Supplementary information

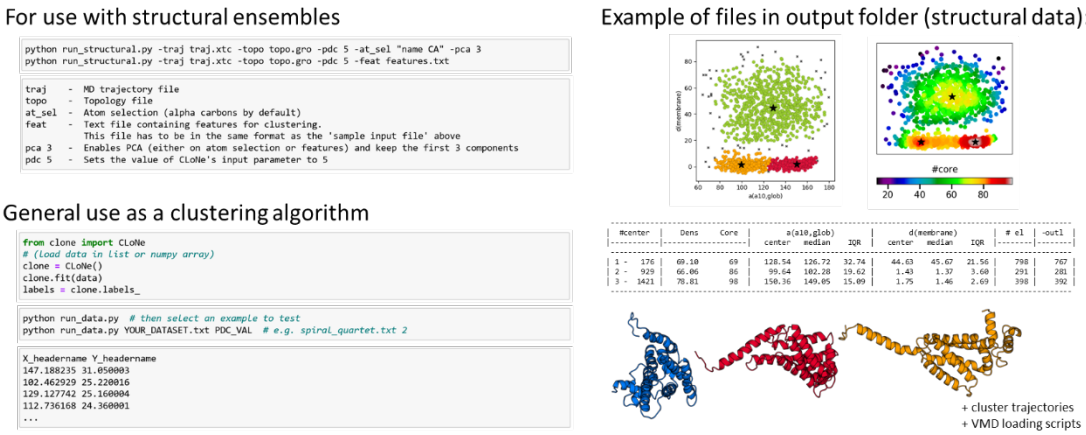
CLoNe: Automated clustering based on local density neighborhoods for application to biomolecular structural ensembles

Träger Sylvain[1,2], Tamò Giorgio[1,2], Aydin Deniz[1,2], Fonti Giulia[1,2], Audagnotto Martina[1,2] and Dal Peraro Matteo[*,1,2]

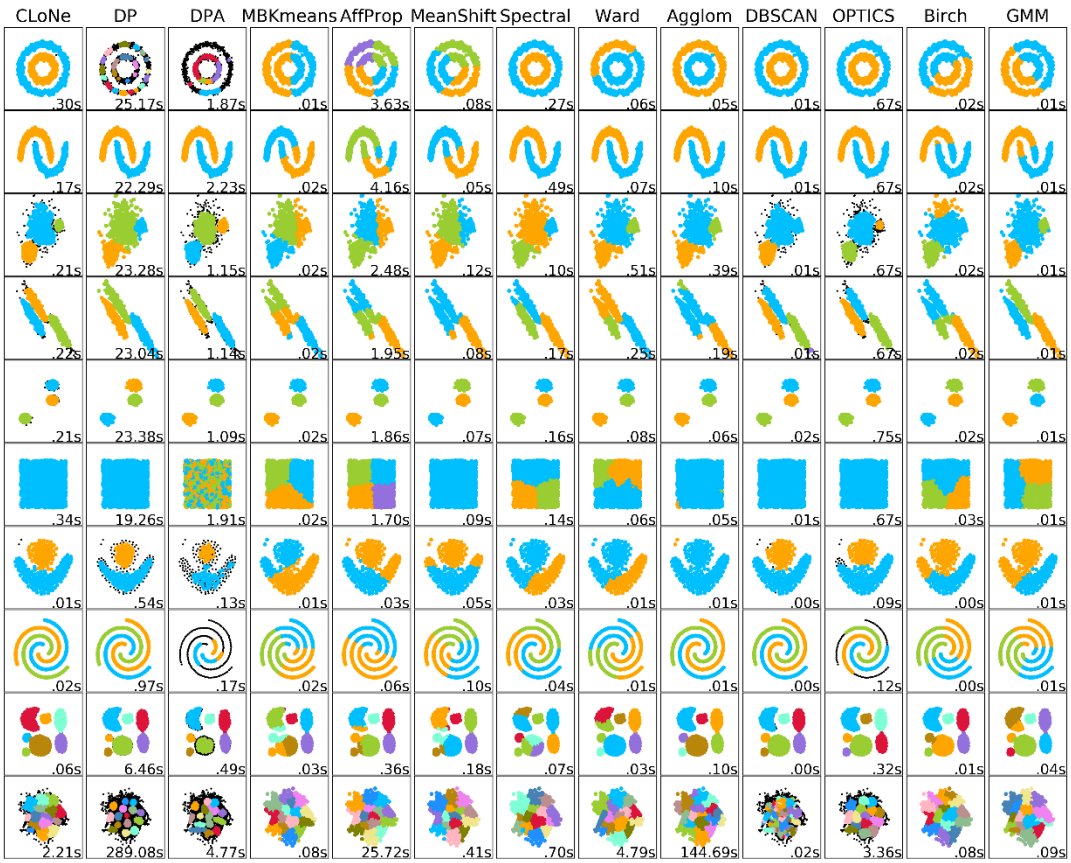
[1]Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

*To whom correspondence should be addressed.

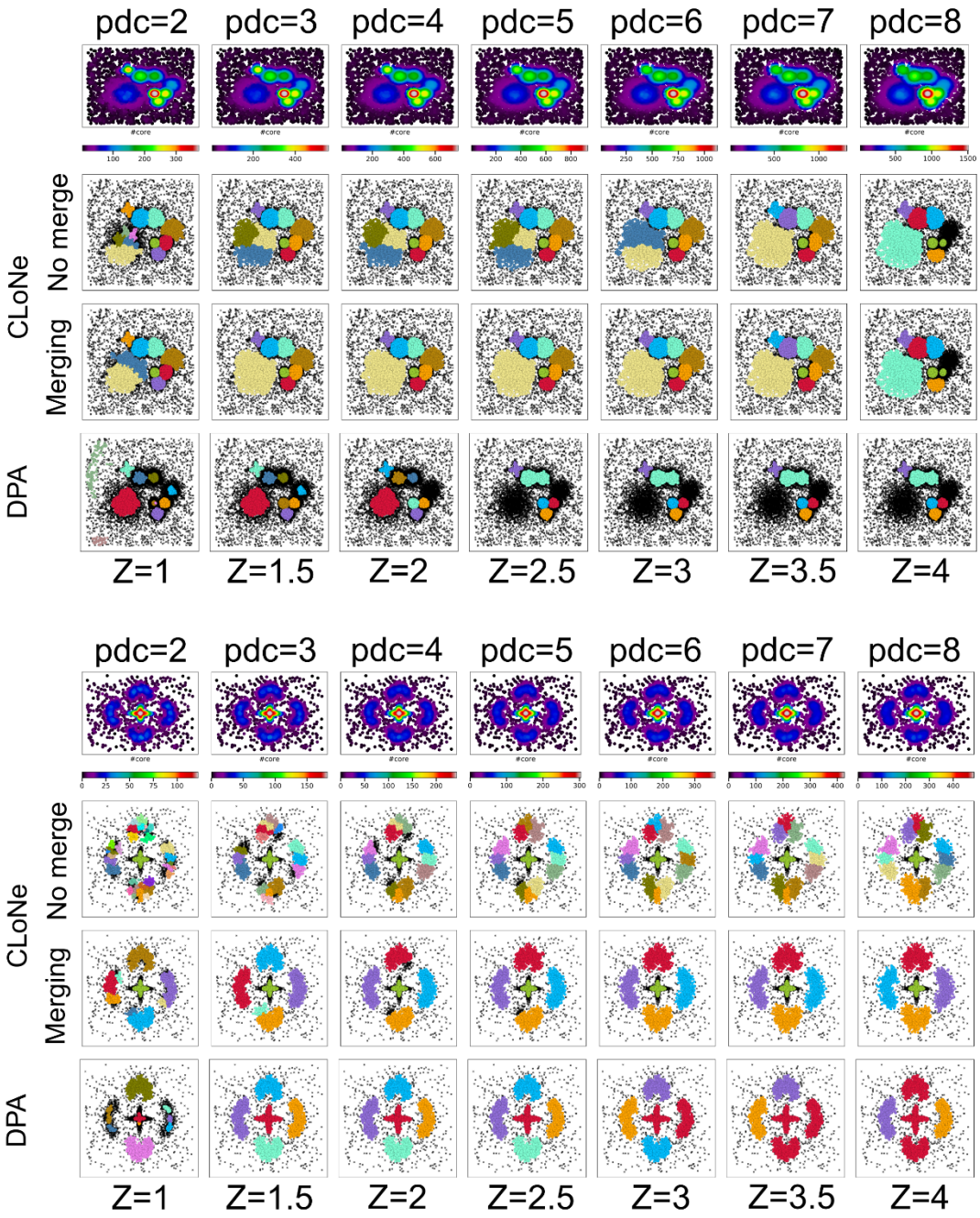
Contact: matteo.dalperaro@epfl.ch



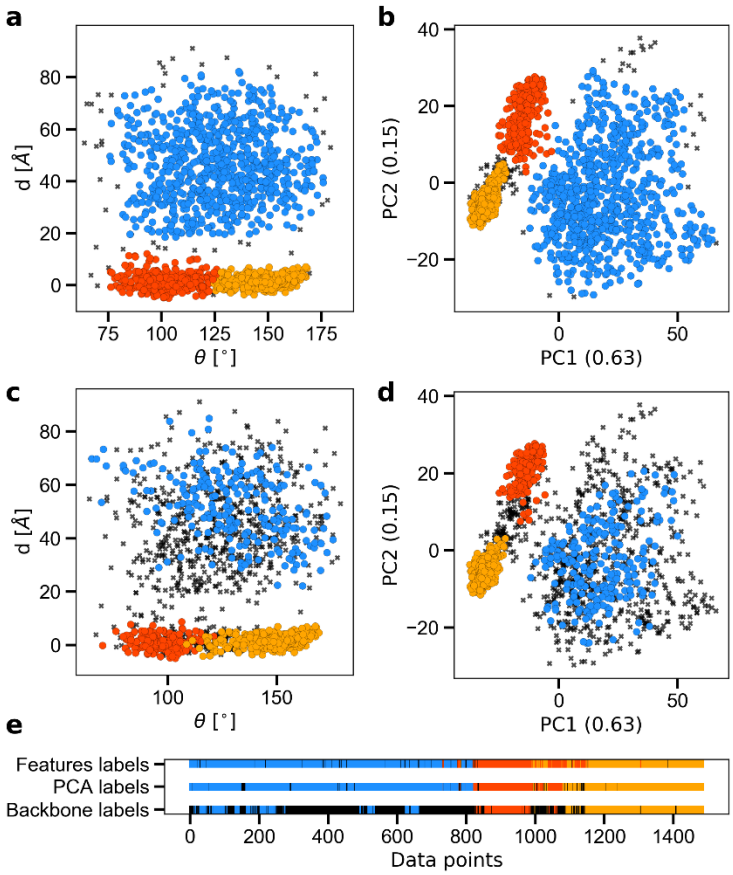
Supplementary Figure 1 | Basic usage, sample input and output of CLoNe. For more details, please see our lab's webpage or the GitHub repository (lbm.epfl.ch/resources and github.com/LBM-EPFL/CLoNe). This figure serves as a general overview for the use of CLoNe and what output files you will obtain. On the upper left, the commands to run CLoNe on structural data are shown as well as a short explanation of the different inputs. The compatible formats for MD trajectory and topology files are those compatible with the MDTraj (McGibbon et al., 2015) library, as it is used to load coordinates and to extract cluster centers and related trajectories. Similarly, the syntax used for atom selection is the one of MDTraj. The right-hand side of the figure shows some of the output generated by CLoNe. In addition to the plots similar to those shown in the main text, there is a separate file summarising the results and detailing statistical information on the clusters. The frames identified as cluster centers are shown below said summary. Not shown in this figure are the Tcl loading scripts to load either the cluster centers or the cluster trajectories into VMD (Humphrey et al., 1996) for further visual inspection. On the lower left is shown the minimal code to run CLoNe as a general clustering algorithm as well as a separate script to run either a selection of benchmark datasets or one provided by the user along with a value for CLoNe's input parameter. A sample input file is shown for the data to cluster. The same file format has to be used for structural features, as header names are used for some output files.



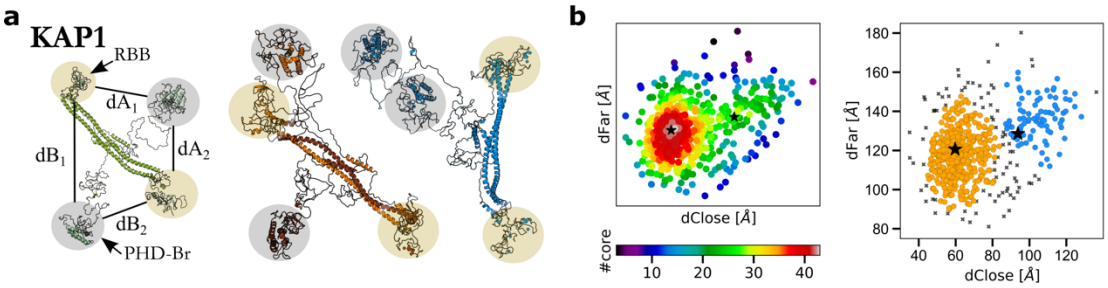
Supplementary Figure 2 | Comparing CLoNe to other clustering algorithms on selected datasets. The base code and parameters for each clustering algorithm is based on scikit-learn (Pedregosa et al., 2011): https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html, except for the last four rows and first three columns. Every column corresponds to a different clustering scheme. In order of left to right, there are CLoNe, Density Peaks, Density Peaks Advanced, minibatch K-means, affinity propagation, mean shift, spectral clustering, Ward hierarchical clustering, average linkage agglomerative clustering, DBSCAN, OPTICS, Birch and Gaussian mixture model. Each row corresponds to a different benchmark case. From top to bottom, there are noisy circles, noisy moons, varied, anisotropic blobs, blobs, no structure, flame, spiral, aggregation and s4. The last four rows correspond to other relevant benchmark from (Chang and Yeung, 2008; Fränti and Sieranoja, 2018; Fu and Medico, 2007; Gionis et al., 2007), with parameters leading to the best results that we could obtain by scanning values. The runtime of each case is shown on the lower left of each plot in seconds. Of note, results for DP were obtained from a Python implementation of the original Matlab code and runtimes may differ from the latter. Similarly, DPA results come from the original code in Fortran.



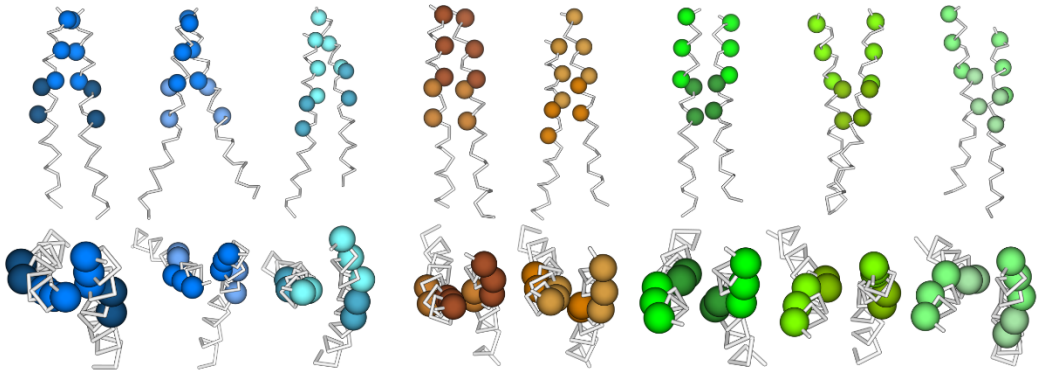
Supplementary Figure 3 | Large range of acceptable values for CLoNe's unique input parameter. The upper dataset is one from *Density Peaks Advanced* (d'Errico et al., 2018), while the lower one is the same as the one in Figure 1 of the main text. Clusters are shown in colors and outliers as smaller black crosses. In each case, the first row shows core cardinalities mapped on every data point as computed by CLoNe. Each column shows the clustering results for a different input value for CLoNe's and DPA's input parameter. In the case of DPA, the range of value matches that of the article. For CLoNe, the upper rows in each case showcase the clustering results without the merging step outlined in the Methods section and the one directly below it exhibit clusters after the merging step.



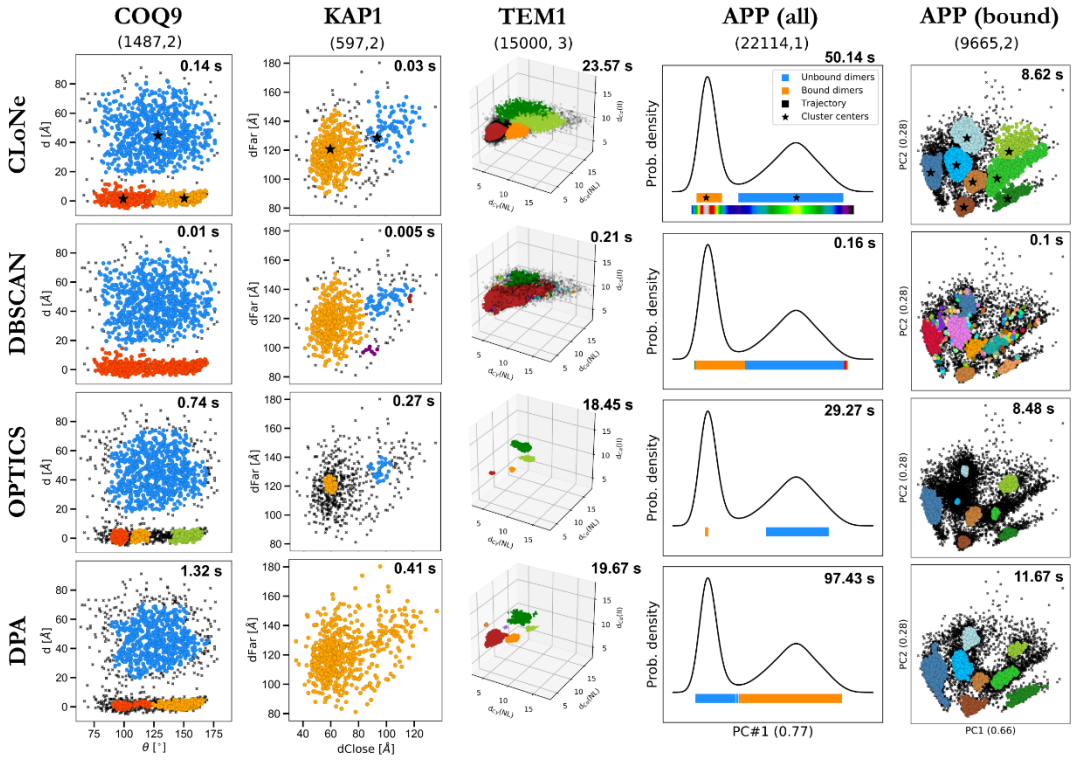
Supplementary Figure 4 | Hypothesis-free and high-dimensional clustering of COQ9. The color code follows the main text, with outliers in black. The feature-driven clustering is shown again in panel **(a)** for reference. We show the same process in panel **(b)** in the case where no prior knowledge would be available (hypothesis-free). To this end, we extracted the X, Y and Z coordinates of each backbone atoms, yielding 687 features in total, and applied PCA to extract the first two principal components (covering 65% and 15% of variability, respectively). Similar clusters to the feature-driven case are detected. To test the ability of CLoNe to handle high dimensional spaces, we applied it directly on the 687 coordinates previously extracted. As this cannot be plotted in a visually aesthetic and comprehensive manner, we color-coded the feature space and principal space in **(c)** and **(d)** according to the obtained labels. A direct comparison of the labels obtained in the three cases is shown in **(e)**.



Supplementary Figure 5 | Inter-domain organization of KAP1. A recent integrative modeling study from our group involved the modelling of a macromolecular dimer known as KRAB-domain associated protein 1 (KAP1) or Tripartite Motif-containing protein 28 (TRIM-28). KAP1 plays a fundamental role in the regulation of gene expression by recruiting several transcription factors and altering chromatin organization (Cheng et al., 2014; Iyengar and Farnham, 2011). Due to its high flexibility and intrinsically disordered nature, its internal architecture remained mostly elusive until recently (Fonti et al., 2019). In this study, around six hundred models were fitted to small-angle X-ray scattering (SAXS) data by minimizing the χ^2 (Grudinin et al., 2017; Hoffmann and Grudinin, 2017). Under the hypothesis that its two PHD-Br domains **(a,** left**)** were key to the shape predicted by SAXS, the distances between either PHD-Br domain and its closest RING/B-box1/B-box2 (RBB) domain was tracked in the ensemble and fed to CLoNe. Two clusters highlighted the internal module organization of KAP1 **(b)**. The first consists of one PHD-Br module close to one RBB module and the other being located further away (a, middle) while the second cluster exhibits conformations with both PHD-Br modules located further away, which may indicate transitional states (**a,** right**)** (Fonti et al., 2019).



Supplementary Figure 6. Cluster centroids extracted from dimerized APP monomers. Each cluster center is shown along with their top view to illustrate the corresponding dimerization motif. The colour code follows that of Figure 5 of the main text, where each dimerization motif is shown as different shades of the same colour. The sidechains were omitted for better visualization.



Supplementary Figure 7. Comparison between CLoNe and other algorithms on all structural datasets. We selected the best performing algorithms highlighted in the introduction and compared their performance on our structural datasets to CLoNe. For each algorithm, we scanned candidate parameter values until we obtained clusters resembling the results of CLoNe were achieved. Outliers are shown in black, and clusters are colored in the same manner as in the main text whenever possible. In the case of TEM1, due to the 3D projection, outliers have been hidden so that the clusters remain visible for OPTICS and DPA. The runtime in each case is shown on the upper right of each panel. DBSCAN handles COQ9, KAP1 and APP well enough, though due to different density levels it misclassifies a few points in the case of KAP1 and is unable to detect more than one membrane-bound state of COQ9. A similar observation can be drawn for TEM1 and APP (bound), but in these cases the clustering results are not satisfactory. OPTICS, as advertised, is able to detect different density threshold and accurately clusters all cases (save an additional cluster in the case of COQ9), but tend to classify most points as outliers in these datasets (25% for COQ9, 84% for KAP1, 95% for TEM1, 51% for APP (all) and 60% APP (bound)). While DPA obtains comparable results to CLoNe for COQ9 and both APP datasets, it is unable to detect a second cluster in KAP1's dataset and identifies two superfluous clusters for TEM1 (6 and 14 data points) in addition to classifying 50% of the frames as outliers. Of note, DPA was unable to process datapoints with identical values in the case of TEM1 and APP (bound), so the corresponding datasets were re-processed to contain only one data point with the same value.

Supplementary Table 1. Cluster data for high dimensional datasets, each containing 16 clusters of 64 elements. In all cases, the 16 clusters are properly identified.

Cluster#	64			128			256			512			1024		
	ρ	#core	#el	ρ	#core	#el	ρ	#core	#el	ρ	#core	#el	ρ	#core	#el
1	26.05	27	64	22.84	32	64	24.08	32	64	25.46	28	64	24.29	32	64
2	26.23	31	64	24.02	31	64	25.91	31	64	25.21	32	64	25.65	30	64
3	26.63	31	64	26.72	32	64	25.84	33	64	24.56	34	64	25.85	31	64
4	28.00	32	64	24.58	33	64	24.29	33	64	26.05	34	64	25.63	31	64
5	26.91	32	64	25.23	33	64	25.11	34	64	26.22	35	64	26.77	30	64
6	27.65	33	64	27.51	35	64	25.90	35	64	25.77	34	64	24.75	33	64
7	25.67	35	64	27.15	36	64	25.63	35	64	27.22	33	64	26.19	32	64
8	27.61	35	64	25.45	37	64	24.91	36	64	25.95	34	64	25.97	33	64
9	28.82	35	64	25.74	36	64	26.30	35	64	27.13	34	64	26.65	32	64
10	28.22	35	64	26.31	37	64	28.09	36	64	27.49	34	64	25.93	34	64
11	26.40	39	64	28.61	36	64	25.27	36	64	26.24	34	64	26.50	33	64
12	29.07	37	64	27.44	38	64	27.15	35	64	26.20	35	64	27.20	33	64
13	30.51	38	64	30.76	36	64	28.06	34	64	27.32	35	64	25.60	34	64
14	30.46	37	64	27.61	39	64	26.53	36	64	26.54	35	64	26.22	36	64
15	27.07	39	64	28.85	38	64	26.86	36	64	27.71	36	64	27.58	35	64
16	28.70	40	64	32.84	40	64	28.85	37	64	27.55	37	64	30.43	35	64

Supplementary Table 2. Statistics for clusters and centers from structural ensembles and related features.

COQ9	*ρ*	**#core**	**elements**	**- outliers**	***θ***	***d***
Diffusing states	69.10	69	798	767	128.54	44.63
Membrane-bound #1	66.06	86	291	281	99.64	1.43
Membrane-bound #2	78.80	98	398	392	150.36	1.75

KAP1	*ρ*	**#core**	**elements**	**- outliers**	**dClose**	**dFar**
Far – Far	18.89	27	93	90	93.71	128.57
Far – Close	36.81	42	504	420	59.66	120.62

TEM1	*ρ*	**#core**	**elements**	**- outliers**	**$d_{cα}$(I263, I279)**	**$d_{cα}$(N276, L220)**	**$d_{cγ}$(N276, L220)**
Deep pocket	241.23	253	1039	996	12.75	9.91	8.90
Open pocket	395.30	449	2540	2208	9.89	11.16	8.99
Semi-open pocket	491.11	613	1823	1612	8.39	8.82	7.13
Closed pocket	758.34	811	9598	6839	8.83	7.01	4.46

APP (all)	*ρ*	**#core**	**elements**	**- outliers**	**PC#1 (0.77)**
Unbound	1435.77	1635	11746	11600	4.69
Bound	2144.34	2456	10368	9665	-5.28

APP (bound)	*ρ*	**#core**	**elements**	**- outliers**	**d(G700)**	**d(G704)**	**d(G708)**	**d(G709)**	**d(A713)**
GA #1 (light green)	126.63	162	363	352	13.60	12.21	11.20	9.02	8.97
GA #2 (green)	236.01	274	997	967	11.97	10.51	8.65	7.31	7.60
GA #3 (dark green)	190.24	232	430	404	11.77	10.46	9.87	5.10	5.14
Hybrid #1 (brown)	330.83	412	1019	875	9.36	7.90	7.85	7.74	8.74
Hybrid #2 (dark brown)	346.34	418	978	880	7.39	6.70	6.29	6.48	6.79
GGG #1 (light blue)	292.95	347	1136	974	9.94	8.69	9.55	11.82	12.45
GGG #2 (blue)	381.70	428	2574	2219	7.90	7.36	7.68	9.75	11.07
GGG #3 (dark blue)	363.07	428	2168	1745	5.51	5.00	5.31	10.32	11.88

Supplementary references

Chang, H., Yeung, D.-Y., 2008. Robust Path-Based Spectral Clustering. Pattern Recogn 41, 191–203. https://doi.org/10.1016/j.patcog.2007.04.010

Cheng, C.-T., Kuo, C.-Y., Ann, D.K., 2014. KAPtain in charge of multiple missions: Emerging roles of KAP1. World J. Biol. Chem. 5, 308–320. https://doi.org/10.4331/wjbc.v5.i3.308

d'Errico, M., Facco, E., Laio, A., Rodriguez, A., 2018. Automatic topography of high-dimensional data sets by non-parametric Density Peak clustering.

Fonti, G., Marcaida, M.J., Bryan, L.C., Träger, S., Kalantzi, A.S., Helleboid, P.-Y.J., Demurtas, D., Tully, M.D., Grudinin, S., Trono, D., Fierz, B., Peraro, M.D., 2019. KAP1 is an antiparallel dimer with a functional asymmetry. Life Sci. Alliance 2, e201900349. https://doi.org/10.26508/lsa.201900349

Fränti, P., Sieranoja, S., 2018. K-means properties on six clustering benchmark datasets.

Fu, L., Medico, E., 2007. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics 8, 3. https://doi.org/10.1186/1471-2105-8-3

Gionis, A., Mannila, H., Tsaparas, P., 2007. Clustering aggregation. ACM Trans. Knowl. Discov. Data 1, 4–es. https://doi.org/10.1145/1217299.1217303

Grudinin, S., Garkavenko, M., Kazennov, A., 2017. Pepsi-SAXS: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. Acta Crystallogr. Sect. Struct. Biol. 73, 449–464. https://doi.org/10.1107/S2059798317005745

Hoffmann, A., Grudinin, S., 2017. NOLB: Nonlinear Rigid Block Normal-Mode Analysis Method. J. Chem. Theory Comput. 13, 2123–2134. https://doi.org/10.1021/acs.jctc.7b00197

Humphrey, W., Dalke, A., Schulten, K., 1996. VMD: Visual molecular dynamics. J. Mol. Graph. 14, 33–38. https://doi.org/10.1016/0263-7855(96)00018-5

Iyengar, S., Farnham, P.J., 2011. KAP1 Protein: An Enigmatic Master Regulator of the Genome. J. Biol. Chem. 286, 26267–26276. https://doi.org/10.1074/jbc.R111.252569

McGibbon, R.T., Beauchamp, K.A., Harrigan, M.P., Klein, C., Swails, J.M., Hernández, C.X., Schwantes, C.R., Wang, L.-P., Lane, T.J., Pande, V.S., 2015. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys. J. 109, 1528 – 1532. https://doi.org/10.1016/j.bpj.2015.08.015

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Rodriguez, A., Laio, A., 2014. Clustering by fast search and find of density peaks. Science 344, 1492–1496. https://doi.org/10.1126/science.1242072