HHS Public Access

Author manuscript

J Phys Chem B. Author manuscript; available in PMC 2024 July 27.

Published in final edited form as:

J Phys Chem B. 2024 May 02; 128(17): 4063–4075. doi:10.1021/acs.jpcb.3c08492.

One Descriptor to Fold Them All: Harnessing Intuition and Machine Learning to Identify Transferable Lasso Peptide Reaction Coordinates

Gabriel C. A. da Hora,
Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States

Myongin Oh,
Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States

John D. M. Nguyen,
Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States

Jessica M. J. Swanson
Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States

Abstract

Identifying optimal reaction coordinates for complex conformational changes and protein folding remains an outstanding challenge. This study combines collective variable (CV) discovery based on chemical intuition and machine learning with enhanced sampling to converge the folding free energy landscape of lasso peptides, a unique class of natural products with knot-like tertiary structures. This knotted scaffold imparts remarkable stability, making lasso peptides resistant to proteolytic degradation, thermal denaturation, and extreme pH conditions. Although their direct synthesis would enable therapeutic design, it has not yet been possible due to the improbable occurrence of spontaneous lasso folding. Thus, simulations characterizing the folding propensity are needed to identify strategies for increasing access to the lasso architecture by stabilizing the pre-lasso ensemble before isopeptide bond formation. Herein, harmonic linear discriminant analysis (HLDA) is combined with metadynamics-enhanced sampling to discover CVs capable of distinguishing the pre-lasso fold and converging the folding propensity. Intuitive CVs are compared to iterative rounds of HLDA to identify CVs that not only accomplish these goals for one lasso peptide but also seem to be transferable to others, establishing a protocol for the identification of folding reaction coordinates for lasso peptides.

Graphical Abstract

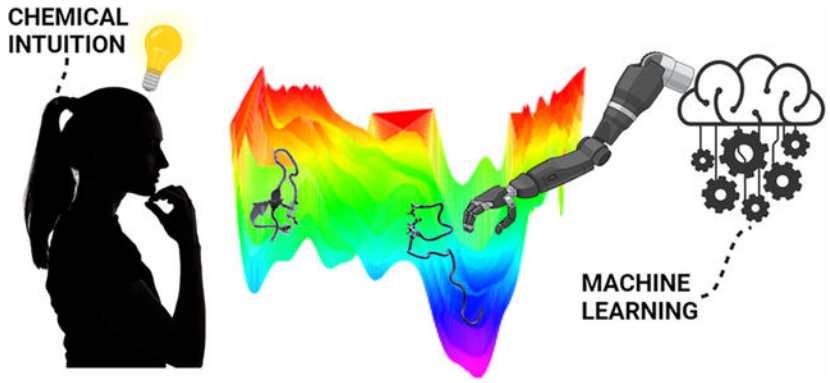
Corresponding Author: Jessica M. J. Swanson – Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States; j.swanson@utah.edu.
Present Address: Myongin Oh - Weill Cornell Medicine, 1300 York Ave, New York, New York 10065, United States

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.3c08492.
Distance and hydrogen bond analyses; HLDA analysis with hydrogen bonds; HLDA analysis with Cα distances; free energy surfaces; comparison between the simulation systems with different CVs; Table S1; and Figures S1–S10 (PDF)

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.3c08492

The authors declare no competing financial interest.



1. INTRODUCTION

Lasso peptides stand out as a captivating category of natural products. Classified as ribosomally synthesized and post-translationally modified peptides (RiPPs), these biologically active molecules exhibit a unique structural motif, where the C-terminal tail threads through a macrocyclic ring formed by N-terminal amino acids to form a knot-like tertiary structure (Figure 1A). This distinctive arrangement imparts extraordinary stability to lasso peptides, rendering them resistant to proteolytic degradation, thermal denaturation, and extreme pH conditions. Because of their significant biological activities, ranging from antibacterial and antiviral properties to enzyme inhibition and anticancer effects, they have gained substantial attention as potential candidates for therapeutic applications.[1,2]

In nature, enzymes enable the formation of lasso peptides via ATP-driven activation of a side-chain carboxylate and the formation of an isopeptide-amide bond to the N-terminus resulting in irreversible cyclization.[3–6] Thus, efforts to synthesize novel lasso peptides using enzymes through, for example, heterologous expression and cell-free biosynthesis are well underway.[7–14] A desired extension of our ability to design lasso-based therapeutics would be an adaptable chemical synthesis platform. Direct synthesis would enable greater flexibility, including the incorporation of non-native moieties and variants. However, synthesizing lasso peptides has proven challenging and consistently led to the formation of the unthreaded variants,[3,15–18] known as the tadpole conformation (Figure 1C). The perceived limitation is accessing the prefolded threaded structure, herein called pre-lasso (Figure 1D), as opposed to the unthreaded structure (pre-tadpole, Figure 1E) in the absence of a stabilizing enzyme.[19]

Recently, interest has grown in exploring whether the pre-lasso conformation can be achieved without an enzyme. Early work investigating the folding of the lasso peptide Microcin J25 (MccJ25) with simulations reported the rapid formation of a left-handed pre-lasso fold (defined by the N-terminus wrapping around the tail in a counterclockwise direction) (Figure 1B).[15] This was a surprising finding as all of the lasso peptides found in nature to date are right-handed. However, revisiting the folding of MccJ25 with updated force fields recently demonstrated that it indeed forms the expected right-handed pre-lasso motif.[20] The earlier observation of the left-handed pre-lasso conformation[15] was attributed to the use of the GROMOS96 43A2 force field,[21] which has since been refined to correct backbone dihedral parameters.[22–25] In addition, the recent simulations revealed that while

the pre-lasso conformation of MccJ25 is metastable, the peptide readily transitions to the entropically favored unfolded and pre-tadpole conformations.[20] This suggests that *de novo* pre-lasso folding is rare and that achieving the lasso motif synthetically will likely require enhanced stabilization of the pre-lasso conformation.

To pursue pre-lasso stabilization effectively, it is important to establish an efficient method for evaluating the relative stability of native and non-native lasso peptide sequences. Given the slow nature of *de novo* pre-lasso folding, enhanced sampling could be of great benefit. Although unbiased approaches, such as replica exchange[26,27] and weighted ensemble,[28] are valuable approaches for broad phase space exploration retaining real dynamics, biased approaches employing collective variables (CVs) can more directly target specific processes or relative probabilities between select ensembles. Thus, CV-based enhanced sampling, in which a bias potential is introduced along predefined reaction coordinates, will be a valuable tool in the assessment of the relative probability of forming the pre-lasso vs pre-tadpole intermediates. Assuming that the barrier for isopeptide bond formation is roughly equivalent and irreversible for these two intermediates, it is their relative probability that determines the likelihood of accessing the lasso vs tadpole final folds. However, the efficacy of CV-based methods heavily depends on the selection of the appropriate CVs. The selected variables must distinguish the relevant structural ensembles (pre-lasso, pre-tadpole, and unfolded in this case) and will ideally capture the slow modes of motion involved in the transition of interest. Poorly chosen CVs can result in slow convergence, insufficient exploration of relevant regions of the phase space, and misleading free energy profiles. Selecting optimal CVs, however, can be highly nontrivial, particularly for complex processes such as protein conformational changes and peptide folding in which many modes of motion are coupled and those most relevant are not readily apparent. For such processes, even differentiating relevant conformations using CVs can be difficult.

In our previous work, we identified parameters for discretizing the phase space of MccJ25.[20] However, an infrequent but non-negligible overlap in the pre-lasso and the pre-tadpole distributions was observed, rendering these parameters ineffective as CVs for distinguishing the two ensembles. Although we additionally developed a triangulation-based algorithm that definitively distinguishes the pre-lasso from the pre-tadpole structures, this algorithm lacks differentiable variables, rendering it incompatible with enhanced free energy sampling techniques. Thus, developing suitable CVs that can discern pre-lasso and pre-tadpole conformations remains an outstanding challenge.

In this study, we attempt to identify differentiable CVs that distinguish the pre-lasso and pre-tadpole ensembles to characterize the folding landscape of lasso peptides. For model peptides, we choose MccJ25,[16,20,29] which is known for its unique β-hairpin secondary structure, as well as two shorter peptides lacking secondary structures, ulleungdin (Uln)[30] and sungsanpin (Sun).[31] Once the isopeptide bonds are formed, each peptide retains the lasso scaffold due to a stopper residue (Tyr20, Tyr13, and Trp14 for MccJ25, Uln, and Sun, respectively) that prevents the tail from unthreading through the ring (Figure 2). We first compare the efficacy of intuitive CVs such as long-lived loop-forming hydrogen bonds and the isopeptide bond distance to machine-learning CVs based on harmonic linear discriminant analysis (HLDA). Although the identified intuitive CVs can access

the pre-lasso ensemble, they do not fully distinguish the pre-lasso from the pre-tadpole conformations. Turning to machine learning, we next employed multiple iterations of HLDA with a range of descriptors and an increasing number of structures in the analyzed ensembles. This process reveals that a linear combination of Cα distances between residues of macrocyclic ring and tail combined with the intuitively selected distance between the two isopeptide-forming residues is effective in characterizing the pre-lasso folding. When employed in well-tempered metadynamics (WTMetaD), this CV combination effectively distinguishes pre-lasso, pre-tadpole, and unfolded ensembles for MccJ25. We then test a potentially transferable protocol for reweighting the Cα ring–tail distances in the HLDA CV for Uln and Sun and verify that it also works for relatively different lasso peptide sequences. We close with a summary of this protocol and perspective on the combined value of intuitive and machine-learning CVs for characterizing the folding landscape of lasso peptides and guiding the design of sequences that stabilize the elusive pre-lasso motif.

2. METHODS

2.1. Well-Tempered Metadynamics.

Metadynamics (MetaD)[32,33] is a widely used enhanced free energy sampling technique that enables the exploration of physical processes occurring at time scales beyond the reach of conventional molecular dynamics (MD). MetaD employs a history-dependent bias in the form of a Gaussian function (eq 1), applied along predefined CVs. This bias modifies the system Hamiltonian periodically, pushing the system away from energy wells and facilitating the exploration of new regions of phase space.

$$V_{\mathrm{G}}(\boldsymbol{s},t)=\int_{t}^{0}\mathrm{d}t' w\exp\left[-\sum_{i=1}^{d}\frac{(\boldsymbol{s}_i(\boldsymbol{R})-\boldsymbol{s}_i(\boldsymbol{R}(t')))^2}{2\sigma_{\mathrm{i}}^2}\right] \tag{1}$$

In eq 1, w is the energy rate defined by the height of the Gaussian (w_0) divided by the frequency of deposition (τ), σ_i denotes the Gaussian width for the i-th CV, and $\boldsymbol{s}_i(\boldsymbol{R}(t'))$ is the value of the i-th CV at time t'. By summing the negative of the added MetaD bias, we can define the underlying free energy surface along the chosen CVs. However, in nontempered MetaD, a constant Gaussian height throughout the simulation leads to free energy calculations oscillating around the true values and, in extreme cases, destabilizing the system due to the addition of too much energy. To address this limitation, the WTMetaD method[34] was introduced. This variant incorporates a time-dependent bias potential, where the Gaussian height exponentially diminishes based on the local bias energy (eq 2). In eq 2, ΔT regulates the rate at which the Gaussian height decreases. This approach has been demonstrated to asymptotically converge the bias potential to a linearly scaled inverse of the free energy,[35] ensuring a more accurate and stable exploration of the system's energy landscape.

$$w(t) = w_0 \tau \exp\left[-\frac{V_G(s,t)}{k_B \Delta T}\right] \quad (2)$$

2.2. Collective Variables: Intuitive vs Machine Learning.

CVs can be valuable in the exploration and characterization of conformational phase space of biomolecules, especially for complex systems such as lasso peptides.[6,20] However, choosing appropriate CVs is not a trivial task, as different CVs may have different advantages and disadvantages depending on the system and the goal of the study. Herein, we explore two main approaches to CV selection: intuitive versus machine-learned.

Intuitive CVs are based on chemical intuition and prior knowledge of the system. They are usually selected based on some physical or geometrical insights into the process of interest. For example, ligand binding has an obvious relationship with the relative distance between the ligand and the receptor. For conformational changes, molecular properties, such as pair distances, bond or dihedral angles, native contacts, or hydrogen bonds are often tested. Intuitive CVs are easier to interpret and implement, and they can capture relevant features of the system that are known or expected from previous studies.[36–38] However, intuitive CVs are unlikely to be optimal. Having been biased by human assumptions, they generally miss influential degrees of freedom, making them less efficient for sampling high-dimensional spaces and less reflective of the thermodynamics and kinetics of the process of interest.[39,40]

In contrast, machine-learning CVs are based on data-driven methods that use machine-learning algorithms to extract relevant features from the input data.[41–43] These methods are usually derived from dimensionality reduction, clustering, or time correlation techniques. Examples include linear methods, such as principal component analysis (PCA),[44–46] linear discriminant analysis (LDA),[47,48] harmonic LDA (HLDA),[49,50] and time-lagged independent component analysis (tICA),[51,52] as well as a number of novel nonlinear methods and algorithms,[41,53] such as diffusion maps[39] and state-predictive information bottleneck (SPIB).[42,54] Machine-learning CVs can overcome some of the limitations of intuitive CVs, such as being biased by human assumptions, missing hidden and often nonlinear correlations, and ultimately being inefficient for sampling high-dimensional spaces.[52,55–58] However, machine-learning CVs may also have some drawbacks, such as being difficult to interpret or implement, requiring large amounts of input data, or being sensitive to noise and outliers.[58,59] In this work, we compare intuitive CVs for a process with seemingly clear geometrical constraints and machine-learning CVs, as obtained with HLDA. HLDA was selected because it aligns with our goal of distinguishing pre-lasso and pre-tadpole ensembles. Unlike PCA and tICA, which focus on preserving configurational variance and maximizing the kinetic content, respectively, HLDA maximizes the separation or distinction between conformational ensembles. This is crucial for our goal of quantifying the relative probabilities of these states. While the above-mentioned nonlinear deep learning methods may also be successful in this goal and should be explored in future studies, we chose a linear method for ease of implementation, direct interpretability, and computational efficiency.

2.3. HLDA.

Linear discriminant analysis[45] is a supervised dimensionality reduction method that projects the input data to a new subspace that maximizes the Fisher objective function J to find a decision boundary with maximal separability among classes

$$J(\boldsymbol{W}) = \frac{\boldsymbol{W}^T \boldsymbol{S}_{\mathrm{B}} \boldsymbol{W}}{\boldsymbol{W}^T \boldsymbol{S}_{\mathrm{W}} \boldsymbol{W}} \tag{3}$$

where $\boldsymbol{S}_{\mathrm{W}}$ is the within-class scatter matrix, $\boldsymbol{S}_{\mathrm{B}}$ is the between-class scatter matrix, and $\boldsymbol{W}$ is the transformation matrix that defines the subspace. The maximization of the objective function is achieved by solving the generalized eigenvalue problem $\boldsymbol{S}_W^{-1}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{W} = \lambda \boldsymbol{W}$. Although commonly used in classification problems, classical LDA poses a major limitation; that is, every pair of classes contributes equally to the between-class variance through the arithmetic mean and thus larger between-class distances dominate.[15] From a data analysis perspective, this seems a bit counterintuitive because classes with smaller variances need to be better defined. From a rare-event perspective, it could be more appropriate to give more weight to states with smaller fluctuations. To resolve this issue, Mendels et al.[49,50] proposed a different measure of the within-class scatter matrix based on the harmonic average as follows

$$\boldsymbol{S}_{\mathrm{W}} = \left(\sum_i \sum_i^{-1} \right)^{-1} \tag{4}$$

where $\sum_i$ is the class scatter matrix for metastable state i. If the system exhibits two metastable states A and B, this leads to the following expression for the collective variable s

$$s(\boldsymbol{R}) = (\boldsymbol{\mu}_{\mathrm{A}} - \boldsymbol{\mu}_{\mathrm{B}})^T \left(\boldsymbol{\Sigma}_{\mathrm{A}}^{-1} + \boldsymbol{\Sigma}_{\mathrm{B}}^{-1}\right) \boldsymbol{d}(\boldsymbol{R}) \tag{5}$$

where μ_i is the mean vector of class i, $\boldsymbol{d}$ is the descriptor vector, and $\boldsymbol{R}$ is the atomic coordinates. In HLDA metadynamics, the system is encouraged to move from one metastable state A to the other B in a perpendicular direction $\boldsymbol{W}$ to the decision boundary that separates the metastable states[18] (Figure 3).

2.4. Simulation Protocol.

The study of the *de novo* folding of the lasso peptide MccJ25 was conducted using MD and the WTMetaD protocol, employing the latest all-atom AMBER ff19SB force field.[60] Starting from the completely linear configurations that were built using the AmberTools20 *LEaP*,[61] the systems were solvated with the OPC water model[62] in periodic octahedron boxes, maintaining a minimum distance of 10 Å between the solute and the box edge. Neutralization of the MccJ25 system was achieved by adding one Na^+ ion, while the other

two peptides were already neutral at pH 7. The systems were then minimized using the steepest descent method, initially with restraints on the protein's heavy atoms (10 kcal mol^{-1} $Å^2$ harmonic positional restraint for 1000 steps), and then gradually releasing these restraints without SHAKE.[63,64]

The systems next underwent a comprehensive equilibration process consisting of eight steps, in which restraints on the backbone atoms were systematically reduced. The SHAKE algorithm[63,64] was introduced to constrain hydrogens in the second step. The first four steps used the NVT ensemble, while the last four switched to NPT. The first five steps were 5 ps each with a time step of 1 fs. The final three stages, each lasting for 10 ps with a time step of 2 fs, decreased the force constant for backbone restraints from 1.0 to 0.5 kcal mol^{-1} $Å^{-2}$ to no positional restraints. The temperature was maintained at 300 K in each simulation using the weak-coupling algorithm with a relaxation time constant of 1 ps. Simultaneously, isotropic pressure scaling was applied to maintain a constant pressure in the NPT simulations, with a pressure relaxation time of 1 ps.[65] The production phase involved multiple replicas for each combination of CVs for each peptide, with an average of 132 ns of biased simulation time per replica, for a total simulation time of ~4.3 μs. The time step was 2 fs. The simulations were conducted until the bias height reached a negligible value, indicating that further simulation time would not contribute significantly to the accumulation of the deposited bias in the system. The Langevin thermostat[66] was employed to maintain the system temperature at 300 K, with a collision frequency of 1 ps^{-1} and a time constant of 1 ps. The pressure was maintained at 1 bar using the Berendsen barostat,[65] with a pressure relaxation time of 1 ps. The particle mesh Ewald (PME) method[67] was employed for long-range electrostatic determinations with an 8 Å nonbonded cutoff. For the treatment of van der Waals interactions, a cutoff of 8 Å was used, consistent with the PME real space cutoff. Coordinates and energies were recorded every 5000 steps.

All WTMetaD simulations were conducted using Amber 2020[61] patched with PLUMED 2.7.1.[68,69] Our previously developed triangulation-based algorithm LATCHED[20] was used to distinguish between pre-lasso and pre-tadpole configurations. Briefly, this algorithm definitively determines if the tail is piercing the ring by defining triangles between each of the two consecutive Cα on residues 1–8 and the tail piercing the Cα and then testing the vectors along the tail to determine if they fall inside any of the triangles. A chirality algorithm additionally distinguishes left-handedness from right-handedness. The code can be accessed at https://github.com/gabedahora/LATCHED. Analyses were performed using the CPPTRAJ program,[70] with 2D free energy surfaces obtained using custom Python scripts. Visualization of structures and trajectories, as well as image generation, was performed using VMD 1.9.3.[71] All of the data and PLUMED input files required to reproduce the results reported in this paper are available on PLUMED-NEST (www.plumed-nest.org), the public repository of the PLUMED consortium, as plumID:23.046.

3. RESULTS AND DISCUSSION

3.1. Intuitive CVs.

In our previous study, we discretized the conformational phase space of MccJ25 using a summation of the ring–tail distances combined with the distance between the isopeptide

bond-forming residues, Gly1 and Glu8.[20] However, these descriptors were unable to fully distinguish the pre-lasso (Figure 1D) from the pre-tadpole (Figure 1E) conformation. Here, our focus is to converge and accurately quantify the folding probability of pre-lasso peptides. To achieve this, we aim to identify and employ optimal CVs that fulfill specific criteria: first, they must clearly distinguish pre-lasso and pre-tadpole conformations; second, they must enable efficient exploration of the phase space; and third, they must be differentiable, rendering them suitable for enhanced free energy sampling techniques.

In many cases, CVs selected based on chemical intuition can effectively describe the process of interest and enable efficient sampling, at least along the selected coordinates.[36,72–76] Additionally, the physical meaning of intuitive CVs is typically more interpretable than that of machine-learning CVs, making the resulting free energy surfaces easier to understand. Given the clear geometrical requirements of the pre-lasso motif, one might expect that intuitive CVs should work well for lasso peptides. Thus, we first identified and tested many intuitive CVs (ring–tail distances, angles, contacts, etc.) and ultimately selected the two that best distinguished the pre-lasso from the pre-tadpole conformations and seemed to capture important degrees of freedom of the system (Figure 4). The first is motivated by the fact that to form a cyclized lasso or tadpole conformation (Figure 1), the two groups involved in the isopeptide bond must come into close proximity. Thus, a seemingly essential descriptor, as used in our previous work,[20] is the distance between the isopeptide bond-forming atoms (Gly1N and Glu8Cδ in MccJ25). For the second CV, we sought an interaction that captures the formation of another distinct structural feature: the loop of the lasso (Figure 1A). The loop corresponds to the region of a lasso peptide that is not threaded through the macrocyclic ring. We selected the backbone–backbone hydrogen bond between the macrocyclic ring and the first residue of the lasso tail below the ring because this interaction represents the closure of the loop. For MccJ25, this descriptor entails the backbone hydrogen bond between Val6 and Tyr20 (hereafter referred to as 6–20), which was described by a coordination number with a cutoff of 3 Å between the donor and acceptor atoms. Another motivation for employing 6–20 as a CV is that the interaction was observed to be consistent within the pre-lasso metastable ensemble in unbiased simulations initiated from the NMR structure of MccJ25 (PDB: 1Q71)[16] from our previous work[20] (Figure S1). Once this interaction is lost, the peptide begins to unravel and unfold. This observation suggests that 6–20 is potentially a slow mode involved in the pre-lasso unfolding, making it an ideal choice for CV. Lastly, it is also the most probable hydrogen bond between the ring and ring-piercing tail residues (Figure S2).

Utilizing the chosen CVs (Gly1N–Glu8Cδ and 6–20), WTMetaD simulations were initiated from the fully stretched linear structure. Although this conformation is unphysical, it challenges the CVs to find the folded ensemble starting from an unbiased state. Four replicas were run for an approximate duration of 130 ns each. Two replicas were able to form pre-lassos in this time, with one demonstrating the desired diffusive motion between unfolded, pre-tadpole, and pre-lasso states. Interestingly, these CVs more commonly formed left-handed pre-lasso conformations than the right-handed pre-lassos observed in nature. Unfortunately, despite the apparent success of achieving diffusive transitions between unfolded and folded states, the 6–20 and 1–8 CVs failed to distinguish between pre-lasso and pre-tadpole conformations on the free energy surface. This was verified with the

LATCHED triangulation algorithm (see Methods),[20] revealing that both conformations (Figure S3) co-occupy two distinct basins in the free energy landscape (Figure 5). Thus, despite the selected intuitive CVs performing better than all others tested, they do not properly distinguish the key conformational states and cannot be used to quantify the relative probability of forming the pre-lasso and pre-tadpole intermediates. New CVs must be identified. Although intuitive CVs may indeed exist that would effectively accomplish these goals, we next turned to machine learning.

3.2. HLDA with Hydrogen Bond Features.

To identify key descriptors that might not have been included in our intuitive CVs, we applied HLDA initially to a set of 81 hydrogen bonds to distinguish the pre-lasso and pre-tadpole ensembles obtained from unbiased simulations of our previous work.[20] The simulation length used to feed the HLDA was 7.2 μs in total, and there were 22,568 filtered structures, including 2,825 pre-lassos and 19,742 pre-tadpoles.[20] The hydrogen bonds were identified using the *hbond* tool from CPPTRAJ[70] on the ensemble of pre-lasso structures obtained from the unbiased simulations in our previous study.[20] By iteratively selecting the dominant weights in the HLDA analysis (further explained in Supporting Information), three hydrogen bonds were selected as a linear combination to define our second CV for enhanced sampling (Figure 6). These hydrogen bonds include Tyr9–Ile7 (*hb*2), which seems to capture the wrapping of the N-terminus around the tail to form the macrocyclic ring, as well as Ser18-Glu8 (*hb*1) and Val6–Tyr20 (*hb*3), which capture loop formation, with the latter supporting the selection of our intuitive CV described above. Additionally, we retained the distance between Gly1 and Glu8 as the first CV to capture the formation of the isopeptide bond.

$$S_{\mathrm{HLDA}} = 0.79178*hb1 + 0.32043*hb2 + 0.52001*hb3$$

Three replicas of WTMetaD were run with these CVs starting from the linear structure. Again, two replicas demonstrated efficient sampling of pre-lasso, pre-tadpole, and unfolded distributions. Yet, again, an overlap between the pre-lasso and pre-tadpole conformations was observed in the basins ((3 Å < CV1 < 6 Å) and (CV2 < 10)) of the resulting free energy landscape (Figure 7A), indicating that the separation of the two ensembles was not achieved. Interestingly, this CV combination led to the formation of new pre-tadpole structures that resemble pre-lasso conformations due to the presence of a β-hairpin loop (Figure 7B). As these new pre-tadpole structures were not included in the HLDA, it is possible that the resulting CV was unable to distinguish them from the pre-lasso ensemble. In the spirit of iterative CV discovery, wherein CVs enable the exploration of new regions of phase space and thus warrant new CV definitions, we attempted adding the novel pre-tadpole conformations along with newly sampled pre-lasso structures to the ensembles and reanalyzing our HLDA weights. We expanded the number of hydrogen bonds to account for those in the expanded pre-lasso ensemble and obtained another separation in HLDA space (Figure S4).

3.3. HLDA with Ring–Tail Features.

Motivated by our previous work in which a summation of ring–tail Cα distances was reasonably effective in discretizing the phase space of MccJ25 using LATCHED,[20] we revisited HLDA using eight Cα distances between residues of the macrocyclic ring (1–8) and the ring-penetrating Phe19 as new features (Figures 8 and 9). These distances intuitively capture the N-terminus wrapping around the tail, a clear geometric feature in the formation of a lassoed structure. All pre-lasso and pre-tadpole conformations obtained thus far in the simulations described above were used in the new HLDA analysis.

$$S_{\mathrm{HLDA}} = -0.3221d_1 - 0.0361d_2 - 0.1063d_3 - 0.1538d_4 - 0.0913d_5 - 0.5176d_6 - 0.6589d_7 - 0.3867d_8$$

The implementation of this new linear combination as a CV, in conjunction with the distance between Gly1 and Glu8 in WTMetaD, enabled both efficient sampling of the phase space by visiting the pre-lasso and pre-tadpole basins and complete separation of the pre-tadpole and pre-tadpole ensembles (Figures 10 and S5). From five replicas employing different combinations of WTMetaD parameters (height: 1.2, 2.4, and 5 kJ; and bias factor: 8, 15, 40, and 80) four were able to obtain pre-lasso and pre-tadpole structures. Moreover, the pre-lasso structures showed up early in the simulation (between 6.4 and 13.4 ns) and in a good amount after multiple visits to the basin (505 structures). Importantly, unlike the previous approach where a summation of Cα distances (1, 3, 6, and 8–19) did not distinguish between the pre-lasso and pre-tadpole conformations,[20] the use of HLDA to obtain refined weights for the eight distances now provides clear differentiation (Figure 10). This includes the “impostor” pre-lasso region found in basin 9 (Figures 10 and S5–H). The replica with lower height and bias factor values failed to form a pre-lasso structure, indicating that the weak and slow biasing impeded effective steering toward the desired states.

Investigating the relative stabilities of different conformational basins provides additional insights into the pre-lasso and pre-tadpole states. Three distinct pre-lasso basins were obtained. The most stable, basin 1, is characterized by CVs within the range of 2.5–3.0 Å for CV1 (distance 1N-8Cδ) and from −12 to −13 for CV2 (HLDA CV) (Figure 11A) and exhibits a minimum energy of −39 kcal/mol. Conversely, basins 2 and 3 representing looser pre-lasso conformations with longer 1N-8Cδ distances (Figure S5) display higher minimum energies of −34 and −28 kcal/mol, respectively. Notably, the pre-tadpole state (Figure 11B) in basin 4 with a broader region defined by CV1 3.5–3.8 Å and CV2 −32 to −45, boasts the lowest energy of −45 kcal/mol. Between the lowest energies of pre-lasso and pre-tadpole basins, energy barriers are on the order of 10 kcal/mol. Integrating over each of these basins, we quantify the relative probability of forming pre-lasso and pre-tadpole structures. The total normalized probability (population) of basin 1 (tighter pre-lasso state) is 2.03×10^{-3}%, while that of basin 4 (pre-tadpole state) is significantly higher at 2.65×10^{-1}%, making the free energy difference going from pre-tadpole to pre-lasso ($\Delta G = G_{\text{pre-lasso}} - G_{\text{pre-tadpole}}$) = 7.02 kcal/mol, with an error of ±0.02 kcal/mol calculated by bias-corrected bootstrapping.[77] The total probabilities (populations) of basins 2 and 3, representing looser pre-lasso conformations, are 4.76×10^{-9} and 2.10×10^{-9}%, respectively. The free energy differences to basin 4 are 10.63 ± 0.24 kcal/mol from basin 2 and 11.12

± 0.01 kcal/mol from basin 3. These results indicate that the pre-tadpole state (basin 4) is significantly more likely to form than the pre-lasso states (basins 1, 2, and 3). The energy difference is also consistent with our previous estimation of the relative entropy favoring the pre-tadpole ensemble and with the experimental challenges in isolating lasso peptides to date.[4,7,11,18,20,78]

Lastly, we expanded our investigation into the combination of intuitive and machine-learning CVs by testing 6–20 combined with HLDA descriptors. This set of CVs was able to form pre-lasso and pre-tadpole structures, although less efficiently (after 33 ns to obtain a pre-lasso conformation) and with less frequency, indicating longer simulation times needed for convergence (Figure S6 and Table S1). Thus, the combination of the isopeptide bond distance (1N-8Cδ) and the ring–tail HLDA CV was deemed most effective.

3.4. Transferability.

With CVs for MccJ25 identified, we next explored the portability of these CV combinations to other lasso peptides. While MccJ25 has been extensively studied due to its promising biological activities, it is also the only known lasso peptide to contain a β-hairpin. This secondary structure gives MccJ25 additional stability and likely unique folding dynamics. Thus, evaluating the portability of our identified CVs is essential for the development of general-purpose CVs for lasso peptide folding.

Uln and Sun are shorter lasso peptides (15 instead of the 21 residues in MccJ25) that have been reported to inhibit cancer cellular migration.[30,31] Neither peptide has apparent secondary structural features like a β-hairpin, and both adopt the lasso scaffold with the C-terminal tail threading through the macrocyclic ring formed by the N-terminus and a side-chain carboxylate (Figure 2).

Thus, the distance between the two isopeptide bond-forming residues, Gly1N and Asp8γ (1N–8γ), in both Uln and Sun, was chosen as the first CV. To assess the possibility that an intuitive CV works for these shorter sequences, we first revisited the equivalent of the 6–20 interaction by identifying the dominant hydrogen bond formed between a ring residue (residues 1–8) and a piercing tail residue (residues 12–13). As detailed in the SI, this CV combination again faced challenges in distinguishing between pre-lassos and pre-tadpoles in the free energy surface (Figure S9), verifying the limitations of the 6–20 equivalents for other lasso peptides.

In contrast, employing the machine-learned HLDA CV based on Cα combined with the 1N–8γ distance was again effective in characterizing the folding landscape of these shorter lasso peptides. In Uln, all five replicas were able to promote the formation of pre-lasso structures (Table S1). As for Sun, three of the five replicas were able to capture a pre-lasso state, generating an average of 442 pre-lasso structures. Similar to MccJ25, this CV combination was also more efficient, forming pre-lassos earlier in the simulation (27.4 ns for Uln and 5.9 ns for Sun) than the 6–20 equiv (36.2 and 69.3 ns, respectively) (Table S1). The resulting free energy profiles revealed distinct and well-separated basins corresponding to the pre-lasso, pre-tadpole, and unfolded states (Figure 12).

For Uln (Figures 12A and S10A), the pre-lasso state (basin 1) was characterized by 2.6 < 1N–8γ CV < 3.2 Å and −14 < HLDA CV < −16, with an associated minimum free energy of −15 kcal/mol. The most stable pre-tadpole state (basin 2) exhibited a range of 2.6 < 1N–8γ CV < 3.3 Å and −48 < HLDA CV < −53, with a minimum free energy of −19 kcal/mol. Another pre-tadpole state (basin 3) featured 2.6 < 1N–8γ CV < 3.2 Å and −21 < HLDA CV < −35, also with a minimum free energy of −15 kcal/mol. The barrier between these states was approximately 11 kcal/mol, indicating slow transitions between distinct pre-tadpole states. By integrating the wells of basin 1 (pre-lasso) and basin 2 (most stable pre-tadpole), we were able to calculate that the ΔG between these two states was 4.24 ± 0.01 kcal/mol. The total probability (population) of basin 1 was found to be approximately 4.54×10^{-4}%, while that of basin 2 was approximately 5.56×10^{-1}%. When considering basins 3 and B (two different pre-tadpole basins), we found that the ΔG between these two states was 3.29 ± 0.01 kcal/mol. The probability for basin 3 was 2.22×10^{-3}%, indicating that it is less likely to form than the most stable pre-tadpole state, but still significantly more likely than the pre-lasso state. This indicates a clear thermodynamic preference for the pre-tadpole states, consistent with an anticipated challenge in obtaining pre-lassos through the *de novo* folding of native sequences.

For Sun (Figure 12B and S10B), the pre-lasso state (basin 1) was defined by 3.1 < 1N–8γ CV < 3.4 Å and −14 < HLDA CV < −16, with a minimum free energy of −20.1 kcal/mol. The lowest pre-tadpole state (basin 2) exhibited a range of 3.0 < 1N–8γ CV < 3.5 Å and −28 < HLDA CV < −33, with a minimum free energy of −24.1 kcal/mol. Another pre-tadpole state (basin 3) featured 3.1 < 1N–8γ CV < 3.4 Å and −39 < HLDA CV < −43, with a minimum free energy of −23.5 kcal/mol. A third pre-tadpole state (basin 4) was defined by 3.1 < 1N–8γ CV < 3.4 Å and −49 < HLDA CV < −52, with a minimum free energy of −23 kcal/mol. Integrating over the lowest energy basins for pre-lasso (basin 1) and pre-tadpole (basin 2), the total probability of the pre-lasso state is approximately 5.26×10^{-4}%, while that of the pre-tadpole state is significantly higher at 4.33×10^{-2}%. The free energy difference going from pre-tadpole to pre-lasso was ΔG = 2.63 ± 0.03 kcal/mol. Comparing the lowest-energy pre-tadpole group (basin 2) with the other pre-tadpole ensemble located in basin 3, we obtained ΔG = 0.27 ± 0.02 kcal/mol. The probability for the less likely pre-tadpole basins 3–4 was 2.74×10^{-2}%. Comparing across peptides, the tested lasso sequences share several features in their folding landscapes, namely that the formation of the pre-lasso intermediate is less probable and that multiple metastable basins describe the pre-tadpole, and sometimes pre-lasso, ensembles. Collectively, these results demonstrate the portability of the above combination of intuitive (1N-8Cx distance) and machine-learned (Cα HLDA) CVs to two distinct lasso peptides from MccJ25. The distinguishability of states and the efficiency of transitions suggest that this CV combination may provide a universal tool for characterizing diverse lasso peptide folding landscapes.

3.5. Study Limitations.

Although the efficacy of the described CVs on such different lasso peptide sequences is encouraging, their transferability will have to be further tested on new sequences and verified using algorithms such as LATCHED,[20] which definitively distinguish pre-lasso from pre-tadpole conformations. Additionally, experimental analyses testing the importance

of ring–tail interactions and the relative probabilities of forming the lasso and tadpole variants will be important contributions to the field. The improbable nature of pre-lasso folding with the sequences explored herein is consistent with failed attempts to access these peptides in the lasso form experimentally.[15,17,18] As increased stability is achieved, however, experimental comparison with our predicted relative probabilities will be essential to validate the efficacy of these CVs. Lastly, although the goal of the present work was to distinguish folding between the pre-lasso and pre-tadpole ensembles, future work including the unfolded ensemble in the CV-discovery process will be important for a more complete understanding of the lasso folding process, including CVs that may contribute to kinetic limitations.

4. CONCLUSIONS

In this work, chemical intuition and machine learning are combined to identify reaction coordinates for the relative folding propensity of lasso peptides. Direct synthesis is highly sought after for lasso peptides due to the remarkable stability of the knotted scaffold and their size, which occurs between those of small-molecule drugs and biologics. However, synthesis efforts have yet to be successful.[7–14] The expected challenge is that of folding a peptide sequence into a lasso-like form (pre-lasso) before the formation of the isopeptide bond that creates the lasso ring. Thus, methods are needed that can guide the design of lasso sequences, including non-native chemical modifications, with increased pre-lasso stability.[20] This work takes an important step toward establishing such methods by identifying CVs that can both distinguish the desired pre-lassos from the undesired pre-tadpole motifs and efficiently sample the relative probability of reaching these two ensembles *de novo*.

To identify effective CVs, many chemically intuitive variables were first explored. Despite our expectation that the geometric constraints of tying a short peptide into a knot would make intuitive CVs obvious, this was not the case. After testing multiple geometric variables that appear definitive in knot formation, the most promising intuitive CVs (the Val6–Tyr20 loop–tail hydrogen bond and the Gly1N–Glu8δ isopeptide bond distance) failed to clearly distinguish between pre-lasso and pre-tadpole structures. Turning to machine learning, multiple rounds of employing the HLDA first revealed that hydrogen bonds alone, even when optimally weighted, were still insufficient to distinguish between pre-lasso and pre-tadpole ensembles. Thus, alternative features were needed. The features that worked were ring–tail Cα distances, which curiously had previously shown promise for discretizing the lasso folding landscape but also failed to distinguish the two ensembles when in the form of an evenly weighted linear combination.[20] HLDA revealed that once properly weighted, these ring–tail Cα distances combined with the intuitively selected Gly1N–Glu8δ isopeptide bond distance effectively distinguish the pre-lasso and pre-tadpole ensembles. Moreover, when used in WTMetaD enhanced sampling, these CVs enable the convergence of the relative probability of forming the pre-lasso and pre-tadpole intermediates.

As these CVs were identified using only one lasso peptide sequence, MccJ25, it is unclear how useful they will be for lasso peptide design, wherein many sequences must be compared. Thus, we next tested the same CVs on two other lasso sequences, Uln and Sun, both shorter than MccJ25 and distinct in the absence of a β-hairpin loop. First, HLDA was

used to reweight the new set of ring–tail Cα distances. With this single step, which can be performed based on short, unbiased simulations, the same CVs again distinguished and efficiently sampled the pre-lasso and pre-tadpole ensembles. Although this protocol must be tested for new sequences (and verified with the LATCHED tools),[20] these results suggest that the combination of ring–tail and isopeptide bond distances may indeed be universal CVs for studying diverse lasso peptide sequences. Such CVs will be a valuable tool to support recent advances in the computational prediction and design of lasso peptides such as LassoHTP.[79] Looking ahead, our findings contribute to the ongoing debate surrounding intuitive CVs versus machine-learning approaches, suggesting that, in some cases, an amalgamation of the two may be optimal. They also lay the groundwork for more targeted investigations into the diverse world of lasso peptides and their intricate folding pathways, underscoring the importance of stabilizing ring–tail interactions in our quest to fold the elusive pre-lasso motif.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors express gratitude for the support received from NIH NIGMS (R35GM143117) and computational resources provided by Bridges-2 at the Pittsburgh Supercomputing Center through the ACCESS program (allocation MCB200018) supported by NSF (grants #2138259, #2138286, #2138307, #2137603, and #2138296), as well as the Center for High-Performance Computing (CHPC) at the University of Utah. These resources played a crucial role in the successful completion of this research. Special acknowledgment goes to Marcus C. Mifflin and Andrew G. Roberts for their insightful contributions to discussions, enhancing the depth of this study. TOC was created with BioRender.com and an image by Freepik.

REFERENCES

(1). Maksimov MO; Pan SJ; James Link A Lasso peptides: structure, function, biosynthesis, and engineering. Nat. Prod. Rep 2012, 29 (9), 996–1006. [PubMed: 22833149]

(2). Cheng C; Hua ZC Lasso Peptides: Heterologous Production and Potential Medical Application. Front. Bioeng. Biotechnol 2020, 8, No. 571165. [PubMed: 33117783]

(3). Martin-Gómez H; Tulla-Puche J Lasso peptides: chemical approaches and structural elucidation. Org. Biomol Chem 2018, 16 (28), 5065–5080. [PubMed: 29956707]

(4). Hegemann JD Factors Governing the Thermal Stability of Lasso Peptides. ChemBioChem 2020, 21 (1–2), 7–18. [PubMed: 31243865]

(5). Duquesne S; Destoumieux-Garzon D; Zirah S; Goulard C; Peduzzi J; Rebuffat S Two enzymes catalyze the maturation of a lasso peptide in *Escherichia coli*. Chem. Biol 2007, 14 (7), 793–803. [PubMed: 17656316]

(6). Hegemann JD; Zimmermann M; Xie X; Marahiel MA Lasso peptides: an intriguing class of bacterial natural products. Acc. Chem. Res 2015, 48 (7), 1909–1919. [PubMed: 26079760]

(7). Kretsch AM; Gadgil MG; DiCaprio AJ; Barrett SE; Kille BL; Si Y; Zhu L; Mitchell DA Peptidase Activation by a Leader Peptide-Bound RiPP Recognition Element. Biochemistry 2023, 62 (4), 956–967. [PubMed: 36734655]

(8). Si Y; Kretsch AM; Daigh LM; Burk MJ; Mitchell DA Cell-Free Biosynthesis to Evaluate Lasso Peptide Formation and Enzyme-Substrate Tolerance. J. Am. Chem. Soc 2021, 143 (15), 5917–5927. [PubMed: 33823110]

(9). Liu T; Ma X; Yu J; Yang W; Wang G; Wang Z; Ge Y; Song J; Han H; Zhang W; et al. Rational generation of lasso peptides based on biosynthetic gene mutations and site-selective chemical modifications. Chem. Sci 2021, 12 (37), 12353–12364. [PubMed: 34603665]

(10). Mevaere J; Goulard C; Schneider O; Sekurova ON; Ma H; Zirah S; Afonso C; Rebuffat S; Zotchev SB; Li Y An orthogonal system for heterologous expression of actinobacterial lasso peptides in Streptomyces hosts. Sci. Rep 2018, 8 (1), No. 8232. [PubMed: 29844351]

(11). DiCaprio AJ; Firouzbakht A; Hudson GA; Mitchell DA Enzymatic Reconstitution and Biosynthetic Investigation of the Lasso Peptide Fusilassin. J. Am. Chem. Soc 2019, 141 (1), 290–297. [PubMed: 30589265]

(12). Zhu S; Fage CD; Hegemann JD; Mielcarek A; Yan D; Linne U; Marahiel MA The B1 Protein Guides the Biosynthesis of a Lasso Peptide. Sci. Rep 2016, 6 (1), No. 35604. [PubMed: 27752134]

(13). Wang M; Fage CD; He Y; Mi J; Yang Y; Li F; An X; Fan H; Song L; Zhu S; Tong Y Recent Advances and Perspectives on Expanding the Chemical Diversity of Lasso Peptides. Front. Bioeng. Biotechnol 2021, 9, No. 741364. [PubMed: 34631682]

(14). Nakashima Y; Kawakami A; Ogasawara Y; Maeki M; Tokeshi M; Dairi T; Morita H Structure of lasso peptide epimerase MslH reveals metal-dependent acid/base catalytic mechanism. Nat. Commun 2023, 14 (1), No. 4752. [PubMed: 37550286]

(15). Ferguson AL; Zhang S; Dikiy I; Panagiotopoulos AZ; Debenedetti PG; James Link A An experimental and computational investigation of spontaneous lasso formation in microcin J25. Biophys. J 2010, 99 (9), 3056–3065. [PubMed: 21044604]

(16). Rosengren KJ; Clark RJ; Daly NL; Goransson U; Jones A; Craik DJ Microcin J25 has a threaded sidechain-to-backbone ring structure and not a head-to-tail cyclized backbone. J. Am. Chem. Soc 2003, 125 (41), 12464–12474. [PubMed: 14531690]

(17). Lear S; Munshi T; Hudson AS; Hatton C; Clardy J; Mosely JA; Bull TJ; Sit CS; Cobb SL Total chemical synthesis of lassomycin and lassomycin-amide. Org. Biomol. Chem 2016, 14 (19), 4534–4541. [PubMed: 27101411]

(18). Waliczek M; Wierzbicka M; Arkuszewski M; Kijewska M; Jaremko L; Rajagopal P; Szczepski K; Sroczynska A; Jaremko M; Stefanowicz P Attempting to synthesize lasso peptides using high pressure. PLoS One 2020, 15 (6), No. e0234901. [PubMed: 32579565]

(19). Zhu S; Fage CD; Hegemann JD; Mielcarek A; Yan D; Linne U; Marahiel MA The B1 Protein Guides the Biosynthesis of a Lasso Peptide. Sci. Rep 2016, 6, No. 35604. [PubMed: 27752134]

(20). da Hora GCA; Oh M; Mifflin MC; Digal L; Roberts AG; Swanson JMJ Lasso Peptides: Exploring the Folding Landscape of Nature's Smallest Interlocked Motifs. J. Am. Chem. Soc 2024, 146, 4444. [PubMed: 38166378]

(21). Schuler LD; van Gunsteren WF On the choice of dihedral angle potential energy functions for n-alkanes. Mol. Simul 2000, 25 (5), 301–319.

(22). Schuler LD; Daura X; Van Gunsteren WF An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. J. Comput. Chem 2001, 22 (11), 1205–1218.

(23). Soares TA; Daura X; Oostenbrink C; Smith LJ; van Gunsteren WF Validation of the GROMOS force-field parameter set 45Alpha3 against nuclear magnetic resonance data of hen egg lysozyme. J. Biomol. NMR 2004, 30 (4), 407–422. [PubMed: 15630561]

(24). Oostenbrink C; Soares TA; van der Vegt NF; van Gunsteren WF Validation of the 53A6 GROMOS force field. Eur. Biophys J 2005, 34 (4), 273–284. [PubMed: 15803330]

(25). Schmid N; Eichenberger AP; Choutko A; Riniker S; Winger M; Mark AE; van Gunsteren WF Definition and testing of the GROMOS force-field versions 54A7 and 54B7. Eur. Biophys. J 2011, 40 (7), 843–856. [PubMed: 21533652]

(26). Mitsutake A; Sugita Y; Okamoto Y Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers 2001, 60 (2), 96–123. [PubMed: 11455545]

(27). Hansmann UHE Parallel tempering algorithm for conformational studies of biological molecules. Chem. Phys. Lett 1997, 281 (1–3), 140–150.

(28). Zuckerman DM; Chong LT Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. Annu. Rev. Biophys 2017, 46 (1), 43–57. [PubMed: 28301772]

(29). Salomón RA; Farias RN Microcin 25, a novel antimicrobial peptide produced by *Escherichia coli*. J. Bacteriol 1992, 174 (22), 7428–7435. [PubMed: 1429464]

(30). Son S; Jang M; Lee B; Hong YS; Ko SK; Jang JH; Ahn JS Ulleungdin, a Lasso Peptide with Cancer Cell Migration Inhibitory Activity Discovered by the Genome Mining Approach. J. Nat. Prod 2018, 81 (10), 2205–2211. [PubMed: 30251851]

(31). Um S; Kim YJ; Kwon H; Wen H; Kim SH; Kwon HC; Park S; Shin J; Oh DC Sungsanpin, a lasso peptide from a deep-sea streptomycete. J. Nat. Prod 2013, 76 (5), 873–879. [PubMed: 23662937]

(32). Laio A; Gervasio FL Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. Rep. Prog. Phys 2008, 71 (12), No. 126601.

(33). Laio A; Parrinello M Escaping free-energy minima. Proc. Natl. Acad. Sci. U.S.A 2002, 99 (20), 12562–12566. [PubMed: 12271136]

(34). Barducci A; Bussi G; Parrinello M Well-tempered metadynamics: a smoothly converging and tunable free-energy method. Phys. Rev. Lett 2008, 100 (2), No. 020603. [PubMed: 18232845]

(35). Dama JF; Parrinello M; Voth GA Well-tempered metadynamics converges asymptotically. Phys. Rev. Lett 2014, 112 (24), No. 240602. [PubMed: 24996077]

(36). Branduardi G. B. a. D. Free-Energy Calculations with Metadynamics: Theory and Practice; Wiley, 2015; pp 1–49 DOI: 10.1002/9781118889886.ch1.

(37). Yang YI; Shao Q; Zhang J; Yang L; Gao YQ Enhanced sampling in molecular dynamics. J. Chem. Phys 2019, 151 (7), No. 070902. [PubMed: 31438687]

(38). Hummer G; Kevrekidis IG Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. J. Chem. Phys 2003, 118 (23), 10762–10773.

(39). Ferguson AL; Panagiotopoulos AZ; Kevrekidis IG; Debenedetti PG Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. Chem. Phys. Lett 2011, 509 (1–3), 1–11.

(40). Aydin F; Durumeric AEP; da Hora GCA; Nguyen JDM; Oh MI; Swanson JMJ Improving the accuracy and convergence of drug permeation simulations via machine-learned collective variables. J. Chem. Phys 2021, 155 (4), No. 045101. [PubMed: 34340389]

(41). Noé F; De Fabritiis G; Clementi C Machine learning for protein folding and dynamics. Curr. Opin. Struct. Biol 2020, 60, 77–84. [PubMed: 31881449]

(42). Mehdi S; Wang D; Pant S; Tiwary P Accelerating All-Atom Simulations and Gaining Mechanistic Understanding of Biophysical Systems through State Predictive Information Bottleneck. J. Chem. Theory Comput 2022, 18 (5), 3231–3238. [PubMed: 35384668]

(43). Ravindra P; Smith Z; Tiwary P Automatic mutual information noise omission (AMINO): generating order parameters for molecular systems. Mol. Syst. Des. Eng 2020, 5, 339–348, DOI: 10.1039/C9ME00115H.

(44). Amadei A; Linssen AB; Berendsen HJ Essential dynamics of proteins. Proteins 1993, 17 (4), 412–425. [PubMed: 8108382]

(45). Pearson KL III. On lines and planes of closest fit to systems of points in space. London, Edinburgh Dublin Philos. Mag. J. Sci 1901, 2 (11), 559–572.

(46). Ichiye T; Karplus M Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. Proteins 1991, 11 (3), 205–217. [PubMed: 1749773]

(47). Uyar A; Karamyan VT; Dickson A Long-Range Changes in Neurolysin Dynamics Upon Inhibitor Binding. J. Chem. Theory Comput 2018, 14 (1), 444–452. [PubMed: 29179556]

(48). Sakuraba S; Kono H Spotting the difference in molecular dynamics simulations of biomolecules. J. Chem. Phys 2016, 145 (7), No. 074116. [PubMed: 27544096]

(49). Mendels D; Piccini G; Parrinello M Collective Variables from Local Fluctuations. J. Phys. Chem. Lett 2018, 9 (11), 2776–2781. [PubMed: 29733652]

(50). Piccini G; Mendels D; Parrinello M Metadynamics with Discriminants: A Tool for Understanding Chemistry. J. Chem. Theory Comput 2018, 14 (10), 5040–5044. [PubMed: 30222350]

(51). Oh M; da Hora GCA; Swanson JMJ tICA-Metadynamics for Identifying Slow Dynamics in Membrane Permeation. J. Chem. Theory Comput 2023, 19 (23), 8886–8900. [PubMed: 37943658]

(52). Wang Y; Lamim Ribeiro JM; Tiwary P Machine learning approaches for analyzing and enhancing molecular dynamics simulations. Curr. Opin. Struct. Biol 2020, 61, 139–145. [PubMed: 31972477]

(53). Stamati H; Clementi C; Kavraki LE Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. Proteins 2010, 78 (2), 223–235. [PubMed: 19731366]

(54). Wang D; Tiwary P State predictive information bottleneck. J. Chem. Phys 2021, 154 (13), No. 134111. [PubMed: 33832235]

(55). Tsai ST; Fields E; Xu Y; Kuo EJ; Tiwary P Path sampling of recurrent neural networks by incorporating known physics. Nat. Commun 2022, 13 (1), No. 7231. [PubMed: 36433982]

(56). Pant S; Smith Z; Wang Y; Tajkhorshid E; Tiwary P Confronting pitfalls of AI-augmented molecular dynamics using statistical physics. J. Chem. Phys 2020, 153 (23), No. 234118. [PubMed: 33353347]

(57). Bonati L; Rizzi V; Parrinello M Data-Driven Collective Variables for Enhanced Sampling. J. Phys. Chem. Lett 2020, 11 (8), 2998–3004. [PubMed: 32239945]

(58). Sidky H; Chen W; Ferguson AL Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. Mol. Phys 2020, 118 (5), No. e1737742.

(59). Noé F; Tkatchenko A; Muller KR; Clementi C Machine Learning for Molecular Simulation. Annu. Rev. Phys. Chem 2020, 71 (1), 361–390. [PubMed: 32092281]

(60). Tian C; Kasavajhala K; Belfon KAA; Raguette L; Huang H; Migues AN; Bickel J; Wang Y; Pincay J; Wu Q; Simmerling C ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. J. Chem. Theory Comput 2020, 16 (1), 528–552. [PubMed: 31714766]

(61). Case DA; Belfon KAA; Ben-Shalom IY; Brozell SR; Cerutti DS; Cheatham TE III.; Cruzeiro VWD; Darden TA; Duke RE; Giambasu G; Gilson MK; Gohlke H; Goetz AW; Harris R; Izadi S; Izmailov SA; Kasavajhala K; Kovalenko A; Krasny R; Kurtzman T; Lee TS; LeGrand S; Li P; Lin C; Liu J; Luchko T; Luo R; Man V; Merz KM; Miao Y; Mikhailovskii O; Monard G; Nguyen H; Onufriev A; Pan F; Pantano S; Qi R; Roe DR; Roitberg A; Sagui C; Schott-Verdugo S; Shen J; Simmerling CL; Skrynnikov NR; Smith JM; Swails J; Walker RC; Wang J; Wilson L; Wolf RM; Wu X; Xiong Y; Xue Y; York DM; Zhao S; Zhu Q; Kollman PA AMBER 2020; University of California: San Francisco, 2020.

(62). Izadi S; Anandakrishnan R; Onufriev AV Building Water Models: A Different Approach. J. Phys. Chem. Lett 2014, 5 (21), 3863–3871. [PubMed: 25400877]

(63). Ryckaert J-P; Ciccotti G; Berendsen HJC 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.

(64). Forester TR; Smith W SHAKE, rattle, and roll: Efficient constraint algorithms for linked rigid bodies. J. Comput. Chem 1998, 19 (1), 102–111.

(65). Berendsen HJC; Postma JPM; Vangunsteren WF; Dinola A; Haak JR Molecular-Dynamics with Coupling to an External Bath. J. Chem. Phys 1984, 81 (8), 3684–3690.

(66). Loncharich RJ; Brooks BR; Pastor RW Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. Biopolymers 1992, 32 (5), 523–535. [PubMed: 1515543]

(67). Darden T; York D; Pedersen L Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. J. Chem. Phys 1993, 98 (12), 10089–10092.

(68). consortium P Promoting transparency and reproducibility in enhanced molecular simulations. Nat. Methods 2019, 16 (8), 670–673. [PubMed: 31363226]

(69). Tribello GA; Bonomi M; Branduardi D; Camilloni C; Bussi G PLUMED 2: New feathers for an old bird. Comput. Phys. Commun 2014, 185 (2), 604–613.

(70). Roe DR; Cheatham TE 3rd PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J. Chem. Theory Comput 2013, 9 (7), 3084–3095. [PubMed: 26583988]

(71). Humphrey W; Dalke A; Schulten K VMD: visual molecular dynamics. J. Mol. Graphics 1996, 14 (1), 33–38 27–38.

(72). Noé F; Clementi C Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. Curr. Opin. Struct. Biol 2017, 43, 141–147. [PubMed: 28327454]

(73). Chen M Collective variable-based enhanced sampling and machine learning. Eur. Phys. J. B 2021, 94 (10), 211. [PubMed: 34697536]

(74). Sun R; Dama JF; Tan JS; Rose JP; Voth GA Transition-Tempered Metadynamics Is a Promising Tool for Studying the Permeation of Drug-like Molecules through Membranes. J. Chem. Theory Comput 2016, 12 (10), 5157–5169. [PubMed: 27598403]

(75). Granata D; Camilloni C; Vendruscolo M; Laio A Characterization of the free-energy landscapes of proteins by NMR-guided metadynamics. Proc. Natl. Acad. Sci. U.S.A 2013, 110 (17), 6817–6822. [PubMed: 23572592]

(76). Vymětal J; Vondrasek J Metadynamics as a tool for mapping the conformational and free-energy space of peptides–the alanine dipeptide case study. J. Phys. Chem. B 2010, 114 (16), 5632–5642. [PubMed: 20361773]

(77). Efron B Better Bootstrap Confidence Intervals. J. Am. Stat. Assoc 1987, 82 (397), 171–185.

(78). Yang Z; Hajlasz N; Kulik HJ Computational Modeling of Conformer Stability in Benenodin-1, a Thermally Actuated Lasso Peptide Switch. J. Phys. Chem. B 2022, 126 (18), 3398–3406. [PubMed: 35481742]

(79). Juarez RJ; Jiang Y; Tremblay M; Shao Q; Link AJ; Yang ZJ LassoHTP: A High-Throughput Computational Tool for Lasso Peptide Structure Construction and Modeling. J. Chem. Inf. Model 2023, 63 (2), 522–530. [PubMed: 36594886]

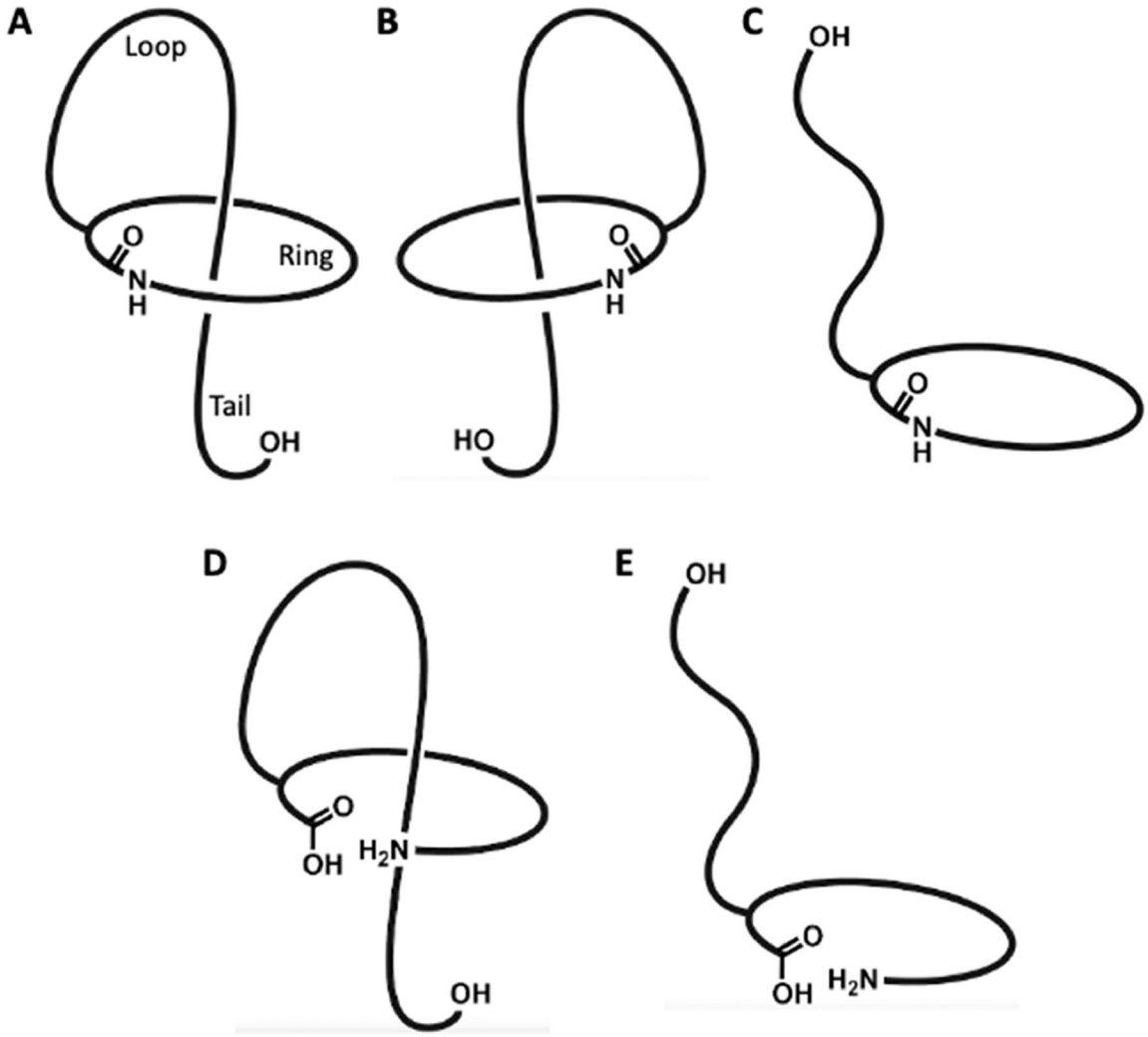


Figure 1.

Following the formation of the isopeptide bond, a lasso peptide can adopt right-handed lasso (A), left-handed lasso (B), or tadpole (C) conformations. The acyclic conformations before isopeptide bond formation are herein referred to as pre-lasso (D) and pre-tadpole (E).

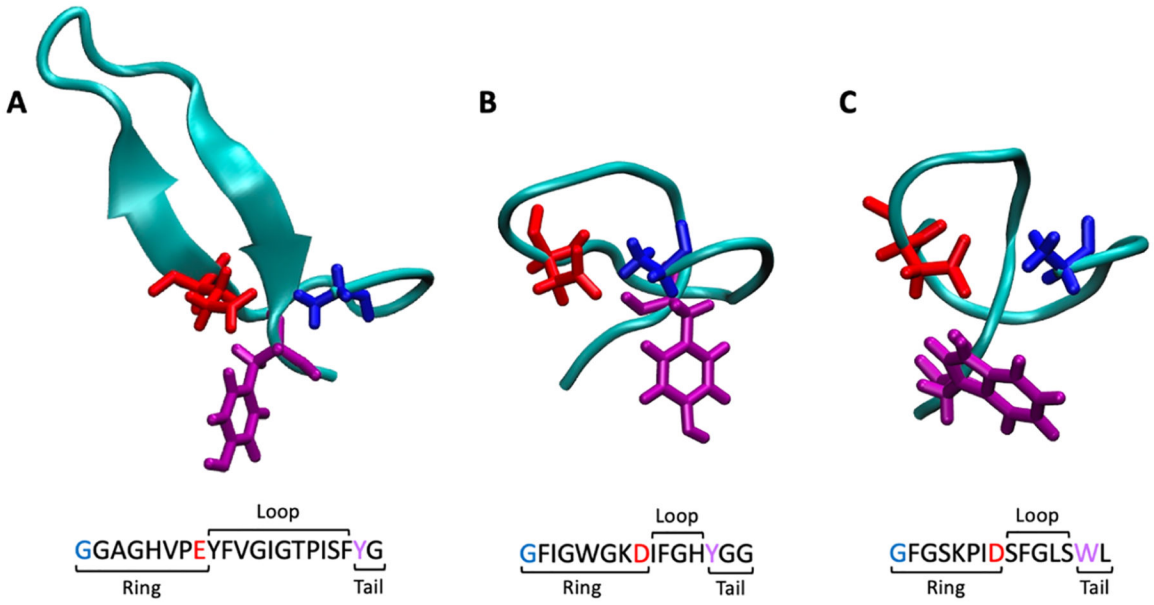


Figure 2.

Pre-lasso structures and sequences of (A) MccJ25, (B) Uln, and (C) Sun. The first and eighth residues that form a ring via isopeptide bond formation are colored blue and red, respectively. The bulky stopper residues that prevent the tail from unthreading are shown in purple.

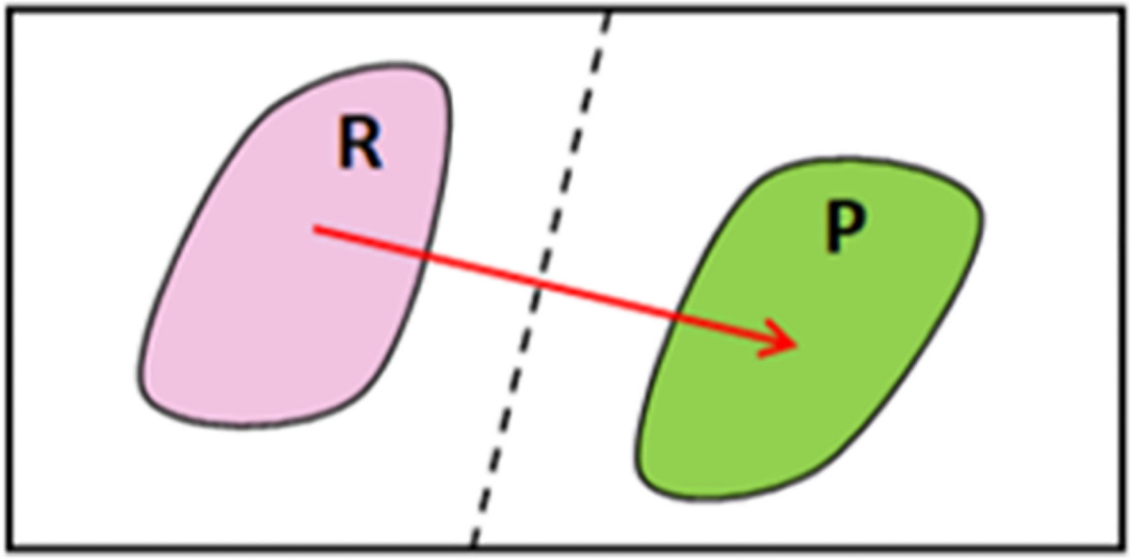


Figure 3.
HLDA decision boundary. Fluctuations from the two metastable states, reactant R (pink) and product P (green), which are separated by a decision boundary (dotted line). The red arrow represents the HLDA subspace onto which the input data is projected.

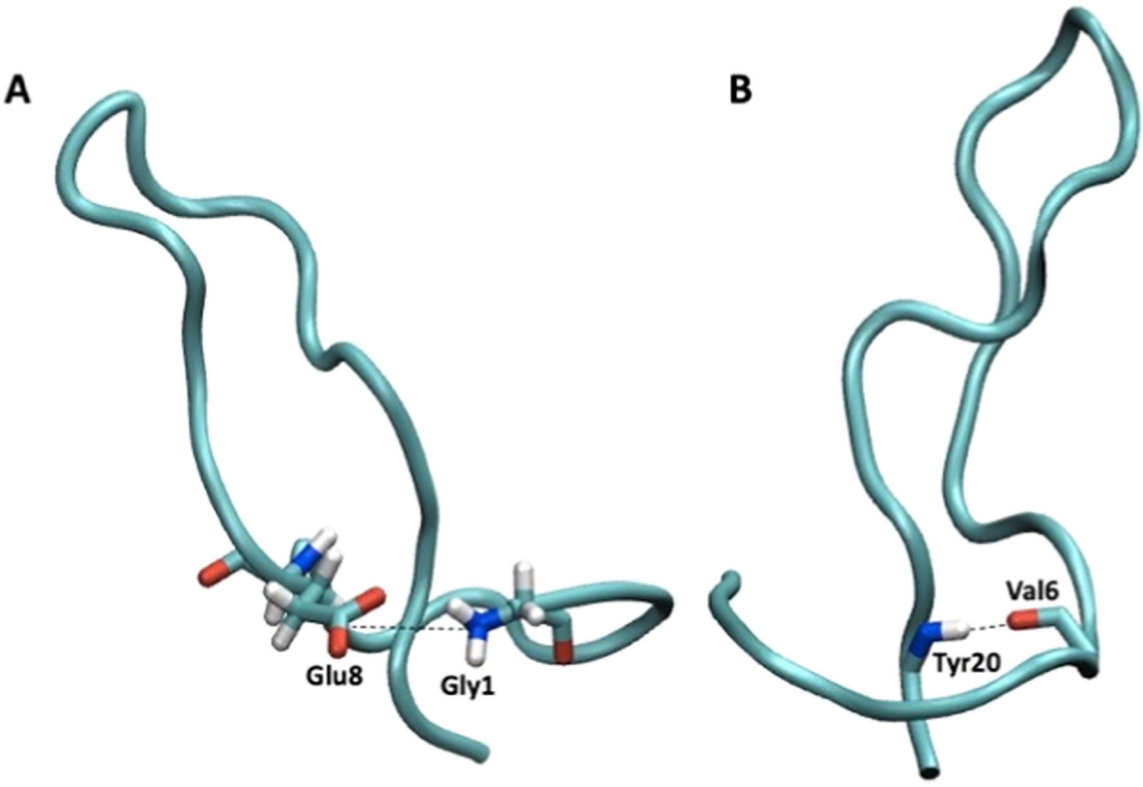


Figure 4.

Intuitive CVs: (A) distance between isopeptide-forming atoms (Gly1N–Glu8δ) and (B) the backbone hydrogen bond between Val6 and Tyr20 (6–20). The backbone is shown as tubes, while Gly1, Glu8, Val6, and Tyr20 are represented by licorice sticks and colored by atom type: C atoms in cyan ribbon, O in red, N in blue, and H in white.

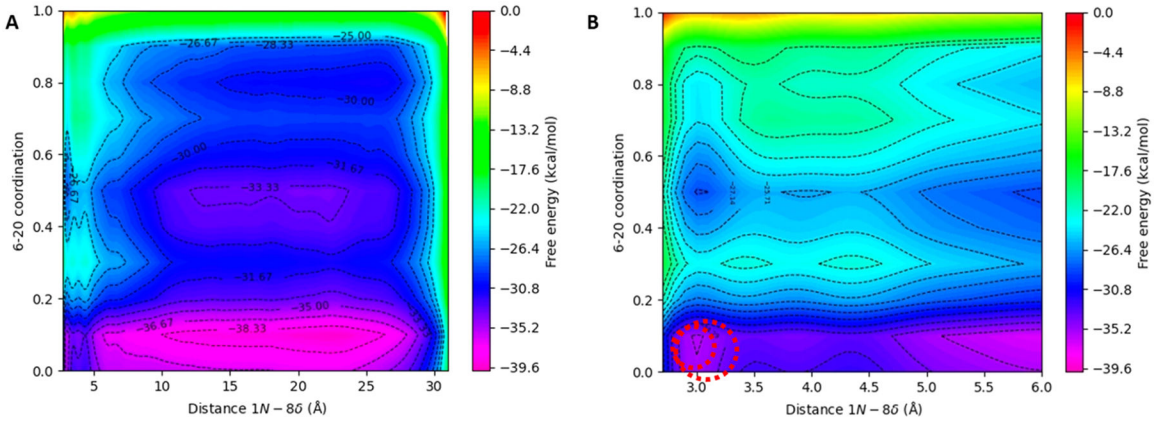


Figure 5.

Lasso folding free energy landscapes as a function of 1N-8Cδ and 6–20 showing (A) the entire region of the sampled phase space and (B) the region where pre-lassos and pre-tadpoles exist (distance 1N-8Cδ < 6 Å). The two circled basins in the region of 1N-8Cδ < 4 Å contain both pre-lassos and pre-tadpoles.

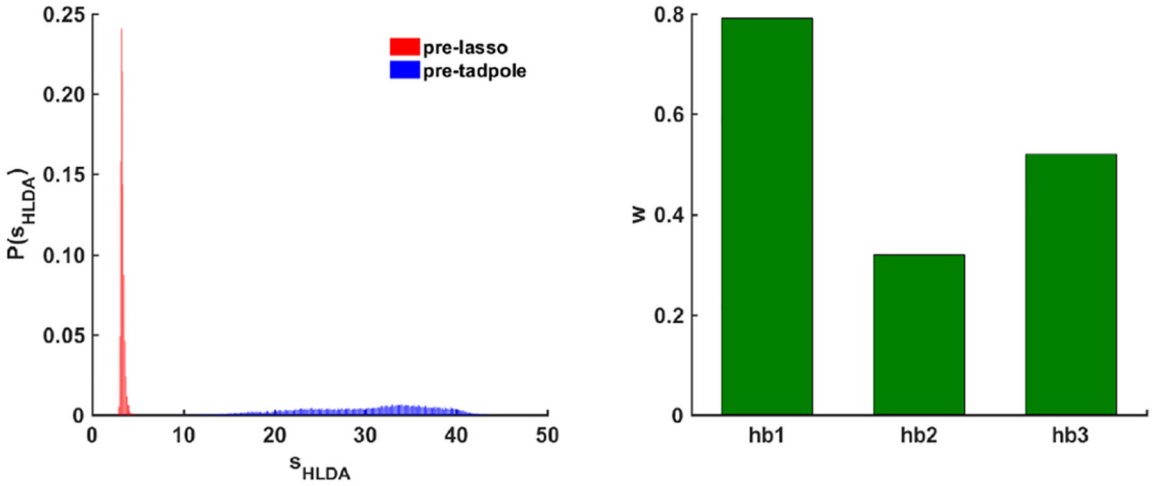


Figure 6.

HLDA analysis of pre-lasso and pre-tadpole states and the corresponding equation with hydrogen bond weights. The probability density functions (left) of pre-lasso (red) and pre-tadpole (blue) structures occupy distinct regions of the HLDA space. The HLDA weights for three selected hydrogen bonds (*hb*1, *hb*2, and *hb*3) are shown by a bar graph (right) consistent with the HLDA CV, S_{HLDA}.

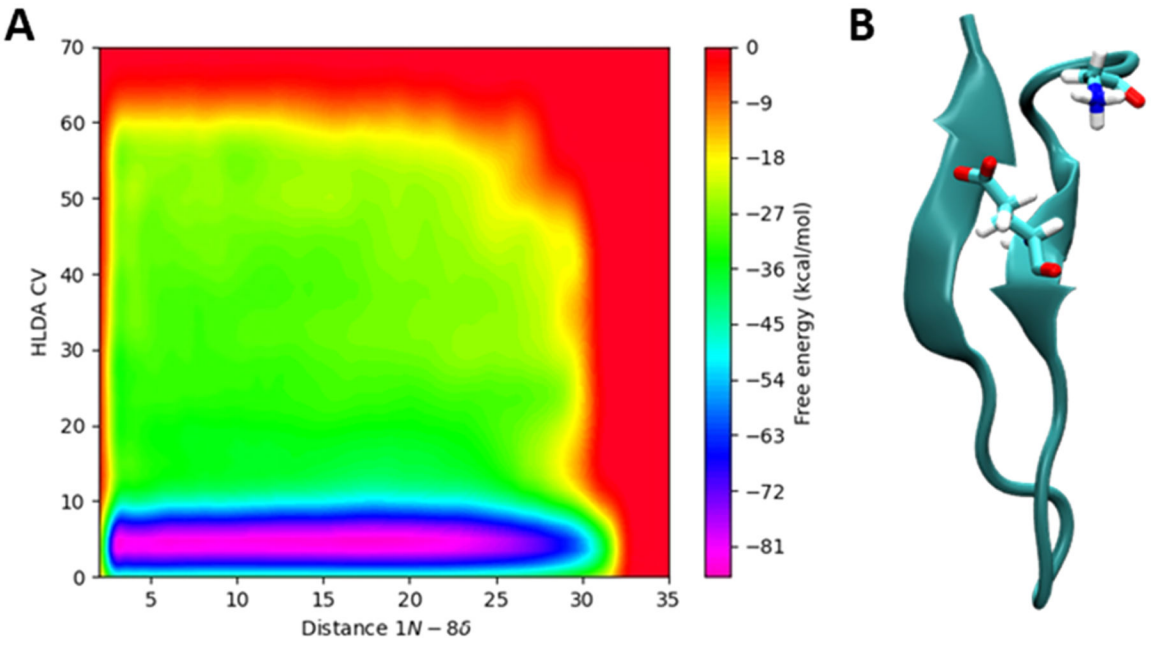


Figure 7.

(A) Lasso folding free energy landscape as a function of 1N-8Cδ and the h-bond HLDA CV focusing on the region where the pre-lassos and pre-tadpoles exist (distance 1N-8Cδ < 6 Å). Again, pre-lassos and pre-tadpoles are found in the same basin. (B) A new pre-tadpole structure that is more similar to the pre-lasso structure, retaining the β-hairpin loop.

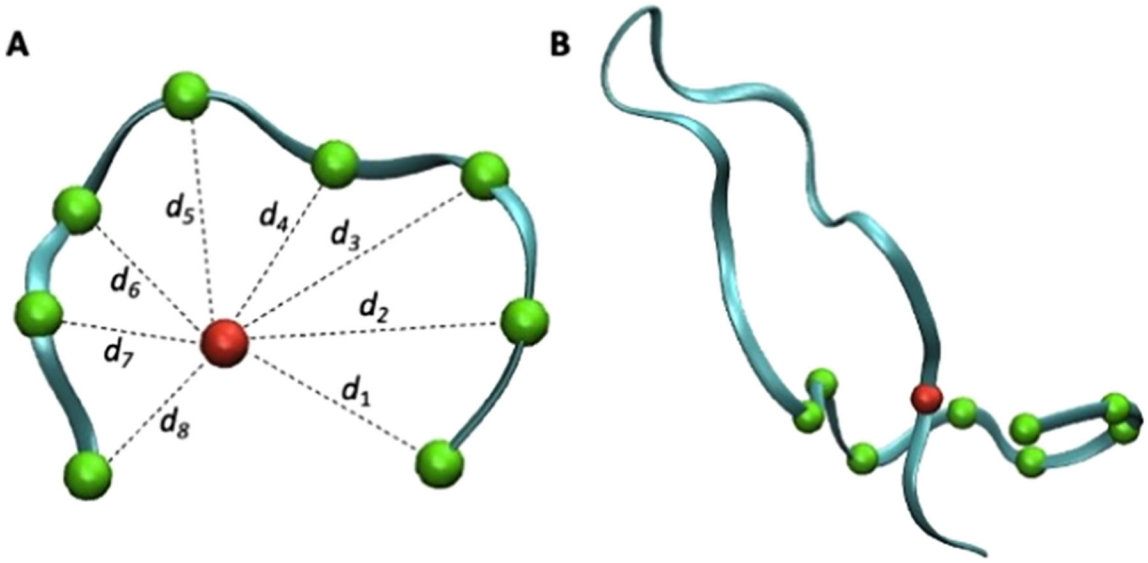


Figure 8.

Cα distances between residues of the ring and Phe19 of MccJ25 viewed from a (A) top–down perspective and (B) side-on perspective. The Cα of the ring and Phe19 are depicted as green and red spheres, respectively, while the protein backbone is shown as a cyan ribbon. In a previous study,[20] the descriptor used only the summation of d_1, d_3, d_6, and d_8.

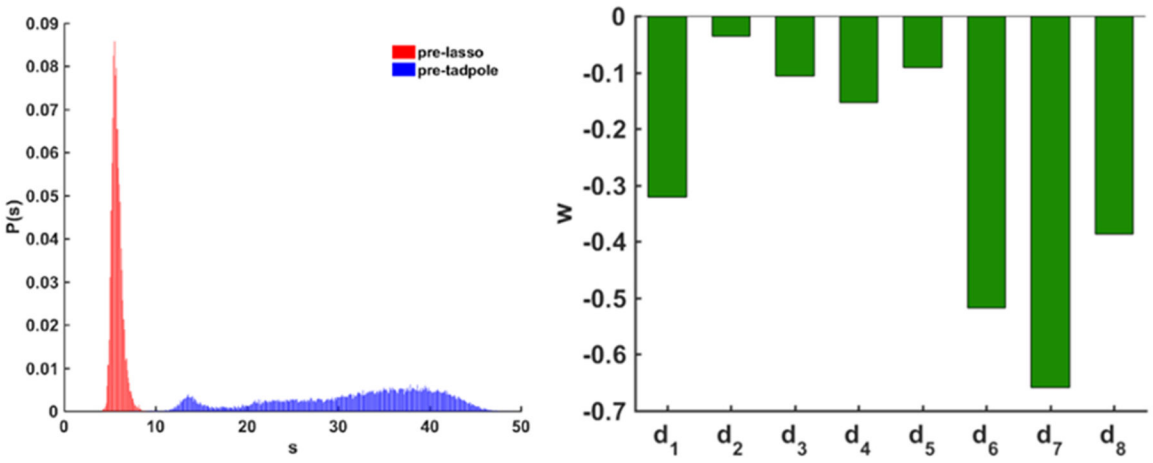


Figure 9.
Revised HLDA analysis of the pre-lasso and pre-tadpole states and the corresponding equation with distance weights.

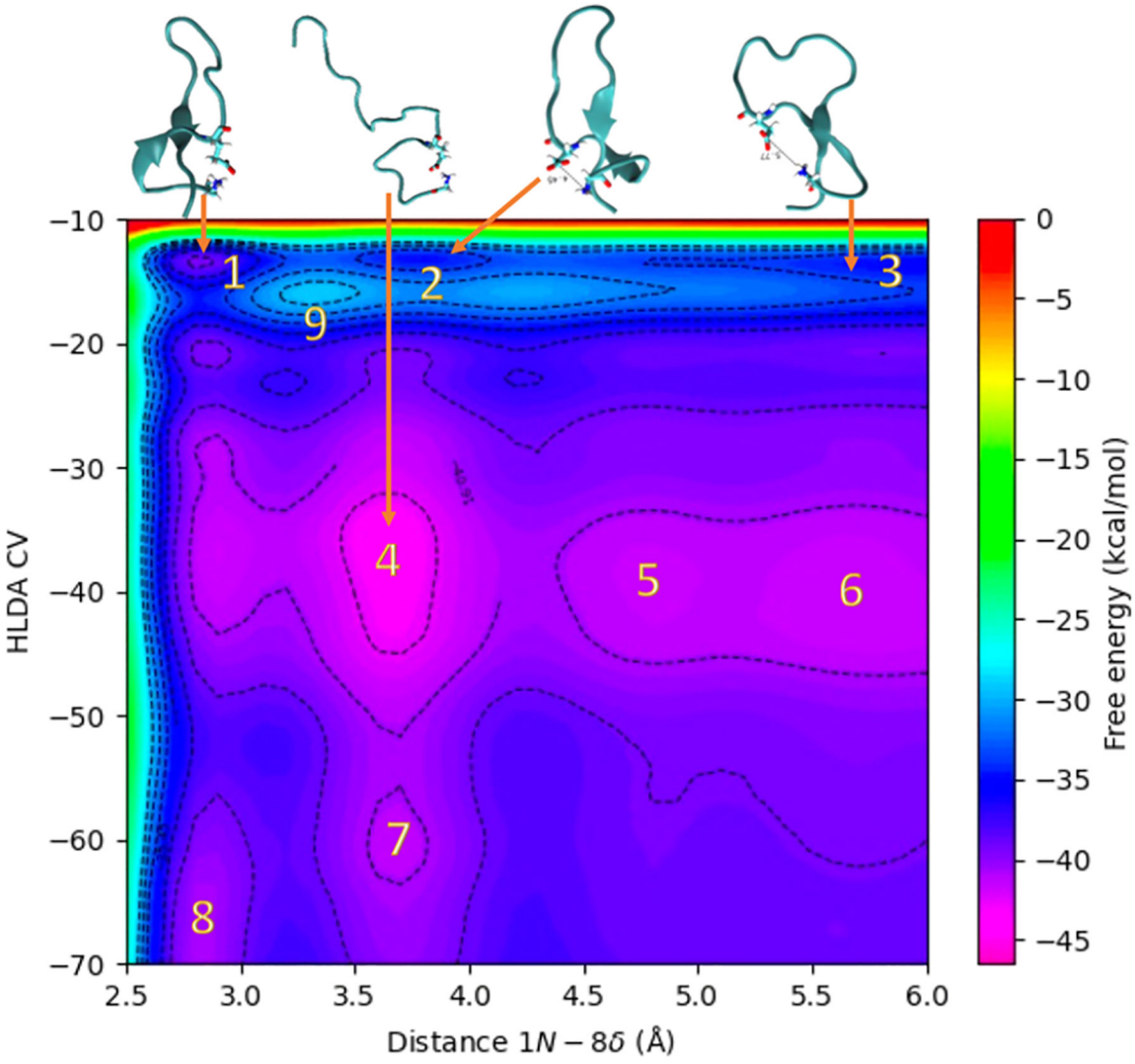


Figure 10.

Lasso folding free energy landscape as a function of 1N-8Cδ and the ring–tail HLDA CV focusing on the region where pre-lassos and pre-tadpoles exist (distance 1N-8Cδ < 6 Å). This time, pre-lassos and pre-tadpoles occupy different regions, as highlighted by basins 1–9. Basins 5–9 are different pre-tadpoles represented in Figure S5D–H, respectively. Peptide backbones are represented as cyan ribbons, and residues 1 and 8 are shown as sticks, with C atoms in cyan, O in red, N in blue, and H in white.

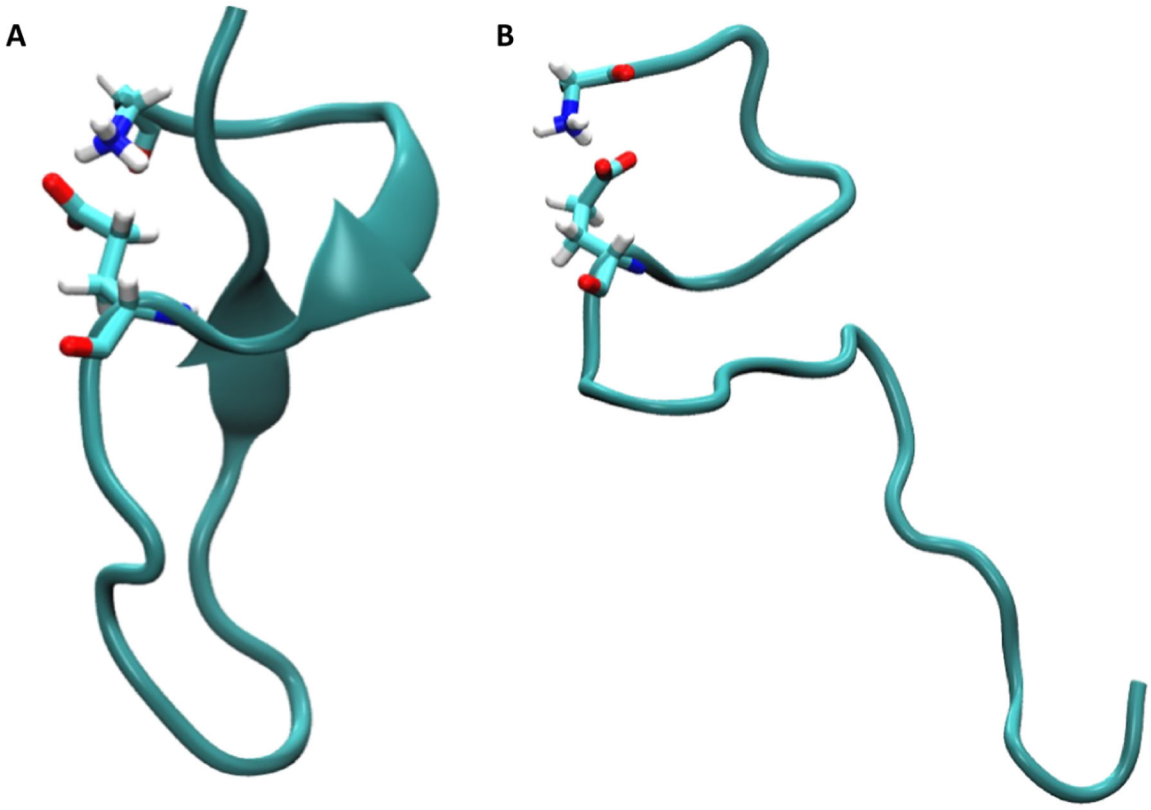


Figure 11.

Structures of peptides found in (A) basin 1 (pre-lassos) and (B) basin 2 (pre-tadpoles). Other relevant structures can be visualized in the SI Figure S5B–H. Peptide backbones are represented as cyan ribbons while residues 1 and 8 are shown in stick with C atoms in cyan, O in red, N in blue, and H in white.

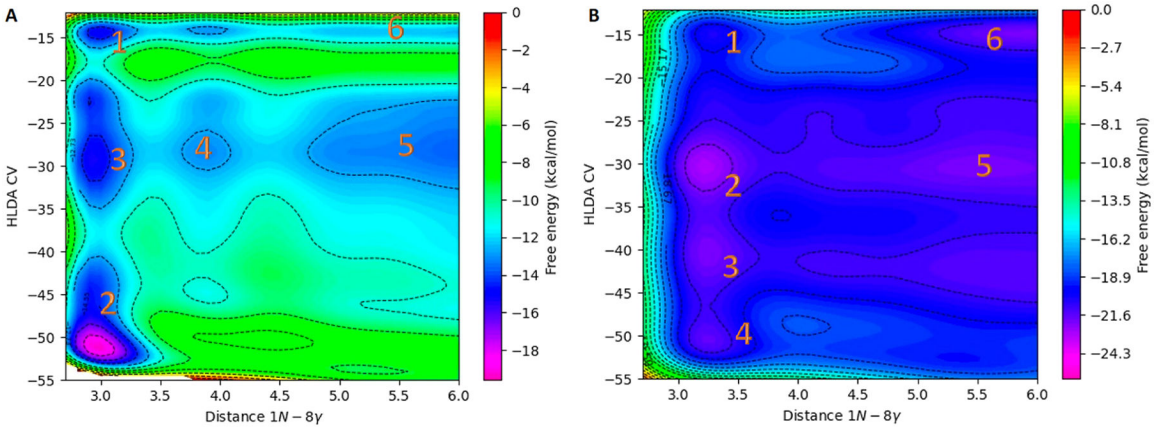


Figure 12.

Lasso folding free energy landscape as a function of 1N-8C g and the ring–tail HLDA CV of (A) Uln and (B) Sun. Basins 1 and 6 highlight the regions where the pre-lasso states are located, while basins 2–5 are regions of pre-tadpole states. Figure S10 shows the conformations for each basin.