Intersection of multiscale MD and NMR

Kevin Sawade

Universität Konstanz

kevin.sawade@uni-konstanz.de

October 9, 2021

Overview

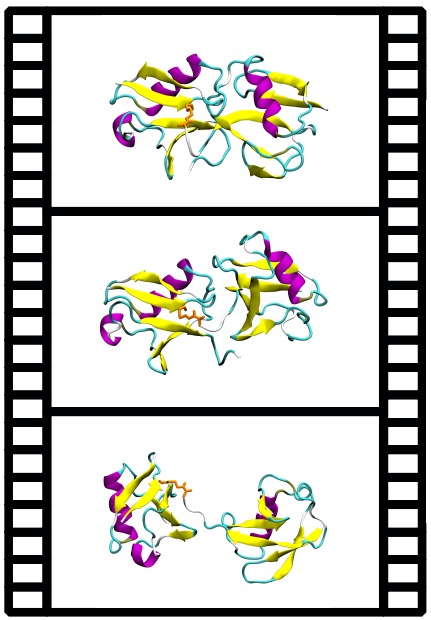
Better sections

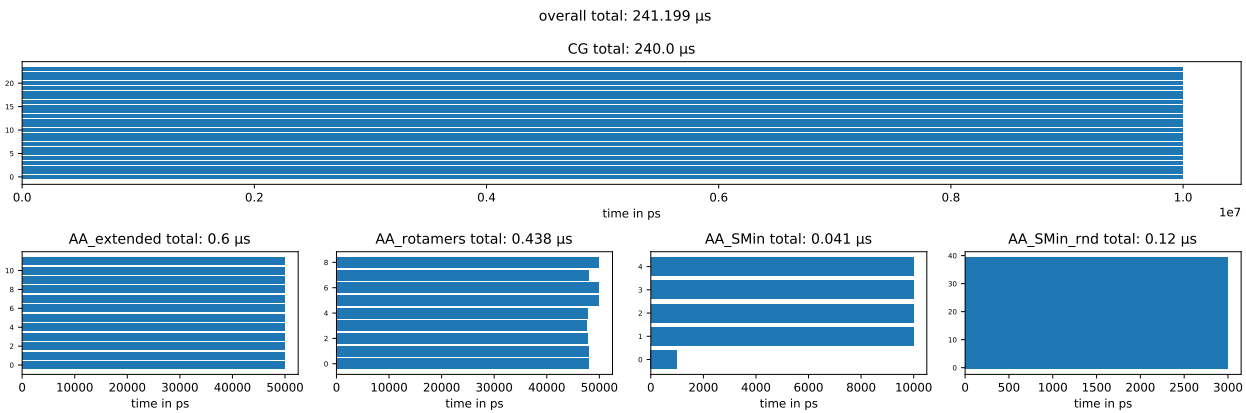
analysis steps: MD, EncoderMap, HDBSCAN, XPLOR
results: best fitting, average of all linear combinations ensemble 15 N cluster RMSD matrix average structure

Data available from previous publications

1. All-atom MD simulations using an altered GROMOS54a7 forcefield. Started from extended conformations.
2. Coarse-grained MD simulations using MARTINI v2.2 ff.
3. All-atom MD simulations started from the 4 lowest sketch-map basins of the CG map using BACKWARD.
4. All-atom MD simulations started from 10 random points around the 4 lowest sketch-map basins of the CG map using BACKWARD.
5. All-atom MD simulations started from χ_3-rotamers of extended structures.

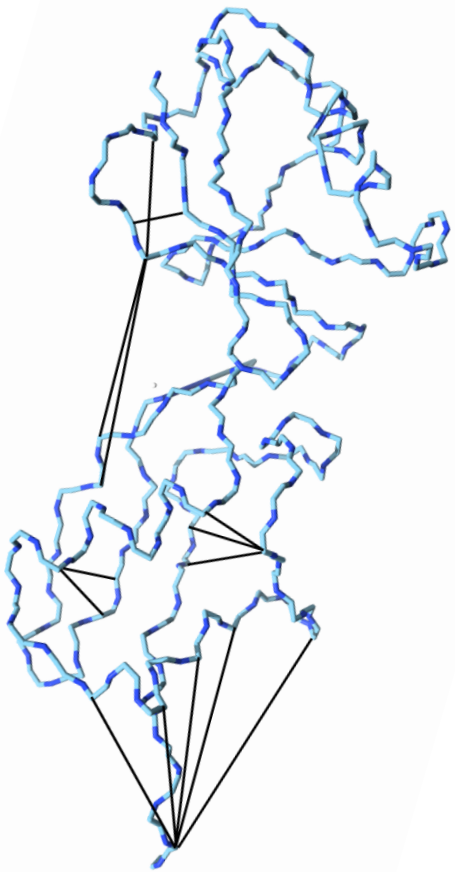
Overview over existing data





Analysis Steps

1. Extract high-dimensional CVs from all simulation frames.



$$\text{distances BA} \quad \begin{matrix} & & \overbrace{\quad\quad\quad}^{\text{distances BA}} \\ & & y_1 \quad \cdots \quad y_{72} \\ \text{distances AB} & \begin{matrix} x_1 \\ \vdots \\ x_{72} \end{matrix} & \begin{pmatrix} d_{a_1,b_1} & \cdots & d_{a_1,b_{72}} \\ \vdots & \ddots & \vdots \\ d_{a_{72},b_1} & \cdots & d_{a_{72},b_{72}} \end{pmatrix} \end{matrix}$$

$$D_{A,B} =$$

$$RWMD_{A,B} = (min(x_1), \ldots min(x_{72}), min(y_1), \ldots min(y_{72}))$$

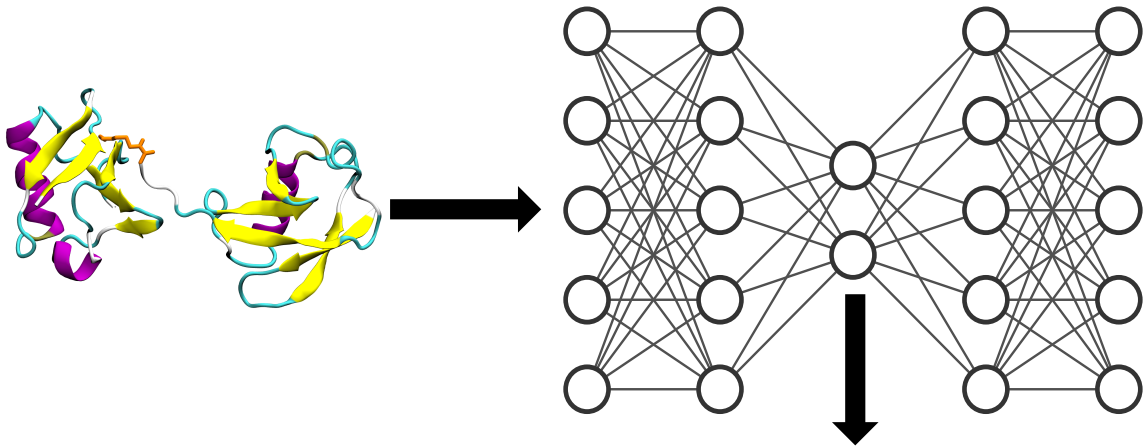
$$D_{B,A} = \begin{matrix} & & \overbrace{\quad\quad\quad}^{\text{distances BA}} \\ & & y_1 \quad \cdots \quad y_{72} \\ \text{distances BA} & \begin{matrix} x_1 \\ \vdots \\ x_{72} \end{matrix} & \begin{pmatrix} d_{b_1,c_1} & \cdots & d_{b_1,c_{72}} \\ \vdots & \ddots & \vdots \\ d_{b_{72},c_1} & \cdots & d_{b_{72},c_{72}} \end{pmatrix} \end{matrix}$$

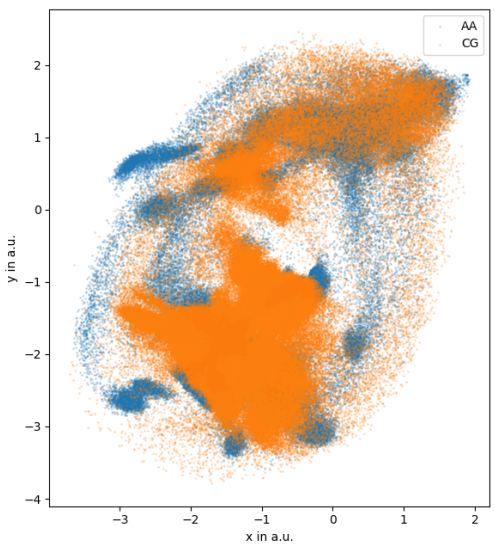
$$RWMD_{B,A} = (min(x_1), \ldots min(x_{72}), min(y_1), \ldots min(y_{72}))$$

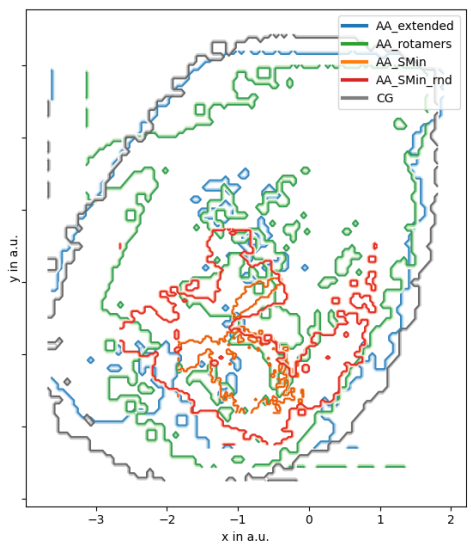
$$RWMD = (RWMD_{A,B} \cup RWMD_{B,A})$$

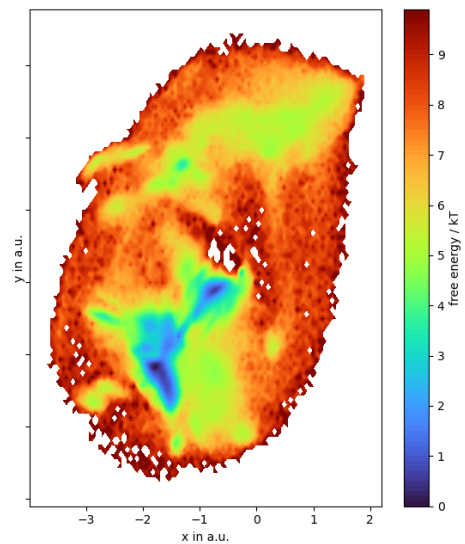
Analysis Steps

2. Run EncoderMap with the high-dimensional CVs as input data.



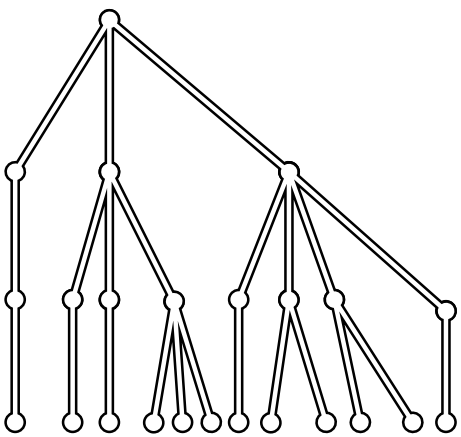


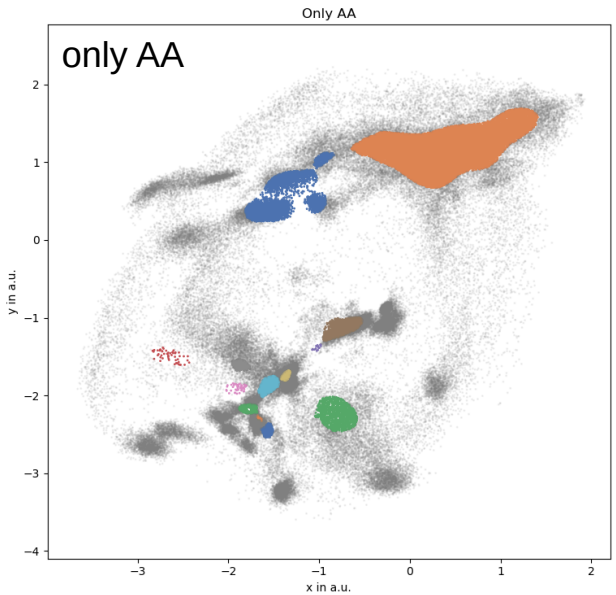


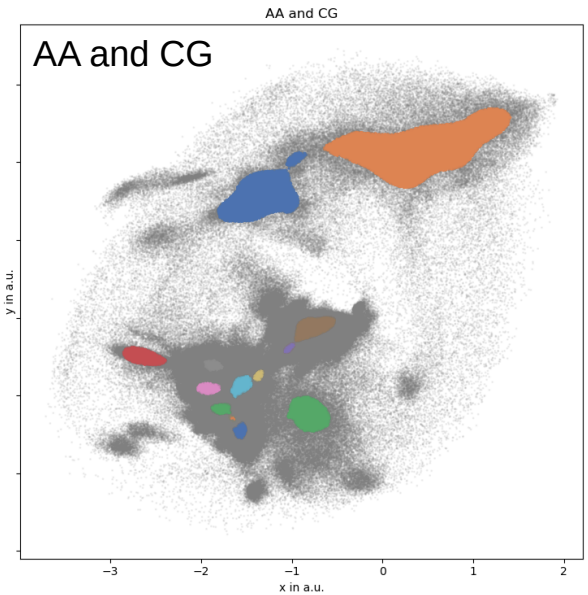


Analysis Steps

3. Cluster data using HDBSCAN.







Analysis Steps

4. Write a code library, that poses as an interface between current python packages and XPLOR. (`https://github.com/kevinsawade/xplor_functions`)
5. Identify possible settings/arguments and define reasonable defaults (solution concentration, probe radius, etc.)

```
psol:
 call_parameters:
  name:
   type: str
   value: psol
   descr: |
    This is the name of the potential term assigned to this PSolPot object
    It can contain any string and can be used to
  restraints:
   type: file
   value: data/diUbi_k6_sPRE_in.tbl
   descr: |
    The location of the spre_tble file that will be passed to XPLOR.
    The spre table file needs to be formatted as such:
    f"assign (resid {resSeq:<2} and name HN)  {sPRE:5.3f} {err:5.3f}"
    So for example: For Ubiquitin the first three lines of that table look like this:
    assign (resid 2  and name HN)   5.510   0.711
    assign (resid 3  and name HN)   1.223   1.816
    assign (resid 4  and name HN)   4.381   0.402
  tauc:
   type: float
   value: 0.2
   descr: correlation time
  probeR:
   type: float
   value: 3.5
   descr: radius of probe molecule
  probeC:
   type: float
   value: 0.24
```

```
def parallel_xplor(ubq_sites, simdir='/home/andrejb/Research/SIMS/2017_*', n_threads='max-2',
                   df_outdir='/home/kevin/projects/tobias_schneider/values_from_every_frame/from_package/',
                   suffix='_df_no_conect.csv', write_csv=True, fix_isopeptides=True, specific_index=None, parallel=False,
                   subsample=5, yaml_file='', testing=False, from_tmp=False, max_len=-1, break_after=False, **kwargs):
    """Runs xplor on many simulations in parallel.

    This function is somewhat specific and there are some hardcoded directories
    in it. It uses MDTraj and OpenMM to load trajectories from Andrej's sim
    directory (/home/andrejb/Research/SIMS/). These trajectories are provided
    in a joblib Parallel/delayed construct to `get_series_from_mdtraj`, which
    results in a list of pandas Series, that are stacked to a long dataframe.

    The dataframe is periodically saved (to not loose anything). Check out the
    function `xplor.delete_old_csvs` to remove the unwanted intermediate csvs,
    produced by this function.

    Args:
        ubq_sites (list): A list of ubiquitination sites, that should be recognized.

    Keyword Args:
        simdir (str, optional): Path to the sims, that contain the ubq_site substring.
            Defaults to '/home/andrejb/Research/SIMS/2017_*'
        n_threads (Union[int, str], optional): The number of threads to run.
            Can be an int, but also 'max' or 'max-2', where 'max' will give
            make this function use the maximum number of cores. 'max-2' will use
            all but 2 cores. Defaults to 'max-2'
        df_outdir (str, optional): Where to save the csv files to. Defaults to
            '/home/kevin/projects/tobias_schneider/values_from_every_frame/from_package/'.
        suffix (str, optional): Suffix of the csv files, used to sort different
            runs. Defaults to '_df_no_conect.csv'.
        write_csv (bool, optional): Whether to write the csv to disk. Defaults
            to True.
        subsample (int, optional): Whether to subsample trajectories. Give an
            int and only use every `subsample`-th frame. Defaults to 5.
        max_len (int, optional): Only go to that maximum length of a trajectory.
            Defaults to -1, wich will use the full length of the trajectories.
        yaml_file (str, optional): Path to a yaml file. If an empty string is provided
            the defaults.yml file from xplor/data is loaded. Defaults to ''.
        from_tmp (bool, optional): Changes the executable of the command to
            work with an ssh interpreter to 134.34.112.158. If set to false,
            the executable will be taken from xplor/scripts.
            Defaults to False
        testing (bool, optional): Adds the '-testing' flag to the command.
            Defaults to False.
        specific_index (Union[int, None], optional): If given, only that index will
            be used in the parallel loop. For Debugging. Defaults to None.
        parallel (bool): Whether to do the calculations in parallel. Defaults to False
```

Analysis Steps

6. Parallelize the calculations for faster throughput.
7. Manually parse .psf files to include 15N relaxation data (very time consuming calculations because every pdb file needs to be changed according to psf atom names).

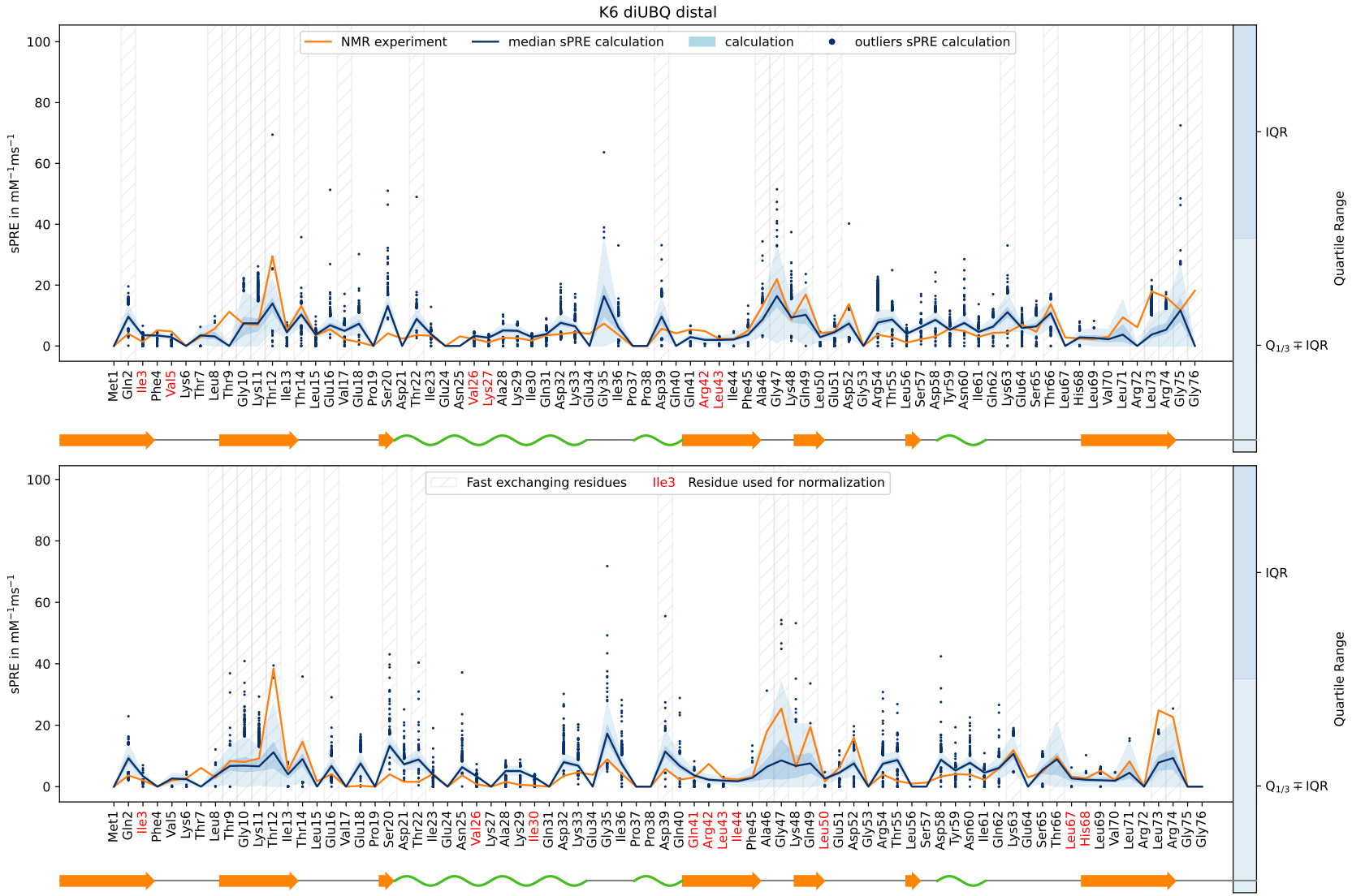
Analysis Steps

8. Normalize sPRE computations.

- Don't consider the fast-exchanging residues.
- Consider proximal and distal unit separately.
- From all simulation frames calculate the variance of the sPRE values for all residues.
- Take the 10 (not fast exchanging) residues with the smallest variances.
- Calculate the factor f_i from $f_i = \frac{v_{i,exp}}{v_{i,sim}}$ for every of these 10 residues.
- Calculate the mean of these ten factors as $F = \frac{\sum f_i}{N}$
- Use the proximal factor F as a factor to normalize the sPRE values of the proximal unit.
- Use the distal factor to normalize sim. sPRE values of distal unit.

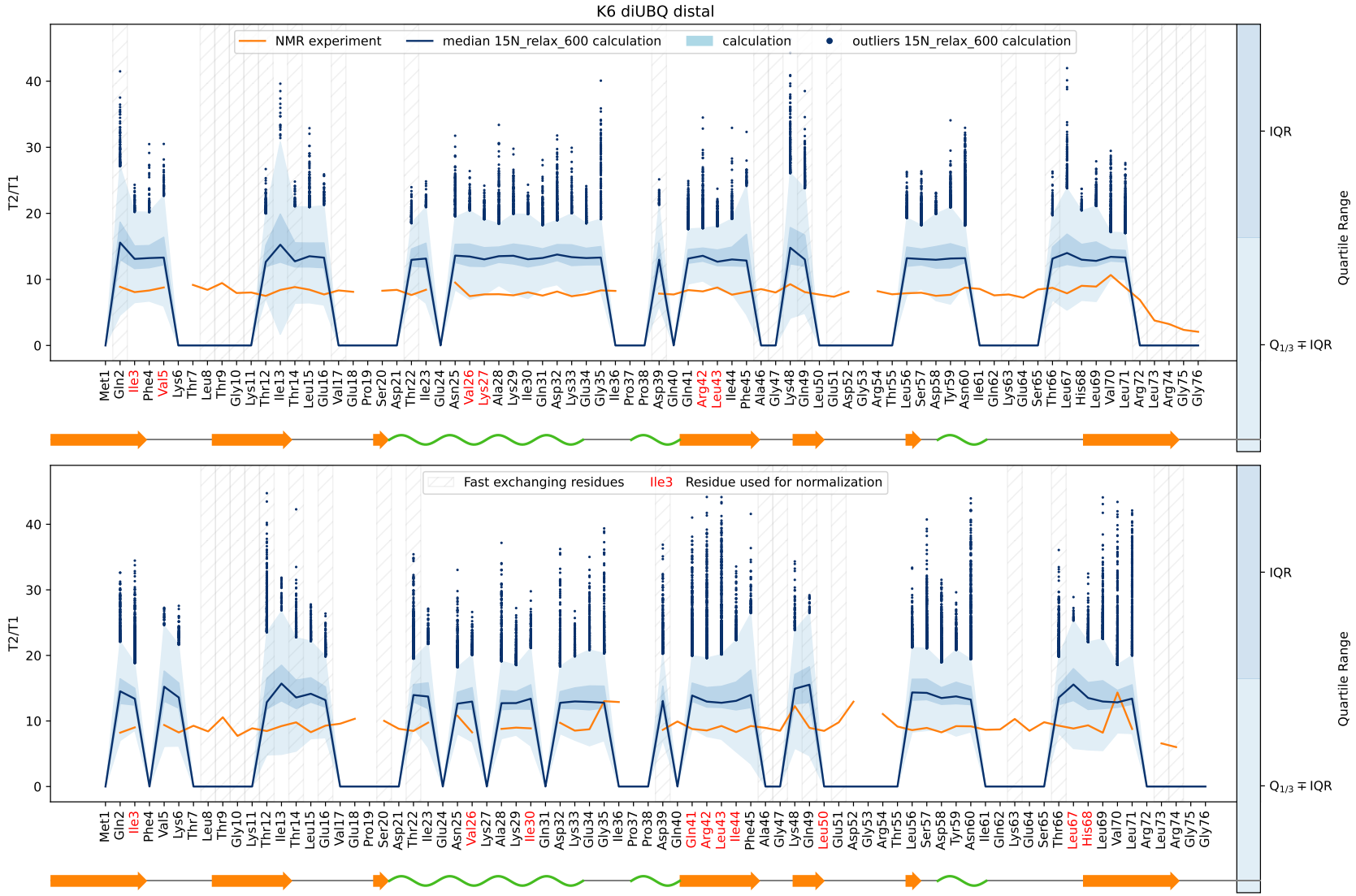
sPRE results

Gorgeous 3



15N results

Bad :(



Calculating cluster coefficients

Solve:

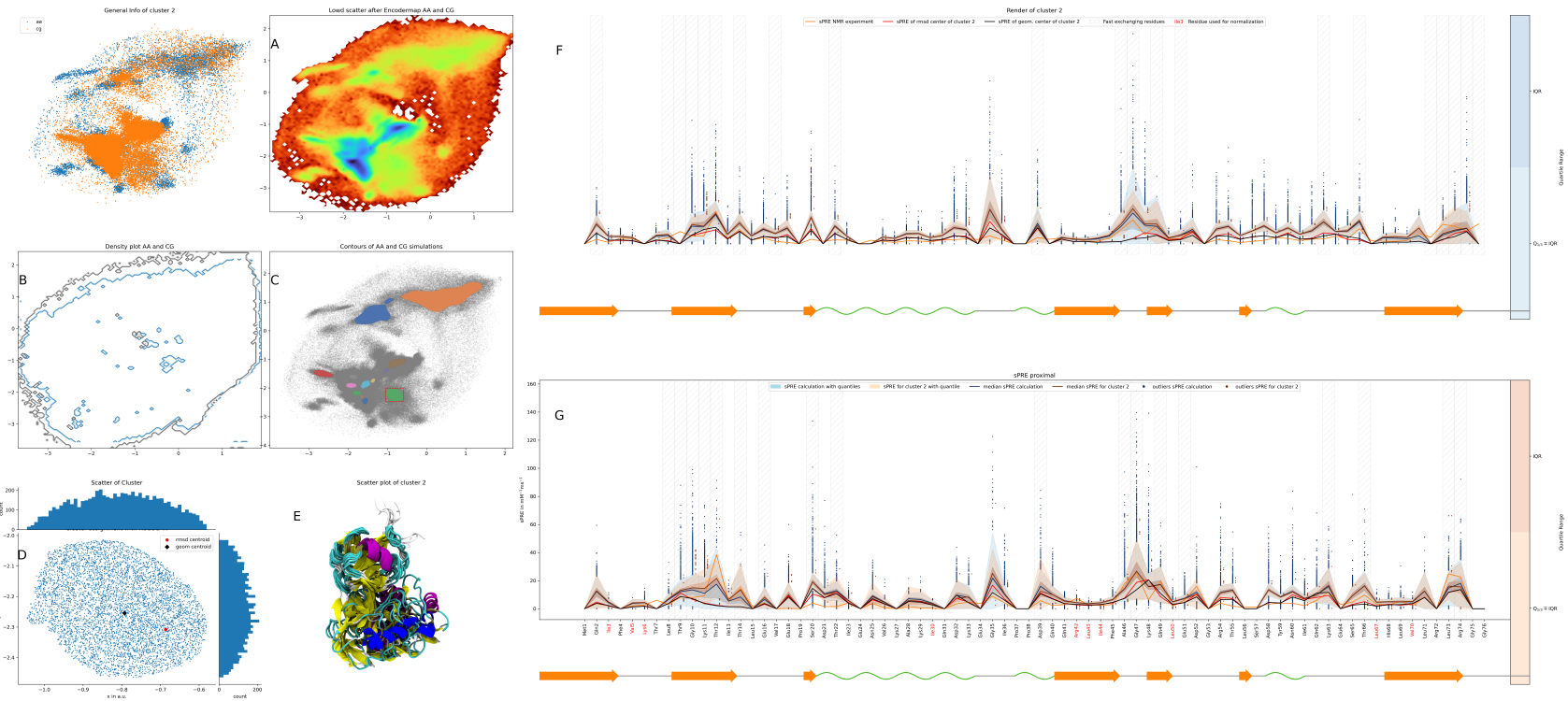
$$\begin{bmatrix} v_{exp,MET1} \\ v_{exp,GLN2} \\ \vdots \\ v_{exp,GLY76} \end{bmatrix} = x_1 \cdot \begin{bmatrix} v_{clu_1,MET1} \\ v_{clu_1,GLN2} \\ \vdots \\ v_{clu_1,GLY76} \end{bmatrix} + x_2 \cdot \begin{bmatrix} v_{clu_2,MET1} \\ v_{clu_2,GLN2} \\ \vdots \\ v_{clu_2,GLY76} \end{bmatrix} + \ldots + x_n \cdot \begin{bmatrix} v_{clu_n,MET1} \\ v_{clu_n,GLN2} \\ \vdots \\ v_{clu_n,GLY76} \end{bmatrix}$$

for $\{x_1, x_2, ..., x_n \in \mathbb{R} | 0 \leq x_n \leq 1\}$
and $\sum_i^n x_i \stackrel{!}{=} 1$

Single cluster

A cluster is extracted from the complete (CG and AA) ensemble but is rendered using only the AA conformations. A cluster is defined by:

- Its contribution to the whole ensemble.
- A coefficient from the linear combination of clusters.



cluster_num	percent of aa frames	percent of cg frames	percent in full ensemble	coefficient in linear combination	mean abs difference of cluster mean to exp values
0	32%	68%	1%	4.6e-16	5.00
1	74%	26%	2%	1.29e-01	4.92
2	18%	82%	1%	7.31e-16	5.14
4	0%	100%	1%	0.e+00	5.32
5	3%	97%	12%	0.e+00	4.97
6	0%	100%	2%	5.84e-01	5.07
7	3%	97%	1%	0.e+00	4.98
8	17%	83%	1%	0.e+00	5.07
9	4%	96%	4%	9.19e-16	5.10
10	1%	99%	4%	1.87e-16	5.06
11	3%	97%	0%	2.87e-01	5.09
12	1%	99%	9%	7.82e-16	5.15

Table: New values

cluster_num	percent of aa frames	percent of cg frames	percent in full ensemble	coefficient in linear combination	mean abs difference of cluster mean to exp values
0	32%	68%	1%	0.e+00	5.00
1	74%	26%	2%	0.e+00	4.92
2	18%	82%	1%	4.57e-15	5.14
4	0%	100%	1%	1.72e-14	5.32
5	3%	97%	12%	0.e+00	4.97
6	0%	100%	2%	4.27e-01	5.07
7	3%	97%	1%	6.33e-15	4.98
8	17%	83%	1%	0.e+00	5.07
9	4%	96%	4%	8.91e-15	5.10
10	1%	99%	4%	2.36e-15	5.06
11	3%	97%	0%	0.e+00	5.09
12	1%	99%	9%	5.7e-15	5.15

Table: Analysis of the K6 clusters.

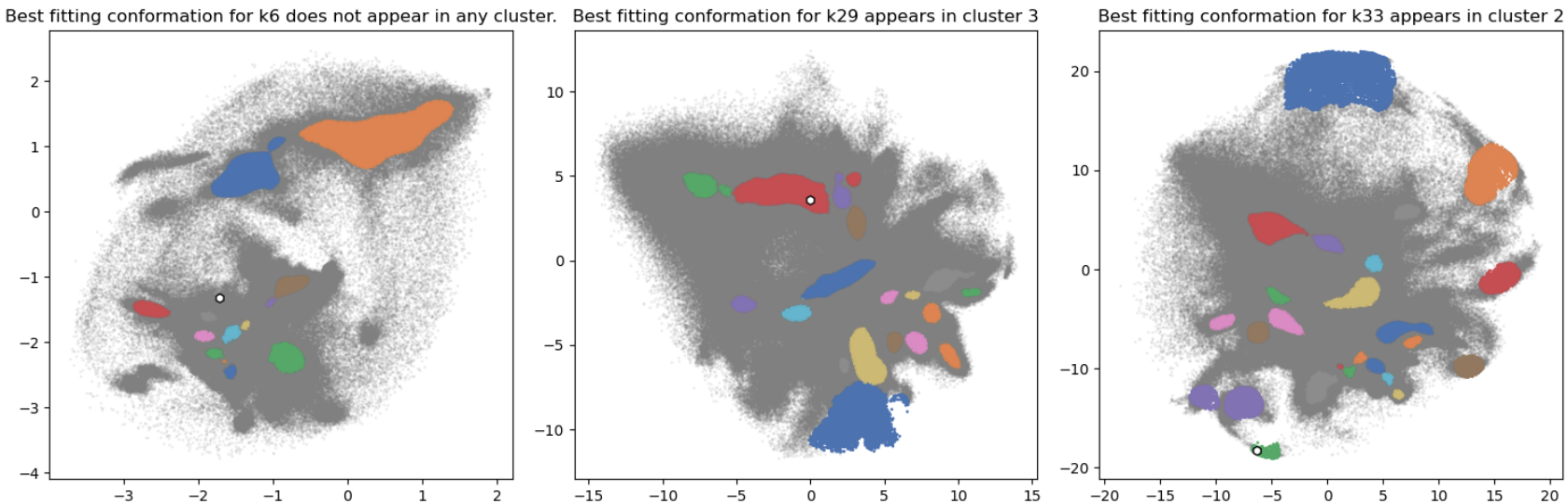
cluster_num	percent of aa frames	percent of cg frames	percent in full ensemble	coefficient in linear combination	mean abs difference of cluster mean to exp values
0	3%	97%	17%	4.18e-01	3.21
1	0%	100%	1%	5.82e-01	3.15
2	15%	85%	1%	2.98e-14	3.79
3	16%	84%	4%	3.31e-15	3.27
4	7%	93%	0%	3.29e-14	4.28
5	2%	98%	1%	5.44e-17	3.19
6	5%	95%	3%	1.21e-14	3.35
8	1%	99%	9%	0.e+00	3.31
9	7%	93%	1%	0.e+00	3.24
10	13%	87%	2%	0.e+00	3.33
12	1%	99%	0%	0.e+00	3.14
13	36%	64%	0%	0.e+00	4.77
14	5%	95%	1%	4.12e-14	4.21
15	7%	93%	1%	2.17e-14	3.60
16	3%	97%	1%	2.86e-14	4.82
17	2%	98%	4%	4.56e-15	3.25
18	2%	98%	0%	2.84e-14	4.16

Table: Analysis of the K29 clusters.

cluster_num	percent of aa frames	percent of cg frames	percent in full ensemble	coefficient in linear combination	mean abs difference of cluster mean to exp values
0	100%	0%	1%	7.43e-15	4.34
1	100%	0%	0%	0.e+00	4.75
2	100%	0%	0%	1.30e-15	4.06
3	100%	0%	0%	0.e+00	4.36
4	81%	19%	1%	0.e+00	4.38
5	1%	99%	0%	0.e+00	4.90
7	3%	97%	0%	8.89e-02	4.41
8	18%	82%	2%	0.e+00	4.47
9	6%	94%	0%	0.e+00	4.53
10	2%	98%	3%	0.e+00	4.36
11	3%	97%	1%	0.e+00	4.55
12	11%	89%	1%	0.e+00	4.21
13	47%	53%	3%	0.e+00	4.16
14	34%	66%	1%	0.e+00	4.15
15	9%	91%	2%	0.e+00	4.12
16	4%	96%	2%	8.96e-02	3.91
17	2%	98%	10%	0.e+00	4.48
18	1%	99%	1%	0.e+00	4.47
19	1%	99%	2%	5.54e-02	4.60
20	1%	99%	8%	0.e+00	4.45
21	2%	98%	2%	0.e+00	4.57
22	1%	99%	3%	0.e+00	4.45
23	4%	96%	0%	0.e+00	4.05

Table: Analysis of the K33 clusters.

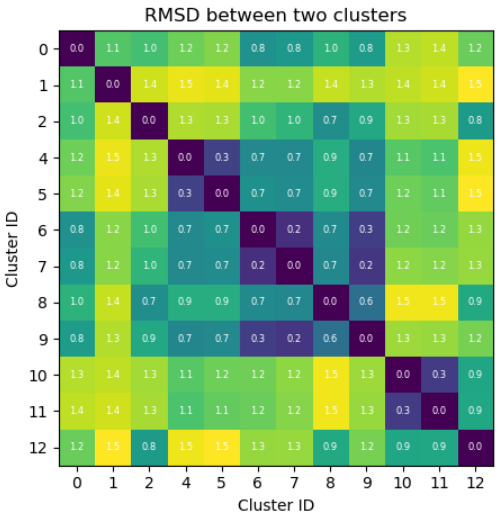
Is the best fitting structure in a cluster?

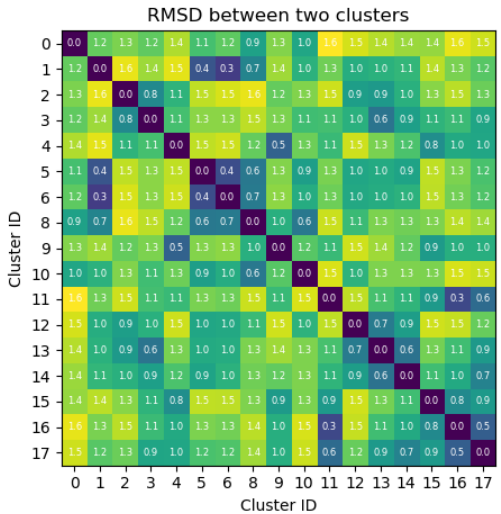


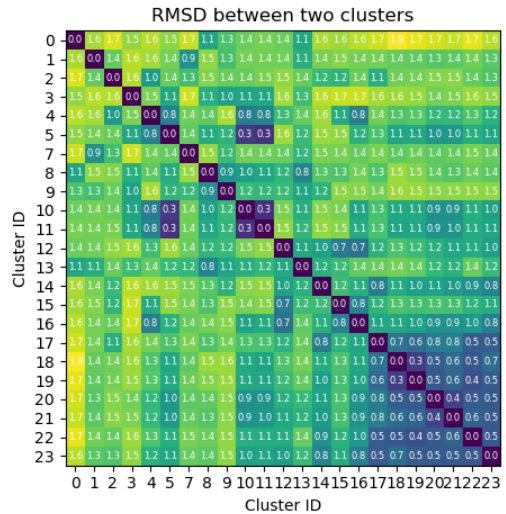
Mean of everything vs linear combination vs exp

todo

RMSD matrices for clusters

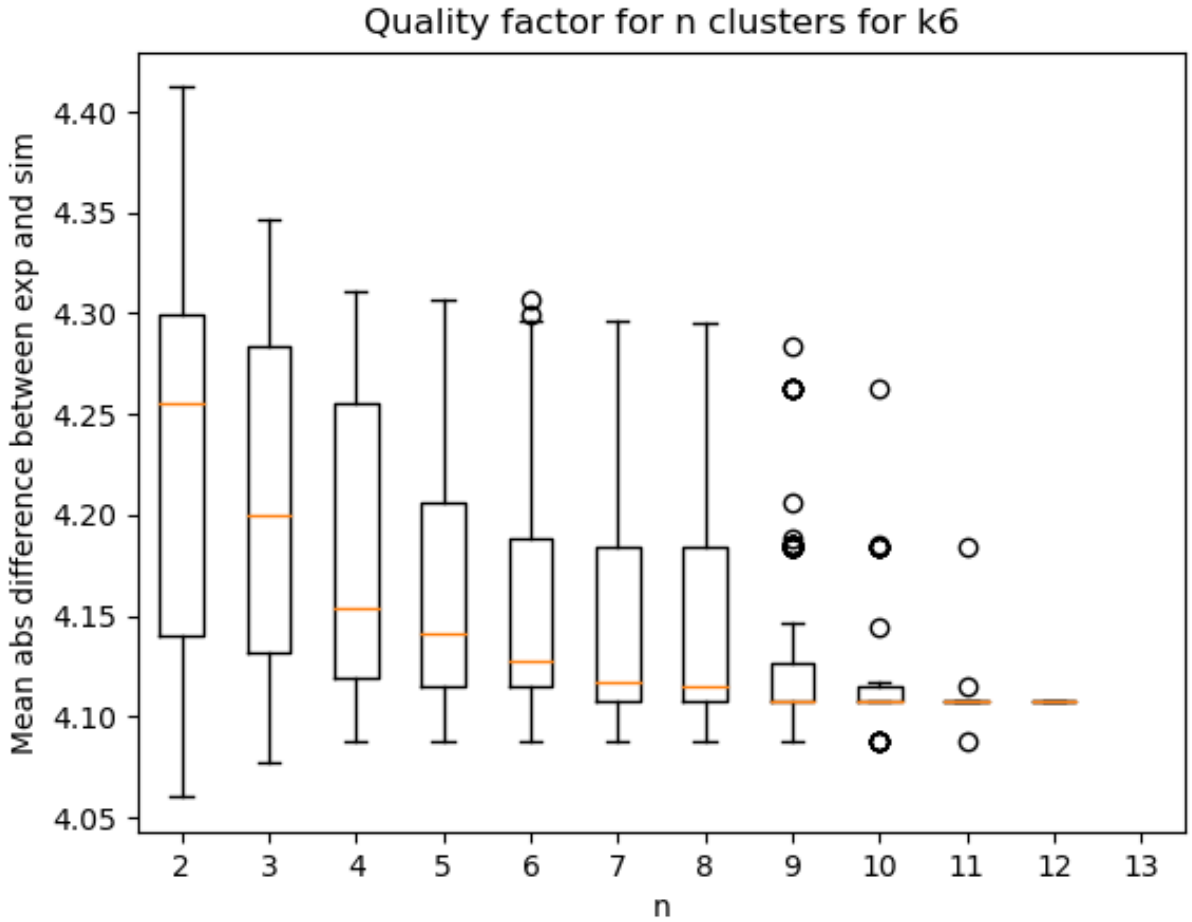






Multiple clusters

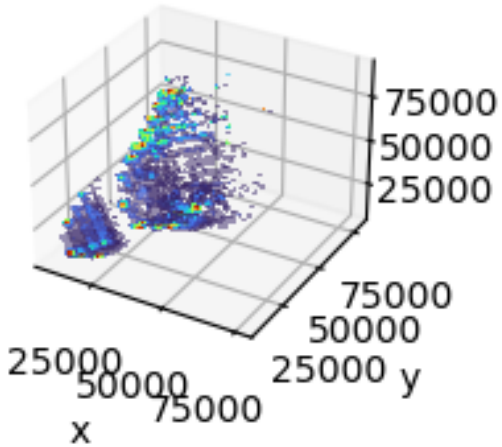
Multiple clusters better represent the nature of the sPRE ensemble. A large coefficient in the linear combination does not necessarily mean that this cluster has similar sPRE values as the experiment.

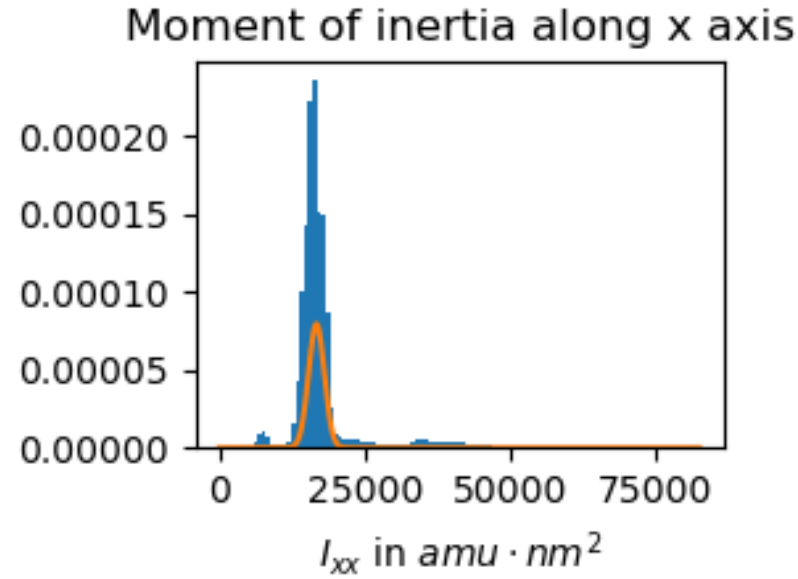


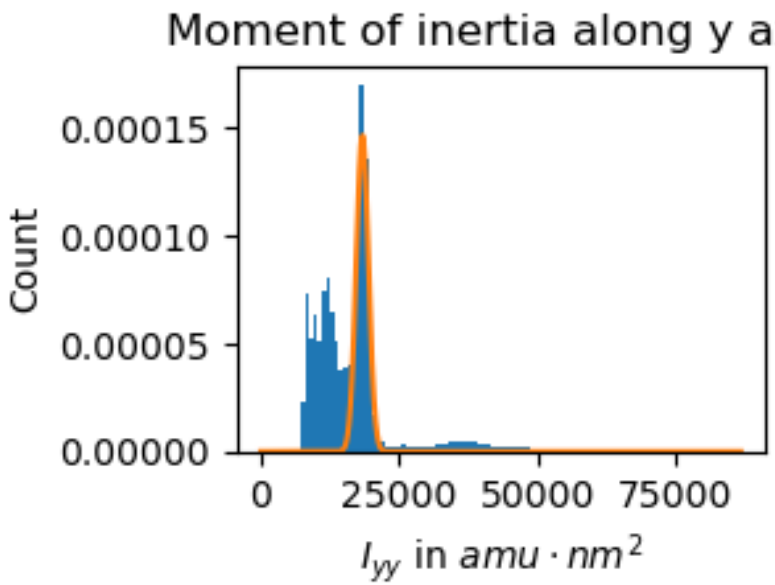
There's always one low-score outlier. Good cluster?

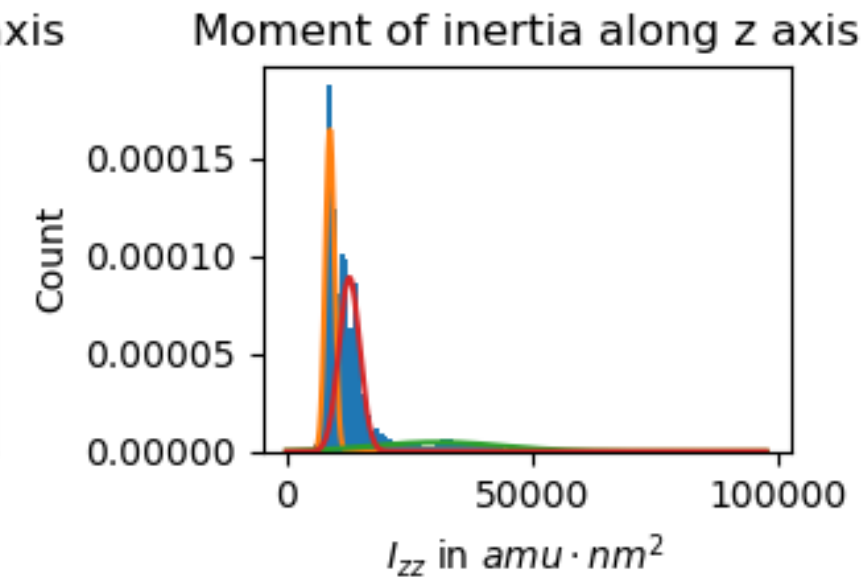
Todo

Test tensors of inertia





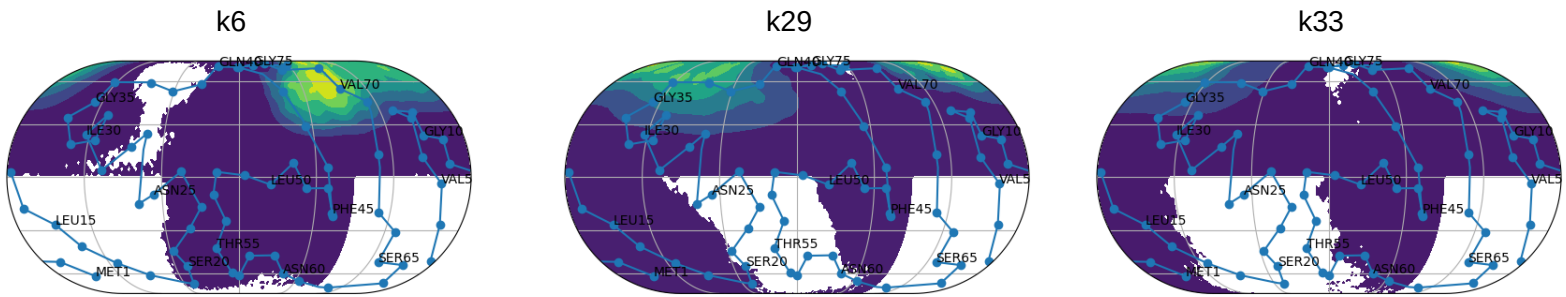




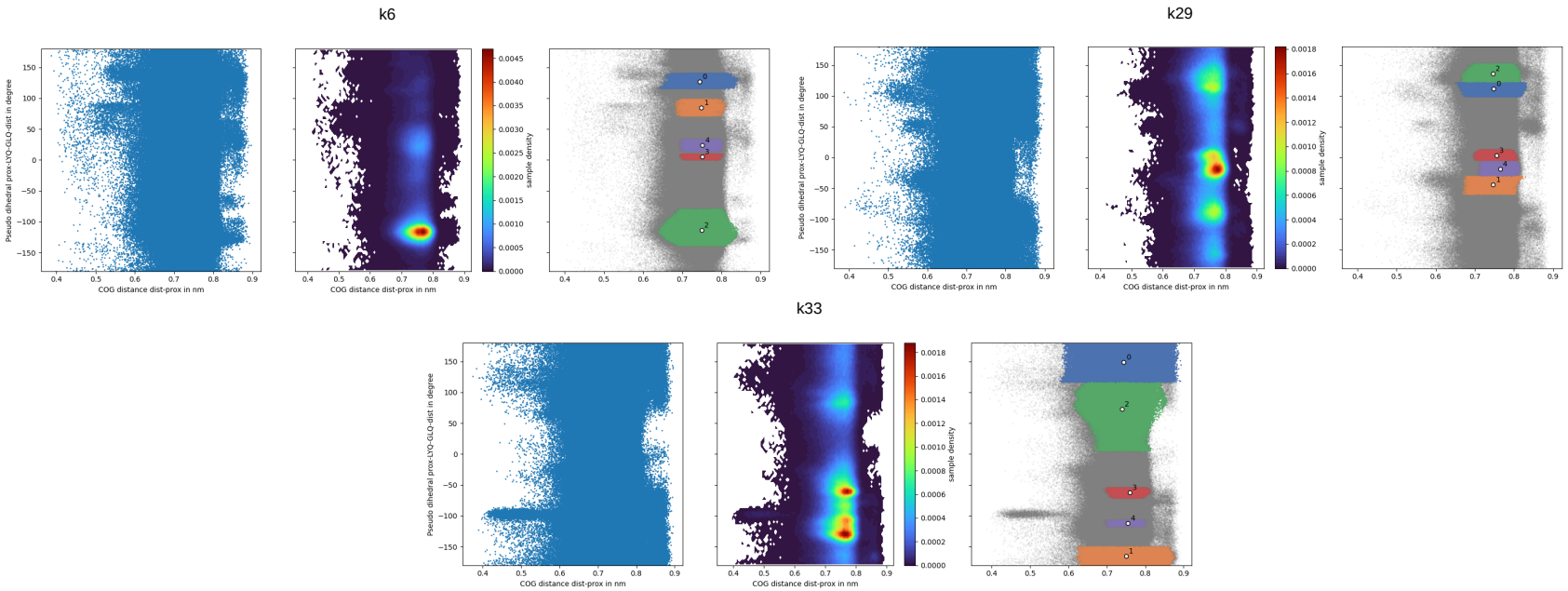
Render of a structures within a I_xx gaussian

todo bad image

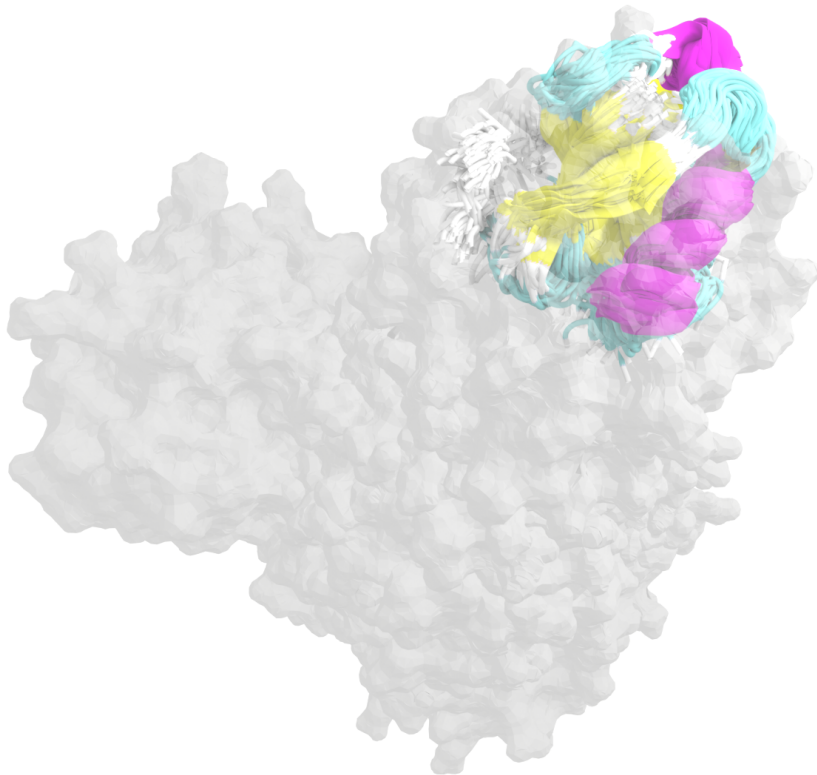
Surface coverage might also be a possibility.



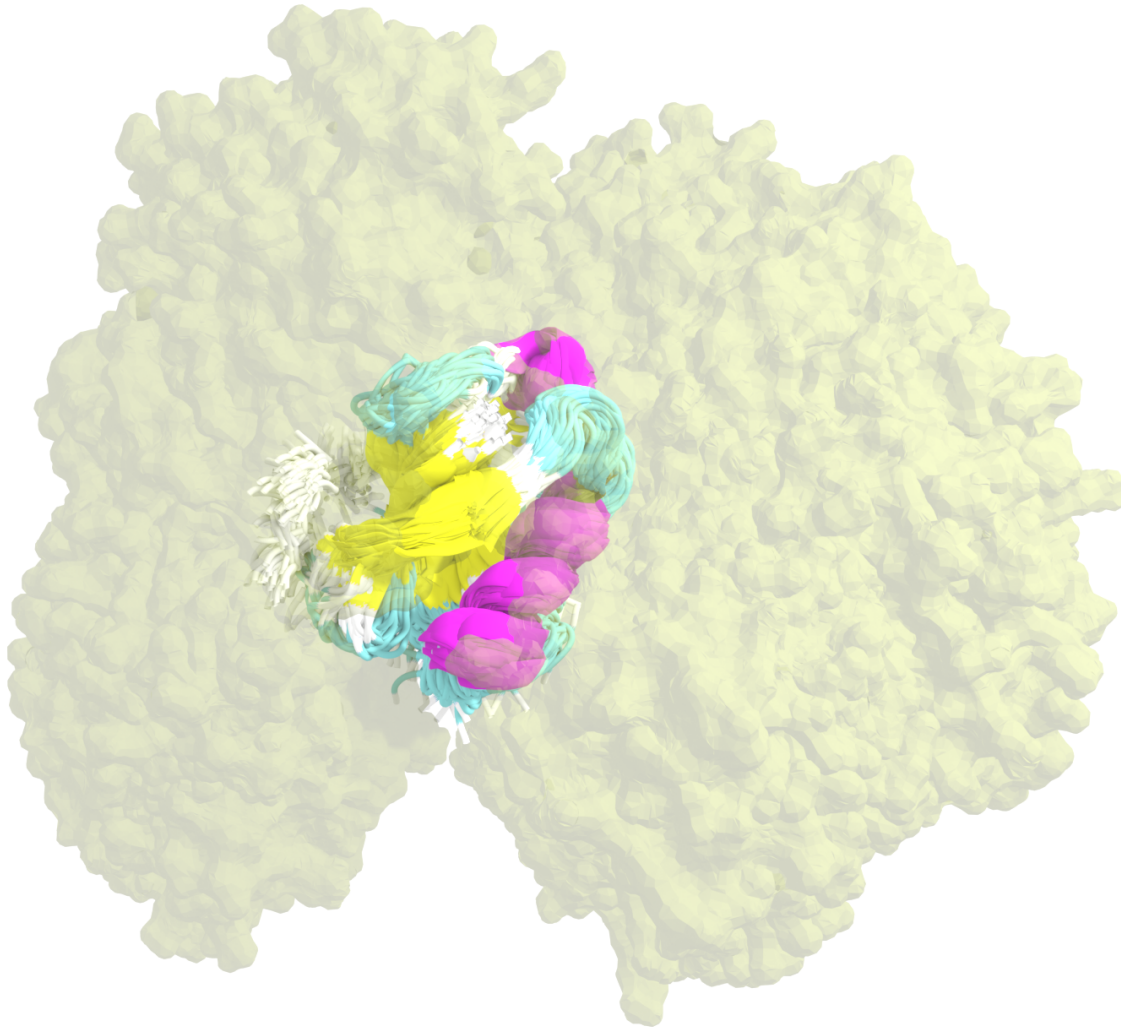
Better: pseudo-dihedral and cog-distance



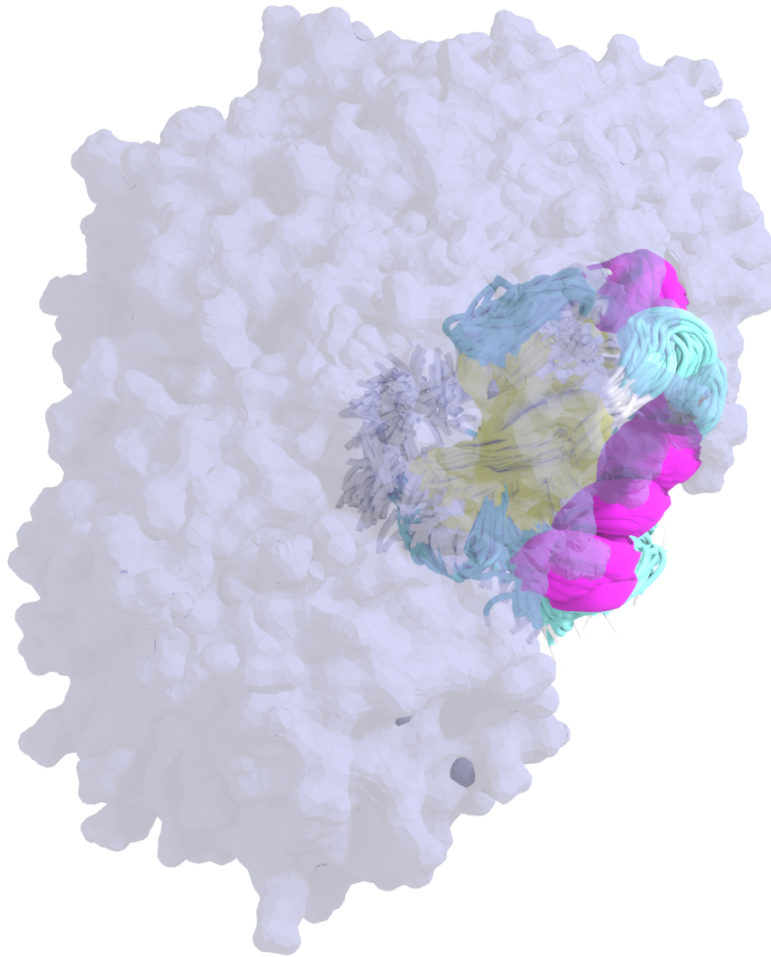
General shape of K6



General shape of K29



General shape of K30



The End