, we used the 90 μ s of MD simulation data to construct a Markov state model (MSM). MSMs are useful for modeling the long-timescale dynamics of molecular systems as a series of memoryless, probabilistic jumps between a discrete set of representative conformations (14, 22, 23). After constructing a transition probability matrix between the discrete conformations, a coarse-grained set of metastable states can be constructed by clustering together multiple, kinetically similar structures (24). The stationary distribution of the transition probability matrix provides an estimate of the true equilibrium state distribution and can be used to reweight the input simulation trajectories toward more accurate equilibrium state populations if the sampling provided by the input trajectories is sufficient (22).

To build the MSM, we selected the orientation angles, the C_2 angle, and the distance between the two monomers at the T188 residue in each chain (T188–T188 distance) as parameters to describe the state space (Fig. 2). Since W184 and M185 in the CA-CTD dimer interface are crucial for dimer formation (10), we inspected the behavior of W184 and other key residues in the CA-CTD dimer interface. Interestingly, a change in the χ_1 dihedral angle of W184 (Fig. 2) is observed between D1 and D2, and for this reason, this angle was also included in the MSM featurization. Both visual inspection and the MSM suggested the presence of two major states, bridged by a metastable intermediate (*SI Appendix, Fig. S3*). After reweighting the input features based on the MSM stationary distribution, the updated energy

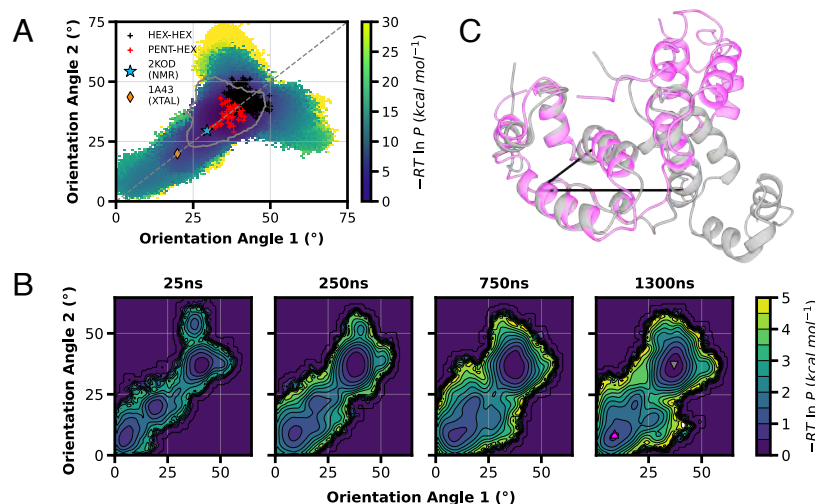


Fig. 3. Detection of alternate HIV-1 capsid protein CTD dimer states using WE simulations. (A) Probability distribution of data from five independent WE simulations (25 μ s of aggregate simulation time) of the WT CA-CTD dimer as a function of orientation angles 1 and 2. Structures of the CA-CTD dimer in solution [PDB ID: 2KOD (6)], the crystal [PDB ID: 1A43 (8)], and the assembled capsid (15) hexamer-hexamer and pentamer-hexamer interfaces are marked by different symbols. Structures along the diagonal dashed line are symmetric. The gray contour outline represents an aggregate of 5 μ s of conventional MD simulations initiated from the NMR structure. (B) Further exploration of stable states using conventional MD simulations (90 μ s of aggregate simulation time), seeded from the WE conformational landscape. Contour plots of the probability distributions that converge to two stable states after 1.3 μ s of simulation time for each of the 68 conventional MD simulations, with D1 and D2 marked by gray and magenta triangles, respectively. Each contour line represents a 0.5 kcal/mol interval. (C) Representative structures of the D1 (gray representation) and D2 (magenta representation) states sampled by the conventional MD simulations in panel (B). The orientation angle of the structurally aligned monomer corresponding to D1 is shown with solid black lines.

landscape (*SI Appendix, Fig. S4*) exhibits a minimally more stable D2 state and a slightly less favorable intermediate, suggesting that transitions within the D2 and intermediate states may be under sampled. In all states, the backbone of each CA-CTD chain remains similar to that of the original D1 monomer within the dimer, while the orientation of the monomers relative to each other changes. The MSM aligns with our visual inspection and initial analysis of the CA-CTD dimer state space.

We observed two main routes for interconversion between the D1 and D2 states in the continuous WE pathways and in the discretized state transitions of the MSM, one of which we refer to as the direct route and the other as the indirect route. While both routes require symmetric reductions in orientation angles, the indirect route features a slower D1 to D2 interconversion that involves the formation of a low C_2 angle metastable intermediate, which is very similar to the crystal structure. The indirect route also involves stepwise changes in the W184 χ_1 angles along with a C_2 angle change, where one W184 χ_1 angle in one of the CA-CTD monomer units flips from -180° to -75° . This is coupled with the C_2 angle changing from 60° to 35° . Only after the other W184 χ_1 angle flips, the dimer is able to fully convert to the D2 conformation (*SI Appendix, Fig. S5* and *Movie S1*). The direct route exhibits a faster D1 to D2 interconversion, with an almost concerted change in the W184 χ_1 dihedral angles. This route does not require a decrease in the C_2 angle, and thus does not visit the metastable intermediate of the indirect route (*SI Appendix, Fig. S5* and *Movie S2*). These atomistic pathways from WE simulations may be useful for targeting or modulating the populations of the CA-CTD dimer states.

The Timescales of D1/D2 Exchange. To compare the timescales of the exchange process between the CA-CTD D1 and D2 states, we used ^{19}F NMR to obtain reference rates of D1/D2 interconversion. While NMR spectroscopy using traditional ^{15}N - ^1H HSQC experiments can differentiate between the D1 and D2 states, characterizing the alternate D2 state (*SI Appendix,*

Fig. S6A) is challenging due to severe resonance overlap. In contrast, the sensitivity of the fluorine-19 nucleus to small structural and electronic changes, along with its inherently large chemical shift range (25–28), overcomes this limitation. Indeed, the two resonances for the individual conformers of the dimer are well resolved in the fluorine spectrum of 4F-Trp CA-CTD (*Fig. 4A*). Also note, the introduction of a fluorine atom at the C-4 position of the W184 indole ring does not perturb the overall structure of the CA-CTD. WT and fluorinated CA-CTD ^1H - ^{15}N HSQC spectra are essentially identical (*SI Appendix, Fig. S6A*), and only minor chemical shift differences are seen for resonances associated with residues close to the W184 fluorine atom (*SI Appendix, Fig. S6B*). Furthermore, we made certain that $\geq 95\%$ of the protein was dimeric, adjusting the concentration of the sample to at least 1.2 mM. The dimerization constant for CA-CTD WT was previously been determined to be 10 μM (6, 10, 29), and for the 4F-Trp CA-CTD, we measured a dimerization constant of 23.2 μM (*SI Appendix, Fig. S7*). Exchange between the D1 and D2 dimers was evaluated using ^{19}F - ^{19}F exchange spectroscopy (EXSY) experiments at multiple mixing times (*Fig. 4B*), and to characterize the kinetics of the D1/D2 exchange process (*Fig. 4C*), the peak intensity ratios were fit to the following equation (30):

$$\frac{I_{12}}{I_{11}}(t_m) = \frac{(1 - e^{-(k_{12}+k_{21}) \cdot t_m}) \cdot k_{12}}{k_{21} + k_{12} \cdot e^{-(k_{12}+k_{21}) \cdot t_m}}, \quad [1]$$

where the forward and backward rate constants are:



To improve the match between simulated and experimentally measured rates, we also carried out WE simulations under nonequilibrium steady-state conditions (ssWE). This entails treating the initial and target states as source and sink states, respectively, and “recycling” trajectories that reach the sink state

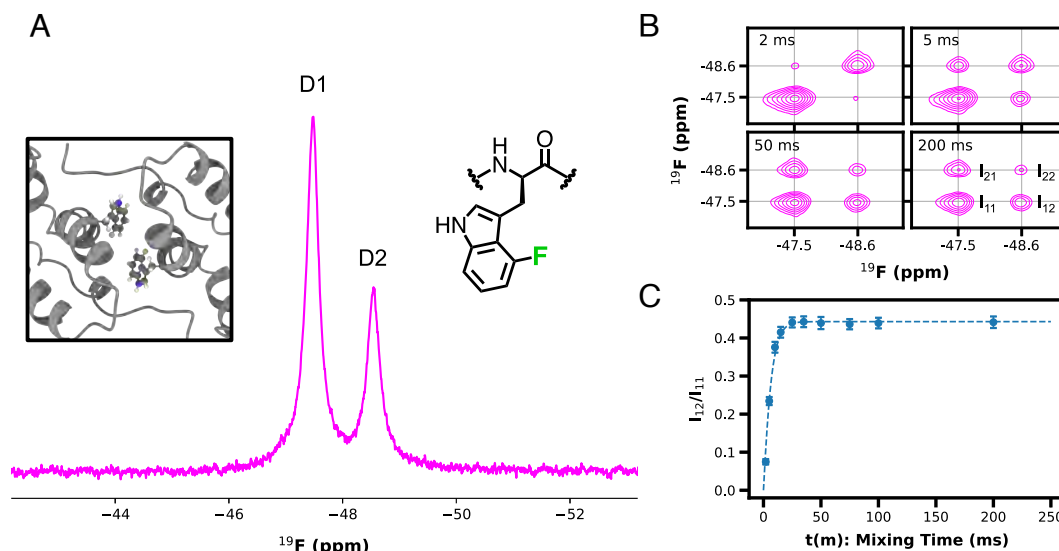


Fig. 4. ^{19}F NMR reports on the D1 and D2 states of the CA-CTD dimer and their rates of interconversion. (A) One-dimensional ^{19}F NMR spectrum of 4F-W184 CA-CTD. Individual resonances for the D1 and D2 species are observed at an approximately 0.65:0.35 ratio (298 K, pH 6.5). The dimer interface [PDB ID: 2KOD (6)] is shown in the boxed region (Trp 184 in stick representation with the F atom colored green), and the chemical structure of 4F-Trp is shown at the *Right*. (B) 2D ^{19}F - ^{19}F exchange (EXSY) spectra for several mixing times. Autopeaks represent magnetization that does not exchange during the mixing time, while the exchange cross peaks arise from magnetization exchange between the D1 and D2 conformations. Cross peak labels are shown on the 200 ms spectrum. (C) Intensity ratios of the I_{12} exchange cross peak over the I_{11} autopeak for multiple mixing times. Data were fit to a two-site exchange model to estimate the forward and backward rate constants for the D1/D2 interconversion process. Error bars were derived from the signal-to-noise ratio of the ^{19}F - ^{19}F EXSY spectra.

by initiating a new trajectory from the source states with the same trajectory weight. However, given that no experimental structure for the D2 state is available, a target sink state could not be strictly defined using an RMSD coordinate. We thus defined the target state using multiple structural features: both orientation angles, the C₂ angle, the T188–T188 distance, and both W184 χ_1 angles. The endpoints of our previous CA-CTD dimer WE simulations were used as starting states for the D1→D2 ssWE simulations, while in the reverse direction, D2→D1 ssWE simulations were seeded from the recycled endpoints of the forward direction D1→D2 ssWE simulations. For the 4F-Trp CA-CTD, we generated a model using our recently developed AMBER ff15ipq (33) force field parameters for fluorinated, aromatic amino acids commonly used with ¹⁹F NMR (34), and ran an independent, replicate set of conventional MD, WE, and ssWE simulations (SI Appendix, Fig. S8).

Interconversion events between the D1 and D2 states in both the forward (k_{12}) and backward (k_{21}) directions were captured in the WE simulations, and rate constants were estimated both directly from WE simulations and from the MSM. As illustrated in Fig. 5, the WE simulation rate estimates for the fluorinated CA-CTD are close to the experimental ¹⁹F NMR values, while the WE simulated rate for the WT CA-CTD is smaller than that of the fluorinated variant, albeit still within the same order of magnitude. The MSM rate constants for both interconversion directions are larger than the experimental reference rate by about an order of magnitude, possibly due to convergence issues (35). Consistent with the experimental data, the simulated D1→D2 rates (k_{12}) are slower than in the reverse D2→D1 (k_{21}) direction (SI Appendix, Table S1). The aggregate simulation time of each individual WE simulation was 6 to 7.5 μ s, with three ssWE simulations in each direction for both the WT and 4F-Trp CA-CTD dimer, performed until convergence of the rate estimate (SI Appendix, Fig. S9A). In summary, the estimated CA-CTD D1/D2 interconversion rates obtained from WE simulations are in excellent agreement with the experimental reference and require less aggregate simulation time than those obtained with an MSM.

To further validate our WE results, we carried out ssWE simulations using alternate progress coordinates and binning schemes, starting from the endpoints of our five conventional MD simulations. These alternate schemes require a longer initial relaxation period, compared to our previously optimized 1D orientation angle binning scheme, whether using fixed bins and the T188–T188 distance as a progress coordinate or a multiregion adaptive binning scheme with a multidimensional progress coordinate of orientation angles and the T188–T188

distance. Eventually, both alternate ssWE simulations converged, yielding stable values for the rate constants (SI Appendix, Fig. S9B), although the T188–T188 distance-based fixed binning simulation required a bin-to-bin transition-based reweighting procedure to push the system toward the true nonequilibrium steady-state distribution (36).

Overall, with all simulation schemes, we observed acceptable agreement between the simulation-acquired D1/D2 forward and backward rate constants and those determined from the experimental data, validating the identity of our candidate D2 state.

Key Differences Between the D1 and D2 States. To identify differences in residue-level, intermonomer interactions that define the D1 and D2 dimer states, we applied both Kullback–Leibler divergence-based metrics and a random forest machine learning model, as implemented in the Key Interactions Finder (KIF) Python package (37). Simulated conformations from the 90 μ s of aggregate conventional MD simulation data were classified as “D1,” “D2,” or “neither,” based on the orientation angles mentioned previously (SI Appendix, Fig. S10A). A complete set of nonbonded intermonomer interactions consisting of hydrogen bonding, hydrophobic, and electrostatic interactions were then calculated. From these features, critical interactions that differentiate the D1 and D2 states of the CA-CTD dimer were selected using the following metrics: Jensen–Shannon distance, mutual information, and feature importance from the random forest classifier (Materials and Methods). We also tested intramonomer interactions, but these features were less differentiating compared to the intermonomer interactions. Overall, we consistently identified the same top features from all three of our scoring metrics (SI Appendix, Fig. S10C).

Differentiating intermonomer interactions between the two CA-CTD dimer states involves a network of hydrogen bonds centered around Y145 and T188, as well as a hydrophobic interaction grid between L151 and L189 (Figs. 6A and 8B). Intriguingly, the Y145 residue of CA resides in the hinge between the CA-NTD and CA-CTD and is important for viral infectivity (5), possibly by influencing the Y145 centered H-bonding network. One of the most distinguishing features between the D1 and D2 conformations is the T188–T188 hydrogen bond, which can form across the CA-CTD dimer interface to possibly stabilize the D2 conformation (SI Appendix, Fig. S10B). The probability distribution of the T188–T188 distance was plotted using differing trajectory lengths of the 68 conventional MD simulations (Fig. 6B). All CA-CTD crystal structures reside within the stable wells in the T188–T188

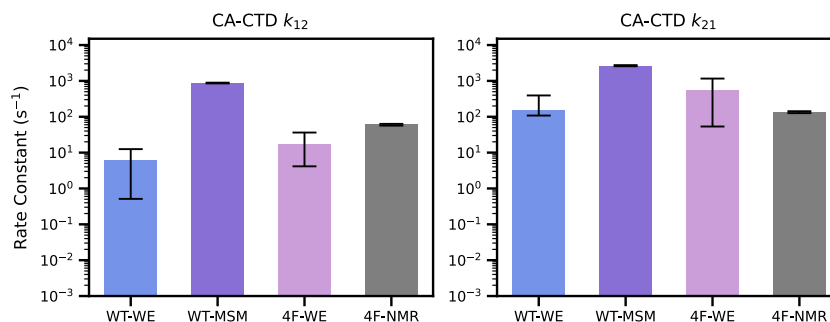


Fig. 5. Rate constants for D1→D2 (k_{12}) and D2→D1 (k_{21}) interconversion processes, calculated from simulations or experimentally measured by ¹⁹F NMR, for the WT and 4F-Trp CA-CTD. Error bars for rate estimates from WE simulations represent 95% credibility regions based on three independent trials using a Bayesian bootstrapping procedure (31). The MSM error estimate represents a 95% CI using Metropolis Markov chain Monte Carlo sampling over the posterior distribution of the transition matrix (32). Error bars for the ¹⁹F NMR rate constants represent one SD of the fitting parameters.

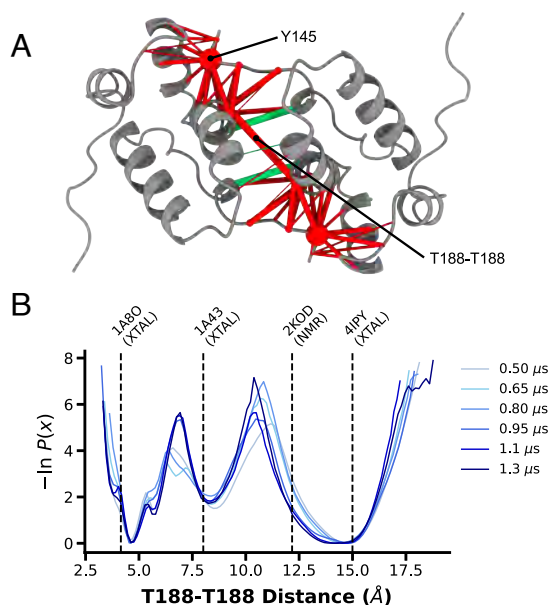


Fig. 6. Key differences between the D1 and D2 states of HIV-1 CA-CTD. (A) Cartoon ribbon representation of the CA-CTD dimer [PDB ID: 2KOD (6)] depicting differentiating nonbonded interactions as cylinders, with cylinder radii reflecting the magnitude of Jensen-Shannon distances between the D1 and D2 states. The red cylinders represent hydrogen bonds while the green cylinders represent hydrophobic interactions. Each C_{α} atom of a residue is depicted by a red sphere, with the size of the sphere representing the contributions of each residue to the set of differentiating nonbonded interactions. (B) Convergence of WE-seeded conventional MD simulations, monitored by probability distributions as a function of the T188-T188 distance between the monomeric units in the dimer. The distances for the experimental CA-CTD dimer structures are indicated by dashed vertical lines, labeled with their respective PDB codes (6–8, 10) and structural determination methods.

distance distribution, while the NMR structure lies between the 1A43 (8) and the 4IPY (7) crystal structures. This is in accord with simulations that start from the NMR structure for which an immediate minor structural adjustment occurs, shifting the conformational ensemble toward the major D1 state. The slight difference in the T188-T188 distances of the NMR structure and D1-like simulation ensemble may arise from the ensemble-averaged nature of the restraints used for the NMR structure determination.

After inspection of the other available Protein Data Bank (PDB) structures of the CA-CTD dimer with respect to this probability landscape, we noticed that the presence or absence

of the first few amino acids in the crystallized protein influences the distance between the two T188 residues across the dimer interface (*SI Appendix, Table S2*). This is consistent with the finding that these first few residues (M144-T148) are involved in the D1/D2 differentiating H-bonding network and that their absence may change the structure and T188-T188 distance of the CA-CTD dimer. In addition, by tracking the distances between E175 and W184, we observed a separation of the CA-CTD dimer conformational space into two separate ensembles (*SI Appendix, Fig. S11*). Given that protonation of E175 has previously been shown to change the populations of D1 and D2 at low pH (5), this differentiation may be related to the pH dependence of dimerization. In summary, we identified important, atomic-level interactions that differentiate the CA-CTD D1 and D2 states and that correlate with previous experimental results.

Disulfide Trapping the D2 State. Having identified the T188-T188 putative hydrogen bond as a key feature that differentiates the CA-CTD D1 and D2 conformations, we attempted to trap the CA-CTD dimer in the D2 state by disulfide bond formation. This was accomplished by substituting the threonine side chain for a cysteine. The T188C protein variant expressed well and was purified. Using size-exclusion chromatography with multiangle light scattering (SEC-MALS), the quaternary state of WT CA-CTD and the T188C variant was assessed at a concentration of 200 μ M. For the WT CA-CTD protein and the T188C variant in reducing conditions, the expected average of the monomer and dimer masses was observed, consistent with a dimerization constant of 10 μ M (6, 10, 29) (*SI Appendix, Fig. S12*). In contrast, the T188C CA-CTD variant in oxidizing conditions was almost exclusively dimeric, indicating the presence of a covalent dimer species. This result was confirmed by mass spectrometry (*SI Appendix, Fig. S13*) and further supported by the ^1H - ^{15}N HSQC spectrum of the oxidized T188C protein (Fig. 7A). Focusing on the W184 H_{ϵ} - N_{ϵ} cross-peaks, both the major D1 and minor D2 resonances are seen in WT CA-CTD and the T188C variant, with the relative amount of the D1 and D2 forms shifting from mostly D1 to mostly D2 as the protein is oxidized, indicating that disulfide bond formation across the dimer interface traps the CA-CTD dimer in a D2-like conformation. Comparing the ^1H - ^{15}N HSQC spectra of WT and the T188C variant, the T186 backbone amide cross-peak is seen to shift from the D1 form to the D2 form (*SI Appendix, Fig. S14*), similarly to previous findings (5). Overall, guided by our simulations, we were able to experimentally trap and structurally verify a disulfide-bonded CA-CTD variant.

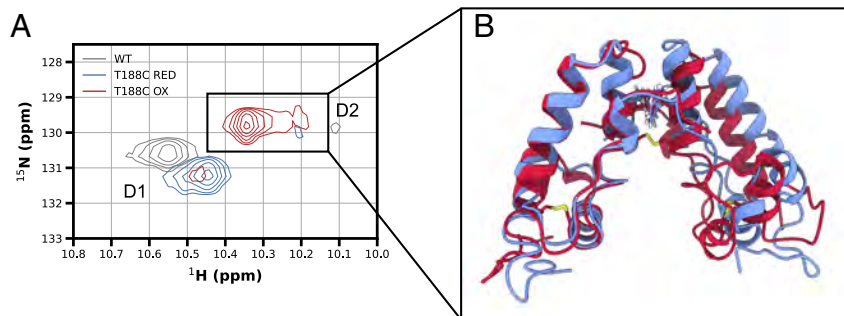


Fig. 7. Disulfide trapping the D2 conformation of CA-CTD. (A) Superposition of the Trp $N_{\epsilon}H$ region of the ^1H - ^{15}N HSQC spectra of WT CA-CTD (gray) and the T188C variant in reduced (blue) and oxidized (red) protein conditions. The D1 and D2 regions are labeled. (B) Superposition of the reduced (blue) and oxidized (red) T188C D2 variant structures in cartoon ribbon representation. The reduced T188C variant is taken from our conventional simulation data seeded from WE simulations and the oxidized T188C variant protein was relaxed from the starting reduced state using a 1 μ s conventional MD simulation.

We used MD simulations to further validate the nature of the disulfide-trapped T188C variant. For the reduced T188C variant, multiple W184 H_e-N_e resonances are present in the ¹H-¹⁵N HSQC spectrum (*SI Appendix, Fig. S15*), and these merge upon oxidation into a single D2 resonance, suggesting that several minor alternate states exist for the reduced T188C variant. Given that the T188–T188 hydrogen bond across the dimer interface is important for stabilizing the D2 conformation, disrupting this hydrogen bond may introduce conformational heterogeneity. Conventional MD simulations for the disulfide-bonded CA-CTD T188C variant, starting from the WT D2 conformation, revealed an immediate tightening of the oxidized T188C CA-CTD dimer interface by 0.6 Å (*SI Appendix, Fig. S16*), possibly caused by a misalignment between the T188–T188 hydrogen bond and the newly formed C188–C188 disulfide bond. Although the simulated structure of the disulfide cross-linked T188C CA-CTD D2 state is slightly different from the original WT D2 structure or the T188C reduced structure (*Fig. 7B*), it is stable and maintained on μs timescales (*SI Appendix, Fig. S16*).

The D2 State is Elusive to More Rapid Computational Methods. We attempted to sample the alternate CA-CTD D2 dimer state using less computationally expensive approaches. However, none of these approaches allowed us to identify this state. Below, we report on two physics-based approaches and a deep-learning approach.

Among the physics-based approaches, we tested Gaussian-accelerated MD (GaMD) simulations (38, 39) and normal mode analysis (NMA) (40, 41). Using GaMD simulations, where a biasing force is added to rapidly surmount barriers, the D2 conformation was not sampled when starting from the NMR structure (6), although a 1A43 (8) crystal structure-like metastable intermediate state was observed (*SI Appendix, Fig. S17*). Using NMA, the slowest normal modes were associated with the initial transition from the NMR structure to the crystal structure-like state, but these modes ultimately failed to describe the full transition path to the D2 alternate state (*SI Appendix, Fig. S18*).

Among the deep-learning approaches, we tested i) AlphaFold2 multimer (42–44) to generate a single structure, and ii) AlphaFold2 combined with subsampling of the input multiple sequence alignment (MSA) to generate a conformational ensemble (45–49). AlphaFold2 multimer yielded a structure that resembled the simulated D1 structure—perhaps reflecting the high prevalence of D1-like models in the AlphaFold2 training set. The combination of AlphaFold2 and MSA subsampling yielded two sets of structures among which the major set of conformations was very similar to the D1-like ensemble and the alternate set of conformations was not similar to the D2 structure, with many predicted structures adopting nonnative monomer configurations (*SI Appendix, Fig. S19*).

Conclusions

Using an integrated NMR and simulation approach, we obtained a validated ensemble of atomistic structures for the alternate D2 conformation of the HIV-1 CA-CTD dimer, which has been “invisible” to traditional structural biology and simulation methods. Specifically, we used WE simulations, overcoming timescale restrictions in conventional MD simulations. On the experimental side, we used ¹⁹F NMR to characterize the alternate dimer conformations of the HIV-1 CA-CTD dimer and measured their interconversion rate, surmounting limitations in spectral resolution that are inherent in traditional ¹H, ¹⁵N, and ¹³C protein NMR. Our rate constants from WE simulations are in excellent agreement with those measured by our ¹⁹F NMR experiments. Among other key interactions, we observed a large change in relative orientation of the CA-CTD monomers involving χ₁ angle flips of W184, which permits conversion between the D1 and D2 conformations. The observation of a similar position of the W184 indole ring for both the D1 and D2 conformations of the CA-CTD dimer is consistent with previous NMR pseudocontact shift results (51) (*SI Appendix, Fig. S20*). In addition, guided by the simulation results, we succeeded in experimentally trapping the alternative D2 structure by disulfide cross-linking.

When modeling the intersubunit CA-CTD D2 conformation in the context of the mature capsid lattice, a highly twisted

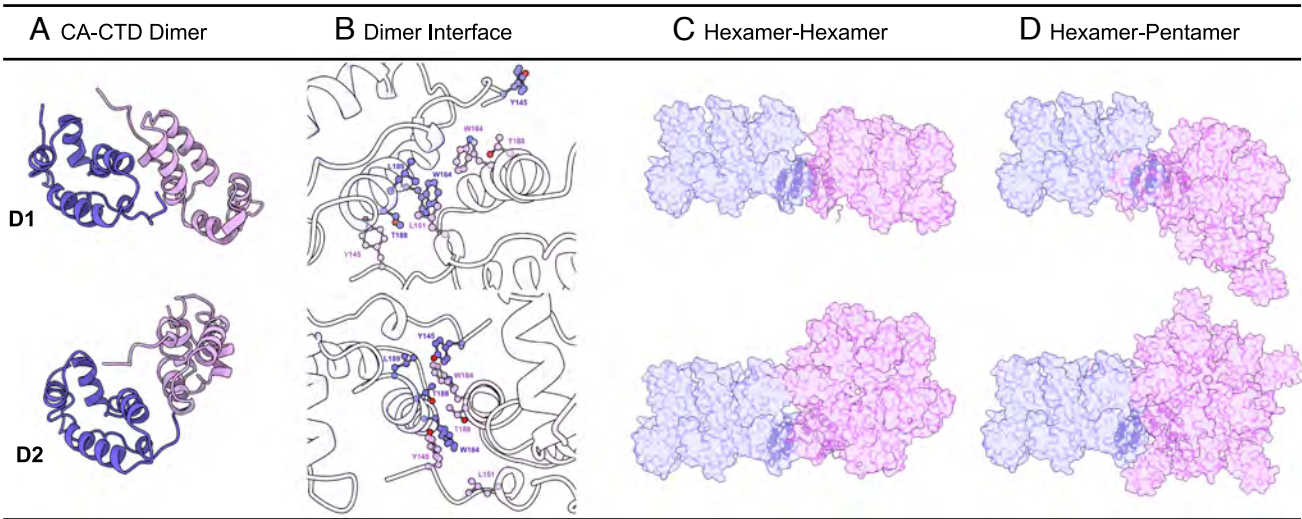


Fig. 8. Structural models of the CA-CTD D1 and D2 states at capsid subunit interfaces in the HIV-1 mature capsid lattice. (A) Representative D1 and D2 structures are shown in ribbon representation. (B) Critical residues and interactions that differentiate the dimer interfaces of D1 and D2. (C) Interfaces between hexameric subunits (11). (D) Interfaces between hexamers and pentamers (50). The assembled capsid subunits are shown in translucent, space-filling representations, with the CA-CTD structure embedded in ribbon representation.

yet viable subunit interface is observed compared to D1 (Fig. 8), which may contribute to the overall plasticity of the lattice, allowing for rearrangements during HIV-1 assembly, disassembly, maturation, or infection. Recent structural studies have demonstrated that modulation of the CA trimer interface, between pentamers and hexamers—not the CA dimer interface as previously suggested (52, 53)—is required to accommodate the highly curved pentameric geometry (50, 54). These findings align with our model of the predominant (D1) CA-CTD dimer interface, which can accommodate hexamer interfaces with other hexamers or pentamers (Fig. 8).

Inositol hexakisphosphate (IP6), a critical HIV-1 capsid assembly cofactor (55), has recently been shown to differentially influence pentameric and hexameric subunit formation (50, 54, 56). IP6 may also be linked to the TVGG-motif-based structural switch in the CA NTD (57). While the exact role of IP6 in regulating the dimer configurations studied here remains unclear, we posit that the IP6-related structural switch of the HIV-1 capsid protein is distinct from the flexible CA-CTD dimer interface. Instead, we hypothesize that the flexible CA-CTD dimer interface is intrinsically linked to the elasticity of the mature, assembled capsid (58, 59), potentially facilitating passage through the nuclear pore as an intact structure (60, 61).

Taken together, our findings open the way to exploit the alternate D2 conformational state for designing capsid assembly inhibitors in the quest for alternative strategies to combat HIV/AIDS.

Materials and Methods

Simulation System Preparation. Coordinates for the fluorinated and WT CA-CTD dimers were prepared from the heavy-atom coordinates of the CA-CTD NMR structure [(62) PDB ID: 2KOD (6)] and the crystal structure [PDB ID: 1A43 (8)]. Hydrogens were added using MolProbity (63) and the protonation states of ionizable residues were chosen to represent the major species present in solution at pH 6.5 to match experimental conditions. The polypeptide chain in the 1A43 (8) crystal structure was extended at the N-terminus to include residues M144 and Y145 for consistency with the 2KOD (6) NMR CA-CTD construct. Using ChimeraX (64), fluorine atoms were substituted individually for hydrogens in Trp184 at the C-4 position of the W184 indole ring, adjusting the C-F bond to the correct value of 1.42 Å. The proteins were solvated in a truncated octahedral box of explicit SPC/E_b (65) water molecules with a 12 Å clearance between the solute and the edge of the box. Neutralizing Na⁺ or Cl[−] ions were added using Joung and Cheatham ion parameters (66). For the oxidized CA-CTD D2 simulations, we added an intermolecular disulfide bond between the C188 of each monomer and an intramolecular disulfide bond between C198 and C218.

Molecular Dynamics Simulations. Simulations of the fluorinated, WT 2KOD, WT 1A43, and T188C CA-CTD dimers were carried out using the GPU-accelerated PMEMD module of the AMBER 22 software package (67–69), with the AMBER ff15ipq protein force field (33), ff15ipq-compatible fluorinated amino acid parameters (34), and the SPC/E_b water model (65). After initial energy minimization, a three-stage equilibration was carried out. First, a 20 ps simulation at constant volume and temperature in the presence of solute heavy-atom positional restraints was executed, using a harmonic potential with a force constant of 1 kcal/(mol · Å²). Next, a 1 ns simulation was carried out at constant temperature and pressure using the same harmonic positional restraints on solute heavy atoms. Finally, an unrestrained 1 ns simulation was run at constant temperature and pressure. Five 1 μs production simulations were carried out at constant temperature and pressure, yielding 20 μs of aggregate simulation time. For the MD simulations of WT CA-CTD starting from the end points of the WE simulation, 68 different initial states were used, with each state being simulated for 1.33 μs, totaling 90 μs of aggregate simulation time.

Simulation temperatures were maintained at 25 °C using a Langevin thermostat with a friction coefficient of 1 ps^{−1}. Simulation pressure was maintained at 1 atm using a Monte Carlo barostat with 100 fs between system volume changes. Van der Waals and short-range electrostatic interactions were truncated at 10 Å, and long-range electrostatic interactions were calculated using the particle mesh Ewald method (70). To enable a 2-fs time step, all CH and NH bonds were constrained to their equilibrium values using the SHAKE algorithm (71). Coordinates for conventional MD simulation analysis and WE simulation analysis were saved every 1 and 10 ps, respectively.

Simulation Analysis. Simulation analyses were performed using the CPPTRAJ (72), MDTraj (73), and MDAAnalysis (74) software packages. The C₂ angle for each CA-CTD monomer was defined with vectors initiated at the center-of-mass of the M144-C218 backbone heavy atoms and the center-of-mass of the K182 backbone heavy atoms. The orientation angle was defined with a vertex at the center-of-mass of the F161-V165 backbone heavy atoms. Two vectors were then defined from this vertex to the center-of-mass of the L189-Q192 backbone heavy atoms, either within the same monomer as the vertex or to the other monomer. The T188–T188 distance was defined as the center-of-mass distance between all side chain heavy atoms of T188. Distances between the CA-CTD E175 and W184 residues across the dimer interface were defined between the center-of-masses of the O_c atoms of E175 and the H_c atom of W184.

Key Interaction Finder Analysis. Critical interactions that permitted differentiation between the D1 and D2 dimer conformers were identified and scored using the Key Interactions Finder Python package (37). For the Jensen–Shannon distance-based feature scores, a continuous distribution was built using a Gaussian kernel density estimation from a histogram of each feature for the D1 or D2 state-labeled data. The random forest classifier model, implemented in scikit-learn (75), was trained to predict the D1 or D2 state, and feature importance was taken directly from the generated model. To avoid overfitting, only 80% of the dataset was used for training and testing in three rounds of stratified 10-fold cross validation. The remaining 20% of the data was held out for evaluating the quality of the final model.

Weighted Ensemble Simulations. All WE simulations were run using the open-source WESTPA 2.0 software package (16) and the GPU-accelerated PMEMD module of the AMBER 22 software package (67–69). All WE simulations were run using the same settings as the production conventional MD simulations and WE simulation data analysis and plotting were carried out using the WEDAP (76) software package. For WE simulations in nonequilibrium, steady-state conditions, trajectories that reached a predefined target state were “recycled” to the initial state, while maintaining the same weight. WE simulations run using this source-sink recycling procedure are referred to as steady-state WE simulations, while WE simulations without the recycling procedure are referred to as equilibrium WE simulations.

All WE simulations were performed using a resampling interval τ of 100 ps. The aggregate simulation time for all reported WE simulations totaled 120 μs. Initially, five replicate equilibrium WE simulations were run for 500 WE iterations (5 μs of aggregate simulation time on four A100 GPUs for 8 d each) to explore the CA-CTD conformational space, starting from the endpoints of the five 1 μs replicates of conventional MD. The larger value of the two CA-CTD orientation angles was used as a one-dimensional progress coordinate. Using the minimal adaptive binning scheme (77), 25 bins were placed between orientation angles of 10° and 70°, with a target count of four trajectories per bin. Three replicate steady-state WE simulations were run from the end points of the equilibrium WE simulations, using the same settings as in the equilibrium WE simulations, except with 30 adaptively placed bins between 0° and 40° orientation angles and no replication of leading and lagging trajectories (specified in WESTPA by using a directionality of 86). In addition, three replicate steady-state WE simulations were run in the opposite direction (D2→D1), initiated from conformations of D2 sampled by the D1→D2 steady-state WE simulations. Each steady-state WE simulation was run for 500 to 600 WE iterations, using eight A100 GPUs for 2.5 to 3.5 d, accumulating 6 to 7.5 μs of aggregate simulation time. The above set of equilibrium and steady-state WE simulations were repeated using the W184 fluorinated CA-CTD variant.

Additional WE simulations were run for WT CA-CTD using the endpoints of five 1 μ s conventional MD simulations and a target count of six trajectories per bin. A WE simulation was carried out using 35 fixed bins along the T188–T188 distance from 4.4 Å to 13.5 Å. After 1,053 WE iterations, the Weighted Ensemble Steady State (WESS) plugin was used to reweight the trajectory ensemble to be more representative of the true nonequilibrium steady state distribution, based on the bin-to-bin transition probabilities (36) along the T188–T188 distance progress coordinate. After reweighting, the WE simulation was extended to 1,295 total iterations for an aggregate simulation time of 16 μ s, using nine A100 GPUs for 7.7 d. A WE simulation was also carried out using a three-region minimal adaptive binning scheme along the T188–T188 distance and the smaller orientation angle value with six bins per region. In the first region, the T188–T188 distance was adaptively binned till the T188–T188 distance was less than 11 Å, after which the focus switched to the orientation angle until the angle value was less than 12°. Finally, the focus switched back to the T188–T188 distance until the target state was reached. This multiregion adaptive binning WE simulation consisted of 9.5 μ s of aggregate simulation time, requiring 4.4 d on nine A100 GPUs.

From the steady-state WE simulations, the rate by which the CA-CTD D1 and D2 states interconvert was calculated based on steady-state trajectories after they approached the respective target state. The D2 state was defined as having both orientation angles < 10°, the C₂ angle > 40°, the T188–T188 distance < 5 Å, and both W184 χ_1 angles between –95° and –40°. The D1 state was defined as having both orientation angles > 40°, the C₂ angle > 40°, and the T188–T188 distance > 15 Å. Since the probability flux between the D1 and D2 conformations at steady state is exactly equal to the inverse of the mean first passage time (78, 79), rate constants over time could be calculated for each WE simulation by tracking the flow of trajectory weight into the target state. The credibility regions of each set of replicate WE simulations were obtained using Bayesian bootstrapping (31).

Construction of Markov State Models. A MSM based on 90 μ s of conventional MD simulation data of CA-CTD was constructed using the Deeptime Python library (80). The features necessary to build the model were the orientation angles, C₂ angle, T188–T188 distance, and W184 χ_1 angles. The χ_1 angles were decomposed into their respective sine and cosine components to alleviate periodicity artifacts. All features first underwent min-max scaling using scikit-learn (75). Other feature sets were also tested, such as the intermonomer distance matrix, and were reduced in dimensionality using time-lagged Independent Component Analysis (tICA) (81) into a smaller set of features that captured 80% of the kinetic variance of the system. The distance matrix-based feature set was unable to successfully differentiate the D1 and D2 CA-CTD dimer conformations. For the optimal set of features, *k*-means clustering was carried out to discretize the simulation space into 250 clusters, and a count matrix was calculated between each cluster to construct the MSM. A maximum-likelihood MSM transition matrix was estimated using a lag time of 20 ns (200 frames), the shortest lag time prior to the plateauing of the implied timescales (SI Appendix, Fig. S21). A coarse-grained transition matrix was built by using PCCA+ (24) to group the 250 clusters into three metastable states (SI Appendix, Fig. S22). The mean first passage time and rate constants between the D1 and D2 states were calculated based on the same state definitions as the WE simulations. Error estimates for the MSM observables are from a Metropolis Markov chain Monte Carlo sampling procedure over the posterior distribution of the transition matrix (32).

Gaussian Accelerated Molecular Dynamics. GaMD (38, 39) were carried out using the igamd flag of the GPU-accelerated PMEMD module in AMBER 18 (67–69, 82). Five independent trajectories were simulated starting from the CA-CTD NMR structure [PDB ID: 2KOD (6)] for 200 ns each, totaling 1 μ s of aggregate simulation time. Each system first underwent the energy minimization and three-stage equilibration described in *Materials and Methods*. The dual boost method was employed, where a biasing force is added to both the total and dihedral potential energies. The threshold energy for applying all boost potentials was set to the upper energy bound. GaMD simulations of the equilibrated CA-CTD systems were carried out in the following five stages: i) a 400 ps of preparatory conventional MD simulation was carried out, without any statistics collected, ii) 2 ns of conventional MD was carried out while collecting potential statistics V_{\max} , V_{\min} , V_{avg} , and σV , iii) 1.6 ns of GaMD was carried out with a fixed

parameter boost potential applied, iv) 50 ns of GaMD was carried out with updating boost parameters, and v), 150 ns of GaMD was carried out with the updated boost parameters, which remain fixed. Reweighting (39) was carried out using a Maclaurin series expansion to the 10th order with 100 bins and maximum energy of 15 kcal/mol based on either the CA-CTD orientation angles, or the T188–T188 distance and the C₂ angle.

Normal Mode Analysis. An anisotropic network model (40), which is a broadly used elastic network model for coarse-grained normal mode analysis (41), was generated using ProDy (83) and visualized using the NMWiz interface of VMD (84). The input coordinates were the C α atoms of residues M144–Q219 from the CA-CTD NMR structure [PDB ID: 2KOD (6)], filtering out the flexible C-terminal G220–L231 residues. Ten normal modes were calculated using a force constant of 1 kcal/(mol · Å²) and a cutoff distance of 15 Å for pairwise interactions.

AlphaFold2. Generating a conformational distribution of structures with subsampled AlphaFold2 (AF2) was based on the approach by Monteiro da Silva et al. (46). The MSA was compiled using MMseqs2 (85), followed by randomly selecting a subset of the MSA, defined by the *max_seq* parameter. The remaining sequences were clustered based on Hamming distance, and the *max_seq* MSA subset was used as the cluster centers. For inference, AF2 uses the cluster centers and a number of random non-cluster-center sequences, defined by the *extra_seq* parameter. Reducing the *max_seq* and *extra_seq* values from their defaults increases the diversity of conformations sampled with AF2 (45, 47–49). Here, AF2 was run using the following pairs of *max_seq*:*extra_seq* values to optimize MSA subsampling: 8:16, 32:64, and 256:512. Ten predictions were generated with independent seeds for 5 models per seed and 3 rounds of recycling per iteration. Dropouts were enabled during inference to sample from the uncertainty of the models and improve overall sampling, setting the dropout rate to 10% for the Evoformer module, and 25% for the structure module (86).

Protein Expression and Purification. DNA encoding the WT CA-CTD (M144–L231) of gag and the T188C variant were subcloned into pET21 (EMD chemicals, Inc. San Diego, CA, USA) using NdeI and XhoI sites and verified by sequencing. CA-CTD proteins were produced in *E. coli* BL21 star (DE3) cells, grown in modified M9 minimal medium containing 1 g/L of ¹⁵NH₄Cl as the nitrogen source. At OD⁶⁰⁰ = 1.0, expression was induced with 0.5 mM IPTG. For uniform 4F-Trp labeling, 25 mg/L of 4F-indole (SigmaAldrich, St. Louis, MO, USA) was added at OD⁶⁰⁰ = 0.6, followed by protein induction at OD⁶⁰⁰ = 1.0 with 0.5 mM IPTG for 16 h at 18 °C. Cells were harvested by centrifugation at 3,000 × g for 20 min at 4 °C and lysed by microfluidization or sonication for 10 min (5 s on and 10 s off with 50% power level) in 50 mL of lysis buffer (25 mM sodium phosphate (pH 5.8) and 0.02% NaN₃). The cell lysate was clarified by centrifugation at 18,000 × g for 50 min at 4 °C. The pH of the supernatant was adjusted to 5.8 with acetic acid, and the conductivity was reduced to below 2.5 mS/cm by adding deionized water. Proteins were fractionated over a 5 ml Hi-Trap SP column (GE Healthcare) using a 0 to 1M NaCl gradient in 25 mM sodium phosphate buffer (pH 5.8), 1 mM DTT, and 0.02% NaN₃. The final purification step involved size-exclusion chromatography over a Superdex 75 26/60 column (GE Healthcare) in 25 mM sodium phosphate buffer (pH 6.5), 1 mM DTT, and 0.02% NaN₃. Molecular masses of purified CA-CTD proteins were confirmed by ESI mass spectrometry (Bruker Daltonics, Billerica, MA, USA).

Oxidation of the T188C CA-CTD protein was achieved using two different approaches, both of which yielded uniformly oxidized protein. Prior to oxidation, the protein was dialyzed into 25 mM sodium phosphate buffer (pH 6.5) with 0.02% NaN₃, and air oxidation was allowed to proceed by leaving the protein sample at 4 °C for over 4 wk. For the other sample, the protein was dialyzed into 20 mM HEPES buffer (pH 6.5) with 100 μ M Cu(II)SO₄ and, after oxidation, all Cu(II) was chelated by adding 5 mM EDTA and removed by dialysis against 25 mM sodium phosphate buffer (pH 6.5) with 0.02% NaN₃.

NMR Spectroscopy. NMR samples contained 120 to 2,000 μ M CA-CTD in 25 mM sodium phosphate buffer (pH 6.5), 0.02% NaN₃, 1 mM DTT (except for the oxidized samples), and 10% D₂O. NMR spectra were recorded at 25 °C on 14.1 and 18.8 T Bruker AVANCE spectrometers equipped with 5 mm triple-

resonance, z-axis gradient cryoprobes. ^1H - ^{15}N HSQC spectra were recorded with an interscan delay of 0.9 s, 8 scans, and 128 complex points in the ^{15}N dimension. All ^{19}F spectra were recorded on a 14.1 T Bruker AVANCE spectrometer, equipped with a CP TXO F/C-H-D triple-resonance, z-axis gradient cryoprobe. ^{19}F chemical shifts were referenced to trifluoroacetic acid. One-dimensional ^{19}F spectra were collected with 1,024 scans, a spectral width of 50 ppm, and a recycle delay of 1.0 s. The carrier frequency was set to -123 ppm for the 4F-Trp CA-CTD dimer. ^{19}F - ^{19}F EXSY experiments were performed using the following mixing times: 2, 5, 10, 15, 25, 35, 50, 75, 100, 200, and 600 ms. For each EXSY experiment, 512 scans were collected in the direct ^{19}F dimension with a spectral width of 9 ppm and 32 complex points were collected in the indirect ^{19}F dimension with a spectral width of 4 ppm. The carrier frequency for both ^{19}F dimensions was set to -125.54 ppm and the recycle delay was set to 1.0 s. D1/D2 interconversion rates were fit to Eq. 1, mapping cross peak intensity (I_{12}/I_{11}) changes for increasing mixing times. The error for each intensity ratio was calculated by error propagation of the peak intensity error, which was derived from the signal-to-noise ratio of the ^{19}F - ^{19}F EXSY spectrum. The error for the rate constants was obtained from the SD of the fitting parameters. Spectra were processed with NMRPipe (87) and analyzed using nmrglue (88) and SciPy (89).

Size-Exclusion Chromatography with Multiangle Light Scattering. Quaternary states of the proteins were assessed at room temperature using an analytical Superdex S75 column with in-line multiangle light-scattering (DAWN HELEOS, Waters-Wyatt Technology Corporation, Santa Barbara, CA), refractive index (OPTILAB DSP, Wyatt Technology, Inc., Santa Barbara, CA, USA), and ultraviolet (Agilent Technologies, Santa Clara, CA) detectors. 100 μL of WT CA-CTD and T188C CA-CTD proteins at a 200 μM concentration were injected into the pre-equilibrated column and eluted with 25 mM sodium phosphate buffer (pH 6.5) and 0.02% $\text{Na}_2\text{S}_2\text{O}_3$ at a flow rate of 0.5 mL/min.

Data, Materials, and Software Availability. CA-CTD D1/D2 key PDB coordinate files, WE simulation files, conventional MD simulation files, MSM building code, code for KIF-based analysis, the GaMD simulation files, and our experimental data and processing scripts have been deposited in Zenodo (<https://doi.org/10.5281/zenodo.14754836>) (90). Some study data are available: All simulation trajectories are available upon request due to large file size limitations.

ACKNOWLEDGMENTS. This work was supported by the NIH Pittsburgh AIDS Research Training program grant T32AI065380 and a University of Pittsburgh Andrew Mellon Predoctoral Fellowship awarded to D.T.Y.; NIH grants P50AI150481 and U54AI170791, as well as NSF grant CHE-1708773 awarded to A.M.G.; and NIH grant R01GM115805 and NSF grant MCB-2112871 awarded to L.T.C. Computational resources were provided through NSF XSEDE grant TG-IO1010161 to D.T.Y., the Texas Advanced Computing Center project MCB23004 to L.T.C., and by the University of Pittsburgh Center for Research Computing, RRID:SCR_022735, through the H2P cluster, which is supported by NSF award number OAC-2117681. We thank Doug Bevan for computer technical support, Mike Delk for NMR instrumental support, Teresa Brosenitsch for critically reading the manuscript, and all members of the Chong and Gronenborn labs, Rieko Ishima, David Case, Sunil Saxena, and the weighted ensemble software development community for helpful discussions. Especially valuable were insightful interactions with Manman Lu, Fatema Bhinderwala, Wenkai Zhu, Christie Monnie, Alex Guseman, Anthony Bogetti, Jeremy Leung, and Marion Silvestrini.

Author affiliations: ^aMolecular Biophysics and Structural Biology Graduate Program, University of Pittsburgh and Carnegie Mellon University, Pittsburgh, PA 15260; ^bDepartment of Structural Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15260; and ^cDepartment of Chemistry, University of Pittsburgh, Pittsburgh, PA 15260

1. E. M. Campbell, T. J. Hope, HIV-1 capsid: The multifaceted key player in HIV-1 infection. *Nat. Rev. Microbiol.* **13**, 471–483 (2015).
2. S. Mattei, B. Glass, W. J. H. Hagen, H. G. Kräusslich, J. A. G. Briggs, The structure and flexibility of conical HIV-1 capsids determined within intact virions. *Science* **354**, 1434–1437 (2016).
3. A. Tan, A. J. Pak, D. R. Morado, G. A. Voth, J. A. G. Briggs, Immature HIV-1 assemblies from Gag dimers leaving partial hexamers at lattice edges as potential substrates for proteolytic maturation. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2020054118 (2021).
4. S. Mattei *et al.*, High-resolution structures of HIV-1 Gag cleavage mutants determine structural switch for virus maturation. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E9401–E9410 (2018).
5. I. J. L. Byeon *et al.*, Motions on the millisecond time scale and multiple conformations of HIV-1 capsid protein: Implications for structural polymorphism of CA assemblies. *J. Am. Chem. Soc.* **134**, 6455–6466 (2012).
6. I. J. L. Byeon *et al.*, Structural convergence between Cryo-EM and NMR reveals intersubunit interactions critical for HIV-1 capsid function. *Cell* **139**, 780–790 (2009).
7. A. Lampel *et al.*, A triclinal crystal structure of the carboxy-terminal domain of HIV-1 capsid protein with four molecules in the asymmetric unit reveals a novel packing interface. *Acta Crystallogr. Sect. A* **69**, 602–606 (2013).
8. D. K. Worthy, H. Wang, S. Yoo, W. I. Sundquist, C. P. Hill, Structures of the HIV-1 capsid protein dimerization domain at 2.6 Å resolution. *Acta Crystallogr. Sect. D* **55**, 85–92 (1999).
9. S. Du *et al.*, Structure of the HIV-1 full-length capsid protein in a conformationally trapped unassembled state induced by small-molecule binding. *Am. J. Mol. Biol.* **406**, 371–386 (2011).
10. T. R. Gamble *et al.*, Structure of the carboxyl-terminal dimerization domain of the HIV-1 capsid protein. *Science* **278**, 849–853 (1997).
11. A. T. Gres *et al.*, X-ray crystal structures of native HIV-1 capsid protein reveal conformational variability. *Science* **349**, 99–103 (2015).
12. G. A. Huber, S. Kim, Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys. J.* **70**, 97–110 (1996).
13. D. M. Zuckerman, L. T. Chong, Weighted ensemble simulation: Review of methodology, applications, and software. *Annu. Rev. Biophys.* **46**, 43–57 (2017).
14. B. E. Husic, V. S. Pande, Markov state models: From an art to a science. *J. Am. Chem. Soc.* **140**, 2386–2396 (2018).
15. T. Ni *et al.*, Structure of native HIV-1 cores and their interactions with IP6 and CypA. *Sci. Adv.* **7**, eabj5715 (2021).
16. J. D. Russo *et al.*, WESTPA 2.0: High-performance upgrades for weighted ensemble simulations and analysis of longer-timescale applications. *J. Chem. Theory Comput.* **18**, acs.jctc.1c01154 (2022).
17. J. Hénin, T. Lelièvre, M. R. Shirts, O. Valsson, L. Delemotte, Enhanced sampling methods for molecular dynamics simulations [article v1.0]. *Living J. Comput. Mol. Sci.* **4**, 1583 (2022).
18. L. T. Chong, A. S. Saglam, D. M. Zuckerman, Path-sampling strategies for simulating rare events in biomolecular systems. *Curr. Opin. Struct. Biol.* **43**, 88–94 (2017).
19. M. C. Zwier, L. T. Chong, Reaching biological timescales with all-atom molecular dynamics simulations. *Curr. Opin. Pharmacol.* **10**, 745–752 (2010).
20. B. W. Zhang, D. Jasnow, D. M. Zuckerman, The “weighted ensemble” path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *J. Chem. Phys.* **132**, 054107 (2010).
21. M. C. Zwier, J. W. Kaus, L. T. Chong, Efficient explicit-solvent molecular dynamics simulations of molecular association kinetics: Methane/methane, na^+/cl^- , methane/benzene, and $\text{k}^+/\text{18-crown-6}$ ether. *J. Chem. Theory Comput.* **7**, 1189–1197 (2011).
22. V. S. Pande, K. Beauchamp, G. R. Bowman, Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52**, 99–105 (2010).
23. J. H. Prinz *et al.*, Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **134**, 174105 (2011).
24. S. Röhlitz, M. Weber, Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification. *Adv. Data Anal. Classif.* **7**, 147–179 (2013).
25. J. T. Gerig, Fluorine NMR of proteins. *Prog. Nucl. Magn. Reson. Spectrosc.* **26**, 293–370 (1994).
26. A. M. Gronenborn, Small, but powerful and attractive: ^{19}F in biomolecular NMR. *Structure* **30**, 6–14 (2022).
27. L. A. Luck, J. J. Falke, Fluorine-19 NMR studies of the D-galactose chemosensory receptor. 1. Sugar binding yields a global structural change. *Biochemistry* **30**, 4248–4256 (1991).
28. B. D. Sykes, W. E. Hull, [13] Fluorine nuclear magnetic resonance studies of proteins. *Methods Enzymol.* **49**, 270–295 (1978).
29. J. L. Newman, E. W. Butcher, D. T. Patel, Y. Mikhaylenko, M. F. Summers, Flexibility in the P2 domain of the HIV-1 Gag polyprotein. *Protein Sci.* **13**, 2101–2107 (2004).
30. W. Zhu, Y. Li, M. Liu, J. Zhu, Y. Yang, Uncorrelated effect of interdomain contact on Pin1 isomerase activity reveals positive catalytic cooperativity. *J. Phys. Chem. Lett.* **10**, 1272–1278 (2019).
31. B. Mostofian, D. M. Zuckerman, Statistical uncertainty analysis for small-sample, high log-variance data: Cautions for bootstrapping and Bayesian bootstrapping. *J. Chem. Theory Comput.* **15**, 3499–3509 (2019).
32. B. Trendelkamp-Schroer, H. Wu, F. Paul, F. Noé, Estimation and uncertainty of reversible Markov models. *J. Chem. Phys.* **143**, 174101 (2015).
33. K. T. Debiec *et al.*, Further along the road less traveled: AMBER ff15ipq, an original protein force field built on a self-consistent physical model. *J. Chem. Theory Comput.* **12**, 3926–3947 (2016).
34. D. T. Yang, A. M. Gronenborn, L. T. Chong, Development and validation of fluorinated, aromatic amino acid parameters for use with the AMBER ff15ipq protein force field. *J. Phys. Chem.* **126**, 2286–2297 (2022).
35. E. Suárez *et al.*, What Markov state models can and cannot do: correlation versus path-based observables in protein-folding models. *J. Chem. Theory Comput.* **17**, 3119–3133 (2021).
36. D. Bhatt, B. W. Zhang, D. M. Zuckerman, Steady-state simulations using weighted ensemble path sampling. *J. Chem. Phys.* **133**, 014110 (2010).
37. R. M. Crean, J. S. G. Slusky, P. M. Kasson, S. C. L. Kamerlin, KIF-Key Interactions Finder: A program to identify the key molecular interactions that regulate protein conformational changes. *J. Chem. Phys.* **158**, 144114 (2023).
38. Y. Miao, V. A. Feher, J. A. McCammon, Gaussian accelerated molecular dynamics: Unconstrained enhanced sampling and free energy calculation. *J. Chem. Theory Comput.* **11**, 3584–3595 (2015).
39. Y. Miao *et al.*, Improved reweighting of accelerated molecular dynamics simulations for free energy calculation. *J. Chem. Theory Comput.* **10**, 2677–2689 (2014).
40. A. R. Atilgan *et al.*, Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80**, 505–515 (2001).
41. I. Bahar, T. R. Lezon, A. Bakan, I. H. Shrivastava, Normal mode analysis of biomolecular structures: Functional mechanisms of membrane proteins. *Chem. Rev.* **110**, 1463–1497 (2010).

42. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
43. M. Mirdita *et al.*, ColabFold: Making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
44. R. Evans *et al.*, Protein complex prediction with AlphaFold-Multimer. bioRxiv [Preprint] (2022). <https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2> (Accessed 1 August 2024).
45. D. del Alamo, D. Sala, H. S. Mchaourab, J. Meiler, Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **11**, e75751 (2022).
46. G. Monteiro *et al.*, High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nat. Commun.* **15**, 2464 (2024).
47. H. K. Wayment-Steele *et al.*, Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).
48. A. Meller, S. Bhakat, S. Solieva, G. R. Bowman, Accelerating cryptic pocket discovery using AlphaFold. *J. Chem. Theory Comput.* **19**, 4355–4363 (2023).
49. B. P. Vani, A. Aranganathan, D. Wang, P. Tiwary, AlphaFold2-RAVE: From sequence to Boltzmann ranking. *J. Chem. Theory Comput.* **19**, 4351–4354 (2023).
50. J. C. V. Stacey *et al.*, Two structural switches in HIV-1 capsid regulate capsid curvature and host factor binding. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2220557120 (2023).
51. W. Zhu, D. T. Yang, A. M. Gronenborn, Ligand-capped cobalt(II) multiplies the value of the double-histidine motif for PCS NMR studies. *J. Am. Chem. Soc.* **145**, 4564–4569 (2023).
52. O. Pornillos, B. K. Ganser-Pornillos, M. Yeager, Atomic-level modelling of the HIV capsid. *Nature* **469**, 424–427 (2011).
53. G. Zhao *et al.*, Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **497**, 643–646 (2013).
54. C. M. Highland, A. Tan, C. L. Ricaña, J. A. G. Briggs, R. A. Dick, Structural insights into HIV-1 polyanion-dependent capsid lattice formation revealed by single particle cryo-EM. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2220545120 (2023).
55. R. A. Dick *et al.*, Inositol phosphates are assembly co-factors for HIV-1. *Nature* **560**, 509–512 (2018).
56. M. Gupta, A. J. Pak, G. A. Voth, Critical mechanistic features of HIV-1 viral capsid assembly. *Sci. Adv.* **9**, eadd7434 (2023).
57. R. T. Schirra *et al.*, A molecular switch modulates assembly and host factor binding of the HIV-1 capsid. *Nat. Struct. Mol. Biol.* **30**, 383–390 (2023).
58. A. Hudait, G. A. Voth, HIV-1 capsid shape, orientation, and entropic elasticity regulate translocation into the nuclear pore complex. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2313737121 (2024).
59. A. Deshpande *et al.*, Elasticity of the HIV-1 core facilitates nuclear entry and infection. *PLoS Pathog.* **20**, 1–24 (2024).
60. C. Li, R. C. Burdick, K. Nagashima, W. S. Hu, V. K. Pathak, HIV-1 cores retain their integrity until minutes before uncoating in the nucleus. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2019467118 (2021).
61. V. Zila, T. G. Müller, B. Müller, H. G. Kräusslich, HIV-1 capsid is the key orchestrator of early viral replication. *PLoS Pathog.* **17**, 1–9 (2021).
62. H. M. Berman *et al.*, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
63. C. J. Williams *et al.*, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
64. E. C. Meng *et al.*, UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).
65. K. Takemura, A. Kitao, Water model tuning for improved reproduction of rotational diffusion and NMR spectral density. *J. Phys. Chem. B* **116**, 6279–6287 (2012).
66. I. S. Joung, T. E. Cheatham, Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **112**, 9020–9041 (2008).
67. D. A. Case *et al.*, The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
68. A. W. Götz *et al.*, Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born. *J. Chem. Theory Comput.* **8**, 1542–1555 (2012).
69. R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, R. C. Walker, Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878–3888 (2013).
70. U. Essmann *et al.*, A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 234505 (1995).
71. J. P. Ryckaert, G. Ciccotti, H. J. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
72. D. R. Roe, T. E. Cheatham, PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
73. R. T. McGibbon *et al.*, MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
74. N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, O. Beckstein, MDAAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32**, 2319–2327 (2011).
75. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
76. D. T. Yang, L. T. Chong, WEDAP: A Python package for streamlined plotting of molecular simulation data. *J. Chem. Inf. Model.* **64**, 5749–5755 (2024).
77. P. A. Torrillo, A. T. Bogetti, L. T. Chong, A minimal, adaptive binning scheme for weighted ensemble simulations. *J. Phys. Chem. A* **125**, 1642–1649 (2021).
78. T. Hill, *Free Energy Transduction and Biochemical Cycle Kinetics* (Dover Publications, 2004).
79. D. Bhatt, D. M. Zuckerman, Beyond microscopic reversibility: Are observable nonequilibrium processes precisely reversible? *J. Chem. Theory Comput.* **7**, 2520–2527 (2011).
80. M. Hoffmann *et al.*, Deeptime: A python library for machine learning dynamical models from time series data. *Mach. Learn. Sci. Technol.* **3**, 015009 (2021).
81. L. Molgedey, H. G. Schuster, Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.* **72**, 3634–3637 (1994).
82. D. A. Case *et al.*, AMBER 2018 (University of California, San Francisco, 2018). <https://ambermd.org>. Accessed 7 January 2020.
83. A. Bakan, L. M. Meireles, I. Bahar, ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).
84. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
85. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
86. B. Wallner, AFsample: Improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics* **39**, btad573 (2023).
87. F. Delaglio *et al.*, NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
88. J. J. Helmus, C. P. Jaroniec, NmrGlue: An open source Python package for the analysis of multidimensional NMR data. *J. Biomol. NMR* **55**, 355–367 (2013).
89. P. Virtanen *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
90. D. Yang, darianyang/d1d2: v1.0.0: Zenodo release. Zenodo. <https://doi.org/10.5281/zenodo.14754836>. Deposited 28 January 2025.