Accepted: 26 October 2024

DOI: 10.1002/pro.5219

RESEARCH ARTICLE

Identifying the minimal sets of distance restraints for FRET-assisted protein structural modeling

Zhuoyi Liu[1,2] | Alex T. Grigas[2,3] | Jacob Sumner[2,3] | Edward Knab[4] | Caitlin M. Davis[4] | Corey S. O'Hern[1,2,3,5,6]

[1]Department of Mechanical Engineering and Materials Science, Yale University, New Haven, Connecticut, USA

[2]Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, Connecticut, USA

[3]Graduate Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA

[4]Department of Chemistry, Yale University, New Haven, Connecticut, USA

[5]Department of Physics, Yale University, New Haven, Connecticut, USA

[6]Department of Applied Physics, Yale University, New Haven, Connecticut, USA

Correspondence

Corey S. O'Hern, Department of Mechanical Engineering and Materials Science, Yale University, New Haven, CT 06520, USA.
Email: corey.ohern@yale.edu

Funding information

National Institutes of Health, Grant/Award Numbers: R35GM151146, T32GM145452, T15LM007056-37

Abstract

Proteins naturally occur in crowded cellular environments and interact with other proteins, nucleic acids, and organelles. Since most previous experimental protein structure determination techniques require that proteins occur in idealized, non-physiological environments, the effects of realistic cellular environments on protein structure are largely unexplored. Recently, Förster resonance energy transfer (FRET) has been shown to be an effective experimental method for investigating protein structure in vivo. Inter-residue distances measured in vivo can be incorporated as restraints in molecular dynamics (MD) simulations to model protein structural dynamics in vivo. Since most FRET studies only obtain inter-residue separations for a small number of amino acid pairs, it is important to determine the minimum number of restraints in the MD simulations that are required to achieve a given root-mean-square deviation (RMSD) from the experimental structural ensemble. Further, what is the optimal method for selecting these inter-residue restraints? Here, we implement several methods for selecting the most important FRET pairs and determine the number of pairs N_r that are needed to induce conformational changes in proteins between two experimentally determined structures. We find that enforcing only a small fraction of restraints, $N_r/N \lesssim 0.08$, where N is the number of amino acids, can induce the conformational changes. These results establish the efficacy of FRET-assisted MD simulations for atomic scale structural modeling of proteins in vivo.

KEYWORDS

cellular crowding, FRET experiments, in vivo protein structure, molecular dynamics simulations

1 | INTRODUCTION

Knowing the three-dimensional structure of proteins enables us to understand the biophysical mechanisms that control protein function, protein–protein interactions, and cell signaling. Nearly all protein structures that have been determined experimentally to date have been characterized under idealized, non-physiological conditions. For example, proteins have been crystallized into non-native, solid phases for x-ray scattering experiments (Smyth and Martin 2000), and proteins have been dissolved into dilute, non-physiological buffers for NMR spectroscopy or cryo-electron microscopy (Carroni and Saibil 2016; Williamson et al. 1985). However, proteins carry out their functions in cellular environments that are significantly different from these in vitro conditions. The cellular environment is crowded with a non-solvent packing fraction of $0.3-0.4$ that includes nucleic acids, carbohydrates, lipids, organelles, and other components (Ellis 2001; Fulton 1982; Zimmerman and Trach 1991). This environment impacts the physical properties of proteins, including the protein's radius of gyration, melting temperature, and rotational diffusion coefficient (Davis et al. 2020; Ebbinghaus et al. 2010; Leeb et al. 2020; Wang et al. 2018). Currently, there are over 200,000 In vitro protein structures deposited in the Protein Data Bank (PDB) (Berman et al. 2000), as well as more than 10^6 computational models predicted by AlphaFold and RosettaFold (Humphreys et al. 2021; Jumper et al. 2021). To date, in vivo structures have been obtained for only three proteins (TTHA1718, GB1, and ubiquitin) using in-cell NMR (Gerez et al. 2022; Sakakibara et al. 2009; Tanaka et al. 2019). In-cell NMR is not widely used since it is difficult to distinguish the isotopic labeled target protein from its environment, which leads to a low signal-to-noise ratio and sensitivity (Ikeya et al. 2019; Luchinat and Banci 2023; Serber and Dötsch 2001). Another experimental method for solving protein structure in vivo is cryo-electron tomography (Cheng et al. 2023); however, this technique requires proteins to be confined to a thickness less than 500 nm, which limits this method to only a subset of cell types and membrane-associated proteins (Dunstone and de Marco 2017; Hylton and Swulius 2021). In addition, there have been numerous computational studies of proteins in vivo. All-atom molecular dynamics (MD) simulations have investigated protein structure in the presence of nearly all components of the cell cytoplasm (Rickard et al. 2020; Stevens et al. 2023). However, current force fields have been calibrated to in vitro protein structures and we do not have accurate potentials for interactions between proteins and nucleic acids, ribosomes, and organelles (Love et al. 2023; Tucker et al. 2022). Thus, we do not yet have a quantitative, atomistic-level understanding of protein structure in cells.

Förster resonance energy transfer (FRET) can be used to determine the separations between donor and acceptor chromophores attached to a pair of amino acids in a given protein by measuring their energy transfer efficiency. The distribution of separations between the two amino acids can then be deduced by calculating the configuration space volumes sampled by the donor and acceptor chromophores (Dimura et al. 2016; Kalinin et al. 2012; Klose et al. 2021). FRET pair labeling is highly specific and offers high accuracy for the inter-residue separations with uncertainties less than 2–4 Å (Agam et al. 2023; Hellenkamp et al. 2018). Additionally, FRET-labeled proteins can be directly expressed or injected into cells, enabling single molecule measurements of protein structure, dynamics, and stability in vivo (Davis and Gruebele 2018; Ebbinghaus et al. 2010; Feng et al. 2019). However, most FRET experimental studies only obtain inter-residue separations for a small number of amino acid pairs in a given protein (Davis and Gruebele 2018; Ebbinghaus et al. 2010; Wang et al. 2018). To investigate the atomic scale structure of proteins in vivo, inter-residue separations obtained from FRET experiments can be incorporated into molecular dynamics (MD) simulations. Current all-atom MD simulations of proteins using in vitro solution conditions have been shown to sample in vitro protein structures obtained from x-ray crystallography and NMR spectroscopy (Lindorff-Larsen et al. 2011, 2012). By including a sufficient number of inter-residue restraints from FRET studies of proteins in vivo into the MD force fields that were developed for in vitro solution conditions, it may be possible to sample the in vivo structural ensemble of proteins.

Several key questions must be answered before FRET-assisted structural modeling of protein structure in vivo can be employed. In particular, given the large number of possible FRET pairs, what is the minimum number of amino acid pairs for which we need distance restraints to achieve a given accuracy for the root-mean-square deviations (RMSD) between the C_α positions in the restrained simulations and those in the experimental structures? However, we do not yet have access to high-resolution in vivo protein structures. Thus, we will first develop the methodology for carrying out restrained MD simulations for protein structural modeling using proteins that are found in multiple conformational states in vitro. The hypothesis is that the method that we use to induce conformational changes in vitro can also be used to study conformational changes in in vivo environments. The initial conformational state will be metastable in the MD force field without restraints, and we will induce a conformational change in the protein by incorporating restraints between amino acid pairs that are satisfied in the target state. To our knowledge, there have only been a few studies aimed at identifying the most important restraints for efficiently moving to the conformational ensemble of the target state (Dimura et al. 2020). For example, currently we do not know the minimal number of restraints and which restraints are necessary to achieve a given RMSD in the C_α positions from the target state, and how random selection of given number of restraints compares to other methods for selecting restraints. To connect the in vitro results to those for in vivo conditions, we will also study one of the few proteins whose structure has been solved in vivo using in-cell NMR spectroscopy.

Here, we select four proteins that each can take on two, distinct conformational states to develop the restrained MD simulation methodology: T4 lysozyme (172L/1L69), phosphoglycerate kinase (2XE6/2Y3I), adenylate kinase (4AKE/1AKE), and tick carboxypeptidase inhibitor (1ZLI/2JTO), where the first and second PDBIDs indicate the initial and target structures, respectively. We initialize the MD simulations with the crystal structure of the initial state, add a given number of C_α distance restraints, and calculate the C_α RMSD relative to the target structure. For the first three proteins (T4 lysozyme [T4L], phosphoglycerate kinase [PGK], and adenylate kinase [AK]) that are metastable in the force field, we test four methods (normal mode analysis, the largest C_α separation method, the largest change in pairwise separation method, and linear discriminant analysis) for selecting the restraints and compare the C_α RMSD to the target structure to that obtained from random selection of the restraints. We also vary the number of restraints to determine the minimum number of restraints needed to achieve a given C_α RMSD from the target. Two of the methods (normal mode analysis and the largest C_α separation method) do not use information about the target structure, whereas the other two methods (the largest change in pairwise separation method and linear discriminant analysis) compare the initial and target structures to identify the most important restraints. We find that for T4L, PGK, and AK, which take on two distinct conformational states in vitro, we can induce the conformational changes using only a small fraction of restraints, $N_r/N \lesssim 0.02$, where N is the number of amino acids in the protein. For the proteins that we considered, this result corresponds to 1–5 restraints, which is a number that can readily be achieved in FRET experiments. In addition, we studied one of the few proteins that has been characterized using in-cell NMR spectroscopy: the B1 domain of protein G (GB1). We need a slightly larger fraction of restraints, $N_r/N \sim 0.08$ (or 5 restraints), to change the protein conformation from the initial in vitro structure (2N9K) to the in-cell NMR structure (7QTS). Using our methods to induce conformational changes from the bound to the unbound conformations of Tick Carboxypeptidase Inhibitor (TCI) and from the in vivo to the in vitro structures of GB1, we show that the fraction of restraints required to induce conformational changes depends on the stability of the target state in the force field. In general, the largest change in pairwise separation method, which has information about the target structure, yields the lowest values for the C_α RMSD for a given N_r. If the restraint selection method does not have information about the target structure, the C_α RMSD is still lower than that for random selection. These results establish the feasibility of FRET-assisted structural modeling of proteins in vivo using restrained MD simulations, but also emphasize the need for improved force fields for MD simulations of proteins in vivo.

2 | MATERIALS AND METHODS

2.1 | Selected proteins

We identified three proteins from the Protein Data Bank (PDB) that can be crystallized into two distinct conformational states in vitro and can be used to develop the restrained MD simulation methodology: T4 lysozyme (T4L), phosphoglycerate kinase (PGK), and adenylate kinase (AK) (see Table 1) (Lallemand et al. 2011; Müller et al. 1996; Müller and Schulz 1992; Zerrad et al. 2011; Zhang et al. 1992, 1995). These proteins have been studied extensively in the context of protein conformational changes (Flores et al. 2006). While the selected targets are in vitro structures, recent studies suggest that the differences between the in vivo and in vitro structures for PGK are largely caused by the hinge motion of the two domains that occurs between the initial and target in vitro structures (Davis et al. 2020; Davis and Gruebele 2018). Previous studies have also suggested that the crowded cellular environment can stabilize the target in vitro structure for AK (Li et al. 2014). For the restrained MD simulations, we require that the initial structure is stable and the target structure is at least metastable over long time scales in MD simulations, which ensures that the unrestrained dynamics does not induce the transition from the initial state to the target (see Figure S1). We also consider one protein that has both an in vitro and in-cell NMR structure, the B1 domain of protein G (GB1) (Gerez et al. 2022; Ikeya et al. 2016). We use the first model from the in vitro NMR bundle as the initial structure and the first model of the in-cell NMR bundle as the target structure (see Figure S6). We find similar results for the restrained and unrestrained MD simulations when we initialize the system using the other NMR models. To investigate the dependence of the restraint selection method on the stability of the target structure, we study heterodimer, tick carboxypeptidase inhibitor (TCI) (Arolas et al. 2005a; Pantoja-Uceda et al. 2008), which is metastable in its bound conformation, but unstable in its unbound conformation (see Figure S7).

TABLE 1 We list the selected proteins including their PDBIDs for the initial and target structures, number of amino acids N, the C_α RMSD (Equation (2)) between the initial and target structures, and the melting temperature T_m.

Protein	Initial structure	Target structure	N	C_α RMSD (Å)	T_m (K)
T4 lysozyme (T4L)	172L	1L69	162	3.95	344 (Baase et al. 2010)
Phosphoglycerate kinase (PGK)	2XE6	2X15	413	4.08	327 (Fiorillo et al. 2018)
Adenylate kinase (AK)	4AKE	1AKE	214	5.96	333 (Chang et al. 2021)
B1 domain of protein G (GB1)	2N9K	7QTS	57	2.86	354 (Campos-Olivas et al. 2002)
Tick carboxypeptidase inhibitor (TCI)	1ZLI	2JTO	77	3.82	>343 (Arolas et al. 2005b)

Note: For GB1, the C_α RMSD is calculated between the first model of the in vitro NMR bundle and the first model of the in-cell NMR bundle.

2.2 | C_α RMSD

To compare the protein structures from the restrained MD simulations (i.e., structure S_i) and the target structure (i.e., structure S_j), we define the C_α separation vector for the βth amino acid,

$$\vec{\Delta}(S_i, S_j; \beta) = \vec{r}_{i,\beta} - \vec{r}_{j,\beta}, \tag{1}$$

where $\vec{r}_{i,\beta}$ is the position of the C_α atom on amino acid β in structure S_i. We define the root-mean-square deviation in the C_α positions between two structures S_i and S_j as

$$\mathrm{RMSD}(S_i, S_j) = \sqrt{\frac{1}{N}\sum_{\beta=1}^{N} \Delta^2(S_i, S_j, \beta)}. \tag{2}$$

When comparing two structures S_i and S_j, we typically align them (i.e., rotate one of them) to achieve the minimum value of the C_α $\mathrm{RMSD}(S_i, S_j)$ for a given S_i and S_j. We can also calculate the C_α RMSD between the restrained and target structures averaged over an ensemble of restrained structures for each amino acid β,

$$\mathrm{RMSD}(\{S_i\}, S_j, \beta) = \sqrt{\frac{1}{N_s}\sum_{i=1}^{N_s} \Delta^2(S_i, S_j, \beta)}, \tag{3}$$

where N_s is the number of protein structures in the restrained ensemble $\{S_i\}$ and S_j represents the target structure.

2.3 | Restraint selection methods

Below, we describe the methods that we employ for selecting the C_α distance restraints. Linear spring restraints will be added to MD simulations of the initial protein structure, where the rest lengths are obtained from the target structure, as discussed below. Three of the restraint selection methods, random selection, normal mode analysis, and the largest C_α separation method, do not use information about the target structure. Two additional methods, the largest change in pairwise separation method and linear discriminant analysis, compare the initial *and* target structure to identify the most effective restraints.

2.3.1 | Random selection

In a protein with N amino acids, there are $N_p = N(N-1)/2$ distinct amino acid pairs. As shown in Figure 1a, the pair separations between C_α atoms can be represented using a symmetric distance matrix $R_{\beta\delta}$, where β, $\delta = 1, \ldots, N$. To establish a baseline for the performance of the restrained MD simulations, we will first consider random selection of the restraints. In this approach, we exclude amino acid pairs that are too close ($R_{\beta\delta} < R_g$, where R_g is the radius of gyration), and pairs for which at least one amino acid is buried with relative solvent accessible surface area $\mathrm{rSASA} < 0.1$ since such pairs will preclude FRET measurements (Grigas et al. 2022; Richards 1974), which reduces the pool of restraints to approximately $N_p/3$ amino acid pairs. For each number of restraints $N_r \ll N_p$ we consider, we select 100 sets of N_r restraints randomly from the pool of allowed pairs.

2.3.2 | Normal mode analysis

For this method, we assume that the normal modes of vibration of the initial structure provide information about how the protein transitions from the initial to the target structure. We follow the methodology used in other recent work aimed at selecting the most effective FRET pairs for restrained MD simulations of proteins (Dimura et al. 2020). First, the method constructs a coarse-grained elastic network from the atomic positions of the initial protein structure and calculates the normal

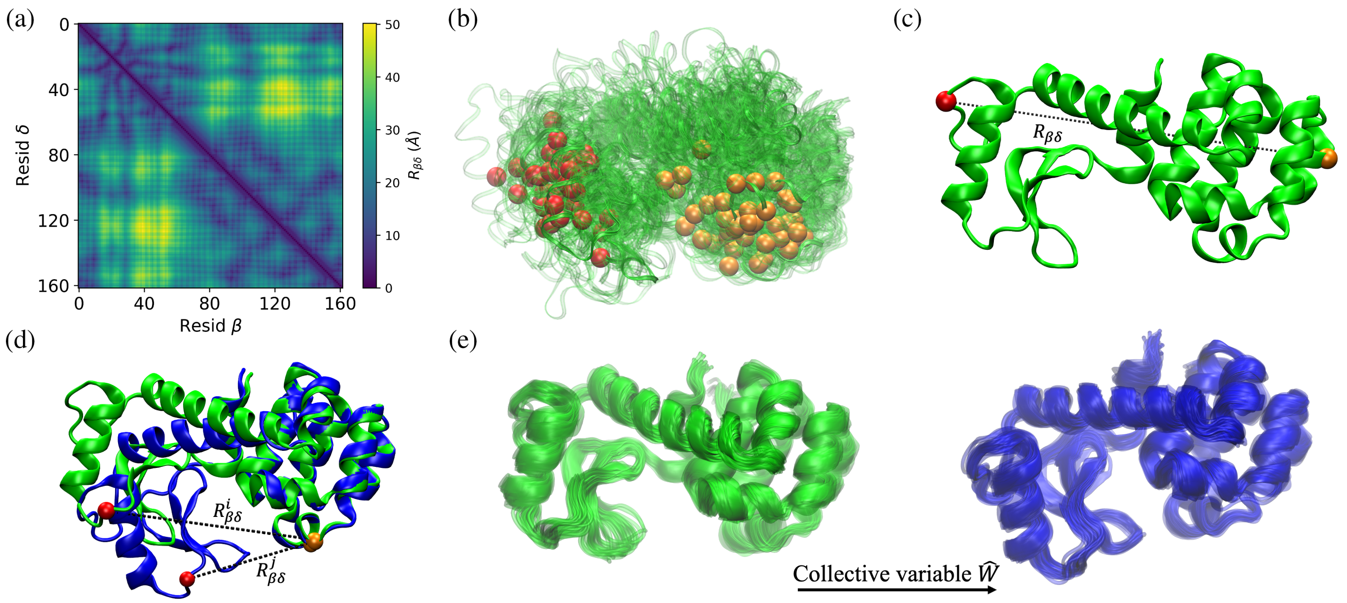

FIGURE 1 Illustration of several distance restraint selection methods, using T4L as the example protein. (a) The symmetric distance matrix $R_{\beta\delta}$, where the color gradient from dark to light indicates increasing C_α separations between amino acids β and δ. (b) For the normal mode analysis method, we generate an ensemble of 10^4 structures (100 structures are shown in green) that have been displaced from the initial structure along a random superposition of normal modes corresponding to the 100 lowest frequencies. A single amino acid pair that minimizes the C_α RMSD among structures in the ensemble is highlighted in red and orange in each structure. (c) We identify the amino acid pairs with the largest C_α separations $R_{\beta\delta}$ in the initial (green) structure. The pair with the maximum $R_{\beta\delta}$ is highlighted in red and orange. (d) We can also select restraints by identifying the largest changes in the pairwise C_α separations between the initial (green) and target (blue) structures. The C_α atoms of the pair with the maximum $\Delta R_{\beta\delta}(S_i,S_j)$ are shown in red and orange. (e) 20 conformations from unrestrained MD simulations starting from the initial and target structures are shown in green and blue, respectively. Using linear discriminant analysis, we identify the collective variable $\widehat{W}$ that maximizes the cross-correlation of the two ensembles.

modes of the network (Ahmed and Gohlke 2006). Then, the initial structure is displaced by a linear combination of the 10 lowest frequency modes with amplitudes that are inversely proportional to the frequency and have proportionality constants $-1<\xi<1$ that are chosen randomly. The normal modes are then recalculated on the displaced structure and the structure is perturbed again using a linear combination of the 10 lowest frequency modes with amplitudes chosen as before. This process of successive displacements along the lowest frequency normal modes is continued until 10^3 structures are obtained (Krüger et al. 2012) and then repeated 10 times with independently generated random mode amplitudes to yield a total of $N=10^4$ structures. 100 of the structures for T4L are shown in Figure 1b.

To identify the amino acid pairs that should be restrained, we seek to minimize the RMSD of the positions of the C_α atoms (Equation (2)) among the protein structures in the ensemble, where the weights for each structure in the ensemble are controlled by the size of the fluctuations in the amino acid pair separations among structures. Deviations in the pair separations between two structures S_i and S_j can be quantified using

$$\chi^2(S_i,S_j)=\sum_{\{(\beta,\delta)\}}\left(\frac{\Delta R_{\beta\delta}(S_i,S_j)}{R^i_{\beta\delta}+R^j_{\beta\delta}}\right)^2, \tag{4}$$

where $\Delta R_{\beta\delta}(S_i,S_j)=|R^i_{\beta\delta}-R^j_{\beta\delta}|$, $\{(\beta,\delta)\}$ is a given set of amino acid pairs, and $R^i_{\beta\delta}$ is the distance between C_α atoms on amino acids β and δ on structure S_i. To estimate the probability of observing a mean-square deviation in the pair separations larger than $\chi^2(S_i,S_j)$, we calculate

$$P(S_i,S_j)=\int_{\chi^2(S_i,S_j)}^{\infty} f(N_m,\chi^2)d\chi^2, \tag{5}$$

where N_m is the number of amino acid pairs in the set $\{(\beta,\delta)\}$ and $f(N_m,\chi^2)$ is the chi-squared distribution with N_m degrees of freedom. To quantify the average C_α RMSD over the ensemble, we calculate

$$\langle\mathrm{RMSD}\rangle=\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{N}P(S_i,S_j)\mathrm{RMSD}(S_i,S_j)}{\sum_{j=1}^{N}P(S_i,S_j)}. \tag{6}$$

To minimize $\langle\mathrm{RMSD}\rangle$, the structures for which the selected amino acid pairs have large deviations in their pair separations (i.e., small $P(S_i,S_j)$) should possess large

C_α RMSD, and the structures for which the selected amino acid pairs have small deviations (i.e., large $P(S_i, S_j)$) should possess small C_α RMSD. To select the set of amino acid pairs that minimize $\langle \text{RMSD} \rangle$, we add one pair to the set of optimal pairs at a time. We start with identifying the single pair (β^*, δ^*) that minimizes $\langle \text{RMSD} \rangle$ and use it as the restraint for $N_r = 1$. We then consider two possible pairs, but fix one of the pairs to be (β^*, δ^*) and find the new pair $(\beta^{**}, \delta^{**})$ in the set $\{(\beta^*, \delta^*), (\beta^{**}, \delta^{**})\}$ that minimizes $\langle \text{RMSD} \rangle$. We use these two pairs as the set of restraints for $N_r = 2$. This process continues until we have $N_r = 1, \ldots, N_{\max}$, where $N_{\max}/N < 0.08$ for all proteins considered.

2.3.3 | Identifying the largest C_α separations in the initial structures

For this method (i.e., the largest C_α separation method) for selecting important restraints, we identify the amino acid pairs with the largest C_α separations, or the maximum $R_{\beta\delta}$ over all pairs $\{(\beta, \delta)\}$ in the initial protein structure (e.g., in Figure 1c, we show the amino acid pair with the largest C_α separation in the x-ray crystal structure for T4L (172L)). After identifying the pair (β, δ) with the largest C_α separation, we find the pair (β', δ') with the second largest C_α separation. This process is then continued to find a set of pairs with the largest C_α separations. However, we seek to identify a minimal set of amino acid pairs without redundant structural information. Thus, we carried out unrestrained MD simulations starting from the initial structure and determined the Pearson correlation between the pairwise C_α separations. In particular, we include (β', δ') in the pool of selected pairs if the Pearson correlation ρ between $R_{\beta\delta}$ and $R_{\beta'\delta'}$ satisfies $|\rho| < 0.9$. We also require that the C_α atoms of the new pair are not already in the pool of selected restraints. Thus, we first add the pair (β, δ) with the largest C_α separation to the pool of restraints ($N_r = 1$). We add (β', δ') with the second largest C_α separation to the pool of restraints ($N_r = 2$) as long as it is uncorrelated with (β, δ) and does not include C_α atom β or δ. If so, we consider the pair (β'', δ'') with the next largest separation. We follow this process until we have sets of restraints with $N_r = 1, \ldots, 5$.

2.3.4 | Identifying the largest change in pairwise C_α separations between the initial and target structures

For this method (i.e., the largest change in pairwise separation method), we assume that the target structure is known. We identify the amino acid pair with the largest change in pairwise C_α separations between the initial and target structures S_i and S_j (i.e., the maximum $\Delta R_{\beta\delta}(S_i, S_j)$). (For example, in Figure 1d, we show the amino acid pair with the largest change in C_α separations between the initial (172L) and target (1L69) x-ray crystal structures for T4L. Note that the amino acids with the largest change in pairwise separation are not the same as those with the largest C_α separation.) To identify a set of non-redundant restraints, we successively implement new restraints in the MD simulations for pairs with the largest deviation in the pair separation from the target. In particular, assume that pair (β, δ) has the largest $\Delta R_{\beta\delta}(S_i, S_j)$ in the unrestrained MD simulations starting from the initial structure. We then carry out MD simulations with $R_{\beta\delta}$ restrained to the target value. We identify the pair (β', δ') with the largest $\Delta R_{\beta\delta}(S_i, S_j)$ and carry out restrained MD restraints enforcing the target values for both $R_{\beta\delta}$ and $R_{\beta'\delta'}$. We also require that the C_α atoms of the new pair are not the same as any of the atoms in the current pool of restraints. We use (β, δ) as the restraint for $N_r = 1$ and use (β, δ) with (β', δ') as the set of restraints for $N_r = 2$. This process continues until we have $N_r = 1, \ldots, N_{\max}$, where $N_{\max}/N < 0.08$ for all proteins we considered.

2.3.5 | Linear discriminant analysis

For the linear discriminant analysis method of selecting restraints, we also assume that the target structure is known. We seek to identify the pairwise separation between C_α atoms that can serve as a collective variable enabling the protein to move from the initial to the target structure (Sittel and Stock 2018). The inspiration for this method is the linear discriminant analysis method that was used to identify the collective variables for small molecule conformational changes (Mendels and de Pablo 2022). Here, we apply the method in the context of large conformational changes in proteins.

Linear discriminant analysis requires a distribution of protein structures, not a single structure. Thus, we first carried out short 20 ns unrestrained MD simulations starting from both the initial and target structures to sample an ensemble of structures near the initial and target structures. In Figure 1e, we show these two distributions of structures for T4L as an example. In this case, we aim to find the collective variable that tracks the hinge closure motion of the two domains. Each protein conformation can be described by the set of all pairwise separations between C_α atoms,

$$\vec{X} = (R_{\beta\delta}, R_{\beta'\delta'}, \ldots)^T, \tag{7}$$

where the length of the vector is the number of distinct amino acid pairs N_p minus the ones that are too close together ($R_{\beta\delta} < R_g$) and the ones where at least one amino acid is buried. We can describe the distribution of protein conformations sampled near the initial and target structures as $\rho_A\left(\vec{X}_A\right)$ and $\rho_A\left(\vec{X}_B\right)$, respectively. We aim to find the direction $\widehat{W}$ in the space of C$_\alpha$ separations such that its projection onto the cross-correlation matrix (between initial and target distributions) is maximized, while its projection onto the covariance matrix for the initial or target distributions is minimized. To determine $\widehat{W}$, we first calculate the mean separations $\left\langle\vec{X}_A\right\rangle$ and $\left\langle\vec{X}_B\right\rangle$ from the distributions of the initial and target structures. Second, we define the (cross-correlation) scattering matrix

$$S_b = \left(\left\langle\vec{X}_A\right\rangle - \left\langle\vec{X}_B\right\rangle\right)\left(\left\langle\vec{X}_A\right\rangle - \left\langle\vec{X}_B\right\rangle\right)^T, \tag{8}$$

which quantifies the covariance of the initial distribution with the target distribution, and define the projection of $\widehat{W}$ onto S_b as $\widehat{W}^T S_b \widehat{W}$. We also quantify the covariance of the initial and target distributions, separately,

$$\Sigma_{A,B} = \left(X_{A,B} - \langle X_{A,B}\rangle\right)\left(X_{A,B} - \langle X_{A,B}\rangle\right)^T. \tag{9}$$

We can then calculate the (direct correlation) scattering matrix,

$$S_w = \Sigma_A + \Sigma_B, \tag{10}$$

with projection $\widehat{W}^T S_w \widehat{W}$. To find the direction $\widehat{W}$ onto which the projections of the two distributions $\rho_A\left(\vec{X}_A\right)$ and $\rho_B\left(\vec{X}_B\right)$ are best separated, we can maximize the Rayleigh ratio,

$$\mathscr{R}\left(\widehat{W}\right) = \frac{\widehat{W}^T S_b \widehat{W}}{\widehat{W}^T S_w \widehat{W}}, \tag{11}$$

which is equivalent to solving for the eigenvector $\widehat{W}_\lambda$ associated with the largest eigenvalue λ of $S_w^{-1}S_b$,

$$S_w^{-1} S_b \widehat{W}_\lambda = \lambda \widehat{W}_\lambda. \tag{12}$$

We identify the most important pairwise distances as those that have the largest weights in $\widehat{W}_\lambda$. For $N_r = 1$, we choose the amino acid pair (β, δ) with the largest weight. For $N_r = 2$, we choose the two amino acid pairs (β, δ) and (β', δ') with the two largest weights, and so on.

2.4 | Restrained MD simulations

To assess the performance of the FRET pair selection methods, we carry out all-atom MD simulations starting from the initial structure and incorporating the C$_\alpha$-C$_\alpha$ separations of the selected pairs from the target structure as the equilibrium lengths of the linear spring restraints. We then monitor the RMSD of the C$_\alpha$ atom positions (Equation (2)) from the target structure as a function of time. Unrestrained and restrained MD simulations were performed using the AMBER99SB-ILDN force field (Best and Hummer 2009; Lindorff-Larsen et al. 2010) in the GROMACS software package (Abraham et al. 2015). The MD simulations were carried out in a periodic dodecahedron-shaped box that is sufficiently large such that the protein surface is at least 20 Å from the box edges. The simulation box was solvated with water molecules modeled using TIP3P at neutral pH and 0.15M NaCl (Jorgensen 1981; Mark and Nilsson 2001). Short-range van der Waals and screened Coulomb interactions were truncated at 1.2 nm, while long-ranged electrostatic interactions were tabulated using the Particle Mesh Ewald summation method. The LINCS algorithm was used to constrain the bond lengths. We performed two energy minimization runs to first relax the protein and then the water molecules and the protein together using the steepest decent algorithm until the maximum net force magnitude on an atom is smaller than 500 kJ mol^{-1} nm^{-1}. We perform NVT simulations of the system for 20 ns at temperature $T = 300$ K using a velocity rescaling thermostat for sampling the canonical ensemble (Bussi et al. 2007). The equations of motion for the atomic coordinates and velocities are integrated using a leapfrog algorithm with a 1 fs time step.

For the restrained simulations, we employ a linear spring potential for each restrained amino acid pair (β,δ),

$$E_r\left(R_{\beta\delta}\right) = \frac{k_r}{2}\left(R_{\beta\delta} - R^0_{\beta\delta}\right)^2, \tag{13}$$

where $R^0_{\beta\delta}$ is the C$_\alpha$-C$_\alpha$ separation for the (β,δ) pair in the target structure and the spring constant $k_r = 5000$ kJ mol^{-1} nm^{-2} is chosen so that the averaged root-mean-square deviation between $R_{\beta\delta}$ and $R^0_{\beta\delta}$ is <0.2 Å (see Figure S2).

For the unrestrained MD simulations for each protein, we ran $N_s = 100$ simulations for 20 ns starting from the initial structure, but with different initial velocities for each atom randomly selected from a Maxwell-Boltzmann distribution at $T = 300$ K. Using these simulations, we can calculate the average C$_\alpha$ RMSD between the initial structures and the target structure, which serves as a baseline for comparison to the results for the restrained MD simulations. For the random restraint selection method at each N_r, we randomly select $N_s = 100$ sets of N_r amino acid pairs and run one restrained MD simulation (for 20 ns) for each set of restraint pairs. For all other restraint selection methods, we carry out $N_s = 50$ simulations for each set of N_r restraints, but with different random initial velocities. The number of samples N_s is chosen so that average C_α RMSD converges at large N_s as shown in Figure S3. For each N_r and for each restraint selection method, we average the C_α RMSD over all N_s simulations. For the restrained MD simulations of the in vitro protein structures, the C_α RMSD converges within the first 5 ns, as shown in Figure S4. Therefore, we only use the data generated from 5 to 20 ns for the calculations of the C_α RMSD. For the in vivo target GB1, we extended the MD simulations to 50 ns and use the time points from 20 to 50 ns to calculate the C_α RMSD (see Figure S6c).

To determine the ability of the restraints to move the protein conformation from the initial to the target structure, we need to measure a reference C_α RMSD (i.e., RMSD_r) that quantifies thermal fluctuations around the target structure. For the proteins T4L, PGK, and AK, we determine RMSD_r by running unrestrained MD simulations starting from the target structure. We set $\text{RMSD}_r = \text{RMSD}(S_i, S_j)$, where S_i is the central structure from the unrestrained MD simulations starting from the target structure, and S_j is the target structure. We find the central structure by identifying the structure that has the largest number of neighboring structures in the ensemble using a cutoff C_α RMSD of 1 Å. If the average C_α RMSD obtained from the restrained MD simulations is below RMSD_r, the restraints were successful in moving the initial structure to the target structure since the differences are comparable to thermal fluctuations observed near the target structure. For the in vitro and in-cell NMR structures of GB1 and the unbound structure of TCI, we set $\text{RMSD}_r = N_m^{-1} \sum_{i>j} \text{RMSD}(S_i, S_j)$, where S_i and S_j are distinct models in the in-cell NMR bundle and N_m is the number of distinct pairs of models.

3 | RESULTS

Does the choice of the amino acid restraints have a significant impact on whether the initial protein structure can be moved to the target structure? As an example, we consider two possible choices for single amino acid restraints in restrained MD simulations of T4L. In Figure 2, we plot the C_α $\text{RMSD}(S_i, S_j)$ as a function of time, where S_i are the structures sampled in the restrained MD simulations and S_j is the target structure, for two possible choices for single amino acid restraints. Although the two restrained MD simulations start from the same initial structure (green) in Figure 2a, the C_α $\text{RMSD}(S_i, S_j)$ display different time dependence. The C_α RMSD from the MD simulations using the (β', δ') restraint rapidly decreases below the reference value, RMSD_r, for the target structure. When we visualize the final frame from this restrained MD simulation in Figure 2c, we find that the two subdomains of T4L are now closer together and there is strong alignment between the final frame (green) and target structure (blue). In contrast, the C_α RMSD does not decrease below RMSD_r for the MD simulations with the (β, δ) restraint, even though the restraint is well-satisfied during the simulations. This example emphasizes the importance of the restraint selection method. Below, we describe the performance of four restraint selection methods in moving an initial structure toward a target structure. We compare the performance of restraint selection methods that have information about the target structure and those that do not, and benchmark their performance to random restraint selection. In addition, we test the methods on moving four proteins from one in vitro structure to another, one protein from an in vitro structure to an in-cell structure, and the same protein from its in-cell structure to its in vitro structure.

3.1 | Performance of restraint selection methods

To quantify the performance of the four restraint selection methods in moving the initial structure to the target structure, in Figure 3, we plot $\langle \text{RMSD}(S_i, S_j) \rangle_{S_i, N_s}$ averaged over structures S_i from the restrained MD simulations and number of samples N_s, where S_j is the target structure, as a function of the number of restraints N_r for several restraint selection methods and three proteins. In Figure 3a, we show the results for T4L. When $N_r = 0$ (i.e., unrestrained MD simulations), $\langle \text{RMSD}(S_i, S_j) \rangle_{S_i, N_s} \sim 4$ Å and the structures sampled in the MD simulations remain far from the target structure. When randomly selecting restraints, $\langle \text{RMSD}(S_i, S_j) \rangle_{S_i, N_s}$ decreases slowly with increasing N_r and does not reach RMSD_r even after five restraints have been included. In contrast, if we use structural information about the target to select the restraints (i.e., using linear discriminant analysis or identifying the largest change in pairwise C_α separations between the initial and target structures), $\langle \text{RMSD}(S_i, S_j) \rangle_{S_i, N_s}$ rapidly decreases to RMSD_r after adding only one restraint. The normal mode analysis and largest C_α separation methods, which do not use information about the target, achieve intermediate results; $\langle \text{RMSD}(S_i, S_j) \rangle_{S_i, N_s}$ decreases to RMSD_r within $N_r = 2$-3 restraints. Thus, both of these methods (normal mode

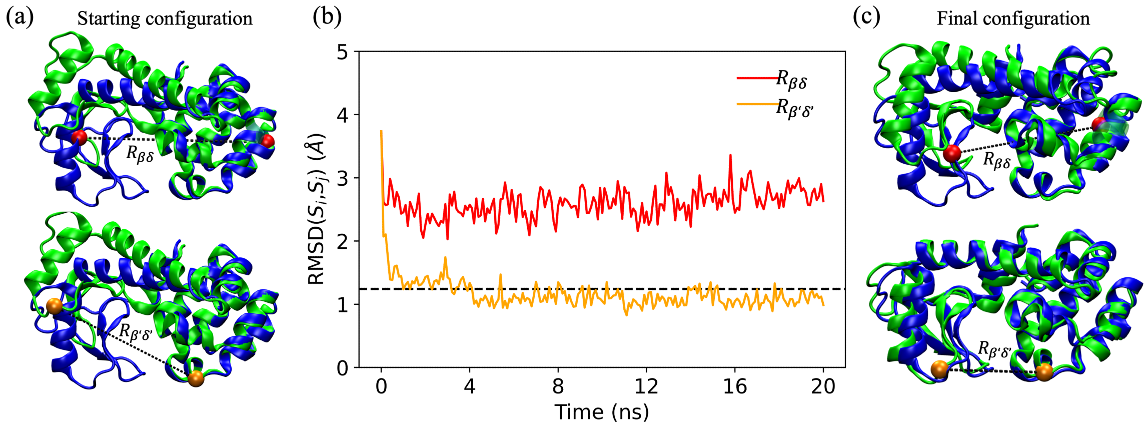

FIGURE 2 (a) The top and bottom green images indicate the same starting structures for restrained MD simulations that enforce two different single restraints: $R_{\beta\delta} = R^0_{\beta\delta}$ (red) and $R_{\beta,\delta} = R^0_{\beta,\delta}$ (orange). The target structures, which are aligned with the starting structures such that the C_α RMSD values are minimized for residues 92–162, are shown in blue. (b) $\mathrm{RMSD}(S_i, S_j)$ as a function of time during restrained MD simulations for the restraints in (a), where the S_i are structures sampled at each time during the restrained MD simulations and S_j is the target structure. The black dashed line indicates RMSD_r of the target. (c) The final frames at 20 ns from the restrained MD simulations with $R_{\beta\delta} = R^0_{\beta\delta}$ (top, red) and $R_{\beta,\delta} = R^0_{\beta,\delta}$ (bottom, orange). The structures from the final frames of the restrained MD simulations (green) are aligned with the target structure (blue) using the same alignment as in (a).

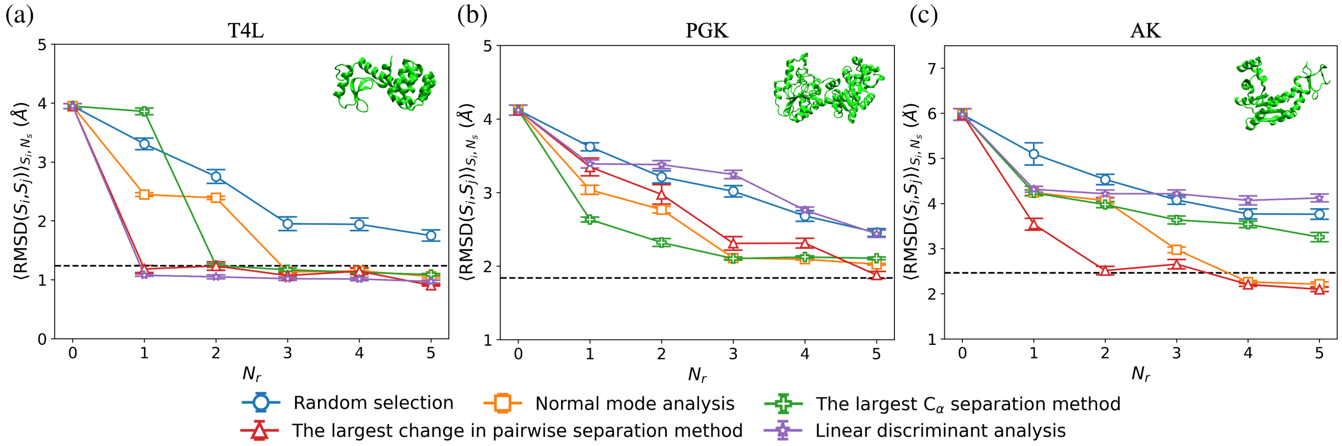

FIGURE 3 The C_α RMSD, $\langle \mathrm{RMSD}(S_i, S_j)\rangle_{S_i,N_s}$, averaged over structures S_i from the restrained MD simulations and number of samples N_s, where S_j is the target structure, is plotted versus the number of restraints N_r for (a) T4L, (b) PGK, and (c) AK. We show results for several restraint selection methods: random selection (blue circles), normal mode analysis (orange squares), the largest C_α separation method (green crosses), the largest change in pairwise separation method (red upward triangles), and linear discriminant analysis (purple stars). The horizontal black dashed line indicates RMSD_r for each target. Snapshots of the initial structures are shown for each protein in the upper right corner of each panel. The error bars give the standard error of $\langle \mathrm{RMSD}(S_i, S_j)\rangle_{S_i}$ from N_s independent simulations.

analysis and the largest C_α separation method) are more efficient than random selection for moving the initial structure to the target for T4L.

To investigate the performance of restraint selection methods in moving an initial structure to the target structure for a wider range of proteins, we performed unrestrained and restrained MD simulations on two larger proteins, PGK (Figure 3b) and AK (Figure 3c), and find similar results for the variation of $\langle \mathrm{RMSD}(S_i, S_j)\rangle_{S_i,N_s}$ with N_r. (Note that for AK, we needed to add a background of intra-domain restraints to recapitulate the B-factor in the "unrestrained" MD simulations as shown in Figure S5.) For example, for random restraint selection, $\langle \mathrm{RMSD}(S_i, S_j)\rangle_{S_i,N_s}$ slowly decreases with increasing N_r, and $\langle \mathrm{RMSD}(S_i, S_j)\rangle_{S_i,N_s} > \mathrm{RMSD}_r$ even for $N_r = 5$. For PGK and AK, all restraint selection methods (except linear discriminant analysis) outperform random restraint selection. Normal mode analysis and the largest C_α separation method, which do not include information about the target, achieve lower values of $\langle \mathrm{RMSD}(S_i, S_j)\rangle_{S_i,N_s}$ than that for random selection, but their relative performance depends on the protein. For

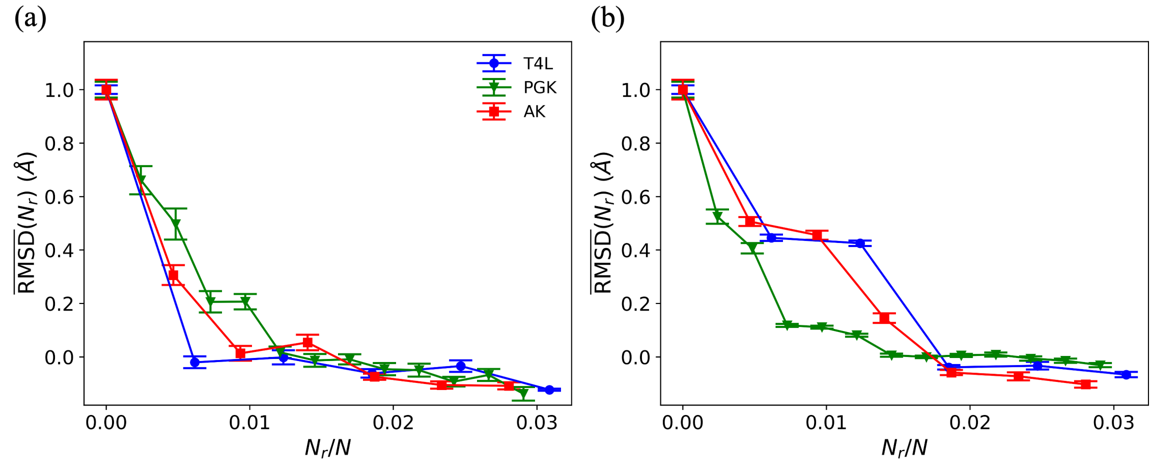

FIGURE 4 The normalized C_α RMSD, $\overline{\mathrm{RMSD}}(N_r)$, is plotted versus N_r/N for T4L (blue circles), PGK (green triangles), and AK (red squares) using (a) the largest change in pairwise C_α separation method and (b) the largest C_α separation method for selecting the restraints.

example, $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}$ decreases faster with N_r for the largest C_α separation method for PGK, whereas $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}$ decreases faster for the normal mode analysis method for AK. As expected, the method of identifying the largest change in pairwise C_α separations between the initial and target structures is the best performing method with $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s} \sim \mathrm{RMSD}_r$ after only a small number of restraints (5 for PGK and 2 for AK). The performance of the restrained MD simulations based on the linear discriminant analysis method of selecting restraints is comparable to that for random restraint selection even though it incorporates structural information about the target. This result likely stems from the fact that a nonlinear method is needed to identify the collective variables in proteins.

3.2 | Minimum fraction of restraints

These results demonstrate that the addition of a small number of pairwise distance restraints in restrained MD simulations can effectively move a protein from an initial structure to a target structure that is more than 4 Å away. However, the minimal number of restraints required for $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s} \sim \mathrm{RMSD}_r$ depends on the protein. What controls the minimum number of restraints? To address this question, we calculate the normalized C_α RMSD,

$$\overline{\mathrm{RMSD}}(N_r) = \frac{\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}(N_r) - \mathrm{RMSD}_r}{\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}(0) - \mathrm{RMSD}_r}, \qquad (14)$$

where $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}(0)$ is $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}$ from unrestrained MD simulations with $N_r = 0$. Thus, $\overline{\mathrm{RMSD}}(N_r) = 1$ indicates that the protein samples the ensemble of initial structures and $\overline{\mathrm{RMSD}}(N_r) = 0$ indicates that the protein samples the target ensemble. In Figure 4, we show that $\overline{\mathrm{RMSD}}(N_r)$ collapses for the three proteins T4L, PGK and AK when plotted versus N_r/N and using the optimal restraint selection method (i.e., the largest change in pairwise separation). In this case, only a small fraction of restraints, less than 1.5% of the protein size, is required to move between the two in vitro protein structures. For the normal mode analysis restraint selection method, which does not consider information about the target, the collapse of $\overline{\mathrm{RMSD}}$ with N_r/N is not as complete, but $\overline{\mathrm{RMSD}} \sim 0$ for $N_r/N \gtrsim 0.02$ for all proteins. Thus, even in the absence of complete knowledge of the target structure, we can induce changes in protein structure from an initial in vitro structure to a target in vitro structure using only a small fraction of amino acid separation restraints.

3.3 | Moving from initial in vitro structure to target in-cell structure for GB1

To investigate whether the proposed methodology can also move an initial in vitro structure to a target in vivo structure, we also carried out unrestrained and restrained MD simulations for the B1 domain of protein G (GB1) (Gerez et al. 2022). We find that inter-bundle $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,S_j} \sim 3.0$ Å between the in vitro NMR models S_i and the in-cell NMR models S_j is rather small, mostly due to small changes in the two top loop regions shown in Figure 5a. Despite this, the C_α RMSD between the in vitro and in-cell NMR bundles is larger than both of the intra-bundle fluctuations. Can we use similar restraint selection methods to those above to move from

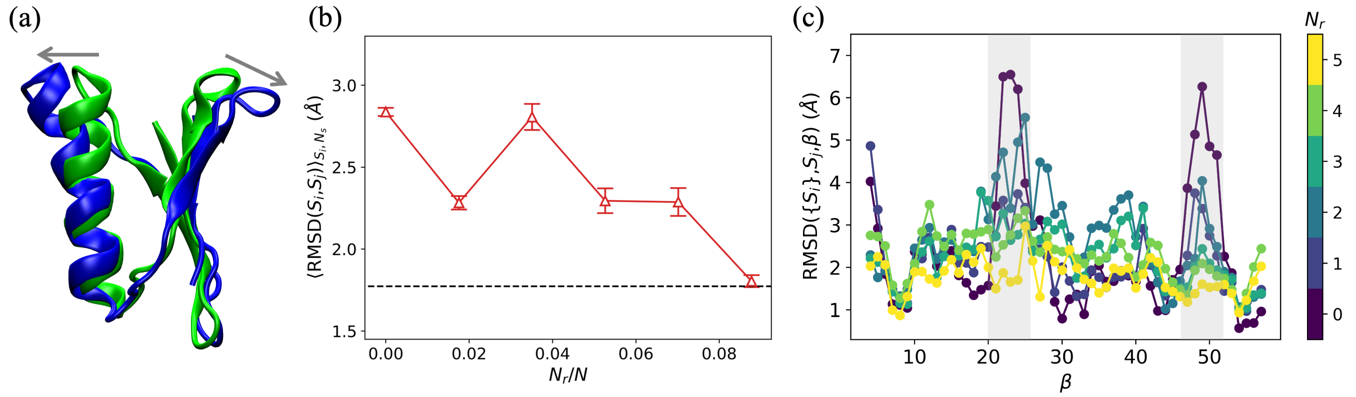

FIGURE 5 (a) A ribbon diagram of the initial in vitro structure (green) of GB1 aligned with its in-cell target structure (blue). The gray arrows indicate the structural changes from the initial structure to the target structure in the two top loop regions. (b) $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}(N_r)$ plotted as a function of N_r/N for restrained MD simulations of GB1 using the largest change in pairwise C_α separation between the initial and target structures method (red upward triangles) for selecting restraints. The horizontal black dashed line indicates RMSD_r for the target. (c) $\mathrm{RMSD}(\{S_i\},S_j,\beta)$ (Equation (3)) is plotted versus amino acid index β, where $\{S_i\}$ are structures sampled in the restrained MD simulations and S_j is the target structure. The color from dark blue to yellow indicates N_r. The two top loop regions in (a) with $21 \leq \beta \leq 26$ and $47 \leq \beta \leq 52$ are highlighted in light gray.

an in vitro structure to an in-cell structure? In contrast to the in vitro target structures for T4L, PGK, and AK, the in-cell target structure for GB1 is not a local minimum of the AMBER99SB-ILDN force field (see Figures S8a,b). As a result, simulating in-cell protein structures is more challenging. In Figure 5b, we implement the largest change in pairwise separation method for identifying restraints and calculate $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}$ as a function of N_r. We find that the target in-cell structures are sampled for $N_r \geq 5$ in restrained MD simulations, which corresponds to $N_r/N \sim 8\%$. Unlike for the in vitro structures, where $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}$ monotonically decreases as N_r increases, $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}$ exhibits non-monotonic behavior with N_r for GB1. To investigate this effect, we calculate the C_α RMSD between the structures from the MD simulations and the target for each amino acid β individually (Equation (3)). As shown in Figure 5c, the deviations in the structures sampled in the unrestrained simulations and target structure occur predominantly in the top two loop regions in Figure 5a with $21 \leq \beta \leq 26$ and $47 \leq \beta \leq 52$. The deviations between the initial and target structures can be decreased in the loop regions to $\lesssim 2$ Å for $N_r = 5$. However, the structural deviations at amino acid positions adjacent to the loop regions, for example, $17 \leq \beta \leq 20$ and $27 \leq \beta \leq 31$, begin to increase as N_r increases. These results suggest that the restraints are competing with the force field when attempting to move the protein to the in-cell structure. The non-monotonic behavior in $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}$ gives rise to a higher fraction of restraints that are necessary to move GB1 from the initial in vitro structure to the in-cell structure.

3.4 | Relation between minimum fraction of restraints and stability of target structure

Several studies have suggested that GB1 is a single-domain protein, not a two-domain protein as for T4L, PGK, and AK (Byeon et al. 2003). It is possible that the larger fraction of restraints needed to induce a conformational change in GB1 (relative to the fraction needed for T4L, PGK, and AK) is related to the fact that it behaves like a single-domain protein. However, the variation in the minimum fraction of restraints can also be related to the stability of the target structure in the force field without restraints. To investigate this effect, we calculated the minimum number of restraints needed to move the "unstable" in vivo structure of GB1 to its "metastable" in vitro structure. We carried out restrained MD simulations of GB1 using restraints selected from the largest change in pairwise separation method. We find that only $\sim 3\%$ of the restraints are needed to move from the in vivo to the in vitro structure for GB1 (Figure 6a), whereas $\sim 8\%$ of the restraints are needed to move from the in vitro to the in vivo structure. We also studied another two-domain protein tick carboxypeptidase inhibitor (TCI) that undergoes a conformational change upon binding to its receptor. (The bound and unbound structures have PDBIDs 1ZLI and 2JTO, respectively.) The unbound NMR structure is unstable in the AMBER99SB-ILDN force field as shown in Figure S7. We find that the fraction of restraints needed to move from the initial bound state to the target unbound state of TCI is significantly higher ($\sim 7\%$) than the fraction of restraints

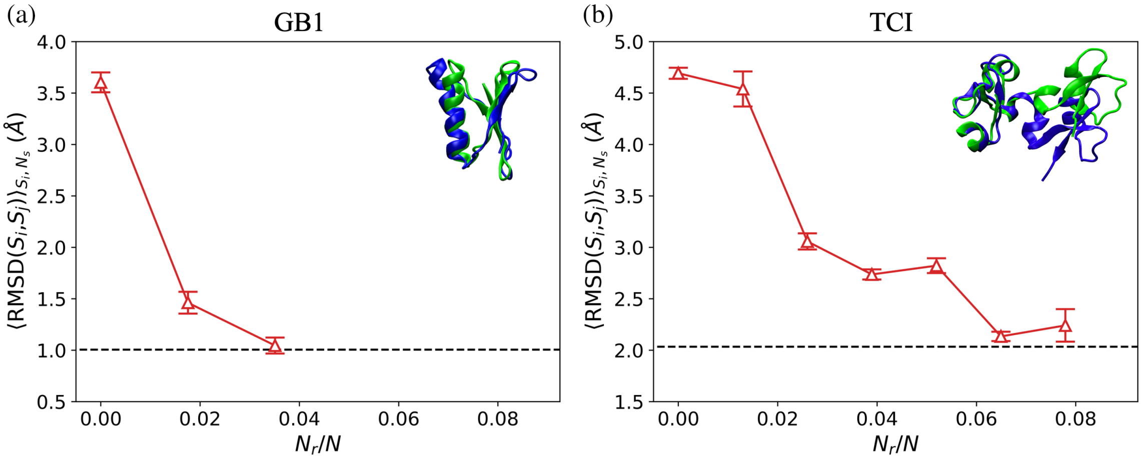

FIGURE 6 $\langle \mathrm{RMSD}(S_i,S_j)\rangle_{S_i,N_s}$ plotted as a function of N_r/N for restrained MD simulations of (a) GB1 and (b) TCI using the largest change in pairwise C$_\alpha$ separation between the initial and target structures method (red upward triangles) for selecting restraints. The horizontal black dashed lines indicate RMSD$_r$ for the targets. Ribbon diagrams of the initial structures (blue) aligned with the target structures (green) for GB1 and TCI are shown in the upper right corners of (a) and (b), respectively.

needed to move from the unstable in vivo state to the stable in vitro state of GB1, even though TCI is a two-domain protein (see Figure 6b). Therefore, the characteristic fraction of restraints needed to induce conformational changes is primarily determined by the stability of the target state in the force field (without restraints).

4 | DISCUSSION

Numerous studies have emphasized that the in-cell environment can strongly influence protein structure and interactions (Ebbinghaus et al. 2010; Ellis 2001; Fulton 1982; Leeb et al. 2020; Speer et al. 2021). However, the complete atomic structure for proteins in the cellular environment has been obtained for only three proteins using NMR spectroscopy (Gerez et al. 2022; Sakakibara et al. 2009; Tanaka et al. 2019). FRET has also been employed to determine a small number of separations between amino acids in several proteins in vivo (Davis and Gruebele 2018; Ebbinghaus et al. 2010; Wang et al. 2018). In contrast, MD simulations, which have been calibrated to structures in the protein data bank, can provide the atomic coordinates of proteins in idealized, in vitro conditions. Properly calibrated force fields that would allow accurate all-atom descriptions of protein conformations in vivo currently do not exist. For example, GTT WW domain and three-helix bundle protein B (PB) fail to refold in all-atom cytoplasm computational models (Rickard et al. 2020; Russell et al. 2023), since the interatomic sticking interactions in current protein force fields increase the stability of non-native states in all-atom cytoplasm models (Rickard et al. 2019; Samuel Russell et al. 2023). The restrained MD simulations of proteins in vivo (using residue separations measured in FRET experiments) can be carried out using current force fields, and thus we do not need to wait for improvements to the force fields in all-atom models of the cytoplasm.

While current all-atom MD simulations allow us to sample in vitro protein structures that match the x-ray crystal or NMR structures, we want to capture the all-atom conformational dynamics of in vivo protein structures. To do this, we can start with an initial in vitro protein structure, and add inter-residue distances measured by FRET as restraints in the MD simulations. However, an important question is how do we determine which amino acid pairs should be measured in the FRET experiments and then incorporated into the restrained MD simulations? In particular, what is the minimal number of restraints needed to achieve a given C$_\alpha$ RMSD to the target structure and what is the optimal method to select these restraints? To answer these questions, we first develop the methodology of restraint selection for proteins with multiple conformational states in vitro. To connect the in vitro results to those for in vivo conditions, we studied GB1, which is one of the few proteins whose structure has been solved in vivo using in-cell NMR spectroscopy. We selected the in vitro structure as the initial structure and the in vivo structure as the target structure. We employed four methods for selecting the restraints, varied the number N_r and type of restraints that are incorporated into the restrained MD simulations, and compared the performance (i.e., C$_\alpha$ RMSD relative to the target structure) of the restraint selection methods to random selection.

We found that for T4L, PGK, and AK, which possess two distinct conformational states that have been solved in vitro and are both stable in the force field without restraints, it only takes a small fraction of restraints ($N_r/N < 2\%$) to induce the conformational changes. The results emphasize that only a limited amount of information about pairwise distances between amino acids is needed to induce protein conformational changes from an initial structure to a target structure. The largest change in pairwise separation method, which uses the target structure information, gives the lowest values of C_α RMSD at each N_r. The normalized $\overline{\text{RMSD}}(N_r)$ for the largest change in pairwise separation method collapses for these three proteins when plotted versus N_r/N, and reaches zero at a small fraction of restraints ($N_r/N < 1.5\%$). The linear discriminant analysis, which also uses the target structure information, exhibits inconsistent performance, sometimes comparable to and sometimes better than random selection. The normal mode analysis and the largest C_α separation method, which do not need target structure information, give lower values of C_α RMSD compared to random selection. The performance of these two methods varies slightly for different proteins. Specifically, for the normal mode analysis, $\overline{\text{RMSD}}(N_r)$ reaches zero when $N_r/N > 2\%$, which is significantly lower than the fraction of restraints used in previous studies of FRET-assisted protein structural modeling (between 5% and 12%) (Dimura et al. 2020). This result is significant since it shows that it is possible to determine the complete protein structure, using information from only $\sim 2\%$ of the C_α separations.

The studies described here provide proof of principle that FRET-assisted MD simulations can improve our understanding of the atomistic structure of proteins in vivo since only a small fraction of restraints are required to lock-in the target structures. We have shown that the characteristic fraction of restraints needed to induce conformational changes is determined by the stability of the target state in the force field without restraints. Since the in vivo structures are not stable minima of the current MD force fields, additional restraints are necessary to reach a given C_α RMSD and the dependence of the C_α RMSD on N_r is non-monotonic. Therefore, an important future direction is to develop force fields for which the in vivo structures are potential energy minima. However, this goal requires a large number of high-quality, all-atom in vivo protein structures. In the absence of many in vivo protein structures, we can also develop in-cell mimetic systems whose crowding and surface sticking interactions have been calibrated to the in vivo studies (Davis et al. 2020; Davis and Gruebele 2018).

The linear discriminant analysis approach of using the direction that maximizes the differences in the initial and final state projections has been used successfully to identify the key pairwise separations that distinguish the initial and target states of small molecules (Mendels and de Pablo 2022). However, while we find that linear discriminant analysis outperforms random selection only for T4L ($N = 162$ amino acids), it performs worse than random selection for PGK ($N = 413$) and AK ($N = 214$). This result suggests that as the number of amino acids increases, the dimensionality of the conformation space increases to such an extent that the linear method cannot capture the key pairwise distances that describe the structural transition. Therefore, nonlinear methods, such as the nudged elastic band method, are needed to calculate the minimum energy pathway between the initial and final states and identify the most important pairwise distances along the pathway (Ghoreishi et al. 2019; Jónsson et al. 1998).

The largest change in pairwise separation method that identifies a minimal set of amino acid pairs without redundant structural information gives the lowest values of C_α RMSD to the target structures. Note that one should not imagine a set of unique, optimal restraints, but an optimal pool of similar restraints that can induce structural changes (see Figure S8). However, this method requires information about both the initial and target structures, and thus it cannot be used to *predict* an in vivo structural ensemble that has not yet been determined. Given that there are several NMR structures for proteins in cellular environments available, this selection method can be used in restrained MD simulations to verify that the simulations can correctly sample the conformational dynamics for these proteins in vivo. In contrast, if we aim to predict the in vivo structure of proteins, we have shown that normal mode analysis, which assumes that the normal modes of vibration of the initial structure provide information about how the protein transitions from the initial structure to the target structure, can be used to select the FRET-pair labeling positions that will enable the restrained MD simulations to sample the target in vivo structure. In this work, which considers large-scale domain motion in proteins, we showed that the normal mode analysis is a successful restraint selection method. However, it is not yet clear whether such motion is relevant for proteins in vivo. Thus, if the normal mode analysis is not efficient in inducing a change from the initial to the target in vivo structure, the largest C_α separation method, or a hybrid method that couples normal mode analysis and the largest C_α separation method, can also be implemented (see Figure S9).

It is important to note that there are several experimental limitations for selecting amino acid pairs for

FRET labeling. For example, the labeled dye molecules should not perturb the protein structure. Thus, in this study, we did not select amino acid pairs that occur within the core. FRET-labeling core amino acids would likely lead to changes in protein structure. In addition, optimal efficiency for FRET-labeling is achieved when the distance d between donor and acceptor molecules is comparable to the characteristic Förster distance R_0 for each dye pair. For most common FRET chromophores, $50\,\text{Å} < R_0 < 60\,\text{Å}$, which is larger than R_g for the proteins we considered in this work. Thus, we excluded amino acid pairs that are separated by $d < R_g$.

As mentioned above, we showed in this study that the normal mode analysis selection method (and largest C_α separation method) can identify a small number of important C_α-C_α distance restraints that can effectively move an initial in vitro structure to a target in vivo structure. However, since the FRET energy transfer efficiency reports on the ensemble of distances between dye molecules (not C_α-C_α distances), quantitative FRET-assisted protein structural modeling requires the mapping between dye–dye distances and C_α-C_α distances (Klose et al. 2021). Therefore, in future studies, we will identify the mapping by incorporating atomic-scale modeling of the dye molecules into the restrained MD simulations.

Finally, our current analysis of restraint selection methods focuses on monomeric, globular proteins. In future studies, we can expand the application of our restraint selection methods to intrinsically disordered proteins (see Figure S10), membrane proteins, protein–protein and protein-nucleic acid complexes. Another interesting future direction is to develop restraint selection methods for triple restraints (co-restraints between 3 residues), using the inter-residue distances obtained from a 3-color FRET experiments (Yoo et al. 2020).

ACKNOWLEDGMENTS

The authors acknowledge support from National Institutes of Health (NIH) Grant No. R35GM151146 (E.K. and C.D.) and NIH Training Grant No. T32GM145452 (Z.L., A.G., and C.O.) and T15LM007056-37 (J.S.), as well as the High Performance Computing facilities operated by Yale's Center for Research Computing.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The code for implementing the five restraints selection methods can be found at https://github.com/lzyttxs/restraint_selection_methods/.

REFERENCES

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1:19–25.

Agam G, Gebhardt C, Popara M, Mächtel R, Folz J, Ambrose B, et al. Reliability and accuracy of single-molecule FRET studies for characterization of structural dynamics and distances in proteins. Nat Methods. 2023;20:523–35.

Ahmed A, Gohlke H. Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. Proteins Struct Funct Bioinf. 2006;63:1038–51.

Arolas JL, Lorenzo J, Rovira A, Castellà J, Aviles FX, Sommerhoff CP. A carboxypeptidase inhibitor from the tick rhipicephalus bursa: isolation, cDNA cloning, recombinant expression, and characterization. J Biol Chem. 2005b;280:3441–8.

Arolas JL, Popowicz GM, Lorenzo J, Sommerhoff CP, Huber R, Aviles FX, et al. The three-dimensional structures of tick carboxypeptidase inhibitor in complex with A/B carboxypeptidases reveal a novel double-headed binding mode. J Mol Biol. 2005a;350:489–98.

Baase WA, Liu L, Tronrud DE, Matthews BW. Lessons from the lysozyme of phage T4. Protein Sci. 2010;19:631–41.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res. 2000;28:235–42.

Best RB, Hummer G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. J Phys Chem B. 2009;113:9004–15.

Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. J Chem Phys. 2007;126:014101.

Byeon IJ, Louis JM, Gronenborn AM. A protein contortionist: Core mutations of GB1 that induce dimerization and domain swapping. J Mol Biol. 2003;333:141–52.

Campos-Olivas R, Aziz R, Helms GL, Evans JN, Gronenborn AM. Placement of 19F into the center of GB1: effects on structure and stability. FEBS Lett. 2002;517:55–60.

Carroni M, Saibil HR. Cryo electron microscopy to determine the structure of macromolecular complexes. Methods. 2016;95:78–85.

Chang J, Zhang C, Cheng H, Tan YW. Rational design of adenylate kinase thermostability through coevolution and sequence divergence analysis. Int J Mol Sci. 2021;22:2768.

Cheng J, Liu T, You X, Zhang F, Sui SF, Wan X, et al. Determining protein structures in cellular lamella at pseudo-atomic resolution by GisSPA. Nat Commun. 2023;14:1282.

Davis CM, Deutsch J, Gruebele M. An in vitro mimic of in-cell solvation for protein folding studies. Protein Sci. 2020;29:1046–54.

Davis CM, Gruebele M. Non-steric interactions predict the trend and steric interactions the offset of protein stability in cells. ChemPhysChem. 2018;19:2290–4.

Dimura M, Peulen TO, Hanke CA, Prakash A, Gohlke H, Seidel CA. Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems. Curr Opin Struct Biol. 2016;40:163–85.

Dimura M, Peulen TO, Sanabria H, Rodnin D, Hemmen K, Hanke CA, et al. Automated and optimally FRET-assisted structural modeling. Nat Commun. 2020;11:5394.

Dunstone MA, de Marco A. Cryo-electron tomography: an ideal method to study membrane-associated proteins. Philos Trans R Soc B Biol Sci. 2017;372:20160210.

Ebbinghaus S, Dhar A, McDonald JD, Gruebele M. Protein folding stability and dynamics imaged in a living cell. Nat Methods. 2010;7:319–23.

Ellis RJ. Macromolecular crowding: obvious but underappreciated. Trends Biochem Sci. 2001;26:597–604.

Feng R, Gruebele M, Davis CM. Quantifying protein dynamics and stability in a living organism. Nat Commun. 2019;10:1179.

Fiorillo A, Petrosino M, Ilari A, Pasquo A, Cipollone A, Maggi M, et al. The phosphoglycerate kinase 1 variants found in carcinoma cells display different catalytic activity and conformational stability compared to the native enzyme. PLoS One. 2018;13:e0199191.

Flores S, Echols N, Milburn D, Hespenheide B, Keating K, Lu J, et al. The database of macromolecular motions: new features added at the decade mark. Nucleic Acids Res. 2006;34: D296–301.

Fulton AB. How crowded is the cytoplasm? Cell. 1982;30:345–7.

Gerez JA, Prymaczok NC, Kadavath H, Ghosh D, Bütikofer M, Fleischmann Y, et al. Protein structure determination in human cells by in-cell NMR and a reporter system to optimize protein delivery or transexpression. Commun Biol. 2022;5:1322.

Ghoreishi D, Cerutti DS, Fallon Z, Simmerling C, Roitberg AE. Fast implementation of the nudged elastic band method in AMBER. J Chem Theory Comput. 2019;15:4699–707.

Grigas AT, Liu Z, Regan L, O'Hern CS. Core packing of well-defined X-ray and NMR structures is the same. Protein Sci. 2022;31:e4373.

Hellenkamp B, Schmid S, Doroshenko O, Opanasyuk O, Kühnemuth R, Rezaei Adariani S, et al. Precision and accuracy of single-molecule FRET measurements–a multi-laboratory benchmark study. Nat Methods. 2018;15:669–76.

Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, et al. Computed structures of core eukaryotic protein complexes. Science. 2021;374(6573):eabm4805.

Hylton RK, Swulius MT. Challenges and triumphs in cryo-electron tomography. iScience. 2021;24:102959.

Ikeya T, Güntert P, Ito Y. Protein structure determination in living cells. Int J Mol Sci. 2019;20:2442.

Ikeya T, Hanashima T, Hosoya S, Shimazaki M, Ikeda S, Mishima M, et al. Improved in-cell structure determination of proteins at near-physiological concentration. Sci Rep. 2016;6: 38312.

Jónsson H, Mills G, Jacobsen KW, editors. Nudged elastic band method for finding minimum energy paths of transitions. Classical and quantum dynamics in condensed phase simulations. Singapore: World Scientific; 1998. p. 385–404.

Jorgensen WL. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. J Am Chem Soc. 1981;103:335–40.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with Alphafold. Nature. 2021;596:583–9.

Kalinin S, Peulen T, Sindbert S, Rothwell PJ, Berger S, Restle T, et al. A toolkit and benchmark study for FRET-restrained high-precision structural modeling. Nat Methods. 2012;9:1218–25.

Klose D, Holla A, Gmeiner C, Nettels D, Ritsch I, Bross N, et al. Resolving distance variations by single-molecule FRET and EPR spectroscopy using rotamer libraries. Biophys J. 2021;20: 4842–58.

Krüger DM, Ahmed A, Gohlke H. NMSim web server: integrated approach for normal mode-based geometric simulations of biologically relevant conformational transitions in proteins. Nucleic Acids Res. 2012;40:W310–6.

Lallemand P, Chaloin L, Roy B, Barman T, Bowler MW, Lionne C. Interaction of human 3-phosphoglycerate kinase with its two substrates: Is substrate antagonism a kinetic advantage? J Mol Biol. 2011;409:742–57.

Leeb S, Sörensen T, Yang F, Mu X, Oliveberg M, Danielsson J. Diffusive protein interactions in human versus bacterial cells. Curr Res Struct Biol. 2020;2:68–78.

Li M, Xu W, Zhang JZ, Xia F. Combined effect of confinement and affinity of crowded environment on conformation switching of adenylate kinase. J Mol Model. 2014;20:2530.

Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. Systematic validation of protein force fields against experimental data. PLoS One. 2012;7:e32131.

Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. Science. 2011;334:517–20.

Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins Struct Funct Bioinf. 2010;78:1950–8.

Love O, Galindo-Murillo R, Zgarbová M, Šponer J, Jurečka P, Cheatham TE. Assessing the current state of Amber force field modifications for DNA-2023 edition. J Chem Theory Comput. 2023;19:4299–307.

Luchinat E, Banci L. In-cell NMR: recent progresses and future challenges. Rend Fis Acc Lincei. 2023;34:653–61.

Mark P, Nilsson L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. J Phys Chem A. 2001;105: 9954–60.

Mendels D, de Pablo JJ. Collective variables for free energy surface tailoring: understanding and modifying functionality in systems dominated by rare events. J Phys Chem Lett. 2022;13: 2830–7.

Müller C, Schlauderer G, Reinstein J, Schulz GE. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure. 1996;4:147–56.

Müller CW, Schulz GE. Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap5A refined at 1.9 Å resolution: a model for a catalytic transition state. J Mol Biol. 1992;224:159–77.

Pantoja-Uceda D, Arolas JL, García P, López-Hernández E, Padró D, Aviles FX, et al. The NMR structure and dynamics of the two-domain tick carboxypeptidase inhibitor reveal flexibility in its free form and stiffness upon binding to human carboxypeptidase B. Biochemistry. 2008;47:7066–78.

Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol. 1974;82:1–14.

Rickard M, Zhang Y, Pogorelov T, Gruebele M. Crowding, sticking, and partial folding of GTT WW domain in a small cytoplasm model. J Phys Chem B. 2020;124:4732–40.

Rickard MM, Zhang Y, Gruebele M, Pogorelov TV. In-cell protein-protein contacts: transient interactions in the crowd. J Phys Chem Lett. 2019;10:5667–73.

Russell PPS, Rickard MM, Boob M, Gruebele M, Pogorelov TV. In silico protein dynamics in the human cytoplasm: partial folding, misfolding, fold switching, and non-native interactions. Protein Sci. 2023;32:e4790.

Sakakibara D, Sasaki A, Ikeya T, Hamatsu J, Hanashima T, Mishima M, et al. Protein structure determination in living cells by in-cell NMR spectroscopy. Nature. 2009;458:102–5.

Samuel Russell PP, Alaeen S, Pogorelov TV. In-cell dynamics: the next focus of all-atom simulations. J Phys Chem B. 2023;127: 9863–72.

Serber Z, Dötsch V. In-cell NMR spectroscopy. Biochemistry. 2001; 40:14317–23.

Sittel F, Stock G. Perspective: identification of collective variables and metastable states of protein dynamics. J Chem Phys. 2018; 149:150901.

Smyth M, Martin J. X-ray crystallography. Mol Pathol. 2000; 53(8):14.

Speer SL, Zheng W, Jiang X, Chu IT, Guseman AJ, Liu M, et al. The intracellular environment affects protein-protein interactions. Proc Natl Acad Sci. 2021;118(11):e2019918118.

Stevens JA, Grünewald F, van Tilburg PM, König M, Gilbert BR, Brier TA, et al. Molecular dynamics simulation of an entire cell. Front Chem. 2023;11:1106495.

Tanaka T, Ikeya T, Kamoshida H, Suemoto Y, Mishima M, Shirakawa M, et al. High-resolution protein 3D structure determination in living eukaryotic cells. Angew Chem Int Ed Engl. 2019;131:7362–6.

Tucker MR, Piana S, Tan D, LeVine MV, Shaw DE. Development of force field parameters for the simulation of single- and double-stranded DNA molecules and DNA-protein complexes. J Phys Chem B. 2022;126:4442–57.

Wang Y, Sukenik S, Davis CM, Gruebele M. Cell volume controls protein stability and compactness of the unfolded state. J Phys Chem B. 2018;122:11762–70.

Williamson MP, Havel TF, Wüthrich K. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. J Mol Biol. 1985;182:295–315.

Yoo J, Kim JY, Louis JM, Gopich IV, Chung HS. Fast three-color single-molecule FRET using statistical inference. Nat Commun. 2020;11:3336.

Zerrad L, Merli A, Schröder GF, Varga A, Gráczer É, Pernot P, et al. A spring-loaded release mechanism regulates domain movement and catalysis in phosphoglycerate kinase. J Biol Chem. 2011;286:14040–8.

Zhang XJ, Baase W, Matthews B. Multiple alanine replacements within α-helix 126–134 of T4 lysozyme have independent, additive effects on both structure and stability. Protein Sci. 1992;1: 761–76.

Zhang XJ, Wozniak JA, Matthews BW. Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. J Mol Biol. 1995;250:527–52.

Zimmerman SB, Trach SO. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. J Mol Biol. 1991;222:599–620.

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: Liu Z, Grigas AT, Sumner J, Knab E, Davis CM, O'Hern CS. Identifying the minimal sets of distance restraints for FRET-assisted protein structural modeling. Protein Science. 2024;33(12):e5219. https://doi.org/10.1002/pro.5219