Quantifying Membrane Structure and Dynamics During Bioproduct Production in *Zymomonas mobilis* by Molecular Simulation

Nitin Kumar Singh[1,2] and Josh V. Vermaas*[1,2,3]

[1]MSU-DOE Plant Research Laboratory, Michigan State University, 612 Wilson Road, East Lansing, MI 48824
[2]DOE Great Lakes Bioenergy Research Center, Michigan State University, 612 Wilson Road, East Lansing, MI 48824
[3]Deparment of Biochemistry and Molecular Biology, Michigan State University, 612 Wilson Road, East Lansing, MI 48824

*Email: vermaasj@msu.edu

Abstract

The conversion of lignocellulosic biomass into biofuels and bioproducts by microbial biorefineries is central to a sustainable chemical industry. *Zymomonas mobilis* is one such biorefinery chassis, and is resistant to ethanol stress, leading to its use in biomass conversion to biofuels and bioproducts. However, *Z. mobilis* growth is often inhibited by organic acids, aldehydes, alcohols, ketones, and amides found in biomass hydrolysate. The resulting slow growth inhibits production, and as a result drives up the price for the resulting products. One hypothesis is that these molecules interact with or disrupt the bacterial membrane, triggering stress responses and hindering growth. To test this hypothesis at the molecular level, we employ all-atom molecular dynamics (MD) simulations to investigate lignocellulose-derived small molecules and their impact on a biologically relevant *Z. mobilis* membrane model. Simulations were conducted across a range of inhibitor concentrations from 0–2.5 mol%, analyzing key membrane properties such as area per lipid (APL), membrane thickness, lipid order parameter ($-S_{CH}$), lateral diffusion coefficient (D_{xy}), and permeability coefficient (Pm). From simulation, we observe altered membrane structure and dynamics at these modest small molecule concentrations commonly found in hydrolysates. Generally, the membranes become thinner, with a higher area per lipid and lower order parameter as the small molecule concentration increases. These trends are stronger for more hydrophobic molecules with greater hydrophobic bulk, as isobutanol, propanol, and propanoic acid showed greater membrane perturbations as the concentration increased compared with other small molecules. Tracking small molecule distributions directly in our equilibrium simulations allows us to determine concentration-dependent free energy profiles for these molecules. While the trends are noisy, generally the barriers to crossing the membrane decrease as the concentration increases, indicating that the membranes become leakier as small molecule concentrations rise. Comparing between native *Z. mobilis* membranes with hopanoids and membranes sharing the same phospholipid composition but without hopanoids, hopanoids stabilize and order the membrane for smaller molecules to maintain membrane structure, but appear insufficient for larger hydrophobic molecules like isobutanol. These findings provide a mechanistic understanding of how small molecules found in biomass degradation streams interact with the *Z. mobilis* membrane, offering valuable insights for future strain engineering efforts to optimize biofuel and bioproduct synthesis from biomass feedstocks by highlighting limits to small molecule tolerance. This knowledge can guide the modification of membrane composition to develop more robust microbes, thereby improving microbial survival and yields in industrial contexts.

Keywords

Zymomonas mobilis; Molecular dynamics; Biofuel; Lignocellulosic inhibitor; Biomass

Introduction

Global efforts towards sustainable and renewable energy and materials have intensified research efforts into converting lignocellulosic biomass from plant or microbial sources, including agricultural or forestry wastes, into biofuels and bioproducts.[1–3] These plant residues are first mechanically and/or chemically pretreated to yield hydrolysates that are a mix of many compounds. Rather than depend on separations technologies to feed individual waste streams, the biorefinery concept[4,5] uses the flexible metabolic pathways found in microbes to biologically funnel these small molecules into specific fuel or product chemicals of commercial interest. However, microbial conversion is challenged by the presence of inhibitory compounds generated during pretreatment and hydrolysis.[6–11] These inhibitors, such as furan aldehydes (e.g., furfural, 5-hydroxymethylfurfural (HMF)), weak organic acids (e.g., acetic acid), and phenolic compounds, significantly hinder the growth and metabolic activity of fermenting microorganisms.[6,12]

Among potential biorefinery platform microorganisms, *Zymomonas mobilis*,[13–17] a natural ethanologenic bacterium distinguished by its high ethanol productivity, ethanol tolerance, and unique Entner-Doudoroff (ED) pathway, has emerged as a leading candidate for industrial biofuel production.[18–22] Recent studies have established *Z. mobilis* as a model organism for biofuel synthesis due to its streamlined genome,[23] well-characterized metabolic network, and genetic tractability.[24] The ED pathway, which operates under anaerobic conditions, enables near theoretical ethanol yields with minimal biomass production,[25] making *Z. mobilis* exceptionally efficient compared to traditional yeast systems.[20,26] Despite its advantages and tolerance to some solvent stressors, *Z. mobilis* remains highly sensitive to lignocellulosic hydrolysate inhibitors, which disrupt cellular integrity, impair enzymatic activity, and suppress sugar utilization, ultimately limiting scaling[27–30] for fuel compounds like ethanol[31] or isobutanol.[32,33] These compounds induce oxidative stress, destabilize membranes, and inhibit glycolysis, with synergistic effects exacerbating toxicity in complex hydrolysates.[31]

While genetics approaches and wide scale screenings can improve resistance to specific small molecules,[34–36] the mechanism by which these small molecules interfere with growth remains unclear, which hinders directed bioengineering to solve these growth and production bottlenecks in *Z. mobilis*. One potential mechanistic hypothesis is that some small molecules alter the membrane structure or dynamics in *Z. mobilis* at low concentrations, triggering stress responses or potentially inducing membrane rupture, analogous to what has been observed for other microbes.[37,38] Molecular simulation through classical molecular dynamics (MD) simulations offer a powerful tool to test the impact of small molecules on membrane structure or dynamics, providing high resolution, time resolved insights into the structural and functional changes induced by inhibitors.[39] By simulating the interactions between *Z. mobilis* membranes and inhibitory compounds, MD simulations enable us to explore binding affinities, membrane perturbations, and potential mechanisms of tolerance that may be difficult to directly probe through other means.

In this study, we employ MD simulations to investigate the molecular interactions of *Z. mobilis* membrane models with key lignocellulosic inhibitors, including ethanol, furfural, HMF, and acetic acid (Figure 1). Our goal is to elucidate the structural and dynamic basis of inhibitor tolerance and identify potential targets for strain improvement. By leveraging computational modeling, we can directly identify changes to membrane structure and dynamics induced by the presence of low inhibitor concentrations on *Z. mobilis* membranes, highlighting which compounds have the greatest perturbative effect. We can also ask hypothetical questions that are difficult to address experimentally by simulating equivalent membranes without the hopanoid compounds that are so prevalent in *Z. mobilis* membranes. These findings can inform future engineering efforts aimed at developing more resilient *Z. mobilis* strains for efficient microbial biorefinery production from lignocellulosic biomass.

Methods

The general approach is to run classical molecular dynamics simulations for *Z. mobilis* membranes in increasing concentrations of small molecules that might perturb the membrane structure, and would be observed in real hydrolysates. We also simulate a hypothetical membrane model for *Z. mobilis* that lacks any hopenoids, in order to compare and contrast the impact these small molecules have on membrane structure and dynamics when exposed to multiple stresses.

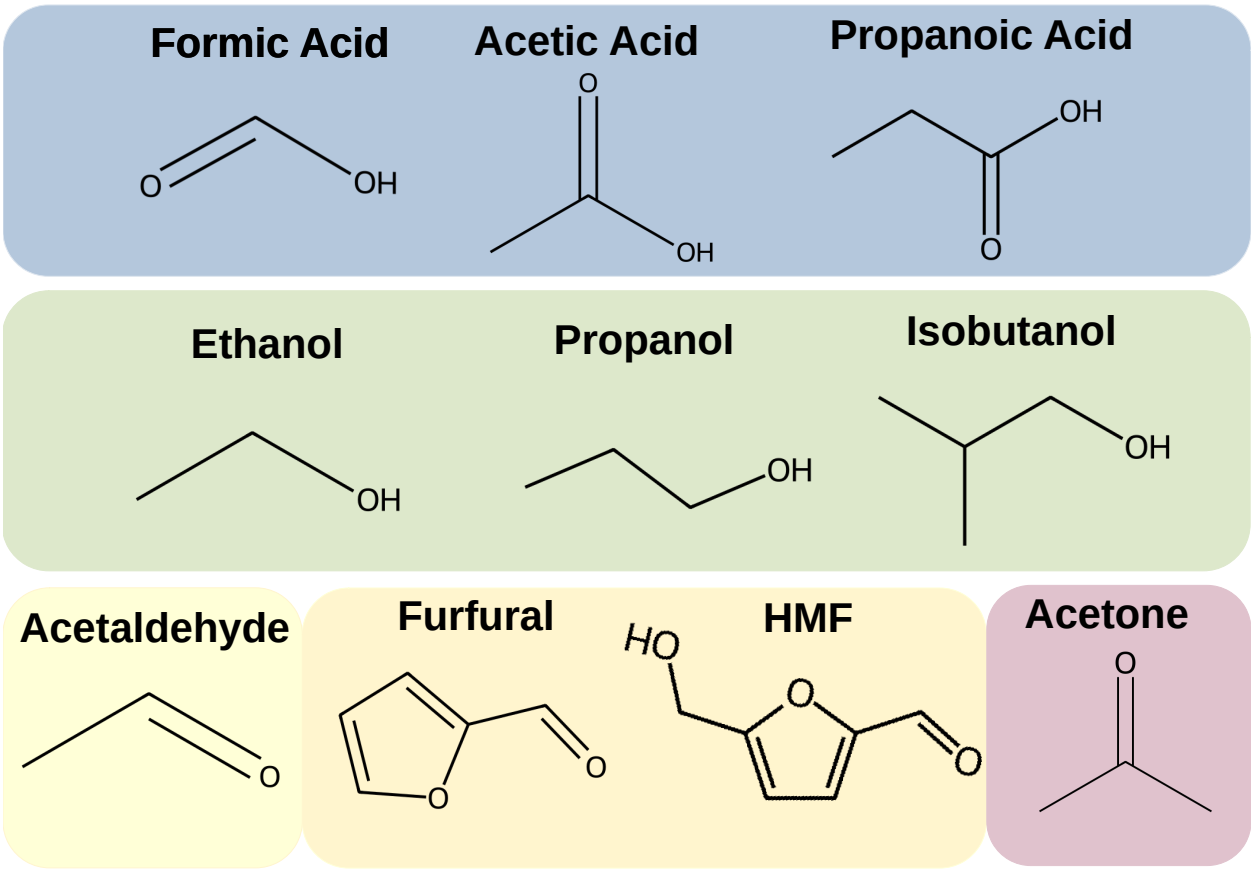


Figure 1: Small molecules considered in this study, grouped by chemical functionality. Blue highlights carboxylic acids (formic acid, acetic acid, and propanoic acid) in their protonated state, green includes alcohols (ethanol, propanol, and isobutanol), yellow contains aldehydes and furan derivatives (acetaldehyde, HMF, and furfural), and mauve represents a ketone (acetone).

Membrane Model and System Preparation

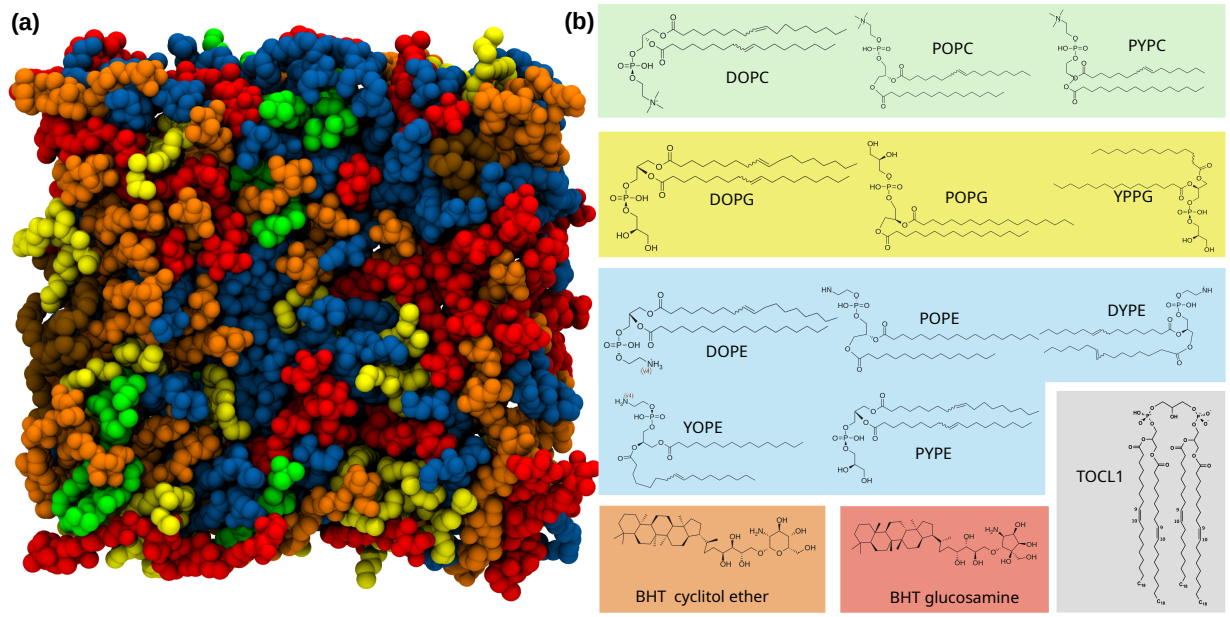


Figure 2: (a) Top view of *Z. mobilis* model membrane structure. The membrane is comprised of phosphatidylcholines (PCs) in green, phosphatidylglycerols (PGs) in yellow, phosphoethanolamines (PEs) in blue, cardiolipin (TOCL1) in grey and hopanoids in orange and red color. (b) The chemical structures of the lipids, colored according to their head groups. The names for these molecules are given in Table 1.

The *Z. mobilis* membrane model was constructed from the phospholipids phosphatidylethanolamine (PE), phosphatidylglycerol (PG), phosphatidylcholine (PC), and cardiolipin (CL), alongside hopanoids (bacteriohopanetetrol cyclitol ether and bacteriohopanetetrol glucosamine), as illustrated in Figure 2. The composition of the model was based on recent lipidomics data,[32] condensing the multitude of lipids identified by lipidomics into a representative lipid mixture as detailed in Table 1. Each membrane leaflet contained 100 molecules, consisting of either a mixture of phospholipids and hopanoids or phospholipids alone. For the hypothetical hopanoid-deficient membrane system, the membrane was generated using the CHARMM-GUI Membrane Builder.[40] Each leaflet contained 100 phospholipid molecules (combining PE, PG, PC, and CL), utilizing double the quantities listed in Table 1 to compensate for the absence of hopanoids.

The hopanoid-rich membrane system was constructed through a similar multi-step process. Initially, the system was built using the CHARMM-GUI Membrane Builder.[40] Because hopanoids are not currently present in the CHARMM-GUI lipid library, cholesterol and ergosterol were used as placeholders for the 50 hopanoid molecules required per leaflet. This 1:1 ratio represents the hopanoid enrichment found in *Zymomonas* strains actively producing ethanol,[41] as hopanoids are essential to microbial fitness under these conditions.[42] Following initial membrane assembly, each cholesterol or ergosterol molecule was replaced with the specific hopanoids found in *Z. mobilis* membranes.[42] Finally, as this replacement process can result in ring penetrations, LongBondEliminator was utilized to resolve these artifacts prior to extended simulation.[43]

To study the interactions of small molecules (Figure 1) with the *Z. mobilis* membrane, simulations were performed at varying concentrations of the small molecules (0, 0.5, 1.0, 1.5, 2.0, and 2.5 mol%). Since the carboxylic acids are less hydrophilic than their charged conjugate base, and that the acids are the primary form that crosses a typical lipid bilayer,[44–46] we only simulate the acid in this study. All small molecules were initially placed in the water phase, as shown in Figure S1. The system was solvated with TIP3P water molecules and was neutralized by adding counterions (Na^+ or Cl^-).

Lipid	Name	Numbers
DOPE 18:1/18:1	1,2-dioleoyl-*sn*-glycero-3-phosphoethanolamine	18
PYPE 16:0/16:1	1-palmitoyl-2-arachidonoyl-*sn*-glycero-3-phosphoethanolamine	4
YOPE 16:1/18:1	1-stearoyl-2-oleoyl-*sn*-glycero-3-phosphoethanolamine	3
POPE 16:0/18:1	1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphoethanolamine	3
DYPE 16:1/16:1	1,2-di-(9*Z*-hexadecenoyl)-*sn*-glycero-3-phosphoethanolamine	1
DOPG 18:1/18:1	1,2-dioleoyl-*sn*-glycero-3-phospho-(1′-rac-glycerol)	8
POPG 16:0/18:1	1-palmitoyl-2-oleoyl-*sn*-glycero-3-phospho-(1′-rac-glycerol)	4
YPPG 16:1/16:0	1-palmitoleoyl-2-palmitoyl-*sn*-glycero-3-phospho-(1′-rac-glycerol)	1
DOPC 18:1/18:1	1,2-dioleoyl-*sn*-glycero-3-phosphocholine	4
POPC 16:0/18:1	1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine	1
PYPC 16:0/16:1	1-palmitoyl-2-palmitoleoyl-*sn*-glycero-3-phosphocholine	1
TOCLI 18:1/18:1/18:1/18:1	1′,3′-bis(1,2-dioleoyl-*sn*-glycero-3-phospho)-*sn*-glycerol	2
	Bacteriohopanetetrol cyclitol ether	25
	Bacteriohopanetetrol glucosamine	25
Total		100

Table 1: Membrane phospholipid and hopanoid composition in wild type *Z. mobilis* adapted from available lipidomics[32] to specify the number of lipids in each leaflet of the lipid bilayer.

Molecular Dynamics Simulations

All MD simulations were performed using the NAMD 3.0.1 simulation engine.[47] The lipid parameters were derived from CHARMM36,[48] while the CGenFF[49] was employed for the small molecules and hopanoids (Bacteriohopanetetrol cyclitol ether and Bacteriohopanetetrol glucosamine). This classical molecular simulation captures partitioning, but not reactivity for molecules like aldehydes present in our molecule set.[50] As is standard for CHARMM36,[48] we used the TIP3P water model,[51] the representative initial system setup is shown in Figure S1. Periodic boundary conditions were applied in all directions to avoid edge effects, with the initial box dimensions of $89 \times 93 \times 100\,\text{Å}^3$.

The systems were first energy-minimized using the steepest descent algorithm to remove any steric clashes. The simulations were performed in a constant pressure and temperature (NPT) ensemble. The temperature was maintained at 300 K using the Langevin thermostat,[52] and the pressure was controlled at 1 bar using the Langevin barostat.[53] The barostat had a piston period of 200 fs and a decay time of 100 fs. Group pressure coupling was used in combination with a flexible cell and a constant cell ratio, as is typical for membrane simulations in NAMD. The flexible cell setting permits the simulation box to adjust both in volume and shape to accommodate anisotropic pressure fluctuations, while the constant ratio between the x- and y- dimensions keeps the membrane aspect ratio constant. A 2 fs time step was applied throughout the simulation, enabled by constraining bond lengths to hydrogen atoms using SETTLE.[54] Long-range electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method[55] with a grid spacing of 1 Å. Van der Waals interactions were truncated at 12 Å with a switching function applied at 10 Å to maintain continuous forces and energies. All the simulations were performed for three independent runs, each run to 1000 ns. All thermodynamic analysis was carried out for the last 800 ns considering the first 200 ns as the equillibration time, and statistics were averaged over the three independent runs.

Membrane Property Analysis

We measure multiple membrane properties from the simulations described above. All properties were calculated for each of the three replicas, and the results were averaged over all replicas. Error estimates were determined as the standard error of the mean across the average of individual simulation replicates. All calculations were performed using in-house Python scripts, using what numpy,[56] scikit,[57] and matplotlib.[58] The following properties were analyzed to characterize the interactions of small molecules with the *Z. mobilis* membrane:

Area per Lipid (APL)

The area per lipid (APL) was determined from the periodic box size of the simulation system, averaged over the equilibrated portion of the trajectory and divided by the fixed number of lipids in a leaflet. While we could have determined the area per lipid for individual lipids, we focus here on the overall area per lipid to track changes in membrane structure when the small molecule concentration rises.

Membrane Thickness

Similarly, we measure the membrane thickness for the entire membrane by measuring the distance along the membrane normal (z) axis between the average position for phosphorus atoms in the upper and lower leaflets.

Lipid Order Parameter (S_{CH})

To facilitate comparisons with potential future NMR experiments, we determine the lipid order parameter, averaged across all lipids all along every acyl chain to measure the overall order as a function of small molecule concentration. The lipid order parameter, S_{CH}, is measured in NMR via carbon-deuterium bond orientations, but we are measuring based on the equivalent C-H bond vector. The angle of this bond compared with the membrane normal axis, θ, is used to quantify lipid order:

$$S_{CH} = \frac{1}{2}\langle 3\cos^2\theta - 1\rangle \tag{1}$$

Lateral Diffusion Coefficient

Measuring overall lipid dynamics is done here by measuring the lateral diffusion of individual lipids within the larger molecular system. The lipid lateral diffusion coefficient (D_{xy}) was calculated from the mean square displacement (MSD) of lipids in the xy-plane using the relation:

$$D_{xy} = \frac{\langle MSD_{xy} \rangle}{4} \tag{2}$$

To compute D_{xy}, the molecular dynamics trajectory was divided into consecutive 10 ns segments. For each segment, the MSD of the lipid phosphorus atoms was calculated relative to the first frame of that segment. A linear regression was then performed on the MSD versus time data for each segment, and the fitted slope was used to determine D_{xy} according to Equation 2. The final diffusion coefficient was reported as the mean of all segment-based values, with the standard error of the mean calculated to represent variability across segments.

Probability Density and Free Energy Profile for Small Molecules Across the Membrane

From our extensive sampling, it is possible to measure the probability distribution for the various small molecules across the membrane. This is done by determining the center of mass for each molecule with respect to the membrane normal, and then histogramming the resulting positions to determine a probability distribution. The probability distribution can then be converted into a free energy profile by the well worn relationship:

$$\Delta G = -RT \ln \frac{p}{p_0} = G_p - G_{p_0} \tag{3}$$

In Eq. 3, p is the probability within a specific bin, while p_0 is the probability in a reference bin within the histogram. We think it is most helpful if the small molecule in solution is considered to be the zero point for free energy, which we accomplish by selecting the probability in solution as p_0.

Flux-based estimation of permeability

The permeability values were calculated directly from the statistics of complete leaflet-to-leaflet crossings observed in atomistic MD trajectories. For a given solute we recorded the total number of full translocation events N in three independent 1000 ns replicas of a hydrated *Z. mobilis* model membrane. A permeability coefficient was then computed following the analysis framework from Venable et al.[59].

$$Pm = \frac{N}{2A\,\Delta t\,C} \tag{4}$$

where the factor 2 accounts for the two bilayer faces, Δt is the aggregate simulation time and C is the instantaneous aqueous concentration extracted from the bulk region and A is the area of the lipid bilayer.

Results

There are two overall goals for the simulations. First, using equilibrium molecular–dynamics simulations, we quantify how membrane structure and dynamics in *Z. mobilis* respond to biomass derived small molecules as their alkyl chain length and concentration increase. Second, we explore the role of hopanoids within the membrane, and how they contribute to exceptional ethanol tolerance by simulating a hypothetical membrane in which all hopanoids are removed. The results addressing these two objectives are presented sequentially in the following sections.

Control Membrane Structure and Dynamics

Before probing structure and dynamics changes within the membrane precipitated by small molecule action, it is important to know what occurs when small molecules are absent. Thus, we quantify the intrinsic behavior of the *Z. mobilis* model membrane in the absence of exogenous stressors, both when hopanoids are

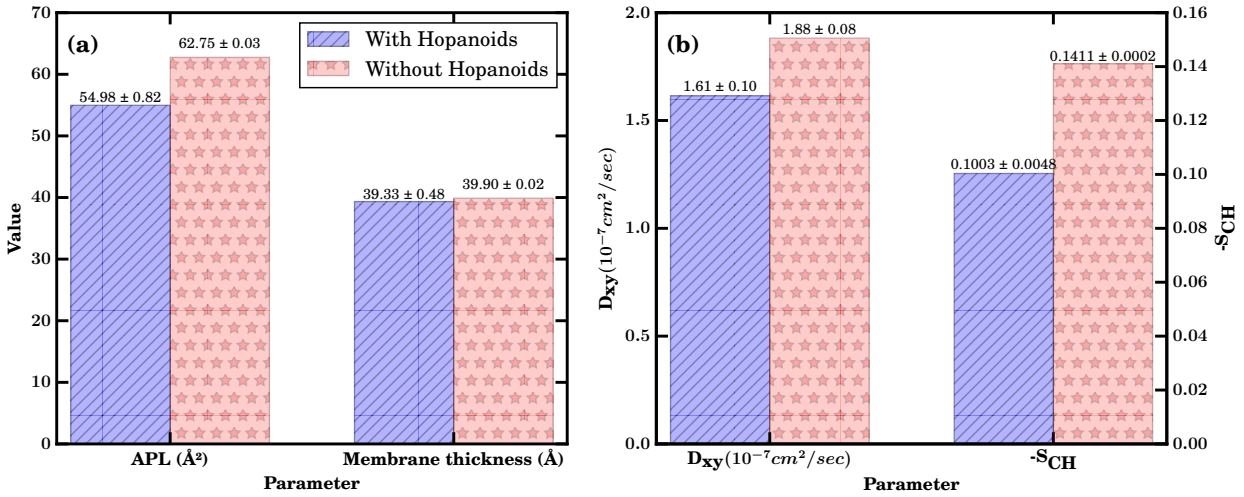


Figure 3: Comparison of control bilayer properties (a) APL and membrane thickness and (b) D_{xy} and $-S_{CH}$; in the presence (blue) and absence (red) of hopanoids. Simulations were performed in the NpT ensemble at 300 K and 1 bar. Each bar represents the mean over the final 800 ns of three independent trajectories, with values are printed above each bar for clarity, and also reproduced in Table S1.

present and absent from the membrane model. The resulting baseline membrane properties in Table S1 and Figure 3 measure average structure and dynamics properties from the individual timeseries from Figures S2 and S3.

Comparing between the standard *Z. mobilis* and the hypothetical hopanoid-free membrane, there are clear differences in membrane structure and dynamics. Removing hopanoids increases the area per lipid by ∼15% and the lipid lateral diffusion coefficient D_{xy} by ∼19%. In our simulations, eliminating hopanoids spreads apart the lipids laterally yet leaves the hydrophobic core thickness essentially unchanged, indicating that hopanoids function primarily as in-plane condensing agents rather than as modulators of bilayer thickness—a behavior reminiscent of other sterol like lipids.[60] Pentacyclic hopanoids intercalate among phospholipid acyl chains to promote tighter packing and reduce free surface area. However, the rigid, nearly cylindrical shape for the ring system present in the hopanoids (Figure 2) means that the tighter packing does not expand the membrane along the membrane normal and keeps the membrane structural orientation intact. One area where this is shown explicitly is in the order parameter analysis, where the removal of the hopanoids lead to ∼40% increase in the $-S_{CH}$ value (Figure 3). The membrane with hopanoid composition is more condensed and less-fluid. Because membrane permeability generally scales with both area per lipid and lipid mobility, the hopanoid-rich state is expected to confer an inherent barrier to small-molecule entry, whereas the hopanoid-depleted control is primed for higher passive uptake.

Membrane Dynamics and Structure Shifts under Stress

Having determined the baseline membrane structure and dynamics in a simple aqueous environment, we studied the impact different classes of small molecule have on the structure and dynamics of *Z. mobilis* membranes. In Figure 4, we report the membrane thickness, area per lipid, D_{xy} (Eq. 2), and $-S_{CH}$ (Eq. 1), as we increase the solute concentration from 0.5% to 2.5 mol%. Figure 4 makes it clear that none of the five hydrolysate derived metabolites cause a significant damage to the *Z. mobilis* bilayer, yet a set of small, internally consistent shifts emerges once we examine each observable in turn.

Among the studied small molecules, furfural, acetic acid, and ethanol produce a measurable, quasi-linear increase in area per lipid with rising mole fraction, with furfural showing the largest perturbation at 2.5 mol%. In contrast, acetone and acetaldehyde do not exhibit statistically significant changes relative to the control, consistent with the confidence intervals shown in Figure 4(a). Even for the most perturbing solutes, the changes remain modest. The mean area per lipid stays within the range of area per lipid values sampled in

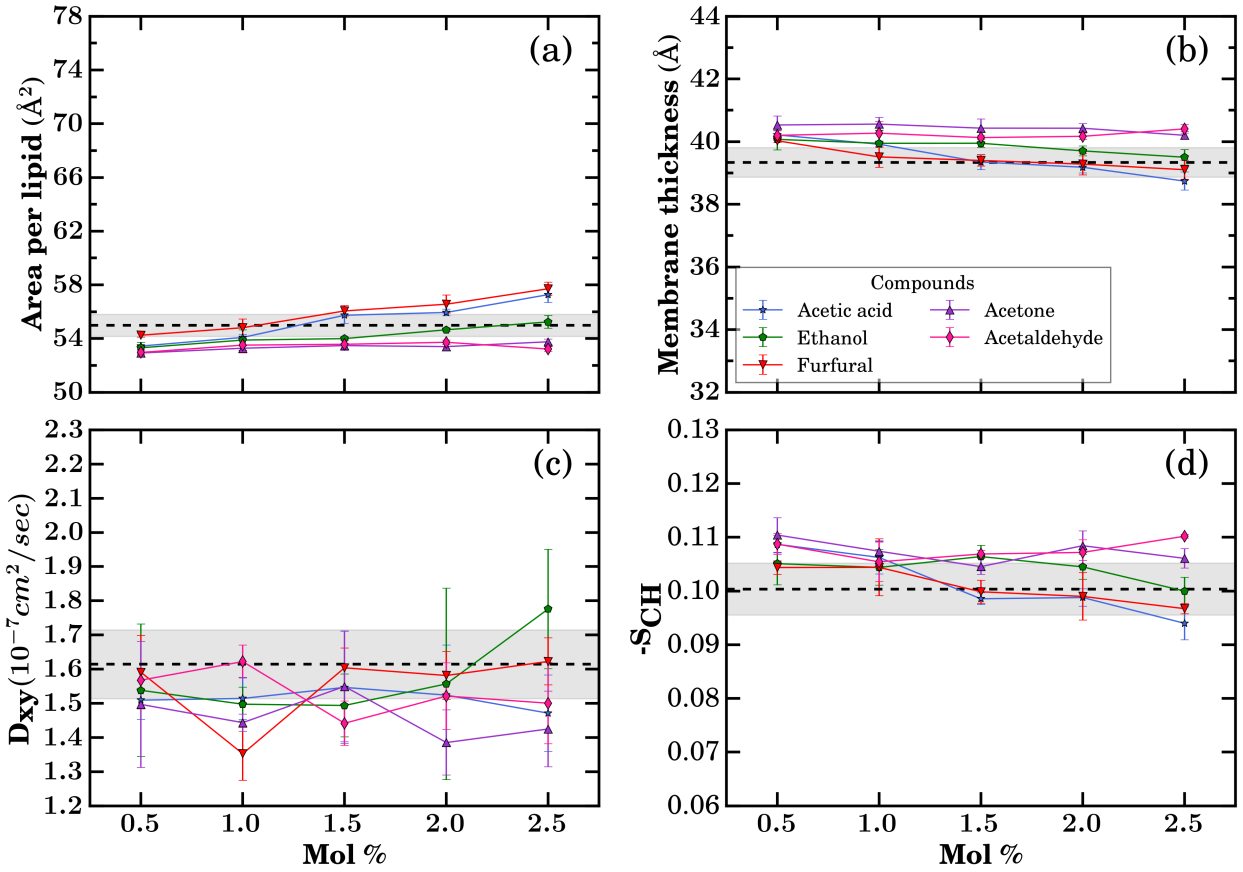


Figure 4: Membrane structure and dynamics properties quantified with increasing concentrations of small molecules from different compound classes introduced in Figure 1: (a) area per lipid, (b) membrane thickness, (c) lateral diffusion coefficient (D_{xy}), and (d) deuterium order parameter ($-S_{CH}$). Simulations were performed in the NpT ensemble at 300 K and 1 bar. Each data point represents the mean over the final 800 ns of three independent trajectories. The black dashed line shows the average values and the grey area shows the standard deviation from control membrane simulation runs when no small molecules were present. The timeseries data for the individual runs of each molecule is shown in Figures S4-S28.

the control simulations without solutes added (grey area in Figure 4), reinforcing that the bilayer remains in the liquid crystalline regime in the studied concentration range.

Membrane thickness displays a related pattern. All five solutes produce a mild concentration-dependent decrease in phosphate-to-phosphate thickness (Figure 4(b)), with the control spanning 39.3 ± 0.5 Å. By 2.5 mol%, acetic acid, ethanol and furfural induced a small but detectable thinning of the membrane.

These structural changes have a far smaller impact on dynamics. Figure 4(c) shows the D_{xy} changing in response to the small molecules. In the solute-free control, D_{xy} is $(1.61 \pm 0.10) \times 10^{-7}\,\mathrm{cm^2\,s^{-1}}$. The D_{xy} value does not show a significant change and it remains consistent across different small molecules within the tested concentrations.

Orientational order parameters $-S_{\mathrm{CH}}$ show small but internally consistent changes (Figure 4(d)). Ethanol induces only minimal tail disorder even at 2.5 mol%, consistent with the high ethanol tolerance observed for *Z. mobilis*. Furfural and acetic acid yield slight decreases in order at higher concentrations, aligning with their observed increases in area per lipid and decreases in membrane thickness (Figure 4(a–b)).

Overall, the analysis portrays a membrane that is remarkably resilient, even at 2.5 mol%. None of the solutes drive any observable outside the range sampled by the control simulations (grey regions in Figure 4). While the hydrolysate components modulate membrane properties in predictable, dose-dependent manner, the perturbations remain well within the tolerance window of *Z. mobilis* membranes (Figure 3). Still, coherent patterns emerge. First, a universal, dose dependent reduction in membrane thickness and a commensurate APL increase. Second, compounds such as furfural and acetic acid causes the maximum perturbation to the membrane among the studied compounds.

Effect of Chain Length on Membrane Dynamics

Whereas Figure 4 scans across multiple small molecule chemistries, converting from alcohols to aldehydes to acids, Figure 5 explores what happens as we add hydrophobic bulk to a subset of these molecules. For acids and alcohols, we are adding carbons to the small molecule as we move from formic acid to acetic acid to propanoic acid or from ethanol to propanol to isobutanol. HMF, while larger than furfural, is also less hydrophobic, which ends up having a strong impact on the trends observed in Figure 5.

Starting from Figure 5(a), the maximum area per lipid within the dataset is far higher than we observed for the control membrane in Figure 3(a). Particularly for the larger, more hydrophobic molecules within the dataset (isobutanol, propanoic acid, and propanol) the area per lipid is so far above the range seen without small molecules present, this represents a substantial membrane structure perturbation. The increased area per lipid is driven by these molecules intercalating near the membrane headgroups and disrupts the lipid packing. Conversely smaller molecules occupy less space near the membrane-water interface, and thus have less dramatic changes at a similar concentration. Regardless of molecular size, the area per lipid increases with increasing molecular concentration. Increasing lipid area with increasing small molecule concentration has been previously observed in prior studies.[61–63]

Membrane thickness is the converse of the area per lipid. Since the size of individual lipids is fixed, a larger spacing between lipids tends to thin membranes. Thus the hydrophobic molecules that show the greatest increase in area per lipid show the greatest decrease in membrane thickness (Figure 5(b)). More broadly, increasing concentrations for these molecules (Figure 1), decrease the membrane thickness, although for most compounds the trend is relatively weak within the studied concentration range.

Higher concentrations of small molecules can disrupt lipid–lipid interactions, leading to bilayer destabilization. Given how important hydrophobic matching is to membrane protein function, the reduced thickness for the most hydrophobic compounds is likely to substantially impact membrane protein function,[64,65] and certainly is a plausible mechanism by which these molecules would stress *Z. mobilis*. Despite these structural changes, the D_{xy} is strikingly resilient: within statistical uncertainty it remains nearly constant across the full concentration range examined (Figure 5(c)). This invariance indicates that lipid mobility, hence the viscous properties of the membrane interior is largely preserved even as the bilayer spreads laterally and becomes thinner.

Returning to structural metrics, the lipid order parameter $-S_{CH}$ quantifies acyl tail rigidity and orientation. Thinner membranes with larger spaces between lipids introduce additional disorder into the membrane, and so the overall trend in Figure 5(d) is for decreased order, mimicking the trend for the membrane thickness. Just like the other structure measures, the largest changes occur for isobutanol, propanol, and propanoic

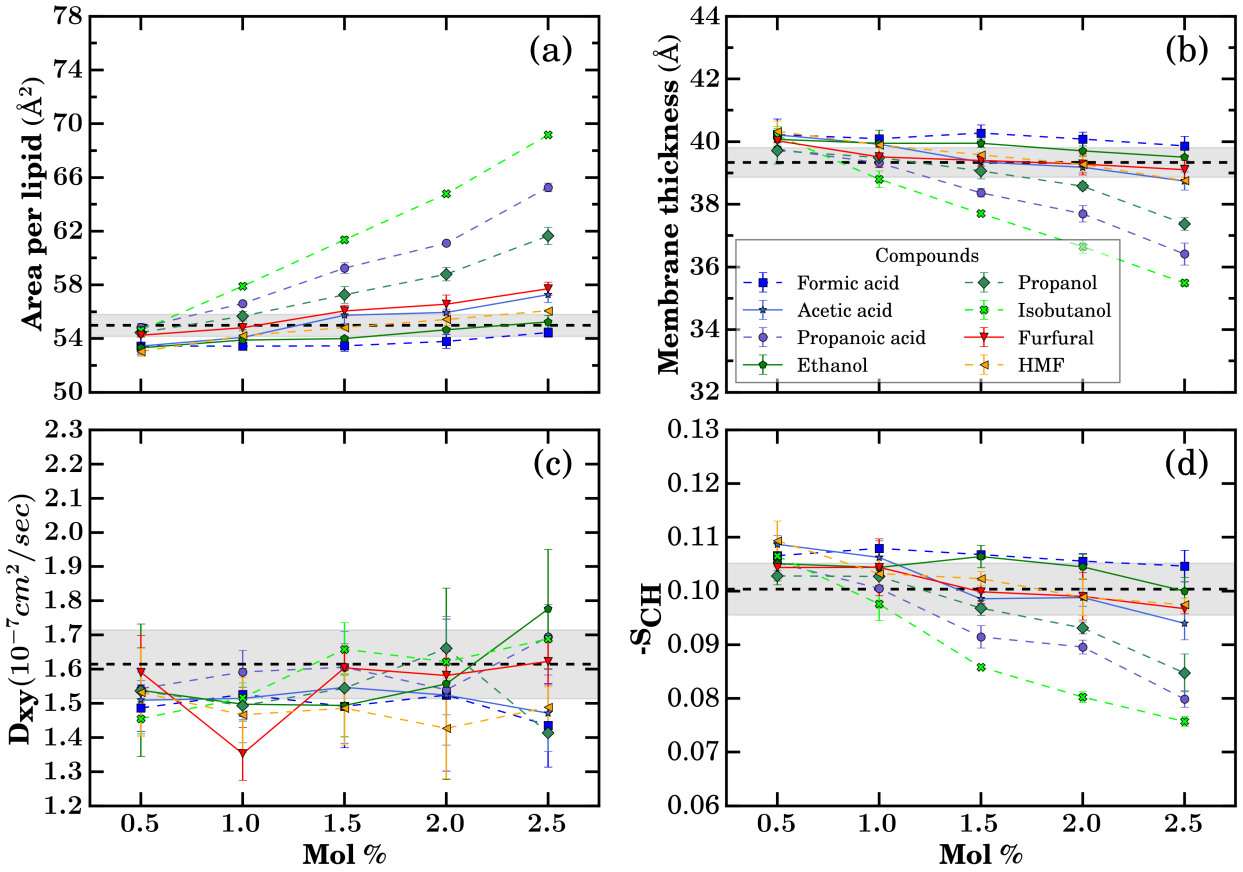


Figure 5: Membrane structure and dynamics properties quantified with increasing alkyl chain length and concentrations of small molecules. (a) Area per lipid, (b) Membrane thickness, (c) Lateral diffusion coefficient (D_{xy}), (d) Deuterium order parameter ($-S_{CH}$). Simulations were performed in the NpT ensemble at 300 K and 1 bar. Each data point represents the mean over the final 800 ns of three independent trajectories. The black dashed line shows the average values and the grey area shows the standard deviation from control membrane simulation runs when no small molecules were present. The timeseries data for the individual runs of each molecule is shown in Figures S4-S18, S29-S48.

acid. The observed shift for these molecules of approximately 0.02 is similar to the shift observed for plant membranes when raising the simulated temperature by 50°C.[66] Unlike temperature, which changes relatively slowly over the course of a day, concentration changes in a hydrolysate stream can change far more rapidly, emphasizing how challenging such an environment is for microbes in general. These findings demonstrate that even within the same class of molecules, variations in the number of carbon atoms and molecular structure significantly influence the nature and extent of their interaction with the lipid membrane. In both carboxylic acids and alcohols, increasing the alkyl chain length generally leads to a greater perturbation of the membrane in terms of lateral expansion and increased fluidity (disorder of lipid tails), and also tending to slight decrease in membrane thickness.

To further assess the association of hopanoids and other lipids within the membrane system, we quantified the lateral distribution of hopanoids and other lipids using the lipid enrichment/depletion index implemented in LiPyphilic.[67] In this analysis, an enrichment index of 1 corresponds to random mixing, values > 1 indicate an increased probability of finding a given lipid type as a nearest neighbour, and values < 1 indicate depletion. We grouped lipids into four classes: hopanoids (HOP), saturated tails (Sat), monounsaturated tails (MU), and polyunsaturated tails (PU), and computed the grouped enrichment index matrices for two representative systems: (i) acetaldehyde at 0.50 mol%, which causes the smallest perturbation in global membrane properties (Figure 4), and (ii) isobutanol at 2.50 mol%, which causes the largest perturbation (Figure 5).

The resulting enrichment matrices (Figure S49 and S50) show that when hopanoids are taken as the reference lipid, all neighbour types (including other hopanoids) have enrichment indices ≈ 0.9 in both systems, indicating a mild depletion of all lipid types in the immediate vicinity of hopanoids relative to random mixing. For example, a Sat–HOP enrichment value of 1.12 would mean that hopanoids are 12% more likely to appear next to a saturated lipid than expected from the bulk composition, whereas a HOP–HOP value of ≈ 0.9 would indicate a slight tendency against hopanoids sitting next to each other. When phospholipids (Sat, MU, PU) are used as reference lipids, hopanoids appear only modestly enriched as neighbours (indices typically ~ 1.1), and the indices for phospholipid–phospholipid contacts lie close to unity and within a relatively narrow range (roughly 0.8–1.2) for all pairs. We therefore do not observe strong hopanoid–hopanoid self-enrichment or any lipid pair with a pronounced enrichment or depletion that would clearly indicate stable hopanoid-rich or hopanoid-poor lateral domains on the simulated length scale. This behaviour is similar for the minimally perturbed acetaldehyde system (Figure S50(a)) and for the more strongly perturbed isobutanol system (Figure S50(b)), suggesting that, in these two representative cases, the lipids remain largely laterally mixed.

Collectively, these enrichment maps, combined with the global APL and thickness data (Figures 4–5), suggest that hopanoids mainly act as agents that condense the membrane in-plane, influencing its average packing and fluctuations, rather than creating distinct thick or thin regions within the simulation boxes and time scales we examined.

Molecular distribution and free energy profile across the membrane

To elucidate the small molecule partitioning behavior within the *Z. mobilis* membrane, we computed the small molecules probability density distributions relative to the bilayer center (Figure S51–S52). From these probabilities, we can further quantify molecular translocation energetics by evaluating the free energy profiles by Boltzmann inversion of the probability distributions (Figure 6). These distribitions provide insight into the preferential localization of each compound and its potential interaction with membrane components and the free energy profiles provide a measure of the energetic barriers associated with membrane traversal and allow for comparative analysis of molecular permeability.

We first study the probability distributions in different classes of molecules, shown in Figure S51. The selected compounds, acetic acid, ethanol, acetone, and acetaldehyde, span a range of chemical classes, enabling a comparative evaluation of how different chemical class compounds influences membrane partitioning and insertion energetics. As shown in Figure S51, trends suggest that most molecules exhibit peak densities at the lipid–water interface, and significantly lower probabilities in the hydrophobic core with the degree of central bilayer penetration strongly correlating with hydrophobicity.

Acetic acid and ethanol showed a sharply peaked distribution at the interface and minimal presence in the membrane core. This pattern reflects the compound's affinity for the polar head group region and aversion to the hydrophobic interior, which is natural for these small molecules with substantial hydrophilic groups. With

increasing concentration, the interfacial peaks in Figure S51 increased slightly, suggesting minor penetration into deeper membrane regions at higher loading. The free energy distributions for acetone and acetaldehyde shows lower peak values at the membrane-water interface as compared to acetic acid and ethanol.

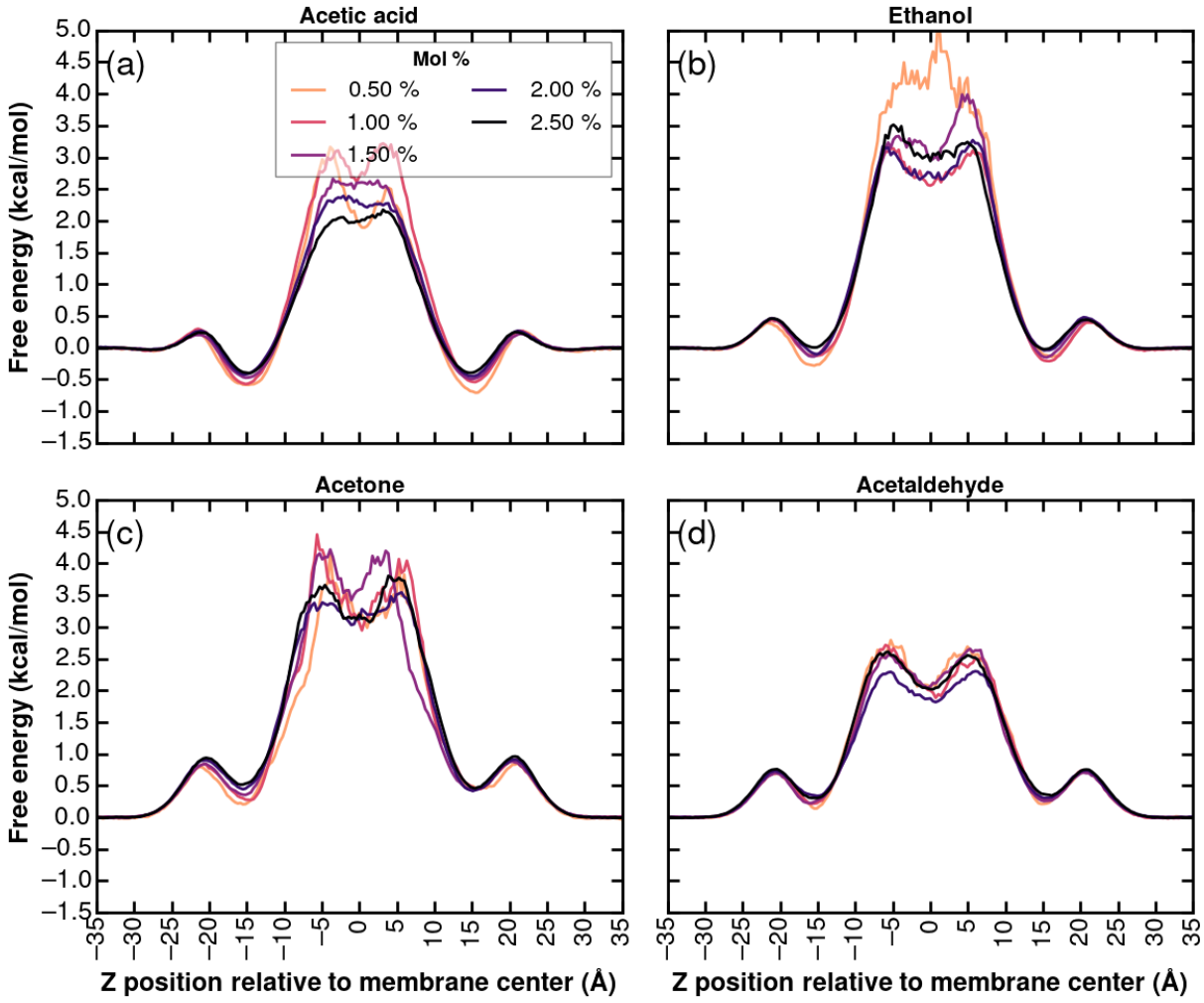


Figure 6: Free energy profiles obtained from the Boltzmann inversion of probability distributions reported in Figure S51 for (a) acetic acid, (b) ethanol, (c) acetone and (d) acetaldehyde.

The probability distributions in Figure S51 can be inverted into free energy profiles to express these probabilities for the small molecule solutes into energy terms. The free energy $\Delta G(z)$ as a function of position z relative to the bilayer center was calculated using Eq. 3. This formulation directly relates the statistical occupancy of solutes at different membrane depths to their underlying thermodynamic stability. The computed free energy profiles (Figure 6) reflect the hydrophobicity trends of these molecules (Acetaldehyde $<$ Acetone $<$ Ethanol $<$ Acetic acid). A consistent feature across all molecules is the presence of a free energy minimum at the bilayer interface, approximately 15–20 Å from the membrane center. This minimum corresponds to a region where solutes experience partial insertion into the bilayer, interacting with both hydrophilic headgroups and partially exposed hydrophobic acyl chains. Because the bilayer and solute loading are symmetric about the membrane midplane, the underlying free-energy profiles are expected to be symmetric, the mild left–right differences seen in Figure 6 arise from finite sampling rather than any built-in structural asymmetry.

At 0.5 mol% acetic acid exhibited a well defined central energy barrier (∼3.5 kcal/mol), which decreased slightly with concentration (Figure 6(a)). The prominent interfacial minima and significant barrier at the

core are consistent with its strong preference for the polar environment and the high energetic cost of insertion into the hydrophobic bilayer center. This pronounced interfacial stabilization in acetic acid arises from its high polarity and capacity for hydrogen bonding. The carboxylic acid group readily forms hydrogen bonds with the phosphate or glycerol moieties of lipid headgroups, promoting strong interfacial association. Ethanol (Figure 6(b)) on the other hand, exhibited a central energy barrier of ∼5 kcal/mol at 0.5 mol% which decreased slightly with concentration.

The profiles of acetone and acetaldehyde (Figure 6(c-d)) also revealed the presence of interfacial energy barrier when approaching the bilayer from the aqueous phase. Acetone and acetaldehyde possess polar carbonyl groups but lack hydrogen bond donors, limiting their ability to form stabilizing interactions with the lipid headgroups during initial insertion. This leads to a small but detectable energy barrier at the onset of the membrane interface, representing the disruption of favorable solute–water interactions without fully compensatory interactions with the membrane.

These interfacial minima also evolve with concentration. For acetic acid and ethanol, increasing the mole percent reduces the interfacial stabilization slightly (Figure 6(a-b)), likely due to saturation of hydrogen bonding sites or local reorganization of the lipid environment. This effect is less pronounced in the other compounds, consistent with their weaker interfacial interactions. Overall, these trends illustrate that the depth of the interfacial free energy well is a direct consequence of solute hydrophilicity and its chemical capacity for hydrogen bonding, and plays a crucial role in determining membrane permeability and accumulation behavior.

Effect of Alkyl Chain Length on Membrane Partitioning and Energetics

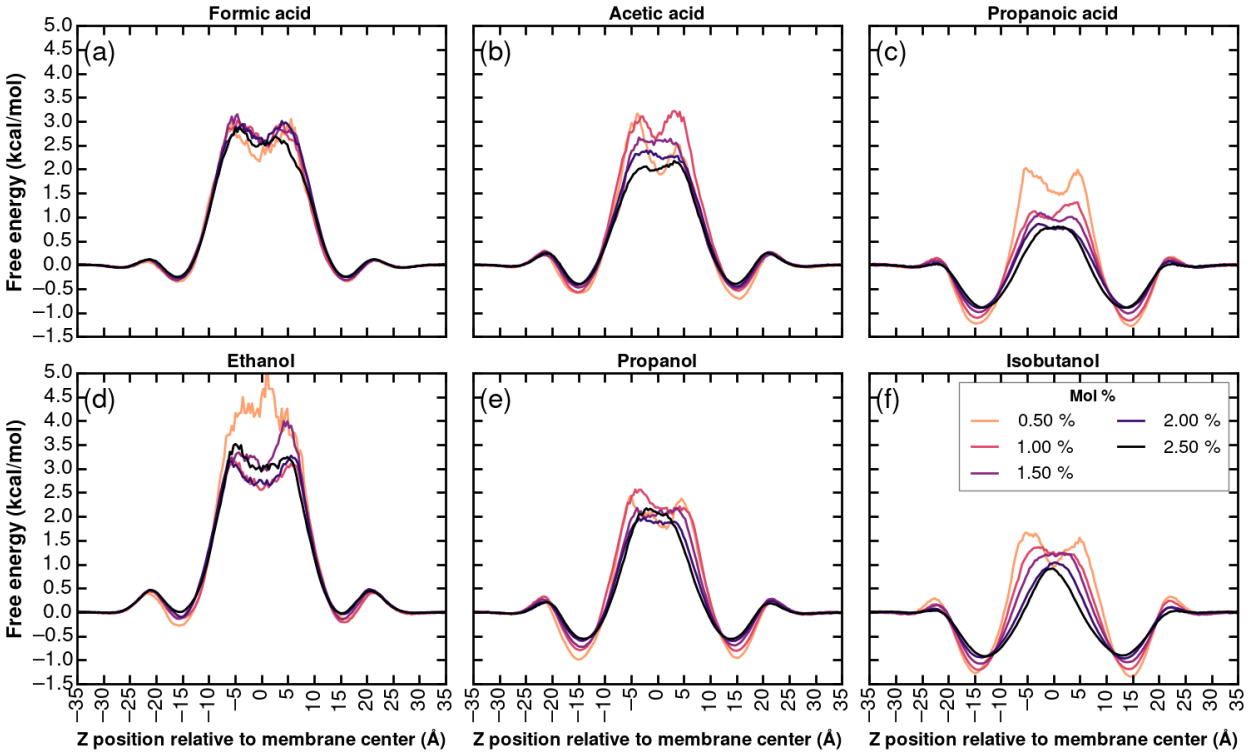


Figure 7: Free energy profiles obtained from the Boltzmann inversion of probability distributions for carboxylic acids (formic, acetic and propanoic) (a-c) and alcohols (ethanol, propanol and isobutanol)(d-f).

To probe how incremental increases in hydrophobicity modulate solute–membrane interactions, we next analyzed a homologous series of carboxylic acids (formic, acetic, propanoic) and alcohols (ethanol, propanol, isobutanol) in concentration range of 0.5–2.5 mol%. The probability density distributions and their Boltzmann-inverted free energy profiles (Figure S52 and Figure 7 respectively) reveal systematic trends in insertion depth,

interfacial stabilization, and central barrier heights as the alkyl chain is lengthened.

Formic acid, acetic acid and propanoic acid remains tightly localized at the lipid–water interface, with a narrow peak at $|z| \approx 15$ Å. For formic acid, its central free energy barrier is ~ 3.5 kcal/mol (Figure 7(a)). Extending to acetic acid, introduces noticeable probability in the upper acyl-chain region (Figure S52), and lowers the core barrier to ~ 2.5 kcal/mol at higher loading of 2.5 mol%(Figure 7(b). Propanoic acid amplifies these shifts and the central barrier falls further to ~ 1.5 kcal/mol at higher loading of 2.5 mol% (Figure 7(c)). Across all three acids, increasing concentration mildly broadens the interfacial maximum and slightly depresses the free energy minima, indicative of concentration-driven lipid perturbation but preserving the hydrophobicity-driven partitioning trend.

A parallel progression is seen for the alcohols. Ethanol exhibits an interfacial peak at $|z| \approx 15$Å and a core barrier of ~ 3.5 kcal/mol at higher loading of 2.5 mol% (Figure 7(d)). Propanol's distribution is broader, with significant density in the upper acyl chains, and its central barrier decreases to ~ 2.5 kcal/mol at higher loading of 2.5 mol% (Figure 7(e)). Isobutanol, the most hydrophobic, shows pronounced occupancy throughout the bilayer core (Figure S52) and a minimal barrier of ~ 1.5 kcal/mol at higher loading of 2.5 mol% (Figure 7(f)). As with the carboxylic acids, higher concentrations introduce slight broadening of the probability profiles and a modest flattening of the free energy wells, reflecting local membrane reorganization at elevated solute loading.

The height of the barrier at the membrane-water interface provides insight into the kinetics of membrane association. Molecules encountering a higher energy barrier (Figure 6 and Figure 7) at the water–membrane interface may exhibit slower rates of insertion and lower interfacial occupancy under equilibrium conditions. This partially explains the broader and more dispersed probability density profiles of acetone and acetaldehyde, as compared to the sharply localized interfacial peak of acetic acid.

In order to address the asymmetry in the free-energy distributions of the inhibitors on the two sides of the membrane, we analyzed the free-energy profiles separately for each of the six independent leaflets per condition (three independent simulations with two leaflets each). For each leaflet, the profiles were folded about the membrane midplane and expressed as a function of the absolute distance from the membrane center, $|z|$. This procedure removes any imposed left–right labeling while preserving leaflet-to-leaflet variability. We then computed the mean free-energy profile and the corresponding standard deviation across all six leaflets. The resulting one-sided free-energy profiles are shown in Figures S74–S82, confirming that any apparent leaflet-level asymmetries are minimal.

With this symmetry established, the homologous-series data underscore a clear structure–function relationship: each additional methylene unit deepens the interfacial free-energy well by approximately 1.0 kcal/mol and lowers the energetic cost of traversing the hydrophobic core, thereby enhancing overall membrane permeability in direct proportion to molecular lipophilicity.

These findings emphasize that the initial steps of membrane insertion specifically the transition from bulk water to the lipid interface, are governed not only by hydrophobicity, but also by the molecular capacity to engage in specific interactions with membrane headgroups. The results provide quantitative insight into solute–membrane interactions and underscore how hydrophobicity modulates both thermodynamic and spatial properties of small molecule insertion in *Z. mobilis* membranes.

Membrane-crossing and small molecule permeabilities

The event based analysis of the permeability values reveals a spread of more than two orders of magnitude in passive permeability that can be traced to subtle variations in chain length, branching and aromaticity. The insights for the permeability calculations also correlate with the membrane dynamics (Figure 4-5), as greater membrane perturbation caused significant change in the permeability values. Permeability values (Figure S73 and Table S3) were calculated by quantifying the number of permeants that traverse the bilayer (Figure 8 and Table S2). Formic, acetic and propanoic acids all display pronounced concentration dependence. For formic acid, crossing events increased from 8 at 0.5 mol% to 96 at 2.5 mol%, with $\log_{10} Pm$ increasing from -1.83 to -1.50 (Table S3) as the number of crossings increased approximately 2x over what would be expected just based on the concentration increase alone. Acetic acid followed the same trend but with greater scatter, suggesting intermittent trapping by hydrogen bonded clusters at the interface. Propanoic acid benefited most from the additional methylene group, reaching $\log_{10} Pm = -0.96$ at 2.5 mol%—around

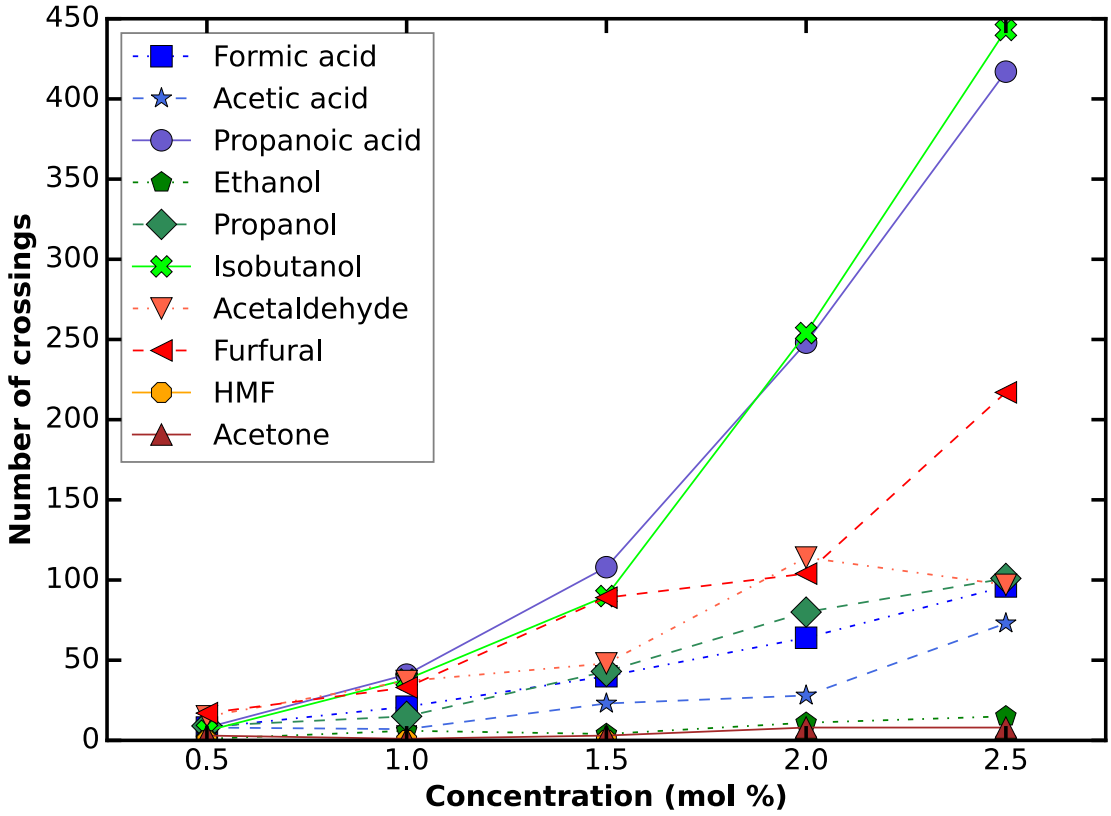


Figure 8: Total crossings for each solute at the range for concentrations from 0.5 mol% to 2.5 mol%, obtained by combining three independent 1000 ns MD trajectories. Data for individual runs is provided in Table S2.

ten fold gain over formic acid at identical loading (Figure S73).

Ethanol, propanol and isobutanol exhibited steeper gains per mole percent added than the acids, reflecting their weaker interfacial hydrogen bonding than the carboxylic acid compounds. Ethanol remained the least permeable alcohol (mean $\log_{10} Pm = -2.46$) because its small non polar patch is outweighed by the desolvation penalty of the hydroxyl group. Linear propanol improved to -1.49 at 2.5 mol%, whereas branched isobutanol underwent 443 crossings in the same window, yielding $\log_{10} Pm = -0.98$. Branching evidently disrupts lipid packing more effectively (Figure 5), lowering the activation barrier (Figure 7) for spontaneous pore like defects through which multiple molecules opportunistically slip. These results directly expose the substantial kinetic hurdle caused by the molecule having to lose its water shell and push methyl groups through tightly packed acyl chains.

Acetaldehyde displayed intermediate permeability, climbing from -1.57 at 0.5 mol% to -1.33 at 2.0 mol% (Table S3) across the loading series. Acetone, however, remained transport limited, fewer than ten full crossings were recorded in any replica, capping $\log_{10} Pm$ at -2.28 (Table S3).

The permeability of furfural and HMF diverged significantly. Furfural's behavior was comparable to unbranched C_3 alcohols; its five-membered ring partitioned readily into the glycerol backbone region, and its $\log_{10} P_m$ increased from -1.57 to -1.20 with concentration. HMF, by contrast, did not register a single translocation event in any of the 1000 ns MD simulations. We attribute this complete impermeability to persistent hydrogen bonds between HMF's hydroxymethyl group and interfacial water molecules, which stabilize surface-bound states and effectively suppress excursions into the bilayer core.

In summary, the permeability ranking, averaged across concentrations, is: propanoic acid $\approx$ isobutanol $>$ furfural $>$ propanol $>$ acetaldehyde $>$ formic acid $\approx$ acetic acid $>$ ethanol $>$ acetone $>$ HMF. The trend shows that permeability generally increases with hydrophobicity, as reflected in the higher transport of propanoic acid and isobutanol compared to smaller or more polar molecules. While it is unsurprising

that lipid bilayers are more permeable to hydrophobic compounds, this has downstream implications for biomaterial and biofuel target selection within a biorefinery, as increasing hydrophobicity may make the molecule a bigger membrane disruptor.

Effect of hopanoids on solvent-stressed membrane properties.

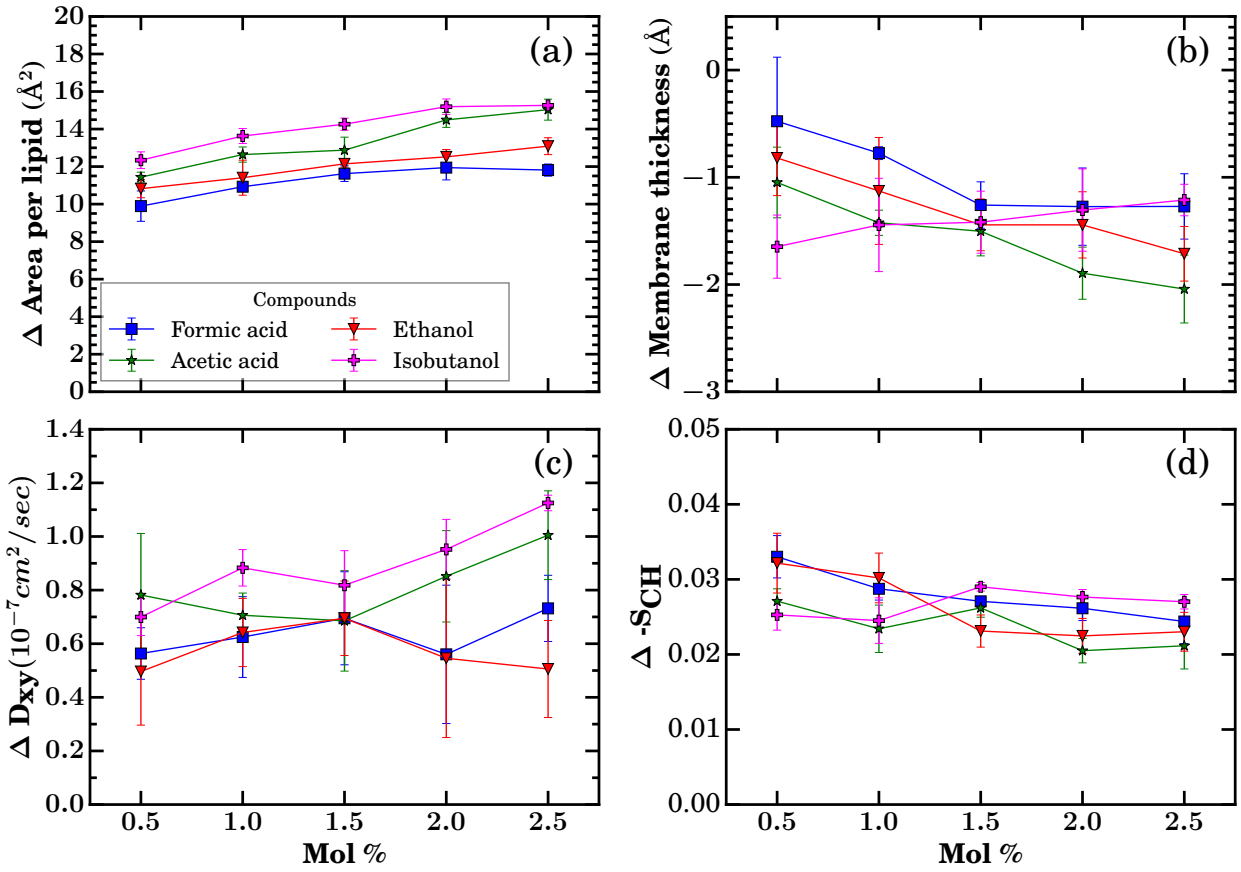


Figure 9: Differences in membrane properties between bilayers without hopanoids and with hopanoids as a function of inhibitor concentration: (a) ΔArea per lipid, (b) Δ membrane thickness, (c) ΔD_{xy}, and (d) $\Delta(-S_{CH})$. For each quantity, $\Delta X = X_{\text{without hopanoids}} - X_{\text{with hopanoids}}$, so positive values indicate that the property is larger in the absence of hopanoids. Data are shown for formic acid, acetic acid, ethanol, and isobutanol over the 0.5–2.5 mol% inhibitor concentration range. Simulations were performed in the NpT ensemble at 300 K and 1 bar. Each data point represents the difference of mean values for the final 800 ns of three independent MD runs. These differences quantify the hopanoid-induced modulation of bilayer structure, dynamics, and ordering. Exact numerical values for the underlying membrane properties are given in Tables S4–S7.

We further extended our study to investigate the stress response of the hypothetical membrane in the absence of hopanoids, for which we already calculated the parameters in Figure 3. Figure 9 shows the concentration-dependent response of bilayers that either lack hopanoids (solid traces) or retain 50% hopanoid present in a typical *Z. mobilis* membrane (dashed traces, Lipid composition of the *Z. mobilis* model membrane is provided in Table 1). Removing hopanoids roughly doubles the slope, i.e. doubles the sensitivity of the bilayer to a given dose of solvent molecules. This clearly demonstrates the value in using hopanoids to stiffen the membrane to mitigate solvent stress, highlighting why *Z. mobilis* can be such a good host for some bioproducts.

In hopanoid-free membranes the area per lipid (Figure 9a) grows in proportion to both inhibitor hydrophobicity and concentration. Formic acid broadens the bilayer by only ~ 3 Å^2 ($\approx 4.5\%$) across the full concentration series, whereas acetic acid and ethanol produce intermediate expansions of 5–7 Å^2 ($\approx 10\%$). Isobutanol is by far the most disruptive: area per lipid rises from ≈ 67 to ≈ 85 Å^2, a ~ 27 Å^2 ($\approx 35\%$) increase. Introducing hopanoids uniformly shifts the area per lipid curves downward by ~ 10 Å^2, consistent with tighter lateral packing. Isobutanol expands the bilayer in the hopanoid rich membrane, but only by ~ 15 Å^2 ($\approx 21\%$). The less hydrophobic inhibitors cause a lesser change to the area per lipid: the net area per lipid change falls to ≤ 4 Å^2 for acetic acid and ≤ 2 Å^2 for ethanol and formic acid (Table S4).

Membrane thickness (Figure 9b) responds inversely to the area per lipid changes. Without hopanoids, on increasing the inhibitor concentration, bilayer contraction is observed (Figure 4 and Figure 5). On removal of hopanoids, the change in membrane thickness was small for formic acid, acetic and ethanol while, isobutanol drives a pronounced ~ 4 Å thinning. The similarity in the membrane thickness values of hopanoid rich and hopanoid less membrane, demonstrate that hopanoids cannot fully resist the deepest-penetrating larger alkyl chain branched alcohol (Table S5).

Changes in lateral diffusion coefficients D_{xy} (Figure 9c) mirror the structural response. The hopanoid-free bilayer starts at $\sim 2.0 \times 10^{-7}\,\mathrm{cm^2\,s^{-1}}$ and accelerates modestly ($\lesssim 10\%$) in the presence of formic acid, ethanol and acetic acid, but by $\sim 25\%$ in the isobutanol series. Hopanoids set a lower baseline mobility ($\sim 1.5 \times 10^{-7}\,\mathrm{cm^2\,s^{-1}}$) and restrict its growth to $\lesssim 10\%$ for for formic acid and acetic acid inhibitors, but ethanol and isobutanol, still reaches $\approx 1.8 \times 10^{-7}\,\mathrm{cm^2\,s^{-1}}$ (Table S6).

The deuterium order parameter S_{CH} (Figure 9d) begins higher in hopanoid-free membranes (~ 0.14 for formic acid) than in hopanoid-containing ones (~ 0.11), indicating that native hopanoids loosen, rather than tighten, the chain backbone at baseline. Nevertheless, hopanoids stabilise the order against further solvent perturbation. Across the 0.5–2.5 mol % range, S_{CH} falls by $\lesssim 0.002$ for formic acid, ~ 0.0051 for ethanol, ~ 0.0147 for acetic acid, and ~ 0.031 for isobutanol. The corresponding drops in hopanoid-free membranes are ~ 0.010, ~ 0.014, ~ 0.021 and ~ 0.035, respectively. Hence hopanoids curtail the rate of disorder accumulation even though their presence confers a slightly more fluid ground state (Table S7).

Taken together, the updated metrics confirm that hopanoids endow the *Z. mobilis* membrane with a quantitative, though not absolute, resilience to lignocellulosic inhibitors. They (i) pre-compress the bilayer laterally, (ii) raise the hydrophobic-core thickness, (iii) lower the intrinsic lateral diffusivity, and (iv) slow the accrual of acyl-chain disorder. The comparable ~ 2-fold slope enhancement seen for the studied inhibitors suggests that hopanoids counter the generic physical effects of small molecules, independent of chemical functional group, by reinforcing lateral packing and vertical cohesion. This buffering is especially valuable at low inhibitor loadings, where small absolute concentration changes would otherwise produce disproportionately large biophysical responses (Figure 9). The magnitude of protection diminishes as the solute's hydrophobicity increase, with isobutanol remaining the most challenging molecule even in a hopanoid-rich lipid membrane system.

Discussion

For lignocellulosic biorefineries to be cost effective, we need strong, reliable microbes that can survive the tough chemical conditions in biomass hydrolysates. While *Z. mobilis* is a promising platform for biofuel production due to its high ethanol productivity and tolerance, its industrial application is often limited by inhibitory compounds. Several studies have investigated *Z. mobilis* responses to stressors such as ethanol, organic acids, and oxygen.[68–70] Recent efforts have shifted toward engineering *Z. mobilis* for isobutanol production, yet the microbe's growth is rapidly arrested by isobutanol at concentrations far below those tolerated for ethanol.[32,33] This pronounced sensitivity now represents the main bottleneck to achieving the titers and yields required for commercial scale isobutanol fermentation in engineered strains.

Batch cultures show that growth rates slows once external isobutanol reaches 8–12 $\mathrm{gL^{-1}}$ and are virtually abolished at 16 $\mathrm{gL^{-1}}$.[33] Multi-omics profiling links this stress to marked membrane swelling, global lipid remodeling, and the appearance of intracellular GFP aggregates at only 0.10–0.15 M.[32] Our atomistic simulations provide the missing mechanistic link: the branched C_4 alcohol diminishes the hydrophobic barrier that is able to repel shorter alcohols, enabling deep-core partitioning that expands the area per lipid, lowers acyl-chain order, and thins the hopanoid-rich bilayer. This pronounced structural disruption

provides a direct physical explanation for the membrane leakage and protein aggregation phenotypes observed experimentally,[71–74] highlighting the need for targeted membrane engineering to overcome this challenge.

A key principle for engineering such resilience lies in leveraging the membrane's native protective components. Our simulations reveal that hopanoids, which are prevalent in the *Z. mobilis* membrane, play a vital, sterol-like condensing role that directly counteracts inhibitor-induced damage. Replacing 50 mol% hopanoids with phospholipids *in silico* increased lateral diffusion by 19% and area-per-lipid by 15% , confirming their sterol-like condensing role. This pre-existing rigidity partially offsets the fluidisation imposed by inhibitors and rationalizes why hopanoid-deficient mutants are hypersensitive to stress *in vivo*.[75,76] Intriguingly, Rivera-Vázquez *et al.*[32] reported that *Z. mobilis* boosts cyclopropane fatty-acid (CFA) content under ethanol but fails to do so under isobutanol, despite up-regulating CFA synthase.[32] Our data illuminate the biophysical stakes: CFA rings would counteract the expansion driven by isobutanol, but their synthesis appears blocked, possibly because isobutanol denatures CFA synthase itself, an idea consistent with our observation that isobutanol partitions near the glycerol backbone where the enzyme acts. Engineering routes that enforce CFA (or hopanoid) over-production may therefore restore bilayer tightness and raise the lethal isobutanol threshold.[77,78]

Extending this molecular lens to the broader inhibitor palette clarifies empirical trends cataloged in hydrolysate-tolerance reviews.[79] Yang*et al.* summarized early high-throughput screens showing that inhibitory strength tracks with compound hydrophobicity across acids, aldehydes, and furans and that combined inhibitors behave additive rather than synergistic effects.[79] Our systematic homologous-series data quantify that trend: each additional methylene lowers the core free-energy barrier for both carboxylic acids and alcohols while simultaneously widening the interfacial free energy well. Weak acids such as acetic acid remain confined to the head-group region and face barriers to the bilayer core, so their principal effect is electrostatic crowding rather than deep structural disruption. Adding methylene units or reducing polarity diminishes that barrier and widens the interfacial well, producing a permeability hierarchy that mirrors hydrophobicity-based toxicity rankings from fermenter screens. These atomistic insights convert qualitative stress phenotypes into quantitative design rules, highlighting levers such as hopanoid-pathway optimisation, stabilisation of CFA synthase, and interface-targeted efflux pumps, which can now be integrated into rational strain-engineering and process-design strategies to create genuinely inhibitor-resistant *Z. mobilis* and robust microbial platforms essential for a circular bioeconomy.

Conclusions

The MD simulations presented in this study provide a detailed characterization of the interactions between small molecules and the *Z. mobilis* membrane. The increased disorder and thinning at higher concentrations suggest potential compromises in membrane integrity, impacting cell viability in the biorefinery to make fuels or materials. The probability density distributions reveal distinct localization patterns, with hydrophilic molecules preferentially residing at the membrane interface and hydrophobic molecules displaying greater penetration into the bilayer core. This suggests that there are fundamental limits to the concentration for organic molecules in solution before the membrane becomes leaky.

The free energy profiles further quantify these trends. We observed that increasing the alkyl chain length reduces the energetic barrier to cross the lipid membrane, demonstrating that permeability through *Z. mobilis* membranes is primarily governed by solute hydrophobicity as it is in other systems. These findings have significant implications for biofuel production, microbial tolerance to fermentation inhibitors, and membrane adaptation strategies. Understanding the molecular determinants of permeability in *Z. mobilis* provides a foundation for future efforts in strain engineering, aiming to enhance robustness and optimize yield in industrial fermentation processes.

Given these insights, future efforts to improve stress tolerance in *Z. mobilis* may employ several adaptive strategies to mitigate the impact of toxic compounds on membrane integrity. These could include modifications to lipid composition to alter bilayer permeability, upregulation of efflux transporters to expel toxic metabolites, and adjustments in membrane properties to reduce interactions with inhibitory compounds. Such adaptations could be leveraged in metabolic engineering efforts to enhance microbial robustness in the biorefinery context to enable robust biofuel and bioproduct formation.

Data Availability

The reduced directory structure that includes analysis scripts, inputs and selected raw outputs used for this publication is available at 10.5281/zenodo.16375964.

Acknowledgements

This material is based upon work supported by the Great Lakes Bioenergy Research Center (GLBRC), U.S. Department of Energy, Office of Science, Biological and Environmental Research Program under Award Number DE-SC0018409. This work utilized computational resources and services provided by the Institute for Cyber-Enabled Research at Michigan State University. Additional computational support was provided by Delta at the National Center for Supercomputing Applications (NCSA) through allocation BIO210061 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS)[80] program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603 and #2138296.

Supporting information

The Supporting Information is in this instance a PDF file with additional analysis that supports the narrative presented here. The SI contains:

- Representative simulation setup of the *Z. mobilis* membrane system.
- The timeseries data for the control membrane in the presence and absence of Hopanoids (Table S1 and Figure S2-S3).
- The timeseries data for the interaction of small molecules with the hopanoid rich membrane (Figure S4-S48).
- Grouped lipid enrichment indices for hopanoids (HOP) and phospholipid tail classes in *Z. mobilis* membranes.
- Probability density distributions (Figure S51 and S52).
- Effect of small molecules on the membrane properties when no Hopanoids were present in the membrane (Figure S53-S72).
- Permeability values of different compounds at five different concentration range (0.5 mol% to 2.5 mol%) calculated by counting the complete leaflet-to-leaflet crossings (Figure S73).
- Number of membrane-crossing events and calculated permeability coefficients ($\log P$, $\mathrm{cm\,s^{-1}}$) for different small molecules at five bulk mole fractions. Translocation events are counted over a period of 1000 ns for three independent runs (Table S3).
- Area per lipid, membrane thickness, Diffusion coefficient and S_{CH} at two inhibitor concentrations with and without the hopanoid in membrane (Table S4-S7).

References

[1] Wu, X.; McLaren, J.; Madl, R.; Wang, D. In *Sustainable Biotechnology*; Singh, O. V., Harvey, S. P., Eds.; Springer Netherlands: Dordrecht, 2010; pp 19–41.

[2] Hoang, A. T.; Ong, H. C.; Fattah, I. M. R.; Chong, C. T.; Cheng, C. K.; Sakthivel, R.; Ok, Y. S. *Fuel Processing Technology* **2021**, *223*, 106997.

[3] Saravanan, A.; Senthil Kumar, P.; Jeevanantham, S.; Karishma, S.; Vo, D.-V. N. *Bioresource Technology* **2022**, *344*, 126203.

[4] Srisawat, P.; Higuchi-Takeuchi, M.; Numata, K. *Polym J* **2022**, *54*, 1139–1151.

[5] Decker, S. R.; Brunecky, R.; Yarbrough, J. M.; Subramanian, V. *Front. Ind. Microbiol.* **2023**, *1*, 1202269.

[6] Klinke, H. B.; Thomsen, A. B.; Ahring, B. K. *Appl Microbiol Biotechnol* **2004**, *66*, 10–26.

[7] Laurens, L. M. L.; Markham, J.; Templeton, D. W.; Christensen, E. D.; Van Wychen, S.; Vadelius, E. W.; Chen-Glasser, M.; Dong, T.; Davis, R.; Pienkos, P. T. *Energy Environ. Sci.* **2017**, *10*, 1716–1738.

[8] Kumar, V.; Yadav, S. K.; Kumar, J.; Ahluwalia, V. *Bioresource Technology* **2020**, *299*, 122633.

[9] Kaur, R.; Tyagi, R. D.; Zhang, X. *Environmental Research* **2020**, *182*, 109094.

[10] Yu, Y.; Wu, J.; Ren, X.; Lau, A.; Rezaei, H.; Takada, M.; Bi, X.; Sokhansanj, S. *Renewable and Sustainable Energy Reviews* **2022**, *154*, 111871.

[11] Balasundaram, G.; Banu, R.; Varjani, S.; Kazmi, A.; Tyagi, V. K. *Chemosphere* **2022**, *291*, 132930.

[12] Almeida, J. R.; Modig, T.; Petersson, A.; Hähn-Hägerdal, B.; Lidén, G.; Gorwa-Grauslund, M. F. *J of Chemical Tech & Biotech* **2007**, *82*, 340–349.

[13] Delgenes, J.; Moletta, R.; Navarro, J. *Enzyme and Microbial Technology* **1996**, *19*, 220–225.

[14] Zaldivar, J.; Martinez, A.; Ingram, L. O. *Biotechnol. Bioeng.* **1999**, *65*, 24–33.

[15] Todhanakasem, T.; Wu, B.; Simeon, S. *World J Microbiol Biotechnol* **2020**, *36*, 112.

[16] Dai, Y.; Wu, B.; Liu, P.; Chen, M.; Song, C.; Gou, Q.; Liu, R.; Xu, Y.; Hu, G.; He, M. *GCB Bioenergy* **2021**, *13*, 1894–1907.

[17] Li, Y.; Xu, Y.; Xue, Y.; Yang, S.; Cheng, Y.; Zhu, W. *Biomass and Bioenergy* **2022**, *161*, 106454.

[18] Rogers, P. L.; Skotnicki, M. L.; Lee, K. J.; Lee, J. H. *Critical Reviews in Biotechnology* **1983**, *1*, 273–288.

[19] Panesar, P. S.; Marwaha, S. S.; Kennedy, J. F. *J of Chemical Tech & Biotech* **2006**, *81*, 623–635.

[20] Rogers, P. L.; Jeon, Y. J.; Lee, K. J.; Lawford, H. G. In *Biofuels*; Olsson, L., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2007; Vol. 108; pp 263–288.

[21] Braga, A.; Gomes, D.; Rainha, J.; Amorim, C.; Cardoso, B. B.; Gudiña, E. J.; Silvério, S. C.; Rodrigues, J. L.; Rodrigues, L. R. *Bioresour. Bioprocess.* **2021**, *8*, 128.

[22] Boismier, E. C.; Aboulnaga, E. A.; TerAvest, M. A. *Current Opinion in Biotechnology* **2025**, *92*, 103257.

[23] Seo, J.-S.; Chong, H.; Park, H. S.; Yoon, K.-O.; Jung, C.; Kim, J. J.; Hong, J. H.; Kim, H.; Kim, J.-H.; Kil, J.-I.; Park, C. J.; Oh, H.-M.; Lee, J.-S.; Jin, S.-J.; Um, H.-W.; Lee, H.-J.; Oh, S.-J.; Kim, J. Y.; Kang, H. L.; Lee, S. Y.; Lee, K. J.; Kang, H. S. *Nat Biotechnol* **2005**, *23*, 63–68.

[24] Yang, S.; Pappas, K. M.; Hauser, L. J.; Land, M. L.; Chen, G.-L.; Hurst, G. B.; Pan, C.; Kouvelis, V. N.; Typas, M. A.; Pelletier, D. A.; Klingeman, D. M.; Chang, Y.-J.; Samatova, N. F.; Brown, S. D. *Nat Biotechnol* **2009**, *27*, 893–894.

[25] Boender, L. G. M.; De Hulster, E. A. F.; Van Maris, A. J. A.; Daran-Lapujade, P. A. S.; Pronk, J. T. *Appl Environ Microbiol* **2009**, *75*, 5607–5614.

[26] Zhao, N.; Bai, Y.; Liu, C.-G.; Zhao, X.-Q.; Xu, J.-F.; Bai, F.-W. *Biotechnology Journal* **2014**, *9*, 362–371.

[27] Ranatunga, T. D.; Jervis, J.; Helm, R. F.; McMillan, J. D.; Wooley, R. J. *Enzyme and Microbial Technology* **2000**, *27*, 240–247.

[28] Agrawal, R.; Verma, A.; Singhania, R. R.; Varjani, S.; Di Dong, C.; Kumar Patel, A. *Bioresource Technology* **2021**, *332*, 125042.

[29] Zhai, R.; Hu, J.; Jin, M. *Biotechnology Advances* **2022**, *61*, 108044.

[30] Guo, H.; Zhao, Y.; Chang, J.-S.; Lee, D.-J. *Bioresource Technology* **2022**, *361*, 127666.

[31] Franden, M. A.; Pienkos, P. T.; Zhang, M. *Journal of Biotechnology* **2009**, *144*, 259–267.

[32] Rivera Vazquez, J.; Trujillo, E.; Williams, J.; She, F.; Getahun, F.; Callaghan, M. M.; Coon, J. J.; Amador-Noguez, D. *Biotechnol Biofuels* **2024**, *17*, 14.

[33] Qiu, M.; Shen, W.; Yan, X.; He, Q.; Cai, D.; Chen, S.; Wei, H.; Knoshaug, E. P.; Zhang, M.; Himmel, M. E.; Yang, S. *Biotechnol Biofuels* **2020**, *13*, 15.

[34] He, M.-x.; Wu, B.; Shui, Z.-x.; Hu, Q.-c.; Wang, W.-g.; Tan, F.-r.; Tang, X.-y.; Zhu, Q.-l.; Pan, K.; Li, Q.; Su, X.-h. *Appl Microbiol Biotechnol* **2012**, *95*, 189–199.

[35] Yang, S.; Franden, M. A.; Brown, S. D.; Chou, Y.-C.; Pienkos, P. T.; Zhang, M. *Biotechnol Biofuels* **2014**, *7*, 140.

[36] Yi, X.; Gu, H.; Gao, Q.; Liu, Z. L.; Bao, J. *Biotechnol Biofuels* **2015**, *8*, 153.

[37] Dombach, J. L.; Quintana, J. L. J.; Nagy, T. A.; Wan, C.; Crooks, A. L.; Yu, H.; Su, C.-C.; Yu, E. W.; Shen, J.; Detweiler, C. S. *PLoS Pathog* **2020**, *16*, e1009119.

[38] Ganesan, N.; Mishra, B.; Felix, L.; Mylonakis, E. *Microbiol Mol Biol Rev* **2023**, *87*, e00037–22.

[39] Hannesschlaeger, C.; Horner, A.; Pohl, P. *Chem. Rev.* **2019**, *119*, 5922–5953.

[40] Lee, J.; Patel, D. S.; Ståhle, J.; Park, S.-J.; Kern, N. R.; Kim, S.; Lee, J.; Cheng, X.; Valvano, M. A.; Holst, O.; Knirel, Y. A.; Qi, Y.; Jo, S.; Klauda, J. B.; Widmalm, G.; Im, W. *Journal of Chemical Theory and Computation* **2019**, *15*, 775–786.

[41] Hermans, M. A.; Neuss, B.; Sahm, H. *J Bacteriol* **1991**, *173*, 5592–5595.

[42] Brenac, L.; Baidoo, E. E.; Keasling, J. D.; Budin, I. *Molecular Microbiology* **2019**, *112*, 1564–1575.

[43] Sarkar, D.; Kulke, M.; Vermaas, J. V. *Biomolecules* **2023**, *13*, 107.

[44] Lee, C. T.; Comer, J.; Herndon, C.; Leung, N.; Pavlova, A.; Swift, R. V.; Tung, C.; Rowley, C. N.; Amaro, R. E.; Chipot, C.; Wang, Y.; Gumbart, J. C. *Journal of Chemical Information and Modeling* **2016**, *56*, 721–733.

[45] Vermaas, J. V.; Dixon, R. A.; Chen, F.; Mansfield, S. D.; Boerjan, W.; Ralph, J.; Crowley, M. F.; Beckham, G. T. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 23117–23123.

[46] Raza, S.; Miller, M.; Hamberger, B.; Vermaas, J. V. *J. Phys. Chem. B* **2023**, *127*, 1144–1157.

[47] Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E. *The Journal of Chemical Physics* **2020**, *153*, 044130.

[48] Yu, Y.; Klauda, J. B. *J. Phys. Chem. B* **2020**, *124*, 6797–6812.

[49] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. *J Comput Chem* **2010**, *31*, 671–690.

[50] Farah, K.; Müller-Plathe, F.; Böhm, M. C. *ChemPhysChem* **2012**, *13*, 1127–1151.

[51] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

[52] Dünweg, B., Landau, D. P., Milchev, A. I., Eds. *Computer Simulations of Surfaces and Interfaces*; Springer Netherlands: Dordrecht, 2003.

[53] Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613–4621.

[54] Miyamoto, S.; Kollman, P. A. *J Comput Chem* **1992**, *13*, 952–962.

[55] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[56] van der Walt, S.; Colbert, S. C.; Varoquaux, G. *Comput. Sci. Eng.* **2011**, *13*, 22–30.

[57] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

[58] Hunter, J. D. *Computing in Science and Engineering* **2007**, *9*, 90–95.

[59] Venable, R. M.; Krämer, A.; Pastor, R. W. *Chem. Rev.* **2019**, *119*, 5954–5997.

[60] Alwarawrah, M.; Dai, J.; Huang, J. *J. Phys. Chem. B* **2010**, *114*, 7516–7523.

[61] Gurtovenko, A. A.; Anwar, J. *J. Phys. Chem. B* **2009**, *113*, 1983–1992.

[62] Sun, D.; Peyear, T. A.; Bennett, W. F. D.; Holcomb, M.; He, S.; Zhu, F.; Lightstone, F. C.; Andersen, O. S.; Ingólfsson, H. I. *J. Med. Chem.* **2020**, *63*, 11809–11818.

[63] Zizzi, E. A.; Cavaglià, M.; Tuszynski, J. A.; Deriu, M. A. *iScience* **2022**, *25*, 103946.

[64] Barnoud, J.; Rossi, G.; Marrink, S. J.; Monticelli, L. *PLoS Comput Biol* **2014**, *10*, e1003873.

[65] Levental, I.; Lyman, E. *Nat Rev Mol Cell Biol* **2023**, *24*, 107–122.

[66] Kulke, M.; Weraduwage, S. M.; Sharkey, T. D.; Vermaas, J. V. *Plant Cell & Environment* **2023**, *46*, 2273–2589.

[67] Smith, P.; Lorenz, C. D. *J. Chem. Theory Comput.* **2021**, *17*, 5907–5919.

[68] Franden, M. A.; Pilath, H. M.; Mohagheghi, A.; Pienkos, P. T.; Zhang, M. *Biotechnol Biofuels* **2013**, *6*, 99.

[69] Jilani, S. B.; Olson, D. G. *Microb Cell Fact* **2023**, *22*, 221.

[70] Ghorbani, M.; Wang, E.; Krämer, A.; Klauda, J. B. *The Journal of Chemical Physics* **2020**, *153*.

[71] He, M.-x.; Wu, B.; Shui, Z.-x.; Hu, Q.-c.; Wang, W.-g.; Tan, F.-r.; Tang, X.-y.; Zhu, Q.-l.; Pan, K.; Li, Q.; Su, X.-h. *Biotechnol Biofuels* **2012**, *5*, 75.

[72] He, M.; Wu, B.; Qin, H.; Ruan, Z.; Tan, F.; Wang, J.; Shui, Z.; Dai, L.; Zhu, Q.; Pan, K.; Tang, X.; Wang, W.; Hu, Q. *Biotechnol Biofuels* **2014**, *7*, 101.

[73] Yang, S.; Pan, C.; Hurst, G. B.; Dice, L.; Davison, B. H.; Brown, S. D. *Front. Microbiol.* **2014**, *5*.

[74] Yang, S.; Mohagheghi, A.; Franden, M. A.; Chou, Y.-C.; Chen, X.; Dowe, N.; Himmel, M. E.; Zhang, M. *Biotechnol Biofuels* **2016**, *9*, 189.

[75] *Advances in Microbial Physiology*; Elsevier, 1993; pp 247–273.

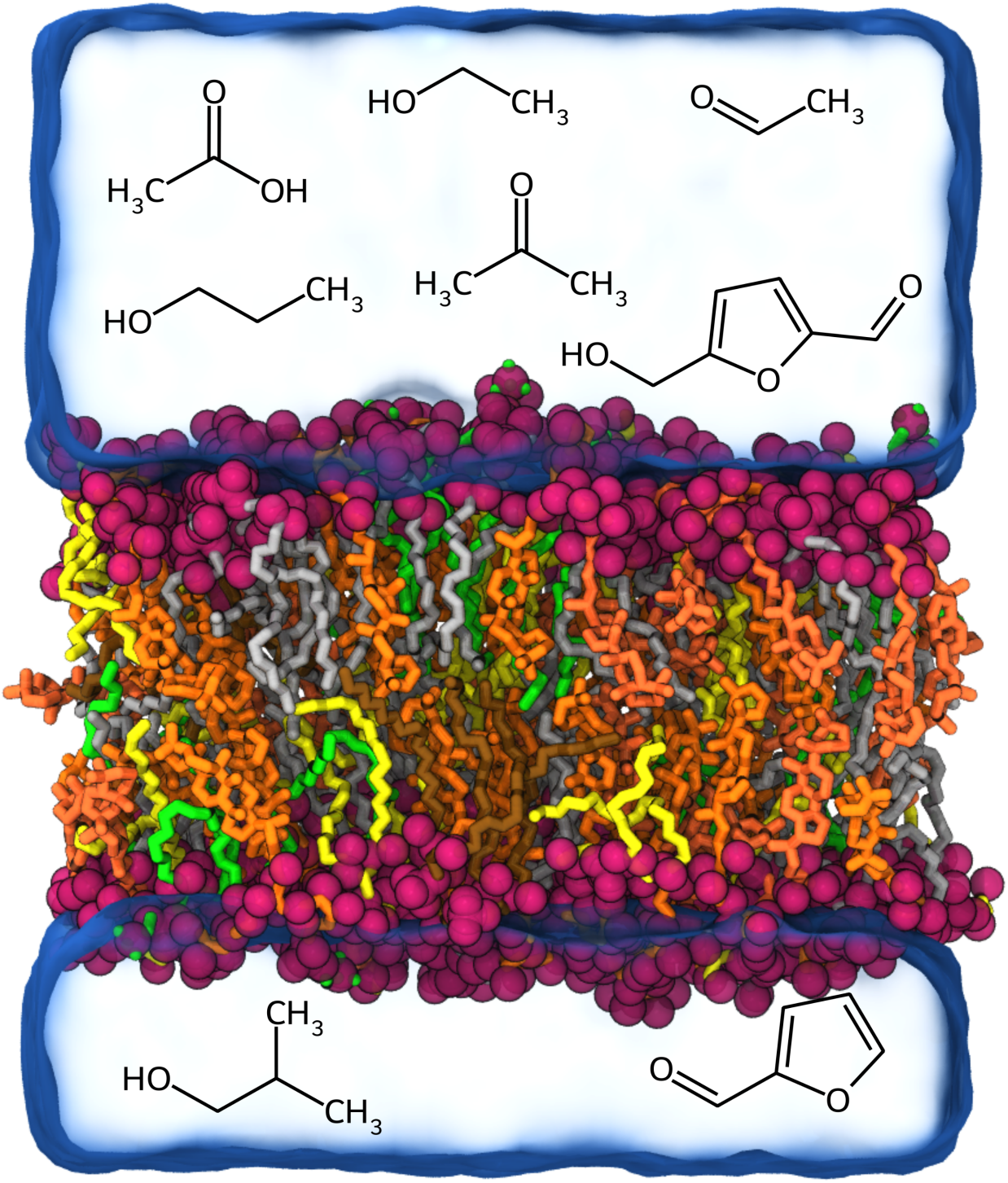
[76] Sáenz, J. P.; Grosser, D.; Bradley, A. S.; Lagny, T. J.; Lavrynenko, O.; Broda, M.; Simons, K. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 11971–11976.

[77] Carey, V. C.; Ingram, L. O. *J Bacteriol* **1983**, *154*, 1291–1300.

[78] Moreau, R. A.; Powell, M. J.; Fett, W. F.; Whitaker, B. D. *Current Microbiology* **1997**, *35*, 124–128.

[79] Yang, X.; Berthold, F.; Berglund, L. A. *Biomacromolecules* **2018**, *19*, 3020–3029.

[80] Boerner, T. J.; Deems, S.; Furlani, T. R.; Knuth, S. L.; Towns, J. ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support. Practice and Experience in Advanced Research Computing. Portland OR USA, 2023; pp 173–176.

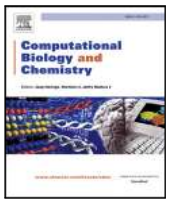


Computational Biology and Chemistry 113 (2024) 108237

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



Statistical analysis of the unique characteristics of secondary structures in proteins

Nitin Kumar Singh [a], Manish Agarwal [b], Mithun Radhakrishna [a,c,*]

[a] Department of Chemical Engineering, Indian Institute of Technology (IIT) Gandhinagar, Palaj, Gujarat 382355, India
[b] Computer Services Centre, Indian Institute of Technology (IIT) Delhi, Hauz Khas, New Delhi, Delhi 110016, India
[c] Center for Biomedical Engineering, Indian Institute of Technology (IIT) Gandhinagar, Palaj, Gujarat 382355, India

ARTICLE INFO

Dataset link: https://github.com/mr2972/helix_uniqueness

Keywords:
Proteins
Secondary structure
α-helix
Amino acids
Machine Learning

ABSTRACT

Protein folding is a complex process influenced by the primary sequence of amino acids. Early studies focused on understanding whether the specificity or the conservation of properties of amino acids was crucial for folding into secondary structures such as α-helices, β-sheets, turns, and coils. However, with the advent of artificial intelligence (AI) and machine learning (ML), the emphasis has shifted towards the precise nature and occurrence of specific amino acids. In our study, we analyzed a large set of proteins from diverse organisms to identify unique features of secondary structures, particularly in terms of the distribution of polar, non-polar, and charged amino acid residues. We found that α-helices tend to have a higher proportion of charged and non-polar groups compared to other secondary structures and that the presence of oppositely charged amino acid residues in helices stabilizes them, facilitating the formation of longer helices. These characteristics are distinct to α-helices. This study offers valuable insights for researchers in the field of protein design, enabling the de-novo creation of short helical peptides for a range of applications. We have also developed a web server for extensive analysis of proteins from different databases. The web server is housed at https://proseqanalyser.iitgn.ac.in/

1. Introduction

Proteins serve as crucial drivers of biological activity, orchestrating numerous cellular processes. Traditionally, it has been understood that a protein's function is intricately tied to its structure. Consequently, unraveling the intricate relationship between sequence, structure, and spatial folding has remained a subject of intense scrutiny for decades (Marks et al., 2012; Anfinsen, 1973, 1962; Watson et al., 2005). Proteins exhibit hierarchical organization, spanning primary, secondary, tertiary, and quaternary levels of structure (Nelson et al., 2008). Despite this complexity, the number of known protein primary structure sequences has surged due to advances in sequencing technology, far outpacing the experimental determination of three-dimensional protein structures (Berman et al., 2000). Experimental techniques such as nuclear magnetic resonance (NMR), X-ray crystallography, and cryo-electron microscopy (cryo-EM) are commonly employed to probe protein structures (Wuethrich, 1989; Ilari and Savino, 2008; Jonic and Vénien-Bryan, 2009). However, each method has its limitations. NMR, for instance, is typically constrained to smaller proteins, those under 50 kDa (Kainosho et al., 2006). Meanwhile, X-ray crystallography may struggle with proteins featuring large flexible regions or undergoing conformational changes (DePristo et al., 2004). Cryo-EM, while promising, presents its own challenges, including computationally intensive tasks like image processing, 3D reconstruction, and model refinement, which demand significant time and computational resources (Bonomi and Vendruscolo, 2019).

In recent years, the rise of machine learning techniques has revolutionized protein structure prediction (Lindorff-Larsen and Kragelund, 2021). Noteworthy methods include Rosetta (Rohl et al., 2004), which employs physics-based energy functions and optimization algorithms; AlphaFold (Jumper et al., 2021), developed by DeepMind, that leverages deep learning models with attention mechanisms for accurate predictions; and others like SPIDER3, RaptorX, and Robetta, that harnesses machine learning algorithms to enhance prediction accuracy and efficiency (Heffernan et al., 2018; Jing et al., 2024; Kim et al., 2004). These methods have showcased remarkable precision, often surpassing traditional approaches in protein structure prediction.

The direct prediction of tertiary structure from the primary sequence has long posed a computational challenge, acknowledged to be NP-complete due to the vast conformational space available (Berger

* Corresponding author.
E-mail address: mithunr@iitgn.ac.in (M. Radhakrishna).

https://doi.org/10.1016/j.compbiolchem.2024.108237
Received 21 June 2024; Received in revised form 11 September 2024; Accepted 30 September 2024
Available online 5 October 2024

and Leighton, 1998). Protein secondary structures are widely recognized as pivotal elements influencing protein folding, structure, and design. Enhanced predictions in this realm have demonstrated efficacy in refining threading methods and are fundamental to numerous ab initio techniques employed in protein structure prediction (Zhang et al., 2018).

Previous research has demonstrated the significance of hydrophobic and charged residues in the stability of proteins and their role in protein–protein interactions. Tripathi et al. (2015), Loladze and Makhatadze (2011) and Strickler et al. (2006) investigated the impact of charge–charge interactions on protein stability, while various other studies have examined the influence of charged amino acids on the surface in regulating protein–protein interactions (Sheinerman and Honig, 2002; Moreira et al., 2007; Glaser et al., 2001). Conversely, the work of Pace et al. (2011), Pace (1992), Privalov and Gill (1988), and Shortle et al. (1990) has concentrated on the role of hydrophobic residues. Although most research has analyzed entire proteins, few studies address the formation of individual secondary structure elements within a protein's globular structure. These studies on secondary structure elements have predominantly relied on identifying patterns in the amino acid sequence that indicate secondary structure elements such as α-helices and β-strands (Dalal et al., 1997). Previous research indicates that a combination of local and non-local structural preferences dictates an amino acid's propensity for a specific secondary structure. The Hydrophobic Polar (HP) model, though simplistic and lacking chemical intricacies of real protein models, offers a framework for studying the role of hydrophobic interactions in protein folding and stability (Dill et al., 1995). Empirical observations often informed the discernment of secondary structure patterns. For instance, α-helical regions typically exhibit hydrophobic periodicities of 3 to 4, while β-strands demonstrate periodicities of 2 (Mandel-Gutfreund and Gregoret, 2002). West and Hecht (1995) demonstrated the significance of binary patterning of polar and non-polar amino acids in α-helix and β-sheet formation. Moreover, statistical analyses of protein databases yielded insights into the frequency and distribution of these patterns across various protein families and organisms. Researchers inferred the likelihood of certain regions adopting specific secondary structure conformations by examining the occurrence of particular amino acid motifs within known structures. To form an α-helix or a β-strand, the linear amino acid sequence must exhibit a specific arrangement of polar and non-polar amino acid residues aligning with the structural repeat characteristic of that secondary structure. Shoemaker et al. (1985) illustrated the feasibility of designing and characterizing straightforward α-helical peptides while highlighting specific side chain interactions. Lyu et al. (1991) produced two α-helical peptides exclusively through ion-pairing between Glumatic acid and Lysine amino acid residues positioned at suitable intervals. While these studies laid the foundation for more advanced methods of secondary structure prediction and provided crucial insights into the principles governing protein folding and stability there are no studies which are conducted to identify the unique characteristics of the different secondary structures on a large set of proteins.

In the current manuscript, we aimed to explore and identify fundamental characteristics shared by different secondary structures, including α-helices, β-sheets, turns, and coils. We investigated a range of parameters, such as hydrophobicity, hydrophilicity, charge density, and the presence of different patterns, such as Like Charge Regions (LCR), within the individual secondary structure elements. To accomplish this, we compiled an extensive dataset comprising proteins sourced from diverse organisms, ensuring a broad representation of structural variations. By leveraging the insights gained from our statistical analysis, we aimed to enhance our understanding of the distinct features associated with α-helices and their differentiation from other structural motifs within proteins.

2. Methodology

1. One thousand PDB IDs were randomly retrieved from the RCSB database (Rose et al., 2010) from each of the following categories: Bacteria, Viruses, and Humans, for analysis. All the PDB IDs are provided in the Supplementary Information.
2. The secondary structure content of the protein was analyzed using the STRIDE program (Heinig and Frishman, 2004). It recognizes secondary structural elements in proteins from their atomic coordinates. It utilizes both hydrogen bond energy and main chain dihedral angles to classify the individual amino acid residues into α-helices, β-strand, turns and coils. All the secondary structure elements belonging to a particular group (such as α-helices, β-strand, turns and coils) were grouped together, and the fractional composition of each of the twenty amino acid residues occurring in these respective secondary structures was computed. For ex., all the α-helical sequences from different proteins were combined into a α-helix dataset.
3. The above datasets were further divided into two classes: the secondary structures, which contain no charged amino acids residue, and the secondary structures, which contain at least one charged amino acid residue.
4. The amino acids were classified into four categories, namely polar (Serine, Threonine, Cysteine, Asparagine and Glutamine), non-polar (Glycine, Alanine, Valine, Leucine, Isoleucine, Proline, Phenylalanine, Methionine, Tyrosine and Tryptophan), positive (Lysine, Arginine and Histidine) and negatively charged (Aspartic acid and Glutamic acid) amino acids (David et al., 2000).
5. Analysis was performed to calculate the distribution of chemical composition, the length of the secondary structures, the fraction of amino acids in each category and charge density. Charge density(λ) was calculated using

$$\lambda = \frac{N^+ + N^-}{N}$$

where N^+ represents the number of positive amino acid residues, N^- the number of negative residues and N is the total number of amino acid residues in a given secondary structure.
6. In order to identify certain patterns, we calculated the distance between the two oppositely charged groups, i.e. "pn-np", "np-np", "pn-pn" or "np-pn" (here "p": positive and "n": negative), for all the secondary structure elements.
7. We also performed the Like Charged Regions (LCRs) analysis for all secondary structure elements to evaluate the length of the LCRs and the charged density in the LCR regions.

3. Results and discussions

3.1. Secondary structure analysis

As outlined in the methodology section, four data sets corresponding to four different secondary structures, namely α-helices, β-strands, turns, and coils, were curated. Fig. 1 shows the fraction of each of the twenty amino acid residues averaged over the secondary structures present in each dataset taken from randomly selected 1000 Bacterial proteins. It reveals that α-helices are characterized by a high content of amino acids such as Alanine, Leucine, Glutamine, Lysine, Arginine, and Glutamic acid, whereas β-strands are characterized by a high proportion of the amino acids Valine, Isoleucine, Phenylalanine, and Tyrosine. Amino acids such as Glycine, Proline, Serine, Asparagine, and Aspartic acid showed high content in coils and turns, representing the unstructured regions in proteins. The same distribution was also calculated for 1000 randomly selected proteins from viruses and Homo sapiens, as shown in Figures S-1 and S-2 of the supplementary information. The distribution closely matches that of bacteria, which further

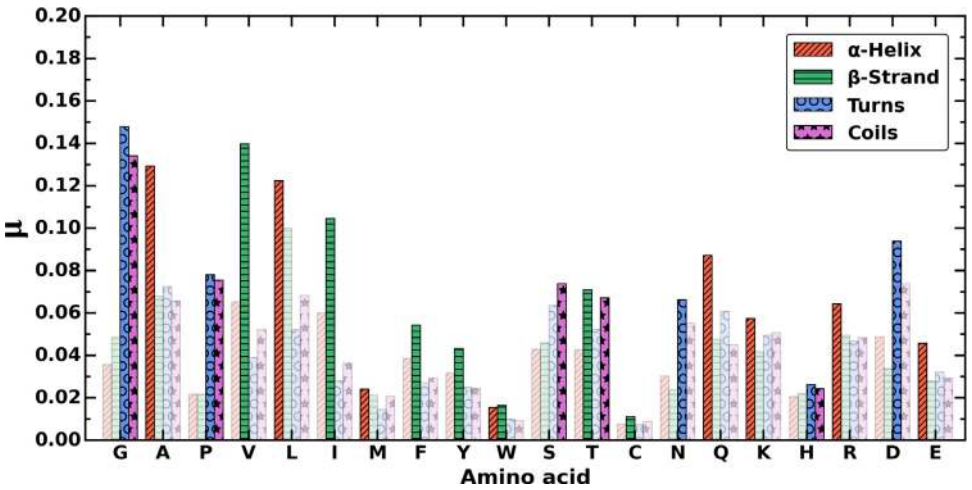


Fig. 1. Mean proportion (μ) of amino acids in different secondary structures in a set of 1000 Bacterial proteins. The bars are adjusted to include a contrast, highlighting the amino acid that is present in a higher proportion within given secondary structures.

emphasizes the conservation of amino acid sequences for different secondary structures during the course of evolution.

Fig. 2 illustrates the results of the amino acid content analysis conducted in accordance with the classification outlined in the methodology section. The x-axis shows the fraction of amino acid residues, and the y-axis shows the probability distribution in different secondary structures. The bar plot showcases the distribution of various amino acid groups, while a line plot connects the midpoints of the histograms, providing an overview of the overall distribution curve.

Fig. 2 reveals several key features associated with each secondary structure. α-helices are distinguished by a higher proportion of non-polar and charged groups compared to other secondary structures, with the distribution of non-polar groups peaking within the range of 0.5 to 0.6 and that of charged amino acid residues peaking within the range of 0.2 to 0.3. Turns and coils, on the other hand, exhibit similar distributions in terms of charged, polar, and non-polar amino acids. A unique characteristic of α-helices is their distinct distribution for both charged and polar amino acid residues, whereas these distributions tend to overlap in β-strands, turns, and coils. Additionally, it is rare to find α-helices and strands where the fraction of charged and polar amino acids exceeds 0.70.

Fig. 3 presents the probability distribution of the charge density for different secondary structures. It is evident from Fig. 3 that α-helices exhibit a bell-shaped distribution, peaking near zero, indicating that the net charge in a helix is generally close to zero. In line with previous studies (Singh et al., 2023), we did not observe any α-helices with an absolute charge density value exceeding 0.5. In contrast, β-strands, coils, and turns display a different distribution, characterized by either a high net positive or negative charge in their secondary structure. These observations highlight the distinct distributional tendencies exhibited by different structural motifs, suggesting underlying structural and functional implications within the analyzed protein sequences. Similar distributions were plotted for viruses and Homo sapiens in Figures S-3 to S-6. It was observed that the distribution closely aligns with that of bacteria, indicating a universal behavior of such distributions across organisms.

We further categorized the secondary structure elements into two groups: (i) secondary structures without any charged amino acid residues, and (ii) secondary structures containing at least one charged amino acid residue. Table 1 presents the analysis for a set of 1000 proteins. The final column displays the ratio of secondary structure elements without charged amino acid residues to those with charged amino acid residues. This ratio was notably lower in α-helices compared to other secondary structures, indicating that the likelihood of forming an α-helical structure without a charged amino acid residue is significantly reduced.

Fig. 4 illustrates the length distribution of all secondary structures in a set of 1000 bacterial proteins. As shown in Fig. 4, α-helices typically have a longer length compared to β-strands, turns, or coils. We also analyzed the length distribution for the two previously mentioned datasets: (i) secondary structures without any charged amino acid residues and (ii) secondary structures with at least one charged amino acid residue. These distributions are depicted in Fig. 5 (a) and (b), respectively. While the distributions of β-strands, turns, and coils exhibit similar patterns for both datasets, the behavior of α-helices is distinct. The distribution indicates that in the absence of any charged amino acid residue, the probability of α-helix formation, peaks in the range of 0 to 10. In contrast, when considering α-helical structures with at least one charged amino acid residue, the peak shifts to a higher range of 10 to 20. Similar calculations were performed for viruses and Homo sapiens in Figures S-7 to S-10. It was observed that the distribution closely aligns with that of bacteria. This data suggests that the presence of charged amino acid residues stabilizes α-helical structures, as also reported in previous studies (Shoemaker et al., 1985; Lyu et al., 1991). However, it is important to note that the net charge in α-helices is close to zero, as explained in Fig. 3.

3.2. Like charged regions (LCR) analysis

LCRs are regions in the amino acid sequence that only contain positively or negatively charged amino acids. Therefore, they represent regions with a net charge. Figure 6 shows the calculation of LCR regions in protein sequences. The red circles represent positively charged amino acid residues, the green circles represent negatively charged amino acid residues, and the blue circles depict neutral amino acid residues. Fig. 7 shows the box plot for the normalized length of LCRs in different secondary structures. It was observed that the α-helical secondary structures have the lowest normalized length of LCRs compared to the other secondary structures. This reinforces the fact that it is difficult to form a α-helix in a region with a net charge. These plots further validate the findings in Fig. 3, which show that the distribution of net charge peaks around zero for α-helices. The box plots also indicate that coils have the highest normalized length of LCRs compared to the other structures, which typically form the unstructured regions of proteins. Similar observations were made for viruses and Homo sapiens as shown in Figures S-11 and S-12.

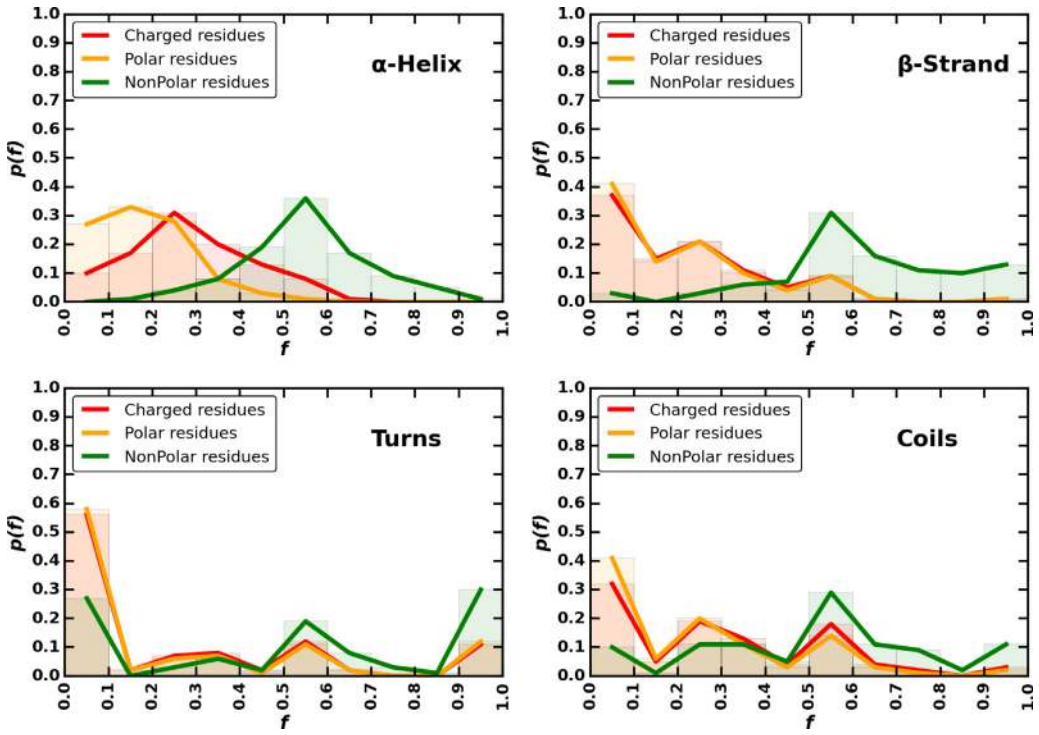


Fig. 2. Probability distribution(p(f)) as a function of the fraction of amino acid residues (f) in different secondary structures (α-helix, β-strand, turns, and coils) in the set of 1000 Bacterial proteins.

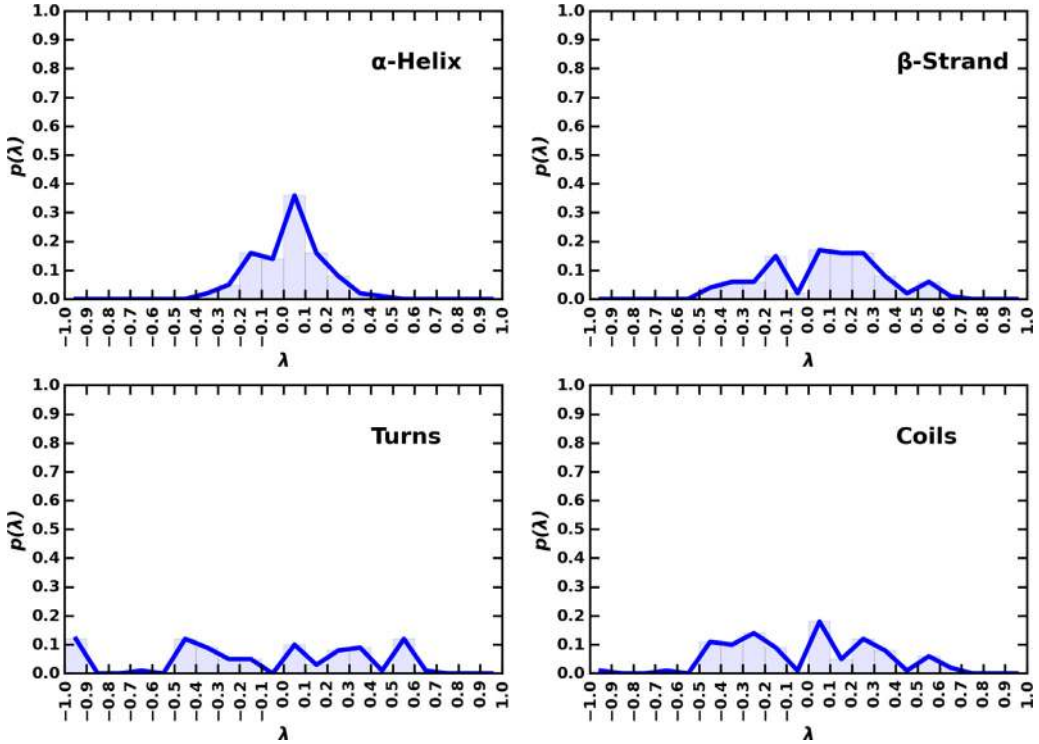


Fig. 3. Probability distribution ($p(\lambda)$ as a function of charge density(λ) in different secondary structures (α-helix, β-strand, turns, and coils) in the set of 1000 Bacterial proteins.

3.3. Distribution of oppositely charged patches

In this section, we carried out a statistical analysis to evaluate the presence of patches of oppositely charged groups of amino acids and their distribution in various secondary structures. The analysis, as shown in Fig. 8, is carried out as follows: firstly, we identified a group of two consecutively occurring oppositely charged amino acid residues i.e., either "pn" or "np" (where "p" represents the positively charged amino acids and "n" represents negatively charged amino acids). Subsequently the distance between this group and the next group of oppositely charged amino acid i.e. "pn" or "np" is calculated.

Fig. 9 shows the probability distribution of this distance. Our analysis revealed that turns and coils did not exhibit any occurrence of such patches, whereas these patches were present in case of β-strands and

Table 1
The number of secondary structures containing charged amino acid residues and those without them, along with their corresponding ratios.

S. No.	Secondary structure	Total secondary structures	Structures with no charged amino acid residues (1)	Structures with charged amino acid residues (2)	Ratio (1/2)
1	α-Helix	20 779	1133	19 646	0.05
2	β-Strand	27 825	10 042	17 783	0.56
3	Coil	51 314	28 940	22 374	1.29
4	Turn	36 726	11 736	24 990	0.46

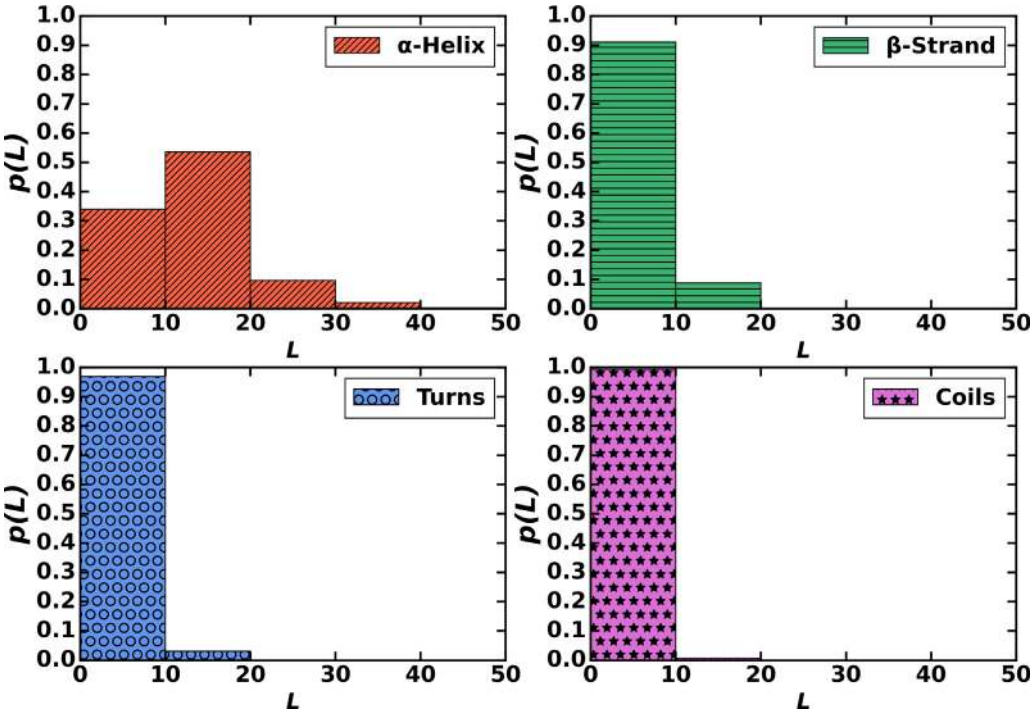


Fig. 4. Length distribution in different secondary structure elements for the set of 1000 Bacterial proteins.

α-helices. However, frequency of occurrence was notably higher in the case of α-helical sequences.

3.4. PSSA webserver

We have developed a webserver, where users can input IDs from various databanks such as RCSB, Uniprot, and Disprot. The analysis is performed on the provided ID to offer valuable insights into the chemical composition of proteins and their respective secondary structures. The web application provides details on the secondary structure content, the proportion of hydrophobic, polar, and charged amino acid residues within each secondary structure, and information on LCRs. The server also provides a contact map to visualize residue proximities within the protein structure, providing insights into protein folding, stability, and interactions. The details about the contact map are mentioned in the supplementary information. The server is housed at https://proseqanalyser.iitgn.ac.in/. We believe that understanding the amino acid composition of individual secondary structures could prove beneficial in de-novo protein design.

Additionally, the secondary structure data from the statistical analysis of 3000 proteins from different organisms were used to train a random forest classifier to identify α-helical sequences among other secondary structures. We integrated the model generated by the Random Forest classifier into our web server, HelixPredictor. This model has been trained on a dataset of secondary structure sequences ranging from 5 to 15 amino acids in length. Users can input amino acid sequences within this range, and the server will indicate whether the provided sequence has a tendency to form an α-helix or not.

4. Comparative analysis across organisms

The Figs. 10–12 demonstrate that the distributions of amino acid residues in the secondary structures of viral, bacterial, and Homo sapiens proteins closely mirror each other, with only minor variations. This consistency suggests a conserved mechanism in the maintenance of secondary structures throughout evolution. The observed similarities across such diverse groups of organisms underscore the fundamental role of various amino acid residues in stabilizing secondary structures, which likely contributes to the preservation of protein function across evolutionary timelines.

5. Conclusions

In this manuscript, we have analyzed a large set of proteins from various organisms and have attempted to identify unique features of the secondary structures in terms of the distribution of polar, non-polar, and charged amino acid residues. Our study finds that α-helices tend to contain a higher proportion of charged and non-polar groups compared to other secondary structures, and that the distribution of charge density typically peaks near zero. The presence of oppositely charged amino acid residues in helices stabilizes them through electrostatic interactions, facilitating the formation of longer helices. These characteristics are distinct to α-helices. Based on these findings, we developed a machine learning algorithm using the Random Forest classifier to distinguish between sequences that have the propensity to form a α-helix. This algorithm is implemented in our web server and

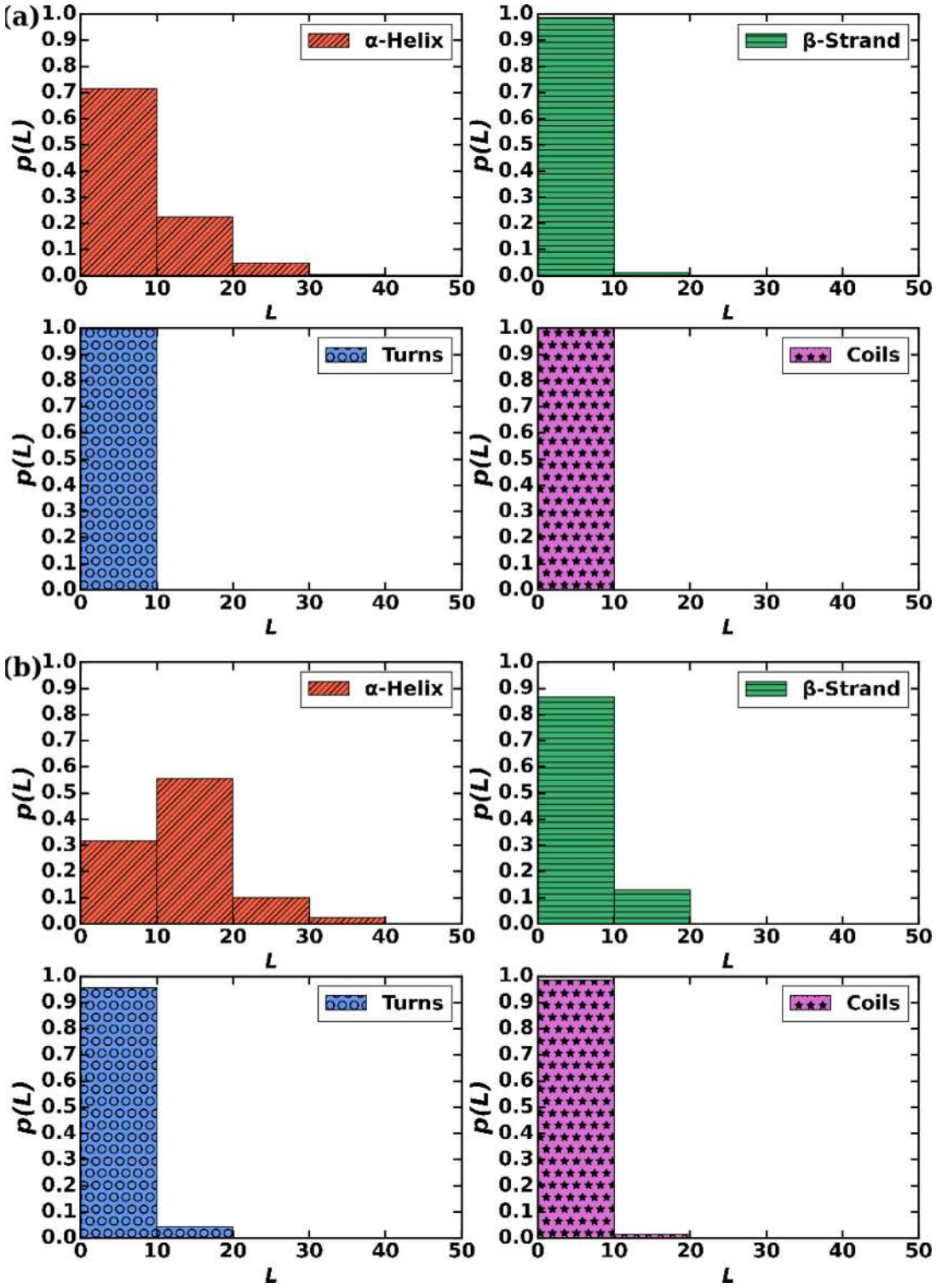


Fig. 5. Length distribution in different secondary structure elements for the set of 1000 Bacterial proteins, (a) without any charged amino acid residues, (b) with at least one charged amino acid residue.

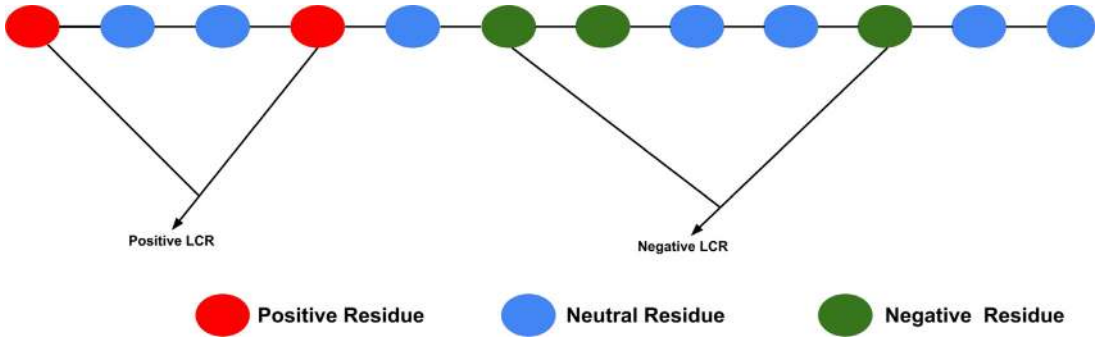


Fig. 6. Pictorial depiction of LCR regions in protein sequences.

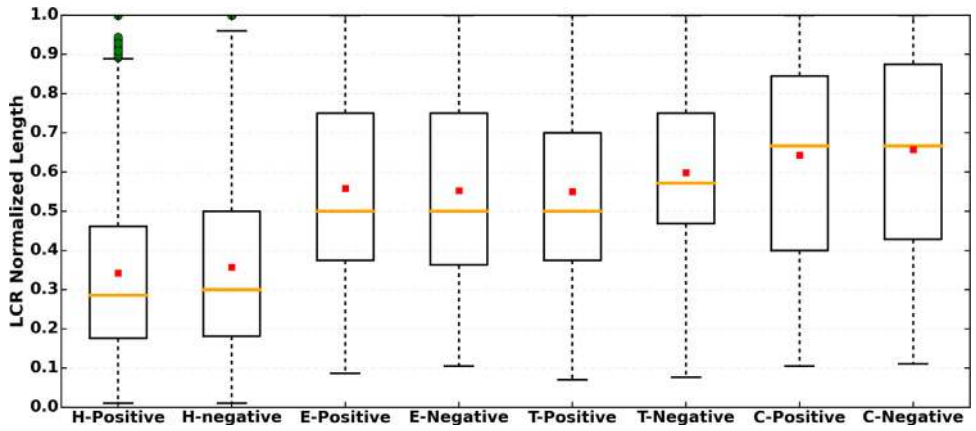


Fig. 7. LCR normalized length in different secondary structures for a set of 1000 Bacterial proteins. Here, X'-positive and X'-negative represent the positive LCR and negative LCR, respectively. X': H = α-helix, E = β-strand, T = Turn and C = Coil.

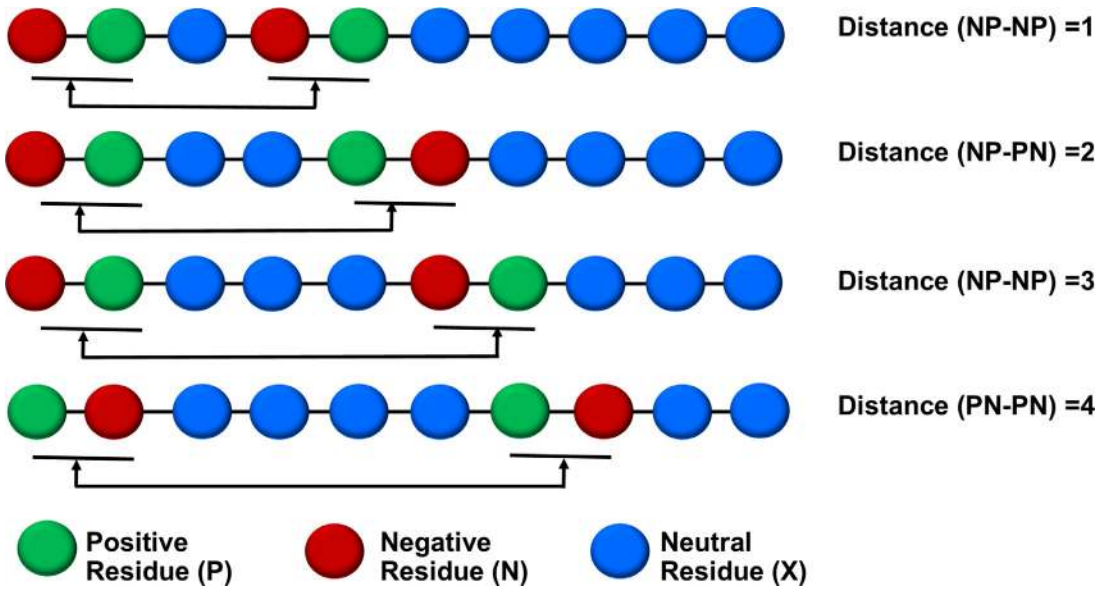


Fig. 8. Calculation of distribution of oppositely charged patches across protein secondary structures. Here, "P" represents positively charged amino acid residues(green), "N" represents negatively charged amino acid residues(red) and "X" represents neutral residues(blue).

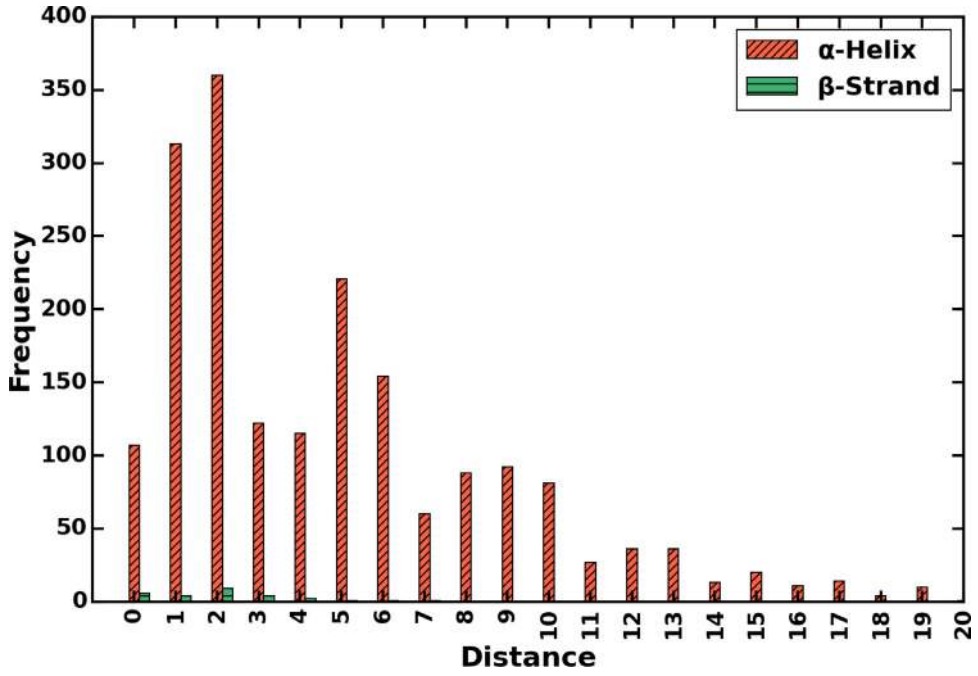


Fig. 9. Distribution of oppositely charged patches across protein secondary structures for a set of 1000 Bacterial proteins.

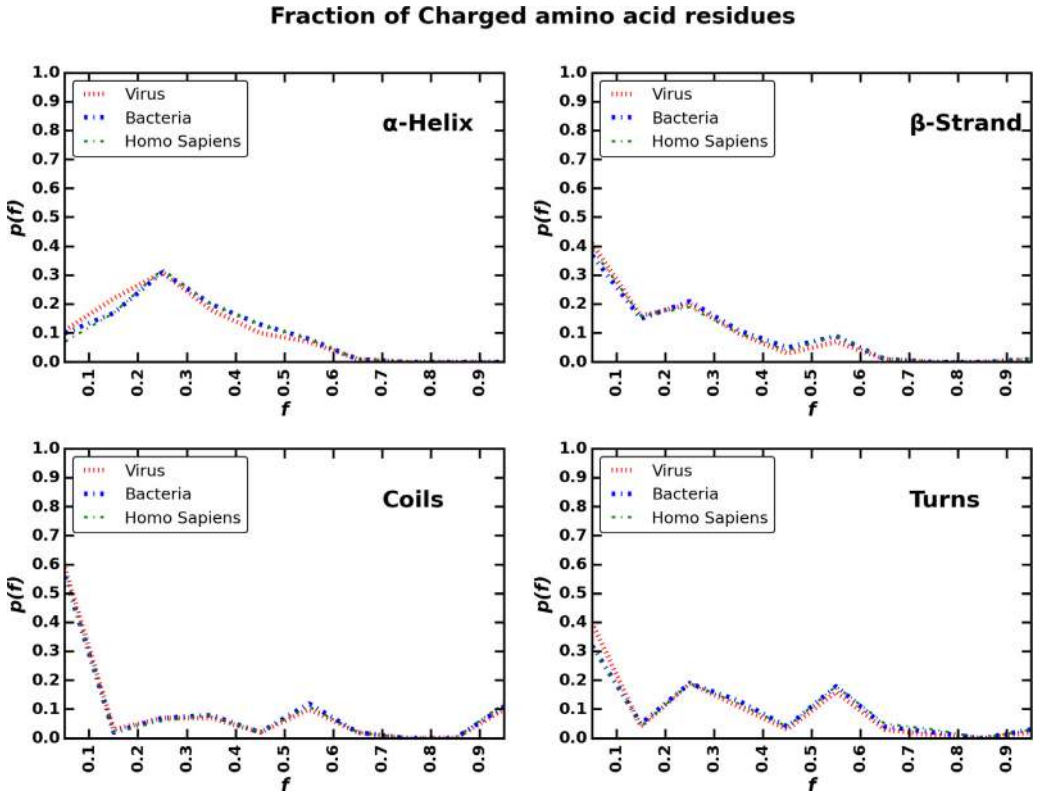


Fig. 10. Probability distribution(p(f)) as a function of the fraction of charged amino acid residues (f) in different secondary structures (α-helix, β-strand, turns, and coils) in 1000 proteins each of Virus, Bacteria and Homo Sapiens.

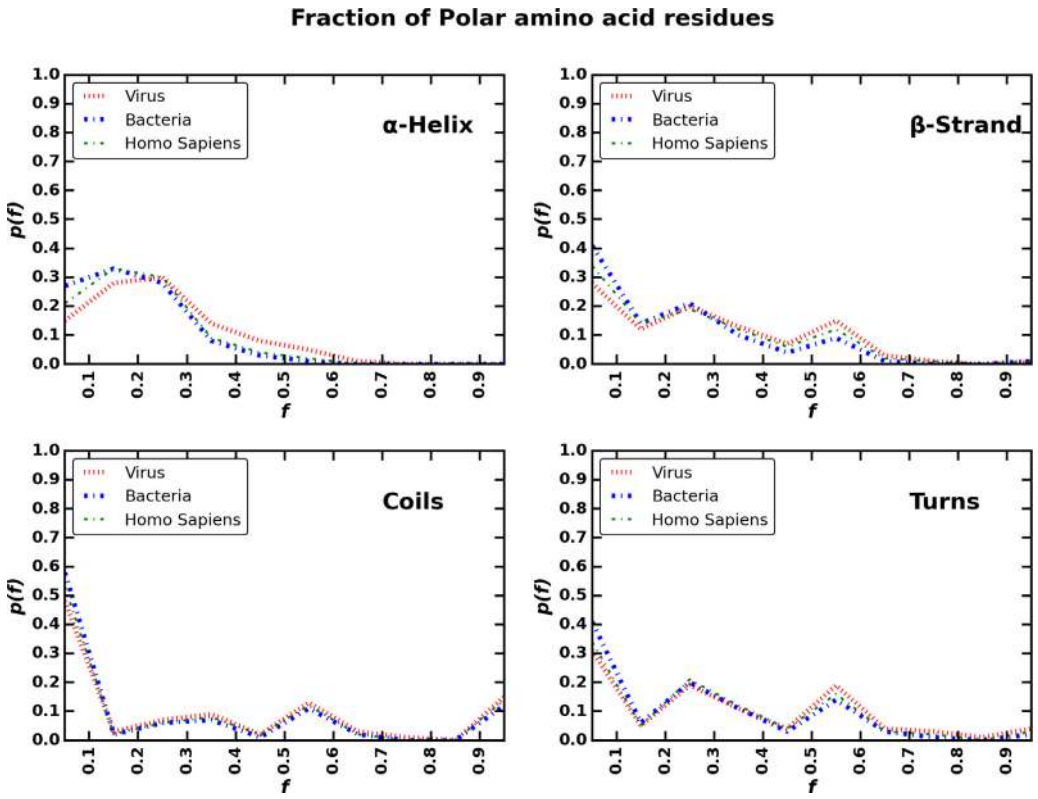


Fig. 11. Probability distribution(p(f)) as a function of the fraction of polar amino acid residues (f) in different secondary structures (α-helix, β-strand, turns, and coils) in 1000 proteins each of Virus, Bacteria and Homo Sapiens.

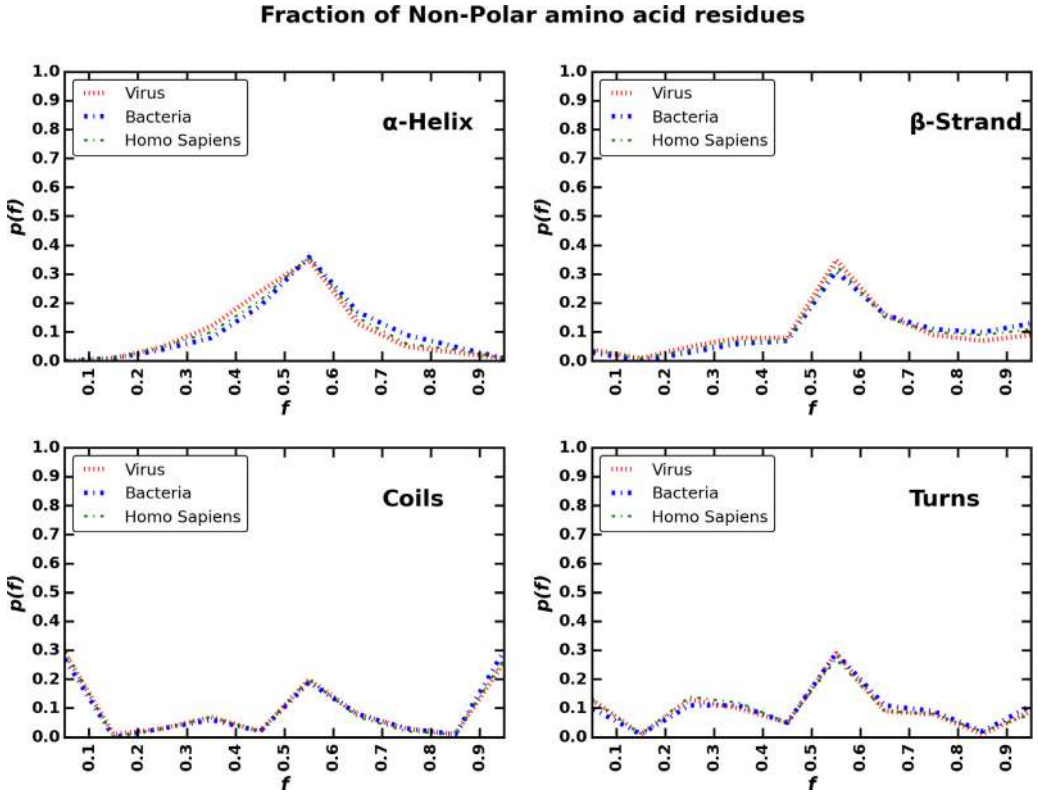


Fig. 12. Probability distribution(p(f)) as a function of the fraction of non-polar amino acid residues (f) in different secondary structures (α-helix, β-strand, turns, and coils) in 1000 proteins each of Virus, Bacteria and Homo Sapiens.

could assist both experimentalists and simulation experts in the de-novo design of short helical peptides for a variety of applications.

CRediT authorship contribution statement

Nitin Kumar Singh: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Manish Agarwal:** Formal analysis, Conceptualization. **Mithun Radhakrishna:** Writing – original draft, Validation, Supervision, Conceptualization.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this publication.

Data availability

All the data and the Python scripts, are available at the Github link: https://github.com/mr2972/helix_uniqueness.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgments

MR wants to thank the Science and Engineering Research Board (SERB), Govt of India, for the Core Research Grant CRG/2022/008633 and Mathematical Research Impact Centric Support (MATRICS) Grant MTR/2022/000664 for funding and the High-Performance Computing Facility at IIT Gandhinagar for support with computational resources. We also acknowledge PARAM ANANTA Supercomputer commissioned by National Supercomputing Mission (NSM) for providing computing resources of HPC System, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST), Government of India. NKS wants to thank the Ministry of Education, Govt of India, for the doctoral fellowship.

Appendix A. Supplementary data

1. Mean proportion (μ) of amino acids in different secondary structures in a set of 1000 Virus and Homo sapiens proteins.
2. Fraction of amino acid residues(f) in different secondary structures (α-helix, β-strand, turns, and coils) in the set of 1000 Virus and Homo sapiens proteins.
3. Probability distribution ($p(\lambda)$ as a function of charge density(λ) in different secondary structures (α-helix, β-strand, turns, and coils) in the set of 1000 Virus and Homo sapiens proteins.
4. Length distribution in different secondary structure elements for the set of 1000 Virus and Homo sapiens proteins.
5. LCR normalized length in different secondary structures for a set of 1000 Virus and Homo sapiens proteins. Here, X'-positive and X'-negative represent the positive LCR and negative LCR, respectively. X': H=α-helix, E=β-strand, T=Turn and C=Coil.
6. Distribution of oppositely charged patches across protein secondary structures for a set of 1000 Virus and Homo sapiens proteins.
7. HelixPred algorithm.
8. Calculation of contact map.
9. PDB IDs of all the proteins analyzed.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiolchem.2024.108237.

References

Anfinsen, C.B., 1962. Some observations on the basic principles of design in protein molecules. Comp. Biochem. Physiol. 4 (2–4), 229–240.
Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Science 181 (4096), 223–230.
Berger, B., Leighton, T., 1998. Protein folding in the hydrophobic-hydrophilic (HP) is NP-complete. In: Proceedings of the Second Annual International Conference on Computational Molecular Biology. pp. 30–39.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28 (1), 235–242.
Bonomi, M., Vendruscolo, M., 2019. Determination of protein structural ensembles using cryo-electron microscopy. Curr. Opin. Struct. Biol. 56, 37–45.
Dalal, S., Balasubramanian, S., Regan, L., 1997. Transmuting α helices and β sheets. Fold. Des. 2 (5), R71–R79.
David, L., Nelson, D.L., Cox, M.M., Stiedemann, L., McGlynn, Jr., M.E., Fay, M.R., 2000. Lehninger principles of biochemistry.
DePristo, M.A., de Bakker, P.I., Blundell, T.L., 2004. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure 12 (5), 831–838.
Dill, K.A., Bromberg, S., Yue, K., Chan, H.S., Ftebig, K.M., Yee, D.P., Thomas, P.D., 1995. Principles of protein folding—a perspective from simple exact models. Protein Sci. 4 (4), 561–602.
Glaser, F., Steinberg, D.M., Vakser, I.A., Ben-Tal, N., 2001. Residue frequencies and pairing preferences at protein–protein interfaces. Proteins: Struct. Funct. Bioinform. 43 (2), 89–102.
Heffernan, R., Paliwal, K., Lyons, J., Singh, J., Yang, Y., Zhou, Y., 2018. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. J. Comput. Chem. 39 (26), 2210–2216.
Heinig, M., Frishman, D., 2004. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res. 32 (suppl_2), W500–W502.
Ilari, A., Savino, C., 2008. Protein structure determination by x-ray crystallography. Bioinform. Data Seq. Anal. Evol. 63–87.
Jing, X., Wu, F., Luo, X., Xu, J., 2024. Single-sequence protein structure prediction by integrating protein language models. Proceedings of the National Academy of Sciences 121 (13), e2308788121.
Jonic, S., Vénien-Bryan, C., 2009. Protein structure determination by electron cryo-microscopy. Curr. Opin. Pharmacol. 9 (5), 636–642.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589.
Kainosho, M., Torizawa, T., Iwashita, Y., Terauchi, T., Mei Ono, A., Güntert, P., 2006. Optimal isotope labelling for NMR protein structure determinations. Nature 440 (7080), 52–57.
Kim, D.E., Chivian, D., Baker, D., 2004. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 32 (suppl_2), W526–W531.
Lindorff-Larsen, K., Kragelund, B.B., 2021. On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. J. Mol. Biol. 433 (20), 167196.
Loladze, V.V., Makhatadze, G.I., 2011. Energetics of charge–charge interactions between residues adjacent in sequence. Proteins: Struct. Funct. Bioinform. 79 (12), 3494–3499.
Lyu, P.C., Sherman, J.C., Chen, A., Kallenbach, N.R., 1991. Alpha-helix stabilization by natural and unnatural amino acids with alkyl side chains. Proc. Natl. Acad. Sci. 88 (12), 5317–5320.
Mandel-Gutfreund, Y., Gregoret, L.M., 2002. On the significance of alternating patterns of polar and non-polar residues in beta-strands. J. Mol. Biol. 323 (3), 453–461.
Marks, D.S., Hopf, T.A., Sander, C., 2012. Protein structure prediction from sequence variation. Nature Biotechnol. 30 (11), 1072–1080.
Moreira, I.S., Fernandes, P.A., Ramos, M.J., 2007. Hot spots—A review of the protein–protein interface determinant amino-acid residues. Proteins: Struct. Funct. Bioinform. 68 (4), 803–812.
Nelson, D.L., Lehninger, A.L., Cox, M.M., 2008. Lehninger Principles of Biochemistry. Macmillan.
Pace, C.N., 1992. Contribution of the hydrophobic effect to globular protein stability. J. Mol. Biol. 226 (1), 29–35.
Pace, C.N., Fu, H., Fryar, K.L., Landua, J., Trevino, S.R., Shirley, B.A., Hendricks, M.M., Iimura, S., Gajiwala, K., Scholtz, J.M., et al., 2011. Contribution of hydrophobic interactions to protein stability. J. Mol. Biol. 408 (3), 514–528.
Privalov, P.L., Gill, S.J., 1988. Stability of protein structure and hydrophobic interaction. Adv. Protein Chem. 39, 191–234.
Rohl, C.A., Strauss, C.E., Misura, K.M., Baker, D., 2004. Protein structure prediction using rosetta. In: Methods in Enzymology, vol. 383, Elsevier, pp. 66–93.
Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlić, A., Quesada, M., Quinn, G.B., Westbrook, J.D., et al., 2010. The RCSB protein data bank: Redesigned web site and web services. Nucleic Acids Res. 39 (suppl_1), D392–D401.

Sheinerman, F.B., Honig, B., 2002. On the role of electrostatic interactions in the design of protein–protein interfaces. J. Mol. Biol. 318 (1), 161–177.
Shoemaker, K.R., Kim, P.S., Brems, D.N., Marqusee, S., York, E.J., Chaiken, I.M., Stewart, J.M., Baldwin, R.L., 1985. Nature of the charged-group effect on the stability of the C-peptide helix. Proc. Natl. Acad. Sci. 82 (8), 2349–2353.
Shortle, D., Stites, W.E., Meeker, A.K., 1990. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. Biochemistry 29 (35), 8033–8041.
Singh, N.K., Agarwal, M., Radhakrishna, M., 2023. Understanding the helical stability of charged peptides. Proteins: Struct. Funct. Bioinform. 91 (2), 268–276.
Strickler, S.S., Gribenko, A.V., Gribenko, A.V., Keiffer, T.R., Tomlinson, J., Reihle, T., Loladze, V.V., Makhatadze, G.I., 2006. Protein stability and surface electrostatics: a charged relationship. Biochemistry 45 (9), 2761–2766.
Tripathi, S., Garcìa, A.E., Makhatadze, G.I., 2015. Alterations of nonconserved residues affect protein stability and folding dynamics through charge–charge interactions. J. Phys. Chem. B 119 (41), 13103–13112.
Watson, J.D., Laskowski, R.A., Thornton, J.M., 2005. Predicting protein function from sequence and structural data. Curr. Opin. Struct. Biol. 15 (3), 275–284.
West, M.W., Hecht, M.H., 1995. Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. Protein Sci. 4 (10), 2032–2039.
Wuethrich, K., 1989. The development of nuclear magnetic resonance spectroscopy as a technique for protein structure determination. Acc. Chem. Res. 22 (1), 36–44.
Zhang, B., Li, J., Lü, Q., 2018. Prediction of 8-state protein secondary structures by a novel deep learning architecture. BMC Bioinform. 19, 1–13.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

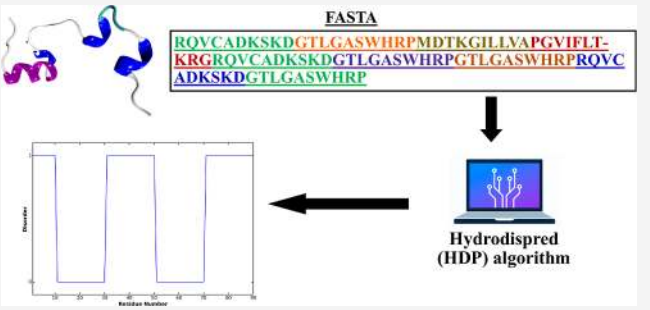
Hydrophobicity—A Single Parameter for the Accurate Prediction of Disordered Regions in Proteins

Nitin Kumar Singh, Pratyasha Bhardwaj, and Mithun Radhakrishna*

Cite This: J. Chem. Inf. Model. 2023, 63, 5375−5383 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The prediction of disordered regions in proteins is crucial for understanding their functions, dynamics, and interactions. Intrinsically disordered proteins (IDPs) play a key role in many biological processes like cell signaling, recognition, and regulation, but experimentally determining these regions can be challenging due to their high mobility. To address this challenge, we present an algorithm called HydroDisPred (HDP). HDP uses a single parameter, the fraction of hydrophobicity (λ) in each segment of the protein, to accurately predict disordered regions. The algorithm was validated using experimental data from the DisProt database and was found to be on par and, in some cases, more effective than the existing algorithms. HDP is a simple and effective method for identifying disordered regions in proteins, and its prediction is not affected by the availability of training data, unlike other ML approaches. The application is housed in the web server and can be accessed through the URL https://proseqanalyser.iitgn.ac.in/hydrodispred/.



INTRODUCTION

Until late 1990s, it was believed that a well-defined three-dimensional structure was essential for protein function.[1] As a result, most experimental and computational studies focused on understanding protein folding and stability.[2−5] However, by late 1990s, it was clear that many proteins that were disordered and lacked a well-defined three-dimensional structure played a key role in various biological functions.[6−8] In fact, proteins with the disorder were found to be involved in a wide range of functions, including transcriptional regulation, protein−protein interactions, and DNA binding.[9−14] These proteins are known as intrinsically disordered proteins and are known to be involved in cancer-associated proteins such as p53, p57, and thyroid-associated protein TC-1, as well as neurodegenerative diseases such as Alzheimer's and Parkinson's, which are caused by the aggregation of amyloid beta (Aβ) and α-synuclein.[15−19]

These proteins can either appear as coils that completely lack a well-defined secondary structure, known as intrinsically disordered proteins (IDPs), or as regions with enhanced spatial mobility within a protein with a defined secondary structure, known as intrinsically disordered regions (IDRs).[20] Given their importance in biological functioning, it is essential to characterize these regions to understand their functional mechanisms and therapeutic applications. However, due to their high spatial and temporal mobility, it is difficult to characterize IDPs and IDRs using traditional methods such as X-ray crystallography and cryo-electron tomography. Other techniques, such as atomic force microscopy (AFM), FRET, and fluorescence correlation spectroscopy, have provided some useful information on the morphology and dynamics of these proteins.[21−23] Integrative modeling techniques that combine data from various experimental sources have also been widely used to predict the structure of these regions. However, effective sampling of IDP structures is challenging due to the many degrees of freedom, making it difficult to predict the structure of disordered regions.

To overcome these difficulties, theoretical approaches, also known as bottom-up approaches, have been applied to identify disordered regions. These approaches are particularly useful for the study of the large proteome and provide insights into the thermodynamic parameters of IDP binding to their partners and the accompanying conformational changes. The design of efficient algorithms based on machine learning, deep learning, and artificial intelligence, including Alpha-Fold and Rosetta, has been successful in predicting the 3D structure of globular proteins.[24−26] However, despite recent advances in structure prediction algorithms, there are still challenges in predicting the disordered regions of proteins, primarily due to the lack of a comprehensive and reliable data set on intrinsically disordered proteins (IDPs).

Received: April 20, 2023
Published: August 15, 2023

https://doi.org/10.1021/acs.jcim.3c00592
J. Chem. Inf. Model. 2023, 63, 5375−5383

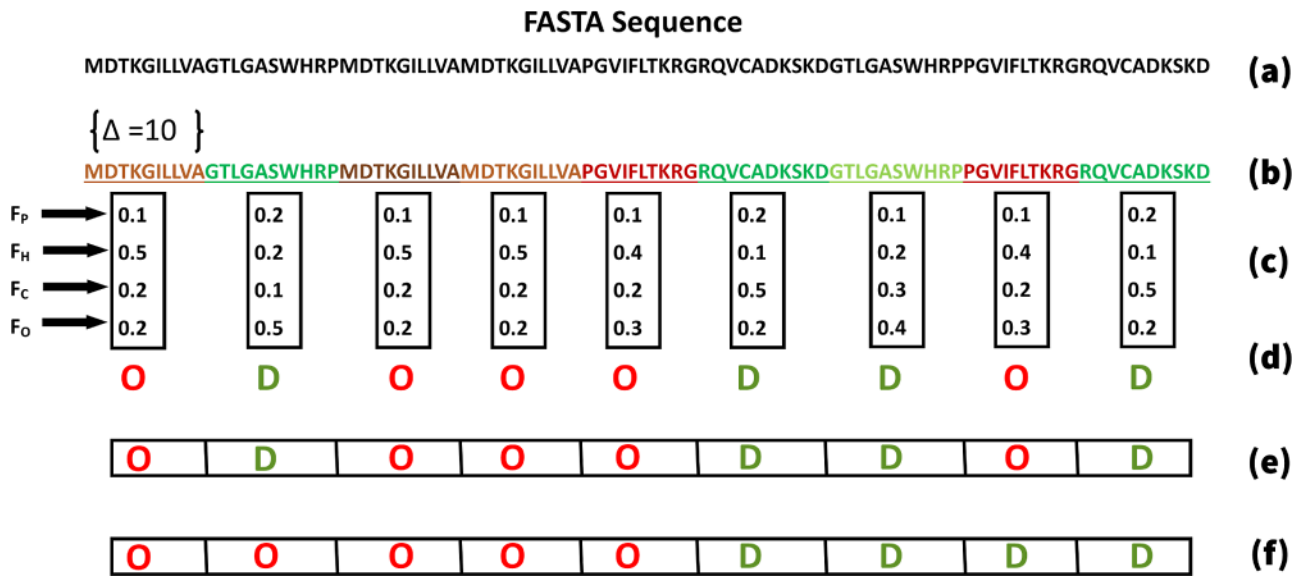


Figure 1. Flowchart depicting the algorithm to identify the disordered regions in the protein sequences. Here, F_H, F_P, F_C, and F_O represent the fraction of hydrophobic, fraction of polar, fraction of charged, and fraction of other residues, respectively.

Many approaches, ranging from simple statistical analysis of the sequence to complex modeling using artificial neural networks, have been employed for prediction. Previous studies of IDPs have shown that the disordered regions are rich in polar and charged amino acid residues and deficient in hydrophobic residues.[8] One of the earliest approaches dates back to the work of Uversky et al., who used the information on the mean hydrophobicity and mean net charge of the peptide to construct a phase diagram that differentiated ordered and disordered regions. The disordered regions were characterized by low mean hydrophobicity and high mean net charge. However, this phase diagram does not accurately predict disordered regions within IDPs. Pappu et al. used the fraction of positive and negative charge to construct phase diagrams that divide the polypeptide into regions of weak and strong polyelectrolyte/polyampholyte.[27] However, it should be noted that these methods do not provide information on the disordered regions within the protein, which is important as natively unfolded proteins may contain up to 80−90% ordered regions, as seen in Sulfotransferase 1A3 (20.3% disorder) and Estradiol 17-β-dehydrogenase 1 (13.1% disorder), among others. The asymmetry in amino acid compositions in terms of hydropathy has been previously applied to predict the transmembrane domains (TMDs) of integral membrane proteins across various organisms and organelles. However, it has to be noted that TMDs are structural proteins with a high helical content and a very high proportion of hydrophobic amino acids, which makes them very distinct from IDRs.[28−31]

With recent advances in algorithm development, web applications have been developed to accurately identify disordered regions. These applications use a combination of machine learning, artificial intelligence, and biophysical models to predict disordered regions. Some examples include PONDER, which uses an artificial neural network, PODDLE,[32] which uses a support vector machine, and IUPred,[33] a biophysical-based predictor based on the interaction energy between amino acids in the polypeptide chain. Most of these methods provide a score between 0 and 1 for each residue, with regions of disorder characterized by a score greater than 0.5. While these algorithms can predict disordered regions satisfactorily for many IDPs, there are cases where the predictions do not match experimental findings and where there is no agreement between the different web applications for certain IDPs.

The main objective of this manuscript is to make the prediction of disordered regions within proteins simple by using a single parameter, i.e., fraction hydrophobicity (λ). Our study shows the HydroDisPred (HDP) algorithm can satisfactorily predict disordered regions in line with the existing web applications. After validating the algorithm for a set of IDPs, we extended the study to predict ordered and disordered regions in nucleoporin proteins, also known as FG NUPs, that are responsible for the bidirectional transport of materials across the nuclear envelope.

Model and Methods. The predictor is designed to perform the analysis of the IDP regions in the proteins based on the fraction of hydrophobic residues in the regions. The algorithm involves characterizing the amino acid residues into four classes, namely hydrophobic, polar, charged, and others. This classification of amino acids is inspired by the work of Ghavami et al.[34] and Rauff et al.[35]

The detailed workflow of the HydroDisPred (HDP) algorithm is explained below.

1. The algorithm takes the FASTA sequence of the protein as the input (Figure 1a).
2. The input sequence is divided into n equal segments, each containing Δ number of residues. It has to be noted that the last segment may or may not contain exactly Δ residues (Figure 1b).
3. The fraction of hydrophobic, polar, charged, and other classes of amino acid residues contained in each segment is calculated based on the classification provided in Table 1 (Figure 1c).
4. Regions that contain the fraction of hydrophobic residues less/more than a critical value λ_C are marked D(disorder)/O(order) (Figure 1d).
5. If ONLY ONE segment of D (disorder) is sandwiched between two ordered (O) regions, then D is treated as O. Similarly, if an ordered (O) segment is sandwiched

https://doi.org/10.1021/acs.jcim.3c00592
J. Chem. Inf. Model. 2023, 63, 5375−5383

Table 1. Classification of Amino Acids

amino acid category	hydrophobic	polar	charged	others
amino acid name	Val, Ile, Leu, Met, Phe, Tyr, and Trp	Ser, Thr, Asn, and Gln	Lys, Arg, Asp, and Glu	Cys, Ala, His, Gly, and Pro

between two D regions, it is also treated as D (Figure 1e).

6. Intrinsically disordered regions are characterized by segments marked as **D** as disordered, and the ordered regions are denoted by **O** (Figure 1f)
7. The selection of Δ (segment length) and λ_C (critical hydrophobic fraction) is extremely critical for the accurate estimation of the disordered regions. After analyzing a vast data set, we have chosen Δ = 10 and λ_C = 0.25.
8. While a smaller value of Δ leads to finite size effects, a larger value of Δ averages out the fluctuations, leading to inaccurate estimation. The algorithm is able to predict the regions within an error of Δ residues.

Figure 1 pictorially demonstrates the workflow of the algorithm.

RESULTS AND DISCUSSION

The following section uses the HDP algorithm described above to predict the intrinsically disordered regions in proteins. Subsequently, this is compared with the experimental prediction given in the DisProt database and the other existing algorithms. Table 2 shows the application of the algorithm for predicting the disordered regions in the viral macrophage inflammatory protein 2 (DISPROT ID: DP00685). DP00685 is a 94-residue-long protein. The protein is divided into equal segments containing 10 residues each (except for the last segment, which only contains four residues). The percentage of various amino acid residues in each group based on classification as given earlier is shown in Table 3. The algorithm identifies two segments where the fraction of hydrophobic residues in that particular segment is less than the critical value λ_C = 0.25, with residues spanning 21−30 and 31−40. Based on these criteria, the algorithm identifies regions 21−40 as the disordered region. Figure 2 compares our prediction with the state-of-the-art algorithms IUPRED2-long and PONDR-VLXT. According to these algorithms, residues with a score greater than 0.5 are classified as disordered regions. While IUPRED2-long does not predict disorderedness in the protein, PONDR-VLXT predicts regions 1−2, 48−52, 11−82, and 92−94 as the disordered regions. Our prediction (21−40) segment is in close agreement with the experimentally predicted regions.[21−34] Although we have compared our prediction with only two state-of-the-art algorithms here, in Figures S1 and S2, we have compared the prediction with a few more algorithms like PONDER-FIT, IUPRED2-SHORT, PONDER-XL1, PONDER-VL3, and ESPRITZ-DISPROT to be thorough.

Table 2. Amino Acid Composition of DP00685 across Various Segments of Size 10: A Breakdown of Hydrophobic, Polar, Charged, and Other Residues along the Sequence

start position	end position	hydrophobic	polar	charged	others	G-P
1	10	0.50	0.10	0.20	0.10	0.10
11	20	0.50	0.30	0.00	0.20	0.00
21	30	0.20	0.20	0.20	0.20	0.20
31	40	0.20	0.10	0.30	0.20	0.20
41	50	0.40	0.30	0.10	0.00	0.20
51	60	0.30	0.40	0.10	0.10	0.10
61	70	0.40	0.10	0.20	0.00	0.30
71	80	0.10	0.20	0.50	0.20	0.00
81	90	0.50	0.20	0.20	0.00	0.10
91	94	0.25	0.25	0.25	0.25	0.00

Figure 3 shows the prediction based on a similar analysis on the protein—membrane fusion protein p14 (DisProt ID: DP01043). DP01043 is the 125 residue protein. Our algorithm predicts the region (1−30) as the disordered region, which is in close agreement with the experimental prediction of (2−31). Both the algorithms IUPRED2-Long and PONDR-VLXT could not predict the disordered region in this protein. Table 3 presents the prediction of disordered regions for an additional 11 intrinsically disordered proteins from the DISPROT database. The results show that our algorithm (HydrDisPred) closely aligns with the experimentally reported values. Compared to other existing algorithms, HDP performs well and, in some cases, even outperforms them. However, it should be noted that there are a few instances, such as for protein DP00148, where none of the algorithms could accurately predict the disordered regions. The predictions from different predictors and HDP are provided in Figure S3.

Disorder Prediction in FG-Nups. In this section, we use the HDP algorithm to predict the disordered regions within a specific class of IDPs known as FG-NUPs. FG-NUPs, or FG-repeats containing nucleoporins, are a subset of nucleoporins that form the nuclear pore complex (NPC) in eukaryotic cells.[36,37] These proteins are characterized by a high proportion of phenylalanine (F) and glycine (G) repeat units in their amino acid sequence, which form a β-propeller structure that interacts with other FG-NUPs and other proteins. FG-NUPs play a crucial role in regulating bidirectional nuclear transport, acting as molecular sieves that perforate the nuclear membrane.

FG-NUPs are a specific class of nucleoporins characterized by three distinct regions: a coil region, an extended region, and a structural region. The coil and extended regions form the disordered regions, while the structural region helps anchor the protein to the inner lining of the nuclear membrane. The coil region is composed of mostly hydrophobic amino acids, including FG repeats, and has a low proportion of charged residues. In contrast, the extended region is rich in charged residues. The transition from the coil to the extended region can be easily identified by calculating the cumulative number of charged residues along the primary sequence, as demonstrated in Figure 4 for NUP 145N and NUP 100. It is clear from the graph that there is a drastic change in the slope in the regions 255−260 for NUP 145N and 608−610 for NUP 100, which helps to differentiate between the coil and extended regions. The HDP algorithm can now be applied to predict where the extended region transitions into the structural region, which is important in understanding the function and interaction of FG-NUPs.

Figures 5 and 6 illustrate the prediction of the disorder-to-order transition in yeast NUPs, NUP100N and NUP145N, respectively, as forecasted by the HDP algorithm. NUPs are distinct from typical IDPs as they have a disordered region

https://doi.org/10.1021/acs.jcim.3c00592
J. Chem. Inf. Model. 2023, 63, 5375−5383

Table 3. Comparative Analysis of Different Proteins Available in the DISPROT Database with Two Existing Web Applications (PONDER-VLXT and IUPred long) with Our Predictor Algorithm[a]

S. no	disprot ID	total number of amino acids	ID region	our prediction	ponder-VLXT	IUPred long
1	DP00842	86	1−86	1−30, 51−86	1−10, 44−86	1−14, 48−86
2	DP01928	289	1−289	31−60, 81−170, 191−240, 261−289	15−70, 77−146, 194−250, 264−289	30−67, 81−113, 123−146, 157−160, 186−214, 220−239, 279−289
3	DP00685	94	21−34	21−40	1−2, 48−52, 71−82, 92−94	
4	DP01043	125	2−31	1−30	1−6, 58−79, 86−116	3−6, 93−114
5	DP00847	112	74−112	51−112	1−2, 4−5, 31−33, 48−57	73−89
6	DP00005	107	1−107	1−20, 41−107	1−56, 64−103	1−17, 36−46, 76−96
7	DP01336	132	42−80	1−20, 41−70, 131−132	1−9, 16−78, 129−131	34−74, 126−132
8	DP02208	88	56−88	31−50, 71−88	1−8, 22−44, 47−88	16−59, 62−88
9	DP01012	117	57−93	41−90	1−3, 34−93, 106−110	66−71
10	DP01481	370	1−94	1−90	0−40, 50−90, 317−344	1−81
11	DP00986	69	2−29	1−20	0−18, 52−54, 57−69	1−22, 44−58
12	DP01983	203	151−203	151−203	61−79, 152−180, 190−203	157−203
13	DP00871	121	1−121		1−6, 15−34, 57−58, 61−88, 101−05	

[a]ID regions represent experimentally determined disorder regions as reported in the DisProt database.

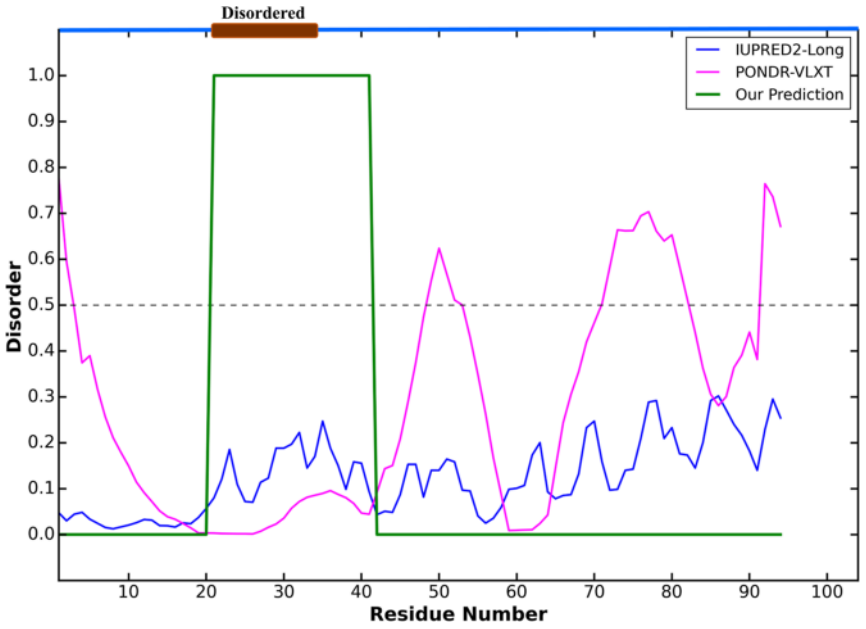


Figure 2. DP00685 disorder analysis. The green line is the prediction from our algorithm, and the magenta and blue curves are the PONDER-VLXT and IUPRED2-long predictions, respectively. The brown patch on the top axis represents the actual disordered region obtained from the DisProt database.

composed of a coil and an extended region connected to the nuclear membrane through a structural region. The transition from disorder to order in NUPs is determined by the last contiguous region of order, while fluctuations in between are ignored. Figures 5 and 6 reveal that the **last contiguous structural region** occurs near residues 790 and 460 in NUP100N and NUP145N, respectively. These values are in close agreement with the experimentally determined values of 800 and 450, respectively. Additionally, the algorithm has been validated for other NUPs, and the data can be found in Figures S4 and S5 in the Supporting Information.

Webserver. We have developed a web server hosted at https://proseqanalyser.iitgn.ac.in/hydrodispred, based on the HDP algorithm, where the user inputs the FASTA sequence in single-line format and generates a plot that illustrates the ordered and disordered regions. The value 1 on the chart represents the disordered regions, and the value 0 represents the ordered regions.

DISCUSSION

Tuning the Optimal Performance of the Algorithm. We conducted a comprehensive study to analyze the impact of varying the bin size (Δ) and the hydrophobicity threshold (λ_C). Three evaluation matrices, namely, accuracy (A), positive predictive value (PPV), also known as precision (P), and negative predictive value (NPV), which are fundamental in quantifying the performance of the predictions, were evaluated. This investigation allowed us to gain insights into the optimal parameter values that yield the best performance for our algorithm (Figure 7).

https://doi.org/10.1021/acs.jcim.3c00592
J. Chem. Inf. Model. 2023, 63, 5375−5383

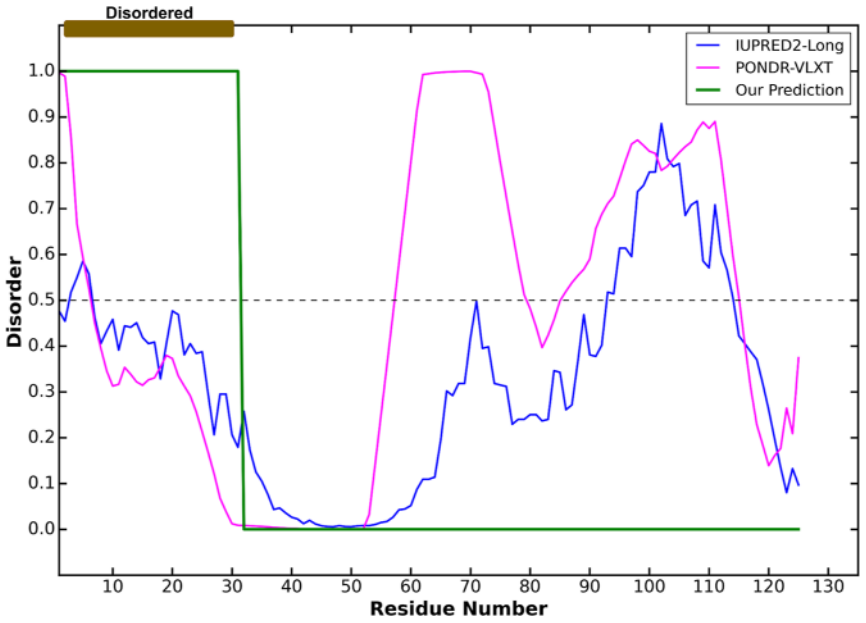


Figure 3. DP01043 disorder prediction. The green line is the prediction from our algorithm, and the magenta and blue curves are the PONDER-VLXT and IUPRED2-long predictions, respectively. The brown patch on the top axis represents the actual disordered region obtained from the DisProt database.

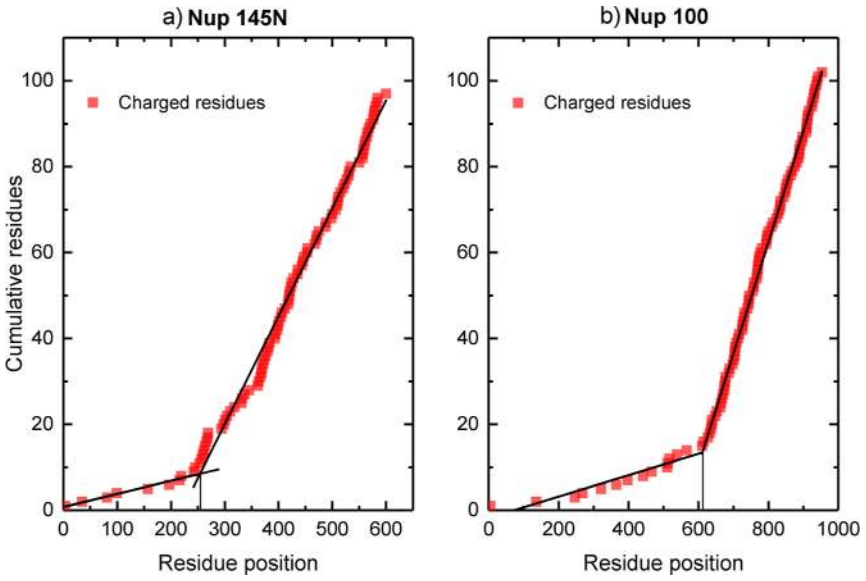


Figure 4. Cumulative number of charged residues along the amino acid sequence as we move away from the center toward the nuclear membrane. A coil to extended transition can be easily identified by a drastic change in the slope.

To clarify, accuracy (A) is defined as $A = \frac{TP + TN}{TP + TN + FP + FN}$, precision (P) is calculated as: $P = \frac{TP}{TP + FP}$, and NPV is calculated as NPV $= \frac{TN}{TN + FN}$, where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. It is worth noting that reliable algorithms should have high values of accuracy, P and NPV.

The highest values of precision and NPV are observed for $\Delta = 10$ and $\lambda_C = 0.25$, as shown in Figure 7. It is evident from Figure 7 that as λ_C increases, accuracy and precision also increase until reaching a certain λ_C value, after which they decline (7a,b). The decrease in performance at higher λ_C values is attributed to an increase in false positive values, and the value of NPV goes to zero at a larger λ_C because the true negatives approach zero at the maximum cutoff. Additionally, the decrease in the values of precision and NPV with increasing Δ is caused by a reduction in the number of true positives and true negatives, respectively.

As described in Figure 1, the minimum length of the IDP region that the algorithm can predict is 2Δ, twice the bin width, unless the IDPs are located at the ends. In such cases, the algorithm can predict regions as small as Δ.

To validate our algorithm, we collected data sets of over 124 IDPs from the DisProt database containing known IDP regions.[38] We calculated the distribution of the length of these IDP regions, as depicted in Figure S8. The graph clearly shows a high proportion of IDRs with a length less than or equal to 20. This observation influenced our decision to choose a bin

https://doi.org/10.1021/acs.jcim.3c00592
J. Chem. Inf. Model. 2023, 63, 5375−5383

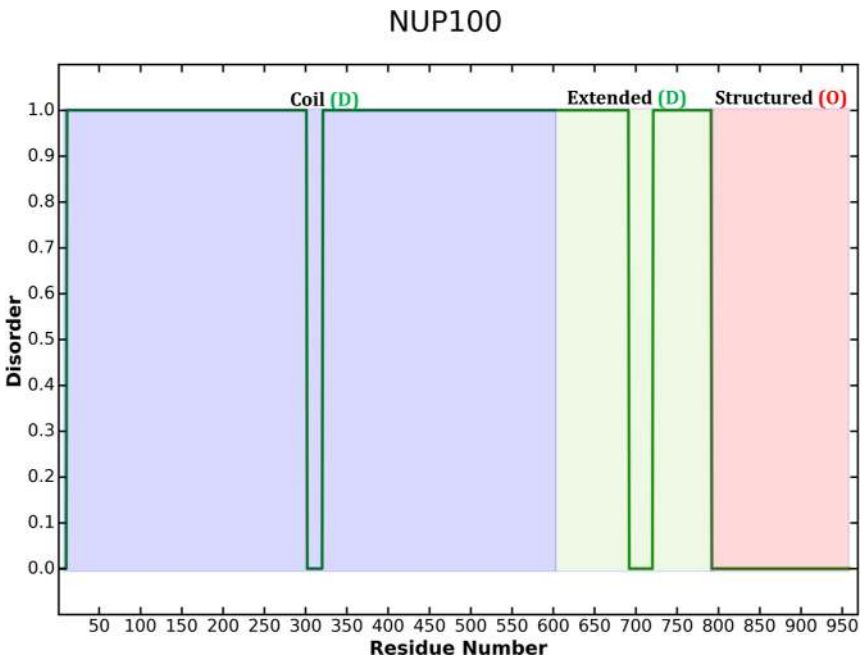


Figure 5. Disorder prediction in NUP100 using the HDP algorithm. HDP analysis predicts that the disorder-to-order transition occurs near residue number 800, which is in close agreement with the experimental findings.

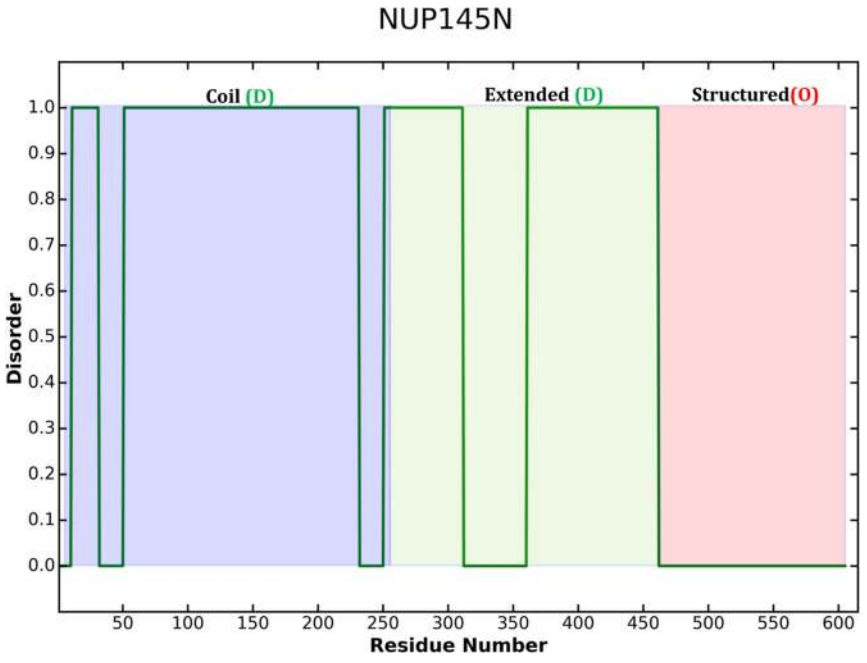


Figure 6. Disorder prediction in NUP145N using the HDP algorithm. HDP analysis predicts that the disorder-to-order transition occurs near residue number 450, which is in close agreement with the experimental findings.

size of 10, as it effectively captured these variations. Additionally, we computed the composition of hydrophobic groups for the IDR regions obtained from the same set of 124 proteins. Figure S9 demonstrates that the distribution peaked around 0.25. Consequently, based on this finding, we selected a bin width of $\Delta = 10$ and a hydrophobicity cutoff of $\lambda_C = 0.25$.

Comparison with Existing Algorithms. We conducted a comparative analysis of the HDP algorithm with existing disorder predictors (Figure 8). These predictors include PONDR-FIT, PONDR-VLXT, PONDR-VL3, PONDR-VSL2, IUPRED-LONG, and IUPRED-SHORT. Despite its simplicity, the HDP algorithm performed on par with the existing machine learning and biophysical model-based predictors.

CONCLUSIONS

In this study, we have presented HydrDisPred (HDP), an algorithm that accurately predicts intrinsically disordered regions in proteins using a single parameter: hydrophobicity. Our approach is based on a statistical analysis of the protein's amino acid sequence, where we divide it into smaller parts and identify regions with an average hydrophobicity below a critical

https://doi.org/10.1021/acs.jcim.3c00592
J. Chem. Inf. Model. 2023, 63, 5375−5383

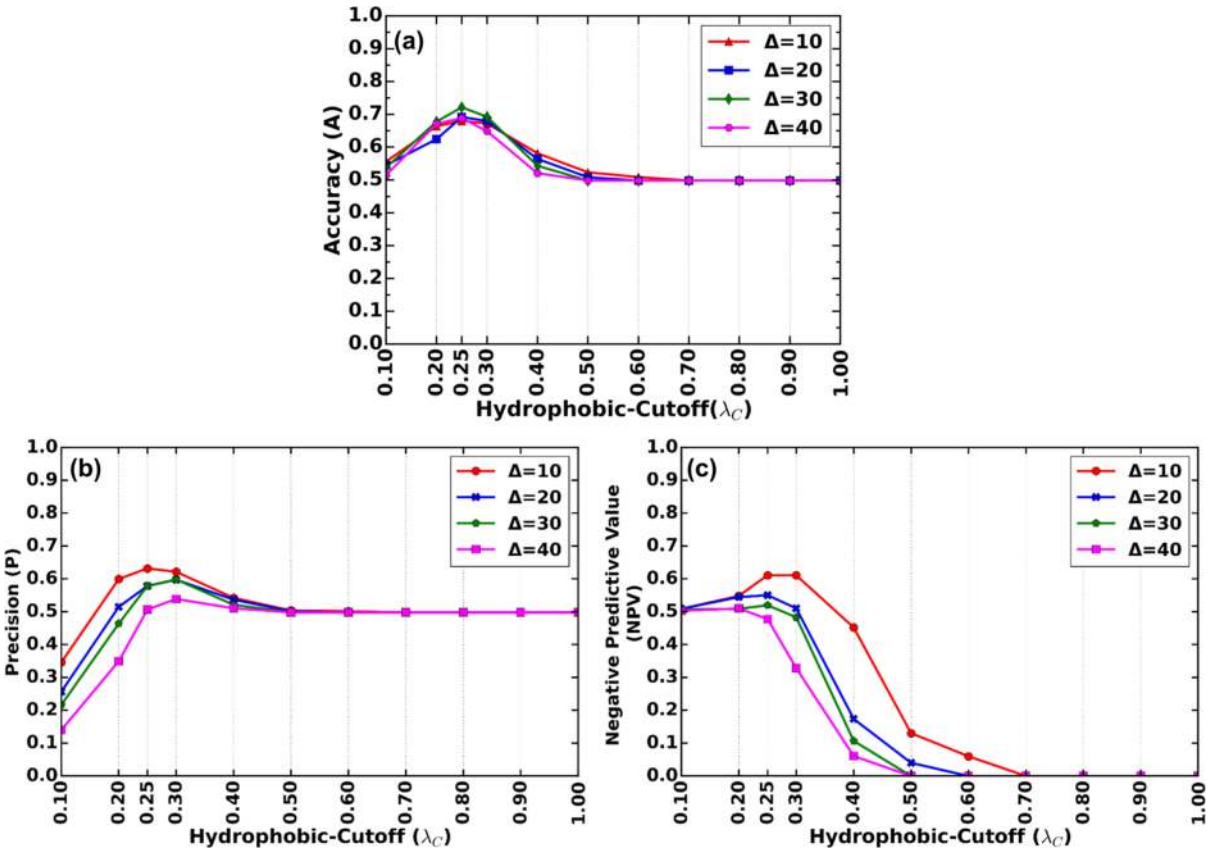


Figure 7. (a) Accuracy (A), (b) precision (P), and (c) negative predictive value (NPV) for the different bin-widths (Δ) and different hydrophobic cutoffs (λ_C) for a set of IDPs studied previously by Almog et al.[38]

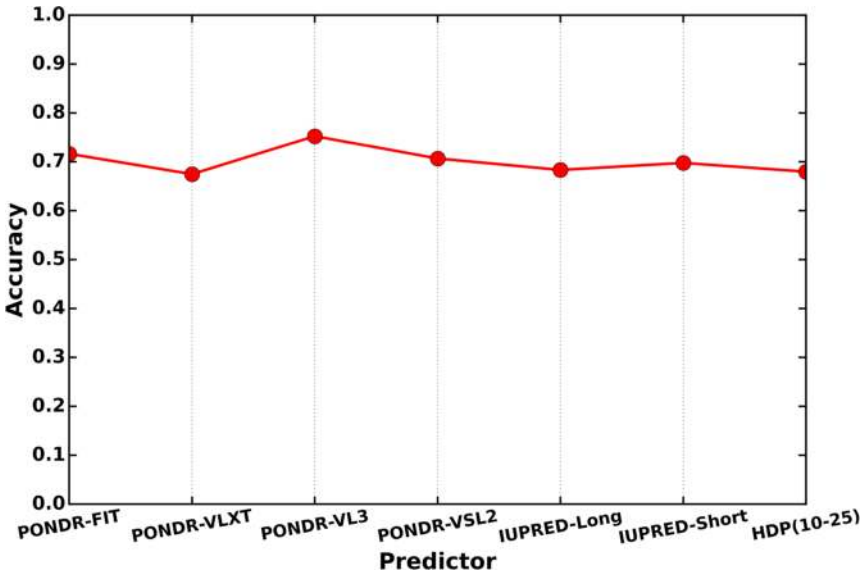


Figure 8. Comparison of the accuracy of different existing predictors with the HDP algorithm. The evaluation was conducted on the protein data set obtained from the DISPROT database as studied by Almog et al.[38]

value of λ_C = 0.25 as intrinsically disordered. We verified our algorithm using a set of known IDPs and FG NUPs and found that our predictions are in good agreement with both computational and experimental studies. One of the key features of HDP is its simplicity and user-friendliness. Our algorithm is based on a simple statistical analysis and is independent of the data set, making it easy to use for researchers in the field of protein disorder prediction. We believe our approach can serve as a useful tool for studying the role of intrinsic disorders in protein function and protein−protein interactions. The algorithm is housed in server https://proseqanalyser.iitgn.ac.in/hydrodispred/ and can be accessed by all researchers and the public at large.

https://doi.org/10.1021/acs.jcim.3c00592
J. Chem. Inf. Model. 2023, 63, 5375−5383

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00592.

DP00685 disorder prediction: the green line is the prediction from our algorithm and various other existing predictors; DP01043 disorder prediction: the green line is the prediction from our algorithm and various other existing predictors; DP00148 disorder prediction: the left graph shows the disorder prediction from the existing software whereas the right graph is from our algorithm, and DP00148 has disordered residues from residue number 378 to 389 and from residue number 428 to 449; disorder prediction in NUP1: previous findings show that the region from 1 to 220 is structured, whereas the other regions are unstructured, and the HDP algorithm shows the results in closer agreement with the previous findings; disorder prediction in NUP2: previous findings show that nucleoporin Nup2p is a natively unfolded protein, and the HDP algorithm also shows that most of the regions are in a disordered state; homepage of HDP website; result page of the HDP website; analysis of the length distribution of intrinsically disordered regions in a set of proteins obtained from the Disprot database; distribution of hydrophobic groups within the intrinsically disordered regions in a protein set obtained from the Disprot database; and plots for comparing the HDP algorithm with existing algorithms for the prediction of IDR regions for proteins listed in Table 3 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Mithun Radhakrishna − *Discipline of Chemical Engineering and Center for Biomedical Engineering, Indian Institute of Technology (IIT) Gandhinagar, Palaj, Gujarat 382355, India*; orcid.org/0000-0001-7127-4744; Email: mithunr@iitgn.ac.in

Authors

Nitin Kumar Singh − *Discipline of Chemical Engineering, Indian Institute of Technology (IIT) Gandhinagar, Palaj, Gujarat 382355, India*

Pratyasha Bhardwaj − *Discipline of Chemical Engineering, Indian Institute of Technology (IIT) Gandhinagar, Palaj, Gujarat 382355, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00592

Notes

The authors declare no competing financial interest.

The HDP algorithm can be tested at https://proseqanalyser.iitgn.ac.in/hydrodispred/. All the data, including the FASTA sequence and the corresponding Python script, are available at the Github link https://github.com/mr2972/IDP_*algorithm*.

■ ACKNOWLEDGMENTS

MR wants to thank the Science and Engineering Research Board (SERB), Govt of India, for the Core Research grant CRG/2022/008633, Mathematical Research Impact Centric Support (MATRICS) grant MTR/2022/000664 for funding, and the High-Performance Computing Facility at IIT Gandhinagar for support with computational resources. We also acknowledge PARAM ANANTA Supercomputer commissioned by National Supercomputing Mission (NSM) for providing computing resources of the HPC System, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and the Department of Science and Technology (DST), Government of India. NKS wants to thank the Ministry of Education, Govt of India, for the doctoral fellowship.

■ REFERENCES

(1) Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.* **1894**, *27*, 2985−2993.

(2) Creighton, T. E. Protein folding. *Biochem. J.* **1990**, *270*, 1−16.

(3) Dobson, C. M. Protein folding and misfolding. *Nature* **2003**, *426*, 884−890.

(4) Levinthal, C. Are there pathways for protein folding? *J. Chim. Phys. Phys.-Chim. Biol.* **1968**, *65*, 44−45.

(5) Karplus, M.; Weaver, D. L. Protein-folding dynamics. *Nature* **1976**, *260*, 404−406.

(6) Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; et al. Intrinsically disordered protein. *J. Mol. Graph. Model.* **2001**, *19*, 26−59.

(7) Dunker, A. K.; Brown, C. J.; Lawson, C. J. D.; Iakoucheva-Sebat, L. M.; Vucetic, S.; Obradovic, Z. The protein trinity: structure/function relationships that include intrinsic disorder. *Sci. World J.* **2002**, *2*, 49−50.

(8) Uversky, V. N.; Gillespie, J. R.; Fink, A. L. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Struct., Funct., Bioinf.* **2000**, *41*, 415−427.

(9) Basu, S.; Bahadur, R. P. A structural perspective of RNA recognition by intrinsically disordered proteins. *Cell. Mol. Life Sci.* **2016**, *73*, 4075−4084.

(10) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradović, Z. Intrinsic disorder and protein function. *Biochemistry* **2002**, *41*, 6573−6582.

(11) Uversky, V. N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **2002**, *11*, 739−756.

(12) Wright, P. E.; Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **1999**, *293*, 321−331.

(13) Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **2002**, *27*, 527−533.

(14) Namba, K. Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells* **2001**, *6*, 1−12.

(15) Dawson, R.; Müller, L.; Dehner, A.; Klein, C.; Kessler, H.; Buchner, J. The N-terminal domain of p53 is natively unfolded. *J. Mol. Biol.* **2003**, *332*, 1131−1141.

(16) Lee, H.; Mok, K. H.; Muhandiram, R.; Park, K.-H.; Suk, J.-E.; Kim, D.-H.; Chang, J.; Sung, Y. C.; Choi, K. Y.; Han, K.-H. Local structural elements in the mostly unstructured transcriptional activation domain of human p53. *J. Biol. Chem.* **2000**, *275*, 29426−29432.

(17) Uversky, V. N. A protein-chameleon: conformational plasticity of α-synuclein, a disordered protein involved in neurodegenerative disorders. *J. Biomol. Struct. Dyn.* **2003**, *21*, 211−234.

(18) Von Bergen, M.; Barghorn, S.; Jeganathan, S.; Mandelkow, E.-M.; Mandelkow, E. Spectroscopic approaches to the conformation of tau protein in solution and in paired helical filaments. *Neurodegener. Dis.* **2006**, *3*, 197−206.

(19) Iakoucheva, L. M.; Brown, C. J.; Lawson, J. D.; Obradović, Z.; Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **2002**, *323*, 573−584.

(20) Dunker, A. K.; Babu, M. M.; Barbar, E.; Blackledge, M.; Bondos, S. E.; Dosztányi, Z.; Dyson, H. J.; Forman-Kay, J.; Fuxreiter, M.; Gsponer, J.; et al. What's in a name? Why these proteins are

intrinsically disordered: Why these proteins are intrinsically disordered. *Intrinsically Disord. Proteins* **2013**, *1*, No. e24157.
(21) Dyson, H. J.; Wright, P. E. Unfolded proteins and protein folding studied by NMR. *Chem. Rev.* **2004**, *104*, 3607−3622.
(22) McCann, J. J.; Zheng, L.; Rohrbeck, D.; Felekyan, S.; Kühnemuth, R.; Sutton, R. B.; Seidel, C. A.; Bowen, M. E. Supertertiary structure of the synaptic MAGuK scaffold proteins is conserved. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 15775−15780.
(23) Kodera, N.; Noshiro, D.; Dora, S. K.; Mori, T.; Habchi, J.; Blocquel, D.; Gruet, A.; Dosnon, M.; Salladini, E.; Bignon, C.; et al. Structural and dynamics analysis of intrinsically disordered proteins by high-speed atomic force microscopy. *Nat. Nanotechnol.* **2021**, *16*, 181−189.
(24) Lindorff-Larsen, K.; Kragelund, B. B. On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. *J. Mol. Biol.* **2021**, *433*, 167196.
(25) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(26) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. *Methods in Enzymology*; Elsevier, 2004; Vol. *383*, pp 66−93.
(27) Holehouse, A. S.; Das, R. K.; Ahad, J. N.; Richardson, M. O.; Pappu, R. V. CIDER: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophys. J.* **2017**, *112*, 16−21.
(28) Rose, G. D. Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature* **1978**, *272*, 586−590.
(29) Sharpe, H. J.; Stevens, T. J.; Munro, S. A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell* **2010**, *142*, 158−169.
(30) Singh, S.; Mittal, A. Transmembrane domain lengths serve as signatures of organismal complexity and viral transport mechanisms. *Sci. Rep.* **2016**, *6*, 22352.
(31) Mittal, A.; Changani, A. M.; Taparia, S. Unique and exclusive peptide signatures directly identify intrinsically disordered proteins from sequences without structural information. *J. Biomol. Struct. Dyn.* **2021**, *39*, 2885−2893.
(32) Shimizu, K. *Protein Structure Prediction*; Springer, 2014; pp 131−145.
(33) Dosztányi, Z.; Csizmok, V.; Tompa, P.; Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **2005**, *21*, 3433−3434.
(34) Ghavami, A.; Veenhoff, L. M.; van der Giessen, E.; Onck, P. R. Probing the disordered domain of the nuclear pore complex through coarse-grained molecular dynamics simulations. *Biophys. J.* **2014**, *107*, 1393−1402.
(35) Ruff, K. M.; Roberts, S.; Chilkoti, A.; Pappu, R. V. Advances in understanding stimulus-responsive phase behavior of intrinsically disordered protein polymers. *J. Mol. Biol.* **2018**, *430*, 4619−4635.
(36) Hoogenboom, B. W.; Hough, L. E.; Lemke, E. A.; Lim, R. Y.; Onck, P. R.; Zilman, A. Physics of the nuclear pore complex: Theory, modeling and experiment. *Phys. Rep.* **2021**, *921*, 1−53.
(37) Denning, D. P.; Patel, S. S.; Uversky, V.; Fink, A. L.; Rexach, M. Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2450−2455.
(38) Almog, G.; Olabode, A. S.; Poon, A. F. Tuning intrinsic disorder predictors for virus proteins. *Virus Evol.* **2021**, *7*, veaa106.

https://doi.org/10.1021/acs.jcim.3c00592
J. Chem. Inf. Model. 2023, 63, 5375−5383

ACS APPLIED BIO MATERIALS

www.acsabm.org

Forum Article

Tuning Electrostatic Interactions To Control Orientation of GFP Protein Adsorption on Silica Surface

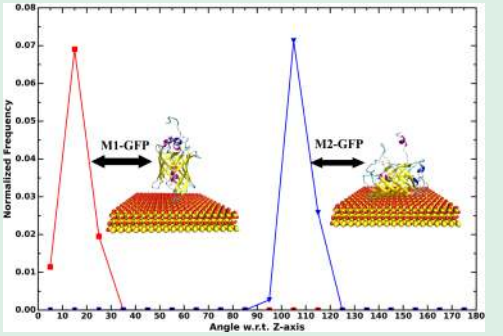
Nitin Kumar Singh, Karthik Pushpavanam, and Mithun Radhakrishna*

Cite This: https://doi.org/10.1021/acsabm.3c00125

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The adsorption of green fluorescent protein (GFP) on silica surfaces has been the subject of growing interest due to its potential applications in various fields, including biotechnology and biomedicine. In this study, we used all-atom molecular dynamics simulations to investigate the charge-driven adsorption of wild type GFP and its supercharged variants on silica surfaces. The results showed that the positively charged variant of GFP adsorbed on the negatively charged silica surface with minimal loss in its secondary structure. Further studies were conducted to understand the role of surface charge distribution on two other positively charged variants of GFP, and the results showed that the orientation of GFP on silica can be easily tuned by careful mutations of the charged amino acid residues on the GFP. This study provides valuable molecular insights into the role of electrostatic-driven adsorption of GFP and highlights the importance of charge interactions in the adsorption process.



KEYWORDS: *green fluorescent protein (GFP), adsorption, silica, molecular dynamics, mutations, electrostatics*

INTRODUCTION

Due to their unique characteristics such as small size, high surface area, and the ability to penetrate biological barriers, nanomaterials have gained traction as promising tools for targeted disease treatment.[1,2] However, when nanomaterials are introduced to biological fluids, they can form a protein corona that may impact their biological and chemical functions in vivo, leading to undesirable effects such as immune response, masking of functions, and increased toxicity.[3,4] Therefore, understanding the interactions governing protein-corona formation on nanomaterials is crucial to developing transformative nanomaterials-based solutions that can have a significant clinical impact.

The aim of this study is to elucidate the design rules governing the interactions between green fluorescent protein (GFP) and its mutant forms adsorbed on inorganic silica surfaces through molecular dynamics simulations. GFP is a suitable model for this study due to its high mutational amenability and negligible impact on size and denaturation, allowing for direct and unambiguous correlations between the mutated protein sequence and the inorganic surface.[5−7] Silica was chosen as the inorganic surface due to its biocompatibility and technological relevance in potential therapeutic applications.[8,9]

Previous research has revealed that the adsorption behavior of lysozyme and bovine serum albumin on silica surfaces is influenced by various factors such as solution pH, salt concentration, protein concentration, and nanoparticle size and surface charge.[10,11] In a 2010 study, Tosaka et al. investigated the adsorption of ribosomal protein L2 on silica surfaces and observed that the electrostatic forces were the primary driving forces in the adsorption of both protein domains on the silica surface.[12] Molecular dynamics simulations have been employed to investigate protein−silica interactions, and these studies have indicated that the adsorption behavior is influenced by the surface charge, hydrophilicity, and hydrophobicity of the silica surface.[13−15]

In 2013, Sobieściak et al.investigated the protein assemblies of His-tagged GFP proteins and found that the clustering of the proteins is a surface-mediated process, and that self-assembled monolayers (SAMs) can guide the formation of various protein arrays.[16] In 2017, Wasserberg et al. examined the interaction between red fluorescence protein (RFP) tagged with histidines and Ni^{2+}:nitrilotriacetic acid (NTA)-terminated thiols placed on an Au(111) slab. They discovered that the binding affinity increased with the length of the His tag, and the orientation could be precisely controlled by the careful placement of His tags.[17]

A recent study by Solveyra et al. investigated the surface curvature and pH dependence of protein adsorption on surfaces. The researchers observed that although pH and ionic strength play a significant role in the adsorption process, electrostatic interactions largely control the process.[18] The adsorption of

Special Issue: Computational Advances in Biomaterials

Received: February 15, 2023
Accepted: June 4, 2023

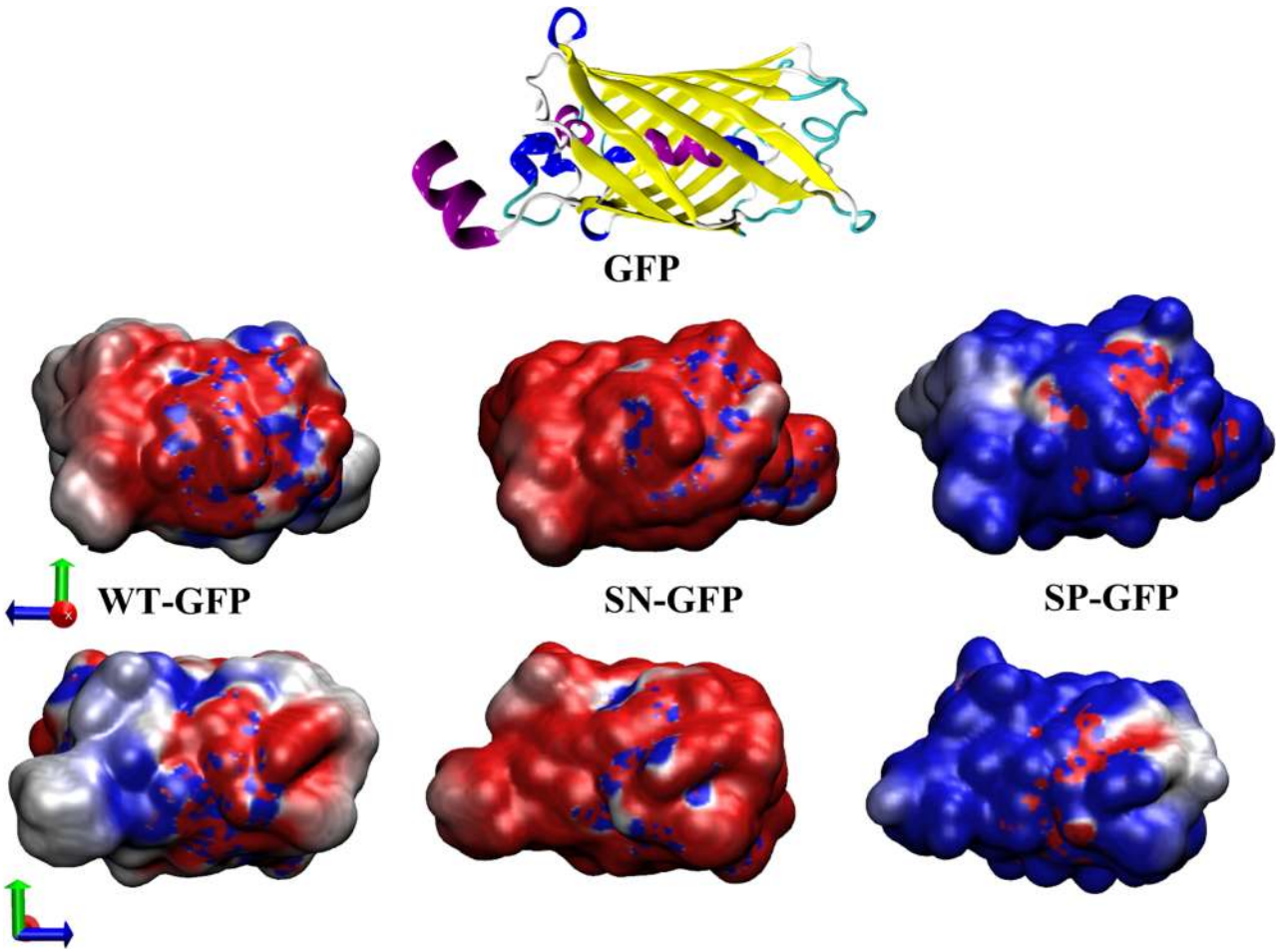


Figure 1. Elecrostatic surfaces for WT-GFP(−8e), SN-GFP (−31e), and SP-GFP (+34e). The calculation was performed using APBS software. Red represents the negative surface, blue represents the positive surface, and white represents the neutral surface.

Table 1. List of GFP Proteins and the Simulation System

no.	system	charge on protein (e)	L_x (Å)	L_y (Å)	L_z (Å)	number of waters	number of ions
1	WT-GFP	−8	90	100	90	24336	8 (Na^+)
2	positive-GFP	+34	90	100	90	24343	34 (Cl^-)
3	negative-GFP	−31	90	100	90	24187	31 (Na^+)

GFP has also been studied on gold clusters. In 2004, Collins et al. investigated the adsorption of histidine-tagged GFPs on gold surfaces and observed the formation of nanocluster films.[19] In addition to studies on the adsorption of GFP on solid surfaces, there have also been studies on GFP interactions with lipid membranes to develop biofunctionalized membranes for use as sensors.[20]

Past studies have been performed with a diverse palette of proteins and surfaces.[21,22] These studies primarily demonstrate the role of electrostatics where an overall increase in protein adsorption is observed with an increase in the difference in surface charge density between the surface and the overall protein charge. Although these studies have considerably enhanced our knowledge of electrostatic interactions, there is mounting evidence suggesting that the orientation of adsorbed proteins on surfaces plays a more significant role in determining their ultimate application (e.g., biosensors).[23,24]

Despite the importance of understanding the bound orientation, experimental acquisition of this information presents a significant challenge. Recently, Fitzkee and colleagues designed single point mutations on GB3 protein and studied their interactions with gold nanoparticles through NMR.[23] They demonstrate that GB3 K13G variant adsorbs faster when compared to the rest of the explored variants on gold nanoparticles in a mixture of GB3 K13G and GB3 which is surprising considering the small side chain of glycine. Although the above study has investigated the role of point mutations and their role in protein orientation on nanoparticle surfaces, there are limited studies that investigate the impact of multiple amino acid mutations on the protein's orientation on the nanoparticle.

Our proposed study aims to investigate the role of electrostatic interactions on the adsorption of GFP and its mutant forms onto a silica surface. The study utilizes molecular dynamics simulations to emphasize the crucial role of surface charge distribution in driving and controlling adsorption. Two new GFP mutant forms, both carrying a net positive charge of +6e, are introduced in the study. The results demonstrate that these mutant forms exhibit entirely different orientations upon adsorption to the silica surface. These findings highlight the potential for manipulating protein orientation on surfaces,

https://doi.org/10.1021/acsabm.3c00125
ACS Appl. Bio Mater. XXXX, XXX, XXX−XXX

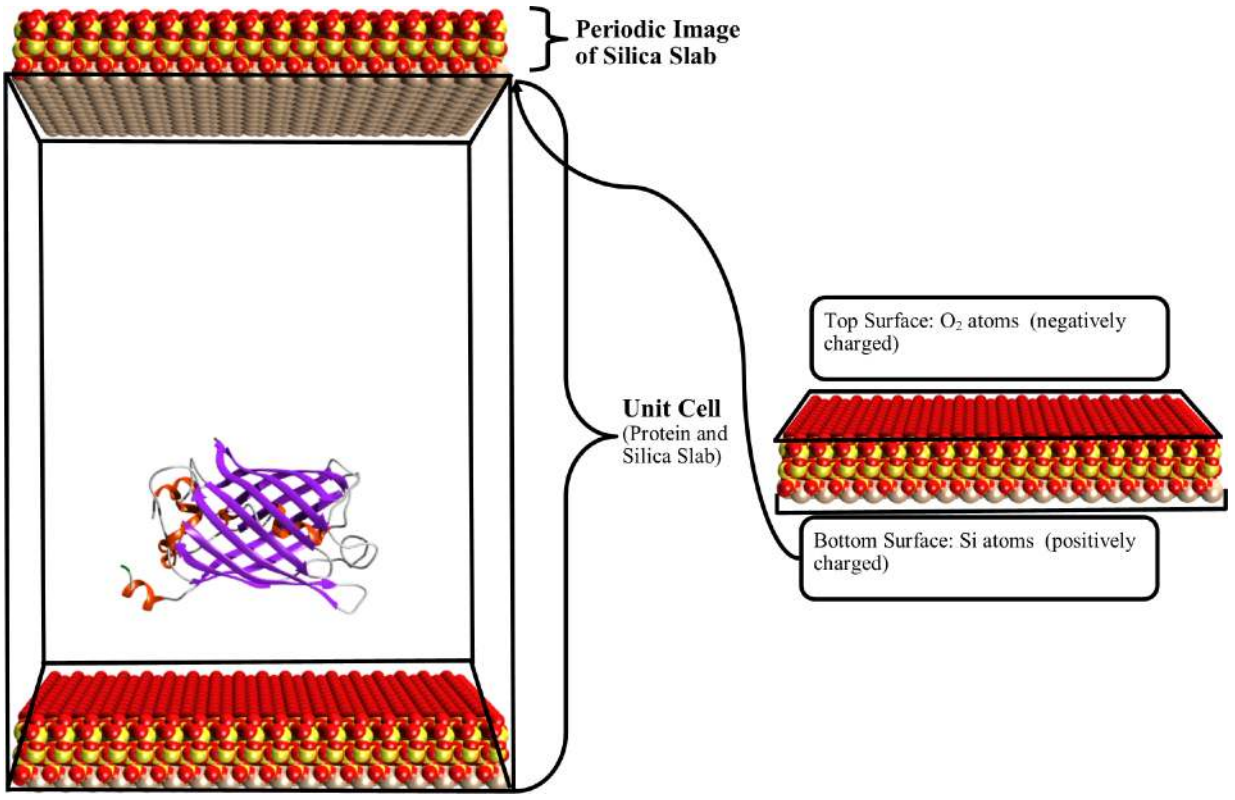


Figure 2. System setup for GFP adsorption study on the silica surface. Oxygen is represented by the red spheres, yellow spheres represent silica atoms, and brown spheres represent the uncoordinated silica atoms at the bottom layer.

which could have significant implications for a range of biotechnology applications, including drug delivery, tissue engineering, and biosensors.

MODEL AND METHODS

The wild-type green fluorescent protein (WT-GFP) was obtained from the Alphafold Protein Structure Database, with the corresponding UniProt ID P42212.[25,26] WT-GFP is a 238-residue protein with a net charge of −8e at a pH of 7. The protein has a β-barrel structure predominantly composed of β-strands that are arranged in an antiparallel manner, forming a β-sheet that wraps around to form the cylindrical shape. To create supercharged variants, multiple mutations were performed on the WT-GFP, resulting in a superpositively charged variant (SP-GFP) with a net charge of +34e and a supernegatively charged variant (SN-GFP) with a net charge of −31e.[27] The complete sequences of all three variants, WT-GFP, SP-GFP, and SN-GFP, are provided in the Supporting Information.[27] The following mutants have been studied extensively in previous experiments. Figure 1 shows the new cartoon representation of tertiary structure and surface charge density of the GFP variants calculated using Adaptive Poisson−Boltzmann Solver (APBS) software.[26]

Bulk Simulations. The bulk simulations were set up by solvating the protein with explicit water modeled using the TIP3P potential using the VMD solvate plugin.[28−30] The system was neutralized by adding either Na^+ or Cl^- ions to maintain electrical neutrality using the VMD autoionize plugin. The simulations were carried out in an initial periodic box of dimensions L_x, L_y, and L_z. The NAMD software package and CHARMM36 force field parameters were used to perform molecular dynamics (MD) simulations.[31,32] The initial system underwent energy minimization for 2500 steps using the conjugate gradient algorithm.[33] The stability of each variant in water was determined through 200 ns of isobaric isothermal (NPT) MD simulations with a time step of 2 fs. The pressure was maintained at 1 bar using the Langevin barostat method with a decay period of 50 ps and damping time of 100 ps, and the temperature was controlled at 300 K using Langevin dynamics.[34] The bonds involving hydrogen atoms were constrained using the SHAKE algorithm.[35] Nonbonded interactions were calculated with a cutoff value of 12 Å and a switching distance of 10 Å, and pairlistdist was set to 14 Å. The long-range electrostatic interactions were calculated using the particle-mesh Ewald summation method.[36] In-house Tcl and Python scripts were used for the analysis. The secondary structure calculation was performed using STRIDE[37] module of VMD. The averaged Radius of gyration is calculated as $\langle R_g \rangle = \left\langle \sqrt{\frac{1}{N}\sum_1^N (r_i - r_{cm})^2} \right\rangle$, where r_i is the position of center of mass of the I^{th} residue, r_{cm} is the center of mass of entire protein, N is the total number of residues, and $\langle\rangle$ represents the ensemble average. Table 1 shows the list of systems, including system size, charge, initial box dimensions (L_x, L_y, L_z), number of water molecules, and number of ions for the MD simulations.

Adsorption on Silica. A silica surface is modeled by using a (101) slab of α-cristabolite. The SiO_2 slab has an intrinsic dipole moment across it because the surface was modeled as ions fixed in space. The slab was cut in a way that the siloxide groups are at the top and the under-coordinated Si species are at the bottom. Even though the slab is neutral and stoichiometric, an electric field is generated above the surfaces, simulating the environment above the charged surfaces that have been observed experimentally. The silica surface is placed at $z = 0$ plane, and the molecular motion of atoms are restricted using the fixed command in NAMD. The protein is solvated and neutralized using ions using the same protocol as described above. The simulated system contains protein, water, counterions, and the silica surface. The force field parameters for the silica slab were obtained from Interface force field of CHARMM using the CHARMM-GUI web application.[38] The system undergoes energy minimization for 2500 steps followed by 200 ns of molecular dynamics simulations carried out in a canonical ensemble (NVT). Figure 2 shows the model silica surface used in the simulation. Molecular dynamics simulations of adsorption of WT-GFP, and its variants were carried out in a the simulation box comprising the protein, water, counterions, and the silica surface with a periodic boundary imposed in all the directions.

https://doi.org/10.1021/acsabm.3c00125
ACS Appl. Bio Mater. XXXX, XXX, XXX−XXX

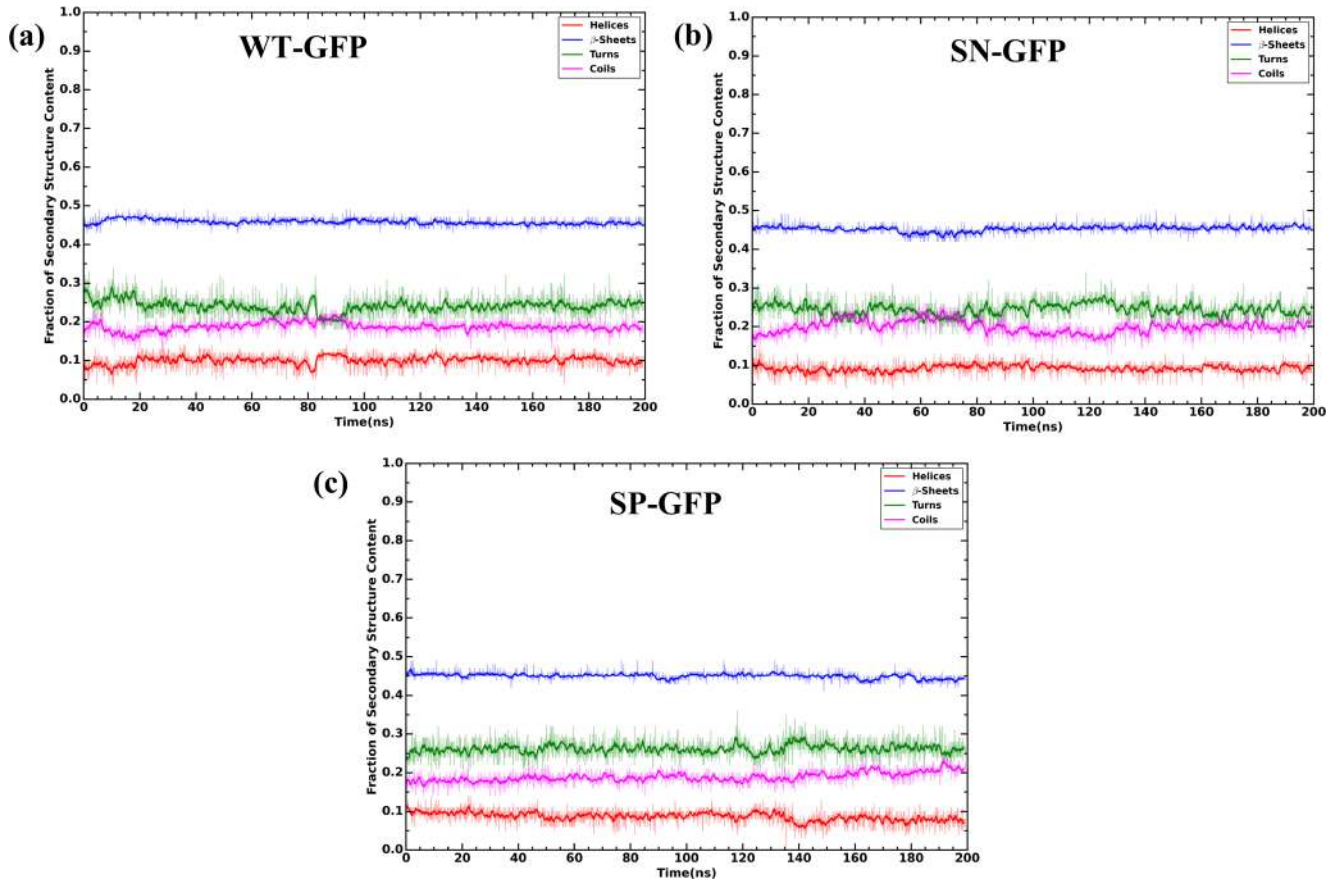


Figure 3. Fraction of secondary structure content for (a) WT-GFP (−8e), (b) SN-GFP (−31e), and (c) SP-GFP (+34e) as a function of time of simulations carried out in an NPT ensemble at 300 K and 1 bar.

RESULTS AND DISCUSSIONS

Bulk Simulations. The secondary structure timeline of WT-GFP and its variants, as shown in Figure 3, was obtained through molecular dynamics simulations in a bulk environment. The timeline clearly illustrates that the tertiary structure of these proteins is predominantly composed of β-sheets (≈46%). Throughout the simulations, there was little to no change in the secondary structure composition of WT-GFP and its supercharged variants, indicating their structural stability in water. To assess any size changes, the radius of gyration (R_g) of all three variants was calculated and no significant variations were observed, further demonstrating the structural stability of these proteins in water at 300 K (Figure 4).

Adsorption of Silica. The silica surface modeled in this study has an overall neutral electrical charge; however, the oxygen groups that are exposed to both the protein and water carry a negative charge of −0.55e. It is worth noting that the bottom layer of the silica surface cut from the crystal has a positive charge due to the presence of Si atoms. As demonstrated in Figure 2 under periodic boundary conditions, the top surface of the periodic image is positively charged.

In Figure 5, snapshots of the equilibrium conformations of WT-GFP and its variants are depicted during adsorption onto a silica surface. WT-GFP and SN-GFP, which carry negative charges (−8e and −31e, respectively), are repelled by the negatively charged silica surface at the z = 0 plane but are attracted to the positively charged top surface of the periodic image. It is worth noting that in experimental conditions, WT-GFP and SN-GFP would likely be found in the supernatant, rather than adsorbed to the silica surface. The adsorption of these proteins on the top surface is a simulation artifact resulting from the positively charged Si atoms present on the bottom layer of the cut silica crystal. During adsorption, the GFP variants WT-GFP and SN-GFP align themselves such that the central axis is parallel to the xy plane, so that the negatively charged amino acid residues on their surface make maximum contact with the positively charged periodic image of the surface. In contrast, the positively charged variant SP-GFP behaves differently upon adsorption. With a positive charge of +34e, SP-GFP is attracted to the negatively charged bottom surface of the silica crystal due to the interactions between the surface's oxygen atoms and the positively charged amino acid residue present on the protein. SP-GFP aligns itself parallel to the xy plane while adsorbing on the bottom surface.

Figure 6 illustrates the secondary structure timeline of WT-GFP and its supercharged variants during adsorption on silica. WT-GFP, with a net charge of −8e, adsorbs on the top surface (i.e., the periodic image of the bottom layer of the silica surface at z = 0). The protein's β-sheet content decreases to approximately 25% from its bulk value of approximately 46%. This decrease in β-sheet content is offset by an increase in coil content from around 20% in bulk to around 40% upon adsorption. This transformation of β-sheets into coils is due to the electrostatic attraction between the negatively charged WT-GFP and positively charged upper surface. Due to the presence of electric field along the $-z$ direction, few of the positively charged amino acid residues are also pulled downward resulting in the decreased β-sheet content. SN-GFP, with a net negative charge

https://doi.org/10.1021/acsabm.3c00125
ACS Appl. Bio Mater. XXXX, XXX, XXX−XXX

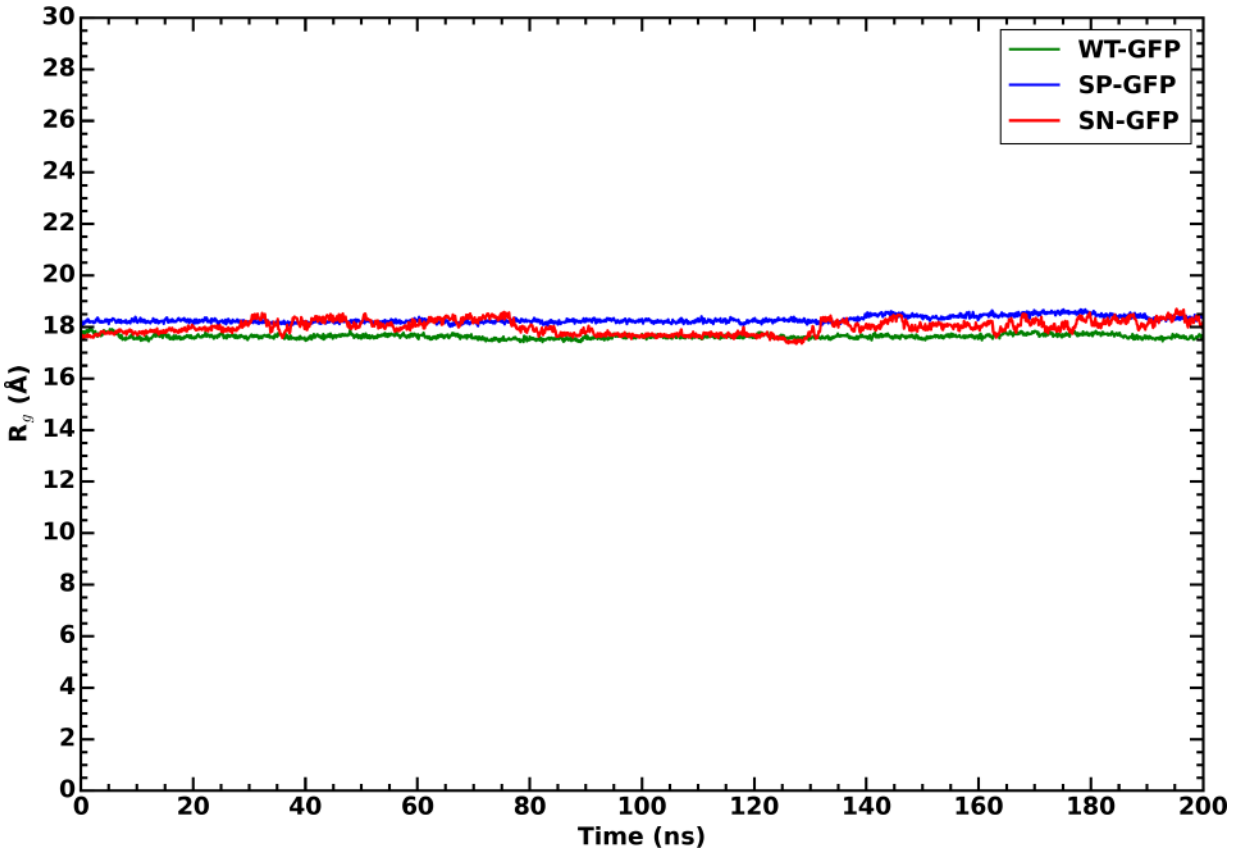


Figure 4. Radius of gyration of WT-GFP (green), SP-GFP (blue), and SN-GFP (red) as a function of time of simulations carried out in an NPT ensemble at 300 K and 1 bar.

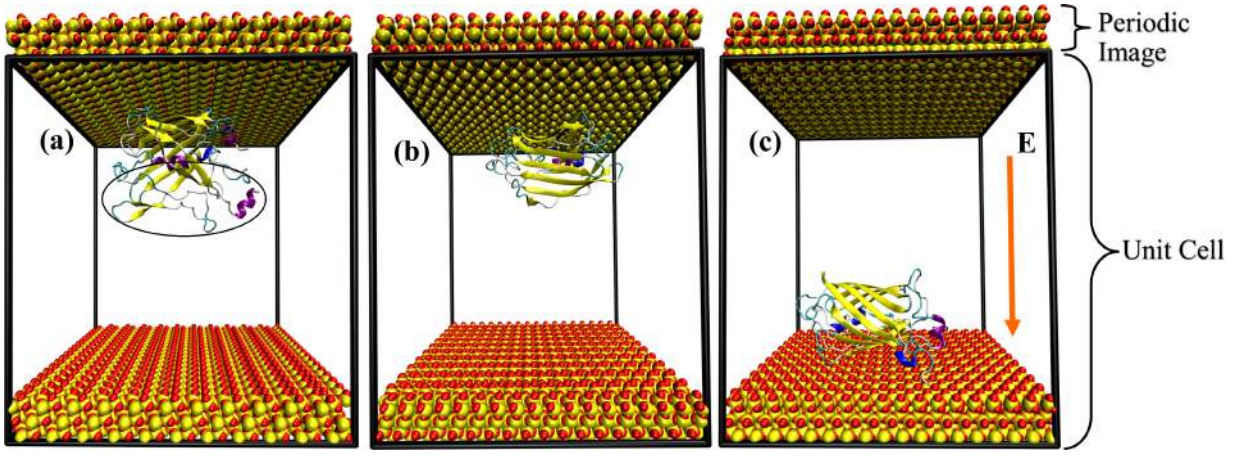


Figure 5. Behavior of (a) WT-GFP, (b) SN-GFP, and (c) SP-GFP in the presence of the silica surface in an NVT ensemble at 300 K. E represents the electric field due to the presence of a periodic image of the silica slab.

of −31e, also adsorbs on the top surface. However, despite having a much higher net negative charge compared to WT-GFP, there is only a slight decrease in the protein's β-sheet content. There is an increase in both coil and turn content. Since many of the positively charged groups present in WT-GFP were mutated to negatively charged amino acid residues in SN-GFP, we did not observe the pulling of residues in the presence of field as observed in the case of WT-GFP. When SP-GFP, which has a net charge of +34e, adsorbs onto the negatively charged silica surface at $z = 0$, a significant amount of its β-sheet structure remains intact as shown in Figure 6(c).

Figure 7 illustrates the radius of gyration (R_g) timeline of WT-GFP and its variants upon adsorption on silica. Unlike their behavior in the bulk state, the three proteins show a difference in their equilibrium values of the radius of gyration upon adsorption. This is due to the partial unfolding of the protein, which transforms its β-sheet structures into coils or turns. The results show that the β-sheet content decreases the most for WT-GFP, which is reflected as an increase in R_g for that specific system compared to SN-GFP and SP-GFP. We have also performed the RMSD and RMSF analysis in order to evaluate the stability of the three proteins, and the results are provided in Supporting Information Figures S-1 and S-2.

https://doi.org/10.1021/acsabm.3c00125
ACS Appl. Bio Mater. XXXX, XXX, XXX−XXX

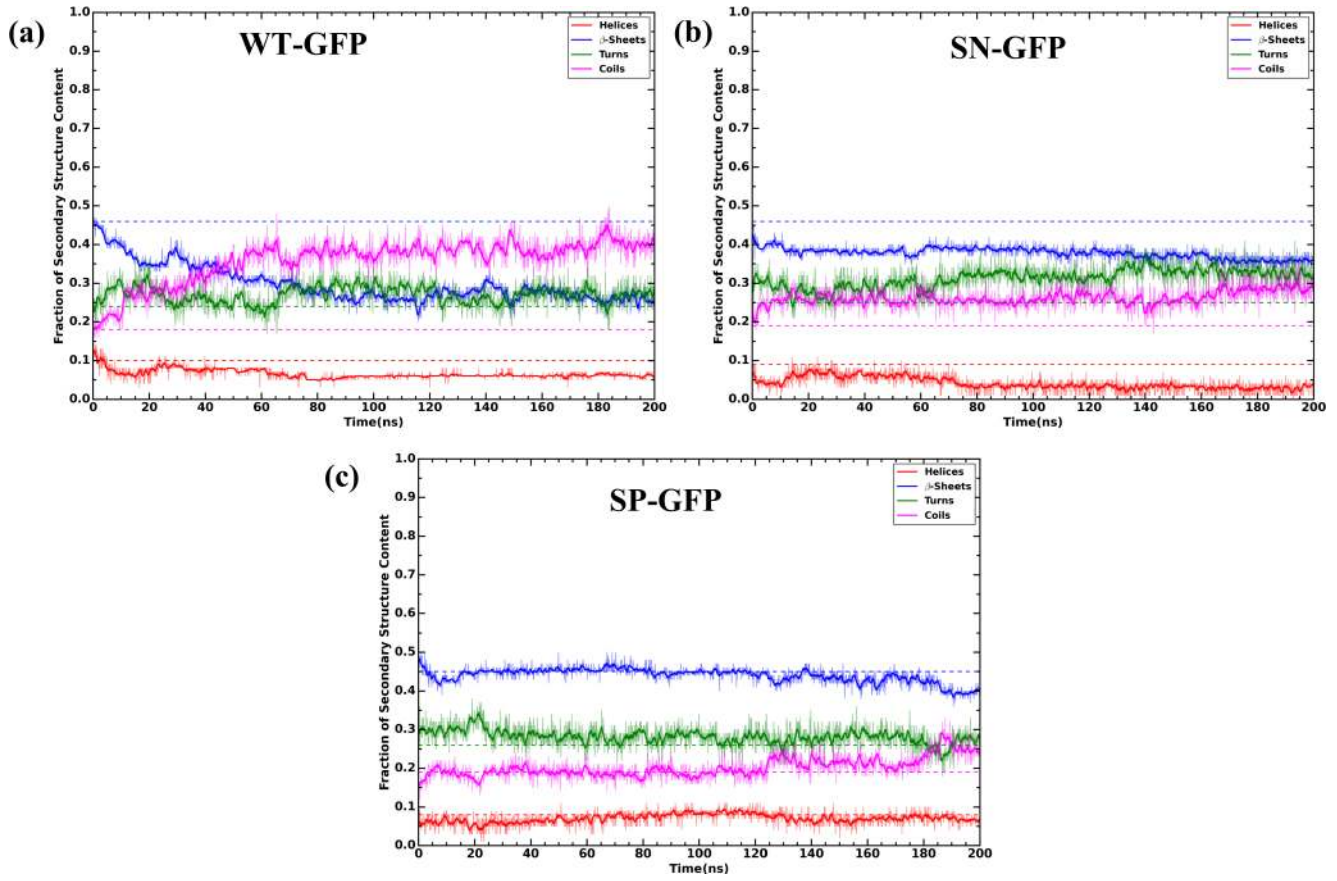


Figure 6. Fraction of secondary structure content for (a) WT-GFP(−8e), (b) SN-GFP (−31e), and (c) SP-GFP (+34e) as a function of time of simulations carried out in an NVT ensemble at 300 K on silica surface. The dotted lines represent the secondary structure content in bulk water for the respective proteins.

Controlling the Orientation of GFP on Silica. The orientation of green fluorescent protein (GFP) when adsorbed on silica surfaces is crucial for various applications, including improving biocompatibility, enhancing performance in biosensors, modifying silica surface properties, and boosting the fluorescence properties and stability of GFP. Wild-type GFP (WT-GFP) and its supernegative variant (SN-GFP) are repelled by silica surfaces at the bottom, whereas the superpositively charged variant (SP-GFP) is strongly attracted to the negatively charged silica surface, adhering in such a way that its central axis is parallel to the *xy* plane with its N-terminus and C-terminus facing sideways. There is increasing interest in fusing other proteins with GFP to visualize and track the localization and movement of target proteins in living cells or organisms. For a successful fusion, the N- and C-terminus of GFP should point away from the surface.

Our study aims to create variants of WT-GFP with the same net charge but distinct orientations when adsorbed on silica surfaces. Using the APBS software, we analyzed the surface charge distribution of WT-GFP, which has a net negative charge of −8e. On the basis of the surface charge distribution and desired orientation, we created two variants of WT-GFP with identical charges, named M1-GFP and M2-GFP. Both variants have a net positive charge of +6e and display variations in their surface charge distribution. Figure 8 and Figure 9 show the structure of M1-GFP and M2-GFP along with the point of mutations on WT-GFP. Table 2 shows the mutation points and the type of mutations performed on the WT-GFP. The sequence comparison of WT-GFP with M1-GFP and M2-GFP is provided in Supporting Information Figures S-3 and S-4.

Molecular dynamics (MD) simulations at 300 K using an NPT ensemble showed that both mutants preserve their secondary structure in the bulk, and their structural content closely resembles that of WT-GFP as shown in Figure 10. The radius of gyration, rmsd, and rmsf values of M1-GFP and M2-GFP closely matched those of WT-GFP, and the results are shown in Supporting Information Figures S-5, S-6, and S-7 respectively. The adsorption of M1-GFP and M2-GFP on the silica surface is driven by the electrostatic interactions between the negatively charged silica surface and the positively charged protein residues. Figures 11 and 12 display the secondary structure timeline of M1-GFP and M2-GFP upon adsorption on the silica surface. M2-GFP shows a slight decrease in its β-sheet content, while M1-GFP experiences a larger decrease in its β-sheet content, but both proteins are able to retain their structural stability upon adsorption.

However, the orientation of M1-GFP and M2-GFP upon adsorption differs dramatically. M1-GFP adsorbs with its axis perpendicular to the *xy* plane, whereas M2-GFP adsorbs with its axis parallel to the *xy* plane. To quantify the orientation, we calculated a parameter $\langle\theta\rangle$, representing the ensemble average of the angle between the central axis of the protein and the normal to the *z*-surface. Figure 13 shows the normalized probability distribution of $\langle\theta\rangle$ for SP-GFP, M1-GFP, and M2-GFP in bulk and upon adsorption on the silica surface. In the bulk, we observe a uniform distribution, as expected in an isotropic system. However, upon adsorption on the silica surface, we

https://doi.org/10.1021/acsabm.3c00125
ACS Appl. Bio Mater. XXXX, XXX, XXX−XXX

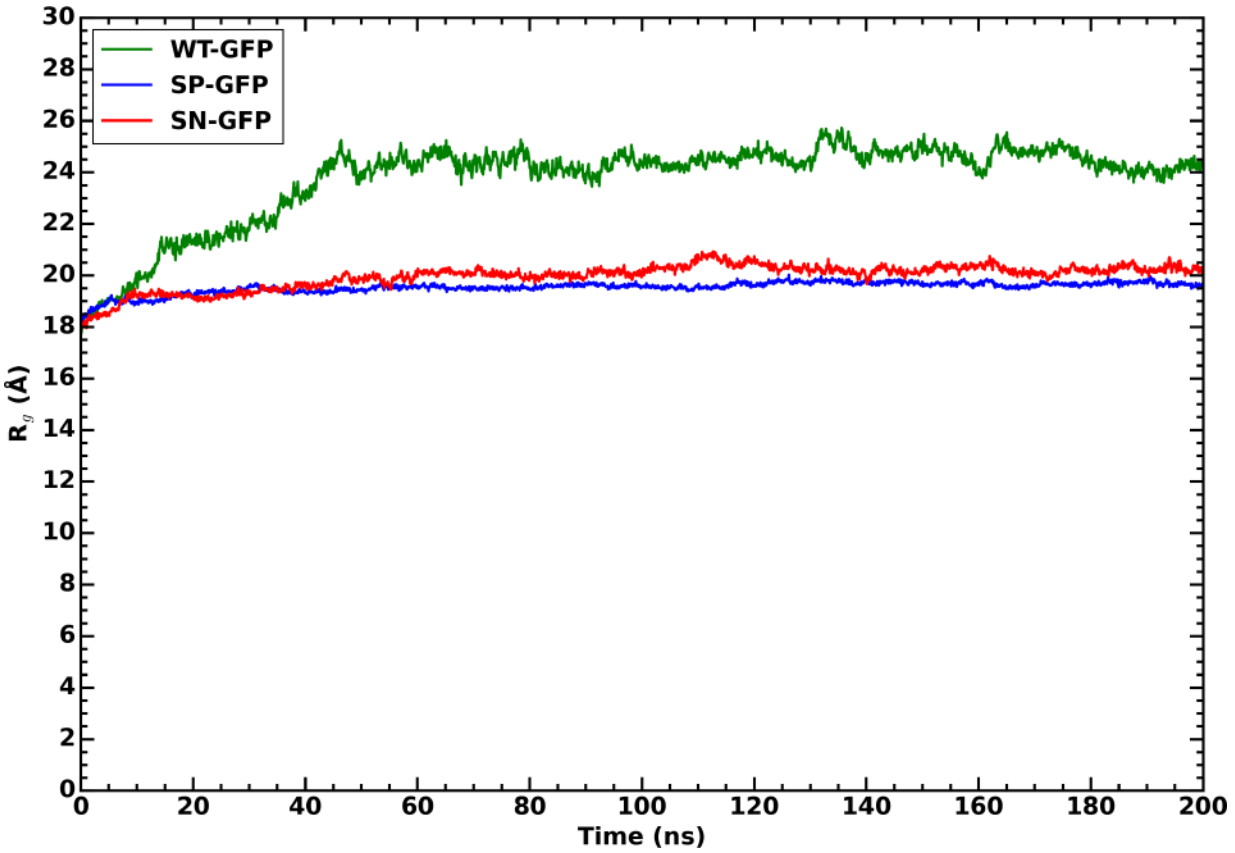


Figure 7. Radius of gyration of WT-GFP (green), SP-GFP (blue), and SN-GFP (red) as a function of time of simulations carried out in an NVT ensemble at 300 K on a silica surface.

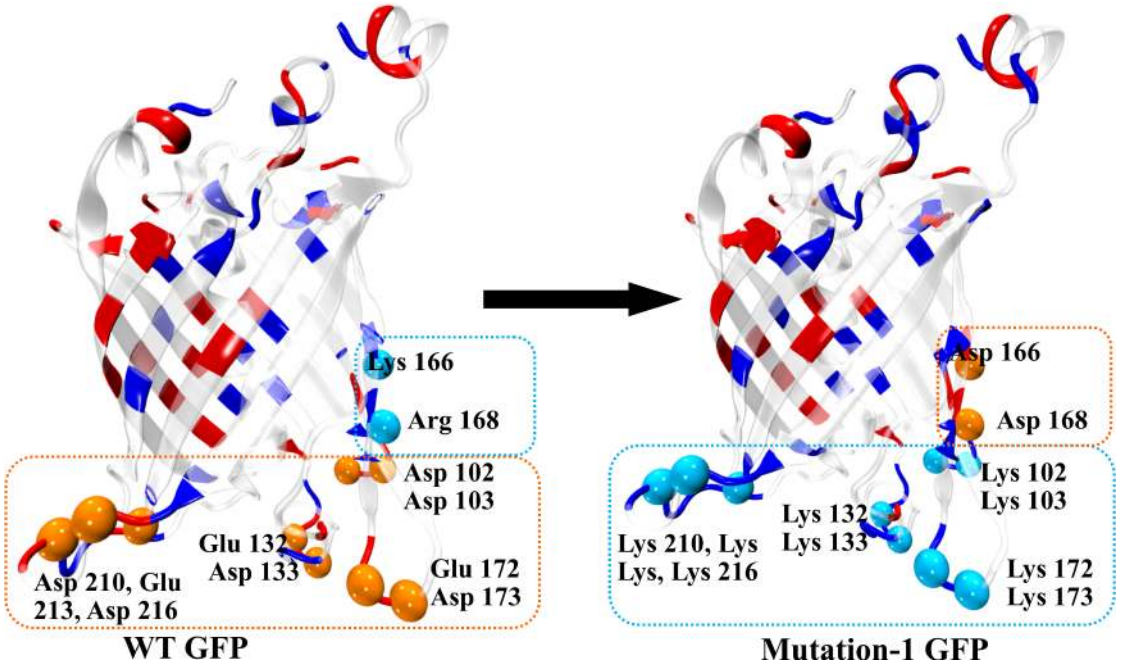


Figure 8. Mutations points on WT-GFP for mutation-1 (M1-GFP).

observe a different behavior: while SP-GFP and M2-GFP adsorb perpendicular to the surface norm (i.e., $\langle\theta\rangle \approx 75°$), M1-GFP adsorbs parallel to the surface norm (i.e., $\langle\theta\rangle \approx 15°$). Our findings demonstrate that controlling the surface charge distribution can regulate the adsorption orientation of GFP variants on silica surfaces.

https://doi.org/10.1021/acsabm.3c00125
ACS Appl. Bio Mater. XXXX, XXX, XXX−XXX

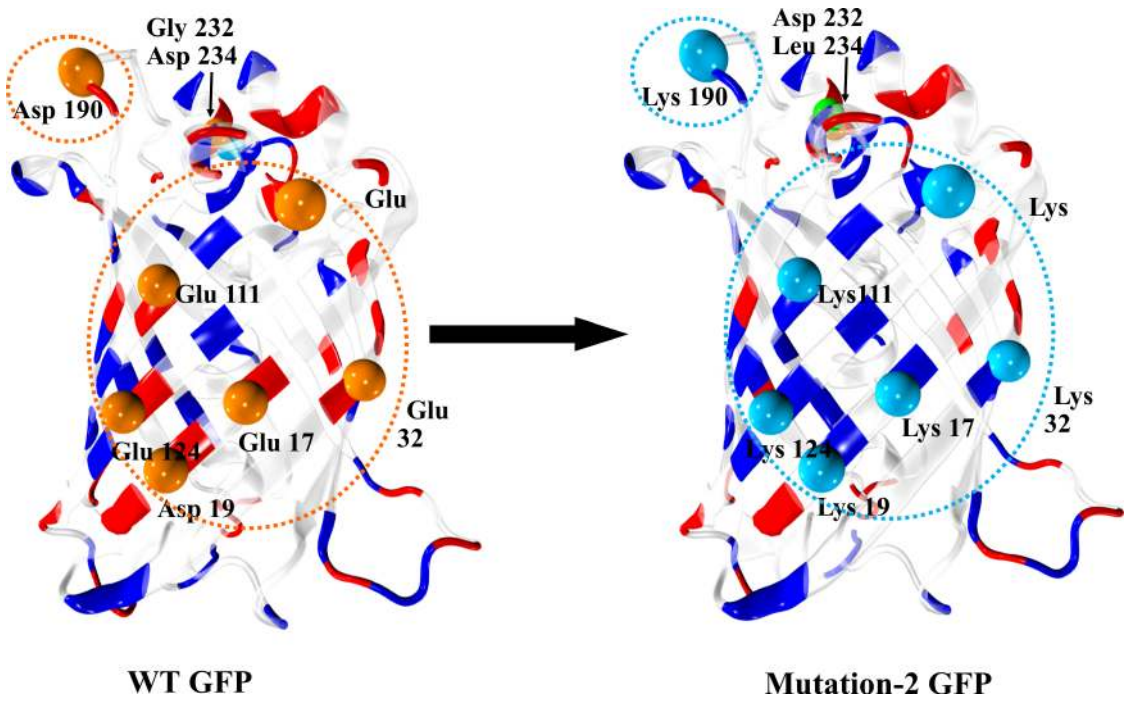


Figure 9. Mutations points on WT-GFP for mutation-2 (M2-GFP).

Table 2. Mutation Points on WT-GFP for M1-GFP and M2-GFP

protein	residue number	WT-GFP	mutation
M1-GFP	102	Asp	Lys
	103	Asp	Lys
	132	Glu	Lys
	133	Asp	Lys
	172	Glu	Lys
	132	Asp	Lys
	210	Asp	Lys
	213	Glu	Lys
	216	Asp	Lys
	166	Lys	Asp
	168	Arg	Asp
M2-GFP	17	Glu	Lys
	19	Asp	Lys
	32	Asp	Lys
	111	Glu	Lys
	115	Glu	Lys
	124	Glu	Lys
	190	Asp	Lys
	232	Gly	Asp
	234	Asp	Leu

DISCUSSION

Artifact of the Electric Field. The adsorption of negatively charged WT-GFP and SN-GFP on the top surface which is positively charged is an artifact of the simulation setup as mentioned in the manuscript. We want to clarify that the majority of the loss in the β-sheet content for WT-GFP arises due to the unfolding of positively charged residues ranging from 148 to 187 that face the negatively charged surface at the bottom. This is as a result of pulling of these residues due to the electric field. This artifact is absent in the case of SN-GFP since these residues are mutated to negatively charged amino acid residues. Therefore, we do not see the pulling of these residues by electric field in the case of SN-GFP even though it contains a higher net negative charge compared to WT-GFP.

To further validate this, we carried out additional simulations by increasing the length of the simulation box in the z direction. An increase in the z direction leads to reduction in the electric field, i.e., in the range as z tends to infinity, the protein adsorbed at the top does not feel the negative surface placed at the bottom. Figure 14 and Figure S-13, respectively, show the β-sheet content and the radius of gyration of WT-GFP adsorbed on the top surface as a function of z. It is clearly evident that as z increases, there is an increase in the β-sheet content which proves that the majority of the unfolding was caused by the artifact of the electric field at smaller box sizes.

Adsorption of Multiple Proteins. In order to evaluate the effect of interprotein interaction on the adsorption behavior, we conducted additional simulations with systems comprising two and four proteins. Specifically, we simulated the adsorption of four proteins, each on M1-GFP and M2-GFP, on a silica surface. We observed that, qualitatively, there was no change in the adsorption orientation when compared to that of the single protein systems. To validate this observation, we deliberately started the simulation with a random initial orientation far away from equilibrium. The proteins exhibited a similar behavior to that observed in the single protein system, dramatically changing their orientation. Supporting Information Figures S-16 and S-17 show the initial and final configuration of multiple M1-GFP and M2-GFP adsorbed on the silica surface. In future work, we would like to investigate the self-assembly pattern of multiple proteins on silica surfaces.

CONCLUSION

In summary, this study aimed to investigate the impact of electrostatic interactions on the adsorption behavior of GFP proteins on silica surfaces. Through molecular dynamics simulations, the study demonstrated that while negatively charged forms of GFP are repelled from negatively charged

https://doi.org/10.1021/acsabm.3c00125
ACS Appl. Bio Mater. XXXX, XXX, XXX−XXX

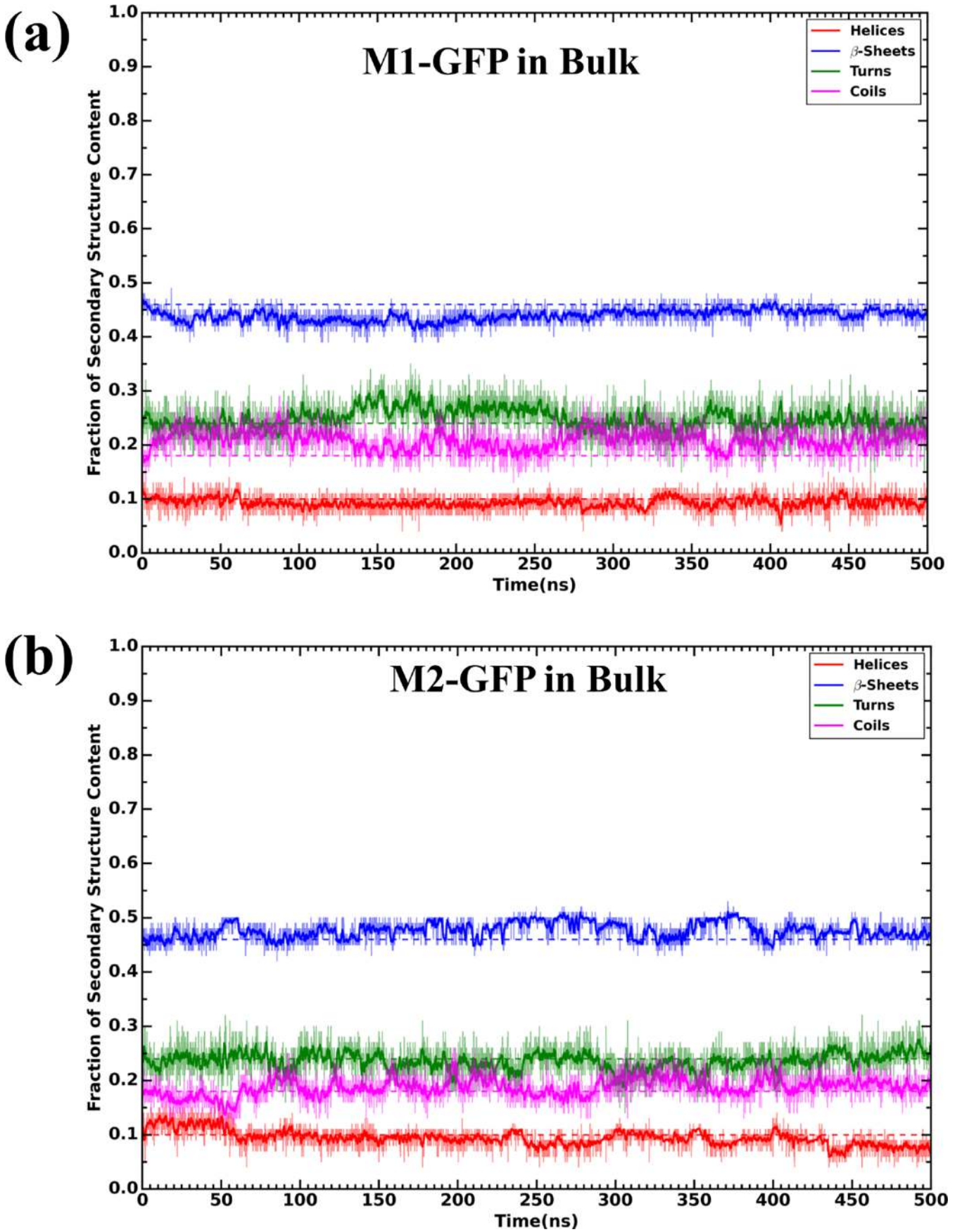


Figure 10. Fraction of secondary structure content for (a) M1-GFP, (b) M2-GFP in bulk water in NPT ensemble at 300 K. The dotted lines represent the secondary structure content of the WT-GFP in bulk water.

https://doi.org/10.1021/acsabm.3c00125
ACS Appl. Bio Mater. XXXX, XXX, XXX−XXX

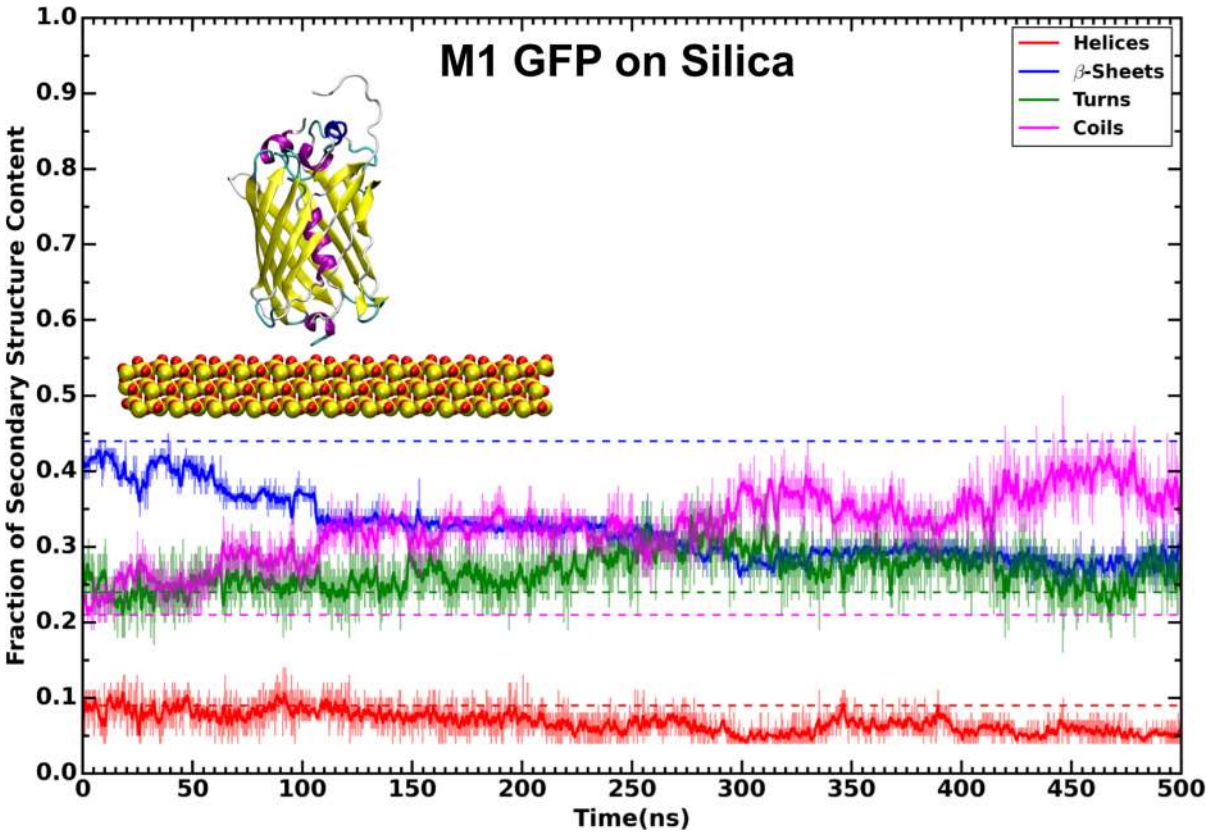


Figure 11. Fraction of secondary structure content for M1-GFP in the presence of a silica surface in an NVT ensemble at 300 K. The dotted lines represent the secondary structure content content of M1-GFP in the bulk water simulations.

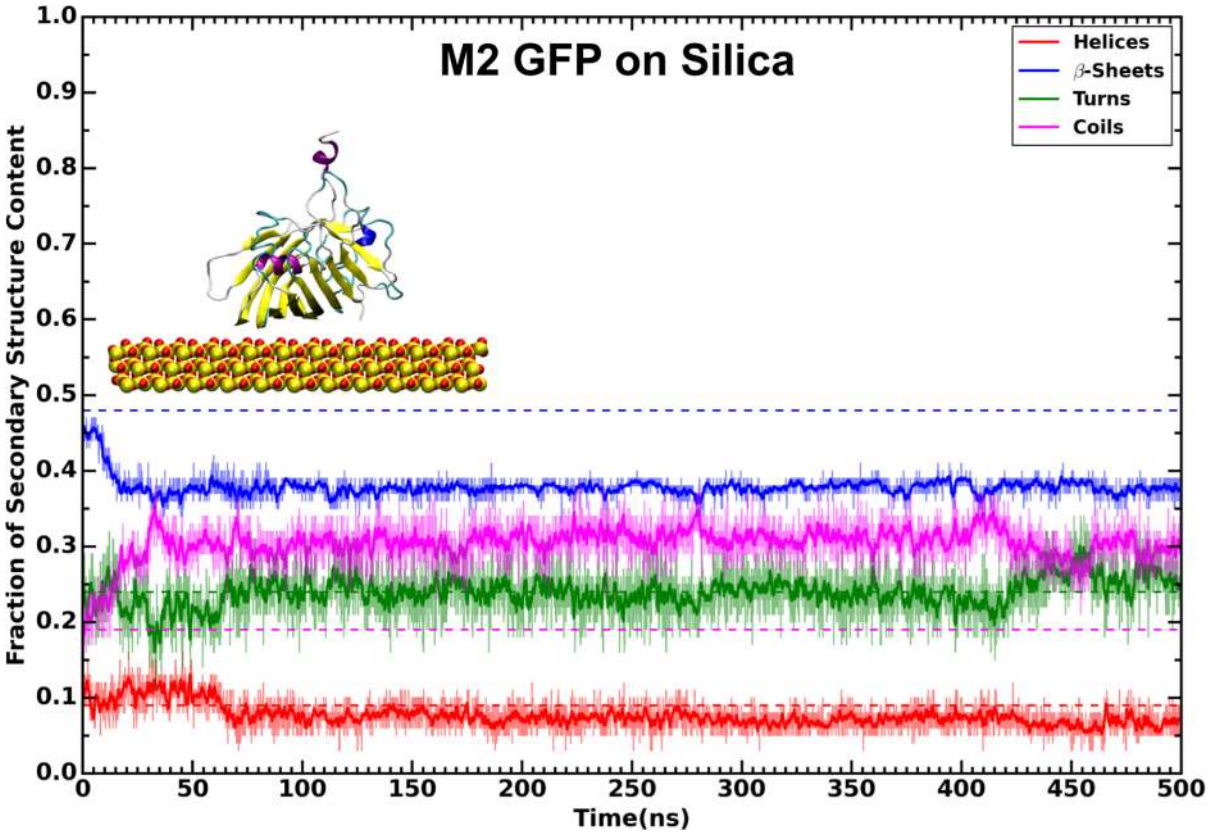


Figure 12. Fraction of secondary structure content for M2-GFP in the presence of a silica surface in an NVT ensemble at 300 K. The dotted lines represent the secondary structure content of M2-GFP in the bulk water simulations.

https://doi.org/10.1021/acsabm.3c00125
ACS Appl. Bio Mater. XXXX, XXX, XXX−XXX

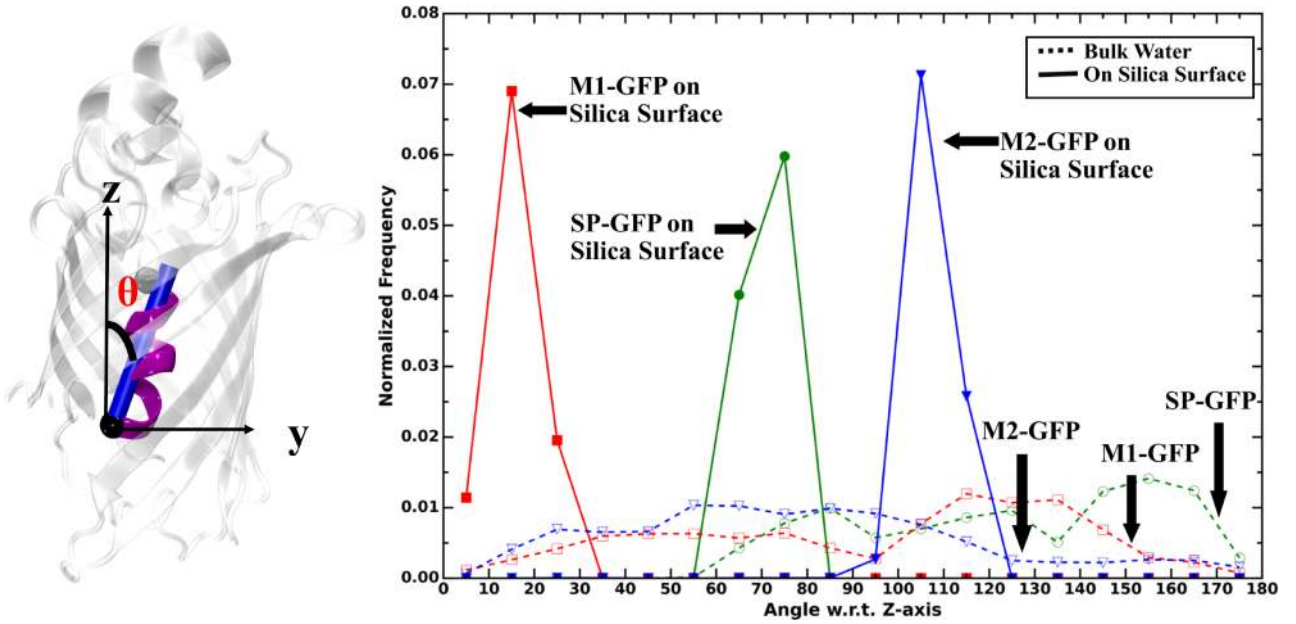


Figure 13. Orientation of super positive (SP), mutation-1 (M1), and mutation-2 (M2) GFP proteins with respect to the *z*-axis in bulk water and on the silica surface.

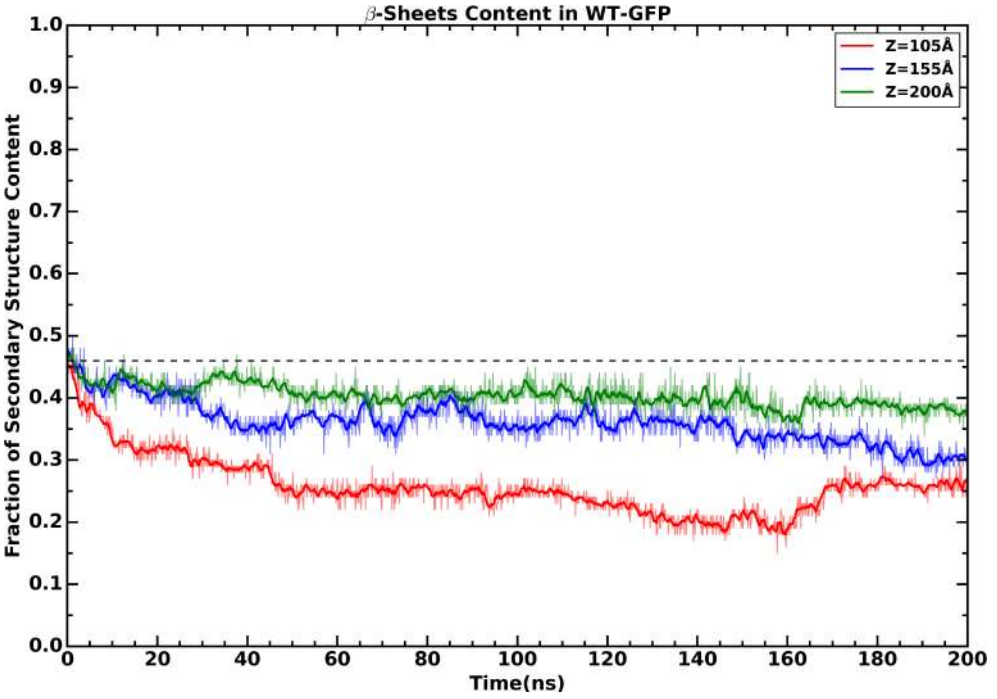


Figure 14. β-Sheet content of WT-GFP in different simulation boxes with z = 105, 155, and 200 Å as a function of time of simulations carried out in an NVT ensemble at 300 K on a silica surface. The dotted line represents the secondary structure content in the bulk water simulation.

silica surfaces, the superpositive mutant form (SP-GFP) is adsorbed on the surface with minimal loss of secondary structure content, with its central axis parallel to the *xy* plane. Furthermore, the study showed that surface charge distribution, rather than net charge, plays a critical role in determining the orientation of adsorbed proteins on the silica surface. By creating two new GFP mutant forms with a net positive charge of +6e, the study found that they both adsorb on the silica surface but in completely different orientations. These findings highlight the potential of controlling protein orientation on surfaces for various biotechnology applications, such as biosensors, drug delivery, and tissue engineering. The molecular dynamics simulations performed in this study provided valuable insights into the adsorption behavior of proteins on silica surfaces and could be useful for designing new materials with specific protein adsorption properties.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsabm.3c00125.

https://doi.org/10.1021/acsabm.3c00125
ACS Appl. Bio Mater. XXXX, XXX, XXX−XXX

Sequences for WT-GFP, SP-GFP, SN-GFP, M1-GFP, and M2-GFP. Figure S-1: RMSD of WT-GFP (green), SP-GFP (blue), and SN-GFP (red) as a function of time of simulations carried out in an NVT ensemble at 300 K on a silica surface. Figure S-2: RMSF of WT-GFP (green), SP-GFP (blue), and SN-GFP (red) as a function of time of simulations carried out in an NVT ensemble at 300 K on a silica surface. Figure S-3: Sequence comparison chart of WT-GFP and M1-GFP. The blue and red patches at the top of the sequence represents the mutations, and the gray patches represent the conserved regions. Figure S-4: Sequence comparison chart of WT-GFP and M2-GFP. The blue and red patches at the top of the sequence represent the mutations, and the gray patches represent the conserved regions. Figure S-5: Radius of gyration of M1-GFP and M2-GFP in bulk water in an NPT ensemble at 300 K and 1 bar. The black dotted line represents the RMSD of WT-GFP in bulk water. Figure S-6: RMSD of M1-GFP and M2-GFP in bulk water in an NPT ensemble at 300 K and 1 bar. The black dotted line represents the RMSD of WT-GFP in bulk water. Figure S-7: RMSF of M1-GFP and M2-GFP in bulk water in an NPT ensemble at 300 K and 1 bar. The black dotted line represent the RMSD of WT-GFP in bulk water. Figure S-8: Total number of contacts for the protein with the silica surface. Figure S-9: Contact map showing the interacting residues of SP-GFP with the silica surface. Figure S-10: Contact map showing the interacting residues of M1-GFP with the silica surface. Figure S-11: Contact map showing the interacting residues of M2-GFP with the silica surface. Figure S-12: Sequence comparison chart of SN-GFP and WT-GFP. The blue patches at the top of sequence represent the mutations, and the gray patches represent the conserved regions. In the SN-GFP sequence most of the lysine residues (K) are converted to negatively charged residues, aspartic acid (D) or glutamic acid (E). Figure S-13: Radius of gyration of WT-GFP in different simulation boxes with z = 105,155, and 200 Å as a function of time of simulations carried out in an NVT ensemble at 300 K on a silica surface. Figure S-14: M1 and M2 orientation on a silica surface. Figure S-15: Adsorption behavior of GFP proteins in the presence of two proteins on a silica surface in an NVT ensemble at 300 K. (a) Adsorption behavior of SP-GFP in the presence of SN-GFP. (b) Adsorption behavior of SP-GFP in the presence of another SP-GFP. (c) Adsorption behavior of SP-GFP in the presence of WT-GFP. Figure S-16: Multiple M1-GFP adsorption on a silica surface in an NVT ensemble at 300 K. Figure S-17: Multiple M2-GFP adsorption on a silica surface in an NVT ensemble at 300 K (PDF)

Orientation of M1-GFP on adsorption on the silica surface (MP4)

Orientation of M2-GFP on adsorption on the silica surface (MP4)

AUTHOR INFORMATION

Corresponding Author

Mithun Radhakrishna − *Discipline of Chemical Engineering Engineering and Center for Biomedical Engineering, Indian Institute of Technology (IIT) Gandhinagar, Palaj, Gujarat 382355, India;* orcid.org/0000-0001-7127-4744; Email: mithunr@iitgn.ac.in

Authors

Nitin Kumar Singh − *Discipline of Chemical Engineering Engineering, Indian Institute of Technology (IIT) Gandhinagar, Palaj, Gujarat 382355, India*

Karthik Pushpavanam − *Discipline of Chemical Engineering Engineering, Indian Institute of Technology (IIT) Gandhinagar, Palaj, Gujarat 382355, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsabm.3c00125

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

M.R., K.P., and NKS thank the HPC facility at IIT Gandhinagar and PARAM Ananta NSM facility for support with computational resources. N.K.S. thanks the Ministry of Education, Govt. of India, for a doctoral fellowship.

REFERENCES

(1) Wong, X. Y.; Sena-Torralba, A.; Alvarez-Diduk, R.; Muthoosamy, K.; Merkoçi, A. Nanomaterials for nanotheranostics: tuning their properties according to disease needs. *ACS Nano* **2020**, *14*, 2585−2627.

(2) Ding, Q.; Cui, J.; Shen, H.; He, C.; Wang, X.; Shen, S. G.; Lin, K. Advances of nanomaterial applications in oral and maxillofacial tissue regeneration and disease treatment. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology* **2021**, *13*, No. e1669.

(3) Umapathi, A.; Kumawat, M.; Daima, H. K. Engineered nanomaterials for biomedical applications and their toxicity: a review. *Environmental Chemistry Letters* **2022**, *20*, 445−468.

(4) Yao, J.; Wang, H.; Chen, M.; Yang, M. Recent advances in graphene-based nanomaterials: properties, toxicity and applications in chemistry, biology and medicine. *Microchimica Acta* **2019**, *186*, 1−25.

(5) Misteli, T.; Spector, D. L. Applications of the green fluorescent protein in cell biology and biotechnology. *Nature biotechnology* **1997**, *15*, 961−964.

(6) Gerdes, H.-H.; Kaether, C. Green fluorescent protein: applications in cell biology. *FEBS letters* **1996**, *389*, 44−47.

(7) Zimmer, M. Green fluorescent protein (GFP): applications, structure, and related photophysical behavior. *Chem. Rev.* **2002**, *102*, 759−782.

(8) Luckarift, H. R.; Spain, J. C.; Naik, R. R.; Stone, M. O. Enzyme immobilization in a biomimetic silica support. *Nature biotechnology* **2004**, *22*, 211−213.

(9) Carlsson, N.; Gustafsson, H.; Thörn, C.; Olsson, L.; Holmberg, K.; Åkerman, B. Enzymes immobilized in mesoporous silica: a physical−chemical perspective. *Advances in colloid and interface science* **2014**, *205*, 339−360.

(10) Kubiak, K.; Mulheran, P. A. Molecular dynamics simulations of hen egg white lysozyme adsorption at a charged solid surface. *J. Phys. Chem. B* **2009**, *113*, 12189−12200.

(11) Kubiak-Ossowska, K.; Cwieka, M.; Kaczynska, A.; Jachimska, B.; Mulheran, P. A. Lysozyme adsorption at a silica surface using simulation and experiment: effects of pH on protein layer structure. *Phys. Chem. Chem. Phys.* **2015**, *17*, 24070−24077.

(12) Tosaka, R.; Yamamoto, H.; Ohdomari, I.; Watanabe, T. Adsorption mechanism of ribosomal protein L2 onto a silica surface: a molecular dynamics simulation study. *Langmuir* **2010**, *26*, 9950−9955.

(13) Kubiak-Ossowska, K.; Jachimska, B.; Mulheran, P. A. How negatively charged proteins adsorb to negatively charged surfaces: a molecular dynamics study of BSA adsorption on silica. *J. Phys. Chem. B* **2016**, *120*, 10463−10468.

(14) Mathé, C.; Devineau, S.; Aude, J.-C.; Lagniel, G.; Chédin, S.; Legros, V.; Mathon, M.-H.; Renault, J.-P.; Pin, S.; Boulard, Y.; Labarre, J. Structural determinants for protein adsorption/non-adsorption to silica surface. *PloS one* **2013**, *8*, No. e81346.
(15) Benavidez, T. E.; Torrente, D.; Marucho, M.; Garcia, C. D. Adsorption of soft and hard proteins onto OTCEs under the influence of an external electric field. *Langmuir* **2015**, *31*, 2455−2462.
(16) Sobieściak, T. D.; Zielenkiewicz, P. Non-specific clustering of histidine tagged green fluorescent protein mediated by surface interactions: The collective effect in the protein-adsorption behaviour. *RSC Adv.* **2013**, *3*, 10479−10486.
(17) Wasserberg, D.; Cabanas-Danés, J.; Prangsma, J.; O'Mahony, S.; Cazade, P.-A.; Tromp, E.; Blum, C.; Thompson, D.; Huskens, J.; Subramaniam, V.; Jonkheijm, P. Controlling protein surface orientation by strategic placement of oligo-histidine tags. *ACS Nano* **2017**, *11*, 9068−9083.
(18) Gonzalez Solveyra, E.; Thompson, D. H.; Szleifer, I. Proteins Adsorbing onto Surface-Modified Nanoparticles: Effect of Surface Curvature, pH, and the Interplay of Polymers and Proteins Acid−Base Equilibrium. *Polymers* **2022**, *14*, 739.
(19) Collins, J.; Xirouchaki, C.; Palmer, R.; Heath, J.; Jones, C. Clusters for biology: immobilization of proteins by size-selected metal clusters. *Appl. Surf. Sci.* **2004**, *226*, 197−208.
(20) Prachayasittikul, V.; Isarankura Na Ayudhya, C.; Hilterhaus, L.; Hinz, A.; Tantimongcolwat, T.; Galla, H.-J. Interaction analysis of chimeric metal-binding green fluorescent protein and artificial solid-supported lipid membrane by quartz crystal microbalance and atomic force microscopy. *Biochemical and biophysical research communications* **2005**, *327*, 174−182.
(21) Wang, A.; Perera, Y. R.; Davidson, M. B.; Fitzkee, N. C. Electrostatic interactions and protein competition reveal a dynamic surface in gold nanoparticle−protein adsorption. *J. Phys. Chem. C* **2016**, *120*, 24231−24239.
(22) Adamczyk, Z. Protein adsorption: A quest for a universal mechanism. *Curr. Opin. Colloid Interface Sci.* **2019**, *41*, 50−65.
(23) Xu, J. X.; Alom, M. S.; Yadav, R.; Fitzkee, N. C. Predicting protein function and orientation on a gold nanoparticle surface using a residue-based affinity scale. *Nat. Commun.* **2022**, *13*, 7313.
(24) Li, Y.; Ogorzalek, T. L.; Wei, S.; Zhang, X.; Yang, P.; Jasensky, J.; Brooks, C. L.; Marsh, E. N. G.; Chen, Z. Effect of immobilization site on the orientation and activity of surface-tethered enzymes. *Phys. Chem. Chem. Phys.* **2018**, *20*, 1021−1029.
(25) Bateman, A.; et al. UniProt: a hub for protein information. *Nucleic acids research* **2015**, *43*, D204−D212.
(26) Jurrus, E.; Engel, D.; Star, K.; Monson, K.; Brandi, J.; Felberg, L. E.; Brookes, D. H.; Wilson, L.; Chen, J.; Gohara, D. W.; Nielsen, J. E.; Holst, M. J.; McCammon, J. A.; Baker, N. A.; et al. Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **2018**, *27*, 112−128. Cited By:830.
(27) McNaughton, B. R.; Chapman, A. M.; Blakeley, B. Detection of biopolymer interactions, cancer cells, and pathogens using split-supercharged GFP. US Patent 9,400,249, 2016.
(28) Price, D. J.; Brooks, C. L., III A modified TIP3P water potential for simulation with Ewald summation. *J. Chem. Phys.* **2004**, *121*, 10096−10103.
(29) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(30) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(31) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Schulten, K.; Chipot, C.; Tajkhorshid, E.; et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **2020**, *153*, 044130.
(32) Huang, J.; MacKerell, A. D., Jr CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of computational chemistry* **2013**, *34*, 2135−2145.
(33) Watowich, S. J.; Meyer, E. S.; Hagstrom, R.; Josephs, R. A stable, rapidly converging conjugate gradient method for energy minimization. *Journal of computational chemistry* **1988**, *9*, 650−661.
(34) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant pressure molecular dynamics simulation: the Langevin piston method. *J. Chem. Phys.* **1995**, *103*, 4613−4621.
(35) Andersen, H. C. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* **1983**, *52*, 24−34.
(36) 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*, 10089−10092.
(37) Heinig, M.; Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic acids research* **2004**, *32*, W500−W502.
(38) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of computational chemistry* **2008**, *29*, 1859−1865.

https://doi.org/10.1021/acsabm.3c00125
ACS Appl. Bio Mater. XXXX, XXX, XXX−XXX

Received: 5 July 2022 | Revised: 12 August 2022 | Accepted: 6 September 2022

DOI: 10.1002/prot.26427

RESEARCH ARTICLE

PROTEINS WILEY

Understanding the helical stability of charged peptides

Nitin Kumar Singh[1] | Manish Agarwal[2] | Mithun Radhakrishna[1,3]

[1]Discipline of Chemical Engineering, Indian Institute of Technology (IIT), Gandhinagar, Gujarat, India

[2]Computer Services Centre, Indian Institute of Technology (IIT), Delhi, India

[3]Center for Biomedical Engineering, Indian Institute of Technology (IIT), Gandhinagar, Gujarat, India

Correspondence
Mithun Radhakrishna, Discipline of Chemical Engineering, Indian Institute of Technology (IIT), Gandhinagar, Gujarat, India.
Email: mithunr@iitgn.ac.in

Funding information
Department of Science and Technology, Ministry of Science and Technology, Government of India; Ministry of Education

Abstract

Cationic helical peptides play a crucial role in applications such as anti-microbial and anticancer activity. The activity of these peptides directly correlates with their helicity. In this study, we have performed extensive all-atom molecular dynamics simulations of 25 Lysine–Leucine co-polypeptide sequences of varying charge density (λ) and patterns. Our findings showed that, an increase in the charge density on the peptide leads to a gradual decrease in the helicity up to a critical charge density λ_c. Beyond λ_c, a complete helix to coil transition was observed. The decrease in the helicity is correlated with the increased number of water molecules in first solvation shell, solvent-exposed surface area, and a higher value of the radius of gyration of the peptide.

KEYWORDS
charge density, helix, proteins, secondary structure

1 | INTRODUCTION

The spontaneous folding of the proteins into the secondary structure plays a crucial role in the biological functioning of the protein. Folding into secondary structures is driven by intermolecular interactions, such as hydrogen bonding, electrostatics, hydrophobic effect, and the van der Waals interaction between the amino acids of peptide chain. The secondary structures are generally classified into classes, such as, α-helix, G_{10}-helix, π-helix, β-sheets, turns, and coils based on backbone dihedral angles. Among these, the α-helices and β-sheets are the most prevalent secondary structures.[1] α-helices are motifs with exceptional folding/unfolding property and a rigid rod-framework, and they have been largely used as a building block in the design of molecular assemblies and small amphiphilic peptides.[2–4] These are the most regular forms in the secondary structure of protein, and play a cooperative role in protein folding.[5,6] Some of the well-known proteins, such as hemoglobin, myoglobin, keratin, human cytochrome c oxidase, GINS protein complex, RNA polymerases, human respiratory complex, and so forth possess a very high helical content.[7] The α-helix structure has shown its application in the various domains, such as membrane interactions, antimicrobial activity,[8–11] and anticancer activity.[12–14]Helical motifs have also been used as binding domains to high-value targets, such as BCL2 and HIV protein gp41.[15,16] Along with these, they also have a role in mediating protein interaction with other biomolecules, such as DNA and RNA.

Because of the vast applications and their abundance in the protein structure, α-helices have been studied extensively using both experimental and computational tools.[17–19] Most of the previous studies have focused on understanding the stability of short helical peptides of poly-L-alanine, poly-L-glutamic acid, and poly-L-lysine. Berger et al. studied the helical stability of poly-D,L-alanine using the deuterium exchange method and observed that the exchange rate was catalyzed by both hydrogen and hydroxyl ions.[20] Ferrati et al. have captured the helix-coil transition of alanine peptide that occurs at short time interval using nuclear magnetic resonance (NMR).[21] Levitt et al. performed computer simulations of 13 residue peptide to study the helical behavior in vacuum and solvent, and they observed an unfolding of the peptide in solvent medium.[22] Bixon et al. studied the solvent effect for the stability of the helical structures.[23] Their finding showed that the helical structure was dictated by a fine balance between peptide–water and intra-peptide hydrogen bonding.

Previous studies have focused on the design of α-helices comprising of hydrophobic residues or sequence of positive and negative residues.[24] In one of the very first study addressing the helical stability, Doty et al. studied the stability of the poly-L-Glutamic acid and reported that the helix to coil transition was observed when the peptide sequence contained at least 40% of the charged residues.[25]

Nishigami et al. studied the helical stability of co-polypeptide comprising of Lysine and Glutamic acid. They observed that the maximum helical propensity was found at the boundary of Lysine and Glutamic acid.[26] Further, many studies have focused on the patterning of charged residues and have shown that distance and ordering of positive and negative residues in the sequence plays an important role on the helical stability. These studies have attributed this to the change in the salt bridge forming propensity of the amino acids.[27–29] Furthermore, the presence of the charged residues also affects the transport properties and solubility due to the presence of dipole moment and net charge on the protein molecule.

α-helical anticancer peptides (α-ACPs) have a large proportion of positively charged residues like Lysine and Arginine in abundance, resulting in a net positive charge on helix. The interaction of the α-ACPs with the tumor cell membrane causes apoptosis and its application in the treatment of cancer is well established. The activity of α-ACPs is highly dependent on the helicity, and studies have shown that lower helicity of cationic anticancer peptides is associated with lowering HeLa activity.[30–32]

Many previous studies have shown that the formation of the hydrophobic packing plays a crucial role in folding of globular proteins[33–35] and residues, such as Methionine, Alanine, Leucine, Lysine (uncharged), and Glutamic acid (uncharged) are known to stabilize helix. The helical stability can be controlled by changing pH, ionic strength, temperature, solvent, etc. Although helices are known to contain charged residues, stabilizing the helix when the net charge is either positive or negative is very difficult due to the electrostatic repulsion between the like charge groups of amino acid side chain.[36–40]

In the current study, we have used all-atom molecular dynamics simulations to understand the helical stability of 25 de novo designed co-polypeptide sequences of Lysine and Leucine that contain net positive charges. Our study tries to address three broad questions (a) Can the helix forming ability of hydrophobic residues be leveraged to design peptide sequences containing like charged residues that fold into a helix? (b) If there exists a critical charge density that dictates the helix coil transition? (c) Does patterning of the hydrophobic and charged residues play a role in helical stability? Our findings show that the peptides containing like charges could be stabilized up to a critical charge density (λ_c) beyond which a helix to coil transition was observed. Further, our findings show that at a fixed charge density (λ), the helical stability of peptide was found to be independent of the patterning of the sequence. We believe that our study could also shed some light on the behavior of intrinsically disordered proteins (IDPs).[41,42]

2 | MODEL AND METHODS

All initial configuration of homopolypeptides (polyleucine, polylysine, polyaspartic acid, polyglutamic acid, and polyarginine) were constructed using the Chimera software.[43] Initial configuration of all peptides were designed to be completely α-helical. The co-polypeptides containing Leucine and Lysine of varying charge density and patterning were also constructed as above. Charge density (λ) is defined as $\lambda = \frac{\sum_1^N z_i}{N}$, where z_i is the net charge on the ith amino acid residue and N is the total number of amino acids in the peptide chain. For example, $\lambda = 0$ for homopolypeptide of Leucine (L_{20}) and $\lambda = 1$ for Lysine (K_{20}). The peptides were solvated in a box of $75 \times 75 \times 75$ Å^3 with TIP3P water model and the system was neutralized by adding counter ions using solvate and autoionize plugin of VMD respectively.[44] The forcefield parameters for the peptides were derived from CHARMM36[45] force field and the simulations were carried out using NAMD (version2.13-gpu)[46] simulation engine. The solvated system was minimized for 2500 steps by means of conjugate gradient method. Multiple independent simulations were performed for each of the studied sequence for 1 μs in isothermal isobaric (NPT) ensemble with a time step of 2 fs. The pressure was maintained at 1 bar using Langevin barostat method with decay period of 50 ps and damping time of 100 ps and temperature at 300 K was controlled using Langevin dynamics.[47] The bonds involving hydrogen atoms were constrained using the SHAKE algorithm.[48] Non-bonded interactions were calculated with a cut-off value of 12 Å with a switching distance of 10 Å and pairlistdist was set to 14 Å. The long-range electrostatic interactions were calculated using the particle-mesh Ewald summation (PME)[49] method with PMEGridSpacing of 1 Å. Multiple timestep parameters were applied for the electrostatics and short-range non-bonded interaction evaluation. The number of timesteps between full electrostatics evaluation was set to 2, that is, the PME calculations were done every other step and the nonbondedFreq was set to 1 to calculate the non-bonded forces every step. Scaled 1–4 parameter is applied to exclude pairs of bonded atoms from non-bonded interactions and 1–4 scaling is applied for the electrostatic interaction, required for scaled1–4 parameter. For each of the sequences studied, four to eight set of independent runs for 1 μs were carried out. All the thermodynamics averages were calculated using the last 200 ns of the 1 μs long simulation runs. Data was averaged over all independent runs. The analysis for the simulations were done using in-house Tcl and Python scripts.

2.1 | Calculation of fraction of helical content (α)

The secondary structure timeline of the peptides studied is analyzed using the STRIDE algorithm,[50] which is available as a plugin in VMD. The STRIDE algorithm takes into account the dihedral angles of the peptide backbone along with the hydrogen bonding distance to assign a secondary structure state to individual amino acid residue of the peptide. The fraction of helical content (α) is calculated by the formula $\alpha = \frac{n_{Helix}}{N}$, where n_{Helix} is the number of amino acids residues in the helical conformation and N is the total number of amino acids in the peptide chain.

3 | RESULTS AND DISCUSSION

3.1 | Simulations of homopolypeptides

In this study, we have considered 20 residue long homopolypeptide of Lysine (K_{20}), Arginine (R_{20}), Aspartic acid (D_{20}), Glutamic acid (E_{20}),

and Leucine (L_{20}). Since the main objective of the work was to understand the effect of charge density (λ) on the helical stability, the simulations of these homopolypeptide sequences serve as control set. These set of simulations provide us information at extreme values of λ, that is, $\lambda = 0$ for Leu_{20} and $\lambda = 1$ for Lys_{20}, Asp_{20}, Arg_{20}, and Glu_{20}.

Molecular dynamics simulations were performed in NPT ensemble at 300 K in explicit solvent. Initial set of all peptides were designed to be in α-helix. Figure 1 shows the average helical content of the homopolypeptides as a function of time. The data for the helicity are averaged over an interval of 5 ns. It is evident from Figure 1 that while L_{20} retained its helicity during course of simulation, all the charged homopolypeptides unfolded and lost their helicity. This can be attributed to the electrostatic repulsion between the like charged groups of the amino acid side chains leading to the loss of intra-peptide hydrogen bonds. The electrostatic energy of the charged homopolypeptides is plotted as a function of time in Figure S1. In case of all the four charged peptides, we see a decrease in electrostatic energy during the course of simulation, which also correlates to the unfolding of the helical polypeptide. This decrease in the energy is due to the reduced repulsion between the like charged groups of the polypeptide side chains. Table 1 shows the difference in electrostatic energy of the charged peptides in helical and unfolded state. Further, we also quantified the water–peptide and intra-peptide hydrogen bonds shown in Figure S2. It is very evident that for all the charged polypeptides, the loss of helical content results in the loss of intra-peptide hydrogen bonds in favor of water–peptide hydrogen bonds due to the larger surface area exposed. The pair correlation function between the like charge side chains calculated for the equilibrium runs were compared with the initial α-helical state of the peptide. We see the first peak is shifted to larger distance upon unfolding for all cases indicative of increased distance between side chains resulting in reduced repulsion (Figure S3).

To understand the effect of temperature on charged peptides, K_{20} was chosen as a model peptide due to the application of cationic peptides in various biological functions. Molecular dynamics simulations of K_{20} were performed up to temperatures of 260 K (it is to be noted that the melting point of the TIP3 water model is 145.6 K). Figure S4 shows the fraction of residues in α-helix for the K_{20} peptide at different temperatures, starting from 300 K to 260 K. Lowering temperature did not change the behavior of K_{20} and the peptide unfolded even at the lower temperatures.

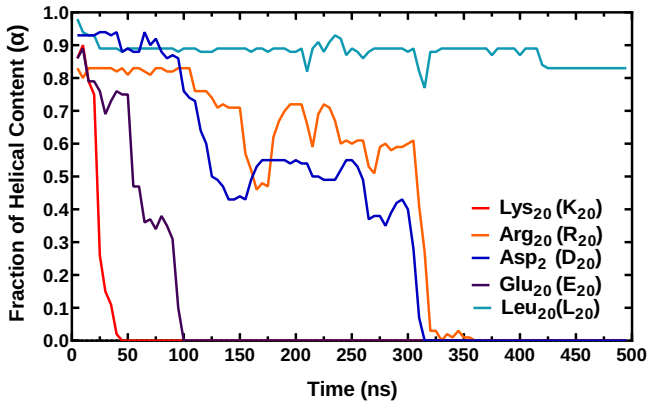


FIGURE 1 Fraction of helical content of homopolypeptides (K_{20}), (R_{20}), (D_{20}), (E_{20}), and (L_{20}) as function of time of simulations carried out in NPT ensemble at 300 K and 1 bar

TABLE 1 Difference in electrostatic energies of helical and unfolded state different charged homopolypeptides

S. No.	Peptide	$\Delta E = E_{folded} - E_{unfolded}$ (Kcal/mol)
1	K_{20}	210.47 ± 87.43
2	R_{20}	163.57 ± 86.90
3	D_{20}	443.07 ± 72.48
4	E_{20}	156.84 ± 74.69

3.2 | Simulations of patterned peptides

From Figure 1, it is evident that charged homopolypeptides lose helicity and hydrophobic polypeptides retain their helicity. However, hydrophobic polypeptides are insoluble in water while charged peptides are soluble. This motivated us to explore various co-polypeptides of Lysine (K) and Leucine (L), and study the effect of charge density (λ) and patterning on the helical stability of peptides. In the current study, we have designed 25 de novo co-polypeptide sequences of λ ranging from 0.16 to 0.83. For this study, we have considered peptides containing 24 amino acid residues. Few sequences considered were of length 21, 25, 27, and 28 to accommodate the patterning. We know that while polyleucine (L_{20}) is stable in the helical form, polylysine (K_{20}) unfolds into a coil. The sequences studied were based on the repeat units $(K_xL_y)_n$ where n is the number of repeat units and x and y are integers. While most of the repeat patterns could be accommodated with N (length of the peptide) = 24, for example, $x = 1$, $y = 1$, $n = 12$. Few sequences, for example (K_1L_4) or (K_4L_1) at $\lambda = 0.2$ and 0.8 respectively, had to have residues $n = 25$ to complete the pattern. While we understand that this is not an exhaustive list of all possible combinations of K and L co-polypeptides, it covers the entire range of λ from low to high. Figure 2 shows the pictorial representation of initial α-helical structure of various sequences studied and Table 2 provides the list of the sequences, their corresponding charge content and their length. Four to eight set of independent simulations up to 1 μs were carried out for each of these sequences and the averages and error bars were calculated based on last 200 ns for each independent run.

Figure 3 shows the averaged fractional helical content (α) and the radius of gyration of all the 25 sequences studied. As we move from left to right, there is an increase in λ with $(K_1L_5)_4$ being at 0.16 to $(K_5L_1)_4$ being at 0.83. At the outset, it is clear that as λ is increased, on average, there is a decrease in the helical content and an increase in the radius of gyration of peptide. Further, we also observe that for $\lambda \geq 0.67$ the helical content drops to zero. Based on the trends observed, the graph can be divided into two regions. The basis for

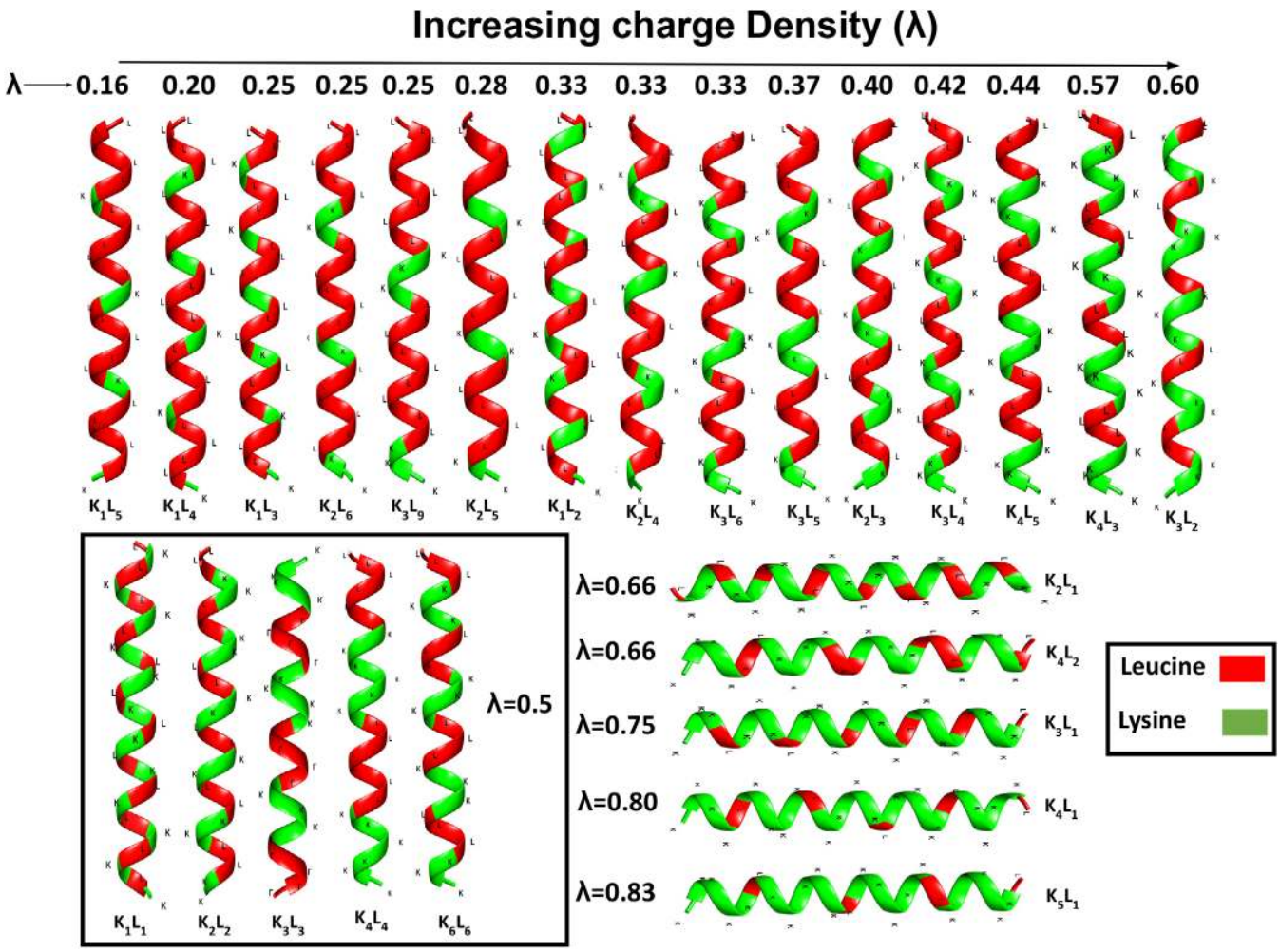


FIGURE 2 Pictorial representation of the initial structure of different co-polypeptide sequences of Lysine and Leucine

classification for these regions could be explained as follows. In region 1 ($\lambda \leq 0.60$), even though there is a gradual decrease in helicity with increasing λ, a fraction of residues still remain in the α-helical state. This is also validated by the fact that there is no dramatic change in radius of gyration of peptides. In Region 2, which comprises of sequences at high charge densities ($\lambda \geq 0.66$), we do not observe any α-helical content in the peptide with all the α-helical residues being transformed into coils. Further, the radius of gyration of peptides in region 2 is much higher than for the peptides in region 1 indicative of unfolding of α-helices. The interface between these two regions is the transition region where it is difficult to characterize the behavior of peptide due to the large fluctuations in the α-helical content. More precisely, the average helical content obtained from each of the independent runs did not converge with few runs displaying some helicity while others displaying no helicity. Furthermore, we also observed refolding of few residues in the coil state back to helix, making the characterization difficult. This behavior is reminiscent of IDPs which change their structure dynamically and previous studies have indicated such behavior among peptide having high charge content albeit for very specific sequences.[24,51] It is important to note that both the K and L residues contributed in almost in equal proportions toward the final helical content of the peptide as shown in Figure S6. It is also to be noted that the loss of the helical content was more pronounced at the ends of the peptide chain as compared with the middle, as shown in Figures S7 and S8.

Figure 4 shows the average helical content of different peptide sequences at a fixed charge density. As the charge density is increased, we see a decrease in helical content. However, at a fix charge content, we did not observe any appreciable change in the helical content for different patterns of sequences.

To quantify the above findings, we calculated the number of water molecules in the first solvation shell, the solvent accessible surface area (SASA), intra-peptide hydrogen bonds and peptide–water hydrogen bonds as shown in Figure 5A,B. For calculating the number of water molecules within the solvation shell, we considered all the water molecules that lie within a radius of 6 Å from the center of mass of the peptide chain. SASA was calculated using inbuilt plugin in VMD with a probe radius of 1.4 Å. The averaged Radius of gyration is calculated as $\langle Rg \rangle = \langle \sqrt{\frac{1}{N}\sum_1^N (r_i - r_{cm})^2} \rangle$, where r_i is the position of center of mass of ith residue, r_{cm} is the center of mass of entire protein, N is the total number of residues and $\langle\rangle$ represents the ensemble average. With an increase in the λ, we observed a direct correlation between

TABLE 2 Twenty-five co-polypeptide sequences of $(K_xL_y)_n$ made of Lysine(K)–Leucine(L) and their corresponding charge content and length

S. No.	x	y	n	N	Sequence	λ
1	1	5	4	24	$(K_1L_5)_4$	0.16
2	1	4	5	25	$(K_1L_4)_5$	0.20
3	1	3	6	24	$(K_1L_3)_6$	0.25
4	2	6	3	24	$(K_2L_6)_3$	0.25
5	3	9	2	24	$(K_3L_9)_2$	0.25
6	2	5	3	21	$(K_2L_5)_3$	0.28
7	1	2	8	24	$(K_1L_2)_8$	0.33
8	2	4	4	24	$(K_2L_4)_4$	0.33
9	3	6	3	27	$(K_3L_6)_3$	0.33
10	3	5	4	24	$(K_3L_5)_4$	0.37
11	2	3	5	25	$(K_2L_3)_5$	0.40
12	3	4	4	28	$(K_3L_4)_3$	0.42
13	4	5	3	27	$(K_4L_5)_3$	0.44
14	1	1	12	24	$(K_1L_1)_{12}$	0.50
15	2	2	6	24	$(K_2L_2)_6$	0.50
16	3	3	4	24	$(K_3L_3)_4$	0.50
17	4	4	3	24	$(K_4L_4)_3$	0.50
18	6	6	2	24	$(K_6L_6)_2$	0.50
19	4	3	4	28	$(K_4L_3)_4$	0.57
20	3	2	5	25	$(K_3L_2)_5$	0.60
21	2	1	8	24	$(K_2L_1)_8$	0.66
22	4	2	4	24	$(K_4L_2)_4$	0.66
23	3	1	6	24	$(K_3L_1)_6$	0.75
24	4	1	5	25	$(K_4L_1)_5$	0.80
25	5	1	4	24	$(K_5L_1)_4$	0.83

number of waters and the SASA. The peptide in its completely helical state has the least SASA and consequently very small number of water molecules are present in the first solvation shell. As the value of λ of the sequence is increased, the loss in the helical content translated to an increase in the coil content of the peptide leading to enhanced SASA and increased numbers of waters in first solvation shell. Figure 5B shows that the peptide loses its hydrogen bonds as the charge density increases in favor of water–peptide hydrogen bonds. Figure 6 shows the electrostatic energy and van der Waal (vdW) energy for all the 25 peptide sequences studied. The electrostatic energy of the completely helical peptide is also shown for reference. With an increase in on, we see an increase in both the electrostatic energy and vdW energy. On comparing the vdW energy of the equilibrated system to completely helical peptide, we do not see an appreciable change in energy up to $\lambda < 0.66$ beyond which the vdW energy of the equilibrated structure increases due to the complete unfolding of the helix. However, we observe a decrease in the electrostatic energy of the equilibrated peptide compared with the completely helical peptide, which can be attributed either to the partial or complete unfolding of peptide minimizing the repulsion between the like charged groups of the peptide.

4 | DISCUSSION

In this section, we would like to provide context to the above simulation findings vis-a-vis experiments. Our simulation findings on peptide sequences $(K_1L_2)_8$, $(K_2L_4)_4$, $(K_3L_6)_3$ with $\lambda = 0.33$ closely matches with the 22 residue sequence studied by Cornut et al.[52] in terms of the charge density. The helical content calculated from the simulations and the secondary structure obtained through circular dichroism

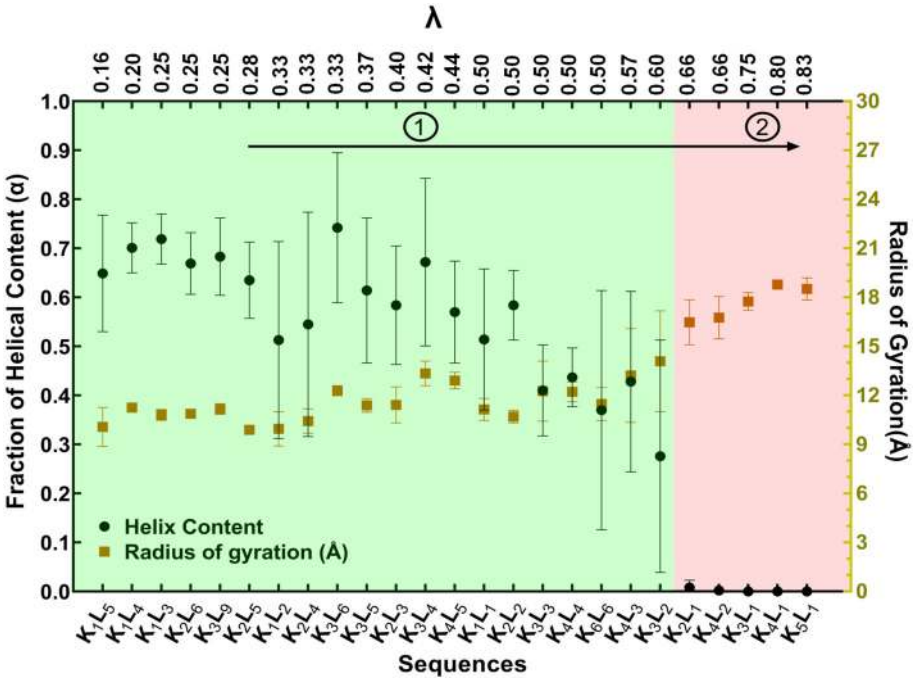


FIGURE 3 Fraction of α-helical content for different sequences. The fraction of the helical content decreased on increasing the charge density on the peptides

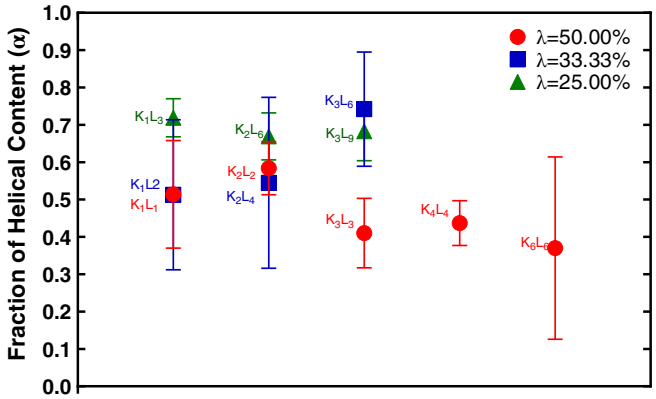


FIGURE 4 Fraction of helical content at fixed charge density λ of 50%, 33%, and 25%

studies are in qualitative agreement. Further Cornut et al also showed that by decreasing the length of the peptide at fixed charged density of $\lambda = 0.32$, there was a dramatic reduction in the helical content which is in line with the previous theoretical and computational studies.[53,54]

The current study emphasizes that the fraction of helical content is inversely proportional to the charge density. At the outset, this results may appear not in agreement with previous experimental findings by Apte et al.[55] and De Grado et al.[56] who observed that peptides with similar charge density formed different secondary structures like α-helices or β-sheets depending on the patterning of Lysine and Leucine residues in the sequence. The results obtained by Apte *et al* for sequences AcLKKLLKLLKKLLKL-OH and Ac-LKLKLKLKLKLKLKL-OH were upon adsorption of these peptides on a gold surface and not in bulk as in the present study. Further, in the work by De Grado *et al*

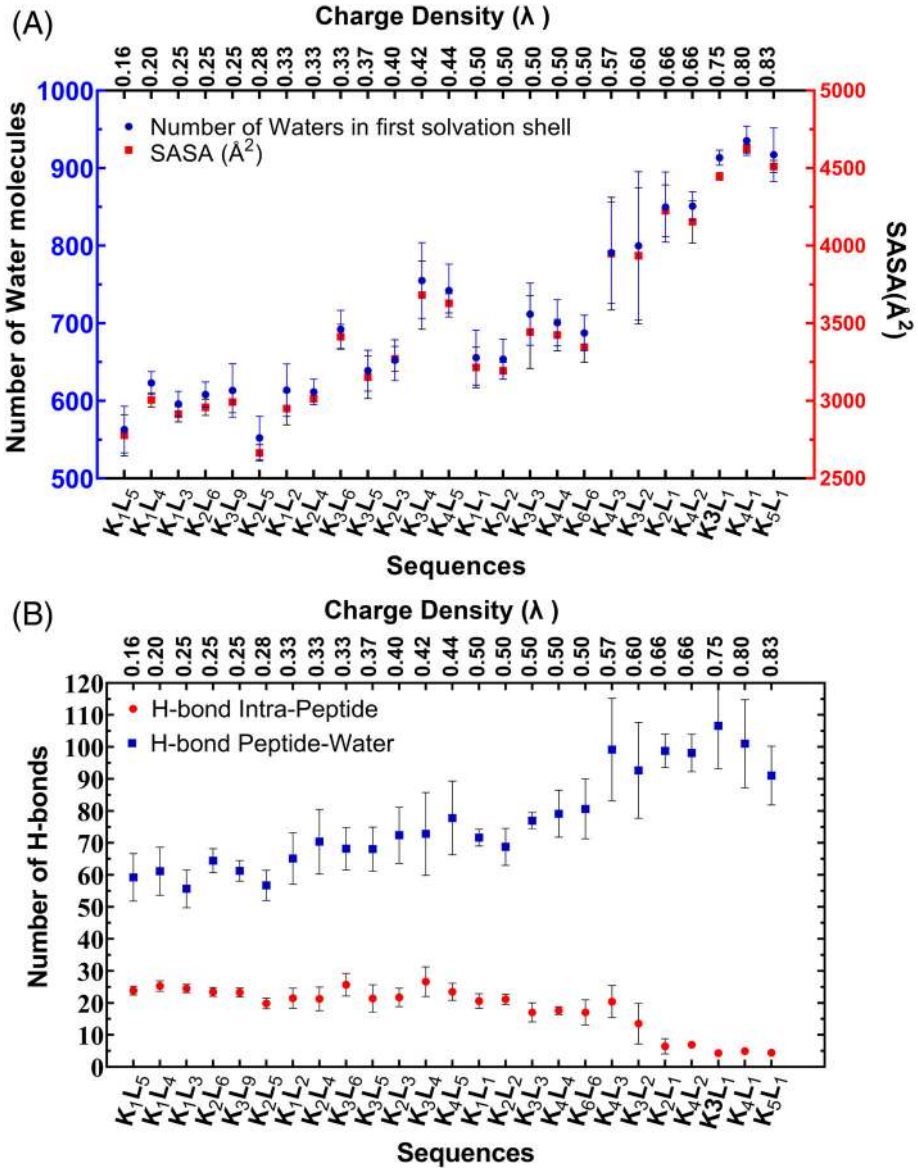


FIGURE 5 (A) Number of water molecules present in the first solvation shell of the protein at a distance within 6 Å and SASA for different peptide sequences. The number of water molecules in the vicinity of peptides increased on increasing the charge density on the sequences (B) Number of intra-peptide hydrogen bonds and water–peptide hydrogen bonds for different peptide sequences. The number of intra-peptide hydrogen bonds decreased whereas the water–peptide hydrogen bonds increased causing the loss of helical content of the peptide

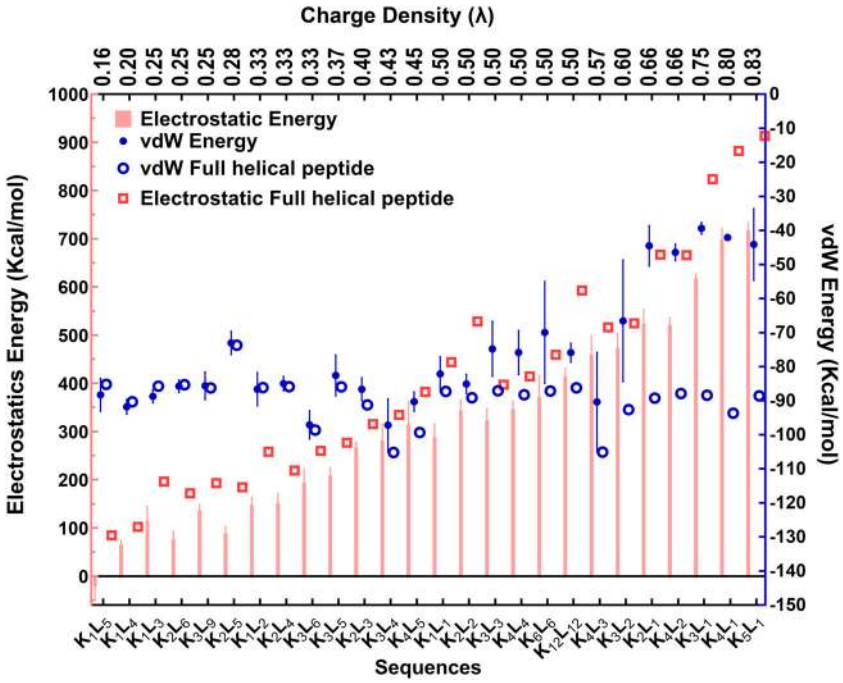


FIGURE 6 Energies for different sequences in NPT ensemble at 300 K and 1 atm. The electrostatic energy increased on increasing the charge content of the sequences and the vdW energy increased because of the unfolding of the helix

whose study predates Apte *et al*, it is clearly stated that the secondary structure observed in the aqueous solution were strongly dependent on the peptide and salt concentration, and also on the length of the peptide. The induction of secondary structures on these peptides was only observed at high peptide concentration where protein–protein interactions stabilize the secondary structures. The current set of simulations was carried out at infinite dilution (presence of one peptide) and in the absence of salt concentration. Therefore, it would not be possible to compare our current study with studies carried out at high salt and peptide concentrations. Further, it has to be noted that simulations carried out using other force fields may shift the value of the critical charge density (λ_c). However, we believe it would not qualitatively affect the critical findings in the manuscript.

5 | CONCLUSION

In summary, we have performed extensive molecular dynamics simulations on homopolypeptides and 25 co-polypeptide sequences of Lysine and Leucine to understand their helical stability. Charged homopolypeptides completely lost their helical content due to the increased electrostatic repulsion between side chains of amino acid groups in the helical state. Co-polypeptides of Lysine and Leucine showed a gradual decrease in the helical content with an increase in the charge density up to $\lambda = 0.6$, beyond which they completely transformed into coils. This unfolding of helices was also quantified by an increase in the radius of gyration of the peptide, increased SASA and increased number of water molecules in the first solvation shell. At a fixed charge density, we did not observe a significant change in the helical content for different peptide sequences. Further, the peptide electrostatic energy decreased compared with a completely helical peptide indicating that the unfolding is driven by the electrostatic repulsion among the charged amino acid residues.

ACKNOWLEDGMENTS

Mithun Radhakrishna wants to thank the Technology Mission Division (TMD), for funding and the High Performance Computing Facility at IIT Gandhinagar for support with computational resources. Manish Agarwal wants to thank the High Performance Computing Facility at IIT Delhi. Nitin Kumar Singh wants to thank the Ministry of Education, Govt of India for the doctoral fellowship.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26427.

DATA AVAILABILITY STATEMENT

All data is made available in the manuscript.

REFERENCES

1. Nelson DL, Lehninger AL, Cox MM. *Lehninger Principles of Biochemistry*. Macmillan; 2008.
2. Yakimov A, Afanaseva A, Khodorkovskiy M, Petukhov M. Design of stable α-helical peptides and thermostable proteins in biotechnology and biomedicine. *Acta Nat*. 2016;8:70-81.
3. Milletti F. Cell-penetrating peptides: Classes, origin, and current landscape. *Drug Discov Today*. 2012;17:850-860.
4. Fischer R, Fotin-Mleczek M, Hufnagel H, Brock R. Break on through to the other side—biophysics and cell biology shed light on cell-penetrating peptides. *Chembiochem*. 2005;6:2126-2142.

5. Fujiwara K, Toda H, Ikeguchi M. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. *BMC Struct Biol*. 2012;12:18.
6. Ihalainen JA, Paoli B, Muff S, et al. α-helix folding in the presence of structural constraints. *Proc Natl Acad Sci USA*. 2008;105:9588-9593.
7. Paoli M, Liddington R, Tame J, Wilkinson A, Dodson G. Crystal structure of T state haemoglobin with oxygen bound at all four haems. *J Mol Biol*. 1996;256:775-792.
8. Li G, Xia X, Long Y, Li J, Wu J, Zhu Y. Research progresses and applications of antimicrobial peptides. *Chin J Anim Nutr*. 2014;26:17-25.
9. Mihajlovic M, Lazaridis T. Charge distribution and imperfect amphipathicity affect pore formation by antimicrobial peptides. *Biochim Biophys Acta*. 2012;1818:1274-1283.
10. Huang Y-B, He L-Y, Jiang H-Y, Chen Y-X. Role of helicity on the anticancer mechanism of action of cationic-helical peptides. *Int J Mol Sci*. 2012;13:6849-6862.
11. Dan N, Setua S, Kashyap VK, et al. Antibody-drug conjugates for cancer therapy: chemistry to clinical implications. *Pharmaceuticals*. 2018;11:32.
12. Qiao X, Wang Y, Yu H. Progress in the mechanisms of anticancer peptides. *Sheng Wu Gong Cheng Xue Bao*. 2019;35:1391-1400.
13. Stefureac R, Long Y-T, Kraatz H-B, Howard P, Lee JS. Transport of α-helical peptides through α-hemolysin and aerolysin pores. *Biochemistry*. 2006;45:9172-9179.
14. Fernández-Vidal M, Jayasinghe S, Ladokhin AS, White SH. Folding amphipathic helices into membranes: amphiphilicity trumps hydrophobicity. *J Mol Biol*. 2007;370:459-470.
15. de Araujo A, Lim J, Wu K-C, et al. Bicyclic helical peptides as dual inhibitors selective for Bcl2A1 and Mcl-1 proteins. *J Med Chem*. 2018;61:2962-2972.
16. Eckert DM, Kim PS. Design of potent inhibitors of HIV-1 entry from the gp41 N-peptide region. *Proc Natl Acad Sci USA*. 2001;98:11187-11192.
17. Finkelstein A, Badretdinov AY, Ptitsyn O. Physical reasons for secondary structure stability: $_{\alpha}$-helices in short peptides. *Proteins Struct Funct Bioinform*. 1991;10:287-299.
18. Doig AJ. Recent advances in helix–coil theory. *Biophys Chem*. 2002;101:281-293.
19. Rohl CA, Baldwin RL. *Methods in enzymology*. Vol 295. Elsevier; 1998:1-26.
20. Berger A, Linderstrøm-Lang K. Deuterium exchange of poly-DL-alanine in aqueous solution. *Arch Biochem Biophys*. 1957;69:106-118.
21. Ferretti JA, Paolillo L. Nuclear magnetic resonance investigation of the helix to random coil transformation in poly-α-amino acids. I poly-L-alanine. *Biopolymers*. 1969;7:155-171.
22. Daggett V, Levitt M. Molecular dynamics simulations of helix denaturation. *J Mol Biol*. 1992;223:1121-1138.
23. Bixon M, Lifson S. Solvent effects on the helix–coil transition in polypeptides. *Biopolymers*. 1966;4:815-821.
24. Dalgicdir C, Globisch C, Peter C, Sayar M. Tipping the scale from disorder to alpha-helix: folding of amphiphilic peptides in the presence of macroscopic and molecular interfaces. *PLoS Comput Biol*. 2015;11:e1004328.
25. Doty P, Wada A, Yang JT, Blout ER. Polypeptides. VIII. Molecular configurations of poly-L-glutamic acid in water-dioxane solution. *J Polym Sci*. 1957;23:851-861.
26. Nishigami H, Kang J, Terada R-I, Kino H, Yamasaki K, Tateno M. Is it possible for short peptide composed of positively-and negatively-charged "hydrophilic" amino acid residue-clusters to form metastable "hydrophobic" packing? *Phys Chem Chem Phys*. 2019;21:9683-9693.
27. Marqusee S, Baldwin RL. Helix stabilization by Glu- ...Lys+ salt bridges in short peptides of de novo design. *Proc Natl Acad Sci USA*. 1987;84:8898-8902.
28. Meuzelaar H, Vreede J, Woutersen S. Influence of Glu/Arg, asp/Arg, and Glu/Lys salt bridges on α-helical stability and folding kinetics. *Biophys J*. 2016;110:2328-2341.
29. Maxfield FR, Scheraga HA. The effect of neighboring charges on the helix forming ability of charged amino acids in proteins. *Macromolecules*. 1975;8:491-493.
30. Xie M, Liu D, Yang Y. Anti-cancer peptides: classification, mechanism of action, reconstruction and modification. *Open Biol*. 2020;10:200004.
31. Boohaker RJ, Lee MW, Vishnubhotla P, Perez JM, Khaled AR. The use of therapeutic peptides to target and to kill cancer cells. *Curr Med Chem*. 2012;19:3794-3804.
32. Huang Y, Feng Q, Yan Q, Hao X, Chen Y. Alpha-helical cationic anticancer peptides: a promising candidate for novel anticancer drugs. *Mini Rev Med Chem*. 2015;15:73-81.
33. Huang C-Y, Klemke JW, Getahun Z, DeGrado WF, Gai F. Temperature-dependent helix- coil transition of an alanine based peptide. *J Am Chem Soc*. 2001;123:9235-9238.
34. Aftabuddin M, Kundu S. Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophys J*. 2007;93:225-231.
35. Moret M, Santana M, Zebende G, Pascutti P. Self-similarity and protein compactness. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2009;80:041908.
36. Jiang Z, Vasil AI, Hale JD, Hancock RE, Vasil ML, Hodges RS. Effects of net charge and the number of positively charged residues on the biological activity of amphipathic alpha-helical cationic antimicrobial peptides. *Biopolymers*. 2008;90:369-383.
37. Chou PY, Scheraga HA. Calorimetric measurement of enthalpy change in the isothermal helix–coil transition of poly-L-lysine in aqueous solution. *Biopolymers*. 1971;10:657-680.
38. Epand RF, Scheraga HA. The helix–coil transition of poly-L-lysine in methanol–water solvent mixtures. *Biopolymers*. 1968;6:1383-1386.
39. Nakazawa T, Ban S, Okuda Y, Masuya M, Mitsutake A, Okamoto Y. A pH-dependent variation in alpha-helix structure of the S-peptide of ribonuclease a studied by Monte Carlo simulated annealing. *Biopolymers*. 2002;63:273-279.
40. Sacquin-Mora S. Fold and flexibility: what can proteins' mechanical properties tell us about their folding nucleus? *J R Soc Interface*. 2015;12:20150876.
41. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*. 2000;41:415-427.
42. Holehouse AS, Ahad J, Das RK, Pappu RV. CIDER: classification of intrinsically disordered ensemble regions. *Biophys J*. 2015;108:228a.
43. Pettersen EF, Goddard TD, Huang CC, et al. UCSF chimera—a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25:1605-1612.
44. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph*. 1996;14:33-38.
45. Huang J, MacKerell AD Jr. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem*. 2013;34:2135-2145.
46. Phillips JC, Hardy DJ, Maia JD, et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys*. 2020;153:044130.
47. Feller SE, Zhang Y, Pastor RW, Brooks BR. Constant pressure molecular dynamics simulation: the Langevin piston method. *J Chem Phys*. 1995;103:4613-4621.
48. Andersen HC. Rattle: a "velocity" version of the shake algorithm for molecular dynamics calculations. *J Comput Phys*. 1983;52:24-34.
49. Darden T, York D, Pedersen L. Particle mesh Ewald: an N log (N) method for Ewald sums in large systems. *J Chem Phys*. 1993;98:10089-10092.

50. Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res*. 2004;32:W500-W502.
51. Sung S-S. Folding simulations of alanine-based peptides with lysine residues. *Biophys J*. 1995;68:826-834.
52. Cornut I, Büttner K, Dasseux J-L, Dufourcq J. The amphipathic alpha-helix concept: application to the de novo design of ideally amphipathic Leu, Lys peptides with hemolytic activity higher than that of melittin. *FEBS Lett*. 1994;349:29-33.
53. Ghosh K, Dill KA. Computing protein stabilities from their chain lengths. *Proc Natl Acad Sci USA*. 2009;106:10649-10654.
54. Zimm BH, Bragg J. Theory of the phase transition between helix and random coil in polypeptide chains. *J Chem Phys*. 1959;31:526-535.
55. Apte JS, Gamble LJ, Castner DG, Campbell CT. Kinetics of leucine-lysine peptide adsorption and desorption at-CH3 and-COOH terminated alkylthiolate monolayers. *Biointerphases*. 2010;5:97-104.
56. DeGrado W, Lear J. Induction of peptide conformation at apolar water interfaces. 1. A study with model peptides of defined hydrophobic periodicity. *J Am Chem Soc*. 1985;107:7684-7689.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Singh NK, Agarwal M, Radhakrishna M. Understanding the helical stability of charged peptides. *Proteins*. 2022;1-9. doi:10.1002/prot.26427