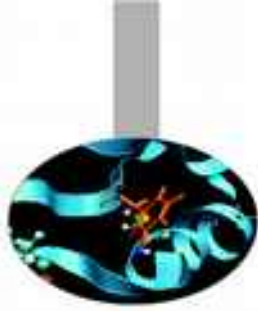
CINECA SCAI
SuperComputing Applications and Innovation



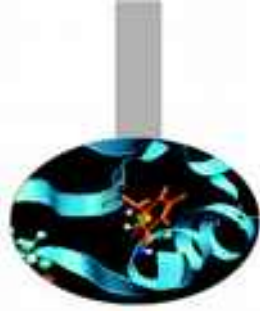
# Introduction to Molecular Simulations

Alessandro Grottesi

PRACE

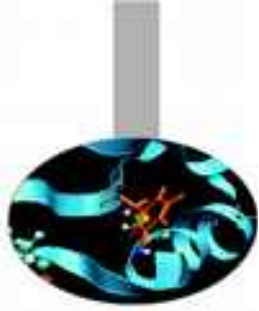
CINECA

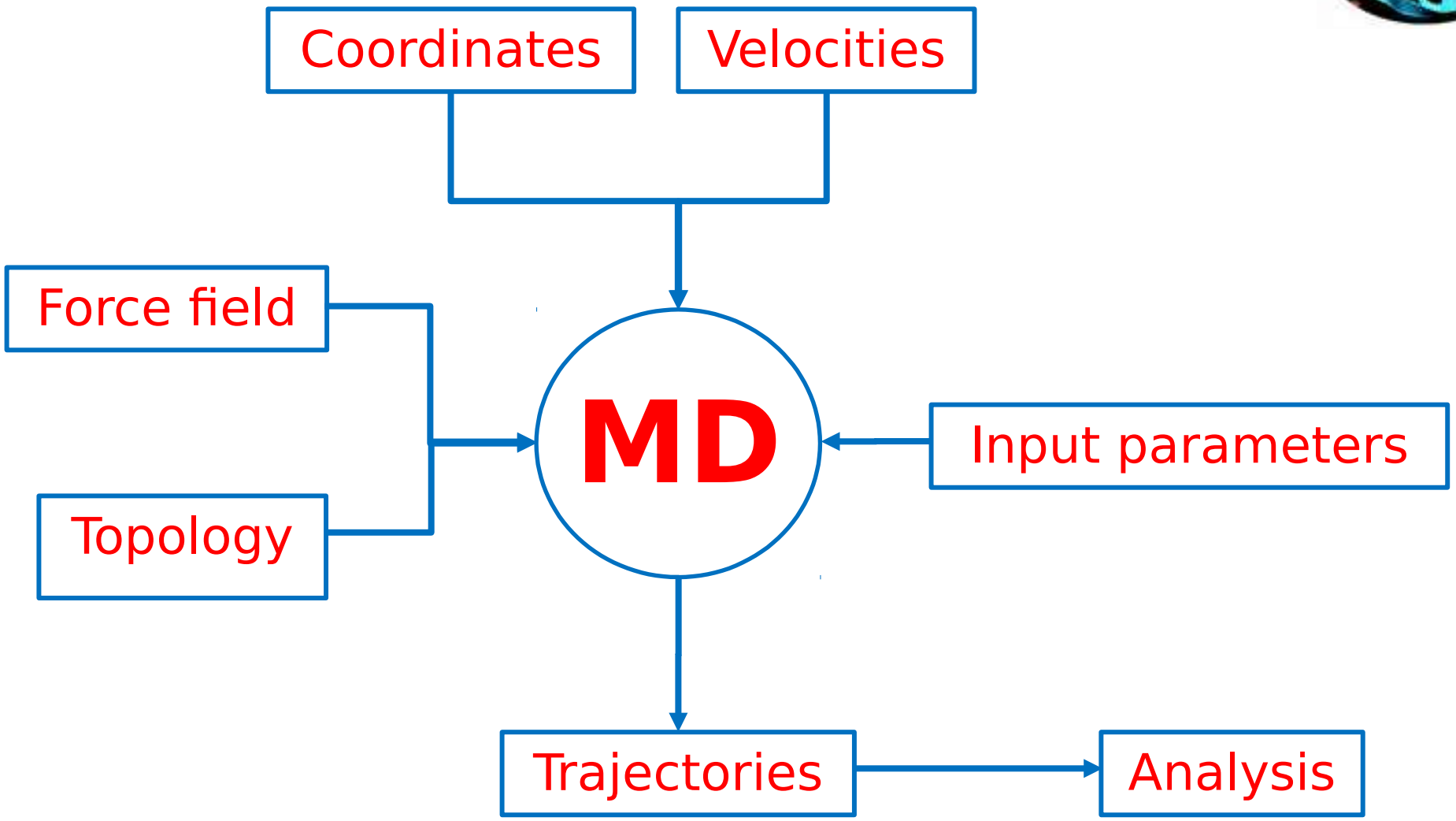
# Outlook

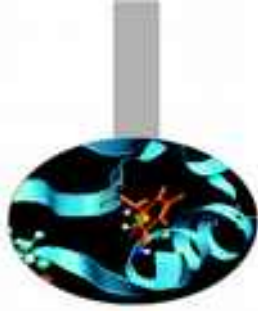


1. Classic Molecular Dynamics
2. Focus on Biomolecules: data insights
3. Setting up a simulation: details and outputs
4. Analysis of data

# MD ingredients

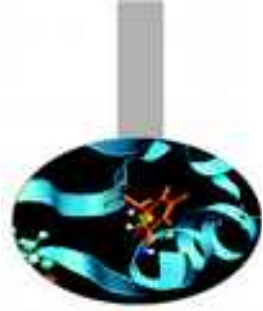






# Coordinates

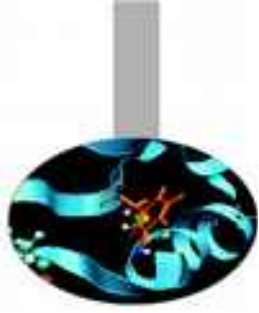
# Data formats for Molecular simulations



Mostly used in classical MD:

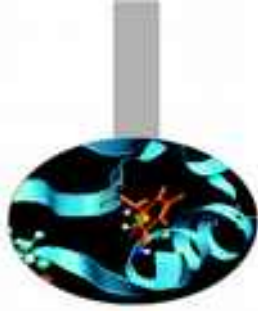
- PDB format
- GROMOS format
- XPLOR
- XYZ
- DCD
- CRD

# PDB data



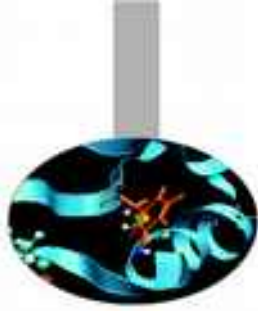
```
TITLE     GROningen Mixture of Alchemy and Childrens'' Stories
REMARK    THIS IS A SIMULATION BOX
CRYST1  175.540  175.540  180.000  90.00  90.00  90.00 P 1           1
MODEL        1
ATOM      1  N   LEU     1      52.630  70.190  49.490  1.00  0.00
ATOM      2  H1  LEU     1      53.370  70.540  50.100  1.00  0.00
ATOM      3  H2  LEU     1      52.130  69.510  50.050  1.00  0.00
ATOM      4  CA  LEU     1      51.710  71.290  49.110  1.00  0.00
ATOM      5  CB  LEU     1      51.140  71.930  50.380  1.00  0.00
ATOM      6  CG  LEU     1      50.170  71.010  51.130  1.00  0.00
ATOM      7  CD1 LEU     1      49.950  71.550  52.550  1.00  0.00
ATOM      8  CD2 LEU     1      48.820  70.920  50.410  1.00  0.00
ATOM      9  C   LEU     1      52.250  72.410  48.180  1.00  0.00
ATOM     10  O   LEU     1      51.500  72.970  47.380  1.00  0.00
ATOM     11  N   GLN     2      53.540  72.710  48.310  1.00  0.00
ATOM     12  H   GLN     2      54.110  72.210  48.960  1.00  0.00
ATOM     13  CA  GLN     2      54.220  73.760  47.510  1.00  0.00
ATOM     14  CB  GLN     2      54.880  74.800  48.420  1.00  0.00
ATOM     15  CG  GLN     2      53.910  75.550  49.350  1.00  0.00
ATOM     16  CD  GLN     2      52.960  76.480  48.590  1.00  0.00
ATOM     17  OE1 GLN     2      53.340  77.220  47.690  1.00  0.00
ATOM     18  NE2 GLN     2      51.720  76.500  49.020  1.00  0.00
ATOM     19 1HE2 GLN     2      51.430  75.950  49.830  1.00  0.00
ATOM     20 2HE2 GLN     2      50.990  76.900  48.440  1.00  0.00
ATOM     21  C   GLN     2      55.280  73.140  46.590  1.00  0.00
ATOM     22  O   GLN     2      56.280  72.600  47.060  1.00  0.00
```

# GROMOS data



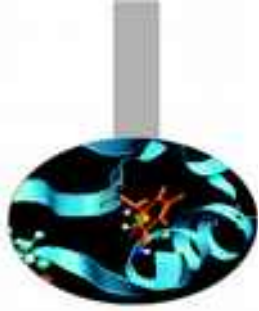
```
GROningen Mixture of Alchemy and Childrens'' Stories
31934
    1LEU      N    1   5.263   7.019   4.949
    1LEU     H1    2   5.337   7.054   5.010
    1LEU     H2    3   5.213   6.951   5.005
    1LEU     CA    4   5.171   7.129   4.911
    1LEU     CB    5   5.114   7.193   5.038
    1LEU     CG    6   5.017   7.101   5.113
    1LEU    CD1    7   4.995   7.155   5.255
    1LEU    CD2    8   4.882   7.092   5.041
    1LEU      C    9   5.225   7.241   4.818
    1LEU      O   10   5.150   7.297   4.738
    2GLN      N   11   5.354   7.271   4.831
    2GLN      H   12   5.411   7.221   4.896
    2GLN     CA   13   5.422   7.376   4.751
    2GLN     CB   14   5.488   7.480   4.842
    2GLN     CG   15   5.391   7.555   4.935
    2GLN     CD   16   5.296   7.648   4.859
    2GLN    OE1   17   5.334   7.722   4.769
    2GLN    NE2   18   5.172   7.650   4.902
    2GLN   HE21   19   5.143   7.595   4.983
    2GLN   HE22   20   5.099   7.690   4.844
    2GLN      C   21   5.528   7.314   4.659
    2GLN      O   22   5.628   7.260   4.706
```

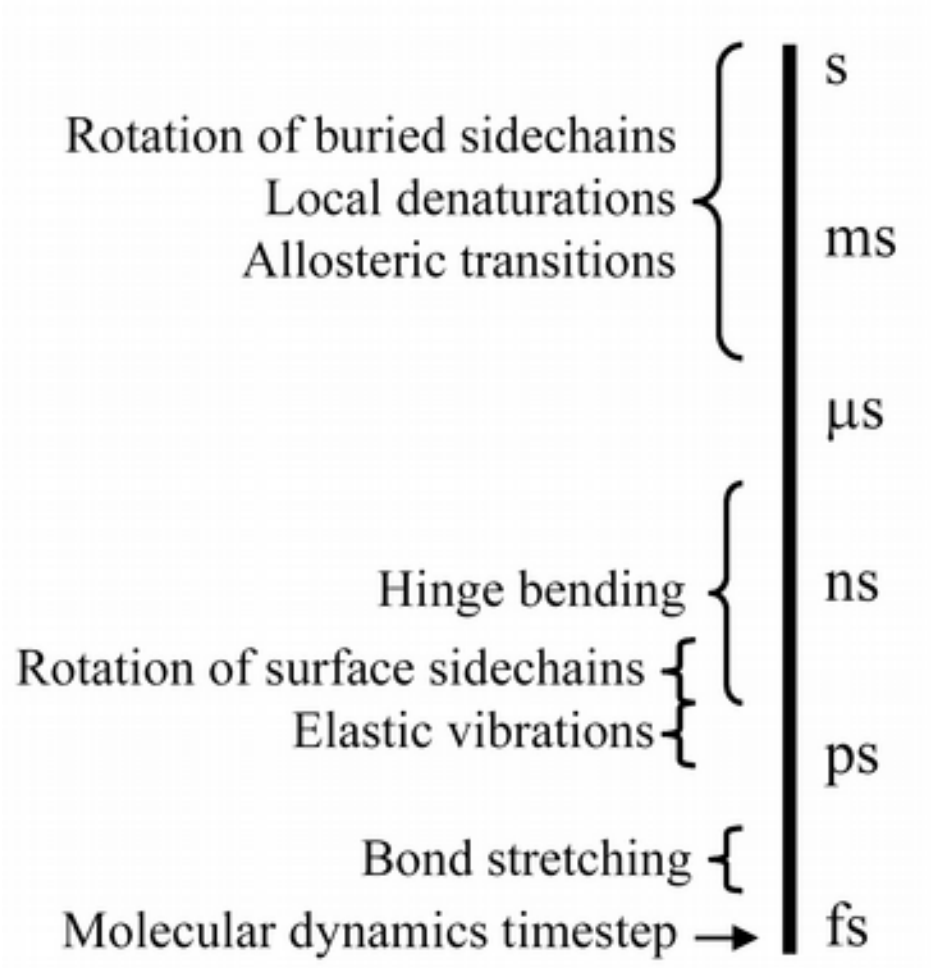
# XYZ format



```
31934
 generated by VMD
  N          52.630001       70.190002       49.490002
  H1         53.369999       70.540001       50.099998
  H2         52.130001       69.510002       50.049999
  CA         51.709999       71.290001       49.110001
  CB         51.139999       71.930000       50.380001
  CG         50.169998       71.010002       51.130001
  CD1        49.950001       71.550003       52.549999
  CD2        48.820000       70.919998       50.410000
  C          52.250000       72.410004       48.180000
  O          51.500000       72.970001       47.380001
  N          53.540001       72.709999       48.310001
  H          54.110001       72.209999       48.959999
  CA         54.220001       73.760002       47.509998
  CB         54.880001       74.800003       48.419998
  CG         53.910000       75.550003       49.349998
  CD         52.959999       76.480003       48.590000
  OE1        53.340000       77.220001       47.689999
  NE2        51.720001       76.500000       49.020000
 1HE2        51.430000       75.949997       49.830002
 2HE2        50.990002       76.900002       48.439999
  C          55.279999       73.139999       46.590000
  O          56.279999       72.599998       47.060001
  N          54.990002       73.169998       45.290001
  H          54.119999       73.540001       44.959999
  CA         55.869999       72.620003       44.230000
```

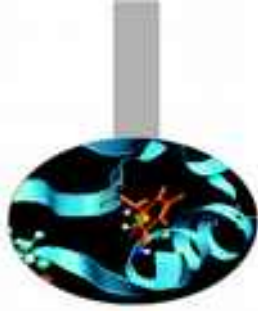
# Timescale






- Protein Folding - milliseconds/seconds ($10^{-3}$-1s)
- Ligand Binding - micro/milliseconds ($10^{-6}$-$10^{-3}$ s)
- Enzyme catalysis - micro/milliseconds ($10^{-6}$-$10^{-3}$ s)
- Conformational transitions - pico/nanoseconds ($10^{-12}$-$10^{-9}$ s)
- Collective vibrations -
- 1 picosecond ($10^{-12}$ s)
- Bond vibrations -
- 1 femtosecond ($10^{-15}$ s)

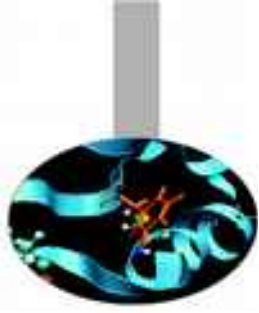
# Equation of motion



The equations that describe the temporal evolution of a physical system is called **equation of motion**. There are different equations of motions, which characterize the motion with different levels of approximation:

- Time-dependent Schrödinger's Equation
  - for quantum-mechanical system
- **Newton's Equation**
  - **for classical-mechanical system**
- Langevin's Equation
  - for stochastic system

# Force field



$$V(\mathbf{r}_1,\mathbf{r}_2,...,\mathbf{r}_n) = \sum_{bond} \frac{1}{2} k_{b_n} (b_n - b_{0\,n})^2 + \sum_{angle} \frac{1}{2} k_{\theta_n} (\theta_n - \theta_{0\,n})^2 +$$

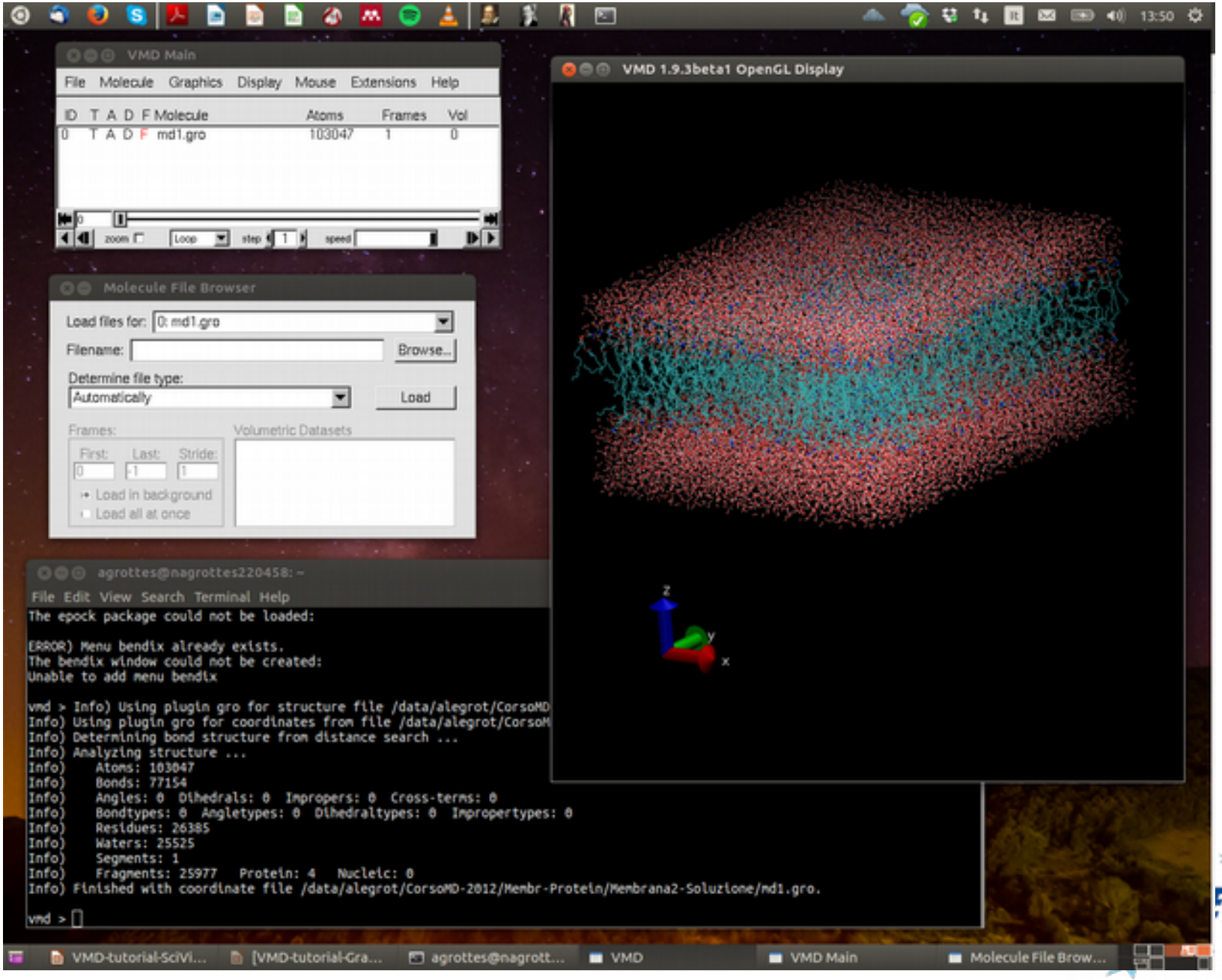
$$+ \sum_{\substack{improper \\ dihedral}} \frac{1}{2} k_{\xi_n} (\xi_n - \xi_{0\,n})^2 + \sum_{dihedral} k_{\phi_n} \left[1 + \cos(m_n \phi_n - \delta_n)\right] +$$

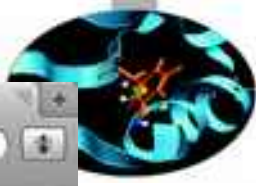
$$+ \sum_{\substack{nonbonded \\ pairs(ij)}} \left\{ \left[ \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}} \right] + \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{\varepsilon_r r_{ij}} \right\}$$

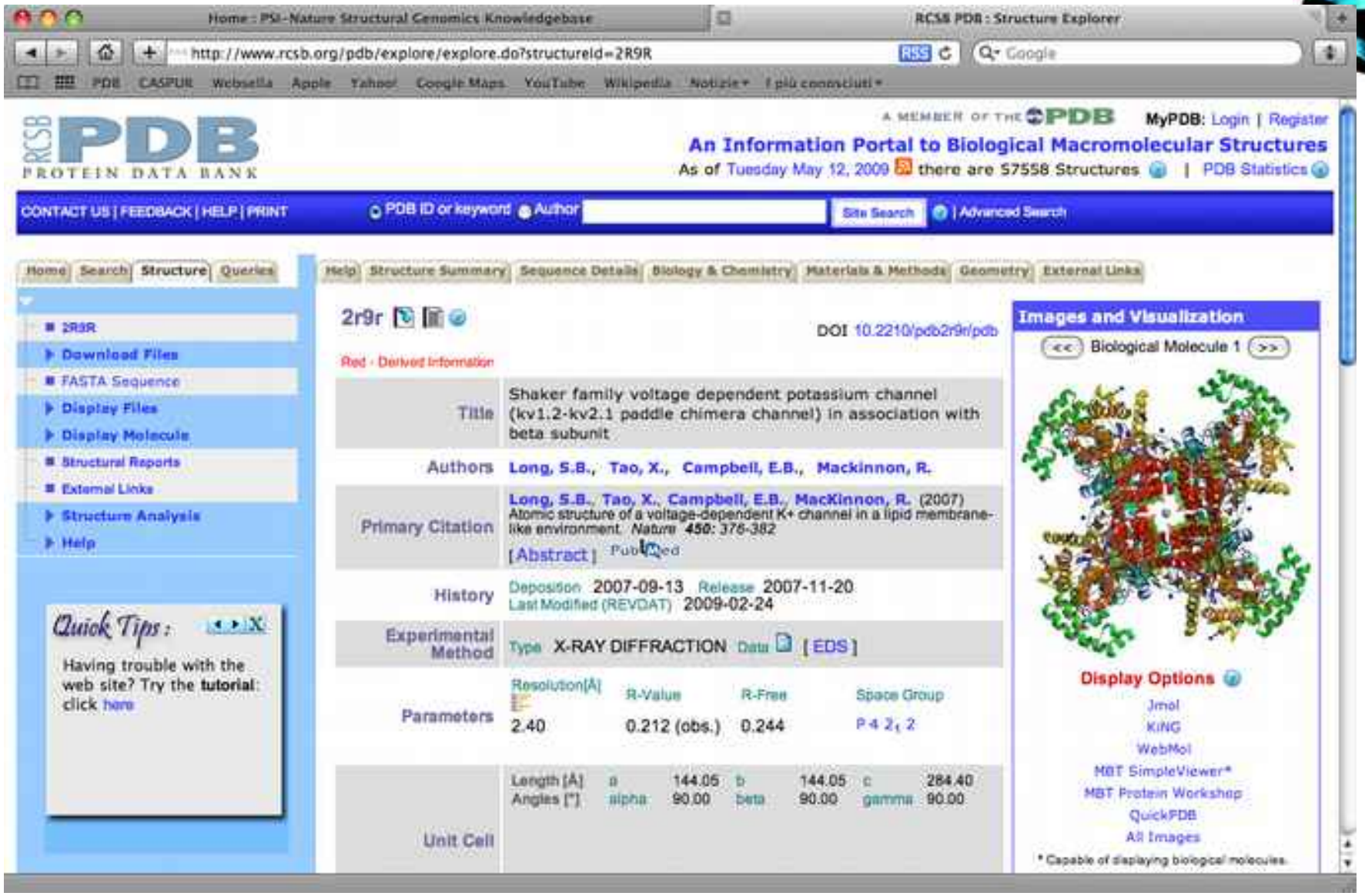
The potential energy function, together with the parameters required to describe the behavior of different kinds of atoms and bonds ($k_b$, $k_\theta$, $k_\xi$, $C_{ij}$, …), is called: **force field.**

Several force fields are currently used and the choice depends from the studied system. Some force field are better suited for nucleic acids, for example, while others for membrane proteins

# MD set up

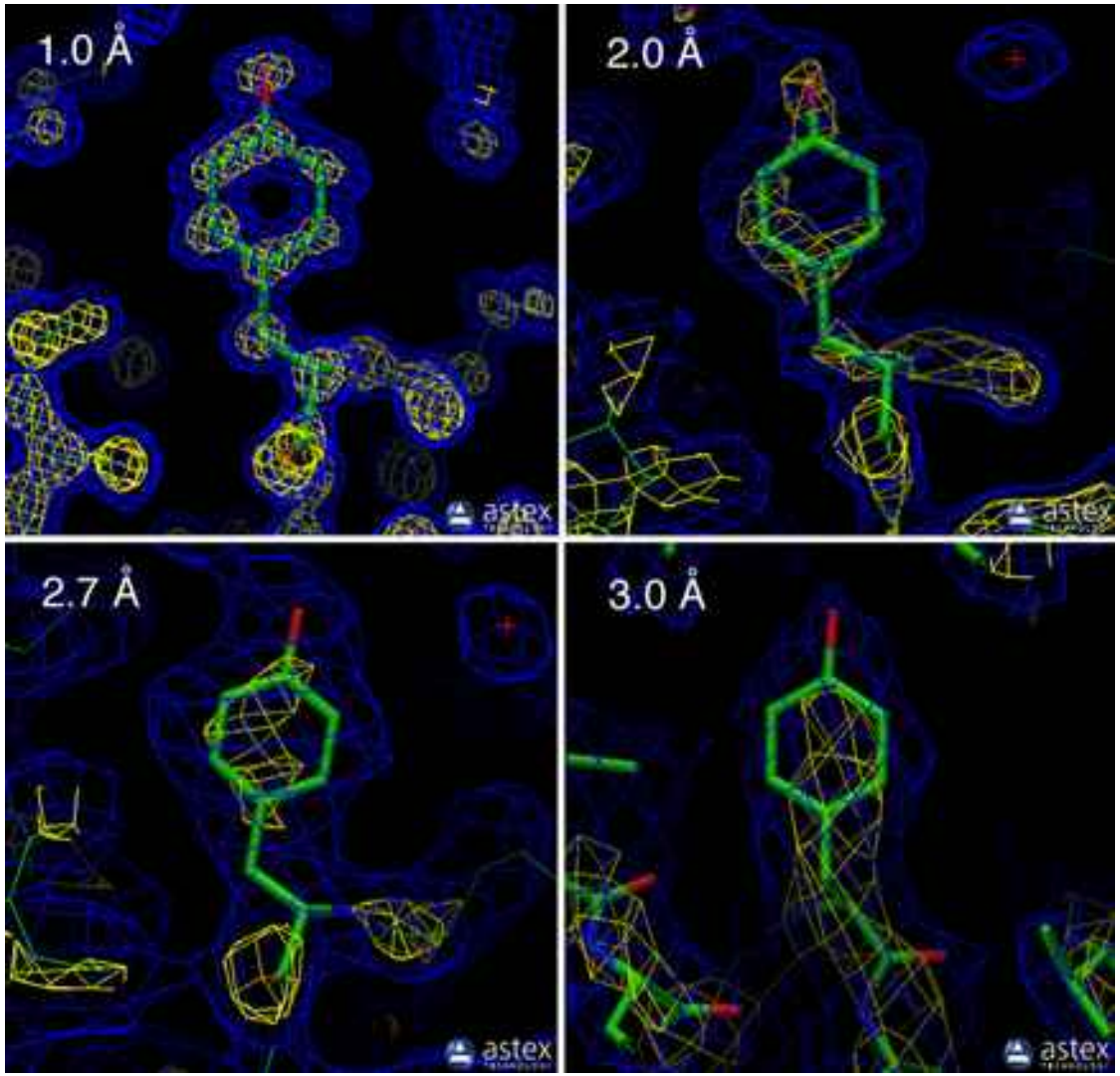

VMD Main
File Molecule Graphics Display Mouse Extensions Help
ID T A D F Molecule Atoms Frames Vol
0 T A D F md1.gro 103047 1 0
Molecule File Browser
Load files for: 0: md1.gro
Filename:
Browse...
Determine file type:
Automatically
Load
Frames:
First: Last: Stride:
Load in background
Load all at once
Volumetric Datasets
VMD 1.9.3beta1 OpenGL Display
agrottes@nagrottes220458: ~
File Edit View Search Terminal Help
The epock package could not be loaded:
ERROR) Menu bendix already exists.
The bendix window could not be created:
Unable to add menu bendix
vmd > Info) Using plugin gro for structure file /data/alegrot/CorsoMD
Info) Using plugin gro for coordinates from file /data/alegrot/CorsoM
Info) Determining bond structure from distance search ...
Info) Analyzing structure ...
Info) Atoms: 103047
Info) Bonds: 77154
Info) Angles: 0 Dihedrals: 0 Impropers: 0 Cross-terms: 0
Info) Bondtypes: 0 Angletypes: 0 Dihedraltypes: 0 Impropertypes: 0
Info) Residues: 26385
Info) Waters: 25525
Info) Segments: 1
Info) Fragments: 25977 Protein: 4 Nucleic: 0
Info) Finished with coordinate file /data/alegrot/CorsoMD-2012/Membr-Protein/Membrana2-Soluzione/md1.gro.
vmd >
VMD-tutorial-SciVi...
[VMD-tutorial-Gra...
agrottes@nagrott...
VMD
VMD Main
Molecule File Brow...






http://www.rcsb.org/pdb

# Initial coordinates: X-Ray vs. NMR

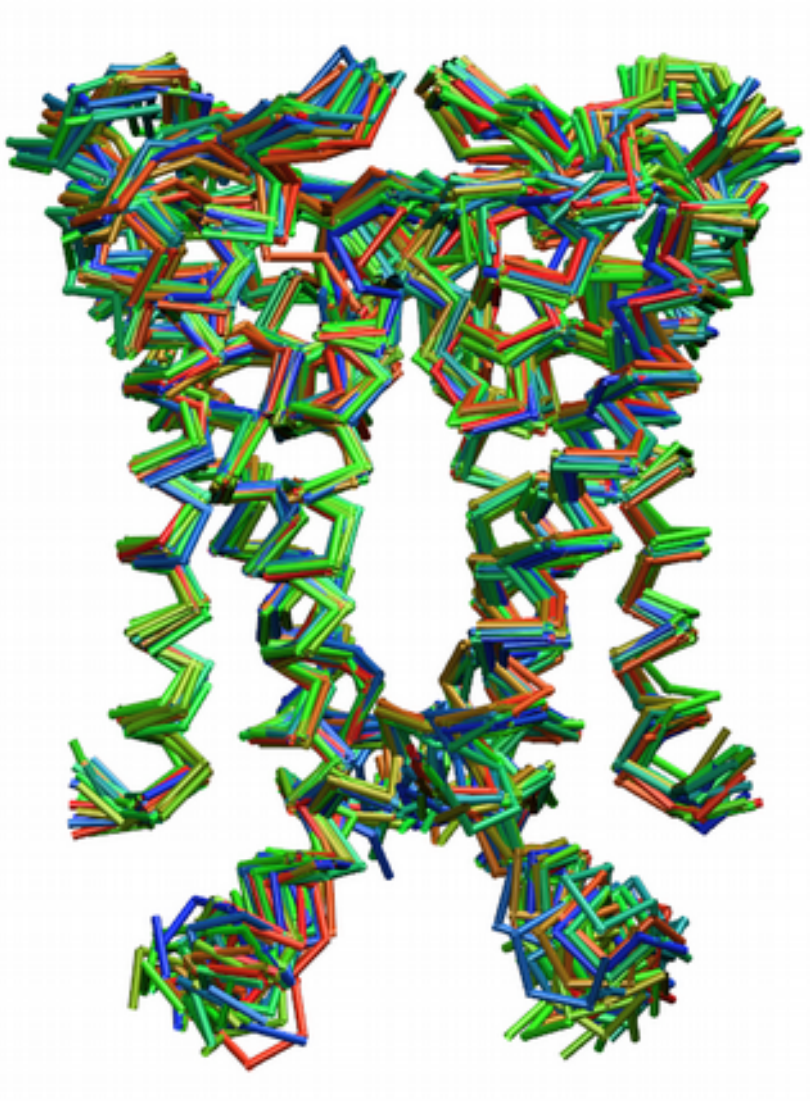



Higher X-ray resolution allows to use a more reliable starting structure in terms of amino-acids stereo-chemistry and accuracy of atomic positions

Error on initial position of protein atoms determines local structural alterations of the protein structure

X-ray resolutions smaller than 2 Å are much more reliable, although difficult to achieve. Generally, a resolution in the range 2 < R < 3 Å are acceptable. Beyond 3 Å the uncertainty of the initial position may cause artefacts in the MD simulation

# Initial coordinates: X-Ray vs. NMR

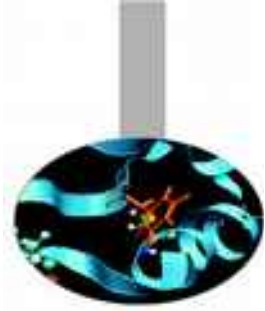


KcsA Potassium channel

(PDB code: 2K1E)

NMR determined structure provide information in a more realistic physiological environment as compared to X-ray determined structures although this could result in lower quality of initial coordinates and incertainties in the position of atomic coordinates.

# Workflow for running MD simulations in GROMACS



Generate a topology — pdb2gmx

Generate simulation box — editconf

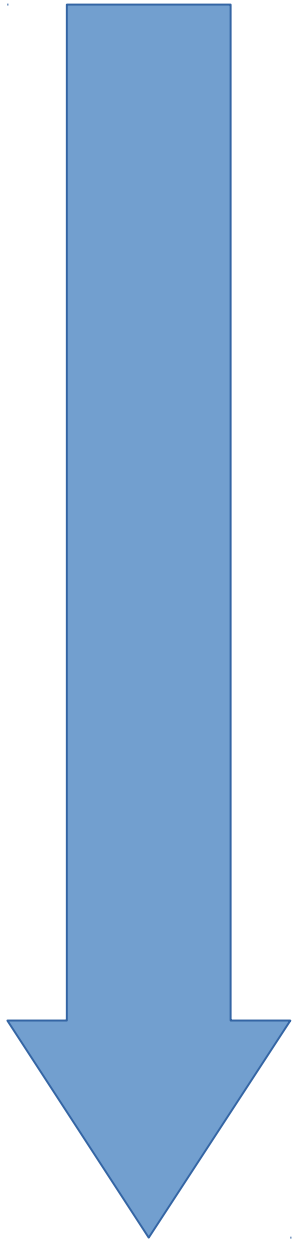
Solvate the system — genbox, genion

Generate input file for mdrun — grompp

topol.tpr

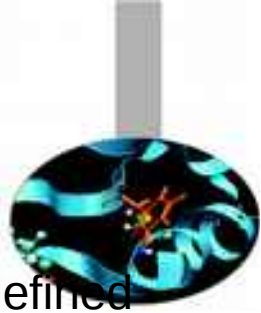
Actually run the simulation — mdrun

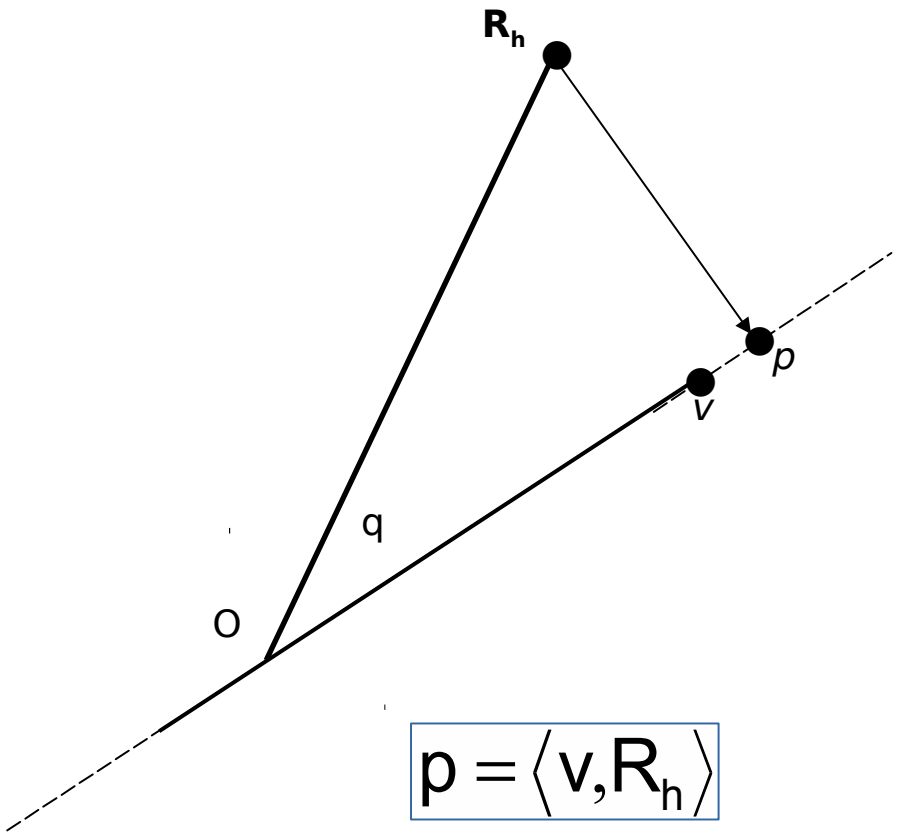
Analysis of trajectory files — trjconv, g_rms, g_covar, g_anaeig, g_energy, g_rmsf, g_rdf



CINECA

# PCA: how it works






$$p = \langle v, R_h \rangle$$

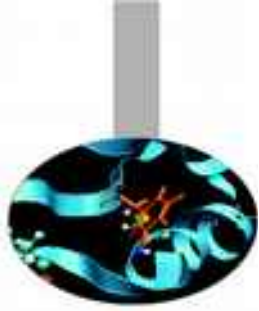
p is the projection of vector $\mathbf{R}_h$ onto unity vector **v** (dot product between **v** and $\mathbf{R}_h$)

Let's assume our simulation is defined by the vector Rn, that simply consists of the set of cartesian coordinates of Ca atoms at a given time step.

Question: what is the unity vector so that projection of $\mathbf{R}_h$ on vector **v** is the largest possible?

Answer: it is the vector **v**, so that the variance of the projected point p of Rn onto v is the largest possible

# Eigenvalue equation



The average of projected points onto **v** is:

$$\mu(v) = \langle v, \bar{x} \rangle$$

$$Cv = \lambda v$$

Eigenvectors

Eigenvalues

Variance of projected point onto **v** is:

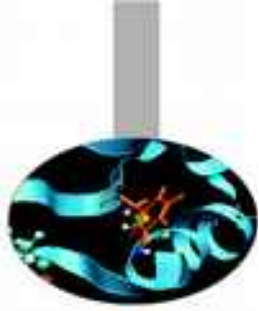
$$\sigma^2(v) = \langle Cv, v \rangle$$

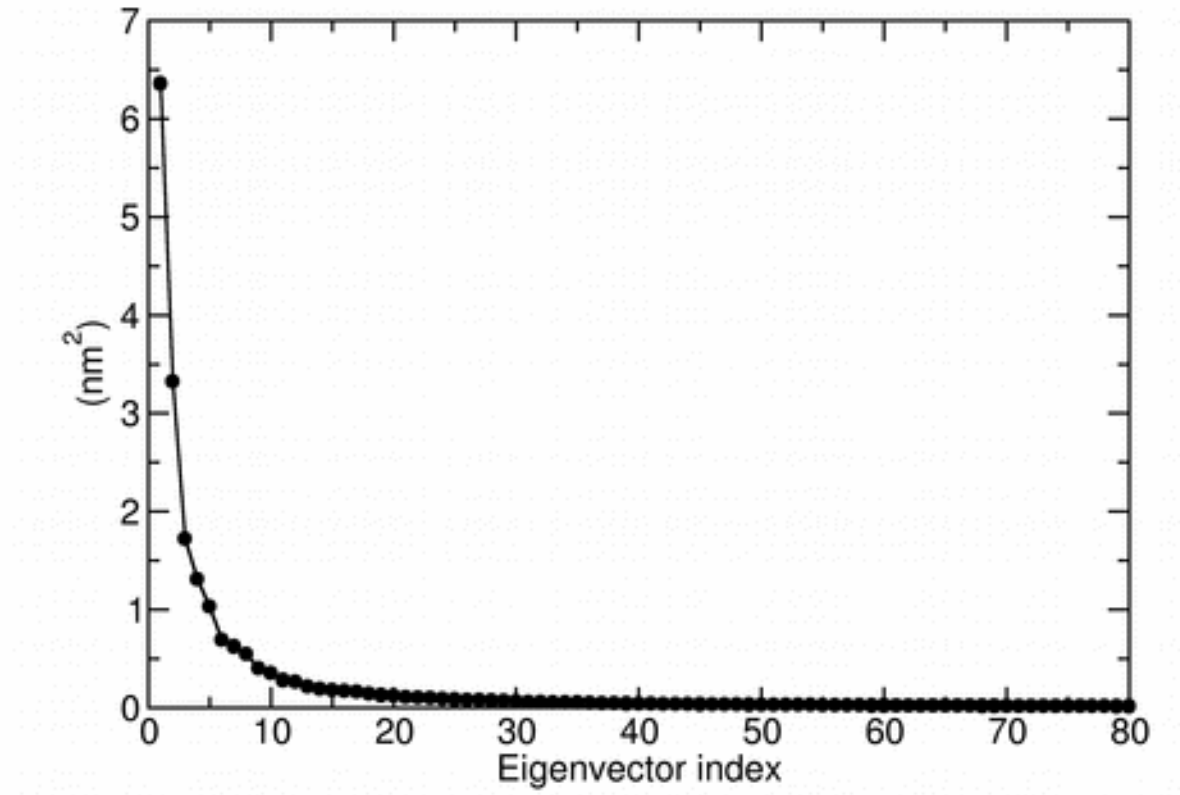
Variance of projected points along vector ***v*** can be expressed in terms of dot product between ***v*** and ***Cv***.

Eigenvectors represents direction where the $\sigma^2$ returns its maximum value.

Moreover, it can be shown that $\sigma^2$ are numerically equivalent to calculated eigenvalues

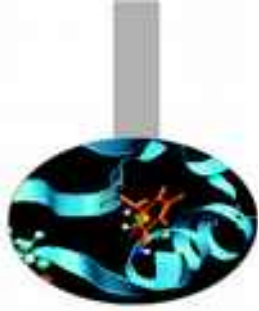
# Essential Dynamics of Proteins






Eigenvalues are sorted in descending order: the first one corresponds to the maximum variance of the projected points. The corresponding eigenvectors are the best principal components of associated eigenvalues.

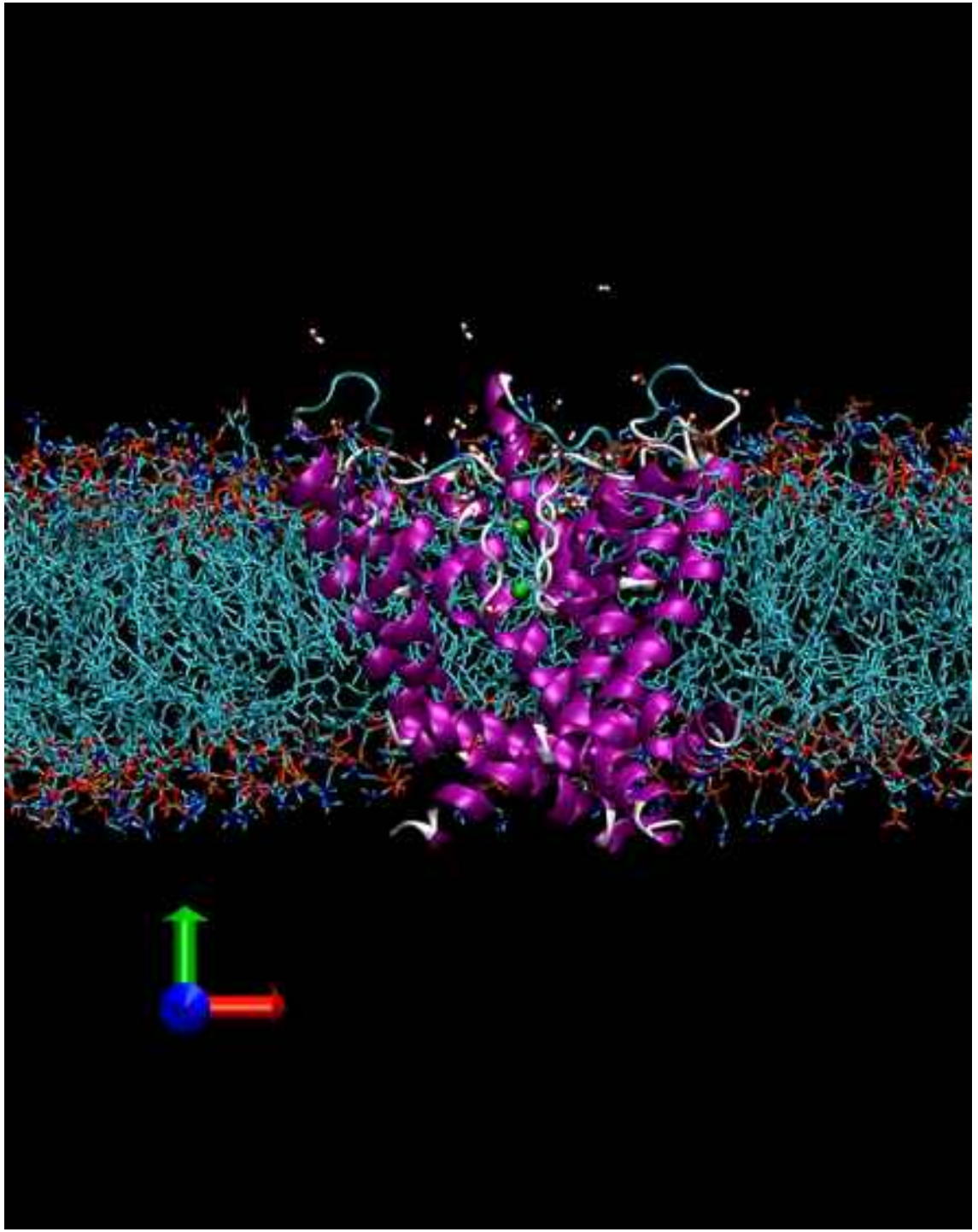
CINECA SCAI
SuperComputing Applications and Innovation



# Data visualizzation for Molecular Simulations

PRACE

CINECA



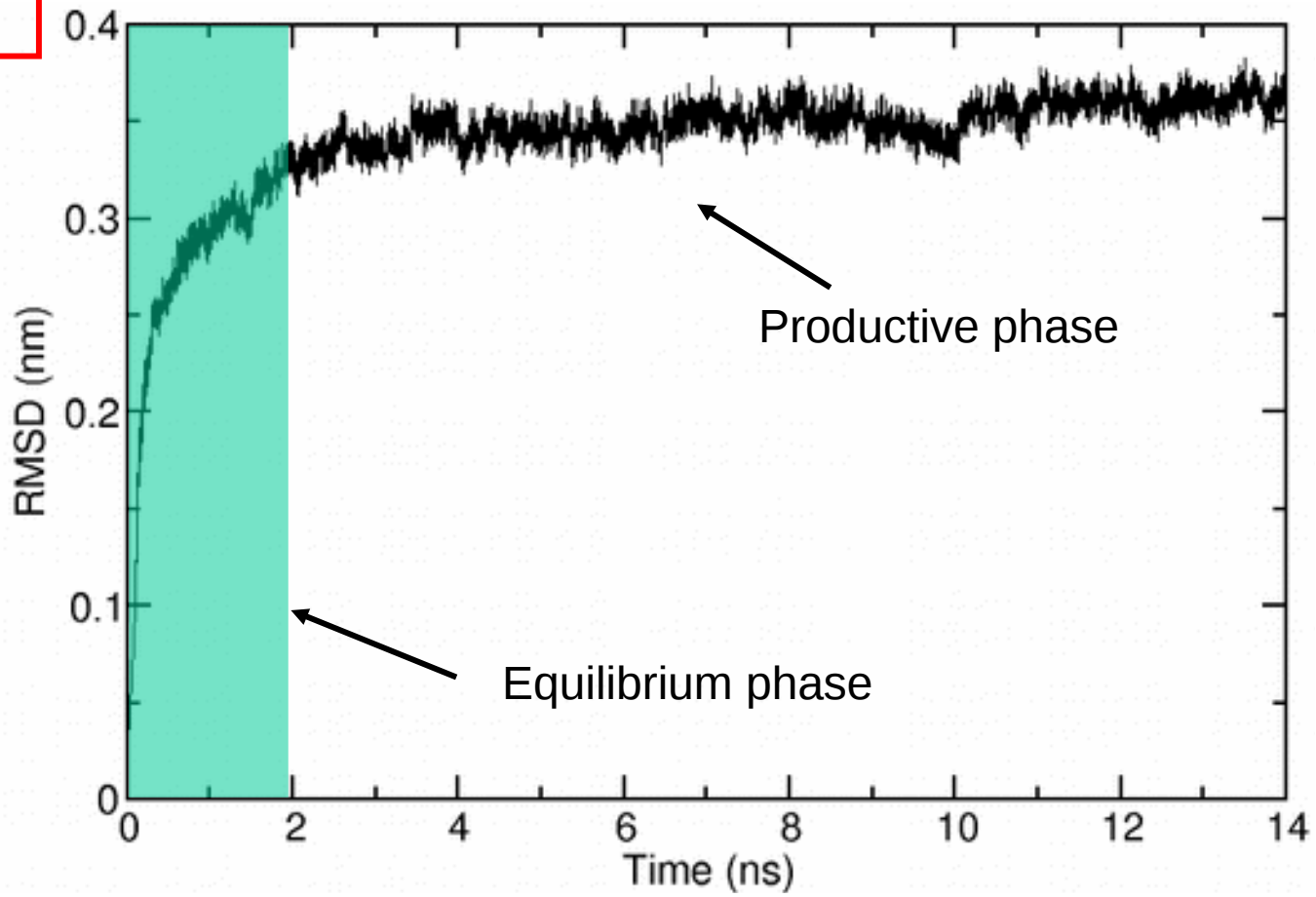
# Molecular Dynamics Simulations

https://youtu.be/gNSeN7NMJRA

# Did we reach equilibrium…?

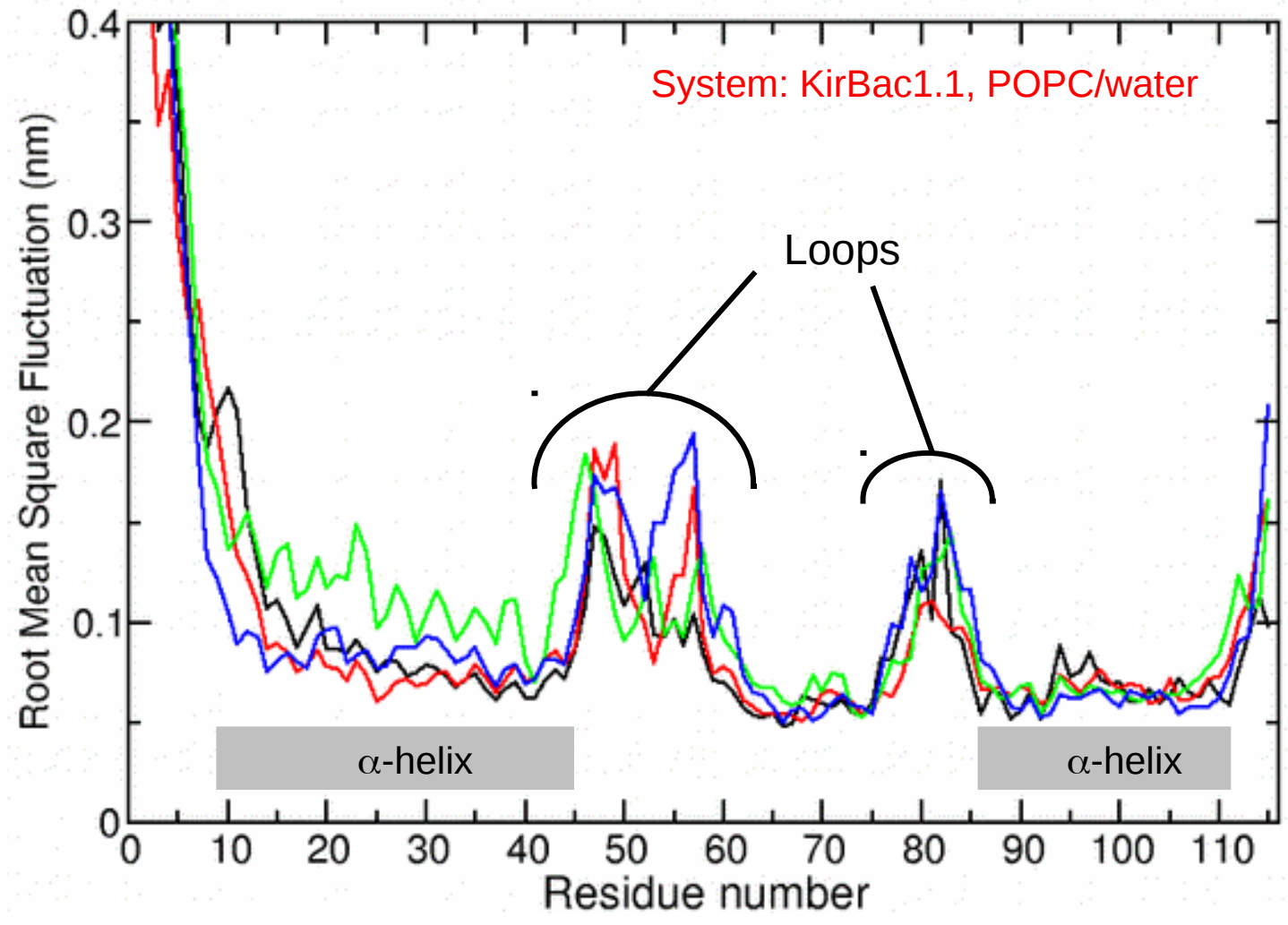
$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(r_i - r_0\right)^2}$$

We need to make sure that all the chemical and physical properties of the system have reached an equilibrium, where their averages do not longer change as a function of time. A simple way to test this is by measuring the RMSD (root mean square deviation) of Cα carbon atoms position with respect to start.



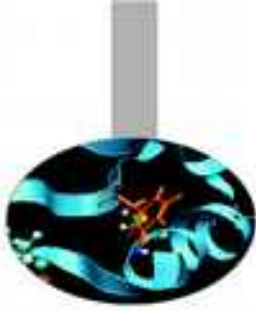
# Measuring chain flexibility

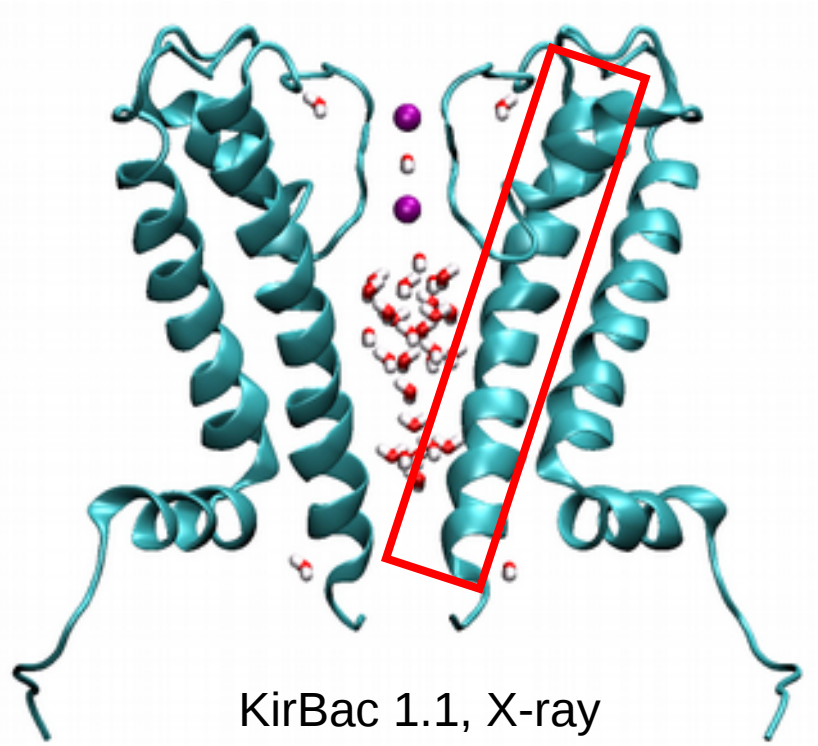
$$\mathrm{RMSF} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(r_i - \langle r \rangle\right)^2}$$



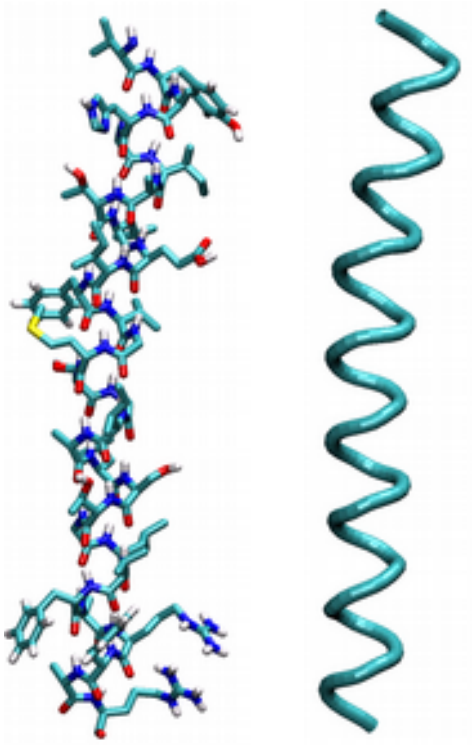

RMSF is a simple tool to measure the rigidity of the polypeptide chain. It calculates the deviations of C-alpha atoms coordinates from their average position. The flexibility pattern reflects the location of secondary structure elements in the protein structure.

# g_anaeig: the flag –filt

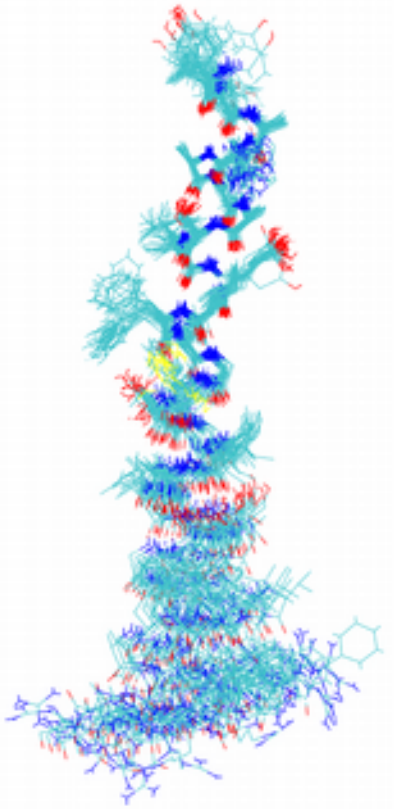




KirBac 1.1, X-ray



X-ray structure

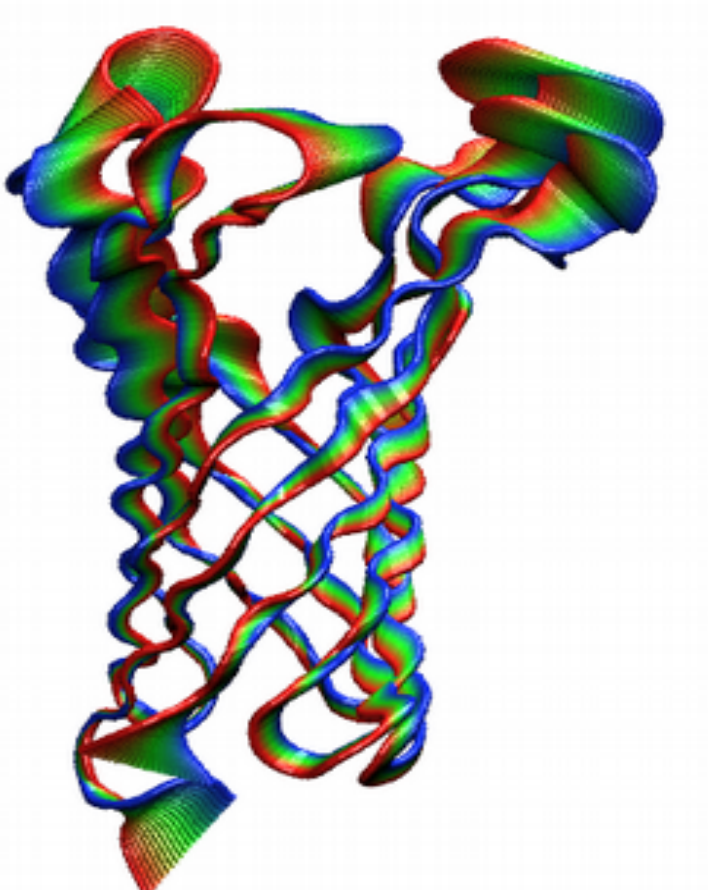


MD simulation

# M2 helix KirBac 1.1, raw data

# M2 helix KirBac 1.1 first eigenvector

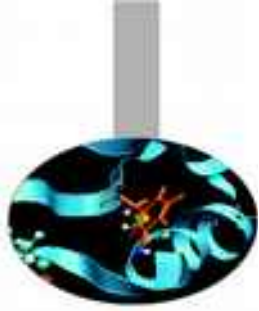
# Visualizzation of trajectories

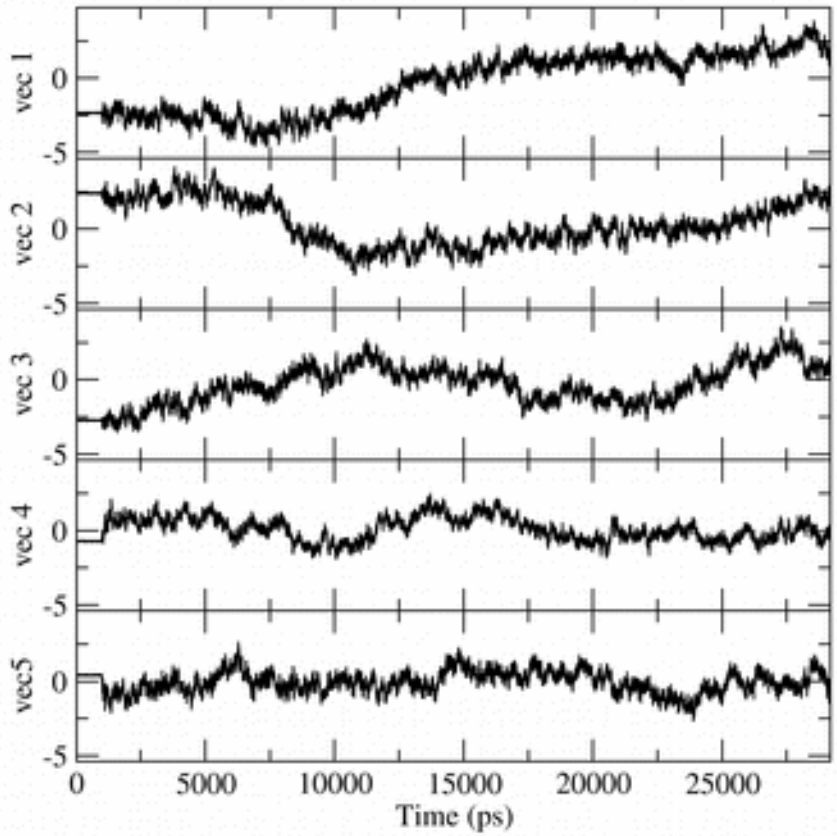


Tube representation of a filtered trajectory onto the first and second eigenvectors of the atomic fluctuation covariance matrix of porin OmpA

Picture produced with RasMol

# g_anaeig: output of flag –proj

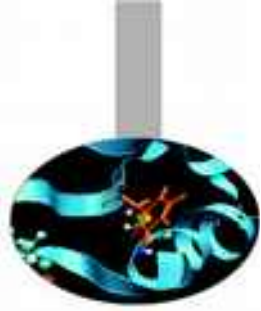


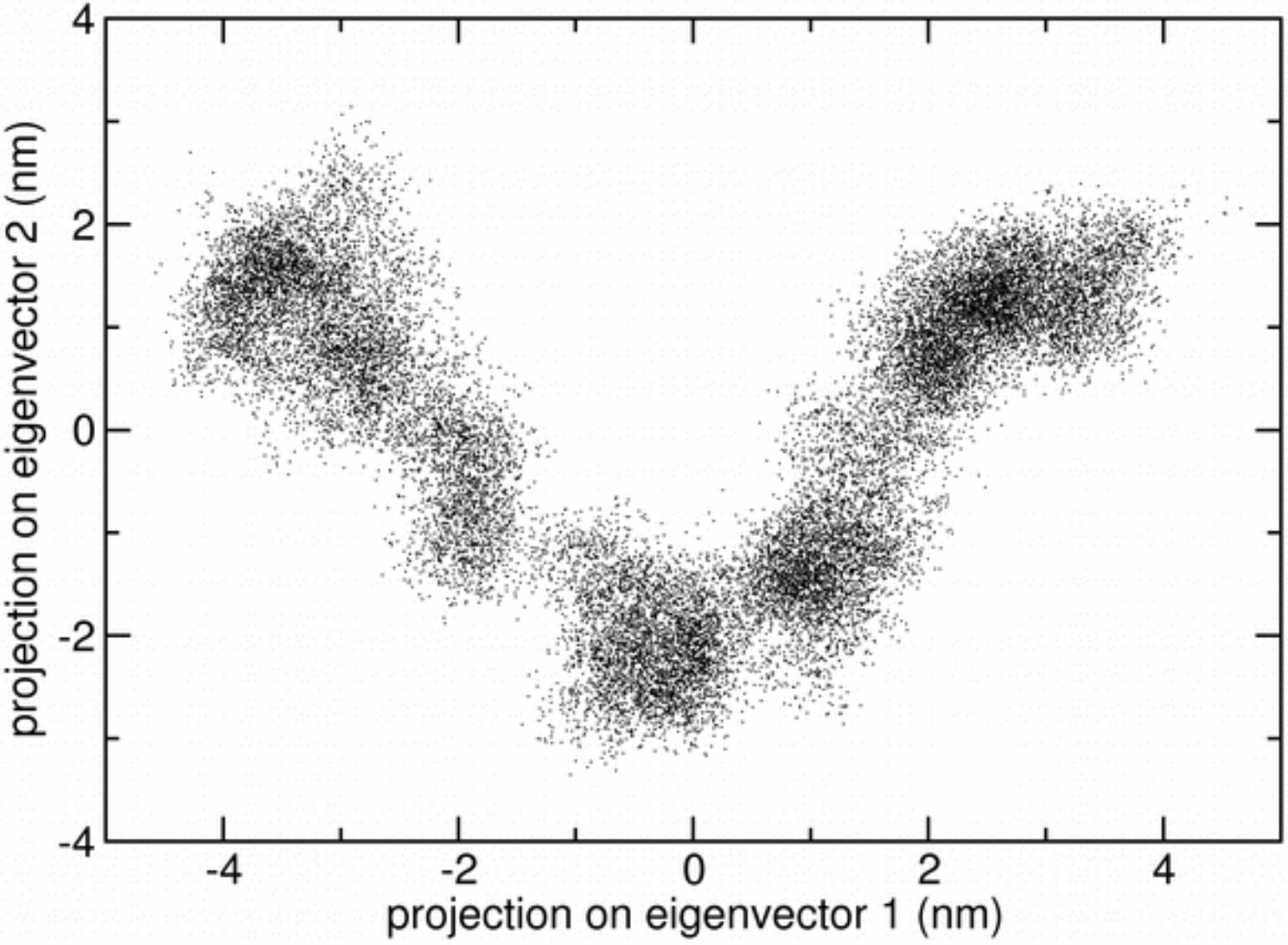


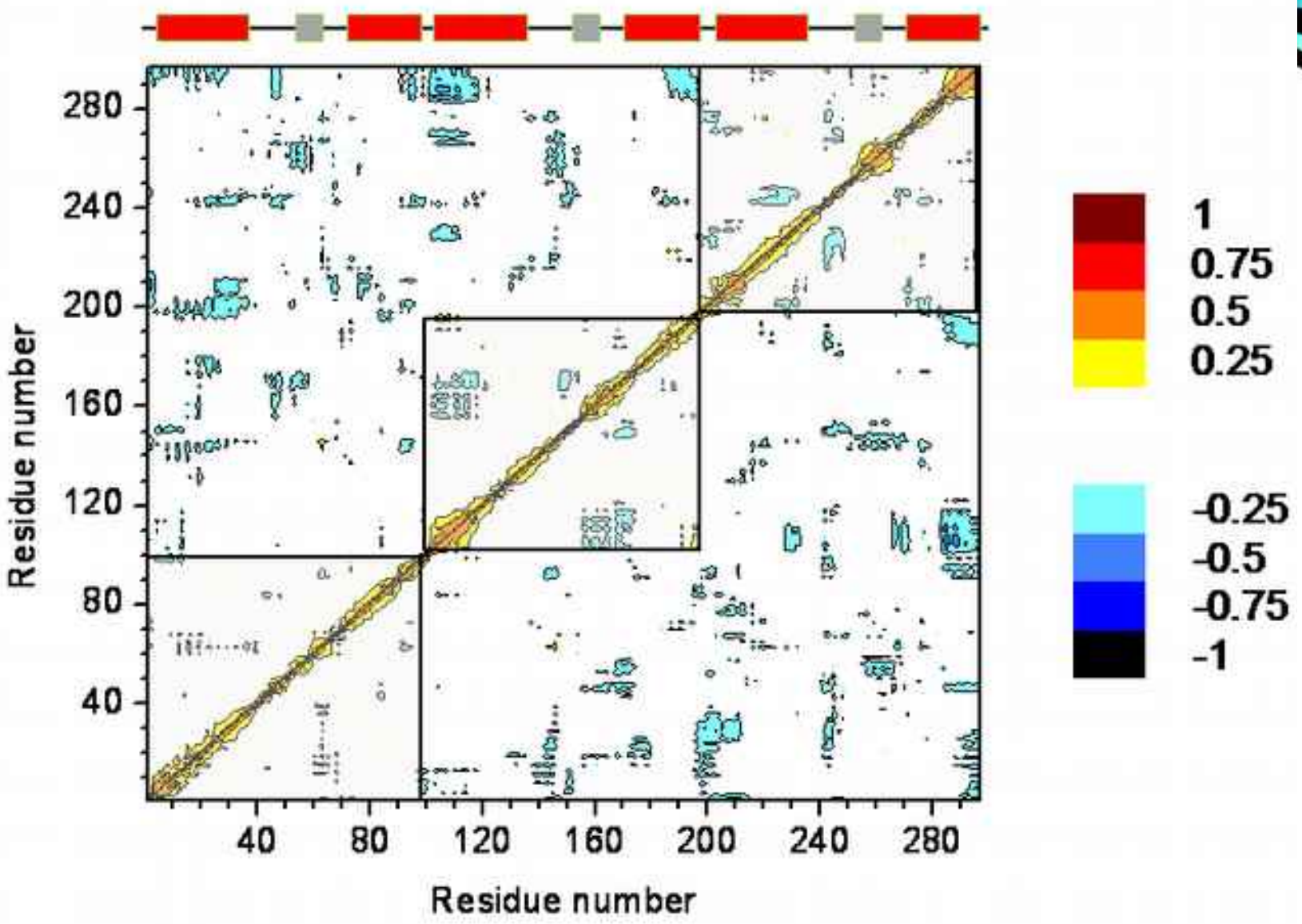

By default, 8 eigenvectors are considered for output using g_anaeig. This option can be set by using the flags `-first` and `-end`

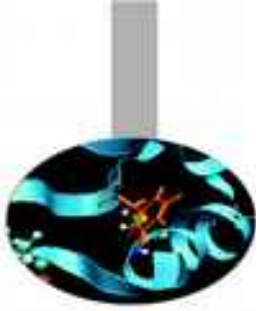
g_anaeig -f trajectory.xtc -v eigenvec.trr -eig eigenval.xvg -s reference.gro
-proj proj.xvg -first 1 -last 5

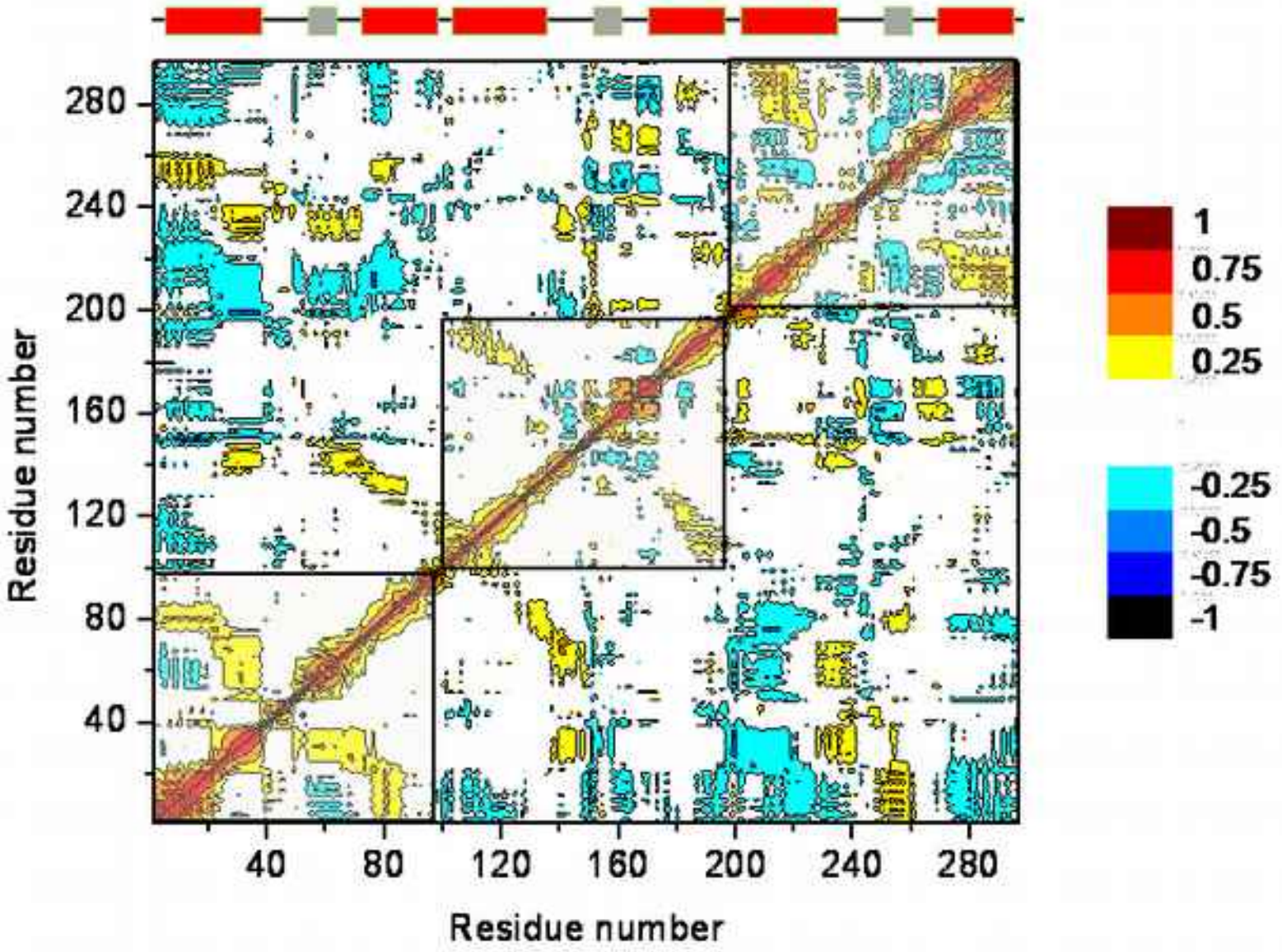
# Graphic representation of classic MD simulations






Residue number
280
240
200
160
120
80
40
40
80
120
160
200
240
280
Residue number
1
0.75
0.5
0.25
-0.25
-0.5
-0.75
-1




Residue number
280
240
200
160
120
80
40
40
80
120
160
200
240
280
Residue number
1
0.75
0.5
0.25
-0.25
-0.5
-0.75
-1