#ToDos Check tenses? Past or present Stick with atomistic or all-atom

Combining MD simulations with sPRE calculations for protein structure prediciton

Abstract

The following pages contain the analysis basis (available MD data), the methods and, results of applying XPLOR NIH to three differently linked (K6, K29, K33) ubiquitin dimers. These sections are meant to appear in the supplementary information of an upcoming paper. The figures can be used and included in the main part as one sees fit.

Available data

Previous work

CG data

MD data at coarse-grained (CG) resolution was taken from [1]. This dataset comprised the ubiquitination sites K6, K29 and K33 where the data for every protein is comprised of 10 replicas of 10 ms each. The MARTINI v2.2 force field was applied to the non-bonded interaction, whereas bonded interactions have been modeled using the IDEN method. These forcefields had to be altered to accommodate the isopeptide bond between the proximal subunit's lysine and the distal subunit's C-terminal glycine (Figure 1). The CG data of all 3 ubiquitylation sites encompass approximately 300 ms.

Back-mapped atomistic data

These simulations were used as input-data for a dimensionality reduction algorithm called sketch-map. Berg at al. used suitable high-dimensional collective variables to project the protein's phase space into 2D. In such representations similarity of any two conformations is represented in the distance in sketch-space. Naturally, basins of higher density represent statistically significant conformational sub-states of low free energy. Susequently, new all-atom simulations were started by selecting points in sketch-space from both the center of these basins (4 simulations per ubiquitination site with 10 ns each) and from randomly chosen points in the vicinity of the basin (40 simulations per ubiquitination site with 3 ns each). Martini's `backward` script, paired with energy minimization to settle initial high-energy structures, yielded atomistic starting structures for these simulations.

Simulations started from extended structures

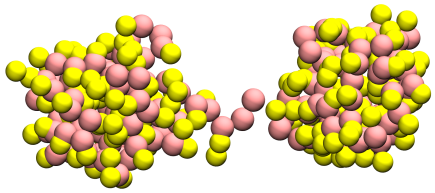

Figure 1: Coarse grained representation of K6-diUbi. Pink beads are unified backbone beads (C_α, N, and C). Yellow beads are unified sidechain beads (one bead for all non-GLY residues). The proximal subunit (right) offers its LYS6 residue to be ubiquitinated by the distal (left) unit's GLY76 residue.

Additionally, to the back-mapped atomistic simulations, further atomistic data were taken from Berg et al. The starting structures of these simulations were created by slightly altering the position between the proximal and distal subunit of the di-ubiquitin proteins. As a reference for the monomer, the 1UBQ crystal structure from the protein database was used [2]. For every ubiquitination site, 12 simulations with 50 ns each are available. The atomistic simulations employed the GROMOS54a7 forcefield which was adapted to accommodate for the isopeptide bond, which was parametrized from regular peptide bonds.

MD simulations conducted for this work

The dataset of atomistic simulations was complemented with additional simulations started from starting structures created by rotating the sidechain χ_3 angle of the ubiquitinated lysine residue in 40 degree steps. The rotamer simulations were intended to accelerate the exploration of the phase space available to the diubiquitin proteins. This dataset encompasses 9 simulations per ubiquitination site with approximately 50 ns each.

Combination of CG and AA data

This might be interesting info for the main part

The simultaneous usage of CG and AA data for this project has been proven to be especially advantageous. Normally, all-atom simulations sample a protein's phase space at a fraction of the speed available to CG simulations (approx. 3500 $\frac{ns}{day}$ for CG and approx. 8 $\frac{ns}{day}$ for AA in the case for diUbi). This disadvantage

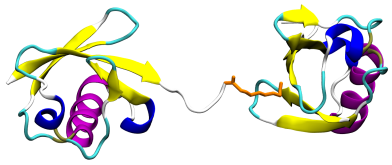

Figure 2: All-atom representation of K6-diUbi. The structure in this figure was used as a basis for Figure 1 and the location of proximal and distal units are identical. Secondary structure has been visualized as cartoon representation and colored accordingly. The residues of the isopeptide bons (GLY76 (left) and LYS6 (right)) are colored orange.

led to sophisticated methods to accelerate sampling. Some of these so-called advanced sampling methods introduce a bias in the form of added potentials, which leads to unavoidable skew in the MD ensemble. Even starting multiple short AA simulations introduces a bias towards the starting structures. One could run long simulations to reduce this starting structure bias, but sooner or later will find themselves in a Catch-22 situation, where any efforts to accelerate the sampling will lead to additional bias. Pure CG simulations will not be able to break free from this problem, as they simply do not offer the resolution of AA simulations. We broke this vicious cycle by combining CG and AA data. Long CG simulations pave the way and define the general shape of the free energy surface and short AA simulations increase the resolution in the interesting parts. This can be seen in Figure 4.

All in all a single ubiquitination site is characterized by 100 ms of CG data and 1.21 ms of AA data.

Methods

Generally, the analysis steps were performed as follows: 1. Extract high-dimensional collective variables. 2. Project into 2D space using Encodermap. 3. Use HDBSCAN to extract statistically-significant sub-states of ensemble. 4. Calculate sPRE values for the all-atom conformations using XPLOR NIH (task has been parallelized). 5. Normalize the sPRE data from XPLOR NIH. 6. Identify some clusters that represent the ensemble available to di-ubiquitin using sPRE and HDBSCAN .

Extact high-dimensional collective variables

Understanding the dynamics of a protein can be difficult. Of course, its possible to look at a nicely rendered movie of how the protein might move. However, making assessments general conformations i.e. states is impossible from hour-long MD movies. For that we chose to use Encodermap to easily visualize overall protein dynamics as a 2D map. But before creating this map a set of collective variables needed to be defined. Collective variables represent a type of data that is somehow aligned with the raw xyz positional data of a molecular dynamics trajectories. With CVs complex processes can be broken down into meaningful variables. One such example might be the simplification of receptor-ligand docking processes by using a receptor-ligand distance as a collective variable. The torsion angles of the backbone are another widely used example of collective variables. From the ϕ and ψ dihedral angles a ramachandran plot can be created to visualize and simplify the complex structure of proteins. In this paper we faced the challenge to unify the CG and AA data, meaning CVs needed to be identical for both types of MD data. Torsional angles could not be calculated from CG data as the backbone N, C, and CA atoms are unified within one CG bead. Distance matrices between all CA atoms would yield a $\binom{15}{2} = 11476$ dimensional dataset. We chose to use the approach also used by Berg at al. employing residue-wise minimal distances, which makes the input data 152-dimensional.

Encodermap dimensionality reduction

The high-dimensional dataset consisted of the 152 residue-wise minimal distances for each MD frame (CG or atomistic). The approximately XXXX frames per ubiquitination site, made it unfeasible to employ non-machine-learning dimensionality reduction algorithms. They could not cope with the number of data (for example: multidimensional-scaling would require an $n_{frames} \times n_{frames}$ distance matrix, which would require impossible amounts of memory. Sketch-map requires one to preselect points of interest which undergo dimensionality reduction. The remaining points are placed by minimizing similarity matrices). We chose to use the machine-learning approach of Encodermap's auto-encoder neural network to retrieve a dimensionally reduced representation of all aforementioned simulations [3]. Only this method allows for iterative optimization of a differentiable function that projects the high-dimensional input data into the low-dimensional latent space of the autoencoder bottleneck. The high-dimensional input data was fed into a dense, fully-connected sequential neural network comprised 250, 250, 125, 2, 125, 250, and 250 neurons. Three main cost functions were minimized in the training phase to ensure a good 2D representation.

- Center cost: Mean absolute distance between all points and the coordinate origin. This cost function keeps the 2D points at the coordinate origin. It becomes higher, the further away points are from the coordinate origin.

$$CenterLoss = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{n}$$

- Auto cost: Compare the mean absolute difference between input and output. This function becomes higher the greater the difference between the encoder input and decoder output are.

$$AutoLoss = \frac{\sum_{i=1}^{n} \| \sqrt{y_i^2 - \hat{y}_i^2} \|}{n}$$

- Encodermap cost: Compare the input and the latent space by transforming it with a sigmoid function. This function becomes higher the greater the sigmoid-weighted difference between the pairwise distances of encoder input and encoder output (i.e. latent space) are:

$$EncoderMapLoss = \frac{1}{m} \sum_{i \neq j} \left[SIG_h(R_{ij}) - SIG_l(r_{ij}) \right]^2$$

,

where SIG is a sigmoid function originally devised by Ceriotti et al. for their sketch-map dimensionality reduction algorithm [4].

$$SIG_{\sigma,a,b}(r) = 1 - \left(1 + \left(2^{\frac{a}{b}} - 1 \right) \left(\frac{r}{\sigma} \right)^a \right)^{-\frac{b}{a}}$$

R_{ij} and r_{ij} are the pairwise distance between all points in the high-dimensional input space and the 2-dimensional latent space. They are defined using the pairwise distance matrices $boldsymbol R_{i,j}$:

$$\boldsymbol{R}_{i,j} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & & \\ \vdots & & \ddots & \\ r_{n,1} & & & r_{n,n} \end{pmatrix}$$

without the diagonal elements (which are always 0) and vectorizing them:

$$R_{ij} = \mathrm{vec}(\boldsymbol{R}_{i,j,i \neq j}) = \begin{pmatrix} r_{1,2} \\ r_{1,3} \\ \vdots \\ r_{1,n} \\ r_{2,1} \\ \vdots \\ r_{n,n-1} \end{pmatrix}$$

The sigmoid function which transforms the high-dimensional input space and the low-dimensional latent space was defined by the following parameters:

σ_{highd}	A	B	σ_{lowd}	a	b
5.9	12	4	5.9	2	4

In the resulting 2D projection every point in xy space represents one conformation sampled by the MD simulations. The closeness of any two points correlates to the structural similarity of these two conformations.

HDBSCAN clustering

The hierarchical density-based clustering algorithm HDBSCAN has been tried and tested for molecular dynamics systems and especially the low-dimensional projections of such [5]. Thus, this established method was chosen. After the 2D projections were obtained for every ubiquitination site, the xy-values were fed into HDBSCAN using the 'leaf' algorithm, with a minimal cluster size of 2500, resulting in 13, 18, and 24 selected clusters for K6, K29, and K33, respectively. The clusters are enumerated based on the number of included points/conformations, with cluster 0 being the largest cluster. Another notable point is that HDBSCAN offers a great advantage by distinguishing between clusters and noise. Noise - speaking from an ensemble viewpoint - is comprised of unique structures that carry no significance and thus can be disregarded.

Look at a specific cluster

As an example cluster 0 of K6 diUbi consists of 131969 conformations sampled by XXX independent CG and XXX independent AA simulations and makes up roughly 12% of the complete K6 diUbi ensemble. Only 3% of the conformations in this cluster are from all-atom simulations (Figure 5). Visualizing 10 of the approximately 4000 all-atom conformations shows, that they are indeed structurally similar (Figure 6).

XPLOR calculations

The NMR prediction and refinement program XPLOR NIH in version 3.3 was used to calculate the sPRE and 15N-relaxation time NMR observables. XPLOR's new python functionality was used to call XPLOR functions from within python. XPLOR's intended workflow follows the steps of structure loading, setting up potentials, setting up the dynamics using the IVM (internal variable module) and module and refining the loaded structure by minimizing the potentials iteratively. As we already had an extensive set of structures we dropped the refinement steps and let XPLOR calculate the initial potential values for every simulation frame. To accelerate this undertaking, we parallelized this computation by using

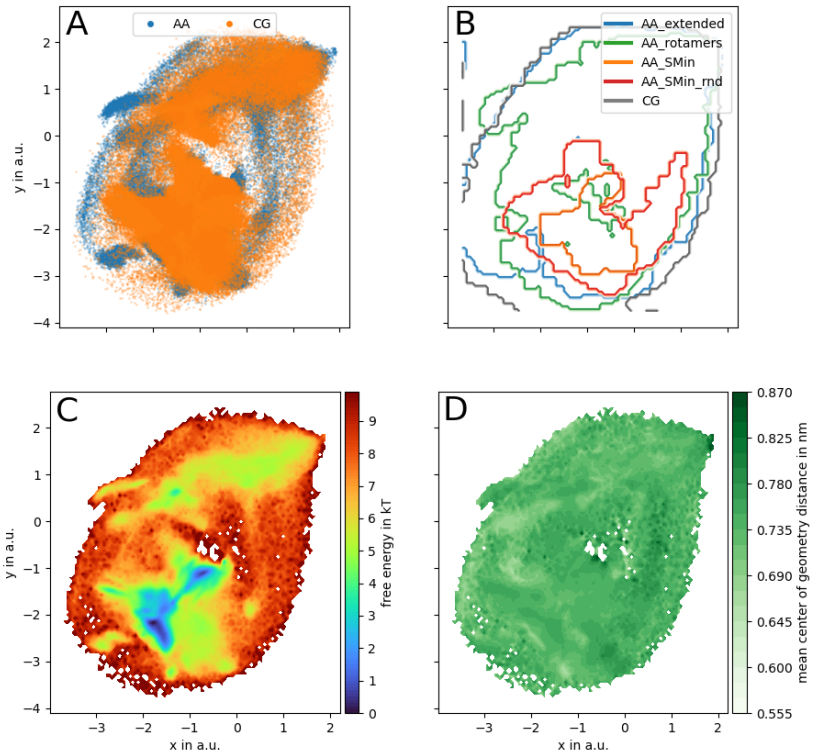

Figure 3: 2D projections of all CG and AA conformations of K6 diUbi. When the low-dimensional coordinates are plotted as a scatter plot, every point can be associated with one conformation sampled by a MD simulation (A). Orange points belong to CG simulations, blue points to AA. The closeness of two points. In (B) the areas occupied by specific MD simulation trajectories is visualized. Here, it can be seen, that the long AA extended simulations sampled more unique conformations, when compared to the limited regions of the sketch-map minima and sketch-map minima random sims. It should be noted, that the AA_rotamers were a more sucessfull strategy in creating unique conformations, when compared to SMin and SMIN_rnd. However, the AA_rotamer simulations are inherently starting-structure biased and might not yield a representative i.e. biased ensemble. That's why the long (and with than more unbiased) CG simulations are used to obtain a more correctly weighted ensemble. The Encodermap projections correlate to many other structural properties of the system. So can the distance between the center of geometries of the Ub subunits be used to color the projection and distinct regions can be observed (D).

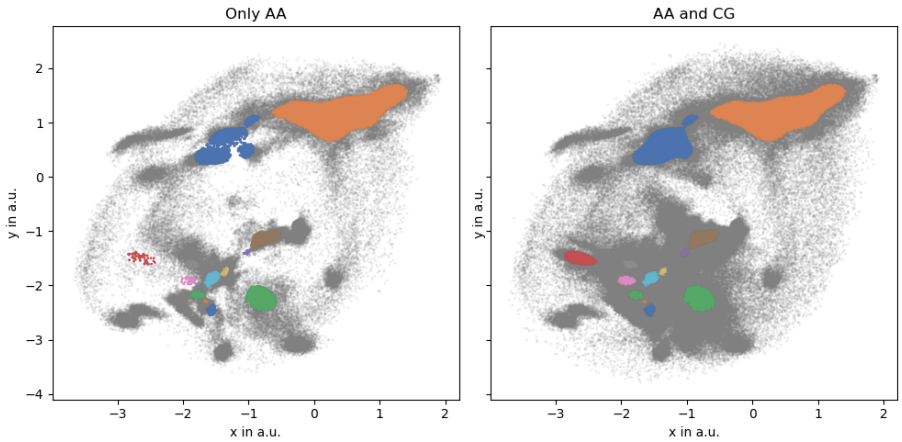

Figure 4: The Encodermap projection of K6 colored according to cluster membership. The advantage of using combined CG and AA data becomes apparent, when the red cluster is investigated. In a combined dataset this cluster contains more than 2500 structurally similar conformations in a dense basin (i.e. statistically significant sub-state of the K6 diUbi system), but multiple AA simulations were not able to sample these structures sufficiently, which would have removed these structures from the list of candidates for further analysis.

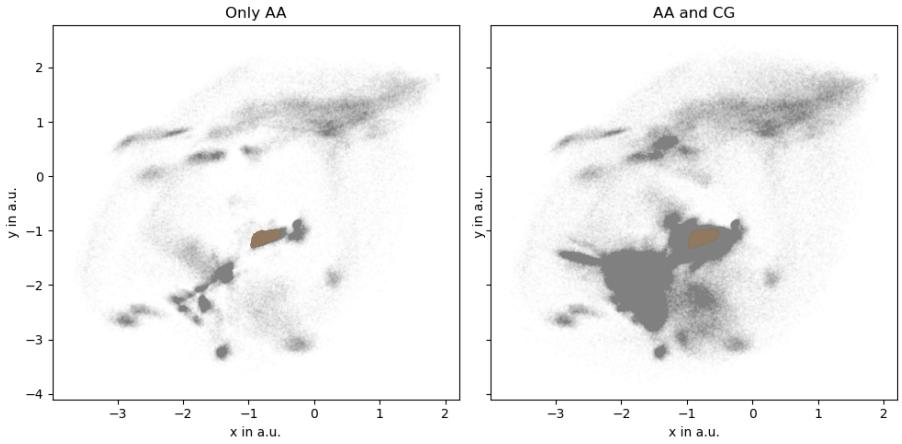

Figure 5: Encodermap projections of K6 diUbi, with cluster 0 highlighted.

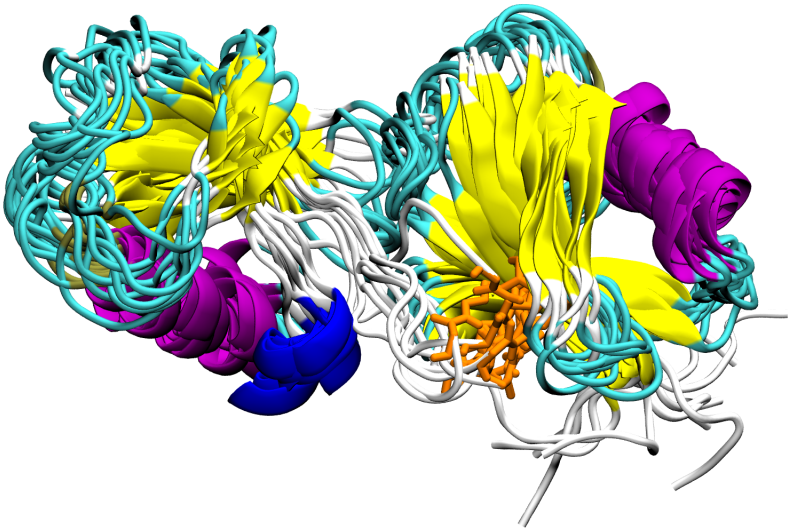

Figure 6: 10 conformations from cluster 0 of K6 diUbi. The structures are rendered as cartoon representation and colored accordingly. The isopeptide LYS and GLY residues are visualized as orange licorice (proximal subunit on the right, distal subunit on the left). It is notable, that the structural difference in the residues taking part in the isopeptide bond is larger, when compared to the protein backbone. The α-helices seem to overlap more precisely.

python's joblib process-based multitasking. Every worker was provided with a number of simulation frames it should run the XPLOR protcol on and then carried out the task of saving structure files and calling XPLOR.

15-N relaxation times

As XPLOR NIH did not recognize the isopeptide bond in the provided pdb files (XPLOR's pdbTool does not possess, contrary to the online documentation, a `readConect` method), we created a workaround that involved creating psf files for the three proteins with XPLOR's pdb2psf program and manually adding the required bond. Some atoms in the pdb files then needed to be manually adjusted as they did not follow conventional naming (e.g. the hydrogens at the N-terminus are most commonly labelled as H1, H2, and H3. XPLOR's pdb2psf program labelled these atoms as HT1, HT2, HT3). With these manual adjustments XPLOR's relaxratiopot potential produced different output, compared to a protocol which did not use this workaround.

IMAGE

Code availability

The code developed for this publication will be made publicly available at: https://github.com/kevinsawade/xplor_functions

Normalization of sPRE values The sPRE values from the XPLOR runs needed to be normalized before they could be quantitatively compared to experimental findings. We adjusted the normalization method described by Gong et al. for our system as follows [6]:

The simulated XPLOR values for the proximal and distal subunit build a $76 \times N$ matrix, respectively. Here, N is the number of aggregated frames over all trajectories for a specific di-ubiquitin protein. As ubiquitin has 76 residues, that number is fixed.

$$\boldsymbol{sPRE}_{sim}^{76\times N} = \begin{bmatrix} sPRE_{MET1,1} & sPRE_{MET1,2} & \dots & sPRE_{MET1,n} \\ sPRE_{GLN2,1} & sPRE_{GLN2,2} & & \\ sPRE_{ILE3,1} & sPRE_{ILE3,2} & & \\ \vdots & & \ddots & \\ sPRE_{GLY76,1} & sPRE_{GLY76,2} & \dots & sPRE_{GLY76,n} \end{bmatrix}$$

From New-Mexico experiments we identified some residues to exhibit a fast proton exchange rate with the solvent and thus do not yield representative sPRE values.

This section needs to be checked and maybe extended with more info about new mexico

These residues have been removed from the considerations for the normalizations. Furthermore, the proximal and distal unit were considered separately, resulting in 6 factors, two for K6, K29, and K33-linked di-ubiquitin, respectively. The per-residue variance over the simulation frames was calculated yielding a vector of variance:

$$\text{var}\left(\boldsymbol{sPRE}_{sim}^{76\times N}\right) = \begin{bmatrix} \text{var}(sPRE_{MET1,1} & sPRE_{MET1,2} & \dots & sPRE_{MET1,n}) \\ \text{var}(sPRE_{GLN2,1} & sPRE_{GLN2,2} & & \\ \text{var}(sPRE_{ILE3,1} & sPRE_{ILE3,2} & & \\ \vdots & & \ddots & \\ \text{var}(sPRE_{GLY76,1} & sPRE_{GLY76,2} & \dots & sPRE_{GLY76,n}) \end{bmatrix}$$

Instead of selecting one singular deeply buried - with the most reliable sPRE value - residue, we chose to select the 10 residues with the smallest variance. For each of these 10 residues a normalization factor $f_{i,ubq,pos}$, where i is the residue, *ubq* the ubiquitination site and *pos* either proximal or distal, was calculated:

$$f_{i,ubq,pos} = \frac{v_{i,ubq,pos,exp}}{v_{i,ubq,pos,sim}}$$

where $v_{i,ubq,pos,exp}$ is the experimental sPRE value for residue i and $v_{i,ubq,pos,sim}$ is the mean simulated sPRE for residue i over N simulation frames. The normalization factors $F_{ubq,pos}$, which were used to normalize all simulated sPRE values were obtained as the mean of the 10 normalization factors $f_{i,ubq,pos}$:

$$F_{ubq,pos} = \frac{\sum_i^{10} f_{i,ubq,pos}}{10}$$

Metric of comparison: mean abs difference As a metric of difference between one set of sPRE-values and another we chose the mean absolute difference, which can in general be obtained as:

$$MAD = \frac{\sum_i^n \|x_i - \hat{x}_i\|}{n}$$

We exclusively used the MAD to compare experimental and simulated sPRE values, so the equation can be re-written as:

$$MAD = \frac{\sum_i^{152} \|x_{exp,i} - x_{sim,i}\|}{152}$$

With this measure, we could compare any simulation frame from the atomistic simulations to the experimental values (Figure 7).

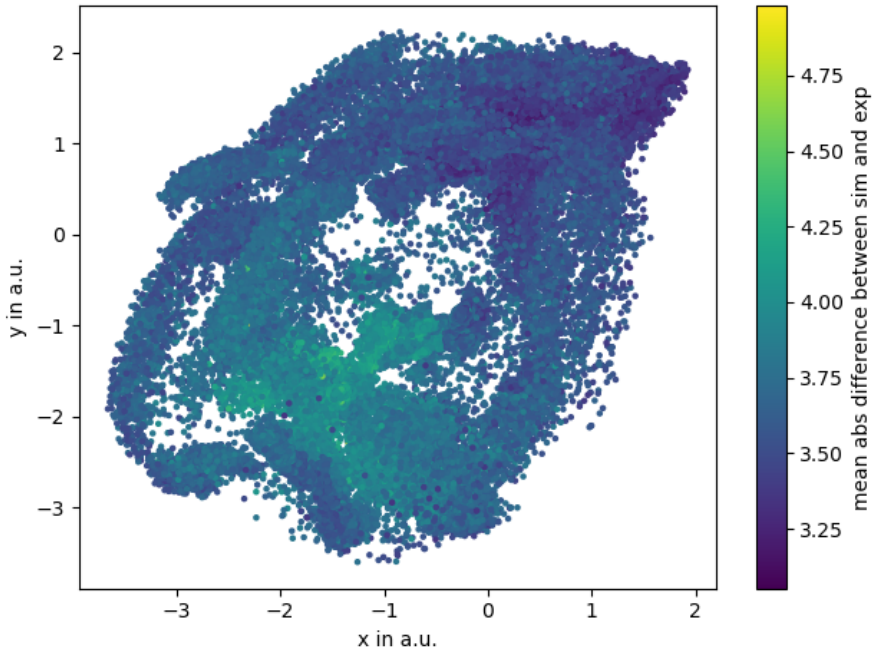

Figure 7: The low-dimensional Encodermap projection (x and y values) colored by the mean abs difference between a conformation from MD simulations and the experimental sPRE values. Certain regions with lower and higher mean abs difference can clearly be destinguished.

Identification of clusters

As a final step we conducted a linear combination to identify only a few clusters per ubiquitination site that should be able to represent the overall conformation of the di-ubiquitin proteins in a dynamic system. Starting from the experimental sPRE-values ($sPRE_{exp}$), we set up this system of equations:

$$\begin{bmatrix} sPRE_{exp,MET1,prox} \\ SPRE_{exp,GLN2,prox} \\ \vdots \\ sPRE_{exp,GLY76,dist} \end{bmatrix} = x_1 \cdot \begin{bmatrix} sPRE_{clu_1,MET1,prox} \\ SPRE_{clu_1,GLN2,prox} \\ \vdots \\ sPRE_{clu_1,GLY76,dist} \end{bmatrix} + \cdots + x_n \cdot \begin{bmatrix} sPRE_{clu_n,MET1,prox} \\ SPRE_{clu_n,GLN2,prox} \\ \vdots \\ sPRE_{clu_n,GLY76,dist} \end{bmatrix}$$

Here, the sPRE-values are represented as column-vectors with proximal and distal values unified. The residues identified by the New Mexico experiment to exhibit a fast exchange have been excluded from these considerations, which makes the column vectors shorter than 152 entries. For every residue there is one experimental value ($sPRE_{exp,MET1,prox}$) and multiple cluster-values. For every cluster the mean sPRE value of all conformations assigned to that cluster was taken, so that:

$$sPRE_{clu_1,MET1,prox} = \frac{\sum_n^i sPRE_{clu_1,MET1,prox,i}}{n_{\text{points in cluster}}}$$

This equation can be written as:

$$\boldsymbol{sPRE_{exp}} = x_1 \cdot \boldsymbol{sPRE_{clu_1}} + x_2 \cdot \boldsymbol{sPRE_{clu_2}} + \ldots x_n \cdot \boldsymbol{sPRE_{clu_n}}$$

The objective function, that was minimized can be written as:

$$f(\boldsymbol{x}) = \| \sum (\boldsymbol{x} \times \boldsymbol{sPRE_{sim}}) - \boldsymbol{sPRE_{exp}} \|$$

Here, $\boldsymbol{x}$ is a vector of coefficients and $\boldsymbol{sPRE_{sim}}$ is a matrix of the simulated sPRE values of the n clusters. The minimization was carried out with boundary conditions, so that $\forall x_i \in \mathbb{R}_{0 \leq x \leq 1}$ and $\sum_n^i x_i = 1$.

$$\min_{\boldsymbol{x} \in \mathbb{R}^n_{\geq 0}} f(\boldsymbol{x})$$

The results of this linear combination can be taken from Figure 8. The cluster id follows the number of points inside this cluster, so that cluster 0 contains the most individual frames (N frames). Every cluster contains a mixed number of CG and atomistic frames and contributes differently to the overall structural

ensemble. Clusters with many frames from multiple simulations are more often visited during our sampling and thus the protein spends more time in the general conformation of that cluster. The coefficients of the linear combination roughly follow that trend. Clusters with higher representation in the ensemble have higher coefficients assigned to them. Meaning that these clusters contribute to the experimental sPRE values. For K33 and K6 the two highest occupied clusters also yield the highest coefficients (although the sequence is inverted for K6). In the case of K29 it's only cluster 0, that significantly, contributes the the ensemble and the sPRE linear combination. All three di-ubiquitin proteins also have a cluster with a lower occupation assigned with a coefficient larger than 0. Furthermore, for all three ubiquitylation sites, the main clusters (2 for K6 and K33, one for K29) and the lower-occupation cluster are in close proximity in the low-dimensional projection (Figure 11). This speaks for a general structural similarity and an underlying dynamic process between the aforementioned clusters.

ubq site	cluster id	N frames	ensemble %	aa %	coefficient
k33	**0**	119908	10	2	0.41
	1	95897	8	1	0.33
	2	31411	3	1	0.00
	3	30973	3	47	0.00
	4	29716	3	2	0.00
	5	29250	2	2	0.00
	6	22921	2	1	0.00
	7	21385	2	18	0.00
	8	21253	2	4	0.00
	9	20462	2	9	0.05
	10	9469	1	1	0.00
	11	9425	1	34	0.00
	12	9010	1	100	0.00
	13	7593	1	81	0.00
	14	6179	1	11	0.00
	15	6174	1	3	0.00
	16	5877	0	3	0.00
	17	5140	0	6	0.00
	18	4838	0	100	0.00
	19	4735	0	1	0.22
	20	4326	0	1	0.00
	21	3231	0	4	0.00
	22	3050	0	100	0.00
	23	2584	0	100	0.00
sum and final linear combination		Σ 504807	Σ 43		Σ 1.0

Figure 8: Solving the linear combination on the example of K33-linked di-ubiquitin.

ubq site	cluster id	N frames	ensemble %	aa %	coefficient
k29	0	186366	17	3	0.68
	1	99803	9	1	0.00
	2	48921	4	16	0.00
	3	44759	4	2	0.00
	4	28701	3	5	0.00
	5	22909	2	13	0.00
	6	16151	1	7	0.00
	7	14024	1	2	0.00
	8	12681	1	5	0.00
	9	11039	1	15	0.00
	10	7428	1	3	0.00
	11	7142	1	7	0.26
	12	6436	1	0	0.05
	13	5148	0	7	0.00
	14	4662	0	2	0.00
	15	4606	0	36	0.00
	16	4112	0	1	0.01
	17	3379	0	1	0.00
sum and final linear combination		Σ 528267	Σ 48		Σ 1.0

Figure 9: Same results fro K29.

ubq site	cluster id	N frames	ensemble %	aa %	coefficient
k6	0	131969	12	3	0.21
	1	96726	9	1	0.47
	2	43256	4	4	0.06
	3	39818	4	1	0.00
	4	27491	2	74	0.00
	5	17686	2	0	0.00
	6	15884	1	32	0.00
	7	13093	1	17	0.00
	8	12241	1	3	0.00
	9	7014	1	1	0.00
	10	6479	1	18	0.00
	11	5848	1	0	0.26
	12	2955	0	3	0.00
sum and final linear combination		Σ 420460	Σ 38		Σ 1.0

Figure 10: Saem results for K6.

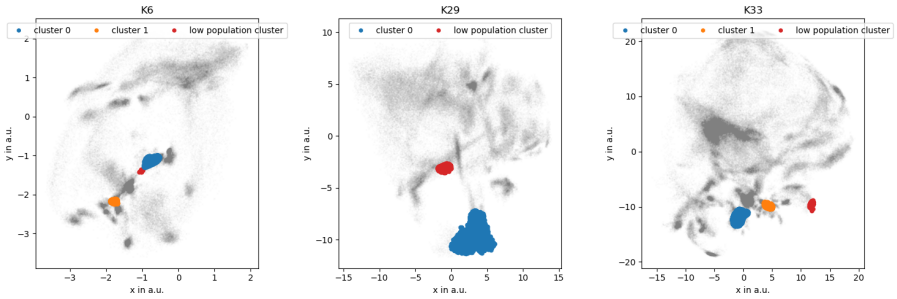

Figure 11: Main clusters (0 (orange) for all three, 1 (blue) only for K6 and K33) and the lower-occupation clusters (red) for the three di-ubiquitin proteins.

Results

As a result we present the aforementioned clusters as candidates for conformations that are not only observable in experiment, but also are the predominant structures in MD simulations.

Cluster combination

The methods section introduced a way of combining different clusters linearly, minimizing the difference of this combination and the experimental sPRE values. Doing so yields a better representation of the sPRE values than calculating a median of all simulated MD frames (Figure 12). This also ties in nicely with the enforced sampling we did to achieve such an exhausting dynamic library of di-ubiquitin structures. Due to starting many short simulations the free-energy surface of the di-ubiquitin proteins is expected to be shallower than the true free-energy landscape. Some basins might be deeper in real-life. However, the basins, that exhibit a large population from our sampling, are the most likely candidates to yield local minima of the free-energy landscape. And these candidates have also been chosen by the linear combination of sPRE values (Figure 13). Furthermore, the difference between experiment and the applied linear combination is smaller, than the difference between the median of all structures and the experimental values (Figure 14).

Visualization of cluster combinations

For K6, the clusters are grouped around lysine 6 covering a defined area of the proximal subunit (Figure 15). The red cluster (for K6 the low-population (red) cluster is cluster 11) can be found (similarly to the dimensionally-reduced projection) between the other two. We propose a combination (using the coefficients of the linear combination and the weights of the respective clusters

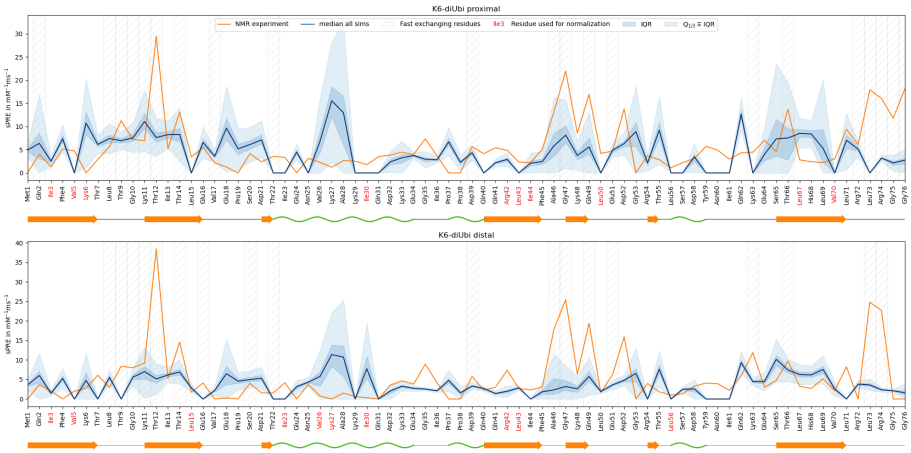

Figure 12: Experimental and simulated sPRE values. The proximal and distal subunit are plotted in the upper and lower part of the figure, respectively. The y-axis is labelled with the residue name and number, as well with the secondary structure information of the crystal structure 1UBQ. Some residues are highlighted in red lettering. These 10 residues per subunit (proximal and distal) exhibit the lowest variance over all simulations and were used to normalize the sPRE values. Very faint hatched bars indicate residues that have been identified by New-Mexico experiments to exhibit a fast exchange. The sPRE values from these residues should be taken with a grain of salt. The orange line plots are the experimental sPRE values for the respective residues. The simulated values are visualized via the median (dark blue line), the inter quartile range (IQR, blue area) and the IQR minusplus the Q1, Q3 quartiles (light blue area) The difference between experimental and median of all can be taken form Figure 14.

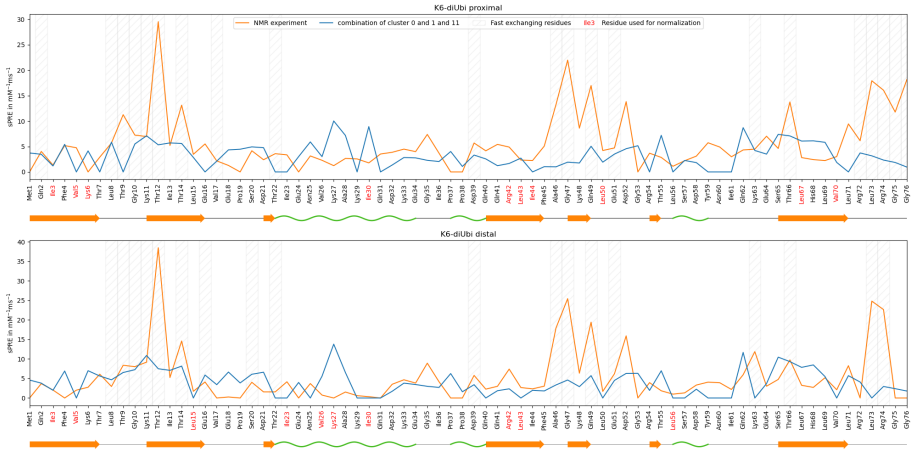

Figure 13: Experminetal and optimized sPRE values. Similar to Figure 12, but this time, the dark blue line is the combination of the 3 selected clusters using the respective coefficients to weight these clusters.

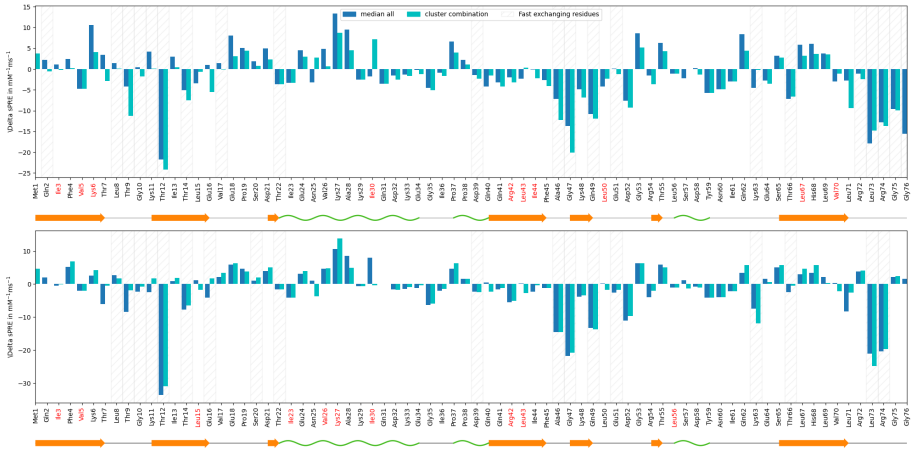

Figure 14: Difference between experimental sPRE values and the median of all simulation frames and the cluster combination of clusters 0, 1, and 11. The labelling of the y-axis is similar to Figure 12. The bar height represents the Δ sPRE between the experimental values and the median of all (dark bue) and the difference between the experimental values and the cluster combination (cyan). Especially in regions where the sPRE values are more trustworthy (slower exchange, no hatched bars), the differences become small.

in the combined CG and atomistic ensemble) as preferable conformations of K6-ubiquitinated di-ubiquitin.

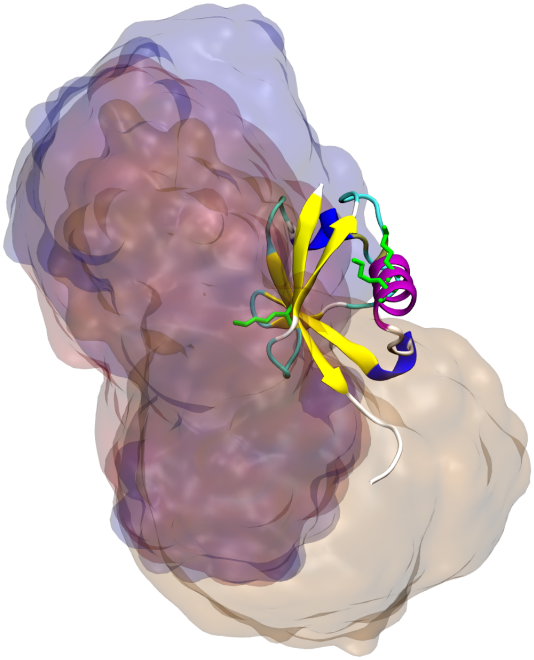

Figure 15: Resulting clusters for K6. The proximal subunit has been rendered as a cartoon representaion with according secondary structure coloring. The lysine reisudes K6, K29 and K33 are highlighted in green. The K6-linked di-ubiquitin protein resulted in three main clusters. The clusters are rendered in the same colors as Figure 11.

Visualization as polar plots

To better understand the arrangement of the two subunits in the clusters, we chose to represent the isopeptide as a polar plot. These images can be understood, when one imagines the proximal subunit of the protein as a sphere. Based on the conformation of the di-ubiquitin, the distal unit will cover different portions of the surface of the proximal subunit. The center of the sphere coincides with the center of geometry of the proximal unit. As visual guidelines, the residuenames are added to the sphere's surface. The latitudinal and longitudinal positions of the residues can be directly obtained by transforming the xyz position of the Cα atom of a respective residue into spherical coordinates and disregarding r

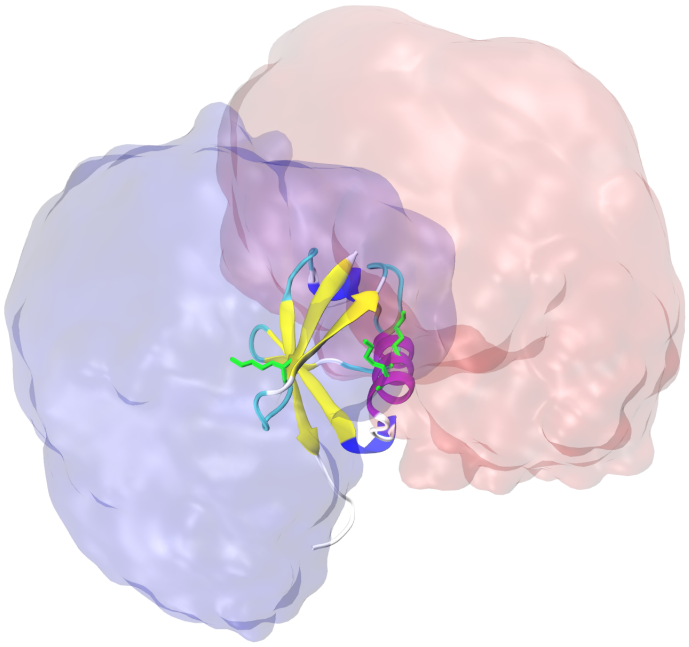

Figure 16: Resulting clusters for K29. In comparison to Figure 15, the proximal ubiquitin subunit has been pitched upward. This becomes apparent, when he pitch of the α helix is compared. The blue cluster is, similar to K6, cluster 0. Contrary to K6, K29 does not exhibit a second cluster with high probability. However, cluster 11 (red), with a coefficient of 0.26 should also be considerd, as it might present a second stable state for K29-ubiquitinated di-ubiquitin.

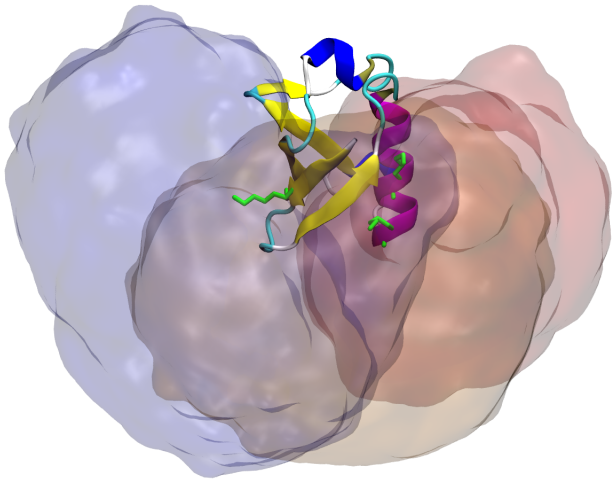

Figure 17: Resulting clusters for K33. In this render, the proximal subunit has been pitched downward, even more. This view shows, that the main clusters for K33 (0 blue, 1 orange and 19 red) cover a specific side of the proximal subunit./label{k33_redners}

the distance between point and coordinate center.

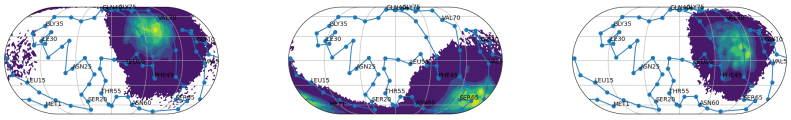

Figure 18: Polar plots for the selected clusters for K6-ubiquitinated di-ubiquitin. Form left to right: cluster 0, cluster 1, cluster 11. Cluster 0 and 1 clearly assume distinct conformations covering specific regions of the proximal subunit's surface.

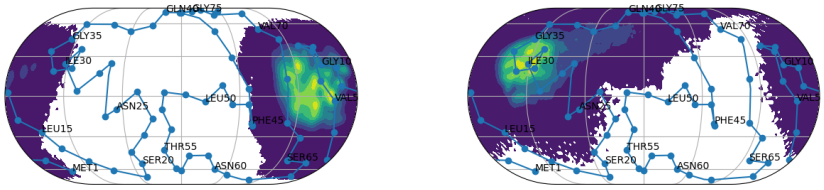

Figure 19: Polar plots for K29. From left to right: cluster 0, cluster 11.

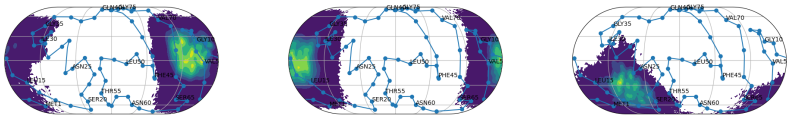

Figure 20: Polar plots for K33. From left to right: cluster 0, custer 1, cluster 19.

Conclusion

In this work, we presented a new method of combining experimental NMR spectroscopy and simulated CG and atomistic structural data. We successfully applied these methods on three di-ubiquitin proteins and were able to reveal a synergistic interplay between tem, neglecting the short-comings of each individual

method and reach a high degree of refinement. As atomisitc MD simulations take time to compute and might be biased, we complemented these data with less comutationally demanding CG simulations and layed ground for ensemble-wide evaluations, such as the assessment of stable states and the relative frequency of these states. Using solvent paramagnetic relaxation techniques, we extended the timespan of the available data even further. We developed new normalization algorihtms and parallelization techniques for XPLOR NIH to reach a level of refinement where both XPLOR normalization and ensemble weights predict identical protein structures f the exhaustive MD sampling.

Appendix

sPRE values for K29 and K33

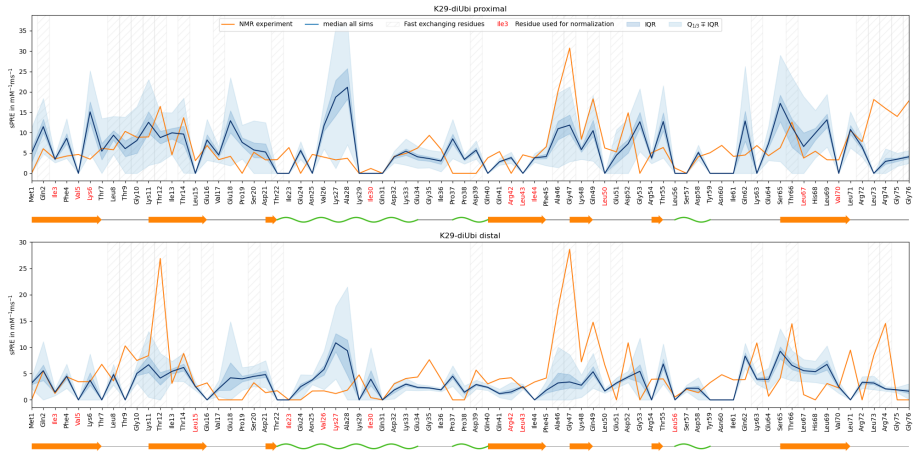

Figure 21: K29 confidence all

Surface coverage for K29 and K33

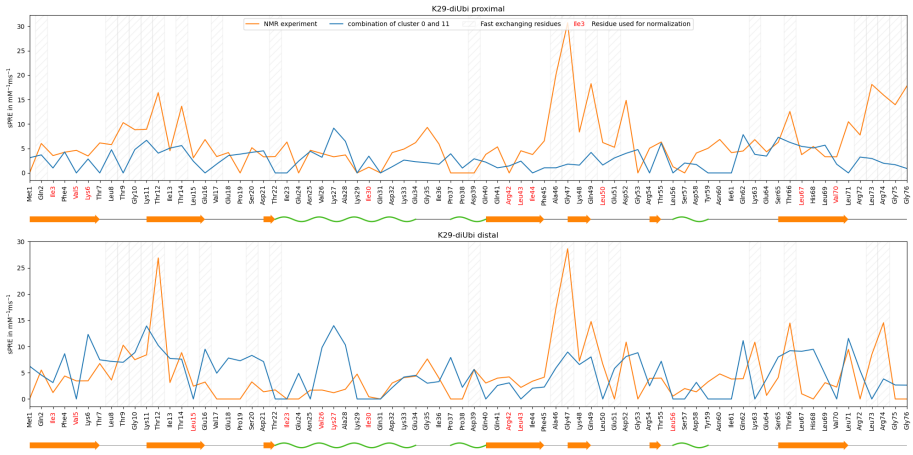

Figure 22: K29 combination clusters 0 and 11.

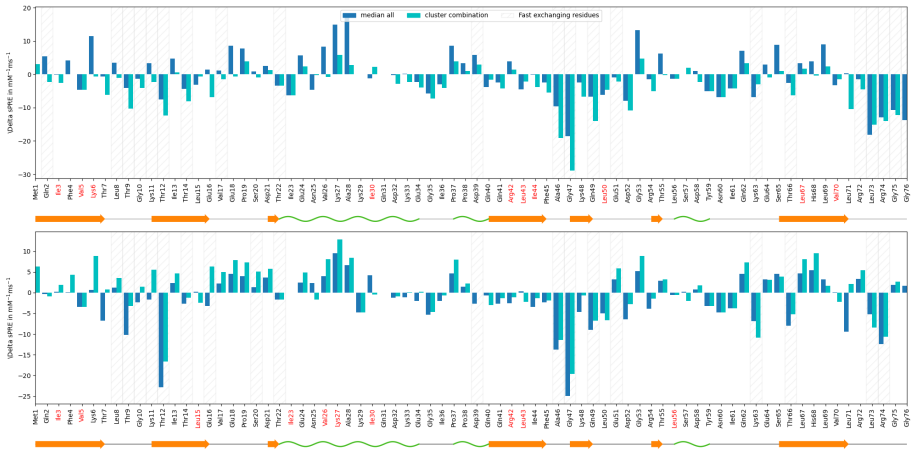

Figure 23: K29 correlation median all (confidence) and combination of clusters 0 and 1.

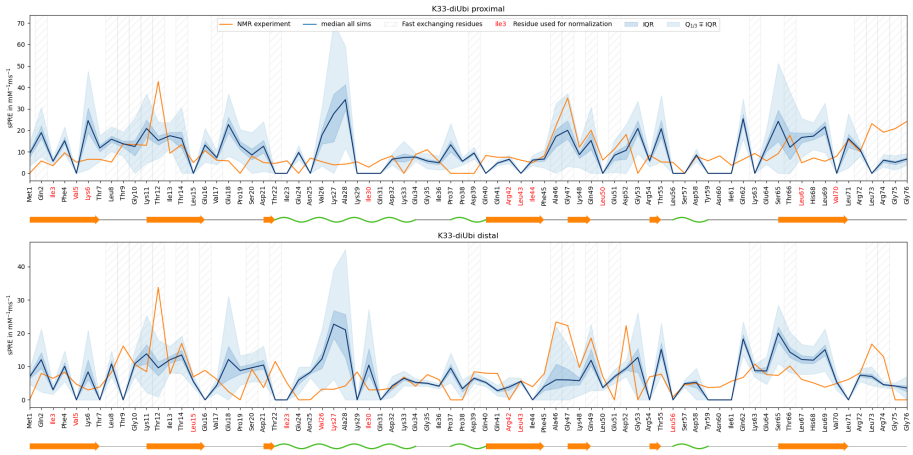

Figure 24: K33 confidence all

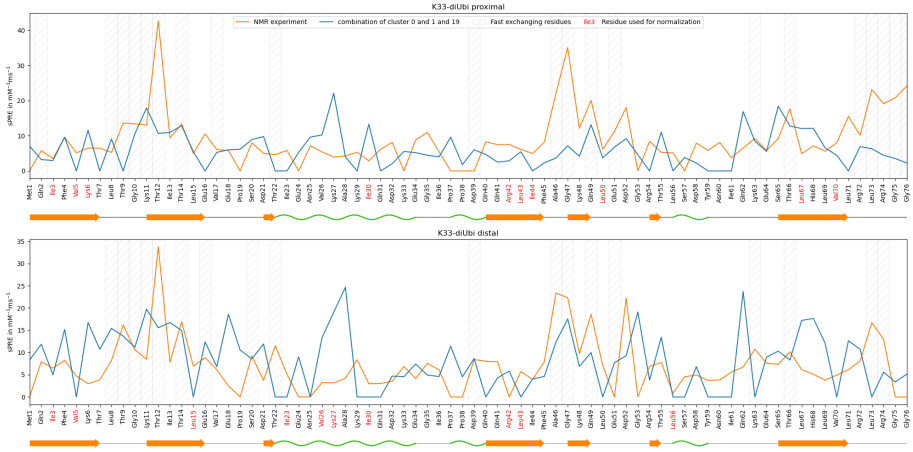

Figure 25: K33 combination clusters 0, 1, and 19.

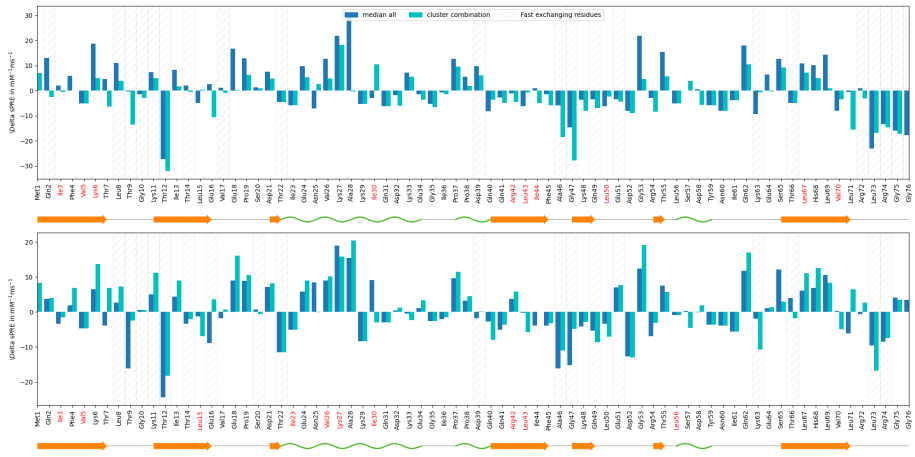

Figure 26: K33 correlation median all (confidence) and combination of clusters 0 and 1.

References

[1] A. Berg, O. Kukharenko, M. Scheffner, and C. Peter, "Towards a molecular basis of ubiquitin signaling: A dual-scale simulation study of ubiquitin dimers," *PLoS computational biology*, vol. 14, no. 11, p. e1006589, 2018.

[2] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook, "Structure of ubiquitin refined at 1.8 Åresolution," *Journal of molecular biology*, vol. 194, no. 3, pp. 531–544, 1987.

[3] T. Lemke and C. Peter, "EncoderMap: Dimensionality reduction and generation of molecule conformations," *Journal of chemical theory and computation*, vol. 15, no. 2, pp. 1209–1215, 2019.

[4] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," *Proceedings of the National Academy of Sciences*, vol. 108, no. 32, pp. 13023–13028, 2011.

[5] L. McInnes, J. Healy, and S. Astels, "Hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[6] Z. Gong, C. D. Schwieters, and C. Tang, "Theory and practice of using solvent paramagnetic relaxation enhancement to characterize protein conformational dynamics," *Methods*, vol. 148, pp. 48–56, 2018.