Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Revealing pH-dependent antimicrobial peptide, GL13K, characteristics: A constant pH molecular dynamics study†

M.N. Hamidabad,[a] and R.A. Mansbach*[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Antimicrobial peptides (AMPs) are a promising potential solution to combat antimicrobial resistance (AMR) due to their positive charges, which induce selective interaction with negatively charged bacterial membranes. Certain AMPs exhibit a low hemolytic index and non-forming fibrils at physiological pH (7.4), enhancing their therapeutic potential. In this study, we employ constant pH molecular dynamics (CpHMD) simulations to investigate the pH-dependent behavior of a 13-residue-long positively charged AMP, GL13K, focusing on the deprotonation states of lysine residues in a single GL13K AMP and their impact on its structural dynamics. We show that the last lysine located near the C-terminus (LYS11) has a significant deprotonation ratio difference with other lysine residues at the reported average experimental side chain pKa value. We observe that increasing the pH results in eventual decrease of the radius of gyration and end-to-end radius and dominance of random coil structure in metastable states. Overall, our study shows the pH-dependent conformational dynamics and pK_a variations of lysine residues in the GL13K antimicrobial peptide, providing critical insights into its structural behavior in solution. These findings establish a foundation for further exploration of GL13K in multi-peptide systems, advancing its potential development as an antimicrobial agent.

1 Introduction

Antimicrobial peptides (AMPs) are a diverse group of short proteins integral to the innate immune system that can also be designed and developed synthetically as drug candidates with desirable properties[1]. They exhibit broad-spectrum activity against a variety of bacteria via several different mechanisms, including direct membrane disruption[2]. The efficacy of AMPs is associated with their physicochemical properties, particularly their hydrophobicity and charge state, which can be significantly influenced by environmental factors such as local ion concentration and local pH. Additionally, the membrane-disruptive AMPs' mechanism of action is directly linked to their structural conformation in the vicinity of the pathogenic membrane, which is likewise highly dependent on their charge states[3] and thus on the local environment imposed by the membrane. The potential interrelation between the charge state of an AMP and its conformational structure is a critical area for investigation. On the one hand, the charge state may significantly alter the conformational structure of the AMPs; on the other hand, the conformations control the intra-molecular interactions, which may change the local environment surrounding a residue, resulting in e.g. different pK_a values of residues of the same type[4].

In this work, we focus on characterizing pH-dependent charge and conformational states of the synthetic 13-residue-long AMP, GL13K (sequence *GKIIKLKASLKLL*). This AMP inhibits the gram-negative endotoxin, lipopolysaccharide (LPS), and exhibits both Gram-negative and Gram-positive bactericidal activity while maintaining low hemolyticity[5]. It was derived from the salivary protein parotid secretory protein (BPIFA2, formerly known as PSP) by isolating residues 141 to 153 and replacing glutamine, asparagine, and aspartic acid in positions 2, 5, and 11 with lysine residues, which have a standard pK_a of 10.5 in aqueous solution. These substitutions confer upon the peptide a strong pH dependence and increase the net charge of the peptide from +1 to +5 at physiological pH (7.4). The pH dependence of the peptide's charge state leads to pH-dependent conformational and aggregative behavior, which appears linked to its capacity to disrupt membranes of both Gram-negative and Gram-positive bacteria, while not impacting host red blood cell membranes. Specifically, at neutral pH, in solution, GL13K predominantly takes on random coil conformations, while at high pH, it shows an increasing propensity towards β sheet formation[6,7]. Similarly, it has been observed that near eukaryotic membrane mimics, such as those modeled by DOPC liposomes, it adopts primarily random-coil conformations, while, in the presence of gram-negative inner membrane/gram-positive plasma membrane mimics, such as those modeled by neg-

[a] *Physics Department, Concordia University, Montréal, QC, H4B 1R6, Canada, E-mail: re.mansbach@concordia.ca*

† Supplementary Information available: DOI: 00.0000/00000000.

atively charged DOPG liposomes, it adopts β sheet structures[8].

Because the conformational changes GL13K undergoes are accompanied by aggregative behavior (non-interacting random coil to strongly aggregative β sheet formation), both conformational changes and self-assembly propensity appear linked to its disruptive mechanisms, which appear to be a combination of pore-forming and micellization[6,8]. It has the ability to self-assemble into multiple types of supramolecular structures[9]. More precisely, while GL13K can form some crystalline β-sheets at an interface even at pH 7.4, it does not self-assemble into fibrils under these conditions[10]; it only forms fibrils at higher pH levels in solution[11] and when interacting with lipopolysaccharide, lipoteichoic acid (components of Gram-positive bacterial cell walls), or the cell wall of Gram-positive bacteria[12]. It has been suggested that the lysine residue located on the non-polar face (Figure 1-b, LYS11) controls fibril formation through the inhibition of peptide self-association when charged (at neutral pH) and the removal of this inhibition at higher pH, when it becomes uncharged[10].

Taken all together, this means that a crucial piece of groundwork in a better understanding of the complex process of GL13K conformational shift and aggregative behavior is a comprehensive calculation of the pH-dependence of its lysine residues and how they control its single-peptide conformations. Such single-molecule details are remarkably difficult to capture experimentally, which leads to a computational approach. Molecular dynamics (MD) simulations are a powerful tool for the study of peptides, such as molecular modeling of peptides[13], peptide-eukaryotic membrane interaction studies[14], peptide-prokaryotic membrane interaction studies[15], and designing new antimicrobial and anticancer peptides[16]. Although MD simulation is a powerful computational tool, it has some considerable limitations; in traditional MD simulations, the charge state of all amino acids in the sequence is set to the most probable value at the beginning of the simulation and during the simulations is fixed[17]. This limitation motivated us to use constant pH molecular dynamics to investigate the effects of pH on GL13K, in particular the role of lysine residues as titratable sites at different pH levels.

Constant pH molecular dynamics (CpHMD) is an advanced computational method used in molecular dynamics simulations to model pH-dependent processes in biological systems[18]. The most common method of implementing it utilizes the λ-dynamics technique to represent the dynamic protonation states of a titratable site using a continuous variable (λ), enabling sampling of multiple protonation states of titratable groups within a peptide in response to a fixed environmental pH. This method has been used to investigate pH effects on peptides in many biologically-relevant cases, including illuminating the pH-dependence of the structures of the amyloid-β protein, a key peptide in the pathology of Alzheimer's disease[19], , and the influenza virus fusion process involving influenza fusion peptide insertion into the membrane of the host endosome[20].

In this article, we employ constant pH molecular dynamics (CpHMD) to examine the protonation states of GL13K lysine residues and their impact on the conformational dynamics of a single peptide. In Section 3.1, we investigate the pK_a values of the four lysine residues, to test the hypothesis that *different lysine residues in the GL13K sequence have different pK_a values* according to their differing local environments. Accepting this hypothesis would suggest a differential impact of the lysines in the sequence on the control of crystalline β-sheets formation of GL13K at different pH levels and would support the idea of a single key lysine controlling aggregation. Next, in Section 3.2, we focus on the lysine residue located on the hydrophobic face (LYS11) (figure 1-b) and examine its protonation state. Specifically, we test the following hypothesis: *Residue 11 has a lower pK_a value than the other lysine residues.* A lower pK_a value would lead to deprotonation of the lysine on the hydrophobic face first and would support the hypothesis that it is the primary controlling factor of fibril formation. To our surprise, however, our results did not support this hypothesis, and thus we generated and tested a new hypothesis: *Residue 11 has a higher pK_a value than the other lysine residues.* To characterize the interplay between pH and accessible configurational space, we investigate the effects of pH level on β ladder content, end-to-end distances, and radius of gyration (Section 3.3) Finally, to better characterize the holistic interaction between pH and conformation, we explore metastable states using Markov state models (MSM) and compare the conformational landscapes of isolated GL13K across different pH levels (Sections A.1, and 3.4).

2 Methods

2.1 Constant pH Molecular Dynamics Simulations

We conduct all-atom constant pH molecular dynamics (CpHMD) simulations of a single peptide in a water box using the λ-dynamics package for GROMACS version 2021[17,24,25], ph-builder[26], with the Charmm36m force field modified for use with this package[27]. We model the initial structure of GL13K using the PEP-FOLD4 web-server[21]. We model the peptide with C-terminus amidation, meaning a neutral free carboxyl group, for consistency with previous simulations and experimental conditions[7,28]. All lysine residues are modeled as titratable, while other residues are modeled as fixed-charge, including the N-terminal +1 charge, as none of the protonation states of the other residues are expected to change much, even at a pH of 12.5. We run five independent replicates for ten different pH levels, 8, 8.5, 9, 9.5, 10, 10.5, 11, 11.5, 12, and 12.5, for a total of 50 simulations. These pH levels were chosen to center on the pK_a of lysine in water (10.5).

We place a single GL13K peptide in a cubic simulation box with an edge length of approximately 8 nm, ensuring a minimum padding of 2 nm from each face of the box such that the peptide does not see its periodic image. A steepest descent algorithm is used for energy minimization with a 0.01 step size and tolerance of $< 1000.0 \frac{kJ}{mol\ nm}$ to a maximum of 5000 steps. Each replica undergoes 10 nanoseconds (ns) of standard NVT and NPT equilibration, followed by a 100 ns final production run under NPT conditions. For analysis purposes, the last 90 ns (post-equilibration) is used. The starting state of all titratable residues (lysine residues) is positively charged (λ=0). We model the water with the TIP3P water model[29], a 3-site rigid water model. In all simulations, we use physiological salt concentration (0.15 mM) of NaCl. The simulation temperature is set to 300 K. Electrostatic interactions

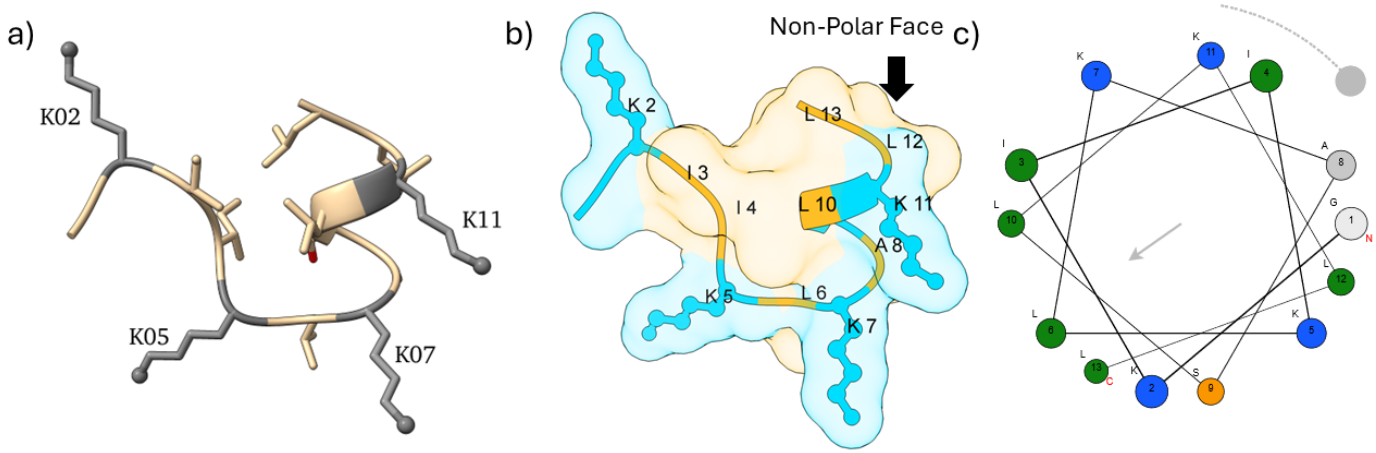

Fig. 1 Structure of GL13K and location of titratable residues. (a) Initial structure of GL13K peptide predicted by PEP-FOLD4 web-server[21]. Lysine residues are colored in gray with the side-chain Nitrogen (Nitrogen of ε-amino group as the titratable site) represented by a black ball. (b) The hydrophobic and hydrophilic surfaces of this structure are colored in gold and blue, respectively. The hydrophobic amino acids and Lysine residues are labeled. Balls and sticks represent lysine side chains. (c) Helical wheel of GL13K is plotted with charged polar residues in blue, uncharged polar residues in orange, hydrophobic residues with bulky side-chains in green, and small non-polar residues in light gray[22]. Subplots (a) and (b) were rendered with UCSF ChimeraX[23], developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

are modeled with the particle mesh Ewald method[30] with a real-space cutoff of 1.2 nm and a 0.14 nm Fourier grid spacing, and van der Waals interactions as LJ potentials with a 1.2 nm cutoff. The v-rescale thermostat[31] is used with a time constant of 0.5 ps. The C-rescale barostat[32] is used to keep the pressure at 1 bar with a 5 ps period. Integration is performed with the leapfrog integrator[33] with a timestep of 2 fs. Bond length constraints to hydrogens are enforced with the LINCS algorithm[34] to fourth order and a single correction iteration. For sampling, we record the data every 0.02 ns, resulting in 4500 samples for each lysine residue per replica at each pH level. In total, for each lysine residue, we have 22500 samples over five replicates at each pH level.

2.1.1 Average Deprotonated State and its relation to Henderson–Hasselbalch equation

In the λ-dynamics package for Gromacs introduced by Aho *et al.*[17], a non-terminal lysine residue has one titratable site (ε-amino group) attached to the furthest carbon (ε-carbon) from the α-carbon. In order to calculate the pK_a values of this titratable site for each lysine residue, we need to calculate the average fraction of the deprotonated state ($\overline{S_{dep}}$) across different replicas and various pH levels. The $\overline{S_{dep}}$ is calculated as

$$\overline{S_{dep}} = \frac{N_{dep}}{N_{dep} + N_{pro}} \tag{1}$$

where N_{dep} and N_{pro} are the total number of frames in which the site is deprotonated and protonated, respectively, as determined from the λ trajectory. Following the recommendations of the package authors[17,27] we define the site as deprotonated if $\lambda > 0.8$ and protonated if $\lambda < 0.2$.

The relation between the pH of a solution, the pK_a of an acid, and the ratio of the concentrations of its conjugate base to the acid is described by the Henderson-Hasselbalch equation[35],

$$pH = pK_a + \log\left(\frac{[A^-]}{[HA]}\right). \tag{2}$$

Since the ratio of concentrations can be assumed to be the same as the ratio of deprotonated to protonated frames, the fraction of the deprotonated state as a function of pK_a may be written,

$$\overline{S_{dep}} = \frac{10^{pH - pK_a}}{10^{(pH - pK_a)} + 1}. \tag{3}$$

2.1.2 Shape Analysis

To quantify GL13K conformations, we calculate a variety of shape parameters using the MDTraj scientific package[36]. We compute the radius of gyration (R_g), acylindricity (a_{cyl}), and asphericity (a_{sph}), which are all quantities derived from the principal moments $\xi_x^2, \xi_y^2, \xi_z^2$ of the gyration tensor. Additionally, we calculate the end-to-end distance (r_{e2e}) by finding the distance between α carbon of the first and the last residues.

Radius of gyration is defined as

$$R_g^2 = \xi_x^2 + \xi_y^2 + \xi_z^2, \tag{4}$$

and is a measure of the overall compactness (low R_g) or extendedness (high R_g) of the protein.

Asphericity is calculated as

$$a_{sph} = \xi_z^2 - \frac{1}{2}\left(\xi_x^2 + \xi_y^2\right) \tag{5}$$

which shows how much the shape of a peptide deviates from a perfect sphere; a value of zero indicates a perfect sphere, while a value of one indicates a perfect line.

Acylindricity is calculated as

$$a_{cyl} = \xi_y^2 - \xi_z^2 \tag{6}$$

which shows how much the shape of a peptide deviates from a perfect cylinder; a value of zero indicates a perfect cylinder.

End-to-end distance is described as

$$|r_{e2e}| = || R_{C\alpha}|_{res\ 1} - R_{C\alpha}|_{res\ 13}| \tag{7}$$

which measures the straight-line distance between the two terminal C_α atoms. Unlike the radius of gyration, this does not quantify the overall shape of the protein; a small r_{e2e} implies proximity of the termini but does not distinguish between a flat hairpin and a rounded loop.

2.1.3 Secondary Structure Analysis

For further quantification of GL13K's conformational tendencies, we assess the secondary structure of the peptide. We used the DSSP package from the MDTraj Python package[37] to identify the secondary structure of each residue at each timestep according to backbone angles and side chain positions. At each timestep i, the fraction $SSAF_S$ of a residue with assigned type S was calculated as,

$$SSAF_S(i) = \frac{\mathscr{O}_S(i)}{L}, \tag{8}$$

where $\mathscr{O}_S(i)$, the occurrence of identifier S at the ith timestep, is the number of times an identifier S was found in timestep i and $L = 13$ is the length of the peptide. We investigate general secondary structure formation using the coarse-grained DSSP assignments of $S \in \{H, E, C\}$ for α-helix, β sheet, and random coil, and we further investigate the different types of β sheet structures by using the more detailed version of DSSP, in which extended strand and isolated β-bridge fractions are defined by "E" and "B" identifiers.

2.1.4 Markov State Model

A Markov State Model (MSM) is a computational method used to analyze and model the dynamics of complex systems, such as protein folding, to find slow processes that are interpreted as transitions between stable states. These stable states are long-lived (e.g. metastable) conformational states (i.e. free energy wells) where the peptide remains for a significant time before transitioning to another state by crossing a free energy barrier. The construction of an MSM assumes that one can represent the system with a model that possesses the Markov property, meaning the future state of an MSM depends only on its current state, not its past states. To construct a traditional MSM, there are six essential steps[38]:

1. **Feature Selection**: Feature selection consists of identifying kinetically-relevant phase-space variables to describe the conformations of the system at each time step, such as peptide backbone dihedrals or distances between backbone atoms. Typically, such variables should be independent of system symmetries, i.e. rotation and translation.

2. **Dimensionality Reduction**: Dimensionality reduction, using methods like PCA or TICA, transforms high-dimensional features into a lower-dimensional space of collective variables (CVs) that is both easier to visualize and more robust for MSM construction while preserving important system dynamics. Thanks to collective motions, the dynamics of protein and peptide systems can be described by just a few dynamically or kinetically relevant variables[39].

3. **Discretization**: Discretization groups data described by CVs into kinetically relevant microstates using standard clustering techniques like k-means.

4. **MSM Estimation**: In the estimation step, one builds the transition probability matrix that describes the MSM between microstates based on the assumption of Markovianity.

5. **MSM Validation**: Via implied timescales and Chapman-Kolmogorov tests[40], one ensures that the assumption of Markovianity holds and that the model accurately captures the system kinetics.

6. **Coarse Graining the Microstates:** Coarse-graining groups microstates into macrostates using methods like Perron-Cluster Cluster Analysis (PCCA) or Robust Perron Cluster Analysis (PCCA+)[41,42], simplifying the MSM while retaining key dynamical properties for easier interpretation of metastable states.

Before constructing the MSM model for each pH level individually, we removed water and ions from the system and fixed the periodic boundary conditions (PBC) by making the peptide whole.

We considered backbone torsions, backbone atom distances, and backbone atom positions as potential features for constructing the MSM. Backbone torsions were selected as the feature because they displayed the highest VAMP-2 score[43]. For dimensionality reduction, we used two different methods, Time-lagged Independent Component Analysis (TICA)[44] and principal component analysis (PCA)[45]. PCA was used on all data simultaneously, to extract general principal components describing the overall conformational variance of the peptide as the pH varied. TICA CVs were determined separately for each pH level to better resolve specific kinetic details.

In the case of TICA, determining an initial suitable lag time for each pH model is a challenging task, and to our knowledge, there is no single method for selecting the lag time. Consequently, we tried a series of lag times and observed the sample free energy landscape built by projecting the trajectories into the independent components (ICs), and plotting it with the first two ICs. Then, we observed the projected trajectories over time for each IC to identify the possible transitions between local minima. Additionally, we calculated the VAMP-2 score for that lag time to ensure the validity of the projected feature into ICs. The smallest lag time that could satisfy these requirements was selected for each pH level. These requirements are (a) it captures the slow processes of the system, validated by visual inspection of the first n ICs over time, and generates a reasonable two-dimensional FES using IC1 and IC2, (b) follows Markovian behavior by inspecting the implied timescale. For discretization, we used the k-means clustering method[46] to group the data based on similarities of features.

For finding the optimal value for the number of clusters, we computed VAMP-2 scores across various cluster sizes and plotted them with confidence intervals. We chose the optimal number of clusters as the minimum number for which the plot of the VAMP-2 score displays a plateau.

For MSM estimation at each pH level for both PCA and TICA CVs, we calculated the implied timescale (ITS) and plotted it versus the lag time to estimate the optimal value of the lag time for constructing the MSM. As is standard in the literature[47], we chose the smallest lag time for which the implied timescale no longer changes, ensuring that it is above the $ITS = \tau$ equivalency line to ensure Markovian behavior. Also, the chosen lag time should separate the slow processes from the rapid ones clearly. In figures S13, S14, S15, S16, S17, and S18, we plotted the timescales of slowest process and timescale separation between them. For MSM validation, we used the Chapman-Kolmogorov test[40]. Finally, we grouped the microstates from previous steps into macrostates using the PCCA+ method[41,42], and we calculated the transition states and the mean first passage time (MFPT) between these states. We constructed the MSM models using the PyEMMA package, version 2.5.7[48] and Deeptime package, version 0.4.5[49].

3 Results and Discussion

We performed constant pH molecular dynamics of GL13K at 10 different pH levels to understand the deprotonation states of GL13K and their effects on peptide conformational behaviour. We investigated the differences of the four lysine residue pK_a values, specifically, if the pK_a value of the last lysine residue is lower than that of the other lysine residues. In what follows, for each section, we explain our hypotheses and explicitly state whether the step was part of hypothesis testing or generation.

The analyses were conducted in accordance with our preregistered plan on the Open Science Framework (OSF; see GL13K constant pH registration[50]). We tested the preregistered hypotheses as outlined in our OSF preregistration, while also exploring additional hypotheses generated during the study, which are discussed in the respective subsections below. The following hypotheses are part of hypothesis testing:

1. Different lysines in the GL13K sequence have different pK_a values.
2. Residue 11 (LYS11) has a lower pK_a than the other lysine residues.
3. Increasing pH will lead to loss of β strand content and decreasing radius of gyration and end-to-end radius.

The following hypothesis is part of hypothesis generation:

1. LYS11 has a higher pK_a value than the rest of the lysine residues.

3.1 LYS11 has higher pK_a value than LYS5 but not LYS2 or LYS7

Initially, we start by expressing the first null and alternative hypotheses as outlined in the OSF plan as part of hypothesis testing:

- **Null Hypothesis (H_0)**: Different lysine residues in the GL13K sequence possess equal pK_a values.

$$H_0 : pK_a|_{LYS2} = pK_a|_{LYS5} = pK_a|_{LYS7} = pK_a|_{LYS11}$$

- **Alternative Hypothesis (H_1)**: Different lysines in the GL13K sequence possess different pKa values.

$$H_1 : pK_a|_{LYS2} \neq pK_a|_{LYS5} \neq pK_a|_{LYS7} \neq pK_a|_{LYS11}$$

For rejecting or accepting the null hypothesis, we calculate the deprotonation fraction for each lysine residue at each pH level across 5 replicas according to equation 1. The pK_a is extracted for each lysine by fitting the Henderson-Hasselbalch equation to the average deprotonation at each pH level according to equation 3. Additionally, to quantify the precision and uncertainty of our extracted pK_a values, we estimate the variability by extracting the maximum and minimum possible pK_a values given the deprotonation data for each replica. If the pK_a values of different residues do not overlap to within these ranges of uncertainty, we interpret it as the pK_a values differing. Additionally, we quantify the magnitude of the difference as a measure of the effect size.

Figure 2 shows the calculated pK_a values and uncertainty level for each lysine residue. Except for LYS5 and LYS11, the rest have a mutual overlapping of pK_a values, so we reject the null hypothesis for the case of a single GL13K peptide in water. For LYS5 and LYS11, we observe a clear separation in uncertainty levels ([9.9678, 10.186] for LYS2 and [10.2326, 10.3554] for LYS11), indicating statistical significance; however, the small difference ($\approx$ 0.0466) with percent difference of average pK_a values between LYS5 and LYS11 defined as

$$\frac{|\overline{pK_a}_{Lys5} - \overline{pK_a}_{Lys11}|}{\frac{\overline{pK_a}_{Lys5} + \overline{pK_a}_{Lys11}}{2}} \times 100 \tag{9}$$

is 1.84 %, which suggests a small effect size.

3.2 LYS11 has a significant deprotonation ratio difference with other lysine residues at lysine side chain pK_a value

As previously mentioned, GL13K does not self-assemble into fibrils at pH 7.4 but does form fibrils at higher pH values. A previously advanced hypothesis for this behavior is that it is controlled by the pK_a value of the lysine residue on the hydrophobic face (LYS11). Therefore, in this section, we focus on this lysine, near the C terminus, employing the next hypothesis:

- **Null Hypothesis (H_0)**: LYS11 located on the hydrophobic face has the same pK_a value as other lysine residues.
- **Alternative Hypothesis (H_1)**: LYS11 located on the hydrophobic face has a lower pK_a value than other lysine residues.

In other words, we investigate whether LYS11 has a lower pK_a value than the rest of the lysine residues (hypothesis testing). To test this hypothesis, at each pH level, we quantify the number of frames where LYS11 is protonated/deprotonated and compare

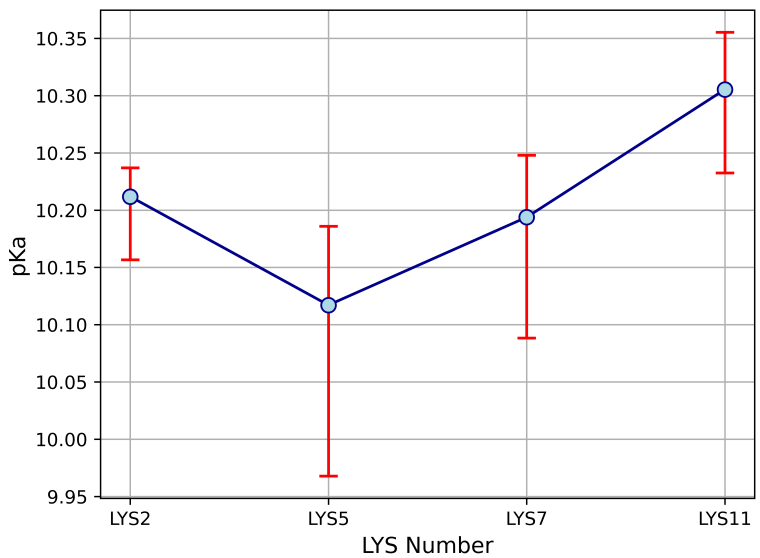

Fig. 2 The pK_a values and uncertainty level of all four lysine residues calculated by fitting the Henderson-Hasselbalch equation to the average deprotonation at each pH level.

it to the number of frames where other lysine residues are protonated/deprotonated. To avoid sampling non-physical states, a biasing potential term in the potential energy function ensures that the majority of λ values lie close to $\lambda = 0$ or $\lambda = 1$[17]. Therefore, after sampling the λ trajectory at each pH level, the resulting distributions are either unimodal (for primarily deprotonated or primarily protonated systems) or bimodal (for systems displaying both protonated and deprotonated character over the course of the simulation). Figure 3 shows these sampled distributions for all four lysine residues at each pH level. To rigorously distinguish the unimodal from the multimodal (bimodal) distributions, we assess the bimodality of each lysine's λ-value data set at each pH level using Hartigans' dip test[51], with a significance threshold of $p < 0.05$. In our study, a significant result indicates a bimodal distribution. The λ values of pH 8 and 12.5 show unimodal behavior for all lysines, and the rest passed Hartigans' dip test (p values reported in Fig. 3 and Table S1). For systems that passed Hartigans' dip test, the bimodal systems, we apply a Gaussian mixture model[52] to characterize the protonated and deprotonated states, yielding a Gaussian distribution per mode. Figure S1 shows the assigned Gaussian distribution functions of each lysine residue per pH level. In the next step, we assess the data set normality using the Quantile to Quantile (QQ) plot[53] and the Shapiro-Wilk test[54]. Figures S2, S3, and S4 show the QQ plot for the data set from the first mode, the second mode of bimodal distribution functions, and unimodal distribution functions, revealing a non-normally distributed function for all these datasets. Additionally, in table 1, p-values of the Shapiro-Wilk test are presented, demonstrating non-normally distributed functions. Moreover, we assess the variance equality of data sets for all the possible comparisons between LYS11 and other lysine residues using the Levene test[55]. In table 2 we present the p-value (p-value threshold is set at 0.05) regarding the Levene test, showing that the majority of comparisons do not have variance equality. At this point, we demonstrated that the data sets do not follow a normal distribution function and do not meet the variance equality condition. Therefore, for comparison between LYS11 deprotonation ratio and other lysine residues, we perform a non-parametric test, the Mann-Whitney U test[56], on the data sets.

To assess whether we can reject the null hypothesis that LYS11 is protonated at the same rate as each of the other lysine residues at each of the 10 pH levels, we perform a number of one-sided Mann-Whitney U tests using an asymptotic method (large sample size) with Bonferroni correction[57]. For the comparison, we have 4 lysine residues and 10 different pH levels, resulting in 30 comparisons. Aiming at a traditional familywise error rate of 0.05, we set the significance level for each test at p = $\frac{0.05}{30}$ = 0.0017. Table S2 presents the one-sided Mann-Whitney p-values for 30 comparisons, but these p-values are consistently large (close to 1), indicating no significant differences between the groups. In addition, we calculated the titration curve of each lysine residue at all different pH levels by fitting the deprotonation ratio into the Henderson-Hasselbach equation described in the methods section (equation 3 and plotting the average pK_a value.

Additionally, we plotted the average deprotonation ratio for each pH level across all replicas in Figure 4. The error bar is calculated as the standard error of independent replicas. In figure S8, we plotted the titration curve of each lysine residue for each replica. From both Figs 2 and 4 we observe that LYS11 not only lacks a lower pK_a value compared to other lysine residues but also exhibits a shift in the other direction, suggesting a potentially higher pK_a value instead, concomitant with especially at pH levels near the pK_a of lysine in water. From these results, we generate a new alternative hypothesis as such: "*LYS11 has a higher pK_a value*

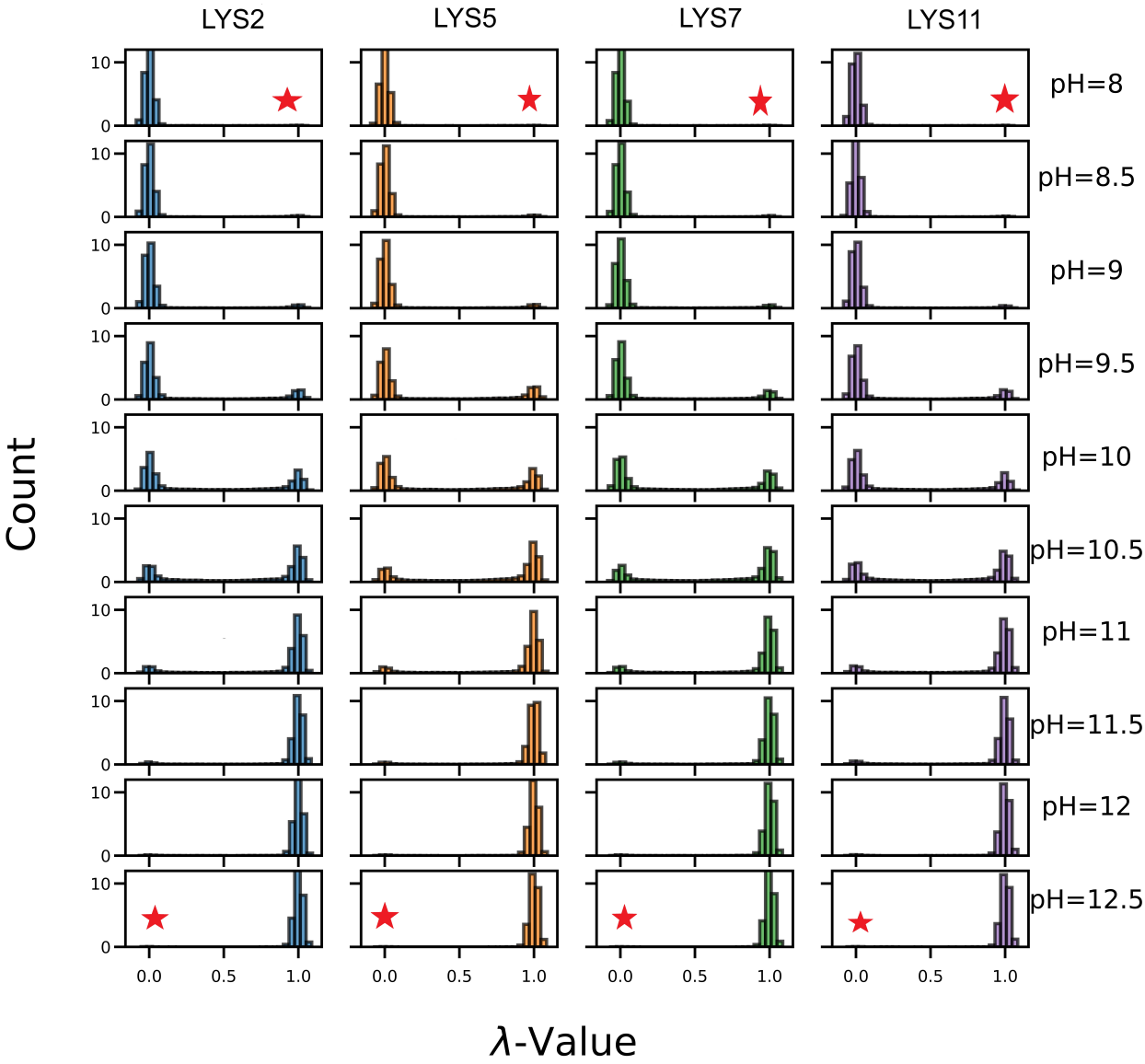

Fig. 3 λ values of 4 lysine residues at 10 different pH levels. The significance threshold of $p < 0.05$ is set to distinguish unimodal from bimodal distribution functions. The unimodal distribution functions are marked with a small red asterisk.

Table 1 P-Values from Shapiro-Wilk test

	pH=8	pH=8.5	pH=9	pH=9.5	pH=10	pH=10.5	pH=11	pH=11.5	pH=12	pH=12.5
LYS2	3E-129	3E-22,2E-33	2E-30,2E-46	4E-67,1E-37	2E-40,1E-81	3E-78,2E-40	1E-38,2E-58	7E-33,1E-37	4E-25,4E-23	2E-127
LYS5	5E-129	7E-24,1E-36	2E-30,1E-47	6E-37,4E-74	1E-83,1E-44	6E-74,8E-38	3E-37,8E-56	4E-34,7E-39	2E-28,3E-24	1E-126
LYS7	4E-129	2E-24,5E-31	5E-31,2E-45	4E-35,8E-65	1E-82,2E-42	1E-39,7E-76	3E-35,6E-58	3E-33,3E-40	7E-30,4E-24	6E-127
LYS11	1E-128	1E-22,2E-25	2E-31,1E-40	2E-34,3E-67	4E-40,5E-77	4E-42,2E-82	1E-39,1E-59	1E-30,1E-45	2E-26,2E-28	4E-128

Table 2 P-Values from Levene Test

	pH=8	pH=8.5	pH=9	pH=9.5	pH=10	pH=10.5	pH=11	pH=11.5	pH=12	pH=12.5
LYS2 to LYS11	1E-2	4E-20	2E-22	1E-1	1E-40	1E-28	6E-3	1E-37	3E-2	3E-2
LYS5 to LYS11	2E-1	5E-50	1E-26	8E-71	1E-90	5E-143	3E-12	9E-30	1E-1	6E-4
LYS7 to LYS11	3E-1	2E-13	1E-16	3E-4	2E-53	1E-75	1E-3	3E-20	2E-1	5E-3

than the rest of the lysine residues" (hypothesis generation). For assessing this hypothesis, we perform the same statistical test with the correct one-sided alternatives. The p-values are reported in Table 3. The results show a significant difference between LYS11 and LYS5 deprotonation ratios at pH levels of 8.5 to 11.5. Generally, LYS11 has a significantly different deprotonation ratio at pH 10, pH 10.5, and pH 11.5 from other lysine residues, specifically at the reported experimental pK_a value of lysine residue[58]. Thus, we reject the null hypothesis in favor of the newly-generated alternative hypothesis that LYS11 has a *higher* pK_a value than other residues as measured by the distributions of λ values.

3.3 Increasing pH stabilizes collapsed conformations partly through the stabilization of transient β strands

Another important aspect of peptides is their conformational shape, including radius of gyration (R_g), end-to-end distance ($e2e$), acylindricity (a_{cyl}), asphericity (a_{sph}), and secondary structure. In this section, we focus on the effects of pH on peptide conformation. Due to lysine residues charge-charge interactions, the peptide may not be able to form a more compact conformation. However, by increasing the pH and crossing the pK_a values of lysine residues, these charges are neutralized, and more compact structures may be seen. Based on these charge-charge interactions, we generate the following hypothesis:

- **Null Hypothesis (H_0)**: Increasing pH does not have any effect on radius of gyration, end-to-end distance, and β-strand content.
- **Alternative Hypothesis (H_1)**: Increasing pH results in a decrease of radius of gyration and end-to-end distance, but an increase of β-strand content.

Figure 5 shows average values of R_g, r_{e2e}, a_{cyl} and a_{sph} at 10 pH levels. The standard error for each pH level is calculated across five independent replicates and indicated as the error bar. We observe little change in the shape parameters with pH. There is a small tendency towards lower radius of gyration and higher variance above pH 10 (cf. also Fig. S9) concomitant with a small tendency towards lower end to end distance and asphericity and higher variance in both (cf. also Figs S10, S11), but no change in acylindricity (cf. also Fig. S12). To quantify the effect size of pH increment on peptide shape descriptors, we calculate the Pearson correlation coefficient between pH and each one of these descriptors. The Pearson correlation coefficient measures the strength and direction of a linear relationship between two variables[59]. In table 4, we present the Pearson correlation coefficient between pH and peptide shape descriptor variables including R_g, $e2e$, a_{cyl}, a_{sph}, and β-strand fraction. We observe a moderate negative linear relation between pH and R_g, $e2e$, and a_{sph} and a weak negative linear relation between pH and a_{cyl}.

Taken together, this indicates that for environmental pH values greater than the pK_a of all lysine residues, when all lysines are nearly entirely deprotonated (cf. Fig. 3), more compact conformations are moderately stabilized; whereas below these pH values, both extended and compact conformations are accessible, with a preference for extended conformations.

Figure 6 shows the average secondary structure of the peptide over different pH levels, focusing on isolated β-bridge and extended strand percentage, with the standard error computed across five independent replicates. In a single peptide, an isolated β-bridge is a minimal β-sheet-like structure, defined by a single hydrogen bond between two residues, creating a localized interaction. On the other hand, an extended strand is a broader β-sheet-like conformation, involving a longer segment of the peptide and potentially the formation of multiple hydrogen bonds between residues. In this figure, all systems across various pH levels exhibit approximately between 90% to 95% coil structure, aligning closely with prior computational studies on the isolated GL13K peptide in solution[7], although the earlier report indicated a slightly higher percentage of helical and β-strand structures in uncharged GL13K, GL13K in its fully deprotonated state, and charged GL13K, GL13K in its fully protonated state, respectively. It is also important to note that the majority of helical structures are formed at a pH level below pH 10, where all four lysines are primarily found in the protonated (charged) state (cf. Fig. 3, top four rows), but β-strand structures are exclusively observed at pH levels of 10.5 and above, where all four lysines are primarily found in the deprotonated (uncharged) state (cf. Fig. 3, bottom five rows). Remarkably, at the reported experimental pK_a value of the lysine side chain (10.5), we observed the emergence of β-sheet structures for the first time, suggesting a critical role of lysine's protonation state in stabilizing these secondary structural elements. In figures S5, S6, and S7, we plotted the average helical, strand, and coil-like fraction per residue at each pH level, respectively. Specifically, in the figure S6, we observed that the β-sheet contents are mainly formed between LYS7 and LYS11, suggesting the important role of the partial deprotonation of those two residues in stabilizing intra-peptide β-sheet contents.

In table 4, we observe a moderate positive linear relation between pH and the strand fraction of secondary structure.

Based on the results, we argue that we can reject the null hypothesis. However, the Pearson correlation coefficient between pH and strand fraction is approximately 0.3, indicating a weak relation. It is important to mention that a vast number of AMPs are disordered in a non-active environment, such as solution[60].

3.4 Increasing pH results in loss of helical structures and more dominance of random coil structure.

Free energy surfaces (FES) can reveal local minima, barriers between these states, and frequency of visited states of the system. The Helmholtz free energy (F) can be calculated as

$$F = -k_B \times T \times ln(\Omega) \tag{10}$$

where k_B, T, and Ω are the Boltzmann constant, absolute temperature, and number of accessible microstates at a given energy level, respectively. In equation 10, the populated states (high Ω) result in lower free energy.

In this article, we construct and compare free energy surfaces in three different sets of collective variables. We calculate the collective motions of greatest variance across all ten systems using principal components analysis (PCA), employ the first two mo-

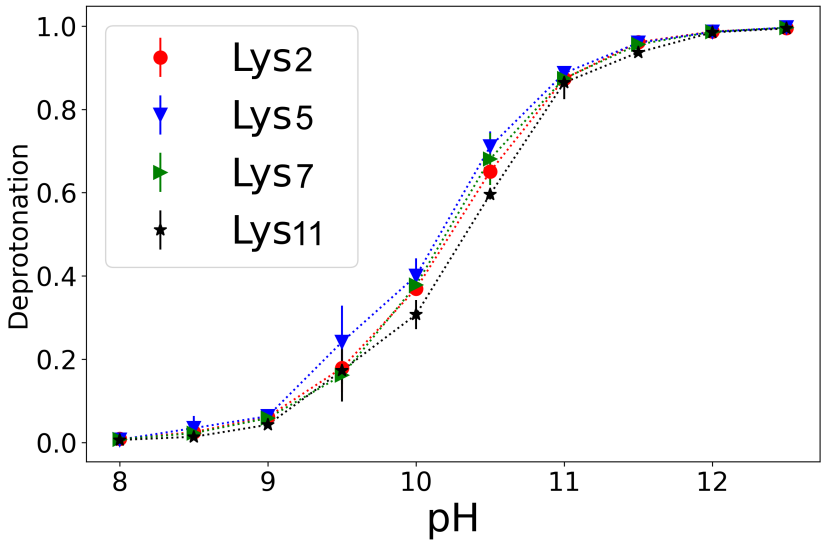

Fig. 4 Titration curves of the lysine residues as titratable sites in GL13K peptide. The error bars indicate the standard error in the means of the five different replicas. Connecting lines are included to guide the eye and do not represent fits.

Table 3 P-Values from one-sided Mann-Whitney Test

	pH=8	pH=8.5	pH=9	pH=9.5	pH=10	pH=10.5	pH=11	pH=11.5	pH=12	pH=12.5
LYS2 to LYS11	7E-2	1E-5	8E-9	2E-1	7E-33	4E-27	3E-1	4E-12	1E-2	1E-1
LYS5 to LYS11	8E-2	2E-13	5E-12	5E-43	8E-70	1E-104	8E-5	2E-10	3E-1	8E-1
LYS7 to LYS11	2E-1	2E-4	4E-5	9E-1	2E-43	1E-61	1E-2	8E-9	2E-1	7E-1

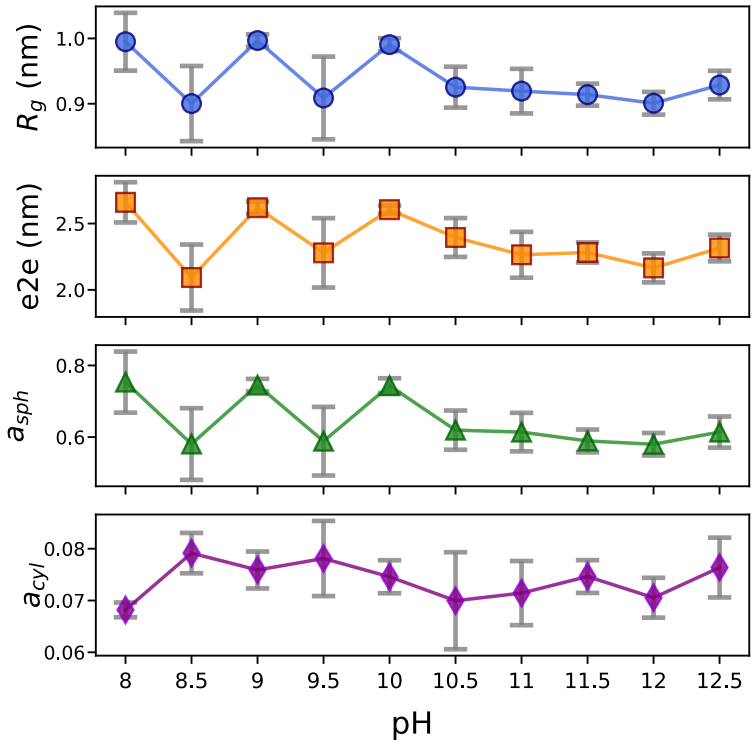

Fig. 5 Average shape parameters calculated for 10 pH levels from top to bottom: radius of gyration (R_g), end-to-end distance ($e2e$), asphericity (a_{sph}), and acylindricity (a_{cyl}). The error bars indicate the standard error in the means of the five different replicas.

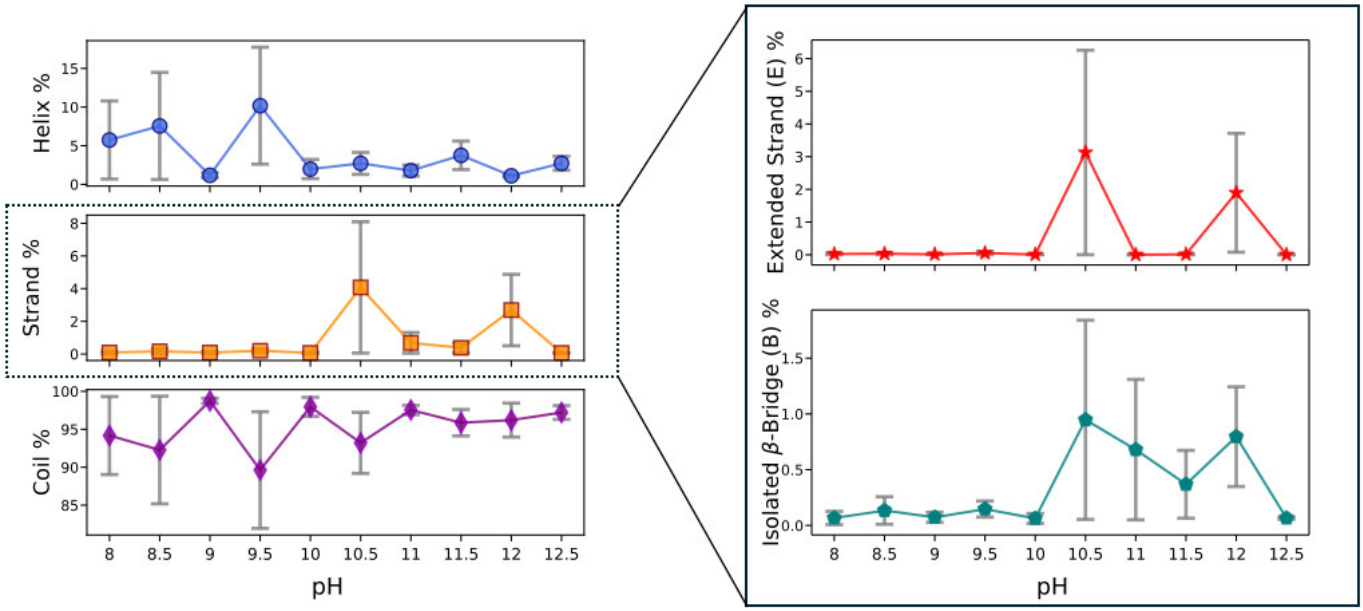

Fig. 6 Secondary structure of peptide at each pH level is calculated. (left) Three categories of secondary structure for each pH level are plotted with standard error as error bars. (right) Two components of the Strand structure, isolated β-bridge and extended strand, are plotted for each pH level.

Table 4 Pearson correlation coefficient

Variable	R_g	e2e	a_{sph}	a_{cyl}	Strand Fraction
pH	-0.48	-0.42	-0.49	-0.23	0.32

ments of the gyration tensor as simple dynamical variables that are known to well-describe the collective motions of short peptides, and calculate for each individual system the top moments of greatest kinetic variance using time-independent component analysis (TICA) in figures 7-a, 7-b, and 7-c, respectively. We applied PCA to a dataset containing all 10 pH level trajectories and applied the obtained transformation to each pH level trajectory individually to be able to compare the constructed free energy surfaces among the pH levels in figure 7-a. In figure 7-a, we observed that FES minima are not always guaranteed to represent unique stable states with clear distant separation resulting in distinct secondary structure, such as pH 9, 10,12. However, this method provides an opportunity to compare the FES of different pH levels with each other. In figure 7-b, we observed that using a simple and the most interpretable approach does not result in capturing all the stable states. However, it clearly shows that increasing the pH, which results in neutralizing the lysine residues, removes the free energy barriers between the states and allows more intermediate visits along the minima paths as illustrated in pH 8, 11.5, and 12.5. Also, by increasing the pH, regions associated with alpha helical and β-sheet structures were destabilized and stabilized, respectively. In figure 7-c, we observed that the TICA transformation on each pH level separately results in the separation of stable states; however, these states are not always identical to figure 7-a, and figure 7-b. At pH 10.5, we only observed a short anti-parallel beta sheet formed in the vicinity of the C-terminus in figures 7-a, and 7-b. This structure was not captured by TICA transformation. The red asterisk in each panel represents the location of the short anti-parallel beta sheet. On the other hand, we observed a short partial alpha helical minima at pH 10.5 using TICA, which were not captured by other methods. We argue that the differences in results arise from the fact that TICA identifies slow, stable conformational states, whereas PCA focuses on capturing the directions of maximum variance, which may not correspond to stable states. This distinction is particularly evident when analyzing features such as 1st and 2nd moments of the gyration tensor in free energy surfaces (FES) constructed using these methods.

4 Conclusions

We performed a constant pH molecular dynamics simulation of a single GL13K antimicrobial peptide in solution at 10 different pH levels from 8 to 12.5. We investigated the pK_a values of 4 lysine residues with the null hypothesis that different lysine residues in the GL13K sequence have the same pK_a values. We showed that the last lysine (LYS11) has a higher pK_a value than the second lysine residue (LYS5), but has an overlap with LYS5 and LYS7 pK_a values. Therefore, we can not reject this null hypothesis for the case of a single GL13K peptide in water. However, it is important to note that the null hypothesis cannot be rejected under the single GL13K in solution conditions. For the multi-peptide condition, it remains unclear whether the null hypothesis holds, as further evidence is needed to determine its validity. Additionally, we showed that the last lysine residue (LYS11) has a significant

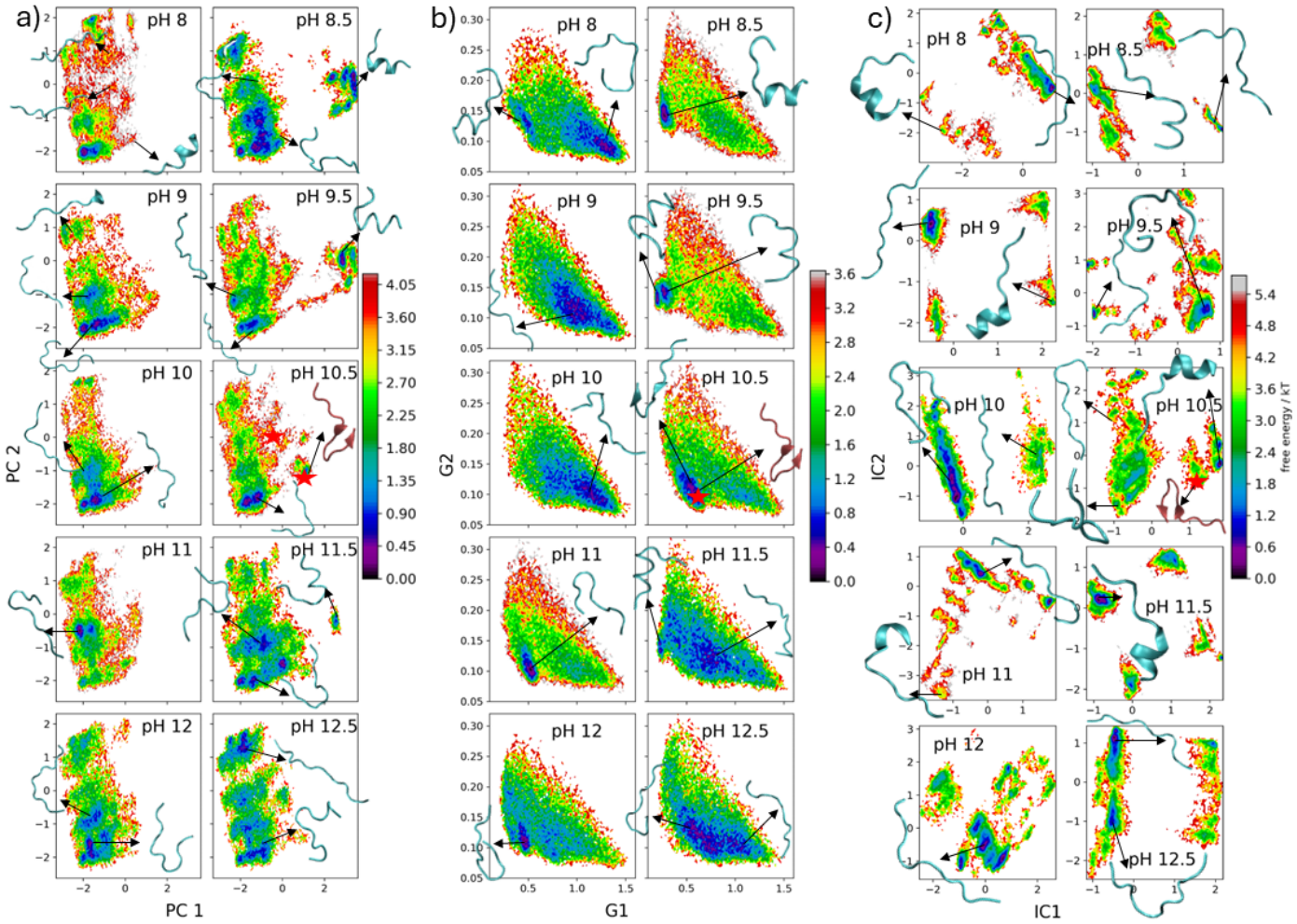

Fig. 7 Two-dimensional free energy surfaces of a single GL13K in solution at different pH levels, (a) applying PCA transformation obtained from all the data across 10 pH levels on individual pH levels. (b) Plotting simple FES using 1st and 2nd moments of the gyration tensor. (c) Applying the TICA transformation to each of the 10 pH levels individually. The red asterisk represents the location of the short anti-parallel beta sheet in these methods.

deprotonation ratio difference with other lysine residues at the reported average experimental side chain pKa value. Consequently, for modeling the protonation/deprotonation of a single GL13K, we can apply the same state for all lysine residues, which is an important aspect for modeling the peptide and its aggregation. We observed that increasing the pH results in a decrease of R_g and r_{e2e} eventually due to removing the lysine charge-charge repulsion. We constructed two-dimensional free energy surfaces using three different methods: PCA, TICA, and simple interpretable CVs. We observe the dominance of the random coil structure of a single GL13K in solution. By increasing the pH, a significant transition from the alpha helical-like in primarily deprotonated states to β-hairpin structure was observed. Furthermore, the high pH, which results in neutralizing lysine side chains, causes the removal of the energy barriers. By comparing peptide conformations at low pH, where protonated states predominate, with those at pH values near the pKa, we identify significant differences that suggest traditional fixed-charge simulations may be inadequate for accurate modeling. Another important observation we had was that the β-hairpin structures in a single peptide mainly occur at residues close to the C-terminus, even though from our previous studies, we observed that uncharged beta ladder content is mostly formed at the N-terminus[7]. We argue that lysines 2,5, and 7 becoming uncharged controls larger-scale aggregation in multi-peptide conditions, while LYS11 can potentially control and contribute to LPS binding through exposing the SER 9 to the lipopolysaccharide molecule[61]. An important area of future work includes the investigation of GL13K pK_a values of different lysine residues in multi-peptide conditions using CpHMD methods with the focus on secondary structure shifts from random coil and helical structures to β-sheet contents at high pH values and revealing more highly populated metastable states of β-strand structures.

Supporting Information

Supporting information is available, including 2 tables and 26 figures.

The datasets containing molecular dynamics trajectories without water, structures, force fields, molecular dynamics input parameters, and analysis files are publicly available at the following DOI:$https://doi.org/10.5281/zenodo.17240858$. Data, including the positions of water molecules, are available from the authors upon reasonable request.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research was supported in part by Discovery Grant #RGPIN-2021-03470 from the National Sciences and Engineering Research Council of Canada. This research was enabled in part by support provided by Calcul Quebec (`www.calculquebec.ca`) and the Digital Research Alliance of Canada (`https://alliancecan.ca`). This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program under grant number CRC-2020-00225.

References

1 Y. Huan, Q. Kong, H. Mou and H. Yi, *Frontiers in microbiology*, 2020, **11**, 582779.

2 N. Chen and C. Jiang, *European Journal of Medicinal Chemistry*, 2023, **255**, 115377.

3 M.-A. Sani and F. Separovic, *Accounts of chemical research*, 2016, **49**, 1130–1138.

4 N. V. Di Russo, D. A. Estrin, M. A. Martí and A. E. Roitberg, *PLoS Computational Biology*, 2012, **8**, e1002761.

5 M. Abdolhosseini, J. B. Sotsky, A. P. Shelar, P. B. M. Joyce and S.-U. Gorr, *Molecular and Cellular Biochemistry*, 2012, **359**, 1–8.

6 N. Harmouche, C. Aisenbrey, F. Porcelli, Y. Xia, S. E. D. Nelson, X. Chen, J. Raya, L. Vermeer, C. Aparicio, G. Veglia, S.-U. Gorr and B. Bechinger, *Biochemistry*, 2017, **56**, 4269–4278.

7 M. N. Hamidabad, N. A. Watson, L. N. Wright and R. A. Mansbach, *ChemBioChem*, 2024, **25**, e202400088.

8 V. Balhara, R. Schmidt, S.-U. Gorr and C. DeWolf, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 2013, **1828**, 2193–2203.

9 Z. Ye and C. Aparicio, *Nanoscale Advances*, 2019, **1**, 4679–4682.

10 H. Youssef and C. E. DeWolf, *Langmuir*, 2020, **36**, 660–665.

11 Z. Ye, X. Zhu, S. Acosta, D. Kumar, T. Sang and C. Aparicio, *Nanoscale*, 2019, **11**, 266–275.

12 Z. Ye and C. Aparicio, *Journal of Peptide Science*, 2022, **28**, e3299.

13 K. Kuczera, *Computational Peptidology*, Springer New York, New York, NY, 2015, vol. 1268, pp. 15–41.

14 K. Hristova, C. E. Dempsey and S. H. White, *Biophysical Journal*, 2001, **80**, 801–811.

15 J. C. Stephani, L. Gerhards, B. Khairalla, I. A. Solov'yov and I. Brand, *ACS Infectious Diseases*, 2024, **10**, 763–778.

16 R. Deb, M. D. T. Torres, M. Boudný, M. Koběrská, F. Cappiello, M. Popper, K. Dvořáková Bendová, M. Drabinová, A. Hanáčková, K. Jeannot, M. Petřík, M. L. Mangoni, G. Balíková Novotná, M. Mráz, C. De La Fuente-Nunez and R. Vácha, *Journal of Medicinal Chemistry*, 2024, **67**, 14040–14061.

17 N. Aho, P. Buslaev, A. Jansen, P. Bauer, G. Groenhof and B. Hess, *Journal of Chemical Theory and Computation*, 2022, **18**, 6148–6160.

18 V. Martins De Oliveira, R. Liu and J. Shen, *Current Opinion in Structural Biology*, 2022, **77**, 102498.

19 T. Albrahadi, C. Hureau and J. A. Platts, *Chemistry – A European Journal*, 2025, **31**, e202500547.

20 D. Lousa, A. R. T. Pinto, S. R. R. Campos, A. M. Baptista, A. S. Veiga, M. A. R. B. Castanho and C. M. Soares, *Scientific Reports*, 2025, **10**, 20082.

21 J. Rey, S. Murail, S. de Vries, P. Derreumaux and P. Tuffery, *Nucleic Acids Research*, 2023, **51**, W432–W437.

22 *Helical Wheels*, `https://clemlab.github.io/helicalwheel/`.

23 E. C. Meng, T. D. Goddard, E. F. Pettersen, G. S. Couch, Z. J.

Pearson, J. H. Morris and T. E. Ferrin, *Protein Science*, 2023, **32**, e4792.
24 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1-2**, 19–25.
25 D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, *Journal of Computational Chemistry*, 2005, **26**, 1701–1718.
26 A. Jansen, N. Aho, G. Groenhof, P. Buslaev and B. Hess, *Journal of Chemical Information and Modeling*, 2024, **64**, 567–574.
27 P. Buslaev, N. Aho, A. Jansen, P. Bauer, B. Hess and G. Groenhof, *Journal of Chemical Theory and Computation*, 2022, **18**, 6134–6147.
28 H. Hirt and S.-U. Gorr, *Antimicrobial Agents and Chemotherapy*, 2013, **57**, 4903–4910.
29 P. Mark and L. Nilsson, *The Journal of Physical Chemistry A*, 2001, **105**, 9954–9960.
30 T. Darden, D. York and L. Pedersen, *The Journal of Chemical Physics*, 1993, **98**, 10089–10092.
31 G. Bussi, D. Donadio and M. Parrinello, *The Journal of Chemical Physics*, 2007, **126**, 014101.
32 M. Bernetti and G. Bussi, *The Journal of Chemical Physics*, 2020, **153**, 114107.
33 R. Hockney, S. Goel and J. Eastwood, *Journal of Computational Physics*, 1974, **14**, 148–158.
34 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *Journal of Computational Chemistry*, 1997, **18**, 1463–1472.
35 H. N. Po and N. M. Senozan, *Journal of Chemical Education*, 2001, **78**, 1499.
36 R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane and V. S. Pande, *Biophysical Journal*, 2015, **109**, 1528 – 1532.
37 R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane and V. S. Pande, *Biophysical Journal*, 2015, **109**, 1528–1532.
38 G. R. Bowman, V. S. Pande and F. Noé, *An introduction to Markov state models and their application to long timescale molecular simulation*, Springer Science & Business Media, 2013, vol. 797.
39 A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis and P. G. Debenedetti, *Chemical Physics Letters*, 2011, **509**, 1–11.
40 J. F. Charles Kingman, *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 1974, **276**, 341–369.
41 C. Schütte, A. Fischer, W. Huisinga and P. Deuflhard, *Journal of Computational Physics*, 1999, **151**, 146–168.
42 S. Röblitz and M. Weber, *Advances in Data Analysis and Classification*, 2013, **7**, 147–179.
43 H. Wu and F. Noé, *Journal of Nonlinear Science*, 2020, **30**, 23–66.
44 L. Molgedey and H. G. Schuster, *Physical Review Letters*, 1994, **72**, 3634–3637.
45 A. Maćkiewicz and W. Ratajczak, *Computers & Geosciences*, 1993, **19**, 303–342.
46 X. Jin and J. Han, in *K-Means Clustering*, ed. C. Sammut and G. I. Webb, Springer US, Boston, MA, 2010, pp. 563–564.
47 N. Nazeer, N. Kooner, A. Ghimire, J. K. Rainey, W. D. Lubell, D. Meneksedag-Erol and M. Ahmed, *Journal of Medicinal Chemistry*, 2024, **67**, 8693–8707.
48 M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz and F. Noé, *Journal of Chemical Theory and Computation*, 2015, **11**, 5525–5542.
49 M. Hoffmann, M. Scherer, T. Hempel, A. Mardt, B. De Silva, B. E. Husic, S. Klus, H. Wu, N. Kutz, S. L. Brunton and F. Noé, *Machine Learning: Science and Technology*, 2022, **3**, 015009.
50 R. Mansbach and Mohammadreza Niknam Hamidabad, *GL13K constant pH*, 2024, `https://osf.io/ew4vm/`.
51 J. A. Hartigan and P. M. Hartigan, *The Annals of Statistics*, 1985, **13**, 100.
52 D. Reynolds, *Encyclopedia of Biometrics*, Springer US, Boston, MA, 2009, pp. 659–663.
53 B. Das and S. I. Resnick, *Stochastic Models*, 2008, **24**, 103–132.
54 S. S. Shapiro and M. B. Wilk, *Biometrika*, 1965, **52**, 591–611.
55 H. Levene, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, 1960, **2**, 278–292.
56 H. B. Mann and D. R. Whitney, *The Annals of Mathematical Statistics*, 1947, **18**, 50–60.
57 M. E. Dewey and E. Seneta, in *Carlo Emilio Bonferroni*, ed. C. C. Heyde, E. Seneta, P. Crépel, S. E. Fienberg and J. Gani, Springer New York, New York, NY, 2001, pp. 411–414.
58 S. Pahari, L. Sun and E. Alexov, *Database*, 2019, **2019**,.
59 K. Pearson, *Proceedings of the Royal Society of London*, 1895, **58**, 240–242.
60 P. Kumar, J. Kizhakkedathu and S. Straus, *Biomolecules*, 2018, **8**, 4.
61 M. Abdolhosseini, S. R. Nandula, J. Song, H. Hirt and S.-U. Gorr, *Peptides*, 2012, **35**, 231–238.

A Supporting Information

Figures S5, S6, and S7 show the average secondary structure and uncertainty level of peptide per residue for helix, strand, and coil fraction over different pH levels, respectively.

A.1 High pH results in dominance of coil structure stable states

Understanding stable states of an antimicrobial peptide reveals the structural behavior of that peptide in the environment. In Figures S23,S24, we built Markov state models using TICA as a dimensionality reduction method and demonstrated the metastable state population percentage in equilibrium for GL13K at each pH level, with the mean first passage time (MFPT) between states calculated in nanoseconds. The relative free energy for each state is calculated. These figures show that a single peptide GL13K majorly takes a random coil structure in solution. For pH levels lower than the experimental pK_a value of the lysine side chain

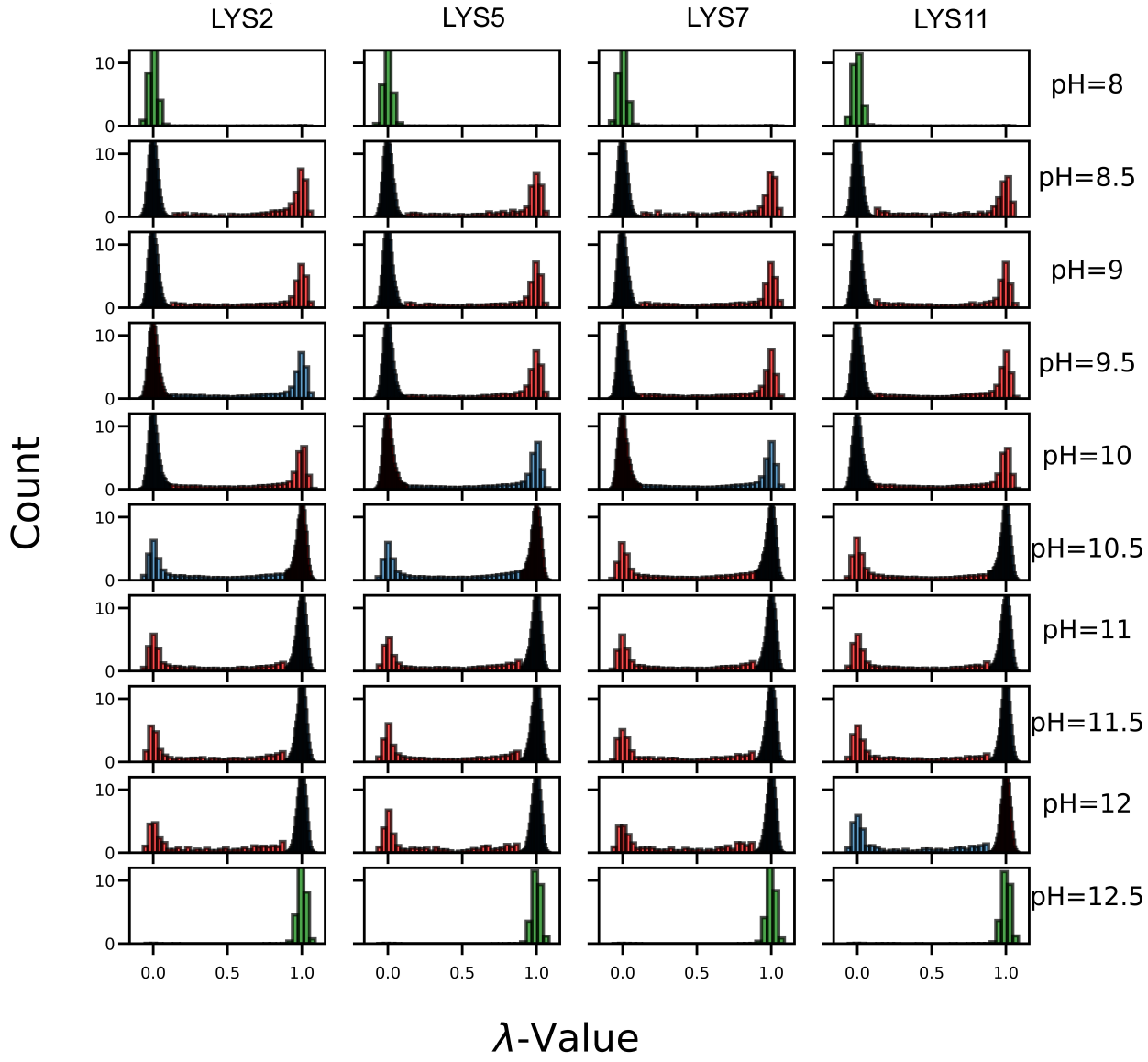

Fig. S1 Assigned Gaussian distribution functions of each lysine residue per pH level.

Table S1 P-Values from Hartigan's Dip Test

	pH=8	pH=8.5	pH=9	pH=9.5	pH=10	pH=10.5	pH=11	pH=11.5	pH=12	pH=12.5
LYS2	4E-1	0	0	0	0	0	0	0	4E-2	9E-1
LYS5	7E-1	0	0	0	0	0	0	0	1E-2	9E-1
LYS7	6E-1	0	0	0	0	0	0	0	3E-2	7E-1
LYS11	9E-1	3E-3	0	0	0	0	0	0	9E-4	8E-1

Table S2 P-Values from one-sided Mann-Whitney Test with wrong alternative hypothesis

	pH=8	pH=8.5	pH=9	pH=9.5	pH=10	pH=10.5	pH=11	pH=11.5	pH=12	pH=12.5
LYS2 to LYS11	9E-1	1	1	7E-1	1	1	7E-1	1	1	9E-1
LYS5 to LYS11	9E-1	1	1	1	1	1	1	1	7E-1	2E-1
LYS7 to LYS11	8E-1	1	1	1E-4	1	1	1	1	8E-1	3E-1

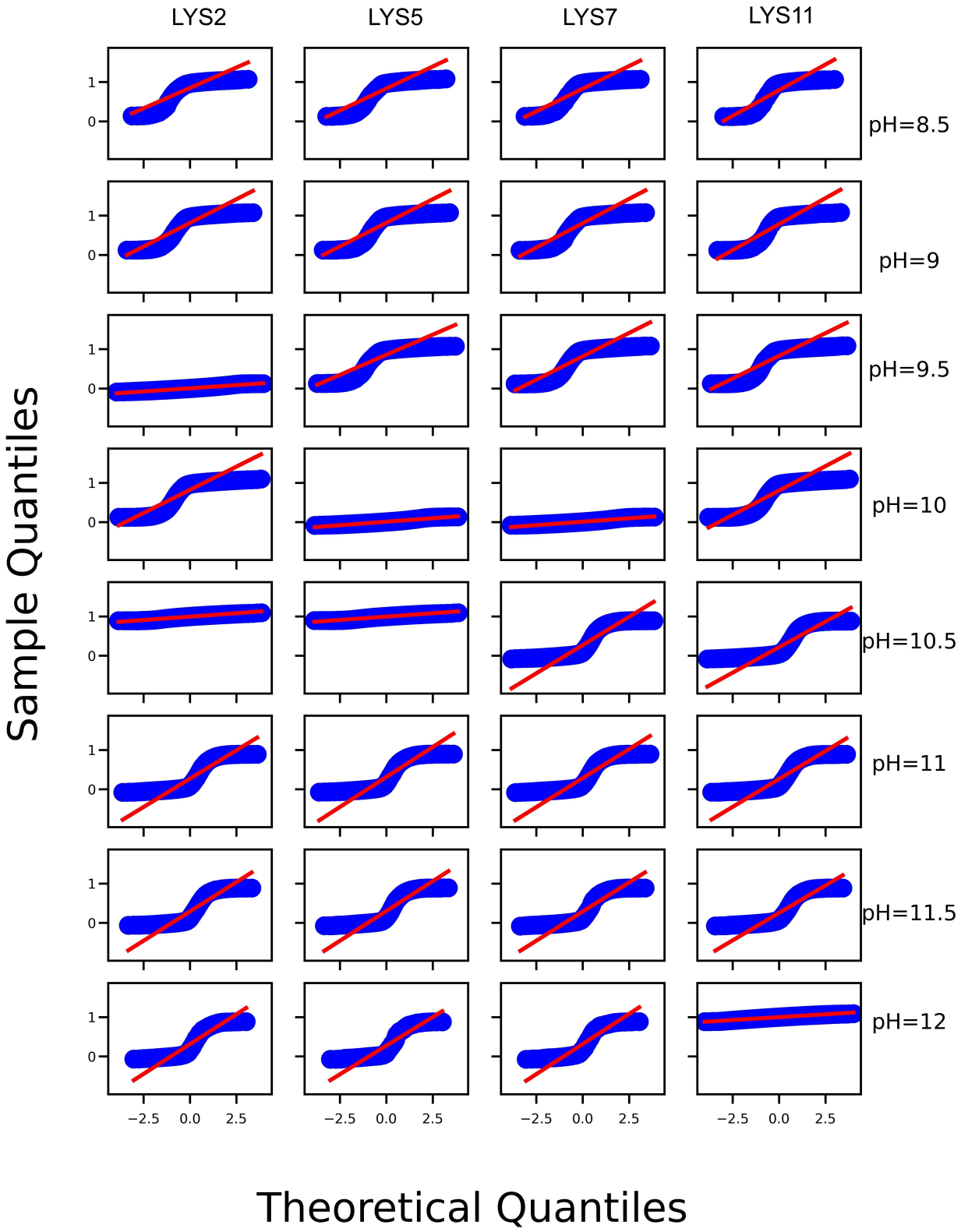

Fig. S2 QQ plot, first mode of bimodal distribution function

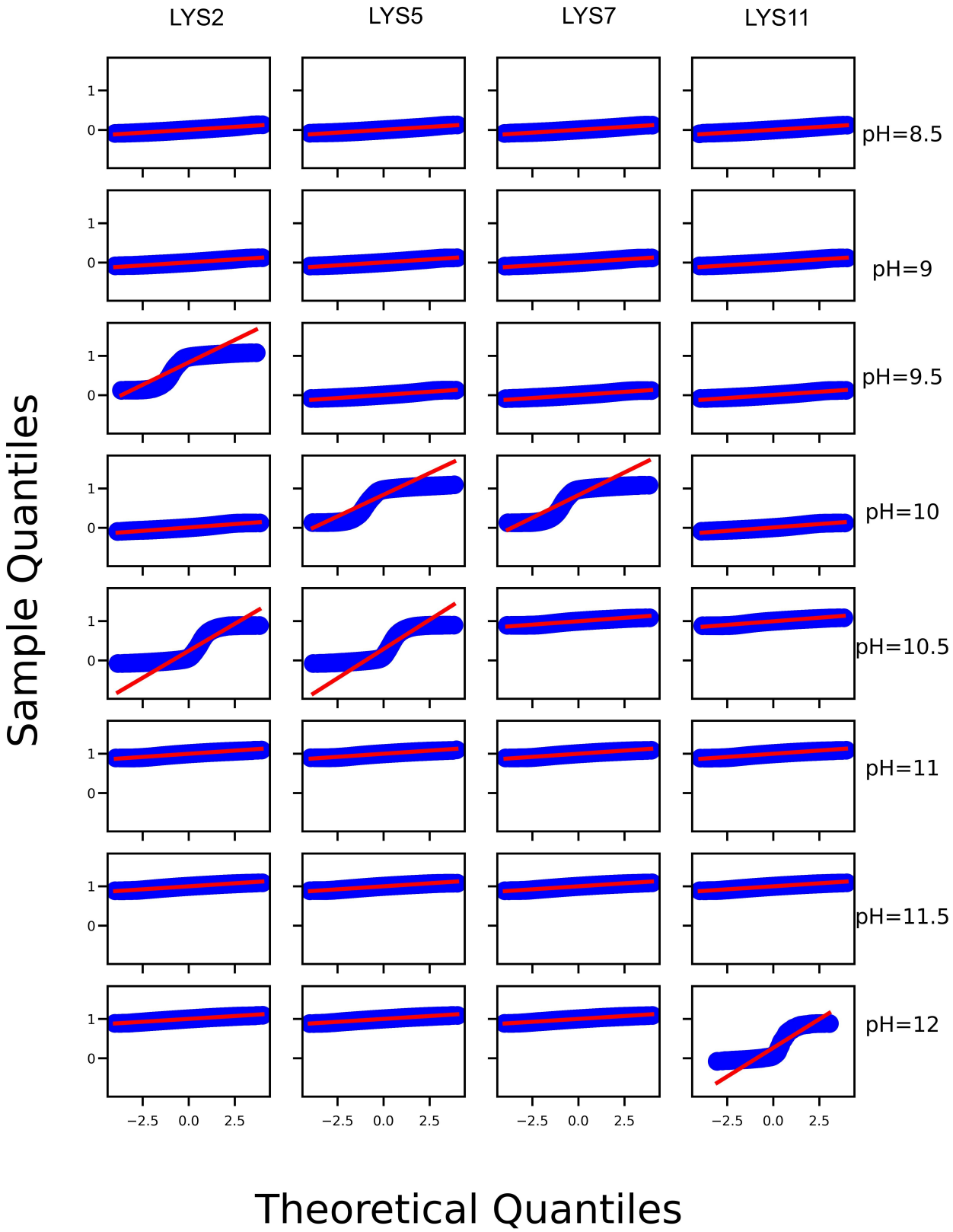

Fig. S3 QQ plot, second mode of bimodal distribution function

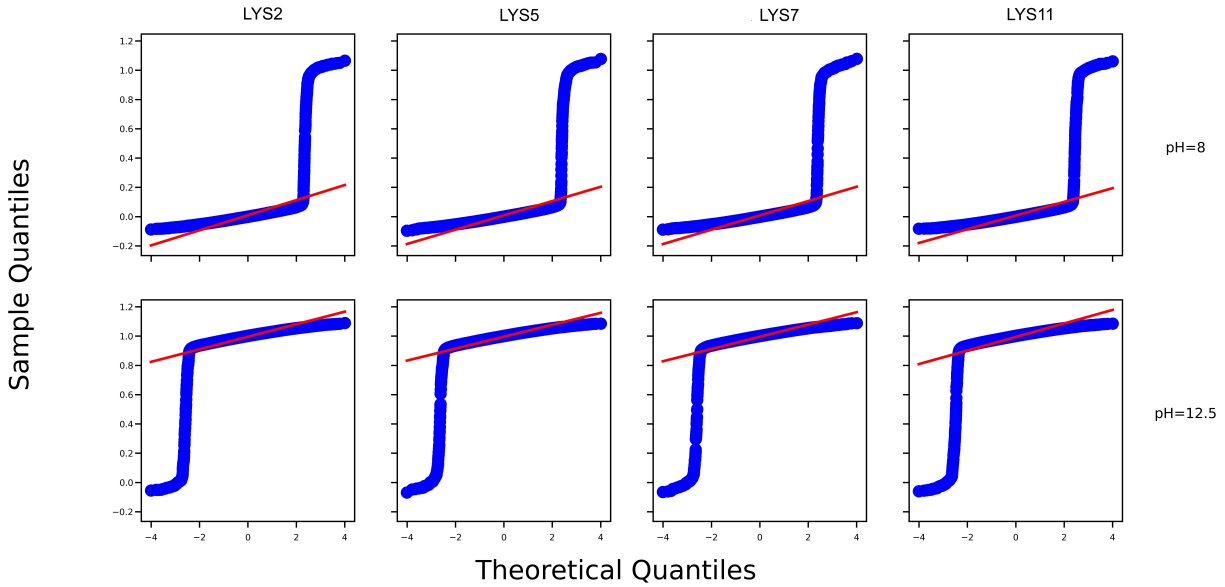

Fig. S4 QQ plot, unimodal distribution function

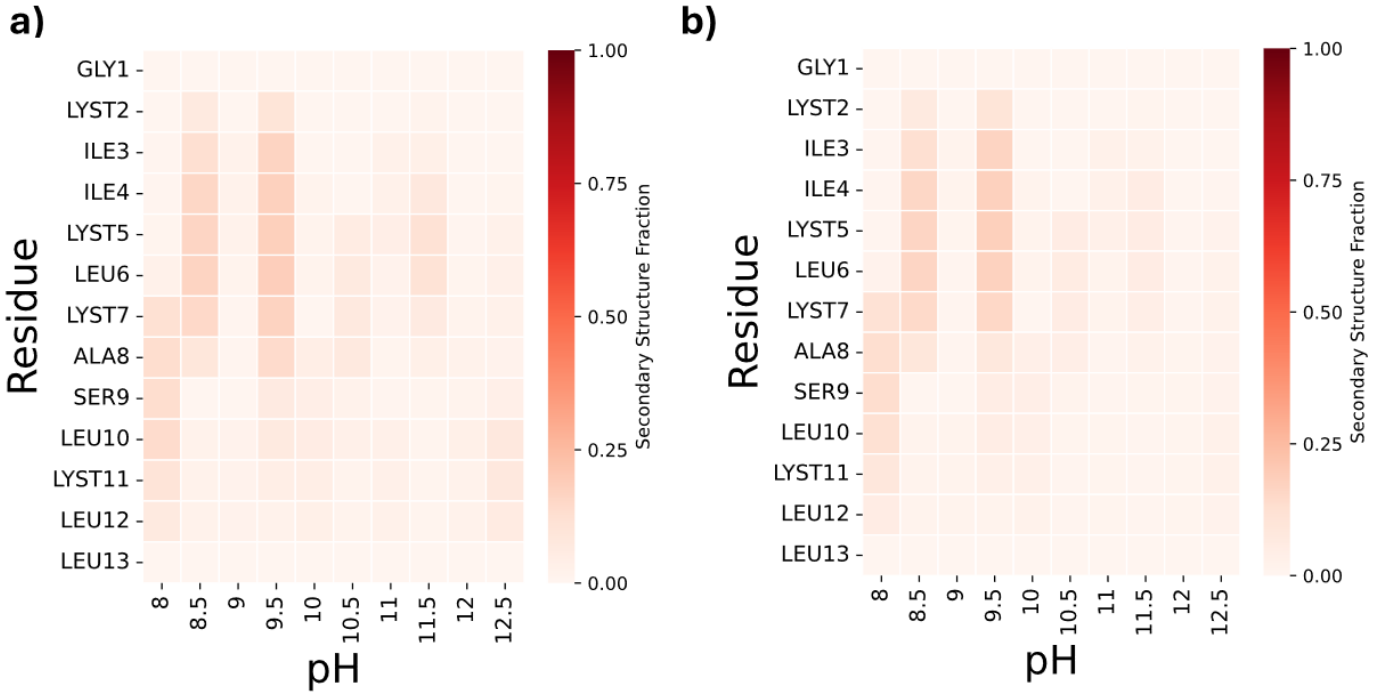

Fig. S5 (a) Average helical structure per residue over 10 pH levels. (b) Error value per residue over 10 pH levels.

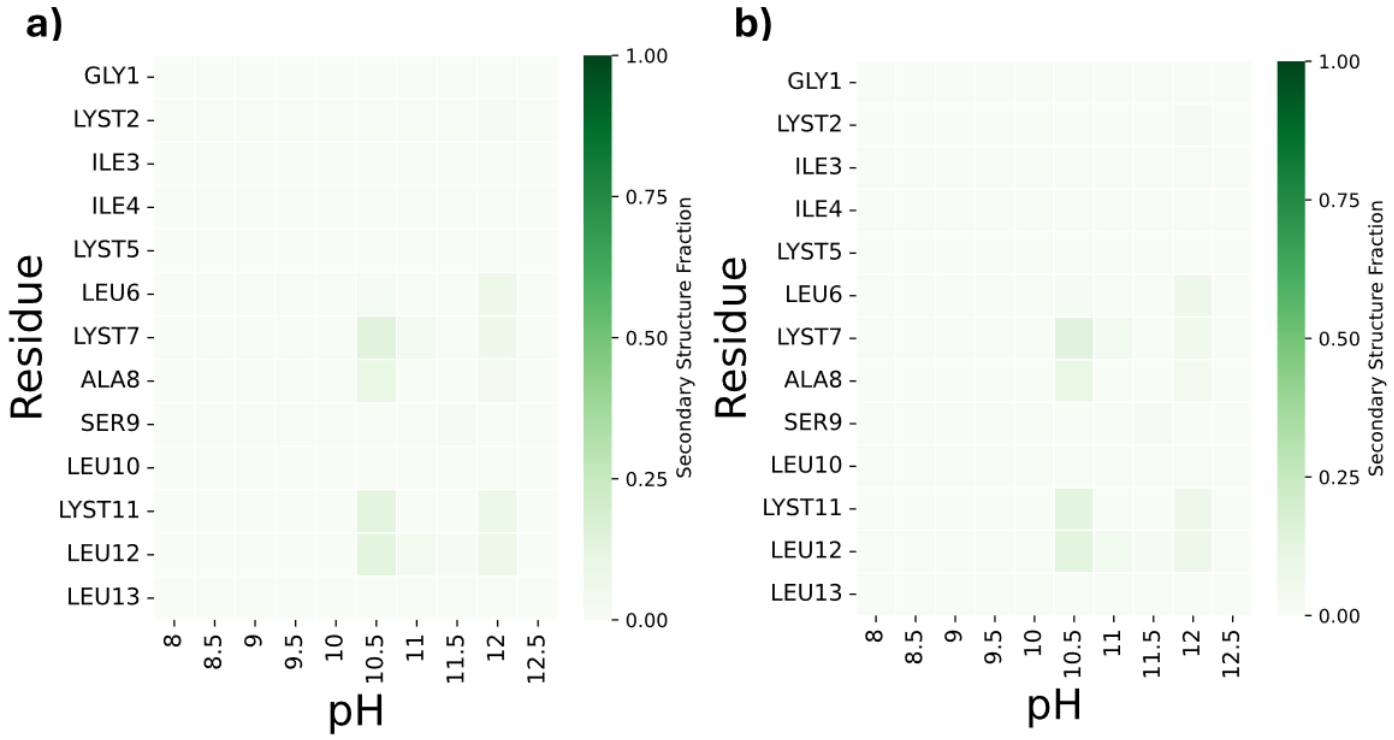

Fig. S6 (a) Average strand structure per residue over 10 pH levels. (b) Error value per residue over 10 pH levels.

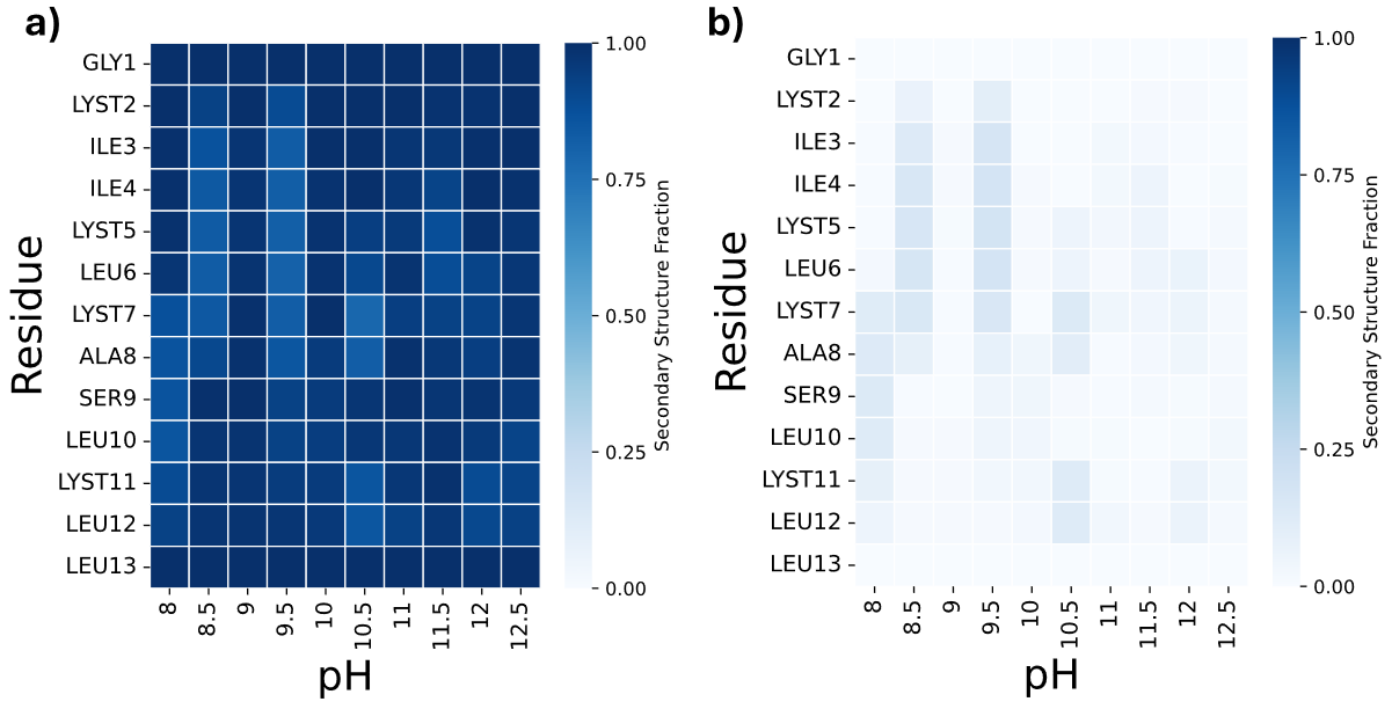

Fig. S7 (a) Average coil structure per residue over 10 pH levels. (b) Error value per residue over 10 pH levels.

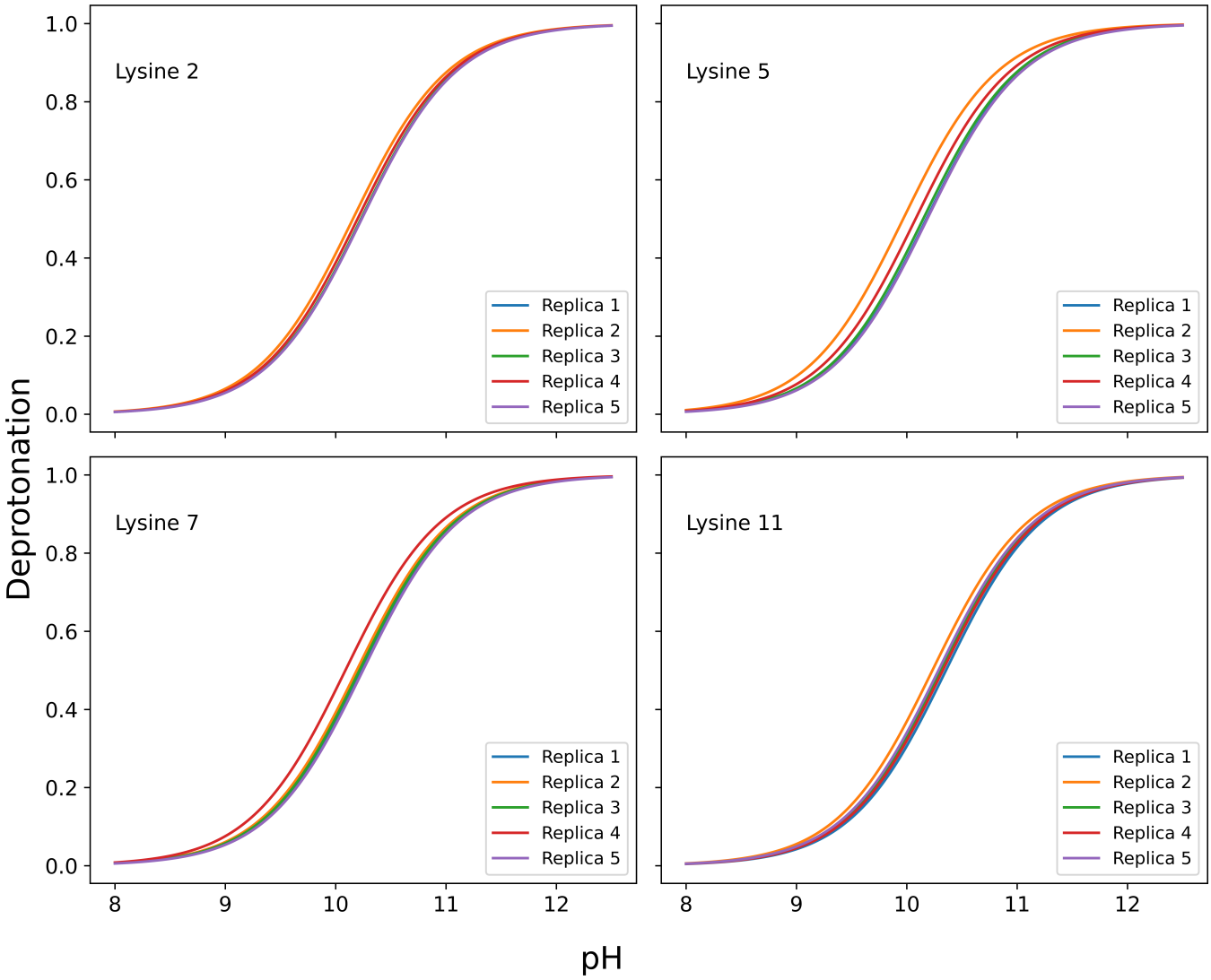

Fig. S8 Calculated pK_a values of each lysine amino acid for each replica by fitting the deprotonated fraction into the Henderson-Hasselbach equation described in the methods section.

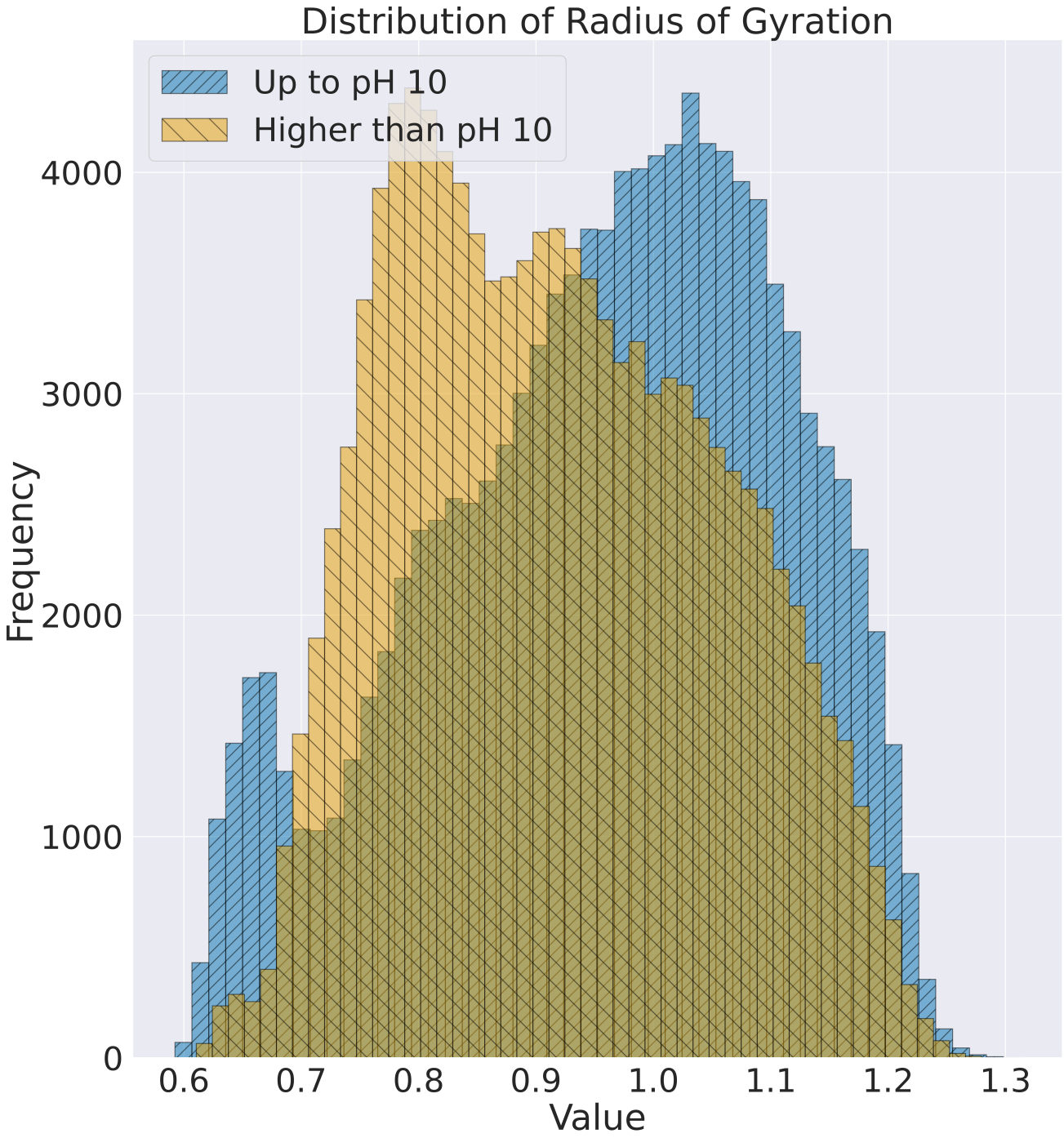

Fig. S9 Distributions of radius of gyration for systems with pH value up to 10 and higher than 10.

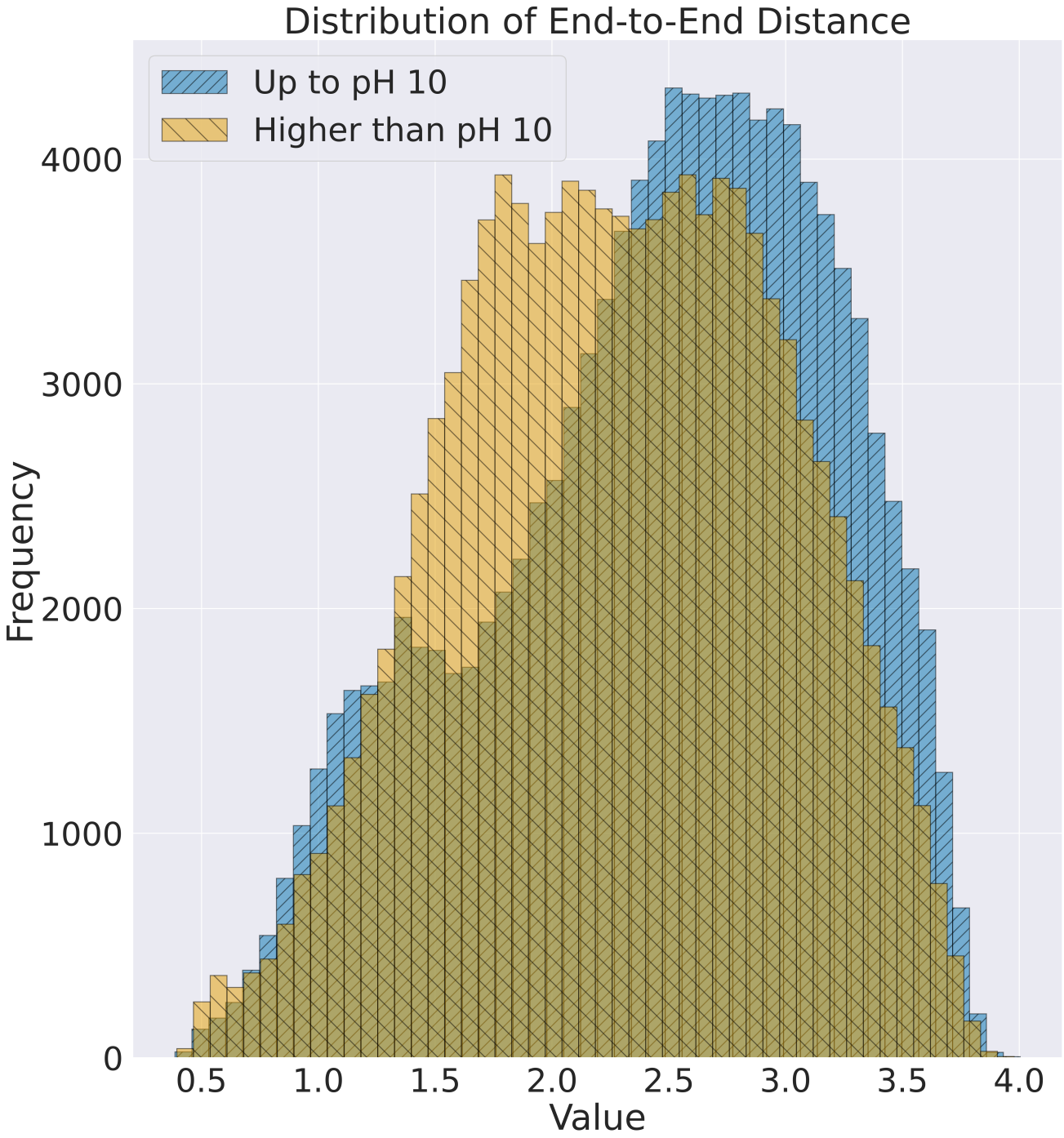

Fig. S10 Distributions of End-to-End distances for systems with pH value up to 10 and higher than 10.

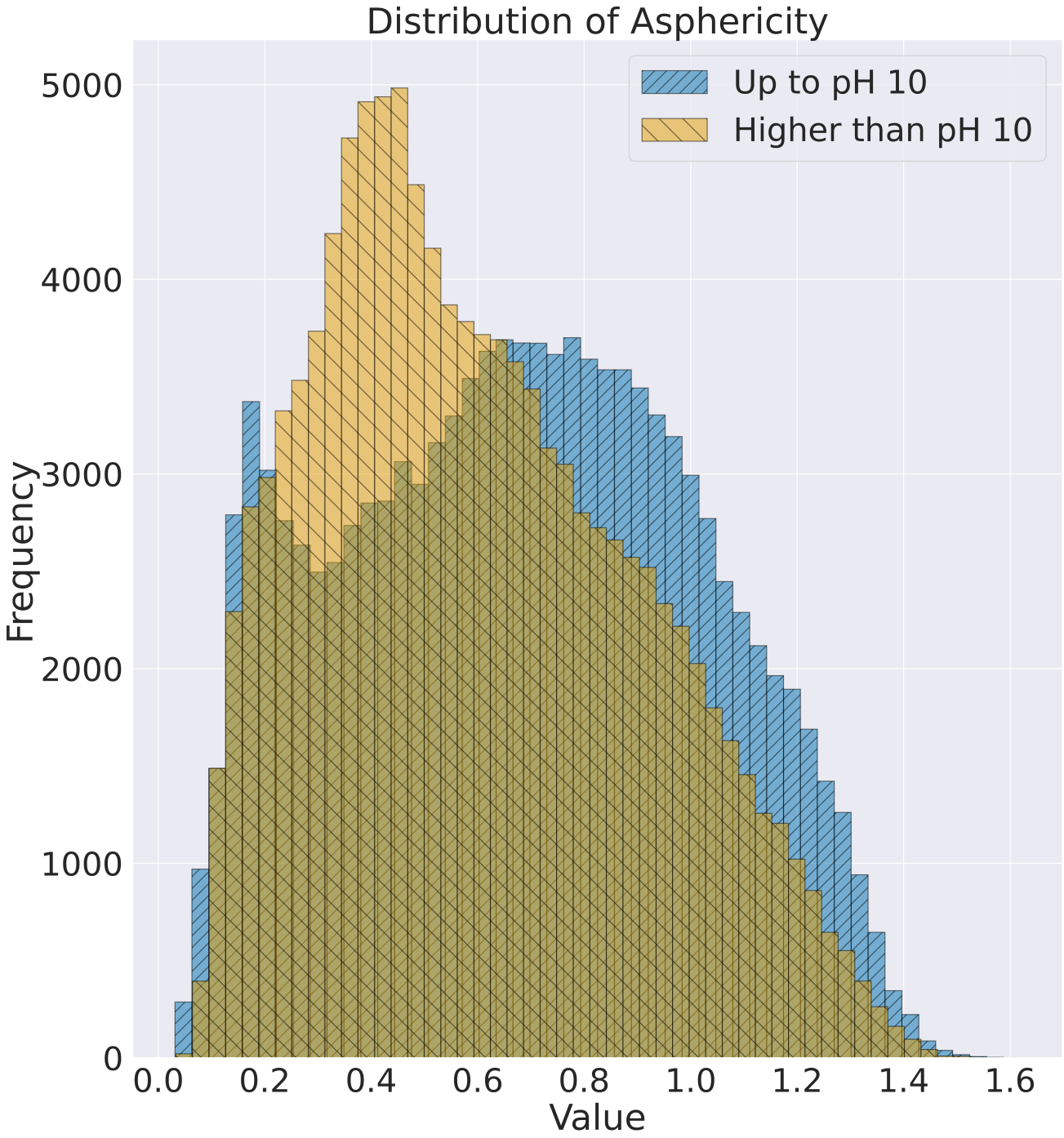

Fig. S11 Distributions of asphericity for systems with pH value up to 10 and higher than 10.

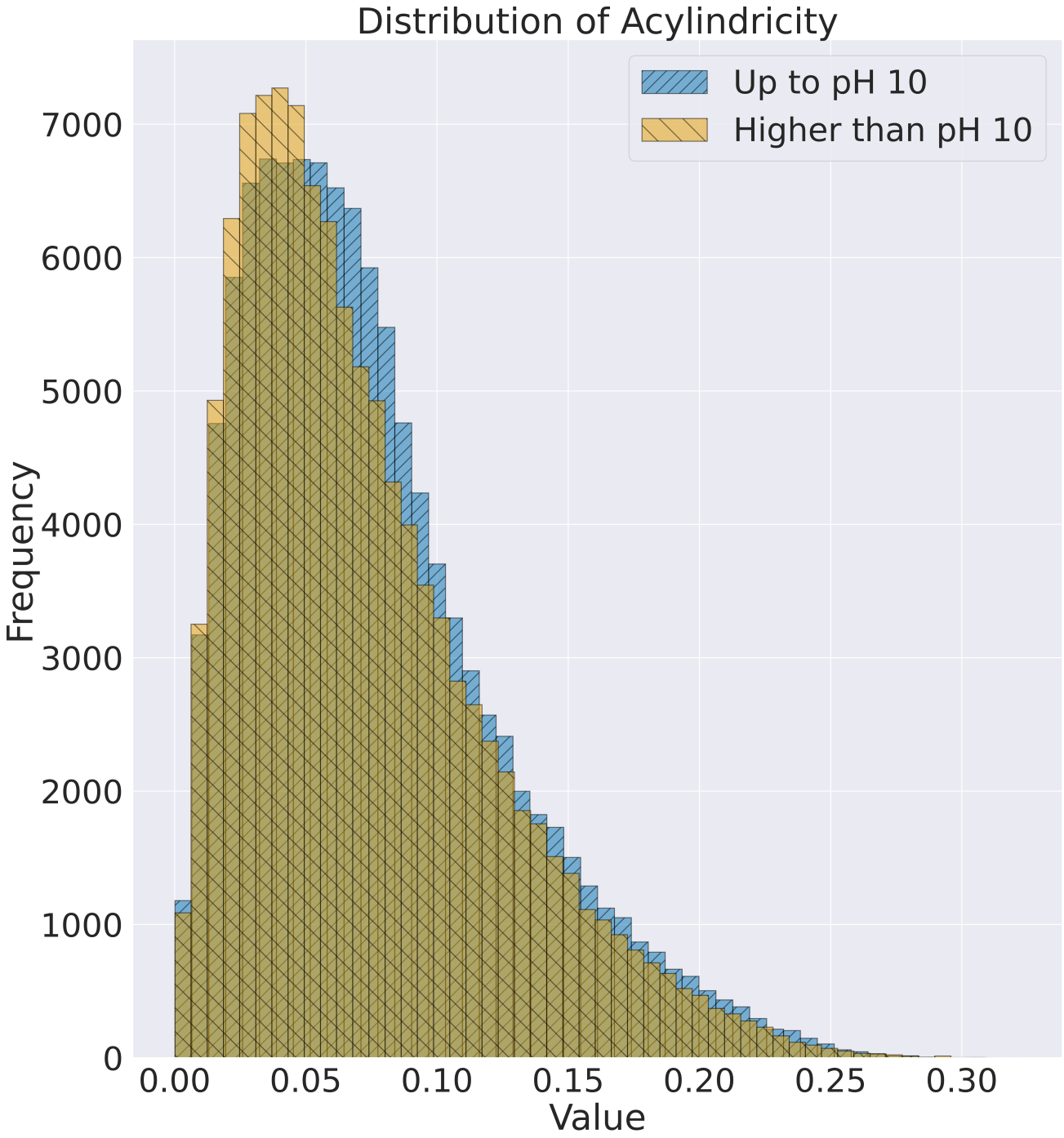

Fig. S12 Distributions of acylindricity for systems with pH value up to 10 and higher than 10.

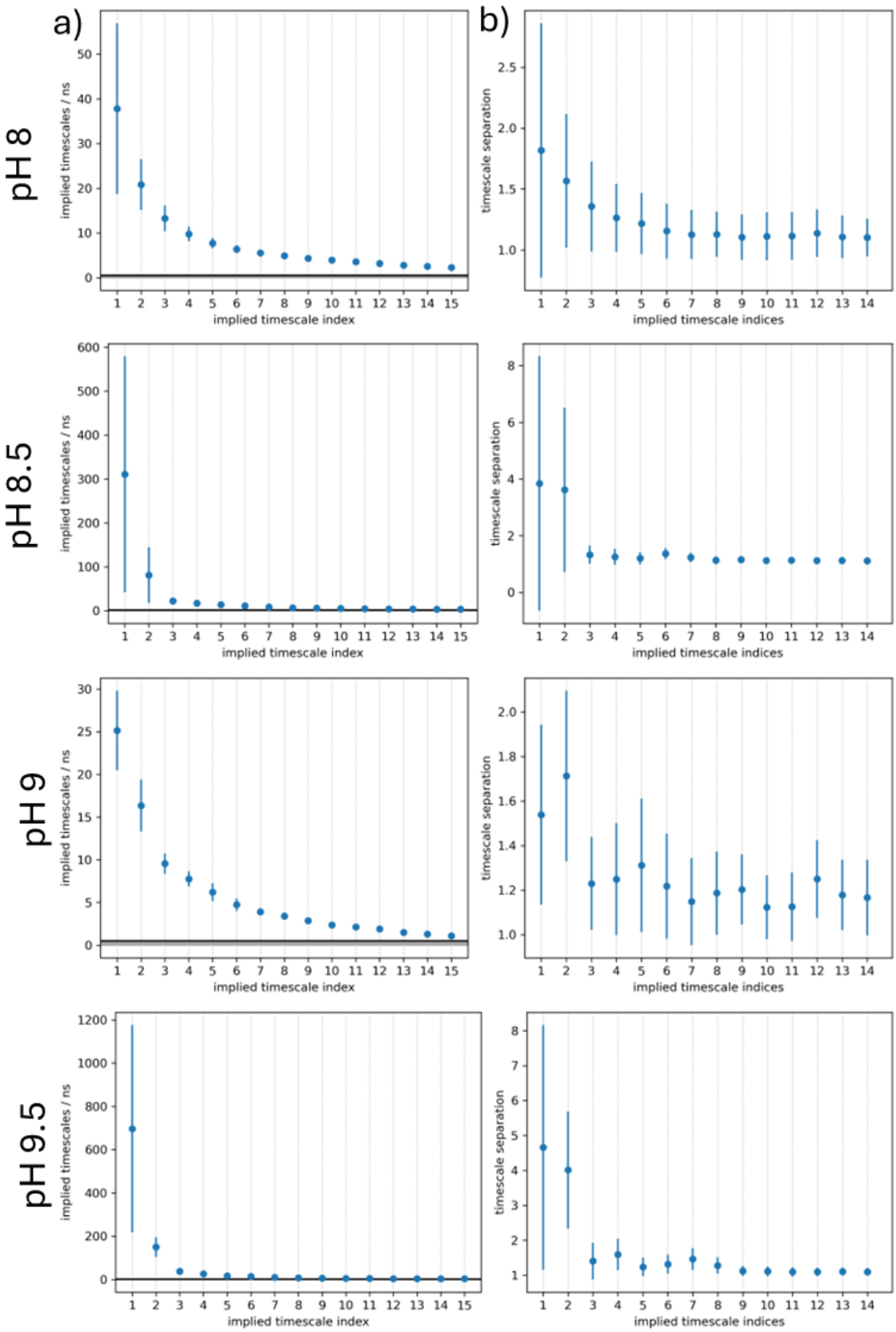

Fig. S13 a) Implied timescale plot using TICA dimensional reduction method, showing relaxation times (y-axis) of dynamic processes versus their indices (x-axis) for varying lag times, with plateaus indicating a Markovian model, under pH conditions of 8 to 9.5. (b) Timescale separation plot using TICA, displaying the ratio of consecutive implied timescales (y-axis) against their indices (x-axis), where larger separations highlight distinct slow and fast processes, ensuring a reliable and interpretable Markov State Model at pH 8 to 9.5.

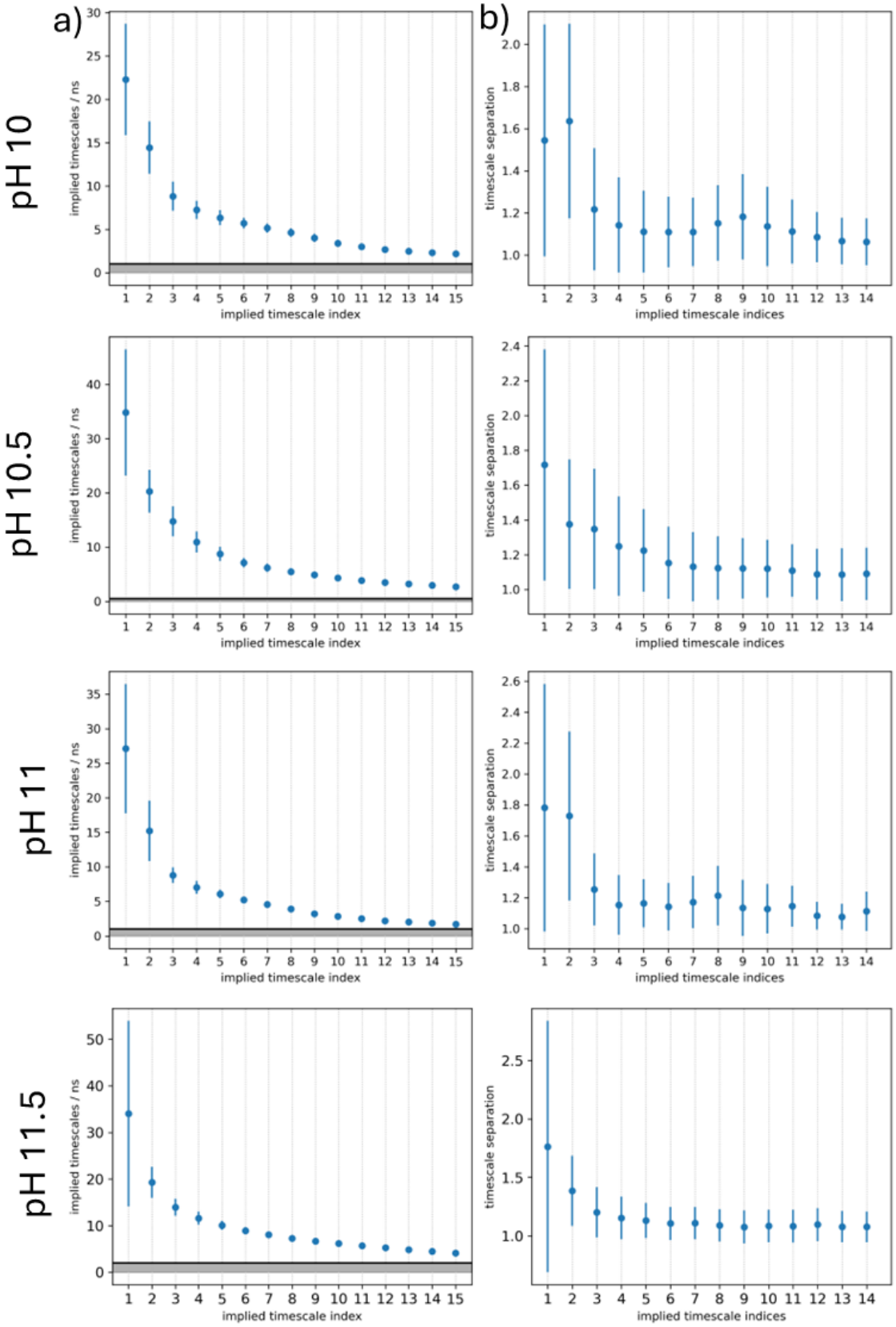

Fig. S14 a) Implied timescale plot using TICA dimensional reduction method, showing relaxation times (y-axis) of dynamic processes versus their indices (x-axis) for varying lag times, with plateaus indicating a Markovian model, under pH conditions of 10 to 11.5. (b) Timescale separation plot displaying the ratio of consecutive implied timescales (y-axis) against their indices (x-axis), where larger separations highlight distinct slow and fast processes, ensuring a reliable and interpretable Markov State Model at pH 10 to 11.5.

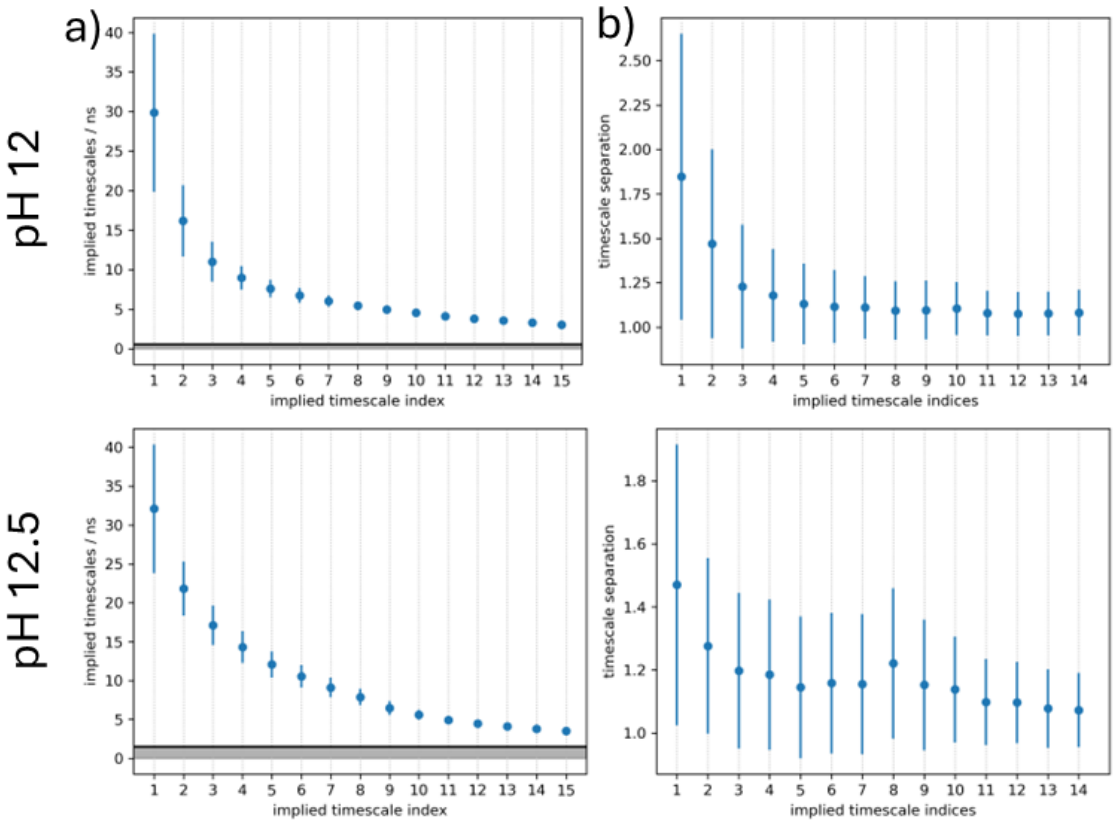

Fig. S15 a) Implied timescale plot using TICA dimensional reduction method, showing relaxation times (y-axis) of dynamic processes versus their indices (x-axis) for varying lag times, with plateaus indicating a Markovian model, under pH conditions of 12 and 12.5. (b) Timescale separation plot displaying the ratio of consecutive implied timescales (y-axis) against their indices (x-axis), where larger separations highlight distinct slow and fast processes, ensuring a reliable and interpretable Markov State Model at pH 12 and 12.5.

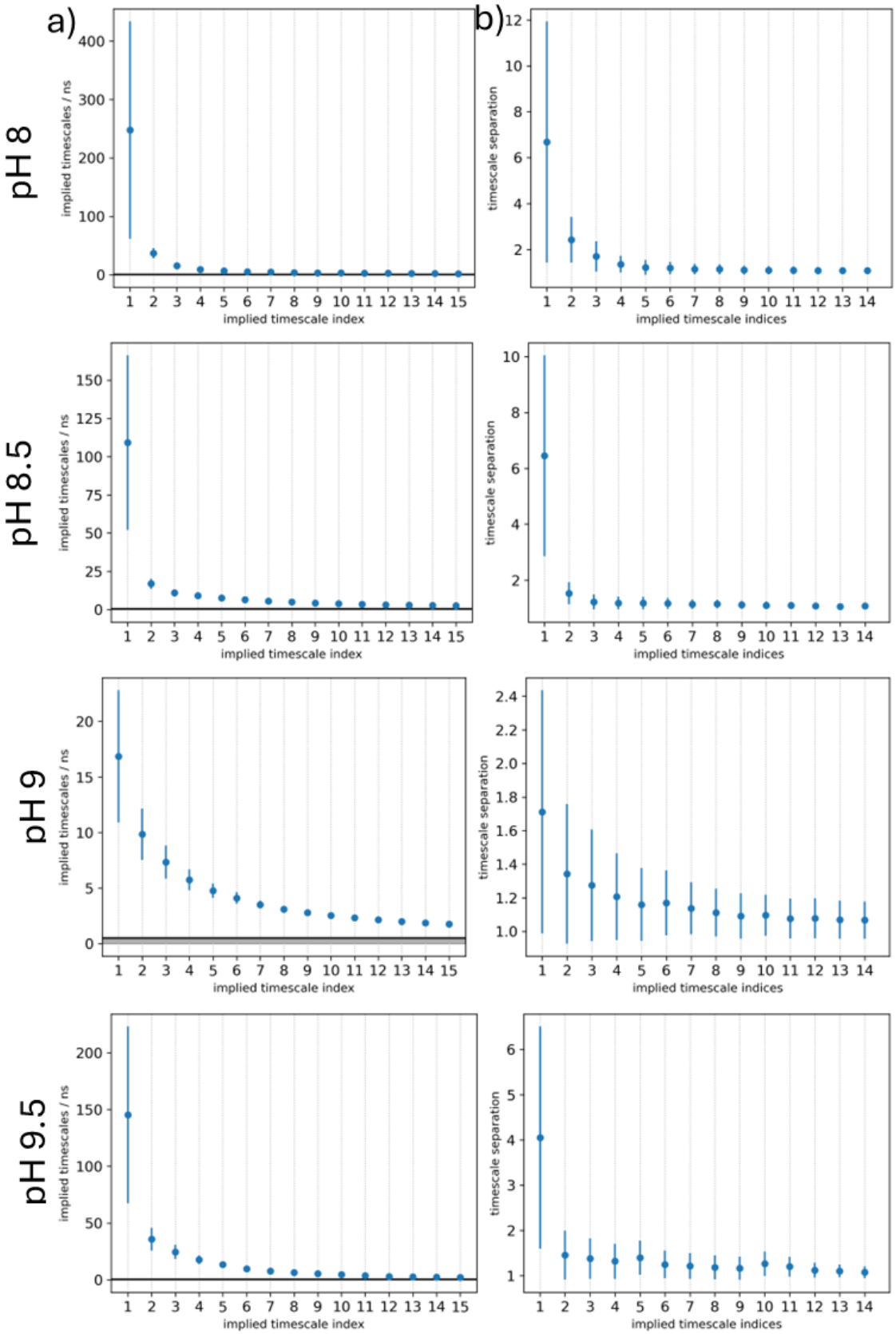

Fig. S16 a) Implied timescale plot using PCA dimensional reduction method, showing relaxation times (y-axis) of dynamic processes versus their indices (x-axis) for varying lag times, with plateaus indicating a Markovian model, under pH conditions of 8 to 9.5. (b) Timescale separation plot displaying the ratio of consecutive implied timescales (y-axis) against their indices (x-axis), where larger separations highlight distinct slow and fast processes, ensuring a reliable and interpretable Markov State Model at pH 8 to 9.5.

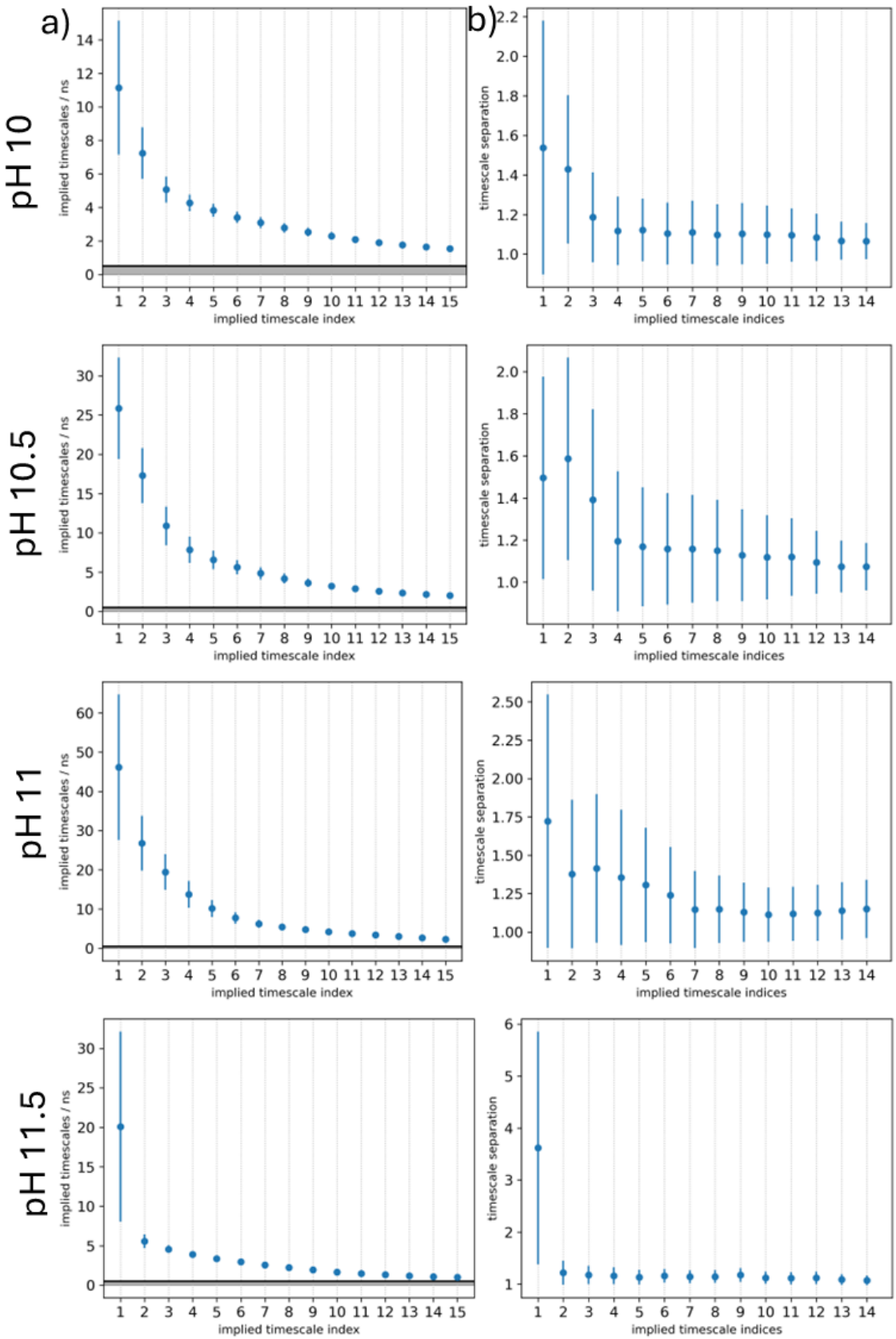

Fig. S17 a) Implied timescale plot using PCA dimensional reduction method, showing relaxation times (y-axis) of dynamic processes versus their indices (x-axis) for varying lag times, with plateaus indicating a Markovian model, under pH conditions of 10 to 11.5. (b) Timescale separation plot displaying the ratio of consecutive implied timescales (y-axis) against their indices (x-axis), where larger separations highlight distinct slow and fast processes, ensuring a reliable and interpretable Markov State Model at pH 10 to 11.5.

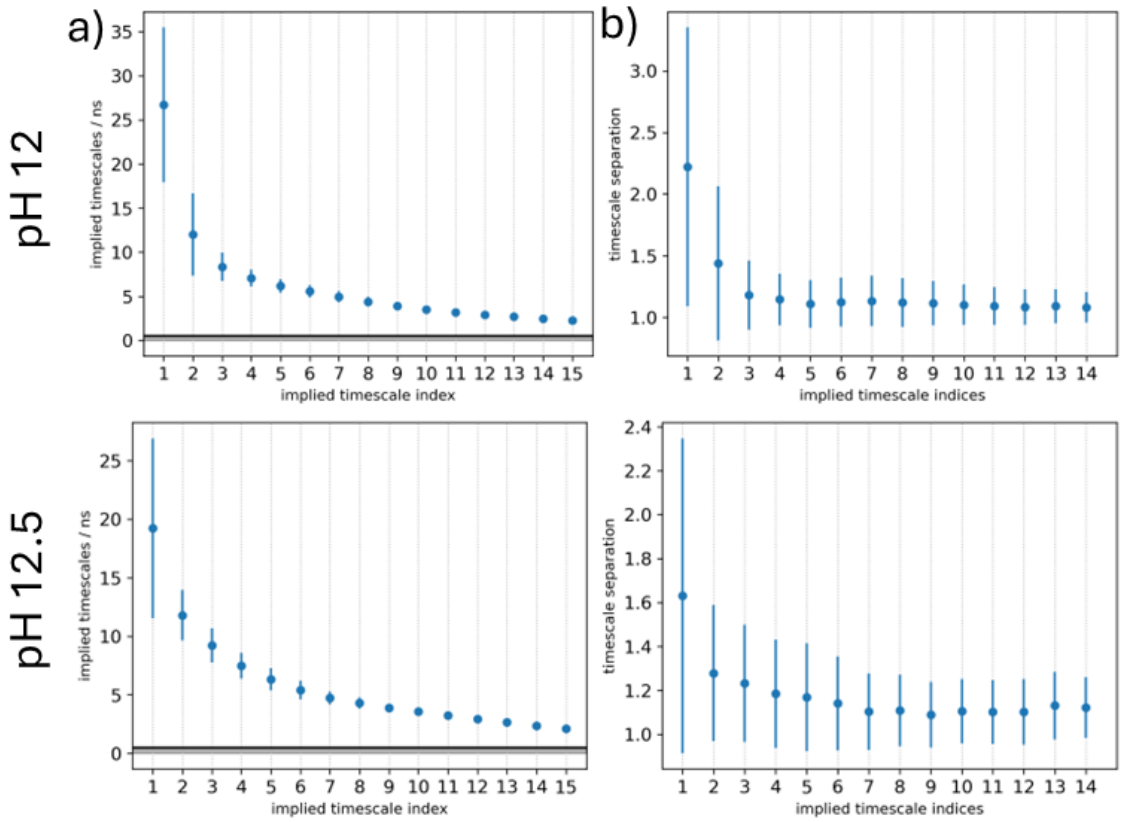

Fig. S18 a) Implied timescale plot using PCA dimensional reduction method, showing relaxation times (y-axis) of dynamic processes versus their indices (x-axis) for varying lag times, with plateaus indicating a Markovian model, under pH conditions of 12 and 12.5. (b) Timescale separation plot displaying the ratio of consecutive implied timescales (y-axis) against their indices (x-axis), where larger separations highlight distinct slow and fast processes, ensuring a reliable and interpretable Markov State Model at pH 12 and 12.5.

(10.5), the helical structure after the random coil is preferred by the peptide; however, by increasing the pH and passing the experimental pK_a value of the lysine side chain, the helical structure fraction decreases significantly. Above pH 10.5, more than 95% of population states at each pH level are occupied by random coil structures. To check the validity of MSM results, we constructed a new MSM model using PCA transformation fitted on all the data across 10 pH levels in figures S25, and S25. We observed that PCA is not able to separate stable states very well, especially in pH levels of 12 and 12.5.

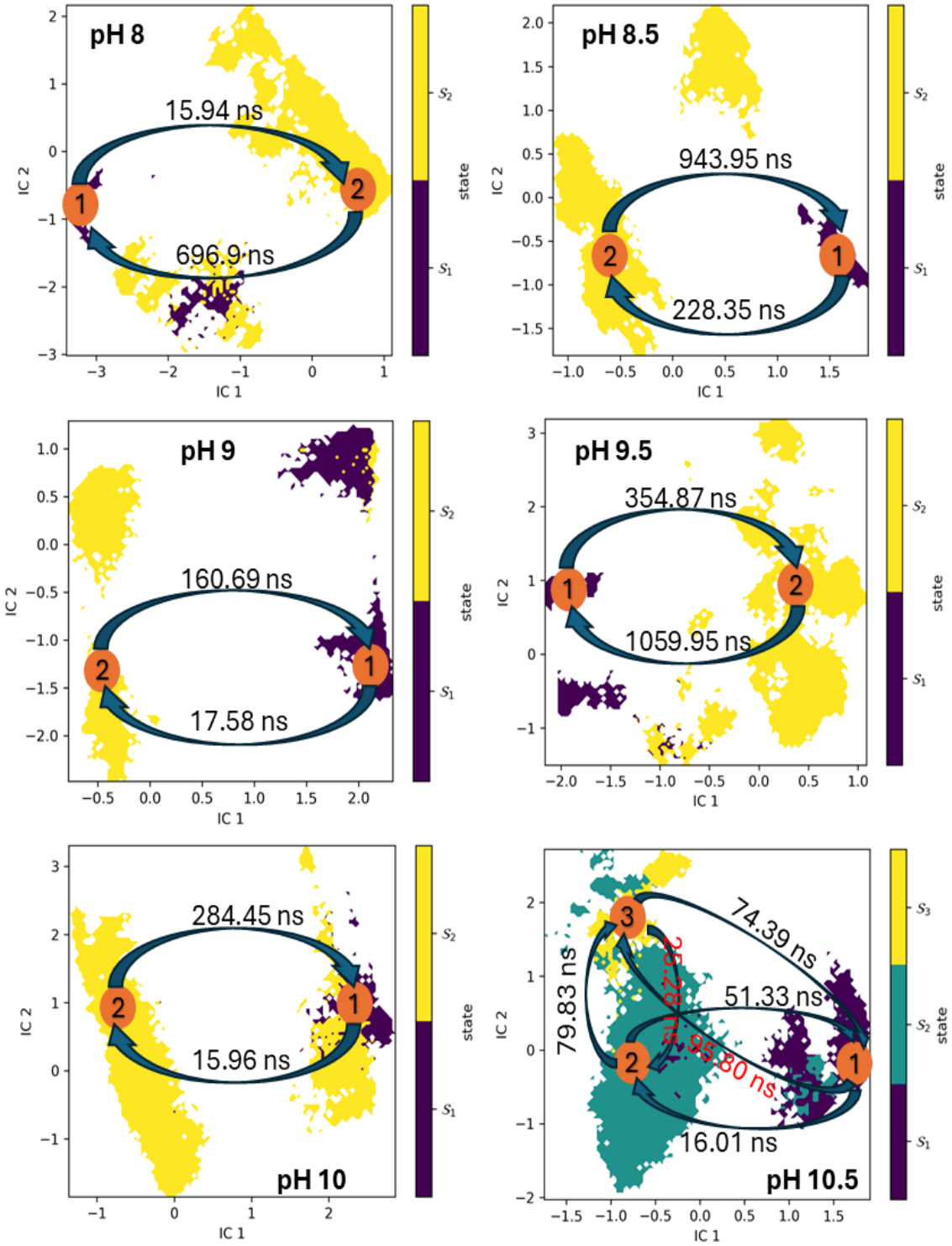

Fig. S19 Free Energy Surface (FES) of the system under pH level of 8 to 10.5 computed using TICA as dimensionality reduction method, colored by metastable states, with mean first passage times (MFPTs) indicated for transitions between states.

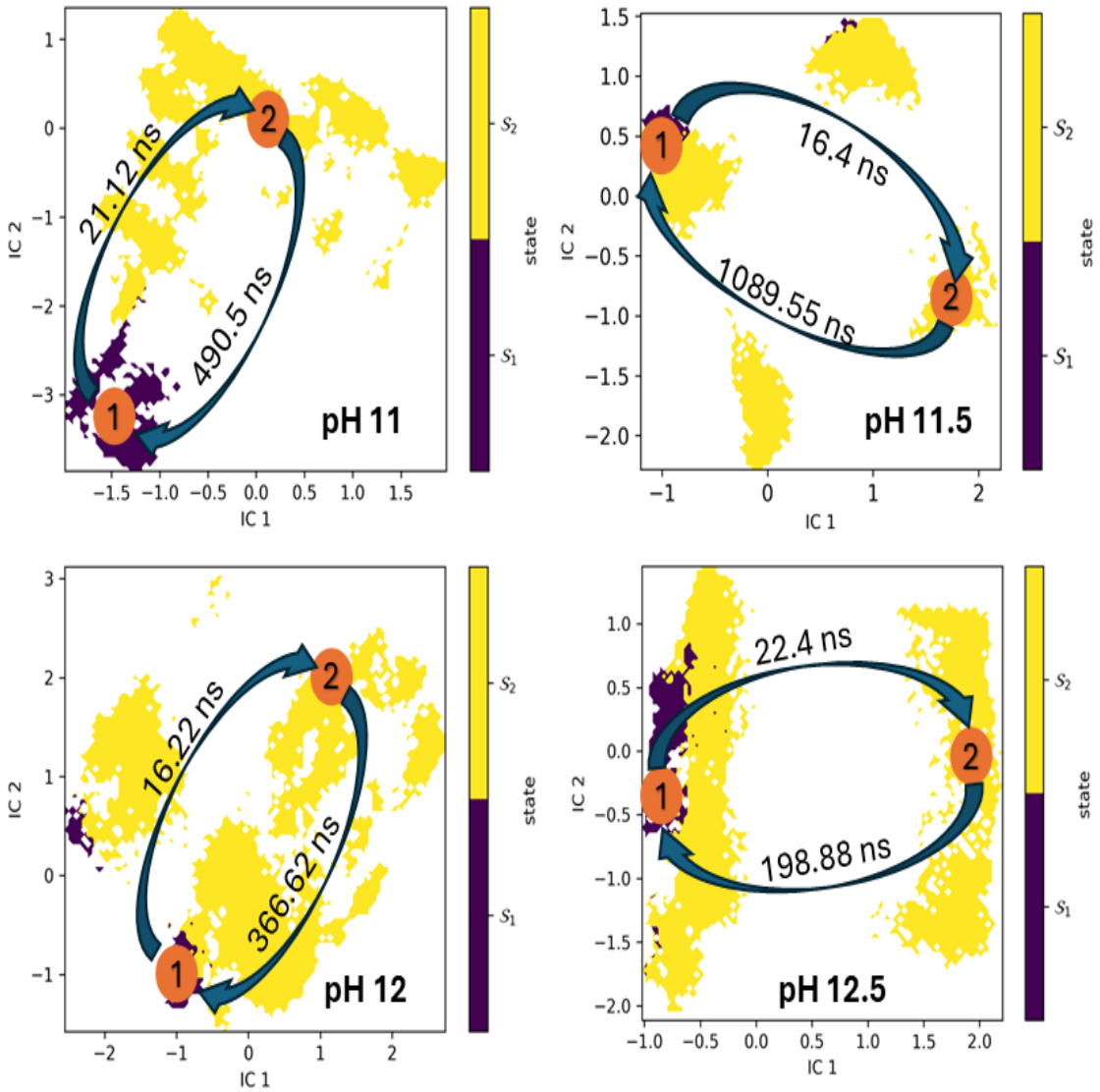

Fig. S20 Free Energy Surface (FES) of the system under pH level of 11 to 12.5 computed using TICA as dimensionality reduction method, colored by metastable states, with mean first passage times (MFPTs) indicated for transitions between states.

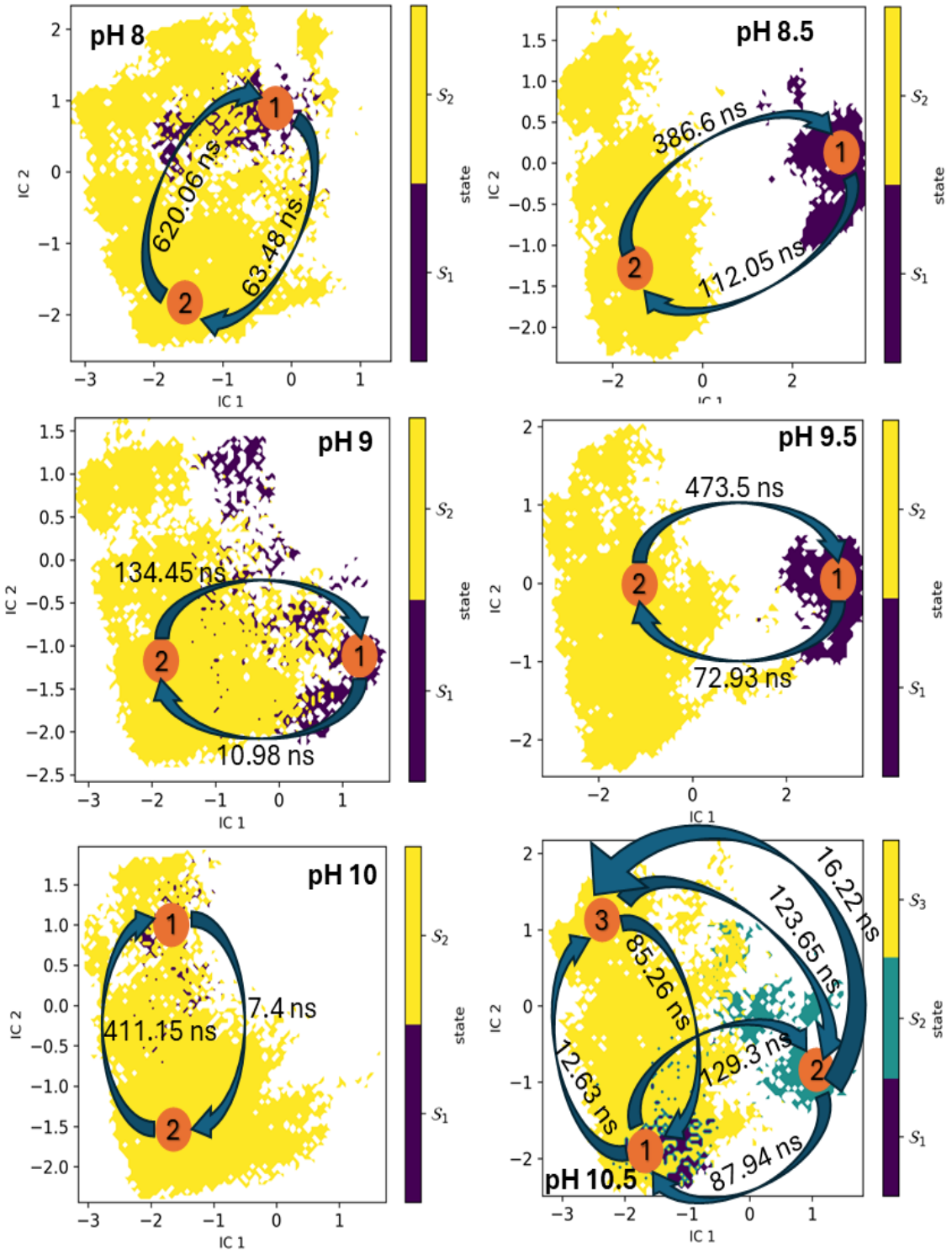

Fig. S21 Free Energy Surface (FES) of the system under pH level of 8 to 10.5 computed using PCA as dimensionality reduction method, colored by metastable states, with mean first passage times (MFPTs) indicated for transitions between states.

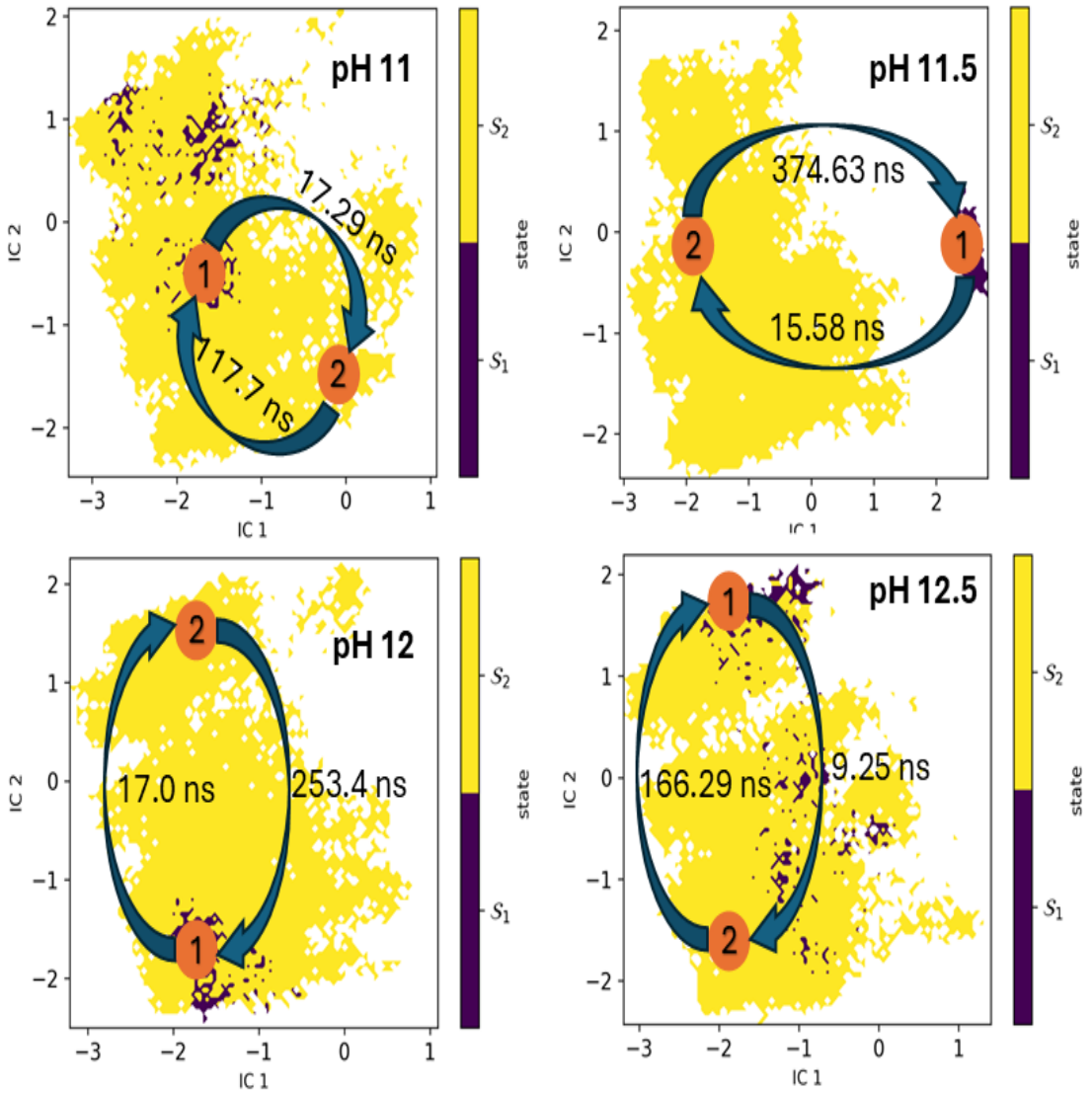

Fig. S22 Free Energy Surface (FES) of the system under pH level of 11 to 12.5 computed using PCA as dimensionality reduction method, colored by metastable states, with mean first passage times (MFPTs) indicated for transitions between states.

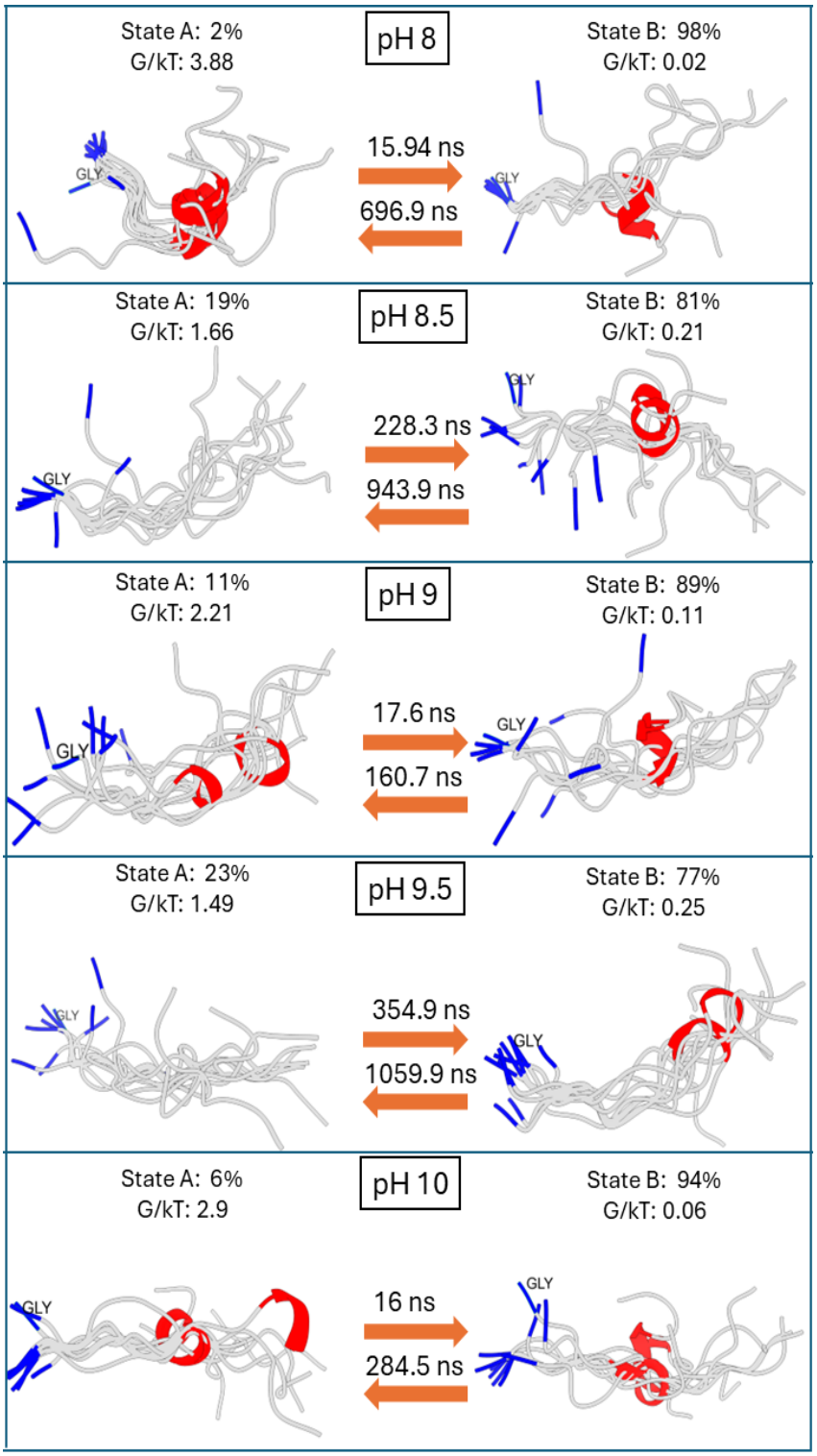

Fig. S23 Metastable state population percentage of GL13K calculated by TICA dimensional reduction method from pH 8 to pH 10 is plotted with the representative structure of each state. These representative structures are based on the probability distributions of microstates identified by PCCA+. The mean first passage time (MFPT) between states is calculated in nanoseconds. The relative free energy for each state is shown under each stable state. The random coil and helical structures are colored in silver and red, respectively. The amino acids colored in blue represent the GLY and N-terminus.

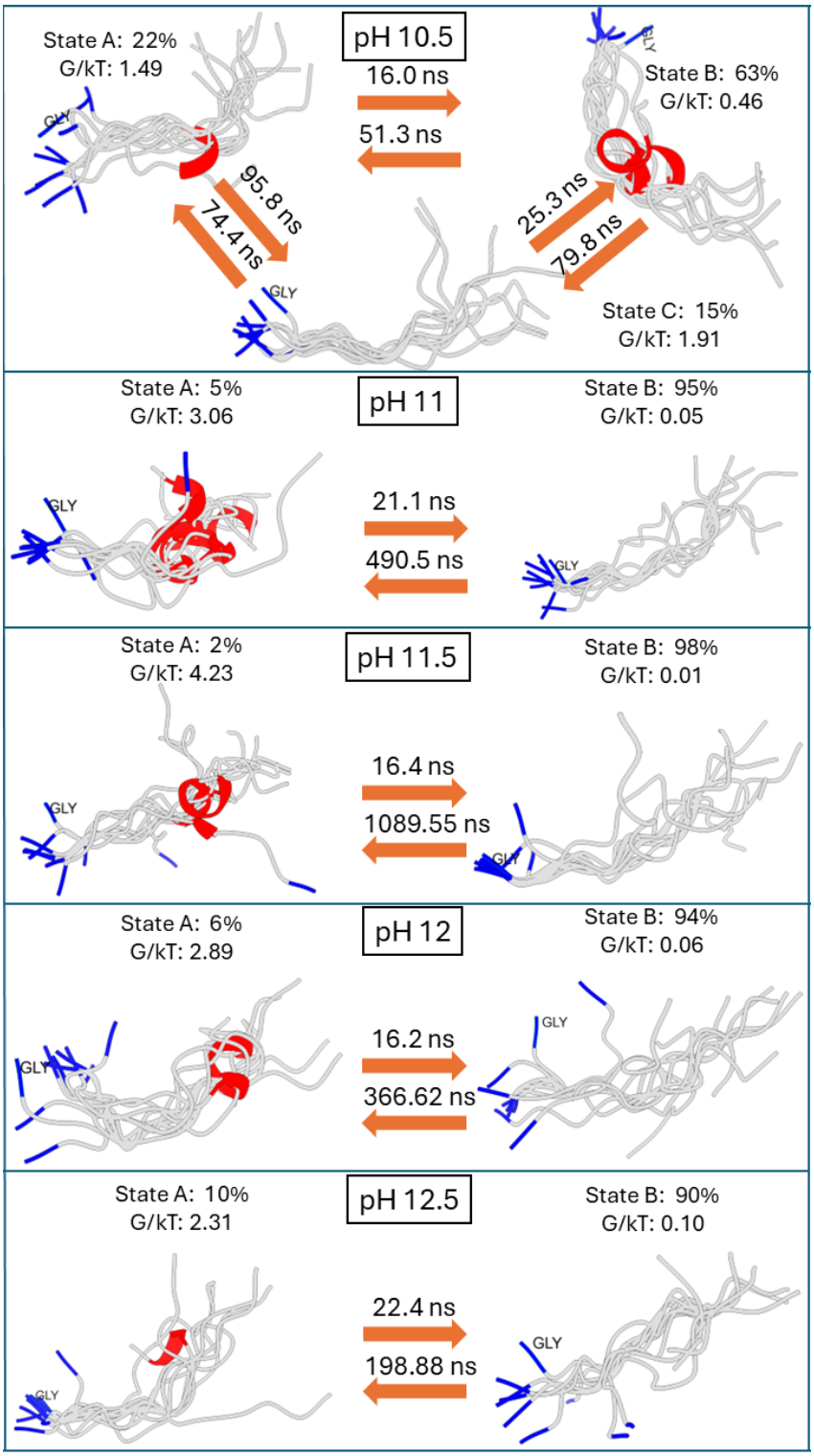

Fig. S24 Metastable state population percentage of GL13K calculated by TICA dimensional reduction method from pH 10.5 to pH 12.5 is plotted with the representative structure of each state. These representative structures are based on the probability distributions of microstates identified by PCCA+. The mean first passage time (MFPT) between states is calculated in nanoseconds. The relative free energy for each state is shown under each stable state. The random coil and helical structures are colored in silver and red, respectively. The amino acids colored in blue represent the GLY and N-terminus.

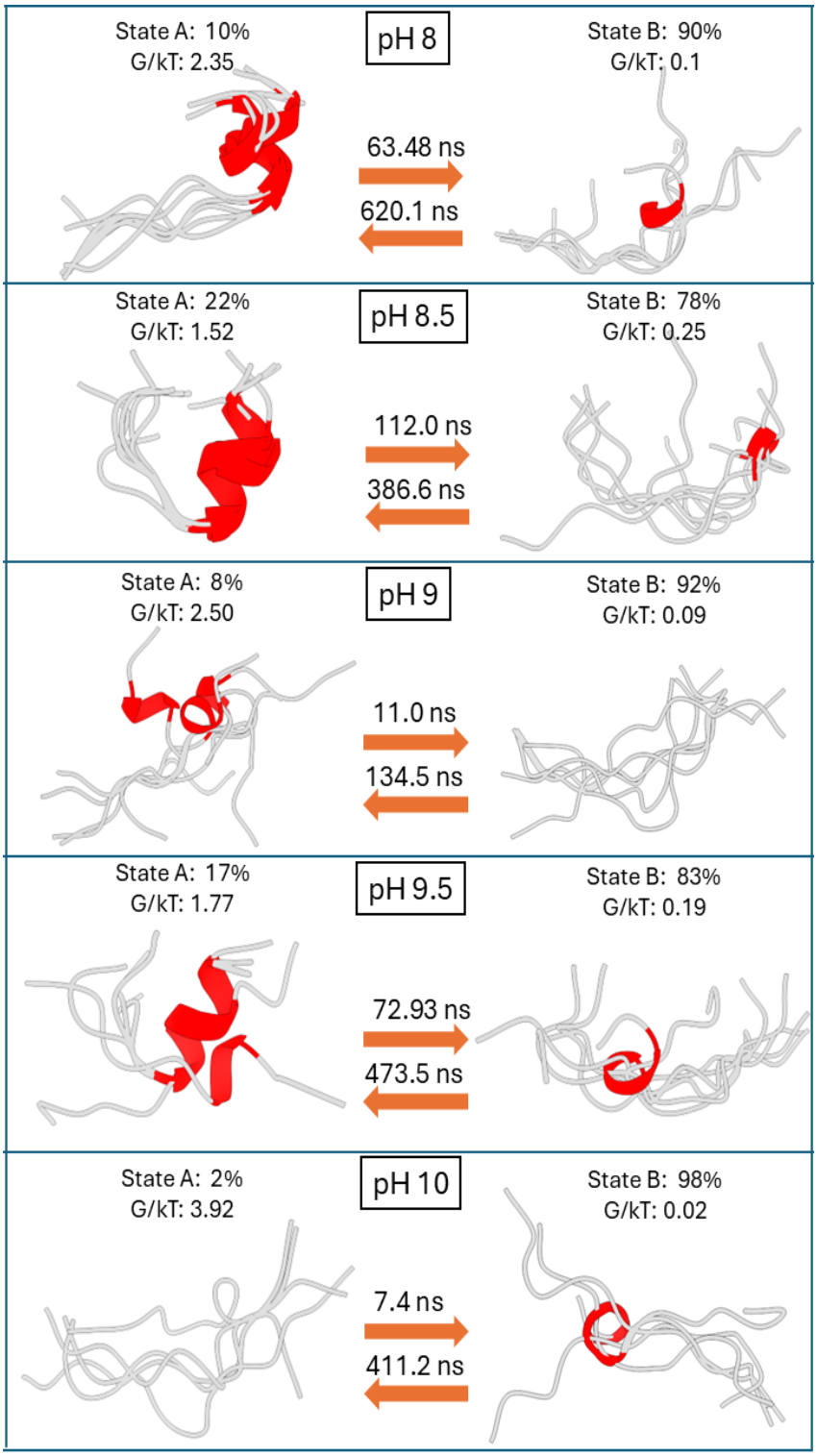

Fig. S25 Metastable state population percentage of GL13K calculated by PCA dimensional reduction method from pH 8 to pH 10 is plotted with the representative structure of each state. These representative structures are based on the probability distributions of microstates identified by PCCA+. The mean first passage time (MFPT) between states is calculated in nanoseconds. The relative free energy for each state is shown under each stable state. The random coil and helical structures are colored in silver and red, respectively.

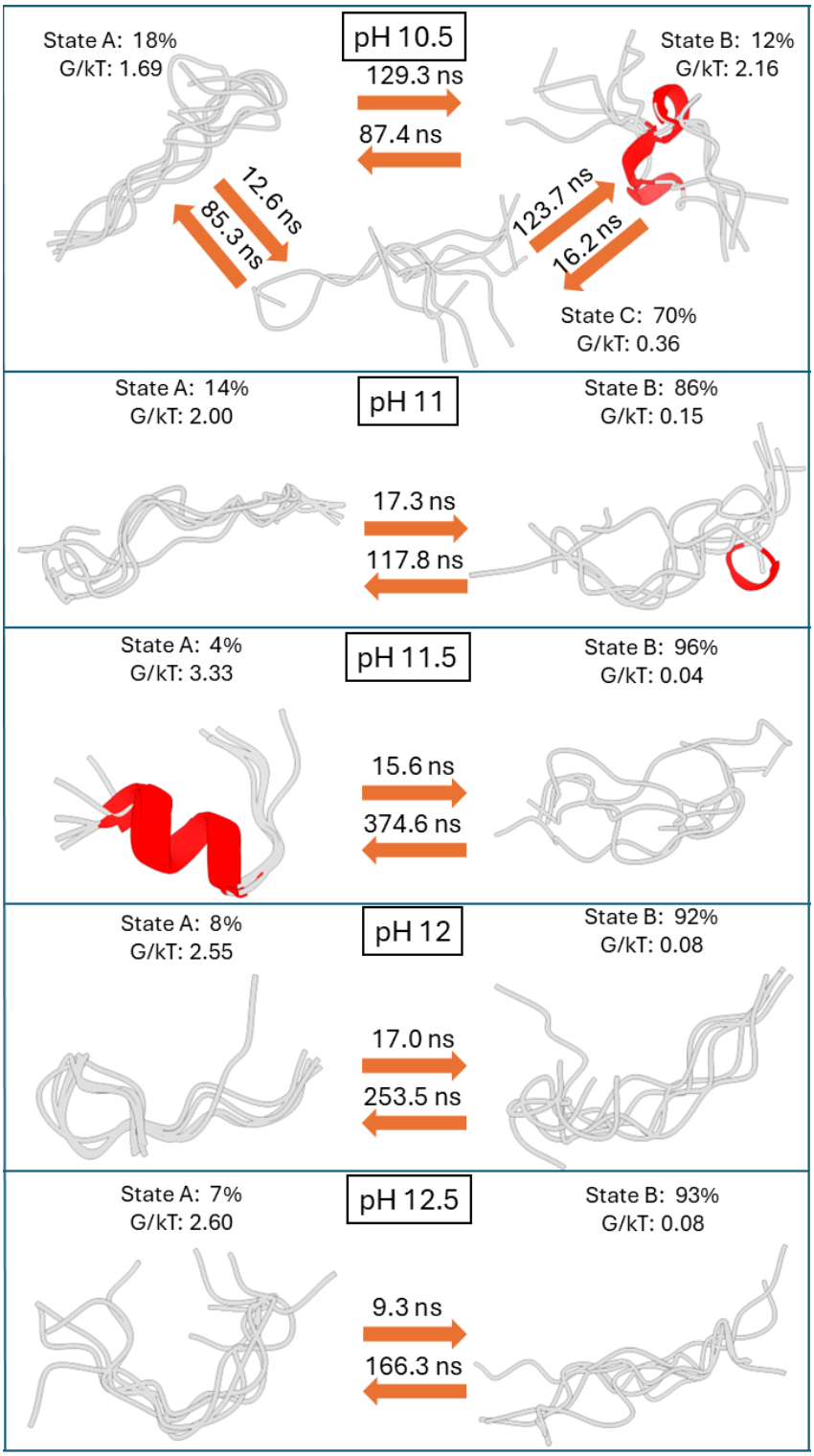

Fig. S26 Metastable state population percentage of GL13K calculated by PCA dimensional reduction method from pH 10.5 to pH 12.5 is plotted with the representative structure of each state. These representative structures are based on the probability distributions of microstates identified by PCCA+. The mean first passage time (MFPT) between states is calculated in nanoseconds. The relative free energy for each state is shown under each stable state. The random coil and helical structures are colored in silver and red, respectively.