Variance of Atomic Coordinates as a Dynamical Metric to Distinguish Proteins and Protein-Protein Interactions in Molecular Dynamics Simulations

*Sanjoy Paul, Sri Rama Koti Ainavarapu, Ravindra Venkatramani**

Department of Chemical Sciences, Tata Institute of Fundamental Research, Mumbai-400 005, India

Keywords: Variance, Energy Landscape, Protein Dynamics, Molecular Dynamics Simulations, Sampling Convergence, Protein-Protein interactions, Ubiquitin Family Proteins

Corresponding Author

Ravindra Venkatramani (email: ravi.venkatramani@tifr.res.in)

Department of Chemical Sciences, Tata Institute of Fundamental Research, Dr. Homi Bhabha Road, Colaba, Mumbai 400005, Maharashtra, India

Abstract

Protein dynamics is a manifestation of the complex trajectories of these biomolecules on a multidimensional rugged potential energy surface (PES) driven by thermal energy. Today computational methods such as atomistic molecular dynamics (MD) simulations can describe thermal protein conformational changes in fully solvated environments over millisecond timescales. Despite these advances, a quantitative assessment of protein dynamics remains a complicated topic, intricately tied to issues such as sampling convergence and the identification of appropriate reaction coordinates/structural features to describe protein conformational states and motions. Here, we present the cumulative variance of atomic coordinate fluctuations (CVCF) along trajectories as an intuitive PES sensitive metric to assess the extent of sampling and protein dynamics captured in MD simulations. We first examine the sampling problem in model 1D and 2D PES to demonstrate that the CVCF when traced as a function of the sampling variable (time in MD simulations) can identify local and global equilibria. Further, even far from global equilibrium, a situation representative of standard MD trajectories of proteins, the CVCF can distinguish different PES and therefore resolve the resultant protein dynamics. We demonstrate the utility of our CVCF analysis by applying it to distinguish the dynamics of structurally homologous proteins from the ubiquitin family (ubiquitin, SUMO1, SUMO2) and ubiquitin protein-protein interactions. Our CVCF analysis reveals that differential side-chain dynamics from the structured part of the protein (the conserved β-grasp fold) present distinct protein PES to distinguish ubiquitin from SUMO isoforms. Upon binding to two functionally distinct protein partners (UBCH5A and UEV), intrinsic ubiquitin dynamics changes to reflect the binding context even though the two proteins have similar binding modes which lead to negligible (sub-angstrom scale) structural changes.

1. Introduction

Atomistic molecular dynamics (MD) simulations provide the ability to describe the thermal motions of solvated biomolecules over the picosecond to millisecond timescales[1–4]. The scale of atomic motions captured in MD range from large scale dynamics such as protein folding[5–10], assembly[11–14], and translocation[15–18] to more subtle structural changes associated with native state fluctuations[19–21] and protein-protein interactions[22–24]. However, a quantitative analysis of protein dynamics extracted from MD simulations remains challenging due to two primary bottlenecks. The first is the identification of atomic motions/reaction coordinates relevant for a particular functional or biophysical context which arises from the large number of coupled protein-solvent degrees of freedom.[25,26] The second pertains to sampling the underlying rugged multidimensional potential energy surface (PES) of proteins. Recent studies have revealed that true convergence of PES sampling in unbiased MD simulations cannot be quantitatively assessed for even the simplest of biomolecules. As such, only self-consistency checks on sampling quality for a given trajectory can be carried out[27–30]. On the other hand, while biased sampling techniques can achieve full sampling along specific reaction coordinates, they rely on identification of relevant collective degrees of freedom and are therefore constrained by the first bottleneck mentioned above.[31] Given the present challenges in understanding dynamics of solvated single proteins using MD, extending the technique to comparatively assess dynamics of different proteins and their complexes in functionally relevant contexts[32–34] is seriously limited. From this perspective, it is useful to explore scalable new metrics and analysis schemes which can help extract more quantitative information and insights on protein dynamics from MD simulations and which have the potential to further the scope of applying MD to biologically relevant contexts.

The variance of atomic coordinate fluctuations has long served as the basis for analyzing protein motions in MD trajectories[35–41]. For instance, in Principal Component Analysis (PCA), the variance-covariance matrix of atomic coordinate fluctuations from MD trajectories is diagonalized to yield orthogonal coordinates with zero covariance termed as principal components (PC). The PCs thus extracted can then be used to carry out quasi-harmonic analysis of MD trajectories yielding native state entropies[35,42,43] and slices of the protein PES[44–46]. PCs have also been used to create reduced dimensional descriptions of protein conformational motions[46]. Notably, Hess has examined both the variance of atomic coordinate fluctuations and associated PCs as measures to assess sampling in model harmonic landscapes and biomolecules.[41] In this study the promise of

variance to indicate convergence of biomolecular trajectories was noted but not explicitly demonstrated beyond the harmonic model potential. Other measures routinely used to analyze MD trajectories are Root Mean Square Fluctuations (RMSF) and Root Mean Square Deviations (RMSD) of atomic coordinates[47–49]. While the former is typically used to provide information on residue level mobility/dynamics at fixed trajectory lengths, the latter is used to track evolution of protein structure along an MD trajectory with respect to a reference structure. On a more quantitative footing, several other measures have been proposed to assess the sampling in MD trajectories such as number of clusters[50], block covariance overlap[41,51], configurational/structural decorrelation time[52], cluster entropy[29], conformational overlap[30] and density overlap[30]. Independently, Monte Carlo (MC) optimization based approaches have been employed to elucidate the global PES of proteins, nucleic acids and atomic clusters. Here, biomolecular structures and atomic phases were enumerated and classified in terms of dis-connectivity graphs[53]. However, such approaches do not capture the dynamics of molecules. Since MD simulations present a complementary approach to capture biomolecular dynamics, it would be desirable to have dynamical metrics for this technique which can sense features of the underlying PES. Linking existing measures which assess sampling in MD simulations to the underlying PES is non-trivial as their readout typically relies on parameters external to the dynamical system of interest. For instance, measures such as cluster entropy or conformational overlap rely on a RMSD based structure clustering parameter. Thus, here we examine a variance based metric which has no tunable external parameters and can provide an assessment of sampling and dynamics captured in MD simulations.

In this study, we introduce the Cumulative Variance of Coordinate Fluctuations (CVCF) along atomistic MD trajectories, as a dynamical metric to examine protein dynamics and sampling convergence in MD simulations. Using model 1D and 2D PES, we first show using MC simulations that CVCF, which traces over the fluctuations of protein atoms as a function of sampling coordinate (number of steps in MC and time in MD simulations), captures both local and global equilibria to distinguish the underlying PES of proteins. For both model PES and protein trajectories, we compare the information content present in CVCF traces with that obtained using cluster based measures to reveal conditions under which a consistent interpretation of data can be obtained. We specifically show that the CVCF metric can assess Boltzmann statistics along a

sampling coordinate. The metric can therefore can be used to define equilibrium measures of protein dynamics (such as spring constants denoting protein flexibility) from MD trajectories when locally converged sections of the trajectory are available. More importantly, we show that independent of convergence to either local or global equilibrium, the values and features of the protein CVCF-trace can provide a comparative assessment of the ruggedness and curvature of the underlying PES sampled by the protein along MD trajectories. Trends in CVCF therefore enable us to compare features of the PES across multiple protein systems using MD simulations. Further, CVCF can be readily decomposed to examine the dynamics for selected subsets of atoms (protein backbone, sidechains, active sites, or atoms at the interface of protein complexes) or for specific coordinates (distances, angles, torsions). We demonstrate some of these attractive features of a CVCF based analysis on multi-microsecond (μs) MD trajectories of structurally homologous ubiquitin family proteins which present a particularly striking example in nature wherein sequence changes and complexation which do not lead to prominent structural changes bring about dramatic functional consequences.

Proteins from the ubiquitin family, which include ubiquitin (Ub) and the small ubiquitin-like protein modifier (SUMO), regulate diverse cellular processes ranging from proteasomal degradation to DNA transcription, replication, and repair in eukaryotes[33,54–57]. These molecular tags alter protein biological states primarily through reversible covalent attachment to substrate lysine residues via an isopeptide bond (ubiquitylation/SUMOylation). Central to the regulatory role of ubiquitin and SUMO proteins as molecular tags in the cellular machinery, is their highly conserved secondary structure and tertiary β-grasp fold. Using a CVCF based analysis, we show that structurally homologous Ub and SUMO proteins can be distinguished on the basis of their distinct sidechain flexibilities in the conserved β-grasp structural fold. Further, we demonstrate that the intrinsic flexibility of Ub changes when the protein is attached to a Ub recognition motif (UEV), which recognizes ubiquitylated cargo, or to a E2 ligase (UBCH5A) which attaches Ub to protein substrates. Surprisingly, we find the changes in Ub dynamics induced by binding to be context sensitive responding in distinctive ways to UEV and UBCH5A which have overlapping binding modes and produce minimal sub-angstrom structural changes in Ub.

2. Results

2.1 CVCF captures both local and global equilibria in model PES

We first illustrate the basic features of CVCF analysis using model 1-D and 2D PES which are sampled using a standard MC algorithm with a Boltzmann weight (sec. 5.1 in Methods). For the diffusion of a particle in a 1D harmonic PES in the high friction limit, an analytical expression (sec. S.1 of ESI) for the time evolution of CVCF can be derived [41]. In this case, at short timescales as the particle samples the PES, the CVCF increases monotonically and ultimately saturates to the value of variance (Eqn. S7) associated with the curvature of the underlying PES (Fig. S1 in ESI). Thus a plateau in CVCF indicates convergence of sampling for a 1D harmonic well. Here, we will show that plateaus in CVCF also indicate convergence (either local or global) for multi-well anharmonic PES. First, we establish our MC protocol and associated CVCF-based analysis of sampling and systematic errors for a particle stochastically moving in a single harmonic well of spring constant k (in units of $k_BT/\text{Å}^2$) for which the final (analytical) converged value of the CVCF, $\sigma^2_{expected} = \frac{k_BT}{k}$ is known (see sec. S.2.1, Fig. S2 A-E of ESI). For this PES, we carried out 5 independent MC runs with identical starting coordinate (x = 0 Å) and an MC step size 100 times smaller than $\sigma_{expected}$. The resultant MC runs were then aligned at the zeroth step to trace the evolution of the average CVCF $<\sigma^2_{CVCF}>$ and standard deviation SD_{CVCF} as a function of the number of MC steps. For the parameters chosen, we find that the $<\sigma^2_{CVCF}>$ plateaus to a value close to $\sigma^2_{expected}$ within sampling error (Fig. S2 B) after 10^9 MC steps. While the SD_{CVCF} values reflect the sampling error for a given number of MC steps, the deviation of the final converged $<\sigma^2_{CVCF}>$ from $\sigma^2_{expected}$ indicates the systematic error in the trajectory. The former depends on the sampling variable (steps in MC runs and time in MD simulations) whereas the latter depends on the resolution of sampling (MC step size or time-step in MD simulations). These basic relations can be easily demonstrated by repeating the analysis by varying the step size, the number of sampling steps, and the number of independent MC runs (sec. S.2.1 of ESI). We next extended the same MC analysis for the sampling problem in an anharmonic 1D inverted Gaussian energy well (sec. S.2.1 of ESI). Here, we find again that the convergence of sampling is captured by a plateau in the $<\sigma^2_{CVCF}>$. However, the CVCF converges to a value (Fig S2) which lies intermediate to the variance for a harmonic fit to the surface (upper bound) and that corresponding to a second order (harmonic) truncation of the Gaussian PES (eqn. S14 of ESI) given by (σ^2_H) . As in the case

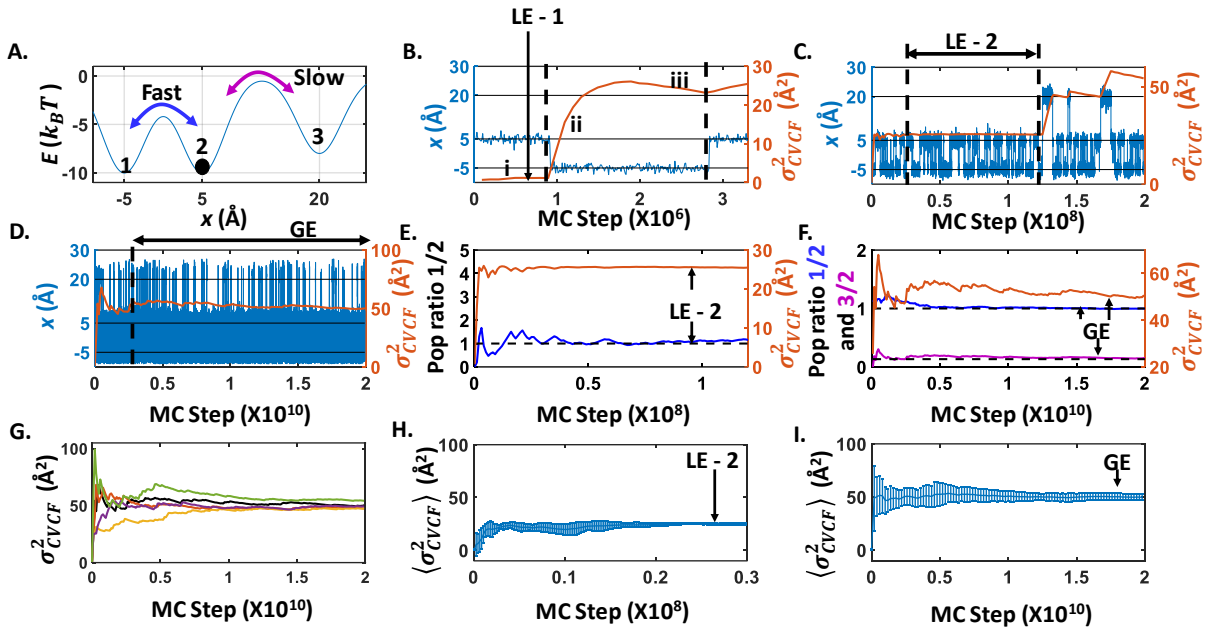

Fig-1: CVCF-trace analysis for 1-D 3-well model PES (A) Model PES depicting 2 closely spaced iso-energetic wells (1 and 2) with a kinetic barrier of ~6 k_BT and a third energetically higher well 3 placed slightly further away from well 2 than well 1 with a higher kinetic barrier of ~9.5 k_BT. (B-D) CVCF-trace is shown along with coordinate fluctuations as a function of MC steps at different resolutions along a specific trajectory. Sections of the trajectories producing distinct CVCF features are marked as regions i, ii, iii in panel B. In panels B and D LE indicates Local equilibrium which manifest as CVCF plateaus and are numbered according to their order of appearance during the MC run. In panel D GE denotes global equilibrium (E) Ratio of populations in the energy well 1 and 2 before it hops into energy well 3 during the initial section of the run attains an almost constant value ~1 as expected from Boltzmann statistics (black dotted horizontal line). (F) Ratio of population in the energy well 1 and 2 (blue) and that in 3 and 2 (purple) evolves to reach a plateau concomitant with CVCF showing expected Boltzmann statistics (black dotted line). (G) CVCF-trace evolution from 5 independent MC runs each initiated at the same position x = 5 Å. Black curve represents the trajectory shown in B-F. (H-I) Evolution of $<\sigma^2_{CVCF}>$ and SD_{CVCF} captured across the set of 5 independent MC trajectories.

of the 1D harmonic well, $<\sigma^2_{CVCF}>$ and s SD_{CVCF} for the anharmonic problem depend on MC step size and the number of MC steps. For a reasonably small choice of MC step size (about an order of magnitude smaller than σ_H), $<\sigma^2_{CVCF}>$ lies within a ~20 % offset from σ^2_H which represents the curvature of the bottom of the anharmonic PES (Fig. S2 F-H in ESI).

We next examine the sampling problem for a 1D PES with three Gaussian wells (Fig. 1A and sec. 5.2 in Methods) to illustrate the features of an anharmonic landscape with multiple kinetic barriers produced in a CVCF trace. Here energy wells 1 and 2 have equal depth (-10 k_BT) whereas energy well 3 was placed relatively higher (-8 k_BT). Energy wells 1 and 2 were separated by a

smaller energy barrier (~6 k_BT) and distance (10 Å) than that between 2 and 3 (~9.5 k_BT barrier and 15 Å separation). For a particle thermally sampling this 1D PES, transitions between energy wells 1 and 2 should occur more frequently than that between energy wells 2 and 3. Therefore we expect that a local equilibrium within the PES section covering wells and 2 will be established, before global equilibrium is established between all three wells. We carried out 5 independent MC runs to sample the 1D PES all starting with an initial position of the particle in energy well 2 (x = 5 Å). In Fig. 1B-D we follow one of these trajectories by monitoring the particle position and CVCF as a function of MC steps. Initially, in this trajectory the particle samples energy well 2 (region i in Fig. 1B) producing coordinate fluctuations around x = 5 Å and a CVCF which quickly attains a plateau indicating local equilibrium (LE-1) within well 2. The particle then transitions to well 1 as seen by a change in the position value to x = -5 Å and a large change in CVCF (region ii in Fig. 1B). As the particle continues to sample well 1 the CVCF shows a gradual decrease which is a signature of local trapping in the well (region iii in Fig. 1B). Subsequently, the CVCF shows a slight rise as the particle transitions back to energy well 2. With increasing number of transitions between wells 1 and 2, the modulations in CVCF become progressively smaller until a second local equilibrium is attained (LE-2 in Fig 1C). The value of the CVCF at this second plateau corresponds to an effective curvature of the PES section covering wells 1 and 2 (Fig. 1A) and persists till the particle transitions to energy well 3 showing coordinate fluctuations around x = 20 Å (Fig. 1C). The CVCF again senses the sampling of a new PES section (energy well 3) by producing a rise followed by progressively decreasing modulations as the particle samples all three wells (Fig. 1C). The CVCF ultimately tends to a third plateau which corresponds to the global equilibrium (GE) wherein all three wells are fully sampled (Fig. 1D). That each plateau in CVCF evolution indeed corresponds to either a LE or GE can be quantified by examining the time evolution of energy well populations along the MC trajectory to check if they follow Boltzmann statistics at the CVCF plateaus (Fig 1E-F). Since energy wells 1 and 2 have equal depth, their population ratio should be 1 according to Boltzmann statistics. For the given MC trajectory leading up to LE-2 (Fig 1E), the particle only samples energy wells 1 and 2 and the population ratio converges to 1 as the CVCF reaches the second plateau (the first CVCF plateau corresponds to equilibrium sampling of energy well 2). As sampling progresses further, the LE is destroyed by the discovery of energy well 3 which is indicated by a large change in CVCF (Fig. 1C). Finally, as the CVCF reaches the third plateau, population ratios of energy well 1/2 (blue curve in Fig. 1F)

and 3/2 (magenta curve in Fig. 1F) converge to values expected from Boltzmann statistics at GE (black dotted lines in Fig. 1E). This analysis shows that the CVCF trends as a function of sampling variable can provide an assessment of LE and discovery of new minima for a PES sampled in MC/MD trajectories. At this point it is useful to compare the information content obtained from CVCF analysis with that using other measures of sampling used for MD trajectories such as cluster number (N_c) and cluster entropy (S_c) [29]. For our model 1D PES, N_C trivially indicates the number of energy wells sampled along a trajectory (Fig. S4 B-C in ESI) and S_c tracks convergence in sampling each cluster/well (Fig. S4 E-F in ESI) along the MD trajectory. As revealed by this analysis, S_c contains the same information as CVCF (Fig. S4 D in ESI) provided the cluster definitions (set by the the clustering threshold *r*) are compatible with the PES. In each of the MC trajectories, the 1st half and 2nd half occupancy comparisons show very similar cluster populations again indicating convergence in sampling for these cases (Fig. S4 G-K in ESI). Thus, while N_c and S_c are dependent on the clustering threshold *r* which can be trivially set for the model PES here but is unknown for proteins, the CVCF does not rely on cluster definitions which makes it an attractive metric for assessing sampling of protein PES.

We next examine the heterogeneity in CVCF values generated under stochastic sampling conditions in Fig 1 G-I, using multiple MC trajectories. As for the case of single wells, we carried out 5 independent MC runs all initiated with identical positions of the particle (x = 5 Å). The CVCF evolution for each of the individual trajectories aligned at the first step is presented in Fig. 1G. The CVCF evolutions initially show a large sampling heterogeneity but ultimately converge to the same GE. Individual trajectories attain LE-2 at different number of MC steps and show plateaus of different lengths due to stochastic sampling of the underlying landscape (Fig. S3 in ESI). For instance, one of the trajectories (green line in Fig 1G) shows fairly steep CVCF increases with a relatively early transition from the LE-2 plateau as it starts sampling well 3 earlier than the other trajectories (Fig S3 of ESI). We note that such heterogeneities exist even for the sampling of simpler harmonic single well potentials (Fig S2 of ESI) and can be quantified (*vide supra*) in terms of systematic and sampling errors provided by $<\sigma^2_{CVCF}>$ and SD_{CVCF} monitored as a function of sampling variable for the 5 aligned trajectories. For instance, just before the earliest LE-2 transition we find $<\sigma^2_{CVCF}>$ = 24.5 Å^2 and SD_{CVCF} = 1.2 Å^2 (Fig. 1 H). Since $<\sigma^2_{CVCF}>$ at this point along the trajectory attains a plateau, we can consider this to be the averaged LE-2 for the set of 5 trajectories.

Subsequently all trajectories are expected to show the same CVCF plateau at the GE and we obtain $<\sigma^2_{CVCF}>$ = 50.5 and SD_{CVCF} = 3.5 Å^2 at the endpoint of the MC runs (Fig. 1 I). The analysis presented here can be extended to multiple dimensions. For instance, we examine the sampling problem using multiple MC runs for a 2D PES with 3 wells (sec S.2.3 of ESI). Here analogous to the 1D case (Fig 1A) the barriers are designed such that transitions between isoenergetic energy wells 1 and 2 are more frequent relative to that between either of energy wells 1/2 and the relatively higher placed well 3. Here again, the CVCF analysis of individual trajectories provides information on local as well as global equilibria as seen by tracking the evolution of coordinate and population ratios of energy wells as a function of MC step (Fig. S5 B-F of ESI). On the other hand, sampling heterogeniety and a collective view of the PES can be obtained by following $<\sigma^2_{CVCF}>$ and SD_{CVCF} along multiple trajectories aligned at the initial value of the sampling variable (Fig S5 G-H).

In summary, the CVCF plotted as a function of the sampling variable in MC/MD trajectories produces characteristic features which can be linked to the underlying PES. Plateaus in the CVCF clearly indicate equilibrium (or at least locally complete) sampling of a PES whereas increases or changes in CVCF indicate transitions to (or incomplete sampling of) new minima in the PES. An increase in CVCF with number of steps indicates the discovery/sampling of new minima/conformational states whereas subsequent progressively smaller increase/decrease in CVCF indicates that a new equilibrium is being established. CVCF values at plateaus (Fig 1) can be related to physically meaningful measures such as spring constant values (*vide infra*) denoting the effective curvature of the section of PES sampled.

2.2 CVCF distinguishes rugged energy landscapes of varying curvatures

Here, we examine the ability of CVCF to distinguish rugged PES which possess multiple minima separated by varying kinetic barriers. As discussed in the last section, heterogeneity arising from stochastic sampling conditions and the incomplete sampling of the PES both present challenges in interpreting CVCF features. Our goal in this section is to inquire whether a CVCF analysis can provide useful insights even when both of the aforementioned limitations are present in a given set of trajectories. For this purpose, we constructed a PES combining 7 inverted Gaussian wells (see sec. 5.2 of Methods and sec. S.2.4 of ESI). The Gaussian wells all have equal width but are well separated and placed at different depths which vary quadratically as a function of the

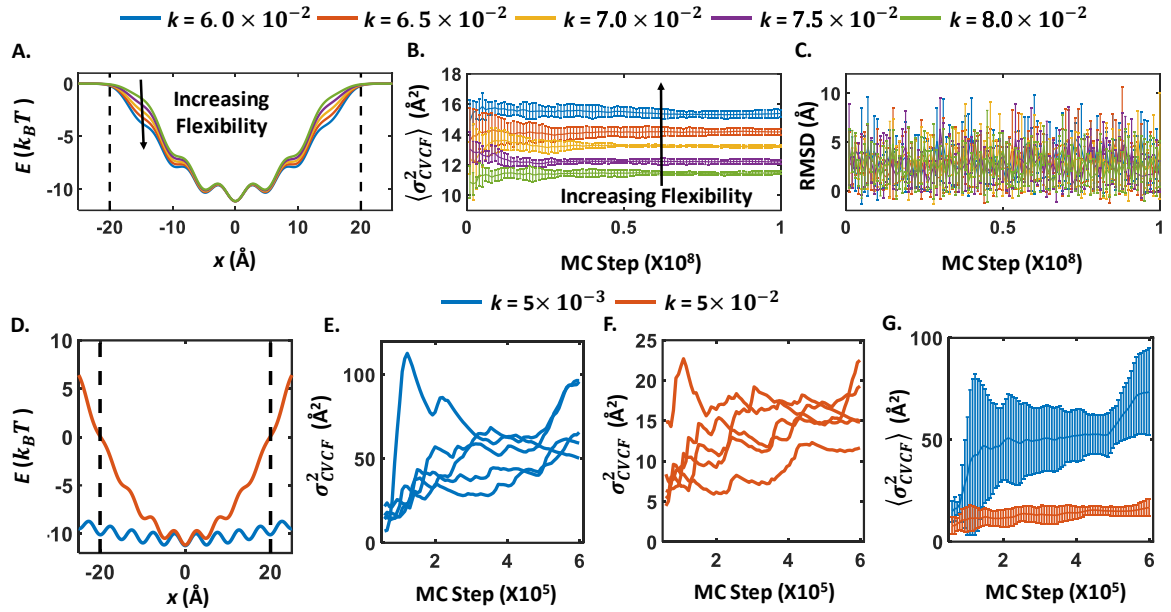

Fig-2: CVCF can distinguish PES of varying curvature. (A) 1D rugged harmonic PES of the form given by eqn. 1.2. The overall curvature of the harmonic surface is tuned by the spring constant k=6 × 10^{-2} (cyan) to 8 × 10^{-2} (olive green) $k_BT/\text{Å}^2$ with an resolution of 5 × 10^{-3} $k_BT/\text{Å}^2$ while other parameters and the minima of the PES are kept fixed across the landscapes. (B) CVCF-trace and (C) RMSD evolution for particle positions sampled across 5 independent MC runs all initiated at the PES minima (colour-coded energy landscape wise). MC sampling is restricted within a range of -20 Å to + 20 Å (marked by black dotted line in A). (D) 2 PES are chosen which differ in their k by an order of magnitude (5 × 10^{-3} versus 5 × 10^{-2} $k_BT/\text{Å}^2$) (E-F) Individual CVCF-traces (color coded) obtained for 5 different MC runs to sample both the landscapes in panel D. (F-G) CVCF evolution distinguishes the PES in panel D at 6 × 10^5 MC steps although for neither of landscapes GE has been attained.

position coordinate around the PES global minimum. The resultant PES can be viewed as a single harmonic well with rugged features created due to the presence of Gaussian modulations (Fig 2A). The curvature k of the PES controls both the equilibrium population ratios as well as the kinetic barriers associated with the energy wells (sec. 5.2 of Methods and sec. S.2.4 of ESI). Sampling this PES using MC, the CVCF-trace as a function of MC step should show a plateau at GE with a value of $1/k$ in the limit of zero systematic and sampling errors and sense changes in k with arbitrary precision. However, under realistic sampling conditions finite errors are expected to limit the resolution of CVCF. Ignoring systematic offsets, the resolution in the CVCF estimate of spring constants is given by $\partial k = \frac{\partial \sigma^2_{CVCF}}{(\sigma^2_{CVCF})^2}\ k_BT/\text{Å}^2$, where $\partial\sigma^2_{CVCF}$ is the sampling error in CVCF (sec S.2.4 of ESI). For instance, using a set of 5 MC runs (10^8 steps each with a step size 0.1 Å) to

sample the PES with $k = 6 \times 10^{-2}\ k_BT/\text{Å}^2$ we obtained a CVCF plateau at GE of 15.4 ± 0.3 Å^2 (blue lines in Fig 2B), which lies within a 10 % offset from the expected value ($1/k$ = 16. 7 Å^2) . For these MC and PES parameters, the CVCF can resolve PES with $\partial k = 1 \times 10^{-3}\ k_BT/\text{Å}^2$ at GE. We explicitly demonstrate the resolution of CVCF on a set of 4 PES with *k* values in the range 6×10^{-2} to $8 \times 10^{-2}\ k_BT/\text{Å}^2$ and differing by $5 \times 10^{-3}\ k_BT/\text{Å}^2$ (Fig. 2A). A set of 5 MC simulations were performed for sampling each of the 5 PES and $<\sigma^2_{CVCF}>$ with SD_{CVCF} across the 5 trajectories were plotted as function of number of MC steps. In each case, a CVCF plateau corresponding to GE is obtained after 10^8 MC steps and CVCF clearly distinguishes all the energy landscapes (Fig. 2B), with $<\sigma^2_{CVCF}>$ decreasing with increasing *k* (Eqn. 5). On the other hand, for a set of PES with *k* values differing by less than the threshold resolution (Fig. S6) set by the sampling error, the distribution of CVCF traces overlap and the underlying PES cannot be distinguished at these MC sampling durations (Fig. S6 of ESI). We note that in stark contrast, commonly used measures of equilibration such as RMSD fail to distinguish any of the set of PES and convey no information on extent of sampling (Fig. 2C and Fig S6 of ESI).

We next examine the value of CVCF analysis far away from GE. As discussed in the previous subsection, the early MC step evolution of CVCF bears signatures of the ruggedness of the PES arising from Gaussian energy wells. We demonstrate this feature using sets of MC runs with varying step sizes (10^{-1} - 10^{-3} Å), to sample PES with *k* values of 6×10^{-2} and $8 \times 10^{-2}\ k_BT/\text{Å}^2$ (Fig. S7 of ESI). For each of the cases, the individual CVCF traces are heterogenous showing multiple plateaus and rise/decay features which sense the LE and transitions between the multiple wells of the PES. Clearly the sampling heterogeneity in the CVCF traces far from GE is too large to be able to resolve the two PES. However, the CVCF can distinguish PES at early timescale well before GE is attained if separation of *k* becomes large relative to the ruggedness present in the PES. For instance, sampling of two PES with *k* values differing by an order of magnitude (Fig 2D) in short MC trajectories far from GE produces CVCF traces which are drifting to different extents and show differing amounts of sampling errors (Fig 2E-G). Nevertheless, these two CVCF-traces are well resolved even for these MC trajectories which are far from attaining GE (Fig. 2E-G). Thus, the CVCF traces can resolve PES curvatures far from GE and the resolution of the CVCF improves as the sampling approaches GE.

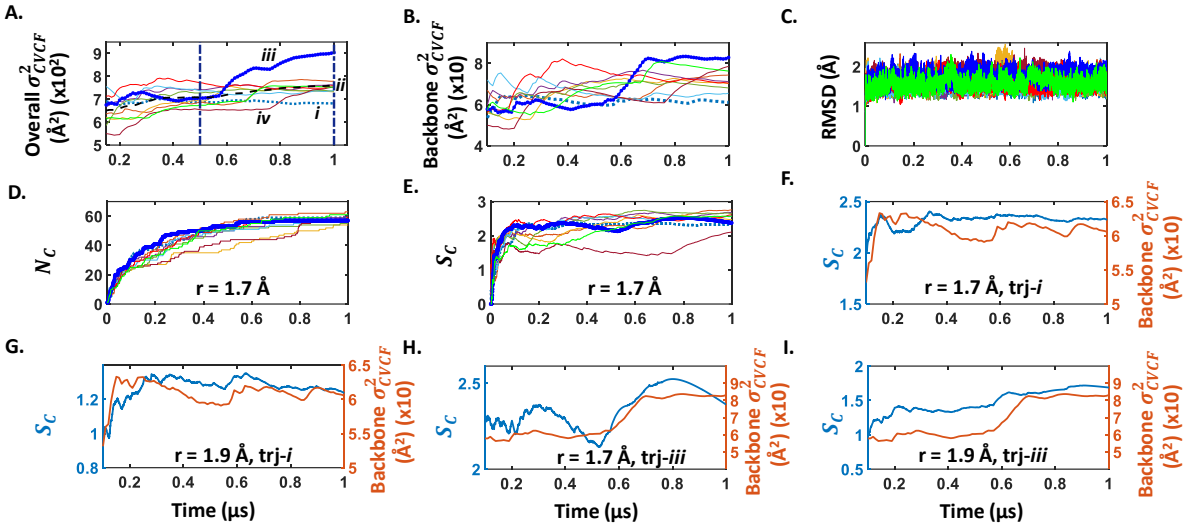

Fig-3: CVCF evolution in protein MD trajectories: (A-B) Individual CVCF–traces (all heavy atoms in panel A and backbone heavy atoms in panel B) for 10 independent μs MD trajectories of Ub. The dashed cyan line (labelled as *i*) represents CVCF evolution in the most stable section of the PES for the set of 10 trajectories and the black dotted line represents the average CVCF across 10 trajectories. (C) RMSD of protein structures (all heavy atoms) obtained along 10 independent μs MD trajectories relative to their starting structures (first frame of each trajectory). (D) Evolution of N_c and (E) S_c along each of the set of 10 MD trajectories using a RMSD clustering threshold of r=1.7 Å. (G-I) Comparison of CVCF-trace evolution with that of S_c for 2 specific MD trajectories i and iii using 2 different values of the clustering threshold r .

2.3 CVCF analysis reveals features of PES sampled in microsecond MD Trajectories of ubiquitin

The CVCF analysis presented in previous two sections can be extended to MD simulations of proteins where the sampling variable is the trajectory time. To demonstrate this, we carried out 10 independent 1 μs NVT MD trajectories of ubiquitin (Ub) in water, all initiated from the same phase space point (same coordinates and velocities) following an initial equilibration protocol (sec. 5.3 and 5.4 in Methods). In analogy with the analytical framework introduced for MC simulations, we plot the CVCF trace ($<\sigma^2_{CVCF}>$ along with standard deviation SD_{CVCF}) of all Ub heavy atoms captured along the 10 trajectories aligned at their initial time-points (Fig 3A). The trends in the CVCF are predictably heterogeneous revealing the complexity of the multi-dimensional PES even for small globular proteins such as Ub. Nevertheless, it is encouraging that the same features observed in MC trajectories of model PES are observed in the CVCF traces of Ub MD trajectories. While no two CVCF traces show completely overlapping trends, we note that 8 out of 10 trajectories attain plateaus within 800 ns or earlier (trajectory *i* and trajectory set *ii* in Fig 3A). The

exceptions are trajectories *iii* and *iv* which initially show LE plateaus (for instance around 500 ns) but later show diverging CVCF traces indicating the discovery of new minima at the end of the 1 μs runs. At the other extreme, trajectory *i* shows a CVCF plateau which extends over the entire 1 μs indicating that the protein is trapped in a single deep well which represents the global minima for our set of 10 trajectories. All trajectories, barring trajectory *i*, exhibit multiple CVCF LE plateaus alternating with steep increases during the 1 μs runs. For a similar CVCF-trace analysis of only the Ub backbone heavy atoms, we find that the variance drops by an order of magnitude. However, the backbone CVCF also displays heterogeneity with roughly 8 out of 10 trajectories attaining an LE plateau by 600 ns which persists for the rest of the trajectory. In contrast, the commonly used RMSD of protein heavy atoms (with respect to their position in the first frame) over the same set of trajectories shows highly overlapping trends (Fig. 3C) which converge within the first few nanoseconds (ns). We next compared the Ub backbone dynamics as reported by CVCF with cluster analysis measures[29,30] N_c and S_c (Fig. 3D-E and Fig. S8 of ESI) with RMSD clustering thresholds ranging from r=1.5-2.1 Å (see sec. 5.5 in 'Methods'). In Fig 3D with r=1.7 Å the N_c values initially increase rapidly but later (after 800 ns) saturate to a value around 50 for all 10 MD trajectories. However, the interpretation of the Ub trajectories using the CVCF-analysis is seen to differ from that using the cluster measures with r=1.7 Å.[29] For instance, the increase in N_c values over the first 800 ns indicates a continuous discovery of new clusters over this time period which is at odds with the CVCF plateau observed for trajectory *i* which suggests an LE during the same period (Fig 3B). Further, Unlike the case of our model PES calculations (sec 2.1) the CVCF evolution was not found to consistently mirror that of S_c captured for r=1.7 Å (Fig 3E). While CVCF has no tunable parameters, the trends in both N_c and S_c are sensitive to the clustering threshold r. Even for an relatively modest increase for r in the range 1.5-2.1 Å, N_c values at the end of the 1 μs drop from around 200 to 5 and the features of S_c change dramatically. In fact, the agreement between CVCF and S_c trends for individual Ub trajectories appear to be dependent on threshold r (Fig. S9-S12 of ESI). For instance, the best agreement (within the range of r values considered here) between the trends in CVCF and S_c for trajectory *i* is found at r = 1.7 Å, whereas for trajectory *iii* it is r = 1.9 Å (Fig 3F-I). We further quantified the extent of Ub conformational space self-consistently explored in the set 10 MD trajectories of Ub using conformational overlap (O_{conf}) and density overlap (O_{dens}) measures as described by Nemec and Hoffman[30]. For clustering thresholds of r = 1.5-1.9 Å, more than 80 % conformational space of Ub has been explored at 1 μs

timescale as indicated by O_{conf} in Fig. S13 (left panel) of ESI. Further, for the range of clustering thresholds considered, the 10 trajectories cover the Ub conformational space (explored within the set) with a probability of 60-90 % as indicated by O_{dens} in Fig. S13 (right panel) of ESI. While these trends convey reasonable fulfilment of the self-consistency checks within the set of 10 Ub trajectory, the sensitivity of the measures N_c, S_c, O_{conf}, O_{dens} to r remains problematic. These metrics yield contrasting interpretations on the extent of sampling and self-consistency checks on a set of trajectories depending on the choice of r (Fig. S13). In contrast, the CVCF offers a parameter free analysis of the set of 10 Ub trajectories, conveying the extent of sampling and heterogeneity associated with the underlying landscape over a specific (1 μs) MD timescale.

The CVCF features captured in the multiple μs trajectories for Ub (a rigid globular protein) demonstrate its sensitivity to the protein PES and enable us to propose intuitive yet robust measures of protein dynamics. Assuming all trajectories are initiated from the same equilibrated phase space point so that their time-points are aligned at the beginning of the production runs, a quasiharmonic spring constant (Eqn. 5) for a protein can be estimated from the $<\sigma^2_{CVCF}>$ values (dotted black trace in Fig. 3A) computed over independent trajectories, provided consistent equilibria is attained over the set. As discussed above, the CVCF values at plateaus (vertical dotted lines in Fig 3A) denote LE (locally complete sampling) and therefore provide physically meaningful spring constant (k_{avg}) values denoting the effective curvature of the PES. For instance, at 500 ns we obtain a consistent set of LE plateaus (Fig 3A vertical dotted lines) for all Ub trajectories so that an average curvature can be defined at this timescale. On the other hand, most but not all trajectories show plateaus at 1 μs which puts constraints on the interpretation of the $<\sigma^2_{CVCF}>$. On the other hand, the SD_{CVCF} values over a set of the trajectories indicate sampling heterogeneity and therefore provide a measure for the ruggedness of the PES sensed by MD over a specific timescale. We also note that independent of convergence to LE/GE, the values and features of protein CVCF provide an assessment of the ruggedness and curvature of the underlying PES sampled by the protein along MD trajectories (see sec 2.2). Trends in CVCF therefore, enable us to compare features of the PES across multiple protein systems as demonstrated in the following sections. Further, using CVCF for selected subsets of atoms or for specific coordinates (distances, angles, torsions), the dynamics of individual protein components (backbone, sidechain, active sites) can be readily decomposed and analyzed. In the following sections we demonstrate the utility of the CVCF analysis by

applying it to study how the flexibility of ubiquitin family proteins is altered by changes in protein sequence or complexation state.

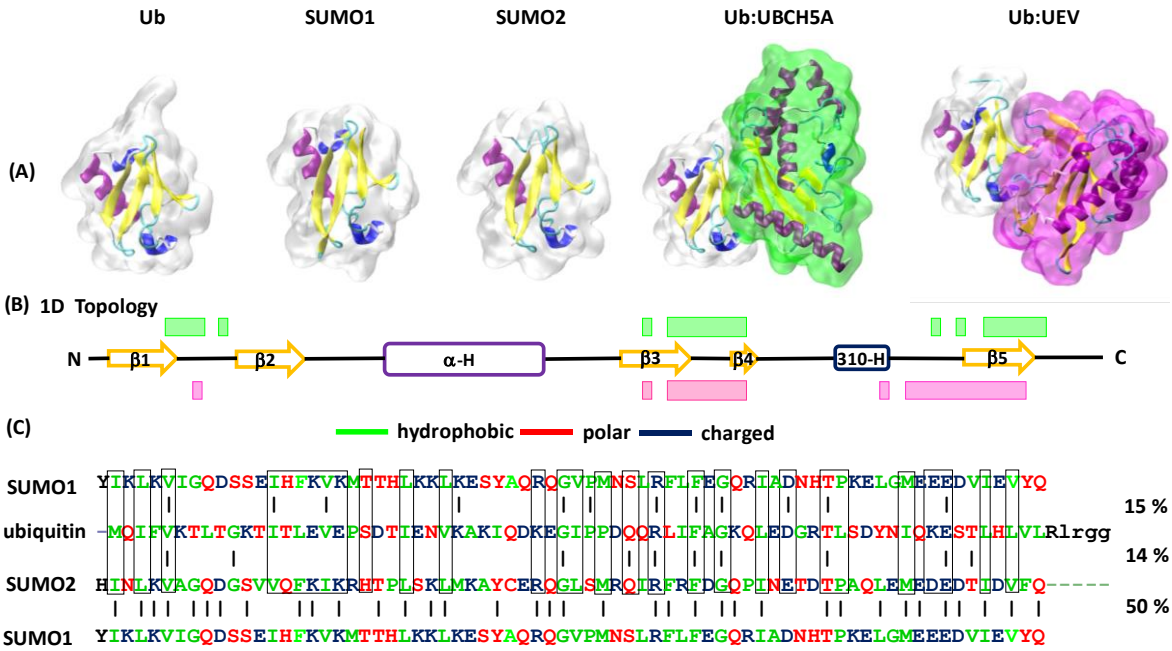

Figure 4: Structural properties and primary sequence information for Ub, SUMO1, and SUMO2. (A) Structures of Ub, SUMO1,SUMO2 and non-covalent complexes of Ub with two protein partners (UEV and UBCH5A). Backbone secondary structure representation shows the conserved β-grasp fold for ubiquitin family proteins. Sidechains are shown using a surface representation. (B) 1D topology for Ub displaying the secondary structure elements: β-sheets (arrows), coils (lines), and helices (cylinders) and the regions contacted by UEV (pink boxes) and UBCH5A (green boxes) along the sequence. (C) Sequence alignment and homology of the core 72 residues forming the β-grasp fold(capitalized) in Ub (1-72), SUMO1 (21-92), and SUMO2 (17-88) based on residue identity (vertical black dashes represent residues that are identical in a pair of proteins) and similarity (boxes indicate residues that are similar across all three proteins). Additionally, amino acid codes in the sequences are colored based on type (green-hydrophobic, red-polar, and blue-charged).

2.4 Sidechain Dynamics of the Structured β-grasp Fold Distinguishes Ub and SUMO Proteins

In this section, we assess the ability of CVCF to capture differences in dynamics of structurally homologous ubiquitin family proteins which include Ub, SUMO1 and SUMO2. The backbone structures of all three proteins are superimposable (RMSD ~ 3.4 Å), comprising of 5 β-strands and a single α-helix arranged in a signature β-grasp fold (Fig. 4A). Interestingly, despite strong structural similarities, the Ub and SUMO proteins have low sequence identity (Fig. 4C), ~20% between Ub and SUMO and ~50% between SUMO variants[58,59], which could distinguish

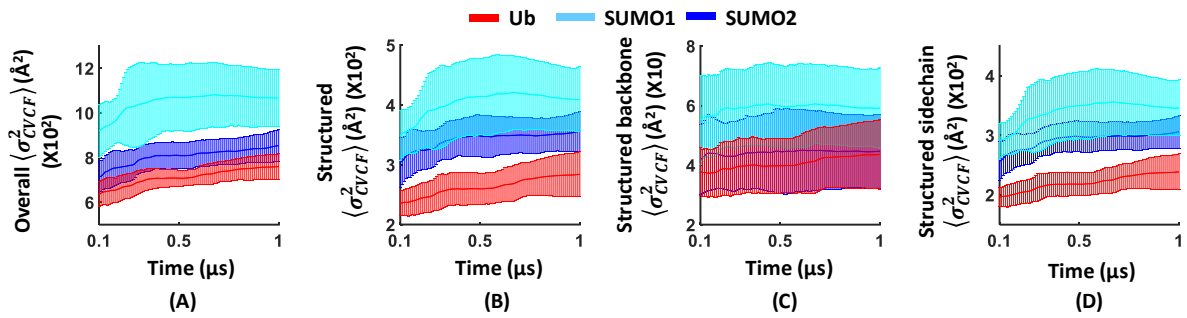

Fig 5: Comparison of the flexibility of ubiquitin family proteins based on CVCF derived from 10 independent μs trajectories. We plot $<\sigma^2_{CVCF}>$ and SD_{CVCF} as a function of time across 10 independent MD trajectories in panels A-D for different subsets of protein atoms. (A) Overall (heavy atom) CVCF, (B) Structured CVCF for heavy atoms in the structured β-grasp region, (C) Structured backbone CVCF for backbone heavy atoms in the structured β-grasp region, (D) Structured sidechain CVCF for sidechain heavy atoms in the structured β-grasp region. Plots of individual CVCF-traces for the set 10 trajectories of Ub and SUMO proteins are shown in the ESI Fig S14. Corresponding normalized CVCF trends in ESI Fig S15.

their cellular role. Recent single-molecule force spectroscopy (SMFS) experiments have suggested that sequence changes and ligand binding both impact the stability and flexibility of ubiquitin family proteins when probed along a specific (N-C termini) direction[60,61]. In this section, we demonstrate that the sequence variations across ubiquitin family proteins which produce minimal changes in structure lead to significant changes in protein dynamics. To this end, we carried out 10 x 1 μs NVT simulations each for solvated SUMO1 and SUMO2 using the same protocol as that for Ub (sec 5.3 in Methods).

As described in the previous sections, we aligned the MD trajectories of the Ub and the SUMO proteins at the same initial equilibrated phase space point and monitored the heavy atom $<\sigma^2_{CVCF}>$ and SD_{CVCF} over the μs timescale (Figure 5A and Fig S14). We find that the overall heavy atom SUMO1 CVCF is well-resolved and distinctly greater than that for Ub and SUMO2 which show a large CVCF overlap. In contrast, the heavy atom CVCF of the conserved β-grasp fold (structured segments excluding β4 as defined in Table S7 of ESI section S.8), show Ub to be distinctly less dynamic (lower CVCF) than the SUMO proteins (higher CVCFs which overlap) on μs timescales (Fig. 5B). We further decomposed the structured region CVCF into backbone (Fig. 5C) and sidechain (Fig. 5D) components. The CVCF of the protein backbone in the signature β-grasp topology is similar and indistinguishable for ubiquitin family proteins in terms of μs protein dynamics. In stark contrast, the sidechain CVCF for the same β-grasp topology cleanly

distinguishes Ub from SUMO proteins (Fig 5D). The trends shown in Fig 5B are preserved even when the CVCF is normalized with respect to number of atoms (ESI Fig. S15). A further examination (ESI Fig. S16) of CVCF trends from individual structured elements (β1, β2, β3, β5 and α-helix) of Ub, SUMO2 and SUMO1 reveals that the dominant contributions to the structured region CVCF comes from the α-helix. In fact, the relative CVCF of individual secondary structure elements for ubiquitin family proteins vary, with only the α-helix and β2 motifs showing Ub to be less flexible than the SUMO proteins (ESI Fig. S16). To summarize, our CVCF analysis shows quantitatively that the poor sequence homology between Ub and SUMO proteins translates into distinctly different structured sidechain dynamics for these proteins.

2.5 Protein-Protein Interactions with different Binding Partners Create Distinct Changes in Intrinsic Ub Dynamics

Here we employ the CVCF analysis to investigate the extent to which Ub dynamics is modified by non-covalent interactions with other protein partners. To this end, we compared Ub dynamics in free form with that when bound to two protein partners (the UEV domain of TSG101 and UBCH5A) which have distinct functional consequences. While the TSG101 UEV domain is a Ub recognition motif, recognizing ubiquitylated cargo, UBCH5A is an E2 ligase which attaches Ub to protein substrates. The Ub binding interface (Fig. 4A and 4B), RMSD changes, and solvent-accessible surface area (SASA) changes upon complexation (ESI section S.8 Tables S8-S9) indicate similar modes of Ub binding for UBCH5A and UEV with minimal structural perturbation relative to free Ub. Binding to UBCH5A and UEV reduces Ub SASA by ~29% and ~38% respectively (ESI section S.8 Table S9). Ub backbone RMSD changes (< 0.6 Å) upon complexation are also less than 15% of differences among ubiquitin family proteins (ESI section S.8 Tables S9 and S10).

Despite similar binding modes and negligible accompanying structural changes (Fig. 4), a CVCF analysis reveals distinct changes in Ub dynamics upon binding to UEV versus UBCH5A (Fig. 6). While the overall heavy atom dynamics of Ub distinctly reduces upon attachment to

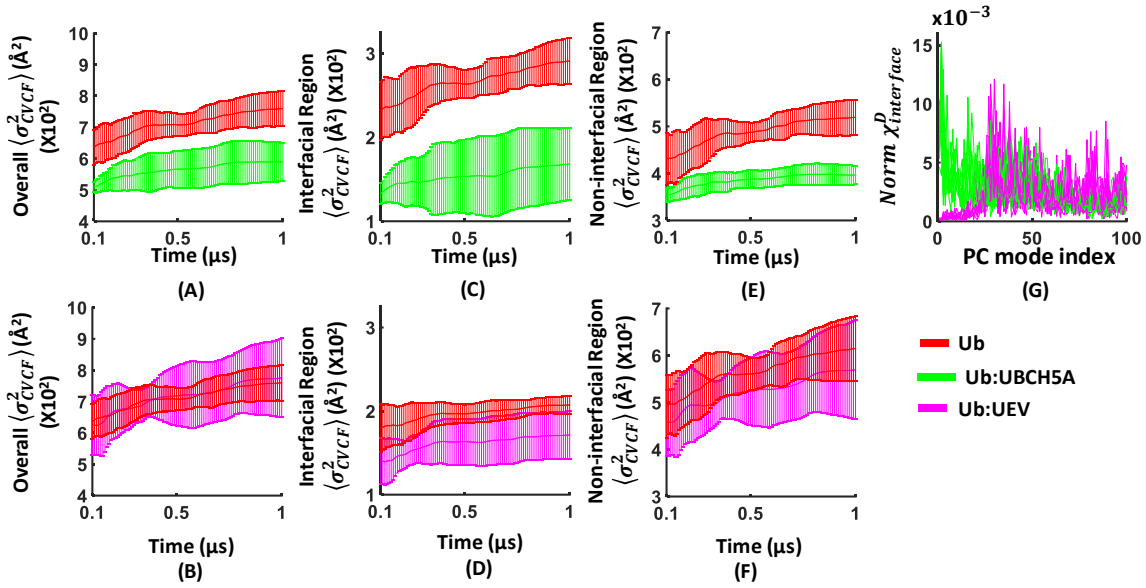

Fig 6: Ub $<\sigma^2_{CVCF}>$ and SD_{CVCF} changes upon binding to protein partners: (A-B) Overall (heavy atom) CVCF for free Ub, Ub:UBCH5A and Ub:UEV, and (C-F) CVCF of Ub heavy atoms at (interfacial regions) and away (non-interfacial regions) from the binding interface in Ub:UBCH5A and Ub:UEV complexes. Panels A-F also show the $<\sigma^2_{CVCF}>$ and SD_{CVCF} of the corresponding regions in free Ub (red). To generate the data shown, Ub atoms within 5.4 Å of the protein partner were considered to be part of the interfacial region (data for other cut-offs in ESI sec. S.5). (G) Normalized (per atom) contribution of interfacial Ub atoms to top 100 lowest frequency PC modes from 10 independent μs trajectories of free Ub. We plot contributions of interfacial atoms which are distinct ($\chi^D_{interface}$) in the UBCH5A and UEV complexes (sec S.8 Table S11). Corresponding plots for all interface atoms and for different interface definitions by varying the contact cutoff are shown in ESI Fig. S17 and Fig. S18

UBCH5A (Fig. 6 A), it does not change upon binding to UEV (overlapping CVCF evolution in Fig 6 B). We next dissected the CVCF trends (Fig. 6 C-F) to examine the effect of UEV/UBCH5A attachment on the dynamics of interfacial (defined in ESI section S.14. Table S5) and non-interfacial Ub heavy atoms. In Ub:UBCH5A, the dynamics of both interfacial and non-interfacial Ub atoms is distinctly lower than that of free Ub (Fig. 6C and 6E). In contrast, UEV binding induces very little change in the dynamics of Ub interfacial atoms and virtually no change in that for non-interfacial atoms (Fig. 6 D-F). The trends reported above are insensitive to variations in the definition of protein-protein interface (Fig. S17). Thus, UEV binding subtly changes equilibrium Ub dynamics only locally at the binding interface, while UBCH5A binding dramatically changes Ub dynamics globally throughout the protein. Notably, both dynamical responses in Ub to UEV and UBCH5A binding occur with minimal perturbation to Ub structure and compactness as indicated above.

We hypothesized that the local versus global extent of changes in protein dynamics created upon binding should be determined by how well connected the binding interface is to the rest of the protein structure. To validate this hypothesis, we computed the contribution of the Ub interface ($Norm\ \chi^{D}_{interface}$, normalized with respect to number of atoms) with UEV/UBCH5A to the PC modes captured in μs trajectories of free Ub (Fig. 6G and Fig. S18 of ESI). Our rationale is that protein atoms with global (long range) connectivity in the protein scaffold should also show significant contributions to the slowest (largest variance) PC modes of free Ub. Indeed, we find that the interface of Ub with UBCH5A is much better connected to the rest of the protein fold, showing larger projections on the global PC modes of free Ub, relative to the interface specific to UEV (Fig. 6G). Thus, Ub dynamics changes upon binding to UEV are expected to be more localized to the interface relative to that upon binding to UBCH5A. The trends in $Norm\ \chi^{D}_{interface}$ remain unchanged by variations in the definition of protein-protein interface (Fig. S18 of ESI).

3. Discussion

The ability of techniques such as MD simulations to quantitatively describe protein dynamics is inherently linked to how well they sample the underlying PES. In this regard, connecting metrics which assess sampling convergence to features of the underlying PES is useful. Further, it is desirable that such metrics be free of parameters extrinsic to the dynamical system as they impose constraints on the interpretation of the data. For instance, metrics such as N_c, S_c, O_{conf}, and O_{dens} depend on the cut-off radius r used for RMSD-based clustering of protein structures. Trends in these metrics are sensitive to the choice of r (sec 2.3), which remains a system unknown[30]. The preceding sections showed that the CVCF is a parameter free dynamical metric which can be connected to features of the underlying PES. We have shown using model PES (sec. 2.1-2.3) that the plateaus (constant values) of a CVCF trace captured as a function of the sampling variable arise when Boltzmann statistics for populating the PES is complete indicating that equilibrium sampling (either LE or GE) has been attained. The CVCF values at such plateaus can be effectively related to the effective curvatures of the underlying PES. We have showed (sec 2.1) that for model PES, the CVCF contains the same information content as the cluster entropy measure S_c proposed by Sawle and Ghosh for assessing sampling in MD simulations if clusters definitions were

consistent with the minima in the PES (solid vs. dashed line in Fig. S4 D).[29] Indeed, the CVCF values at plateaus should be viewed as a measure of system entropy captured after complete sampling of a section of PES is achieved. The correspondence is formally equivalent for the case of a PES with a single harmonic well, in which case the CVCF value obtained for a particle`s position coordinate after attaining equilibrium represents the width of the probability distribution of finding the particle in the well. Further, the CVCF is completely determined in terms of system properties and has no tunable parameters which makes it attractive as a transferrable and objective metric to assess the sampling of proteins. Using multiple trajectories, as assessment of the ruggedness of a PES can be obtained in terms of the heterogeneity observed in CVCF traces of trajectories aligned at their initial phase space point (sec 2.2). Furthermore, such collective CVCF traces can distinguish between two PES even when the sampling is far from GE provided the differences in their curvatures are resolved over the ruggedness of the PES (sec 2.2). Thus, the CVCF-based analytical framework presented here can be used to assess differences in the dynamics of proteins/protein domains captured in MD simulations with an ability to easily tune the resolution and the scope of the analysis by altering the subsets of atoms. These features are demonstrated by an explicit application of the CVCF based analysis to the ubiquitin family proteins to show how these proteins dynamically differ from each other and as they interact with protein partners.

The variance of atomic coordinate is a very conventional and commonly used measure for the analysis of protein dynamics and it is useful to compare the CVCF-based analysis with the routinely used mean-square-fluctuation (MSF) analysis of protein trajectories[48]. Essentially, the CVCF at a single time-point T in the trace of an MD trajectory is the same as the mean-square-fluctuation for a given subset of atoms ($\sigma^2_{CVCF}(t{=}T)$ = MSF(T)). However, the conventional MSF analysis places no rigorous constraints on T, which as can severely limit the utility of this analysis. For instance, typically, the MSF analysis is carried out on a single trajectory of arbitrary length T, decided often by the length afforded by computational expense. Sometimes, an averaging is also carried out over fixed time windows over the trajectory where again the duration of the time windows is a free parameter[41]. In addition to producing results which are sensitive to choice of T, the former analysis completely disregards the heterogeneity present in sampling the PES. The latter analysis, on the other hand averages over dynamics occurring on multiple timescales with ambiguous weights unless a firm link is established between the time windows and equilibrium

sampling. The CVCF analysis on the other hand indicates clear sections of trajectory, plateaus for single protein trajectories or durations of well resolved dynamics for multiple proteins, where a MSF analysis may add further meaningful value in terms of dissecting the information. We demonstrate this feature by a MSF analysis on trajectories of ubiquitin family proteins (Fig. S19-S21 of ESI). As in the case of conventional MSF analysis, we present the fluctuations as a function of residue number. However, unlike the typical implementation which is C_α atom based, we present residue fluctuations which are summed over all atoms pertaining to the subset used in CVCF analysis (sec. 5.4 Methods). Further each set of fluctuations are presented at specific time-points along the microsecond trajectories which provide further information on the dynamics of ubiquitin family proteins as discussed below. We propose this analytical framework as a systematic way to present MSF data for proteins.

Our CVCF–based analysis clearly establishes protein dynamics as a molecular metric which is exquisitely sensitive to both the sequence and the complexation state of ubiquitin family proteins. The data presented show that changes in protein flexibility can arise directly from changes in sequence rather than from differences of the conserved 3D protein fold. A structural superposition of Ub and SUMO backbones (Fig. 4 and ESI section S.8 Table S10) reveals that the conserved β-grasp fold in the proteins is similar but not identical (RMSD ~3.4 Å). However, these differences do not lead to distinct backbone flexibilities for the structured regions of Ub and SUMO (Fig. 5C). Rather differences in sidechain heavy atom flexibilities from the same structured region are able to cleanly resolve ubiquitin family proteins (Fig. 5D). Thus, the different sequences of these proteins provide a means to distinguish protein flexibility through the sidechain packing while preserving the overall secondary structure and tertiary fold which marks their cellular function as molecular tags.

Our results also provide clear evidence of the changes in intrinsic protein flexibility during protein-protein interactions which can be either local or non-local depending on the binding partner. In this study we have examined distinct protein-protein interactions (Fig. 4A) for Ub with activating enzymes (UBCH5A) and with protein domains which sense these molecular tags (UEV). The UEV domain of human TSG101 is a Ub binding domain (UBD) involved in both HIV-1 budding and vacuolar protein sorting (VPS) pathways where it interacts with Ub tagged proteins as they are sorted into vesicles that bud into multivesicular bodies (MVBs)[62]. Our CVCF

analysis (Fig. 6) shows that effect of UEV attachment on Ub flexibility is subtle and only restricted to binding region (Fig 5B-F). Indeed, UBD-Ub interactions are typically very weak (K_d~10-500 μM)[63,64] serving to facilitate reversible UBD-Ub interactions and to prevent non-specific associations with the high cellular concentration (~10 μM for mammalian cells) of free Ub[63,65]. In contrast, for Ub interactions with the E2 enzyme UBCH5A which creates poly-Ub chains,[66] our CVCF analysis reveals dramatic decreases in Ub flexibility globally over the entire protein structure (Fig 6A-E). Previously, Wlodarski and Zagrovic have carried out a comparative analysis[67] between X-ray structures of Ub bound to protein partners and a reference "conformationally selected" free Ub structure from an NMR RDC-refined ensemble. The researchers found that the average RMSD of Ub atoms near the binding interface was larger than that of atoms in the rest of the protein which implies that the conformational changes were local to the binding interface[67]. In contrast, here we find both subtle local and dramatic global changes in intrinsic Ub dynamics, with negligible changes in protein backbone, upon binding with UEV and UBCH5A. We find that the Ub binding interface with the recognition motif UEV is weakly coupled (Fig 6G) with respect to the rest of the protein thereby influencing only Ub flexibility locally. In contrast, the Ub binding interface with the activating enzyme UBCH5A is strongly coupled to the rest of the protein structure leading to a global decrease in Ub flexibility.

4. Conclusions

We have introduced CVCF as a dynamical metric to monitor and assess sampling in MD simulations. Features of the protein CVCF-trace when monitored along a trajectory can provide an assessment of equilibrium sampling of a section of the PES or discovery of new regions of the PES. The CVCF-trace analysis depends only on system properties and does not rely on any tunable external parameters which makes it useful as an objective metric to study protein dynamics captured in MD simulations. Further, the resolution of the CVCF-based analytical framework can be easily tuned to assess comparative dynamics of arbitrary subsets of atoms within proteins and protein complexes. These features are demonstrated by an explicit application of the CVCF based analysis to the ubiquitin family proteins to show how these proteins dynamically differ from each other and as they interact with protein partners. Application of the CVCF-analysis on structurally homologous ubiquitin family proteins reveals that Ub differs from its SUMO homologs in terms

of sidechain dynamics arising from the structured β-grasp region. Binding of Ub to two functionally distinct protein partners with very similar binding modes was found to produce very different changes in Ub dynamics as revealed by the CVCF-analysis.

5. Methods

5.1 Monte Carlo Simulations: For different model PES (Fig. 1 and 2), we used the Metropolis-Hasting algorithm[68,69] to sample the coordinate of a single particle starting with a well-defined initial coordinate within a minima in the landscape. At every MC step, the coordinates were changed randomly within a range corresponding to the MC step size. The energy difference between the new and old positions were compared using a Boltzmann criterion to either accept or reject the move. If the new energy of the particle (E') was lower than that at the old position (E) then the move was accepted otherwise the probability (p) of move acceptance was calculated as:

$$p = e^{-(E'-E)/k_BT} \quad (1)$$

If this probability was higher than a random number between 0 and 1 then the move was accepted otherwise it was rejected. Details of different model PES considered and choice of MC step size and MC trajectory lengths are described below. Further details of specific MC implementations are presented in sec 2.2 of ESI

5.2 Model energy landscapes: We considered 5 different 1D and 2D model PES to demonstrate the features of CVCF. Here we only discuss the PES corresponding to the data shown in the main MS (Figs 1 and 2). Other landscapes and associated MC simulation parameters are described in sec. S.2 of ESI. For the 1-D 3 well energy landscape (Fig. 1A and sec S.2.2 of ESI) the PES has the form

$$\mathrm{E(x)} = -\sum_{i=1}^{3} a_i * \mathrm{e}^{-\left(\frac{\mathrm{x}-x_i}{\sigma_i}\right)^2} \quad (2)$$

where, $a_1 = a_2 = 10\ k_BT$, x_1 = -5 Å, x_2 = 5 Å, $\sigma_1 = \sigma_2 = \sqrt{16}$ Å, $a_3 = 8\ k_BT$, x_3 = 20 Å, $\sigma_3 = \sqrt{16}$ Å. Energy wells 1 and 2 were relatively closer (10 Å) and separated by small activation barrier (~ 6 k_BT). Energy well 3 is placed at x = 20 Å and separated by 15 Å from well 2 and a large activation barrier of ~ 9.5 k_BT. Hence, for these parameters, the transitions between energy wells 1 and 2 are expected to be faster than that between wells 2 and 3. This PES was sampled with a MC step size of 0.1 Å which is ~70 times smaller than the width position distribution function of the particle in each well (sec. S.2.1 of ESI). Sampling was restricted to lie between x = - 9 Å and

x =+ 27 Å in order to avoid random diffusion of the particle on the flat surface away from the minima. A total of 5 independent MC simulations were carried out for this PES all starting from identical initial position of the particle at x = 5 Å (energy well 2).

For the 1-D multi-well rugged energy landscape (Fig. 2A and sec S.2.4 of ESI) the PES has the form

$$\mathrm{E(x)} = -\sum_{i=1}^{7} a_i * \mathrm{e}^{-\left(\frac{\mathrm{x}-x_i}{\sigma_i}\right)^2} \quad (3)$$

where, $a_i = \left(10 - \frac{1}{2}kx_i^2\right)k_BT$ and $\sigma_i = 3$ Å for i=1-7 with x_1 = 15 Å, x_2 = 10 Å, x_3 = 5 Å, x_4 = 0 Å, x_5 = -15 Å, x_6 = -10 Å, x_7 = -5 Å,

The PES has multiple (total 7) inverted Gaussian wells embedded within a harmonic surface. Hence, such a landscape has features of local ruggedness controlled by the width of the Gaussian wells and a global curvature tuned by the spring constant k of the harmonic surface. Here we varied only the global curvatures by changing k and generated multiple surfaces for our MC simulations. In all cases, an MC step size of 0.1 Å was used and 5 independent runs of 10^8 steps were carried out. The σ^2_{CVCF} values obtained at 10^8 MC steps reproduces the global curvature of harmonic parabola (sec s.14 Table-S5). Sampling was restricted between - 20 Å to + 20 Å in order to avoid random diffusion of the particle on the flat surface away from the minimum.

5.3 Molecular Dynamics Simulations: All atom Molecular Dynamics (MD) simulations were carried out using GROMACS version 5.0[70] and the CHARMM36 force field[71,72] with an explicit solvent TIP3P water model[70]. Initial coordinates for Ub and SUMO2 proteins were from crystal structures in the protein data bank: PDB code: 1UBQ and 1WM3 respectively. As no apo SUMO1 crystal structure was available, we derived SUMO1 coordinates from the crystal structure of the protein in complex with a peptide (chain A from PDB code: 4WJQ). For the Ub:UEV complex we used crystal structure coordinates (Chains A and B of PDB code: 1S1Q) after mutating Selenomethionine residues in both chains to methionine. Finally, for Ub:UBCH5A, we used crystal structure coordinates of a complex (Chains A and C of PDB code: 3PTF) which contains 2 Ub and 2 UBCH5A molecules non-covalently bound in the unit cell. Although the second UBCH5A molecule (chain B) also makes contact with the Ub molecule considered, this interaction was found to be weak and to break quickly in short preliminary MD runs. We simulated the 76 residue sequence (1-76) for Ub and the core 72 residue sequence for SUMO proteins (SUMO1:

21-92 and SUMO2:17-88). We added hydrogen atoms to the protein structures and immersed them in a cubic solvent box with 10 Å padding (minimum distance between protein surface atoms and solvent box edge) along the box edges. The system was then neutralized (by adding Na^+ and Cl^- ions) and the coordinates of solvent atoms, ions and hydrogens were optimized (protein heavy atoms coordinates were fixed) using the steepest descent algorithm until the maximum force on atoms was less than 1000 kJ/mol/nm. Periodic boundary condition (PBC) were employed with the Particle Mesh Ewald (PME) method for long range electrostatic interactions[73]. A 1.0 nm short range van der Waals cutoff and 1.0 nm short range electrostatic cutoff were used in the subsequent equilibration and simulation steps. During equilibration, we first optimized the solvated protein systems, while keeping protein heavy atoms fixed, for 10000 steps using the steepest descent algorithm. Then the systems was equilibrated at 300 K temperature (heat equilibration) with protein heavy atoms fixed for 100 ps using the Nosé-Hoover[74,75] thermostat with a 0.5 ps relaxation time constant. Optimization and heat equilibration were repeated again with harmonic constraints of 25 kcal/mol/Å^2 on protein heavy atoms followed by NPT equilibration at temperature 300 K and Parrinello-Rahman[76] isotropic pressure coupling (relaxation time constant 2.0 ps) at 1 bar pressure for 150 ps. The NPT equilibration step was repeated two more times with harmonic constraints of 10 kcal/mol/Å^2 and 5 kcal/mol/Å^2 on protein heavy atoms successively. Finally, unconstrained NPT simulations were performed for 1 ns. The solvated protein coordinates and velocities obtained at the end of this step were used to seed 10 independent μs NVT production runs for each of the 5 protein systems (Ub, SUMO1, SUMO2, Ub:UEV, and Ub:UBCH5A) with the aforementioned temperature coupling method. Production trajectories sampled coordinates every 20 ps and were analyzed using VMD 1.9.1[77].

5.4 Calculation of CVCF and Quasi-Harmonic Spring Constants: The Cumulative Variance of Coordinate Fluctuations (CVCF) is an intuitive descriptor of protein dynamics which is highly sensitive to the underlying PES sampled by a protein in MD trajectories. The MD simulation trajectory provides coordinate of protein atoms as a function of time (snapshots/frames). If $x_i(t)$ is ith atomic coordinate at snapshot *t* along the MD trajectory, then the cumulative variance of coordinate fluctuations (CVCF) sampled in T_S snapshots/frames in a trajectory is defined as:

$$\sigma^2_{CVCF}(T_S) = \sum_{i=1}^{N} \frac{\sum_{t=1}^{T_S}(x_i(t)-\langle x_i(t)\rangle)^2}{T_S-1} = \sum_{i=1}^{N} \frac{1}{T_s-1}\sum_{i=1}^{T_s}\left(x_i - \frac{1}{T_s}\sum_{i=1}^{T_s} x_i\right)^2 \qquad (4)$$

where the index *i* runs over a subset of *N* protein atoms of interest (for instance subsets of heavy, backbone, and sidechain atoms, or interfacial and non-interfacial atoms in protein complexes) from the core 72 residue sequence of ubiquitin family proteins. For the two Ub complexes (Ub:UEV and Ub:UBCH5A), while atomistic simulations of entire complexes were carried out, the CVCF analysis was restricted to Ub atoms (residues 1-72 discarding the last 4 C terminal tail residues). For all protein systems (Fig. 1A) in the study, the CVCF was computed for specific subset of atoms as a function of T_S (sampling time ranging from 100ns to 1 μs). While applying Eqn. 1 for a given subset of atoms, rigid body translation and rotational motions were eliminated by aligning the coordinates of the subset in each frame of the trajectory to their positions in the first frame (t=0 timepoint). If the subsets of atoms for which CVCF is computed differs from that for alignment of frames in the MD trajectory, the resulting variance may include rigid body translations and rotation. While such inconsistent choices of alignment and analysis set do not seem to significantly impact the results presented here for ubiquitin family proteins (Fig S22-S23), it may be an issue for flexible protein systems. In the standard CVCF-trace analysis presented here, we plot average CVCF values along with their standard deviation from 10 independent μs timescale MD simulation trajectories (Fig 5A-D, and 6A-F). In the case of MC simulations of model PES (Figs 2 and 3) the CVCF simply represents the variance of a single coordinate (see ESI section S.2 for details).

The thermal fluctuations of a given subset of protein atoms along an MD trajectory can be described in terms of a single effective quasi-harmonic spring according to the equipartition theorem. Quasi-harmonic spring constants can be derived from the average CVCF values across multiple trajectories at specified time-points where the sampling was locally complete (dotted black lines at CVCF plateaus in Fig 3A).

$$k_{avg}(T_S) = \frac{k_B T}{\langle \sigma^2_{CVCF}(T_S) \rangle} \quad (5)$$

Where T is temperature (set to 300 K in this study) and k_B is the Boltzmann constant. The average variances $\langle \sigma^2_{CVCF}(T_S) \rangle$ can be computed across multiple independent μs trajectories at time-points T_S where consensus LE is attained (black dotted vertical lines in Fig. 3A for Ub). Trajectories for our protein systems showing least standard deviation of CVCF values (among the set of 10 trajectories) which sample the most stable sections of the PES accessible in our simulations are highlighted in various figures (Trajectory i in Fig 3A and thick purple lines in ESI Fig. S14).

5.5 Cluster analysis: For MC sampling of model 3 well PES in Fig. 1A, we demarcated 3 clusters associated with energy wells 1, 2 and 3. The coordinate evolution in each MC trajectory was examined and population of each energy well was counted (criteria: x lies within 2 Å around the energy minima) as a function of increasing MC steps cumulatively. The cluster entropy $S_C = \sum_{i=1}^{N_C} p_i \log(p_i)$ was computed as a function of MC steps (T_S) where N_C is the total number of clusters (here 3) and p_i is the probability of the ith cluster located by sampling up to T_S steps[29]. In case of Ub MD trajectories, we performed the computation of N_C and S_C using a python code provided by Nemec and Hoffman[30] considering protein heavy backbone atoms (no hydrogens) only. For each trajectory, we carried out cluster analysis with different values of the RMSD clustering threshold parameter (r) and the GLOBAL partitioning scheme. Here, two structures belong to the same cluster if their RMSD is less than r. For each trajectory, we carried out cluster analysis at four different value of r : 1.5 Å, 1.7 Å, 1.9 Å and 2.1 Å. In addition to N_C, and S_C, we also computed the conformational overlap (O_{conf})[30] and density overlap (O_{dens})[30] using the same python code provided by Nemec and Hoffman[30]. Using this code, for every (reference) trajectory frame $\kappa \in K$ the number of structures ($e_{r,\kappa l}$) of a certain trajectory $l \in L$ within the r-neighbourhood was monitored in EventCurves. Effectively, the events $e_{r,\kappa l}$ count the number of neighbours for the reference frames κ. With r as the threshold with RMSD(κ, x) ≤ r , which is used to count the number of events, which have an RMSD value smaller than or equal to r between the reference trajectory frame κ and any other simulation frame x of the corresponding trajectory l. $O_{conf}(K, L; r)$ provide information on the number of reference frames $\kappa \in K$ for which one can find at least one occurrence of all trajectories, normalized by the total number of reference frames n_κ ($O_{conf} \in [0, 1]$).

$$O_{conf}(K, L; r) = \frac{1}{n_\kappa} \sum_{\kappa \in K} \delta(\Pi_{l \in L} e_{r,\kappa l}) \,, \quad \delta(x) = \begin{matrix} 1 \; for \; x \neq 0 \\ 0 \; otherwise \end{matrix} \tag{6}$$

$O_{dens}(K, L; r)$ estimates how similar the probability density is for the r-neighbourhoods of different reference frames for different trajectories in the set L with $O_{dens} \in [0, 1]$. This metric can be defined by the average of the ratio between minimal and maximal events ($e_{r,\kappa l}$) over each reference frame κ for each reference trajectory $k \in K$ separately. The value for all reference trajectories K is then the sum over all overlap values for all individual reference trajectories $k \in K$ normalized by the number of reference trajectories n_κ.

$$O_{dens}(K, L; r) = \frac{1}{n_K}\sum_{K \in K}\langle\frac{\min\{e_{r,kl}\,;\, l \in L\}}{\max\{e_{r,kl}\,;\, l \in L\}}\rangle_k \quad (7)$$

For the set of 10 trajectories *L* the $O_{conf}(L, L, r)$ and $O_{dens}(L, L, r)$ measures were computed for five *r* values of 1.1, 1.3, 1.5 Å, 1.7 Å and 1.9 Å and for different durations of the sets of MD trajectories (ranging from 100 ns to 1 μs with an interval of 100 ns). Specifically, for a given time duration and *r* value, O_{conf} and O_{dens} were computed for all possible pairs ($\binom{10}{2}$ = 45) of the set of 10 MD trajectories of Ub of the specified duration with averages and standard deviations computed across possible trajectory pairs.

5.6 Principal Component Analysis: For an N atom system, Principal Component Analysis (PCA) yields a set of 3N-6 orthogonal principal components (PC eigenvectors) and their variances (the eigenvalues) by diagonalizing the 3Nx3N variance-covariance matrix of atomic coordinates with elements $\sigma_{ij}^2(T_S) = \frac{\sum_{t=1}^{T_S}(r_i(t)-\langle r_i(t)\rangle)(r_j(t)-\langle r_j(t)\rangle)}{T_S-1}$ [78,79]. Here < > indicates an average over the MD trajectory and $\boldsymbol{r_i}$ is the position vector of the i^{th} protein atom. The set of PC eigenvectors $\boldsymbol{\xi} = \{\boldsymbol{\xi_1}, \boldsymbol{\xi_2} \ldots \boldsymbol{\xi_{3N-6}}\}$ were sorted in descending order of their corresponding eigenvalues (variances $\boldsymbol{\lambda_i}$) : $\boldsymbol{\lambda_1} > \boldsymbol{\lambda_2} > \ldots\ldots \boldsymbol{\lambda_{3N-6}}$. We carried out PCA of protein heavy (non-hydrogen) atom fluctuations sampled in 10 independent 1 μs trajectories for bound Ub (Ub:UEV and Ub:UBCH5A). Only protein residues 1-72 from the conserved ubiquitin fold were included in the PCA. We used PCA of free Ub trajectories to estimate the connectivity of interfacial amino acid heavy atoms of Ub in the complexes Ub:UEV and Ub:UBCH5A to the rest of the protein matrix. Interfacial amino acids (lists provided in ESI sec S.8 Table S11) of Ub were defined in terms of a distance cut-off (ranging from 3.4-7.4 Å) for sidechain-sidechain contacts in the complex crystal structures. We define the connectivity of Ub atoms interfacing with protein partners (UBCH5A and UEV) to the rest of the protein structure in terms of their contributions to the slow PC modes from MD trajectories of free Ub which is obtained as follows: For a given MD trajectory of a N atom protein system an eigenvector/PC mode can be represented as an N-dimensional vector $\xi_i = (\xi_{i1}\ \xi_{i2}\ \ldots\ldots \boldsymbol{\xi}_{iN})$, where $\boldsymbol{\xi}_{ij}$ is the j^{th} atomic coefficient in PC eigenvector *i*. We determined the total contribution of the Ub interfacial atoms in complexes to the ith PC mode as $\chi_{interface}(i) = \sum \boldsymbol{\xi}_{ki}^2$, where the summation index *k* runs over the sets of interfacial atoms. In our calculations, these we considered

the normalized (with respect to number of atoms) contributions of $\chi^{D}_{interface}$ from sets of interfacial atoms (highlighted in ESI section S.8 Table S11) which are distinct for complexes of ubiquitin with UBCH5A and UEV (Fig. 6G and ESI Fig. S18).

Acknowledgements

S. P., R.V., and S.R.K.A. acknowledge financial support from Tata Institute of Fundamental Research (Department of Atomic Energy, India). The authors would like to thank Dr. Hemachandra Kotamarthi and Dr. Ranja Sarkar for helpful discussion on selection of protein complexes during the initial stages of our study.

References

(1) Lee, E. H.; Hsin, J.; Sotomayor, M.; Comellas, G.; Schulten, K. Discovery Through the Computational Microscope. *Structure* **2009**, *17* (10), 1295–1306.

(2) Noé, F. Beating the Millisecond Barrier in Molecular Dynamics Simulations. *Biophys. J.* **2015**, *108* (2), 228–229.

(3) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. To Milliseconds and beyond: Challenges in the Simulation of Protein Folding. *Curr. Opin. Struct. Biol.* **2013**, *23* (1), 58–65.

(4) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-Timescale Molecular Dynamics Simulations of Protein Structure and Function. *Curr. Opin. Struct. Biol*. **2009**, *19* (2), 120–127.

(5) Bowman, G. R.; Voelz, V. A.; Pande, V. S. Taming the Complexity of Protein Folding. *Curr. Opin. Struct. Biol*. **2011**, *21* (1), 4–11.

(6) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. Long-Time Protein Folding Dynamics from Short-Time Molecular Dynamics Simulations. *Multiscale Model. Simul.* **2006**, *5* (4), 1214–1226.

(7) García, A. E.; Onuchic, J. N. Folding a Protein in a Computer: An Atomic Description of the Folding/Unfolding of Protein A. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (24), 13898–13903.

(8) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. Challenges in Protein-Folding Simulations. *Nat. Phys.* **2010**, *6* (10), 751–758.

(9) Scheraga, H. A.; Khalili, M.; Liwo, A. Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. *Annu. Rev. Phys. Chem.* **2007**, *58* (1), 57–83.

(10) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Atomic-Level Description of Ubiquitin Folding. *Proc. Natl. Acad. Sci.* **2013**, *110* (15), 5915–5920.

(11) Zheng, W.; Tsai, M. Y.; Chen, M.; Wolynes, P. G. Exploring the Aggregation Free Energy Landscape of the Amyloid-β Protein (1-40). *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (42), 11835–11840.

(12) Carballo-Pacheco, M.; Strodel, B. Advances in the Simulation of Protein Aggregation at the Atomistic Scale. *J. Phys. Chem. B* **2016**, *120* (12), 2991–2999.

(13) Ma, B.; Nussinov, R. Simulations as Analytical Tools to Understand Protein Aggregation and Predict Amyloid Conformation. *Curr. Opin. Chem. Biol*. 2006, *10* (5), 445–452.

(14) Morriss-Andrews, A.; Shea, J.-E. Computational Studies of Protein Aggregation: Methods and Applications. *Annu. Rev. Phys. Chem.* **2015**, *66* (1), 643–666.

(15) Gumbart, J.; Schulten, K. Molecular Dynamics Studies of the Archaeal Translocon. *Biophys. J.* **2006**, *90* (7), 2356–2367.

(16) Di Marino, D.; Bonome, E. L.; Tramontano, A.; Chinappi, M. All-Atom Molecular Dynamics Simulation of Protein Translocation through an α-Hemolysin Nanopore. *J. Phys. Chem. Lett.* **2015**, *6* (15), 2963–2968.

(17) Zhang, B.; Miller, T. F. Long-Timescale Dynamics and Regulation of Sec-Facilitated Protein Translocation. *Cell Rep.* **2012**, *2* (4), 927–937.

(18) Zhang, B.; Miller, T. F. Direct Simulation of Early-Stage Sec-Facilitated Protein Translocation. *J. Am. Chem. Soc.* **2012**, *134* (33), 13700–13707.

(19) Kumawat, A.; Chakrabarty, S. Hidden Electrostatic Basis of Dynamic Allostery in a PDZ Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (29), E5825–E5834.

(20) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. Simultaneous Determination of Protein Structure and Dynamics. *Nature* **2005**, *433* (7022),

128–132.

(21) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Shaw, D. E. Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *J. Phys. Chem. B* **2016**, *120* (33), 8313–8320.

(22) Pan, A. C.; Jacobson, D.; Yatsenko, K.; Sritharan, D.; Weinreich, T. M.; Shaw, D. E. Atomic-Level Characterization of Protein–Protein Association. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (10), 4244–4249.

(23) Plattner, N.; Doerr, S.; De Fabritiis, G.; Noé, F. Complete Protein-Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling. *Nat. Chem.* **2017**, *9* (10), 1005–1011.

(24) Dror, R. O.; Mildorf, T. J.; Hilger, D.; Manglik, A.; Borhani, D. W.; Arlow, D. H.; Philippsen, A.; Villanueva, N.; Yang, Z.; Lerch, M. T.; et al. Structural Basis for Nucleotide Exchange in Heterotrimeric G Proteins. *Science* **2015**, *348* (6241), 1361–1365.

(25) McGibbon, R. T.; Husic, B. E.; Pande, V. S. Identification of Simple Reaction Coordinates from Complex Dynamics. *J. Chem. Phys.* **2017**, *146* (4).

(26) Noé, F.; De Fabritiis, G.; Clementi, C. Machine Learning for Protein Folding and Dynamics. *Curr. Opin. Struct. Biol.* **2019**, *60*, 77–84.

(27) Grossfield, A.; Zuckerman, D. M. Chapter 2 Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations. *Annu. Rep. Comput. Chem*. 2009, *5*, 23–48.

(28) Genheden, S.; Ryde, U. Will Molecular Dynamics Simulations of Proteins Ever Reach Equilibrium? *Phys. Chem. Chem. Phys.* **2012**, *14* (24), 8662.

(29) Sawle, L.; Ghosh, K. Convergence of Molecular Dynamics Simulation of Protein Native States: Feasibility vs Self-Consistency Dilemma. *J. Chem. Theory Comput.* **2016**, *12* (2), 861–869.

(30) Nemec, M.; Hoffmann, D. Quantitative Assessment of Molecular Dynamics Sampling for Flexible Systems. *J. Chem. Theory Comput.* **2017**, *13* (2), 400–414.

(31) Tribello, G. A.; Ceriotti, M.; Parrinello, M. Using Sketch-Map Coordinates to Analyze and Bias Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (14), 5196–5201.

(32) Du, Y.; Duc, N. M.; Rasmussen, S. G. F.; Hilger, D.; Kubiak, X.; Wang, L.; Bohon, J.; Kim, H. R.; Wegrecki, M.; Asuru, A.; et al. Assembly of a GPCR-G Protein Complex. *Cell* **2019**, *177* (5), 1232-1242.e11.

(33) Komander, D.; Rape, M. The Ubiquitin Code. *Annu. Rev. Biochem.* **2012**, *81* (1), 203–229.

(34) Rogala, K. B.; Gu, X.; Kedir, J. F.; Abu-Remaileh, M.; Bianchi1, L. F.; Bottino1, A. M. S.; Dueholm1, R.; Niehaus1, A.; Overwijn1, D.; Priso Fils1, A. C.; et al. Structural Basis for the Docking of MTORC1 on the Lysosomal Surface. *Science* **2019**, *366* (6464), 468–475.

(35) Brooks, B. R.; Janežič, D.; Karplus, M. Harmonic Analysis of Large Systems. I. Methodology. *J. Comput. Chem.* **1995**, *16* (12), 1522–1542.

(36) Ichiye, T.; Karplus, M. Collective Motions in Proteins: A Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and Normal Mode Simulations. *Proteins Struct. Funct. Genet.* **1991**, *11* (3), 205–217.

(37) Perahia, D.; Levy, R. M.; Karplus, M. Motions of an ?-Helical Polypeptide: Comparison of Molecular and Harmonic Dynamics. *Biopolymers* **1990**, *29* (4–5), 645–677.

(38) Case, D. A. Normal Mode Analysis of Protein Dynamics. *Curr. Opin. Struct. Biol*. **1994**, *4*, 285-290.

(39) Hensen, U.; Meyer, T.; Haas, J.; Rex, R.; Vriend, G.; Grubmüller, H. Exploring Protein Dynamics Space: The Dynasome as the Missing Link between Protein Structure and Function. *PLoS One* **2012**, **,** *7* (5), e33931.

(40) Hess, B. Similarities between Principal Components of Protein Dynamics and Random Diffusion. *Phys. Rev. E* **2000**, *62* (6), 8438–8448.

(41) Hess, B. Convergence of Sampling in Protein Simulations. *Phys. Rev. E* **2002**, *65* (3), 031910.

(42) Karplus, M.; Kushick, J. N. Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules* **1981**, *14* (2), 325–332.

(43) Andricioaei, I.; Karplus, M. On the Calculation of Entropy from Covariance Matrices of

the Atomic Fluctuations. *J. Chem. Phys.* **2001**, *115* (14), 6289–6292.

(44) Sicard, F.; Senet, P. Reconstructing the Free-Energy Landscape of Met-Enkephalin Using Dihedral Principal Component Analysis and Well-Tempered Metadynamics. *J. Chem. Phys.* **2013**, *138* (23), 235101.

(45) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock, G. Construction of the Free Energy Landscape of Biomolecules via Dihedral Angle Principal Component Analysis. *J. Chem. Phys.* **2008**, *128* (24), 245102.

(46) Lange, O. F.; Grubmüller, H. Can Principal Components Yield a Dimension Reduced Description of Protein Dynamics on Long Time Scales? *J. Phys. Chem. B* **2006**, *110* (45), 22842–22852.

(47) Zhang, D.; Lazim, R. Application of Conventional Molecular Dynamics Simulation in Evaluating the Stability of Apomyoglobin in Urea Solution. *Sci. Rep.* **2017**, *7* (1), 44651.

(48) He, Y.; Maisuradze, G. G.; Yin, Y.; Kachlishvili, K.; Rackovsky, S.; Scheraga, H. A. Sequence-, Structure-, and Dynamics-Based Comparisons of Structurally Homologous CheY-like Proteins. *Proc. Natl. Acad. Sci.* **2017**, *114* (7), 1578-1583.

(49) Maragliano, L.; Cottone, G.; Cordone, L.; Ciccotti, G. Atomic Mean-Square Displacements in Proteins by Molecular Dynamics: A Case for Analysis of Variance. *Biophys. J.* **2004**, *86* (5), 2765-2772.

(50) Lyman, E.; Zuckerman, D. M. Ensemble-Based Convergence Analysis of Biomolecular Trajectories. *Biophys. J.* **2006**, *91* (1), 164–172.

(51) Romo, T. D.; Grossfield, A. Block Covariance Overlap Method and Convergence in Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2011**, *7* (8), 2464–2472.

(52) Lyman, E.; Zuckerman, D. M. On the Structural Convergence of Biomolecular Simulations by Determination of the Effective Sample Size. *J. Phys. Chem. B* **2007**, *111* (44), 12876–12882.

(53) Joseph, J. A.; Röder, K.; Chakraborty, D.; Mantell, R. G.; Wales, D. J. Exploring Biomolecular Energy Landscapes. *Chem. Commun.* **2017**, *53* (52), 6974–6988.

(54) Hochstrasser, M. Origin and Function of Ubiquitin-like Proteins. *Nature* **2009**, *458*

(7237), 422–429.

(55) Jackson, S. P.; Durocher, D. Regulation of DNA Damage Responses by Ubiquitin and SUMO. *Mol. Cell* **2013**, *49* (5), 795–807.

(56) Flotho, A.; Melchior, F. Sumoylation: A Regulatory Protein Modification in Health and Disease. *Annu. Rev. Biochem.* **2013**, *82* (1), 357–385.

(57) Hay, R. T. SUMO: A History of Modification. *Mol. Cell* **2005**, *18* (1), 1–12.

(58) Flotho, A.; Melchior, F. Sumoylation: A Regulatory Protein Modification in Health and Disease. *Annu. Rev. Biochem.* **2013**, *82* (1), 357–385.

(59) Gill, G. SUMO and Ubiquitin in the Nucleus: Different Functions, Similar Mechanisms? *Genes Dev.* **2004**, *18* (17), 2046–2059.

(60) Kotamarthi, H. C.; Yadav, A.; Koti Ainavarapu, S. R. Small Peptide Binding Stiffens the Ubiquitin-like Protein SUMO1. *Biophys. J.* **2015**, *108* (2), 360–367.

(61) Kotamarthi, H. C.; Ainavarapu, S. R. K. Mechanical Unfolding Studies on Single-Domain SUMO and Multi-Domain Periplasmic Binding Proteins. *Biophys. Rev. Lett.* **2017**, *12* (01), 1–10.

(62) Sundquist, W. I.; Schubert, H. L.; Kelly, B. N.; Hill, G. C.; Holton, J. M.; Hill, C. P. Ubiquitin Recognition by the Human TSG101 Protein. *Mol. Cell* **2004**, *13* (6), 783–789.

(63) Hicke, L.; Schubert, H. L.; Hill, C. P. Ubiquitin-Binding Domains. *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (8), 610–621.

(64) Husnjak, K.; Dikic, I. Ubiquitin-Binding Proteins: Decoders of Ubiquitin-Mediated Cellular Functions. *Annu. Rev. Biochem.* **2012**, *81* (1), 291–322.

(65) Haas, A. L.; Bright, P. M. The Immunochemical Detection and Quantitation of Intracellular Ubiquitin-Protein Conjugates. *J. Biol. Chem.* **1985**, *260* (23), 12464–12473.

(66) Bosanac, I.; Phu, L.; Pan, B.; Zilberleyb, I.; Maurer, B.; Dixit, V. M.; Hymowitz, S. G.; Kirkpatrick, D. S. Modulation of K11-Linkage Formation by Variable Loop Residues within UbcH5A. *J. Mol. Biol.* **2011**, *408* (3), 420–431.

(67) Wlodarski, T.; Zagrovic, B. Conformational Selection and Induced Fit Mechanism

Underlie Specificity in Noncovalent Interactions with Ubiquitin. *Proc. Natl. Acad. Sci.* **2009**, *106* (46), 19346–19351.

(68) HASTINGS, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57* (1), 97–109.

(69) Newman, M. E. J.; Barkema, G. Monte Carlo Methods in Statistical Physics. Clarendon Press, 1999.

(70) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

(71) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257–3273.

(72) Huang, J.; MacKerell, A. D. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J. Comput. Chem.* **2013**, *34* (25), 2135–2145.

(73) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An $N \cdot \log(N)$ Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.

(74) Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31* (3), 1695–1697.

(75) Nosé, S. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* **1984**, *52* (2), 255–268.

(76) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190.

(77) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38.

(78) Daidone, I.; Amadei, A. Essential Dynamics: Foundation and Applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2* (5), 762–770.

(79) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins Struct. Funct. Genet.* **1993**, *17* (4), 412–425.