Methods and results for diUbi XPLOR

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 resolution was taken from [1]. This dataset comprised the ubiquitination sites K6, K29 and K33 where every ubiquitination is comprised of 10 replicas with 10 ms each. The MARTINI v2.2 force field was used for the non-bonded interaction, whereas bonded interactions have been modeled using the IDEN method. These forcefields had to be altered to accommodate the 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.

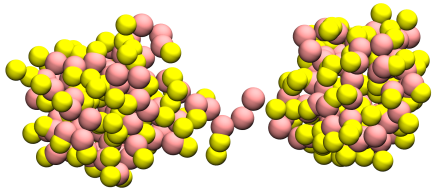

Figure 1: Coarse grained representation of K6-diUbi. Pink beads are unified backbone beads (C_α, N, and C). Yellow beads are unified sidechain beads (one bead for all non-GLY residues). The proximal subunit (right) offers its LYS6 residue to be ubiquitinated by the distal (left) unit's GLY76 residue.

Back-mapped atomistic data

These simulations were used as input-data for a dimensionality reduction algorithm called sketch-map. Here, the high-dimensional phase space of the protein is projected into 2D. The similarity of any two conformations is represented as the distance in sketch-space. Naturally, basins of higher density start to become apparent for statistically significant sub-states of lower free energy. From these basins new all-atom simulations were started from the center of the basin (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) by using Martini's backward script paired with some energy minimization to settle the initial high-energy structure.

Extended and rotamer atomistic data

Besides the back-mapped all atom simulations further atomistic data are available from Berg et al. The starting structures of these simulations are created by slightly altering the position between proximal and distal subunit. As a monomer structure the 1UBQ crystal structure from the protein database was used [2]. For every ubiquitination site, there are 12 simulations with 50 ns available. All atomistic simulations use the GROMOS54a7 forcefield with an addition of the isopeptide bond, which was parametrized from regular peptide bonds.

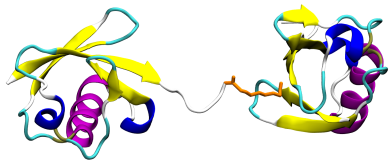

Figure 2: All-atom representation of K6-diUbi. The structure in this figure was used as a basis for Figure 1 and the location of proximal and distal units are identical. Secondary structure has been visualized as cartoon representation and colored accordingly. The residues of the isopeptide bons (GLY76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

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 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

A general overview over the analysis steps: 1. Extract high-dimensional collective variables. 2. Use Encodermap to project into 2D space. 3. Use HDBSCAN to extract statistically-significant sub-states of ensemble. 4. Use XPLOR NIH to calculate the sPRE values for the all-atom conformations (parallelize this task). 5. Normalize the sPRE data from XPLOR. 6. Use sPRE and HDBSCAN to identify some clusters that represent the enxemble available to di-ubiquitin.

Extact high-dimensional collective variables

Collective variables represent a type of data, that is somehow aligned with the raw xyz positional data of a molecular dynamics trajectory. Oftentimes CVs are used to visualize complex dynamic processes. 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 approcah also used by Berg at al. employing residue-wise minimal distances, which makes the input data 152-dimensional.

Encodermap dimensionality reduction

We used the Encodermap's auto-encoder neural network to retrieve a dimensionally reduced representation of all aforementioned simulations [3]. 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. We used the tensorflow python library to minimize three cost functions:

- Center cost: Mean absolute distance between all points and 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

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

- Encodermap cost: Compare the input and the latent space by transforming it with a sigmoid function.

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

,

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 $boldsymbol R_{i,j}$:

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

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 wes using the following parameters:

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

The resulting 2D projection was used as input to HDBSCAN clustering.

HDBSCAN clustering

The hierarchical density-based clustering algorithm HDBSCAN has been tried and tested for molecular dynamics systems and especially the low-diemnsional projections of such [5]. 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 enumbered based on the number of points/conformations in them, with cluster 0 being the largest cluster and so on. Another point to note is, that HDBSCAN offers a great advantage by distinguishing between clusters and noise, which - speaking from an ensemble viewpoint - is comprised of unique structures that carry no significance.

Look at a specific cluster

As an example cluster 0 of K6 diUbi will be further investigated. This cluster consists of 131969 conformations sampled by XXX independent CG and XXX independent AA simulations and thus makes up roughly 12% of the complete

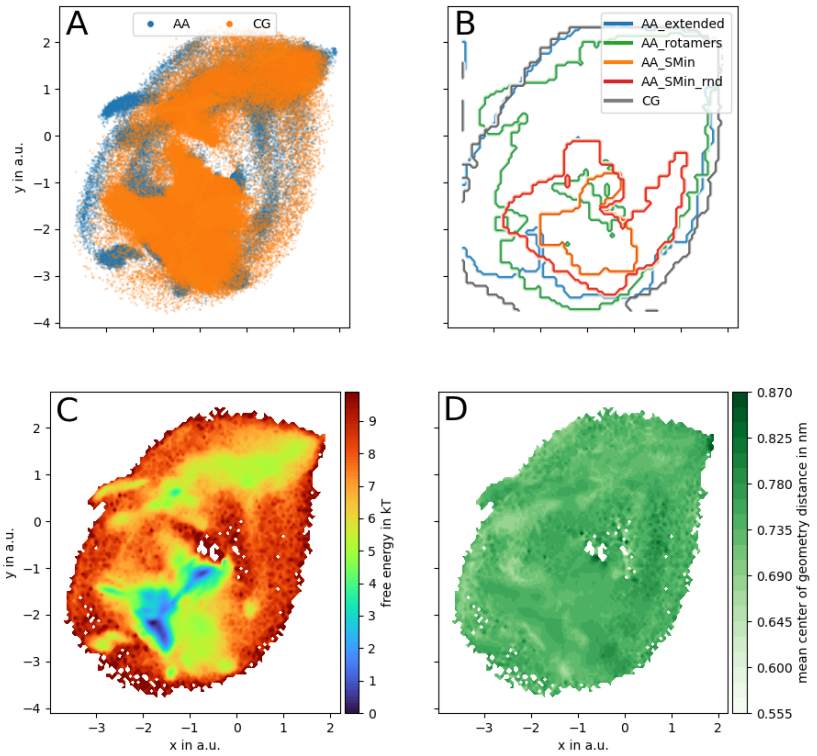

Figure 3: 2D projections of all CG and AA conformations of K6 diUbi. When the low-dimensional coordinates are plotted as a scatter plot, every point can be associated with one conformation sampled by a MD simulation (A). Orange points belong to CG simulations, blue points to AA. The closeness of two 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).

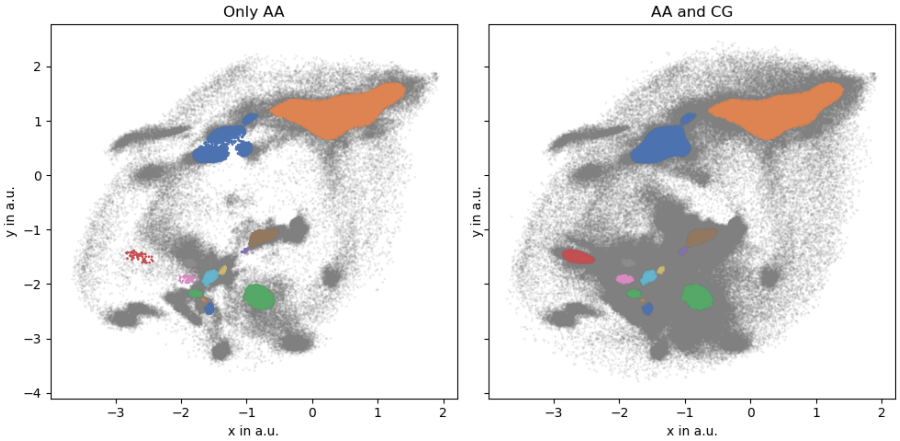

Figure 4: The Encodermap projection of K6 colored according to cluster membership. The advantage of using combined CG and AA data becomes apparent, when the red cluster is investigated. In a combined dataset this cluster contains more than 2500 structurally similar conformations in a dense basin (i.e. 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.

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 and this task succeeded (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 the starting structure. To accelerate this undertaking, we parallelized this computation by using python's joblib process-based multitasking. Every worker was provided with a number of simulation frames it should run XPLOR on and then carried out the task of saving structure files and calling XPLOR.

15-N relaxation times

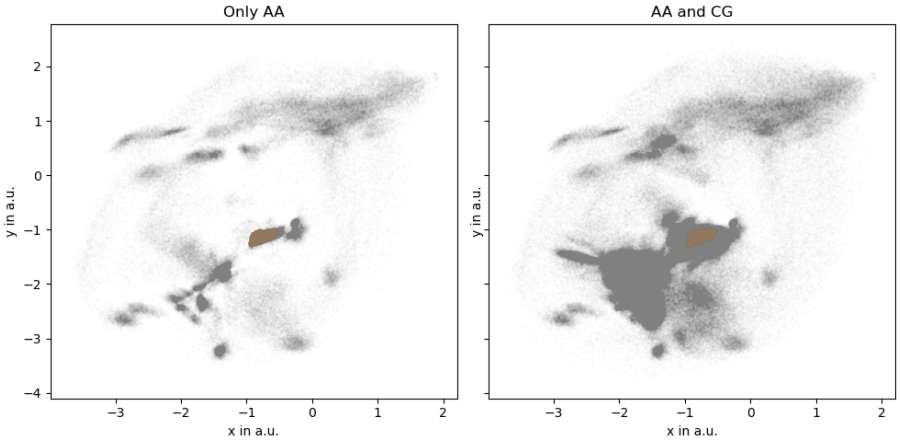

Figure 5: Encodermap projections of K6 diUbi, with cluster 0 highlighted.

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 the old protocol, which did not use the psf files.

IMAGE

Code availability

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

Normalization of XPLOR values The sPRE values from these calculations 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.

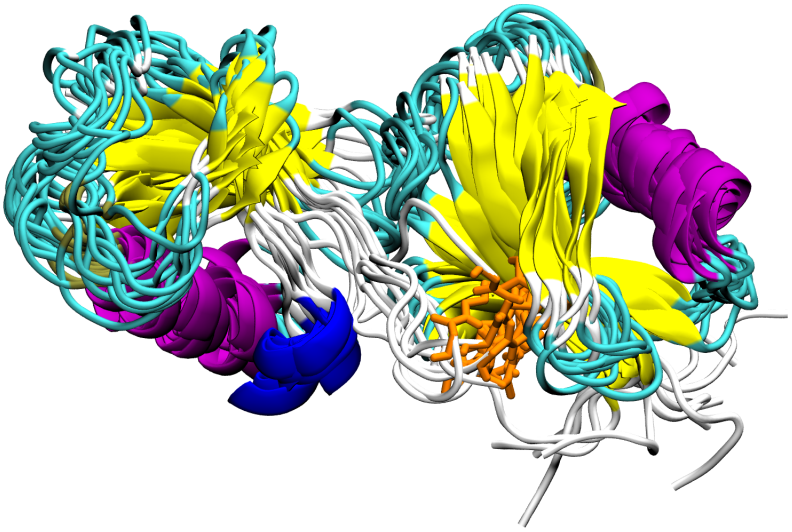

Figure 6: 10 conformations from cluster 0 of K6 diUbi. The structures are rendered as cartoon representation and colored accordingly. The isopeptide LYS and GLY residues are visualized as orange licorice (proximal subunit on the right, distal subunit on the left). It is notable, that the structural difference 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.

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

From New-Mexico experiments we identified some residues to exhibit a fast proteon 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 - and thus returning most reliable sPRE values - 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 $f_{i,ubq,pos}$:

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

[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.