Conformational landscape reduction of a dynamic 29-residue peptide by a perfluoroaromatic small molecule

Ethan D. Evans[1,]*

1) Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 20139, United States.

Abstract

Conformationally dynamic peptides and proteins display both important biochemical properties and present a challenge for computational modeling. Characterizing the accessible structural landscape represents one route to understand their function with molecular level detail. We characterize a self-labeling 29-residue peptide, MP01-Gen4, that undergoes structural alterations in the presence of a perfluoroaromatic reaction partner. Replica exchange molecular dynamics (REMD) shows MP01 to access a broad set of states, that microsecond-long explicit solvent simulations only minimally sample. REMD and structural network analysis find an altered and reduced conformational landscape when MP01 interacts non-covalently or is covalently attached to the perfluoroaromatic small molecule. Residues throughout the peptide, notably at the C-terminus, interact with the small molecule in conformational state-dependent manners. The results help explain and generate hypotheses for experimental observations including the importance of flexibility and the role of the N- and C-terminal regions, both of which are distant from the active cysteine. The simulations highlight the importance of substantial sampling in minimally stabilized, conformationally dynamic systems and supplies a case study for small molecule-mediated, peptide conformational changes.

Introduction

Peptides and miniproteins have served as a staple for computational[1–4] and biochemical[5–8] studies, and have garnered increasing interest for their functional roles in biological systems[9–11]. The wealth of available bioinformatic data has uncovered thousands of peptides and, while still quite challenging, improved functional annotation. Peptides have been shown to serve many roles, they can: act as toxins[12], assist in skeletal muscle formation[13], promote dissociation of RNA-protein P-bodies[10] and, as discussed here, perform chemical reactions[14,15]. However, in many cases, there is little structural information, leaving important mechanistic details unknown.

Molecular dynamics (MD) represents one method to help characterize peptide systems and extract biophysical insight. MD has supplied extensive thermodynamic and kinetic information on many peptides; well-studied examples include the avian pancreatic peptide[16,17], the villin headpiece[18–20], the WW domain[19,21,22], chignolin[23,24] along with the tryptophan zipper[18,25] and Trp-cage[26]. These prototypical systems have been used to test sampling methods[20] and

optimize force fields[19] and have provided detailed information on folding pathways, metastable states and transition kinetics.

Peptides lacking strong stabilizing elements like disulfide bonds may frequently undergo transitions between metastable states. This suggests that to generally understand their biochemical and biophysical properties, a sufficiently large portion of their conformational landscape should be characterized. Lacking knowledge on the set of available states or appropriate order parameters by which to bias a simulation, requires one to perform extensive sampling. Replica exchange molecular dynamics (REMD)[27] presents an enhanced sampling method, requiring minimal knowledge of specific biasing parameters. For REMD, multiple independent copies of the system are simulated over a set of temperatures; exchanges are attempted between pairs at different temperatures, inducing a random walk through different temperature-defined canonical ensembles. This process allows the system to both enter and escape energetic minima faster than conventional MD. The method has seen considerable use for modeling peptides including the Aβ peptide[28,29] and short three-stranded β-sheets[30] and additionally has been extended for distributed computing[31]. REMD thus presents as a viable method to characterize the states accessible to peptides.

Structural alteration upon function (e.g. binding) or modification (e.g. post-translational modification or other covalent labeling) is an important feature often seen in peptides and proteins. For instance, the short p53 peptide undergoes conformational changes upon binding the protein MDM2[32], as does the avian pancreatic polypeptide-derived P007 upon binding to DNA[33]. For covalent modification, phosphorylation has been shown to kinetically stabilize the initially disordered kinase-inducible domain of CREB[34]. Additionally, we have previously demonstrated structural modifications upon the interaction and reaction of a perfluoroaromatic small molecule with several partially helical ~30mer peptides[15,35].

Here, we study the structural landscape of the abiotic, 29-residue peptide MP01-Gen4 that performs a self-labeling nucleophilic aromatic substitution between its reactive cysteine and a perfluoroaromatic small molecule (CA, Figure 1)[14,15]. Experimentally, the system displays a structural ensemble change from a mixed helix-coil state to a helical state, when either non-covalently interacting or covalently reacted with the CA. Residue-specific electron paramagnetic resonance (EPR) spectroscopy showed Gen4 to be conformationally dynamic, with residues in the terminal regions (1–7, 27–29) displaying the largest decrease in dynamics upon labeling[36]. Using both microsecond long explicit solvent MD and implicit solvent REMD, we observed a broad conformational landscape for Gen4. Further, REMD studies showed Gen4 to access an altered and limited set of states when the CA was covalently attached or non-covalently interacting.
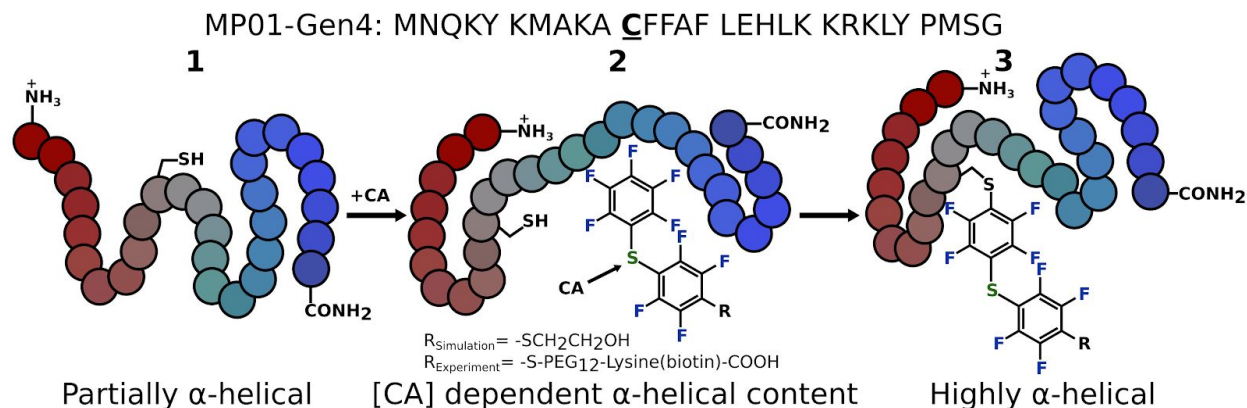
Figure 1. MP01-Gen4 undergoes conformational changes when non-covalently interacting or covalently attached to the CA. Sequence of MP01-Gen4 with active site cysteine underlined in bold (top). MP01 adopts a mixed α-helix/coil in solution (1), a predominantly α-helical structure upon addition of the CA, even when unreacted (2), and after covalent labeling it remains α-helical (3). $R_{Simulation}$ was used in all computational studies.

Computational methods

Molecular dynamics

All simulations (Table S1) were performed using NAMD 2.12[37] with the CHARMM36 force field[38,39]. The abbreviations Gen4, Gen4+CA, and Gen4-CA refer to simulations of MP01-Gen4 alone, MP01-Gen4 non-covalently in the presence of the CA, and MP01-Gen4 covalently attached to the CA respectively. The CA molecule was a truncated version, with the perfluoroaromatic as an arylthioethanol derivative (Figure 1). The CA and a covalent Cys-CA were parameterized for the CHARMM Generalized Force Field (CGenFF)[40,41], the sidechain parameters from the Cys-CA were used to modify those of a CHARMM36 Cys residue. The C-terminus of the peptide was patched as an amide to match experiment. Protein structure files were prepared using the AutoPSF tool in VMD[42]. Simulations were performed using Langevin dynamics with a time step of 2 fs and a damping coefficient of 1 ps^{-1}, with hydrogens constrained via SHAKE[43]. For explicit solvent simulations, pressure was controlled using a Nosé–Hoover Langevin piston[44,45], the particle mesh Ewald method was used to treat interactions with real space electrostatic and Lennard-Jones energies cut off at 12 Å. Explicit solvent simulations were carried out at 297 K, 150 mM NaCl, and TIP3P[46] water molecules at 1 bar (NPT) in a periodic, cubic cell with 12 Å padding and frames written every 2 ps. Production runs were performed after 2000 steps of conjugate gradient-based minimization, heating from 60 to 297 K with 600 simulation steps at each temperature and a 1 ns simulation at 297 K, all with backbone restraints (NVT, NVT, and NPT respectively). The Gen4 explicit solvent starting conformations were acquired following a 500 K NVT simulation, from which two random structures were obtained, possessing no helical or beta sheet secondary structure (calculated by STRIDE[47]).

Replica exchange molecular dynamics (REMD)[27] simulations were performed using 10 replicas in implicit solvent with temperatures exponentially spaced between 300 and 500 K. The simulation set up protocol mirrored that of the explicit solvent simulations but used the Generalized Born/solvent-accessible surface area (BG/SASA) model[48,49], a 16 Å non-bonded cutoff distance and 100 mM ion concentration. Exchanges were attempted every 1000 steps (2 ps) with frames saved every 20 ps. For the Gen4+CA REMD simulation, a flat bottom restraining potential was placed on the distance between the center of mass for the peptide and the CA with no potential applied until a distance of 25 Å, where a harmonic potential (25 kcal mol^{-1} Å^{-2}) started (no biasing potential was used in the explicit solvent Gen4+CA simulation). The number of replicas was set such that the acceptance probability was approximately 25% for all REMD simulations along with overlap of the replica potential energy profiles (Figure S1). The SASA of the first and second half of the three general simulations were compared to assess convergence, all of which showed substantial overlap (Figure S2).

Data analysis

The first 20 ns of all REMD (40 ns for explicit solvent) simulations were treated as an extended equilibrium and removed, leaving for each replica approximately 0.82 μs, 1.17 μs (the two Gen4 simulations, which were combined), 1.06 μs (Gen4-CA), and 1.05 μs (Gen4+CA) for further analysis. Similarly, this left approximately 3.94 μs (Gen4, simulation 1), 3.92 μs (Gen4, simulation 2), 3.40 μs (Gen4+CA, simulation 3), 2.44 μs (Gen4-CA, simulation 4), and 4.09 μs (Gen4-CA, simulation 5) for analysis from the explicit solvent simulations. Two-dimensional free energy surface (FES) plots were obtained using only the 300 K replicas using the plot_free_energy function in PyEMMA[50]. FES plots used the first two principal components following dihedral principal component analysis (dPCA)[51–53] of residues 2–27. Agglomerative clustering (sklearn.cluster.AgglomerativeClustering) with a distance threshold of 50 (above which a new cluster would be created) was used to group structures using the same backbone dihedral angles as used for the FES analysis; cluster centroids were extracted using KMedoids (sklearn_extra.cluster.KMedoids, k = 1) for each of the resulting clusters. Average cluster ϕ and ψ angles for residues 2–27 were calculated and used to obtain differences between clusters with standard deviation error propagation. Distances between the CA and all peptide residues (or specific subcomponents as listed) within a given distance of it were calculated for all frames and clusters (or combinations of clusters) in the Gen4+CA REMD simulation.

Cluster-to-cluster network connectivity was determined using cluster centroid-to-centroid distances, initially representing a fully connected graph. From this graph an approximate solution to the minimum Steiner tree was calculated using structure distances as weights (networkx.algorithms.approximation.steiner_tree). Additional edges between other similar states were added to the resulting graph (Euclidean distance < 4). Simulation-to-simulation cluster comparisons were performed using cluster centroids; from this, a bipartite graph was constructed

linking only clusters with highly similar (Euclidean distance < 3) conformations (networkx.algorithms.bipartite).

Results

Explicit solvent MD of MP01-Gen4 suggests a conformationally dynamic peptide but does not provide sufficient sampling

The two explicit solvent folding simulations of Gen4 found no single, long-lived conformation and did not converge in their sampling. Additionally, they did not adopt similar structures throughout the course of the simulation (Figures S3–S6). The SASA for both simulations fluctuated from near 3000 $Å^2$ to over 4700 $Å^2$, likewise, the radius of gyration (R_g) ranged from under 10 $Å$ to 22 $Å$. The largest fluctuations occurred during structural rearrangements between metastable states or the initial collapse stage. Several large structural changes occured over hundreds or thousands of nanoseconds (from ~0.5 to ~1.4 ns in simulation 1), while others took less than a hundred nanoseconds (conversion of the state depicted at 1750 ns with 2 helical segments to the single helical structures from 2000–3250 ns in simulation 2, Figure S5). In addition to simulation 2's central helical residues, the secondary structure adopted by the peptide fluctuated with time for both simulations, with turns and coils being the most prevalent elements (Figures S4 and S6). The central residues in simulation 2 took on a helical conformation for nearly 1 μs, flanked by more flexible N- and C-terminal coiled regions. Overall, the simulations showed only a handful of metastable conformations following the initial collapse, suggesting that enhanced sampling methods were required to access more of the accessible states.

MP01-Gen4 exhibits numerous conformations using REMD

Implicit solvent REMD of Gen4 uncovered a broad structural landscape with minimal energy barriers between states. Two dimensional FES analysis of the first two principal components of the 300 K frames showed several general energetic wells, some possessing multiple minima. Only small energy barriers were observed between minima, ranging from below 0.3 kcal mol^{-1} to approximately 1.5–1.8 kcal mol^{-1} (Figure 2A). Cluster centroids for the top 20 largest clusters mapped well to the minima observed in the 2D-FES (Figure 2A, white circles). In several cases, multiple centroids (and corresponding clusters) co-located within energetic minima. This suggested that individual mimina possessed structural variants, something not initially obvious in the low dimensional FES representation. The agglomerative clustering required well over 100 clusters to capture the space of structures observed (191 in total), with the top 20 clusters accounting for 25.3% of the population (Figure 2B). The largest two clusters each held ~2% of the population. The population representation of subsequent clusters dropped to below 1% before 20 clusters and less than 0.5% near the 80th largest cluster (accounting for 65.8% of the population). Overall, a substantial portion of structures were spread out over many clusters, defining a broad conformational landscape.
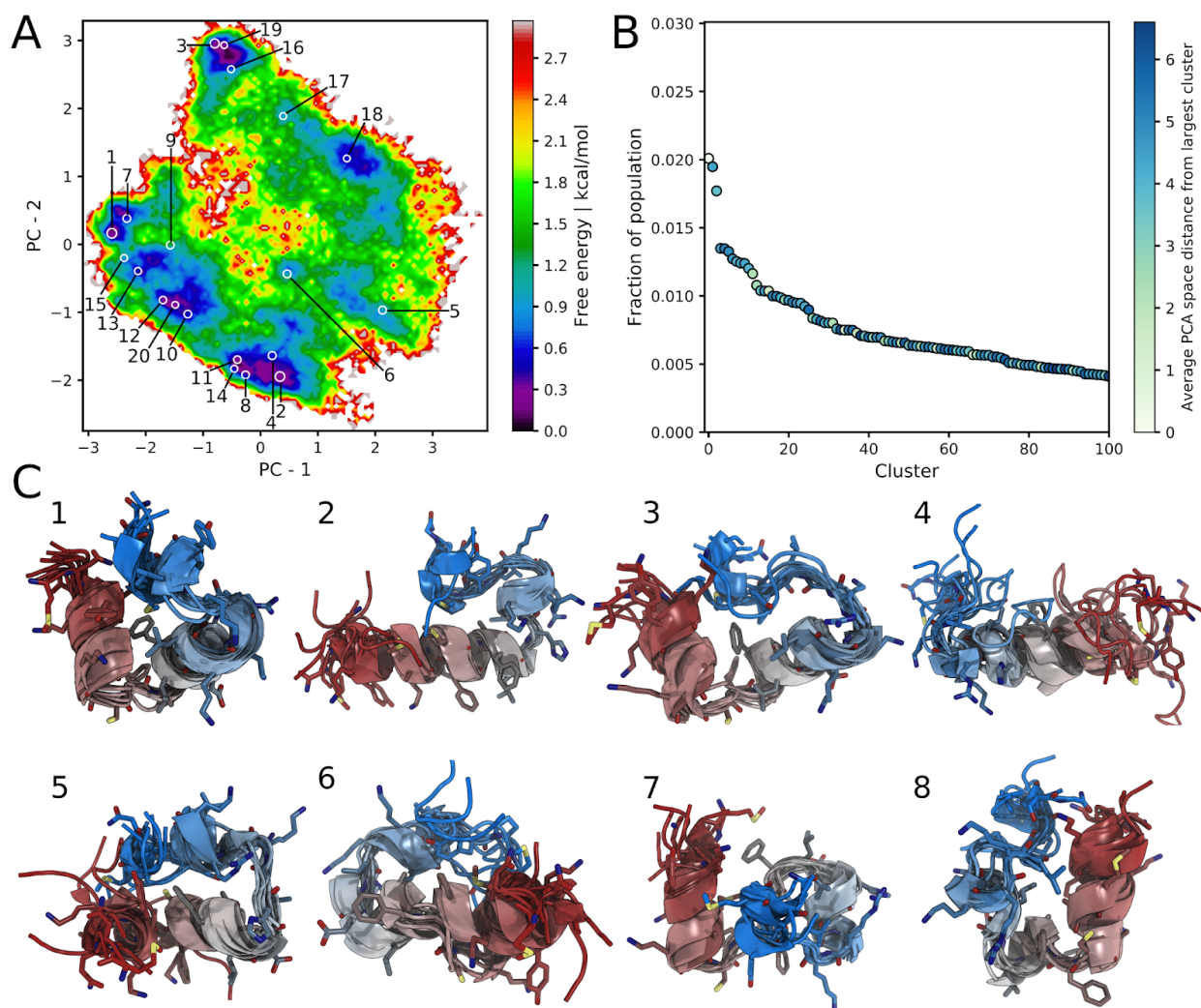
Figure 2. MP01-Gen4 displays a broad structural landscape with minimal energetic barriers. (A) 2D-FES using the first two dPCA components with the locations of the top 20 largest clusters plotted at the coordinates for the cluster centroid. The cluster size is depicted by the size of the circle. (B) Fraction of the structural population accounted for by the largest 100 clusters with colors depicting the 2D-dPCA space distance of a given cluster to the largest cluster. (C) The eight largest clusters, showing centroid structures and additional members. Side chains shown in stick representation for the centroid, the red to blue color pattern maps from the N- to C-terminus.

Cluster structural analysis showed a host of diverse conformations (Figure 2C). Gen4 generally adopted states with helical components linked by small loops. The N- and C-terminal regions generally exhibited more variability for any given cluster (e.g. Cluster 4), this observation led to using only residues 2–27 for clustering. The active site cysteine was observed

in numerous conformations with no clear, predominant structural motif that might lead to the peptide's high reactivity.

Centroid-to-centroid backbone angle differences highlight the breadth of structural similarity as well as differences across the entire Gen4 population. Starting from a fully connected centroid graph, we solved the approximate Steiner tree problem to represent the conformational space (Figure 3A). To do so, the distance between a pair of two centroids (calculated using the backbone angles for residues 2–27) represented the cost of including an edge between the pair in the final minimum spanning tree. This led to connecting similar clusters, thus generating a low dimensional representation for the entire conformational space. To better display the extent of similar clusters, we added edges between highly similar clusters creating an augmented Steiner tree (solid blue lines for edges between similar structures in the initial tree, dashed for those not initially included). This depicted the relationship between all clusters in terms of their similarity in the original high dimensional space, which is not necessarily retained in the dPCA analysis, and helped to organize the conformational space. For Gen4, while the FES showed many broad minima, the augmented Steiner tree revealed that the majority of these structures were close in terms of their backbone Euclidean distances. It highlighted several nodes with a high degree of connectivity and that many of the more populated states were structurally similar to smaller clusters. This was likewise not obvious from the FES-mapped cluster analysis. Further, because it depicted all clusters, this representation localized and contextualized the small clusters in terms of neighboring clusters. It is essential to note that this analysis was only used to map the general space of conformations, not to ascribe a mechanism or path of conformational change.
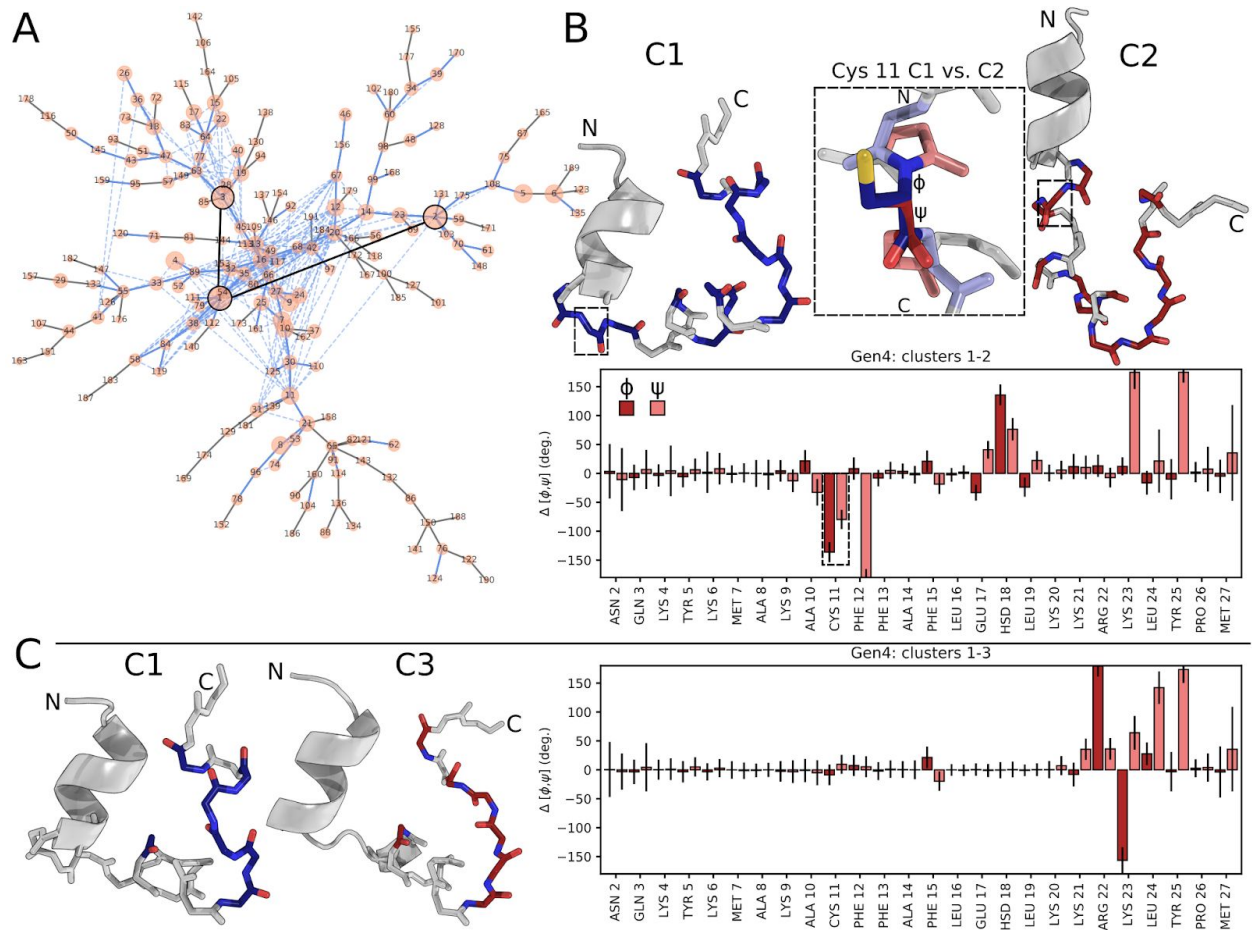
Figure 3. Select structural changes in the backbone torsion angles for MP01-Gen4. (A) Structural similarity network of Gen4 REMD cluster centroids with node size representing the population percent. Solid lines show edges from the original approximate Steiner tree solution, blue lines (solid or dashed) link structurally similar centroids. Bold black lines connecting clusters 1, 2, and 3 are for visualization only and not part of the Steiner tree. (B). Cluster 1 to cluster 2 ϕ and ψ angle differences mapped to centroid structures (top) with Cys 11 comparison (inset) and average cluster angle changes (bottom). (C) Cluster 1 to cluster 3 ϕ and ψ angle differences mapped to centroid structures (left) and average cluster angle changes (right).

Differences between select clusters are attributable to changes in a subset of backbone dihedral angles. We calculated the average cluster-to-cluster ϕ and ψ backbone angle changes between clusters 1 and 2 along with 1 and 3 (Figure 3A bold black lines and circled nodes, B, and C). The nodes of these two pairs were not directly linked in the Steiner tree but were connected by only a small number of edges. Cluster 1 and 2 differed in the central and C-terminal portions of Gen4, with large or small changes in roughly 11 residues. In this comparison, the active site Cys11 showed large changes in both its ϕ and ψ angles with near to a

complete 180° flip in its ϕ angle (Figure 3B, inset). The centroid structures for clusters 1 and 3 appeared visually similar (despite being approximately the same distance away from each other as clusters 1 and 2 in dPCA space); mapping to this, their average cluster angle differences were small up to Lys21 (aside from Phe15, Figure 3C).

Non-covalent interactions between the CA and MP01-Gen4
Explicit solvent MD of unreacted MP01-Gen4 and CA (Gen4+CA) found the CA to substantially interact with the peptide, but the simulation was limited in its extent of conformational sampling. Approximately 3.5 μs of unbiased simulation were performed from a starting state with the CA ~8 Å from Cys11 and using a partially helical Gen4 conformation from the second folding trajectory (Table S1). Over the course of the simulation, the peptide SASA fluctuated around 3500 $Å^2$ with a range between 2800–4200 $Å^2$, less than that observed in the Gen4 simulations (Figure S7). Gen4's R_g remained near 10 Å for the first two microseconds after which larger conformational changes occured with an accompanying increase in R_g. The secondary structure fluctuated with both long- and short-lived states (Figure S8). The CA appeared to interact with all portions of MP01, including both, minimally structured, termini. Multiple residues were observed close to the CA for several segments of the trajectory, notably Tyr5 showed a < 10 Å distance between its C_β and the perfluorophenyl-bridging sulfur (Figure S9). The CA primarily remained near the peptide, with a median shortest distance of 2.36 Å between any atoms of the two (Figure S10). On the occasions the CA diffused away from MP01 (unrelated to crossing a periodic boundary), it appeared to rapidly return; however, the impact of the high CA concentration (~4.6 mM due to the simulation box size) may make this observation overly optimistic.

REMD suggests an altered conformational landscape of Gen4+CA
REMD simulations suggested an altered and restricted conformational landscape for MP01-Gen4 in the presence of the CA. Structure clustering and analysis of the 2D-FES for Gen4+CA projected onto the dPCA space of Gen4 showed a shifted landscape with new states (e.g. near cluster 2) and larger energetic barriers between select states (e.g. the distinct well near cluster 5), reaching to between 2.1 and 2.4 kcal mol^{-1} (Figure 4A). The energetic landscape and cluster positioning seen in the FES was not impacted by the choice of projecting the Gen4+CA data using the Gen4-fit dPCA model (Figure S11). Additionally we verified that similar structures were observed between simulations given coordinates that were in PC space (Figure S12). Despite the shift in states, several of the states accessible to the lone peptide remained accessible (conformations near those of Gen4's cluster 3 or 9, Figure 2A). A bipartite graph comparing the similar clusters of Gen4 to Gen4+CA revealed that a number of states less populated by Gen4 (e.g. clusters 34, 47, and 84) increased in population size for Gen4+CA (mapping to clusters 1, 9 and 6 respectively, Figure S13). In contrast, select larger proportion Gen4 states, including
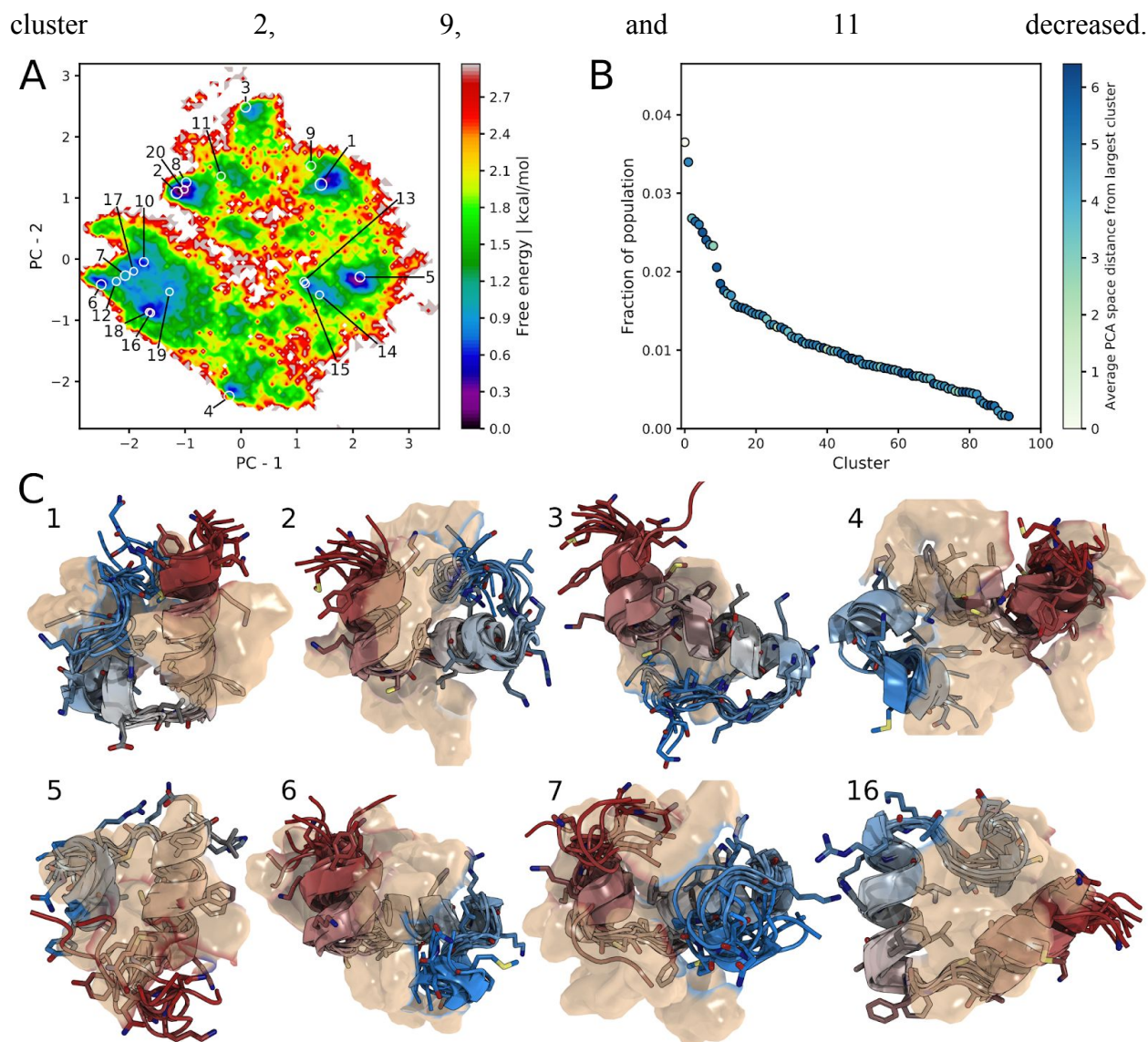
cluster 2, 9, and 11 decreased.

Figure 4. MP01-Gen4+CA populates a reduced set of new and old states with significant CA interactions across the entire peptide. (A) 2D-FES using the first two dPCA components with the locations of the top 20 largest clusters plotted at the coordinates for the cluster centroid. The cluster size is depicted by the size of the circle. (B) Fraction of the structural population accounted for by each cluster, with colors depicting the 2D-dPCA space distance of a given cluster to the largest cluster. (C) General conformations of select clusters and a transparent surface showing the location of the CA for the select structures. Side chains shown in stick representation for the centroid, the red to blue color pattern maps from the N- to C-terminus.

Compared to Gen4 by itself, fewer than half the number of clusters were required to capture the states of Gen4+CA (92 in total), with the largest two clusters accounting for ~3.5% of the population each (Figure 4B). The top 20 clusters now accounted for 42.8% of the population, suggesting the peptide predominantly exists in a smaller number of conformations in

9

the presence of the CA. This observation, along with many clusters possessing helical components, match well with the experimental increase in ɑ-helicity (from a partially coiled state) for the non-reactive, Cys-to-Ser, version of MP01-Gen4 in the presence of CA[15].

Gen4+CA clusters showed helical peptide conformations and that, while the CA interacts with much of the peptide, it does so in a cluster-specific manner. Overall, at 300 K the median, minimum distance between Gen4 and the CA was 2.46 Å with little difference in the distribution for the top 20 clusters and no biasing force applied (Figures S14, S15). The largest clusters displayed a variety of structural arrangements with one or two helical segments frequently accompanied by one or both of the terminal regions in a coiled state (Figure 4C). Overlaying cluster members showed the CA to interact in a number of different conformations and with different parts of the peptide (surface representation, Figure 4C). We further characterized the location of the CA in each cluster by measuring the distance between atoms on the CA (excluding the ethyl alcohol) and peptide atoms for all members and determining the frequency at which a residue was within 3.5 Å from the CA for the 10 largest clusters (Figures 5 and S16). Structurally similar clusters, like cluster 2 and 8, showed more similar CA interaction patterns. Apart from these, most exhibited more unique interactions, for instance cluster 3 only showed eight residues above 20%, four of which (Ala10, Phe12, Phe15 and Met27) were above 80% (Figure 5). Many clusters showed interactions with the C-terminal residues as well as Phe15 and 12, while the most N-terminal residues were less frequently observed to be close (Figure 5 shows cluster centroid structures with a single CA as examples). Measuring the distance from peptide atoms to the electrophilic aromatic carbon on the CA similarly showed cluster-dependent distance profiles (Figure S17). For this analysis, the cut off distance was increased to 4.5 Å, however, even with this, the overall percentages of frames was reduced—a complete analysis of this type likely requires substantially more sampling. To gain a broad understanding of the residue preference of the CA, we analyzed the top 20, 40, 60 and all 92 clusters, finding the ensemble interaction pattern to be mostly captured with just the top 20 clusters (Figure S18). Additionally, Phe12 and 15, Pro26 and Met27 were found to interact the most; in contrast, the polar and charged Glu17, Arg22, Lys20, Lys21, and Lys23 were observed to be close in < 10% of all frames.
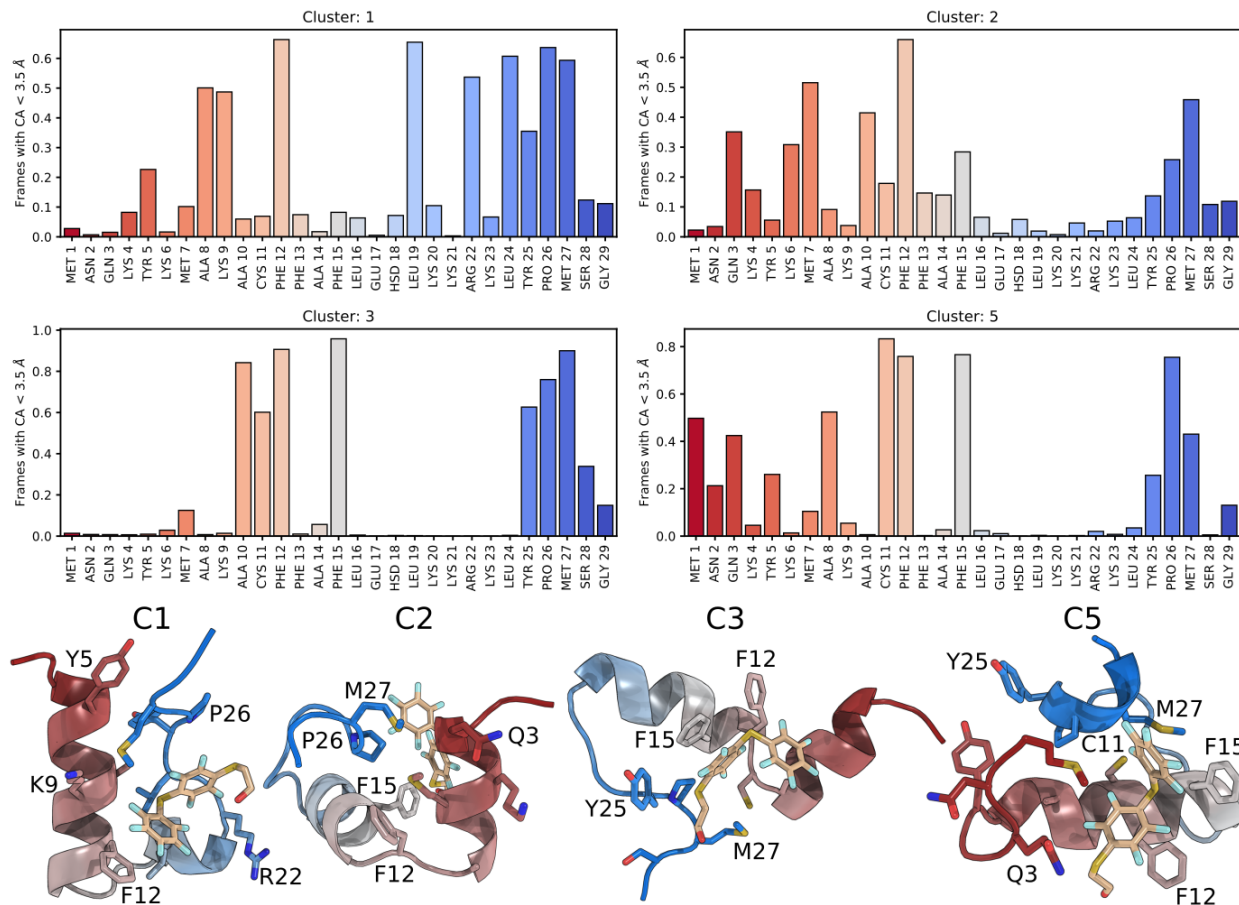
Figure 5. The CA interacts with different parts of MP01-Gen4 in a conformation-dependant manner. Per residue fraction of cluster members (clusters 1, 2, 3 and 5) with a distance less than 3.5 from any atom in the CA not including the ethyl alcohol tail (top). Cluster 1, 2, 3 and 5 centroid structures showing residues (along with the CA) with CA-distance fractions greater than 20% in sticks (select residues labeled, bottom).

Network construction from cluster centroids of the Gen4+CA REMD simulation showed two subnetworks completely linked by small distance edges (the first containing nodes connected to clusters 2 and 3 and the second with nodes linked to cluster 6 and 7, Figure 6A). The results agreed well with cluster positioning on the 2D-FES; for example, clusters 1, 4, and 5 each showed local connections to smaller clusters, but were not attached to the larger subnetworks through any short distance edges, mirroring their separated minima. Similarly, the observed connectivity mapped to average cluster backbone differences, in which small distance (blue) edges connected clusters that differed in only a few dihedral angles (Figure S19).

To further understand differences between clusters and how this may lead to specific CA interaction profiles, we looked at two pairs of nodes (clusters); one connected by two short distance edges (2 and 3) and another 'connected' by several longer distances (2 and 5). Clusters 2 and 3 primarily differed in backbone angles in residues 20–25 as well as Phe12 (Figure 6B).

While the N-terminal portion remained largely unchanged and helical, the change of the C-terminal angles resulted in several residues (Gln3, Lys6, Met7) that heavily interacted with the CA in cluster 2 to almost completely lack CA interactions in cluster 3 (Figure 5). The more distant clusters 2 and 5 show a substantial rearrangement in the majority of backbone angles (Figure 6C) along with a shifted interaction profile in which Met1, Ala8, Cys11, Phe15 and Pro26 increased as Lys6, Met7 and Ala10 decreased in their extent of interaction (Figure 5).
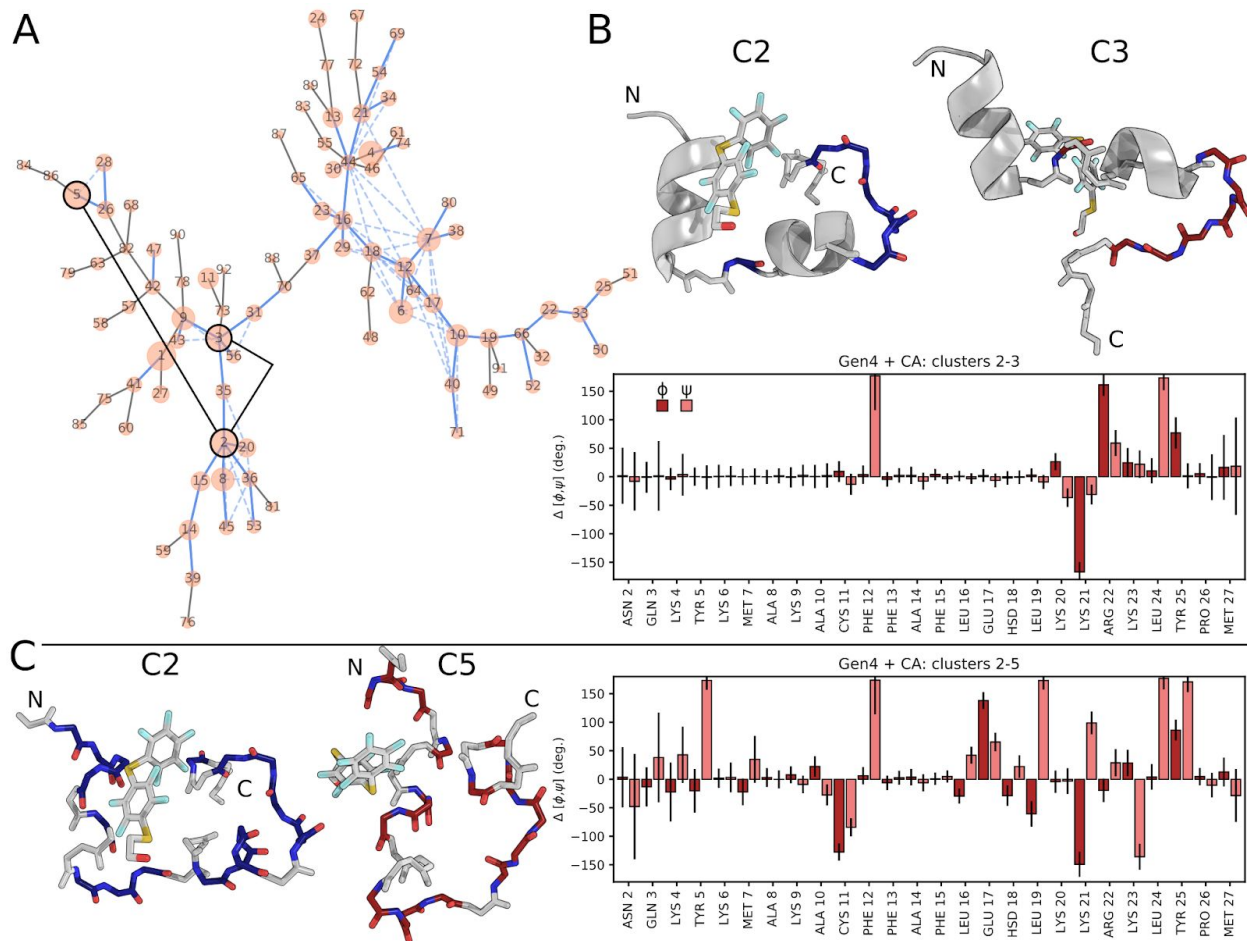

Figure 6. Select structural changes in the backbone torsion angles for MP01-Gen4+CA. (A) Structural similarity network of Gen4+CA REMD cluster centroids with node size representing the population percent. Solid lines show edges from the original approximate Steiner tree solution, blue lines (solid or dashed) link structurally similar centroids. Bold black lines connecting clusters 2, 3, and 5 are for visualization only and not part of the Steiner tree. (B). Cluster 2 to cluster 3 ϕ and ψ angle differences mapped to centroid structures (top) and average cluster angle changes (bottom). (C) Cluster 2 to cluster 5 ϕ and ψ angle differences mapped to centroid structures (left) and average cluster angle changes (right).

Covalent Gen4-CA explicit solvent MD slowly samples a small number of states

Explicit solvent Gen4-CA simulations sampled different conformational regions relative to each other. Simulation 4 showed minimal state residence and more structural alterations than simulation 5. Starting from the same conformation as the Gen4+CA explicit solvent simulation, simulation 4 quickly transitioned to mixed coil-turn states with large and rapid SASA and R_g fluctuations between metastable states lasting tens to low hundreds of nanoseconds (Figures S20 and S21). Simulation 5, which started from a covalently labeled structure in the largest cluster of the Gen4 REMD simulation, slowly transitioned out of its starting conformation (Figure S22). It then adopted several longer lasting states, including one with a central helix and coiled termini at ~1750 ns, it remained in this state despite large fluctuations of the terminal regions (causing spikes in the R_g and SASA) for the remainder of the simulation (Figures S22 and S23). The simulations sampled different parts of the energetic landscape without convergence, again requiring enhanced sampling for a more complete conformational analysis.

REMD suggests a shift toward a more limited conformational landscape upon CA labeling
Two conformational families dominated the energetic and conformational landscape of the Gen4-CA REMD simulation. The slightly deeper minima comprised clusters 1, 4, and 5 while the other housed clusters 6, 7, 8, 12 and 13 (Figure 7A). These were considered structural families due to minimal angular differences between the family members. Clusters 1 and 13 acted as 'hubs' from which the other structures were minimally different. Clusters 6 or 7 could similarly have served as hub nodes; however, we highlight 13 due to the similarity of the 13-to-6 change versus that of 1-to-5 (Met27 ψ angle change) as well as the 13-to-7 change relative to 1-to-4 (Gln3 ψ angle change, Figure S24). The other minima in the 2D-FES were less populated, with many of the states from both the Gen4 and Gen4+CA REMD simulations missing. However, this still suggested there to be many accessible states, separated by relatively small energetic differences. As with the Gen4+CA FES, no major change was observed when the dPCA was fit on the Gen4-CA data instead of projecting to the space defined by dPCA fit on the Gen4 REMD data, e.g. the two main minima with the same clusters were observed (Figure S11). Likewise, proximally located structures in the 2D-FES coordinates showed similar 3D structures when comparing Gen4-CA to Gen4+CA (Figure S12).
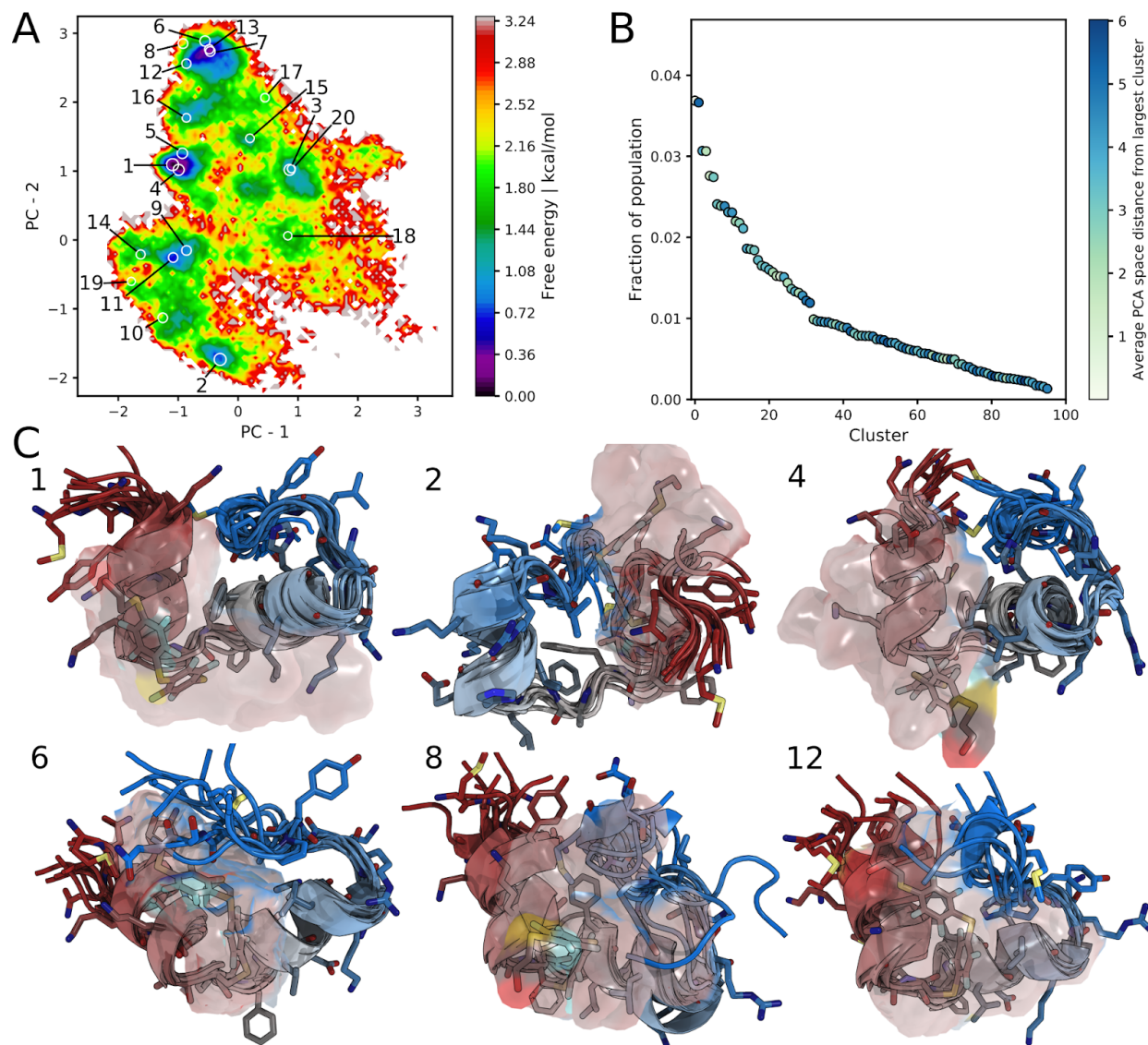
Figure 7. MP01-Gen4-CA accesses a conformationally restricted set of states relative to MP01-Gen4 and MP01-Gen4+CA. (A) 2D-FES projected onto the first two dPCA components from Gen4. The locations of the top 20 largest clusters are plotted at the coordinates for the cluster centroid. The cluster size is depicted by the size of the circle. (B) Fraction of the structural population accounted for by each cluster, with colors depicting the 2D-dPCA space distance of a given cluster to the largest cluster. (C) General conformations of select clusters and a transparent surface showing the location of the CA for the select structures. Side chains, including the Cys-CA shown in stick representation for the centroid, the red to blue color pattern maps from the N- to C-terminus.

The 20 largest clusters contained 47.8% of the population, a small increase relative to the Gen4+CA simulation but in line with the highly similar circular dichroism (CD) spectra of

Gen4-CA and that of the Gen4 Cys-to-Ser mutant with a large concentration of CA. In total, 96 clusters were used to accommodate the complete population (Figure 7B). The structural states populated by the Gen4-CA simulation largely differed from the Gen4+CA structural landscape. Many of the largest Gen4+CA clusters (1, 4–7, 9, 11–19) disappeared with no similar (Euclidean distance < 4) structures in the Gen4-CA simulation (Figure S13). Interestingly, the Gen4-CA cluster 13 family all showed similarity to two clusters from the Gen4 simulation (cluster 63 and 64) while the Gen4-CA cluster 1 family had no similar structures in the free peptide simulation. The majority of conformations displayed by Gen4-CA exhibited two short helical regions flanking a loop or turn with the labeled cysteine and a C-terminal coiled tail MP that wrapped back to interact with the peptide or CA depending on the broad structural family (Figure 7C).

The network structure of the Gen4-CA simulation contained a single large subnetwork, connected by only similar centroid edges (Figure 8A). Within this subnetwork, two highly self-similar groups could be seen, one including clusters 1, 4, 5, 23, 24, 26, 33 and 69 ('family 1') and the other with clusters 6, 7, 8, 12, 13, 17, 21, 22, 27, 30, 46 and 50 ('family 13'). These two groups were connected by only a single similar edge between clusters 8 and 23, linking the families by a series of backbone angle alterations (in purely a graphical, non-mechanistic manner, Figure S25). Together, the two families accounted for 36.6% of all structures, with the cluster 1 family supplying 15.5% while the other gave 21.1%. Nevertheless, there remained a large number of additional clusters, representing a non-trivial portion of the population, in alternative conformations, predominantly connected by non-similar edges.
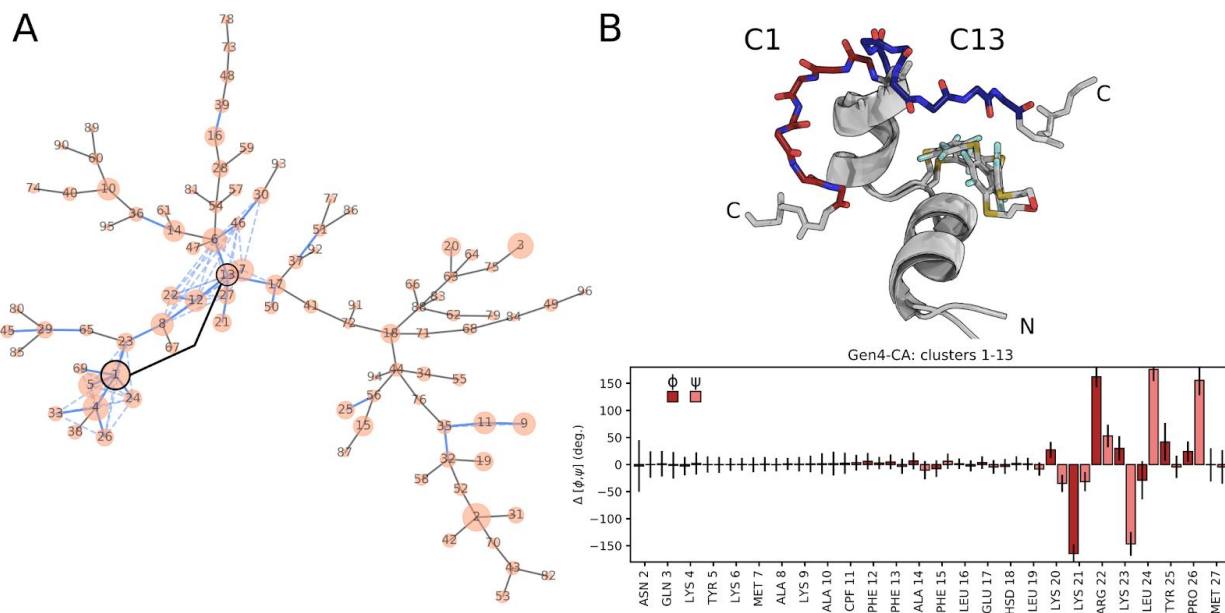

Figure 8. Backbone angle differences between primary clusters of MP01-Gen4-CA. (A) Structural similarity network of Gen4-CA REMD cluster centroids with node size representing the population percent. Solid lines show edges from the original approximate Steiner tree solution, blue lines (solid or dashed) link structurally similar centroids. Bold black lines

connecting clusters 1 and 13 are for visualization only and not part of the Steiner tree. (B). Cluster 1 to cluster 13 ϕ and ψ angle differences mapped to centroid structures with Cys-CA side chain shown in sticks (top) and average cluster angle changes (bottom).

Seeing a larger percentage of the population adopting relatively similar states, we analyzed the average cluster backbone angle differences between the two groups. The primary difference between clusters 1 and 13 resided in large ϕ and ψ angle changes for residues 20–26 (Figure 8B). The main structural effect was to shift the C-terminus from interacting or partially covering the CA in cluster 13 to a state in which it interacted intramolecularly on the opposite side of the peptide. This left the main helix-turn-helix motif intact. From these two 'hub' nodes, other clusters in their respective groups could be accessed with mostly a single angle alteration. These data, together with the cluster and FES analysis suggest that the covalently labeled Gen4-CA displays a reduced conformational landscape relative to both of the previous REMD analyses.

Discussion

We first contextualize the work with the force field used and choice of implicit solvent REMD. A number of documented inaccuracies may have occurred; for instance, implicit solvent models frequently over-stabilize structural features[54] and salt bridges[55], and can inadequately treat solute-solvent nonpolar terms[56]. Similarly, overly compact structures of disordered proteins have been reported using different combinations of force fields and explicit water models[57]. Given the dynamic nature of MP01, the use of more modern force fields like CHARMM36m[58], RSFF2C[59], or a99SB-*disp*[60] and modeling water dispersion with the TIP4P-D water model[57] would likely produce a more accurate description. However, to begin to understand MP01's structural landscape and its relative energies, while lacking knowledge on the number of states, the faster implicit solvent calculations were deemed an necessary tradeoff. Further, given the experimental increase in α-helicity and structural rigidity upon interaction and labeling, inaccuracies of the CHARMM36 force field for modeling dynamic peptides may be mitigated as these two systems experimentally behave more like a structured protein.

Differences from experimental conditions may also have unknowingly altered the results. Under experimental peptide concentrations (≤ 100 µM for reactions) MP01 appears monomeric, however at higher concentrations it begins to form larger aggregates[15]. Thus, using a single copy of MP01 may not fully reflect its solution behavior, where multiple copies may interact. Similarly, we cannot rule out that multiple copies of the CA interact with MP01 or that the complete CA would provide significantly different results. These points are partly mitigated due to the extent of conformational landscape alteration observed in the Gen4+CA REMD simulation and that experimentally a CA variant containing only the perfluoroaromatic attached to a carboxy-PEG$_{12}$-thiol shows no change in reactivity relative to the full CA[14].

A, perhaps, unexpected outcome of our methods to analyze conformational states was the representational power of the similarity-augmented approximate Steiner trees. Networks have a rich history in capturing important aspects of atomistic simulations, including Markov state models (MSMs)[61–64], Gaussian and anisotropic network models[65,66] along with structural similarity-based networks[67]. 1D- or 2D-FES analysis may collapse different regions of conformational space onto close, if not overlapping regions, which may skew the interpretation or hide important variation. Network or graph structure offers a complementary view to the structural space, and depending on the level of abstraction used in clustering (higher abstraction from fewer clusters as more information is combined) could illustrate different biophysical relationships. Further, minimal or no agglomerative clustering with similarity-augmented Steiner trees, combined with ideas from kinetic network models[68] or dynamic graphical models[69] may show use for extracting stationary and dynamic properties.

The REMD simulations revealed a surprisingly large difference between the Gen4+CA and Gen4-CA simulations. Ensemble CD measurements demonstrated both systems to adopt similar helical signatures (using a Cys-to-Ser mutation for a nonreactive MP01), however, the REMD analysis suggests this is achieved in two different manners. While Gen4-CA is best characterized by a large fraction of states possessing a helix-turn-helix and flexible C-terminal tail, Gen4+CA is characterized by its breadth of states, including conformations not seen or minimally populated in the other two REMD simulations (e.g. clusters 4 and 5). Similarly, Gen4+CA shows a pronounced shift in its 2D-FES to regions that are less populated by Gen4 and Gen4-CA (near clusters 1 and 5, Figure 4). The Gen4+CA simulation did, however, serve to increase the population density near cluster 2, 8 and 20 which were not represented by Gen4, but constituted key components of the Gen4-CA landscape. The region near cluster 3 (Gen4+CA) was less favored; with the hypothesis that the C-terminal tail acts like a 'clamp' and covers the CA once reacted, such structural alteration prior to CA interaction would serve to exclude the CA (cluster 13 and related family, Gen4-CA, Figures 7 and 8).

The conformational features from the REMD simulations help to understand and generate hypotheses for the role of the terminal regions. The MD-based observation of interactions between the CA and the terminal regions of Gen4 may suggest that these regions do not only play a role in the peptide's structure (e.g. the N-terminal helix seen in Gen4-CA), but could help to recruit, retain and thus increase the effective local concentration of the CA. Similarly, the Gen4-CA simulation showed that the C-terminal portion made substantial CA interactions in one of the two primary states, hiding the CA from solvent (cluster 13 and associated family, Figure 7 and 8). Given the large number of peptide conformational states, each with unique CA interaction profiles (Figure 5, S9 and S16–18), we suggest that prior to reaction, Gen4 and the CA do not possess a single interaction complex, but many. To continue, we propose that the effect of N- and C-terminal truncations (previously shown to decrease the reaction rate constant[15]) is to decrease the extent of peptide-to-CA interactions, along with possible

intramolecular interactions. This may lead to a decrease in reactivity, but not necessarily abolish it, and help explain the experimental truncation series.

The simulations highlight a decrease in accessible conformational space and suggest a role for flexibility. EPR measurements uncovered a general decrease in dynamics throughout the entire Gen4 peptide upon reaction, with the largest decrease found at the termini. While not an exact comparison, the EPR observed decrease in dynamics may manifest at the ensemble conformational level by the contraction of the observed structural landscape to the two primary structural families (Gen4 versus Gen4-CA). For the two families of Gen4-CA, the major differences involve grouped backbone angle changes, mostly in the C-terminus (Figure 8). Relative to the many, diverse transitions between the host of states of Gen4, Gen4-CA dynamics may be dominated by this reduced set of dynamics and result in the observed EPR measurements. This leads to the importance of flexibility for the reactivity of Gen4. Gen4 was shown to decrease in reactivity with high concentrations of the structure stabilizing additive trimethylamine N-oxide[70] (TMAO, 0.5–1.0 M)[15]. Thus, reducing the flexibility or stabilizing a set of structure with TMAO may serve to decrease MP01's ability to undergo necessary structural rearrangements, for instance, to reach one of the states unique to Gen4+CA.

While many observations explained or possessed experimental correlations, the helical content from the REMD simulations did not map well to experiment (Figure S26). Helices were substantially less prevalent than turn and coil secondary structures, and showed very large standard deviations for each of the lower temperature replicas in the three types of simulation. The forcefield, implicit solvation and extent of sampling are prominent factors that may have led to this, as well as the yet unknown importance of simulating multiple peptides and CAs.

This study presents the characterization of a conformationally dynamic, reactive peptide. There exist many other abiotic peptides that similarly undergo structural alterations upon reaction, as well as others that do not[35]. The prevalence and mechanism of this feature in natural peptides and miniproteins (whether they bind or react with proteins, nucleic acid or small molecules) remains an open question. As does the extent of peptides with defined folds versus those that adopt a broad conformational landscape. The growing realization of the biological importance of peptides and miniproteins leaves significant room and sufficient impetus for future experimental and computational studies into their properties.

Supporting information description
Explicit solvent time courses with snapshots and secondary structure, CA distance plots from select residues and closest residues plots along with REMD simulation comparison graphs can be found in the accompanying supporting information.

Corresponding author

*Ethan D. Evans (eevans@mit.edu, 617-324-4931)

References

(1) Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. Energy Parameters in Polypeptides. VII. Geometric Parameters, Partial Atomic Charges, Nonbonded Interactions, Hydrogen Bond Interactions, and Intrinsic Torsional Potentials for the Naturally Occurring Amino Acids. *J. Phys. Chem.* **1975**, *79* (22), 2361–2381. https://doi.org/10.1021/j100589a006.

(2) McCammon, J. A.; Northrup, S. H.; Karplus, M.; Levy, R. M. Helix–Coil Transitions in a Simple Polypeptide Model. *Biopolymers* **1980**, *19* (11), 2033–2045. https://doi.org/10.1002/bip.1980.360191108.

(3) Bruccoleri, R. E.; Karplus, M. Prediction of the Folding of Short Polypeptide Segments by Uniform Conformational Sampling. *Biopolymers* **1987**, *26* (1), 137–168. https://doi.org/10.1002/bip.360260114.

(4) Daggett, V.; Kollman, P. A.; Kuntz, I. D. Molecular Dynamics Simulations of Small Peptides: Dependence on Dielectric Model and PH. *Biopolymers* **1991**, *31* (3), 285–304. https://doi.org/10.1002/bip.360310304.

(5) Némethy, G.; Leach, S. J.; Scheraga, H. A. The Influence of Amino Acid Side Chains on the Free Energy of Helix-Coil Transitions1. *J. Phys. Chem.* **1966**, *70* (4), 998–1004. https://doi.org/10.1021/j100876a008.

(6) O'Neil, K. T.; DeGrado, W. F. A Thermodynamic Scale for the Helix-Forming Tendencies of the Commonly Occurring Amino Acids. *Science* **1990**, *250* (4981), 646–651. https://doi.org/10.1126/science.2237415.

(7) Scholtz, J. M.; Marqusee, S.; Baldwin, R. L.; York, E. J.; Stewart, J. M.; Santoro, M.; Bolen, D. W. Calorimetric Determination of the Enthalpy Change for the Alpha-Helix to Coil Transition of an Alanine Peptide in Water. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88* (7), 2854–2858.

(8) Muñoz, V.; Serrano, L. Elucidating the Folding Problem of Helical Peptides Using Empirical Parameters. II†. Helix Macrodipole Effects and Rational Modification of the Helical Content of Natural Peptides. *J. Mol. Biol.* **1995**, *245* (3), 275–296. https://doi.org/10.1006/jmbi.1994.0023.

(9) Anderson, D. M.; Anderson, K. M.; Chang, C.-L.; Makarewich, C. A.; Nelson, B. R.; McAnally, J. R.; Kasaragod, P.; Shelton, J. M.; Liou, J.; Bassel-Duby, R.; Olson, E. N. A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle

Performance. *Cell* **2015**, *160* (4), 595–606. https://doi.org/10.1016/j.cell.2015.01.009.

(10) D'Lima, N. G.; Ma, J.; Winkler, L.; Chu, Q.; Loh, K. H.; Corpuz, E. O.; Budnik, B. A.; Lykke-Andersen, J.; Saghatelian, A.; Slavoff, S. A. A Human Microprotein That Interacts with the MRNA Decapping Complex. *Nat. Chem. Biol.* **2017**, *13* (2), 174–180. https://doi.org/10.1038/nchembio.2249.

(11) Sberro, H.; Fremin, B. J.; Zlitni, S.; Edfors, F.; Greenfield, N.; Snyder, M. P.; Pavlopoulos, G. A.; Kyrpides, N. C.; Bhatt, A. S. Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. *Cell* **2019**, *178* (5), 1245-1259.e14. https://doi.org/10.1016/j.cell.2019.07.016.

(12) Sharpe, I. A.; Gehrmann, J.; Loughnan, M. L.; Thomas, L.; Adams, D. A.; Atkins, A.; Palant, E.; Craik, D. J.; Adams, D. J.; Alewood, P. F.; Lewis, R. J. Two New Classes of Conopeptides Inhibit the A1-Adrenoceptor and Noradrenaline Transporter. *Nat. Neurosci.* **2001**, *4* (9), 902–907. https://doi.org/10.1038/nn0901-902.

(13) Bi, P.; Ramirez-Martinez, A.; Li, H.; Cannavino, J.; McAnally, J. R.; Shelton, J. M.; Sánchez-Ortiz, E.; Bassel-Duby, R.; Olson, E. N. Control of Muscle Formation by the Fusogenic Micropeptide Myomixer. *Science* **2017**, *356* (6335), 323–327. https://doi.org/10.1126/science.aam9361.

(14) Evans, E. D.; Pentelute, B. L. Discovery of a 29-Amino-Acid Reactive Abiotic Peptide for Selective Cysteine Arylation. *ACS Chem. Biol.* **2018**, *13* (3), 527–532. https://doi.org/10.1021/acschembio.7b00520.

(15) Evans, E. D.; Gates, Z. P.; Sun, Z.-Y. J.; Mijalis, A. J.; Pentelute, B. L. Conformational Stabilization and Rapid Labeling of a 29-Residue Peptide by a Small Molecule Reaction Partner. *Biochemistry* **2019**, *58* (10), 1343–1353. https://doi.org/10.1021/acs.biochem.8b00940.

(16) Liwo, A.; Wawak, R. J.; Scheraga, H. A.; Pincus, M. R.; Rackovsky, S. Prediction of Protein Conformation on the Basis of a Search for Compact Structures: Test on Avian Pancreatic Polypeptide. *Protein Sci.* **1993**, *2* (10), 1715–1731. https://doi.org/10.1002/pro.5560021016.

(17) Krüger, P.; Strassburger, W.; Wollmer, A.; van Gunsteren, W. F. A Comparison of the Structure and Dynamics of Avian Pancreatic Polypeptide Hormone in Solution and in the Crystal. *Eur. Biophys. J. EBJ* **1985**, *13* (2), 77–88. https://doi.org/10.1007/BF00256528.

(18) Zagrovic, B.; Snow, C. D.; Khaliq, S.; Shirts, M. R.; Pande, V. S. Native-like Mean Structure in the Unfolded Ensemble of Small Proteins. *J. Mol. Biol.* **2002**, *323* (1), 153–164. https://doi.org/10.1016/S0022-2836(02)00888-4.

(19) Mittal, J.; Best, R. B. Tackling Force-Field Bias in Protein Folding Simulations: Folding of Villin HP35 and Pin WW Domains in Explicit Water. *Biophys. J.* **2010**, *99* (3), L26–L28. https://doi.org/10.1016/j.bpj.2010.05.005.

(20) Pan, A. C.; Weinreich, T. M.; Piana, S.; Shaw, D. E. Demonstrating an Order-of-Magnitude Sampling Enhancement in Molecular Dynamics Simulations of Complex Protein Systems. *J. Chem. Theory Comput.* **2016**, *12* (3), 1360–1367. https://doi.org/10.1021/acs.jctc.5b00913.

(21) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain. *Biophys. J.* **2008**, *94* (10), L75–L77. https://doi.org/10.1529/biophysj.108.131565.

(22) Ensign, D. L.; Pande, V. S. The Fip35 WW Domain Folds with Structural and Mechanistic

Heterogeneity in Molecular Dynamics Simulations. *Biophys. J.* **2009**, *96* (8), L53–L55. https://doi.org/10.1016/j.bpj.2009.01.024.

(23) Satoh, D.; Shimizu, K.; Nakamura, S.; Terada, T. Folding Free-Energy Landscape of a 10-Residue Mini-Protein, Chignolin. *FEBS Lett.* **2006**, *580* (14), 3422–3426. https://doi.org/10.1016/j.febslet.2006.05.015.

(24) Okumura, H. Temperature and Pressure Denaturation of Chignolin: Folding and Unfolding Simulation by Multibaric-Multithermal Molecular Dynamics Method. *Proteins* **2012**, *80* (10), 2397–2416. https://doi.org/10.1002/prot.24125.

(25) Singhal, N.; Snow, C. D.; Pande, V. S. Using Path Sampling to Build Better Markovian State Models: Predicting the Folding Rate and Mechanism of a Tryptophan Zipper Beta Hairpin. *J. Chem. Phys.* **2004**, *121* (1), 415–425. https://doi.org/10.1063/1.1738647.

(26) Zhou, R. Trp-Cage: Folding Free Energy Landscape in Explicit Water. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (23), 13280–13285. https://doi.org/10.1073/pnas.2233312100.

(27) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1), 141–151. https://doi.org/10.1016/S0009-2614(99)01123-9.

(28) Anand, P.; Nandel, F. S.; Hansmann, U. H. E. The Alzheimer's β Amyloid (Aβ1–39) Monomer in an Implicit Solvent. *J. Chem. Phys.* **2008**, *128* (16), 165102. https://doi.org/10.1063/1.2907718.

(29) Kim, S.; Takeda, T.; Klimov, D. K. Globular State in the Oligomers Formed by Aβ Peptides. *J. Chem. Phys.* **2010**, *132* (22), 225101. https://doi.org/10.1063/1.3447894.

(30) Rao, F.; Caflisch, A. Replica Exchange Molecular Dynamics Simulations of Reversible Folding. *J. Chem. Phys.* **2003**, *119* (7), 4035–4042. https://doi.org/10.1063/1.1591721.

(31) Rhee, Y. M.; Pande, V. S. Multiplexed-Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. *Biophys. J.* **2003**, *84* (2), 775–786. https://doi.org/10.1016/S0006-3495(03)74897-8.

(32) Yadahalli, S.; Li, J.; Lane, D. P.; Gosavi, S.; Verma, C. S. Characterizing the Conformational Landscape of MDM2-Binding P53 Peptides Using Molecular Dynamics Simulations. *Sci. Rep.* **2017**, *7* (1), 1–12. https://doi.org/10.1038/s41598-017-15930-4.

(33) Yang, L.; Schepartz, A. Relationship between Folding and Function in a Sequence-Specific Miniature DNA-Binding Protein. *Biochemistry* **2005**, *44* (20), 7469–7478. https://doi.org/10.1021/bi050121h.

(34) Stanley, N.; Esteban-Martín, S.; De Fabritiis, G. Kinetic Modulation of a Disordered Protein Domain by Phosphorylation. *Nat. Commun.* **2014**, *5*, 5272. https://doi.org/10.1038/ncomms6272.

(35) Evans, E. D.; Pentelute, B. L. Studies on a Landscape of Perfluoroaromatic-Reactive Peptides. *Org. Biomol. Chem.* **2019**, *17* (7), 1862–1868. https://doi.org/10.1039/C8OB01678J.

(36) Lindemann, W. R.; Evans, E. D.; Mijalis, A. J.; Saouaf, O. M.; Pentelute, B. L.; Ortony, J. H. Quantifying Residue-Specific Conformational Dynamics of a Highly Reactive 29-Mer Peptide. *Sci. Rep.* **2020**, *10* (1), 1–8. https://doi.org/10.1038/s41598-020-59047-7.

(37) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802. https://doi.org/10.1002/jcc.20289.

(38) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D.

Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone φ, ψ and Side-Chain X1 and X2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257–3273. https://doi.org/10.1021/ct300400x.

(39) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616. https://doi.org/10.1021/jp973084f.

(40) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52* (12), 3144–3154. https://doi.org/10.1021/ci300363c.

(41) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52* (12), 3155–3168. https://doi.org/10.1021/ci3003649.

(42) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

(43) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

(44) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant Pressure Molecular Dynamics Algorithms. *J. Chem. Phys.* **1994**, *101* (5), 4177–4189. https://doi.org/10.1063/1.467468.

(45) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J. Chem. Phys.* **1995**, *103* (11), 4613–4621. https://doi.org/10.1063/1.470648.

(46) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(47) Frishman, D.; Argos, P. Knowledge-Based Protein Secondary Structure Assignment. *Proteins Struct. Funct. Bioinforma.* **1995**, *23* (4), 566–579. https://doi.org/10.1002/prot.340230412.

(48) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112* (16), 6127–6129. https://doi.org/10.1021/ja00172a038.

(49) Tanner, D. E.; Phillips, J. C.; Schulten, K. GPU/CPU Algorithm for Generalized Born/Solvent-Accessible Surface Area Implicit Solvent Calculations. *J. Chem. Theory Comput.* **2012**, *8* (7), 2521–2530. https://doi.org/10.1021/ct3003089.

(50) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11* (11), 5525–5542. https://doi.org/10.1021/acs.jctc.5b00743.

(51) Aalten, D. M. F. V.; Groot, B. L. D.; Findlay, J. B. C.; Berendsen, H. J. C.; Amadei, A. A Comparison of Techniques for Calculating Protein Essential Dynamics. *J. Comput. Chem.*

1997, *18* (2), 169–181.
https://doi.org/10.1002/(SICI)1096-987X(19970130)18:2<169::AID-JCC3>3.0.CO;2-T.

(52) Mu, Y.; Nguyen, P. H.; Stock, G. Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *Proteins Struct. Funct. Bioinforma.* **2005**, *58* (1), 45–52. https://doi.org/10.1002/prot.20310.

(53) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. Dihedral Angle Principal Component Analysis of Molecular Dynamics Simulations. *J. Chem. Phys.* **2007**, *126* (24), 244111. https://doi.org/10.1063/1.2746330.

(54) Bursulaya, B. D.; Brooks, C. L. Comparative Study of the Folding Free Energy Landscape of a Three-Stranded β-Sheet Protein with Explicit and Implicit Solvent Models. *J. Phys. Chem. B* **2000**, *104* (51), 12378–12383. https://doi.org/10.1021/jp0027602.

(55) Zhou, R.; Berne, B. J. Can a Continuum Solvent Model Reproduce the Free Energy Landscape of a β-Hairpin Folding in Water? *Proc. Natl. Acad. Sci.* **2002**, *99* (20), 12777–12782. https://doi.org/10.1073/pnas.142430099.

(56) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. On the Nonpolar Hydration Free Energy of Proteins: Surface Area and Continuum Solvent Models for the Solute−Solvent Interaction Energy. *J. Am. Chem. Soc.* **2003**, *125* (31), 9523–9530. https://doi.org/10.1021/ja029833a.

(57) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119* (16), 5113–5123. https://doi.org/10.1021/jp508971m.

(58) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; Groot, B. L. de; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14* (1), 71–73. https://doi.org/10.1038/nmeth.4067.

(59) Kang, W.; Jiang, F.; Wu, Y.-D. Universal Implementation of a Residue-Specific Force Field Based on CMAP Potentials and Free Energy Decomposition. *J. Chem. Theory Comput.* **2018**, *14* (8), 4474–4486. https://doi.org/10.1021/acs.jctc.8b00285.

(60) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci.* **2018**, *115* (21), E4758–E4766. https://doi.org/10.1073/pnas.1800690115.

(61) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo. *J. Comput. Phys.* **1999**, *151* (1), 146–168. https://doi.org/10.1006/jcph.1999.6231.

(62) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States. *J. Chem. Phys.* **2007**, *126* (15), 155102. https://doi.org/10.1063/1.2714539.

(63) Buchete, N.-V.; Hummer, G. Peptide Folding Kinetics from Replica Exchange Molecular Dynamics. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2008**, *77* (3 Pt 1), 030902. https://doi.org/10.1103/PhysRevE.77.030902.

(64) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic Discovery of Metastable States for the Construction of Markov Models of Macromolecular Conformational Dynamics. *J. Chem. Phys.* **2007**, *126* (15), 155101. https://doi.org/10.1063/1.2714538.

(65) Doruker, P.; Atilgan, A. R.; Bahar, I. Dynamics of Proteins Predicted by Molecular

Dynamics Simulations and Analytical Approaches: Application to α-Amylase Inhibitor. *Proteins Struct. Funct. Bioinforma.* **2000**, *40* (3), 512–524. https://doi.org/10.1002/1097-0134(20000815)40:3<512::AID-PROT180>3.0.CO;2-M.

(66) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **2001**, *80* (1), 505–515. https://doi.org/10.1016/S0006-3495(01)76033-X.

(67) Rao, F.; Caflisch, A. The Protein Folding Network. *J. Mol. Biol.* **2004**, *342* (1), 299–306. https://doi.org/10.1016/j.jmb.2004.06.063.

(68) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. Protein Folding Pathways from Replica Exchange Simulations and a Kinetic Network Model. *Proc. Natl. Acad. Sci.* **2005**, *102* (19), 6801–6806. https://doi.org/10.1073/pnas.0408970102.

(69) Olsson, S.; Noé, F. Dynamic Graphical Models of Molecular Kinetics. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (30), 15001–15006. https://doi.org/10.1073/pnas.1901692116.

(70) Zou, Q.; Bennion, B. J.; Daggett, V.; Murphy, K. P. The Molecular Mechanism of Stabilization of Proteins by TMAO and Its Ability to Counteract the Effects of Urea. *J. Am. Chem. Soc.* **2002**, *124* (7), 1192–1202. https://doi.org/10.1021/ja004206b.