Protein structure and dynamics

Tamás Hegedűs

tamas.hegedus@hegelab.org

Dept. of Biophysics and Radiation Biology

Importance of protein dynamics

The atomic level basis of a disease…
The shape of a drug binding site…

There is no single structure
but a conformational ensemble at 37°C

Importance of computational modelling

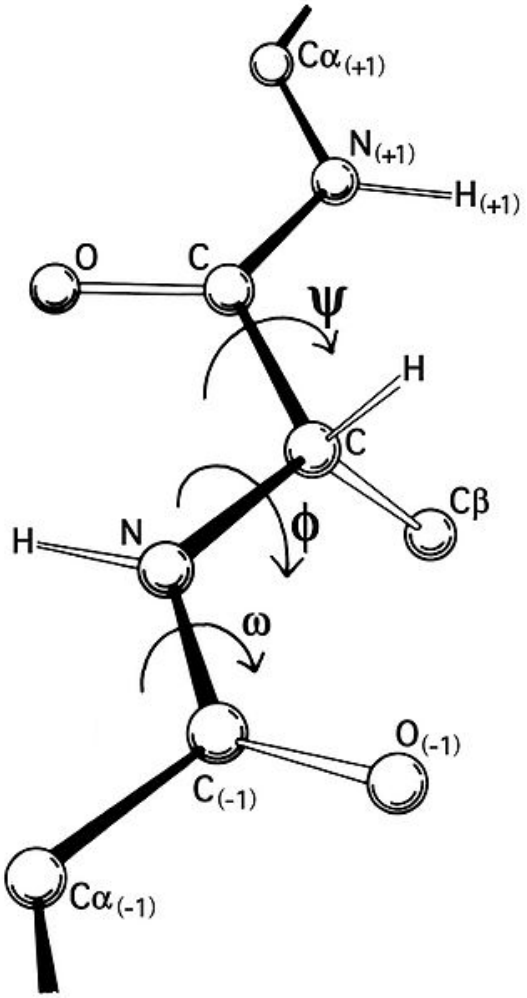
Atomic level information on motions

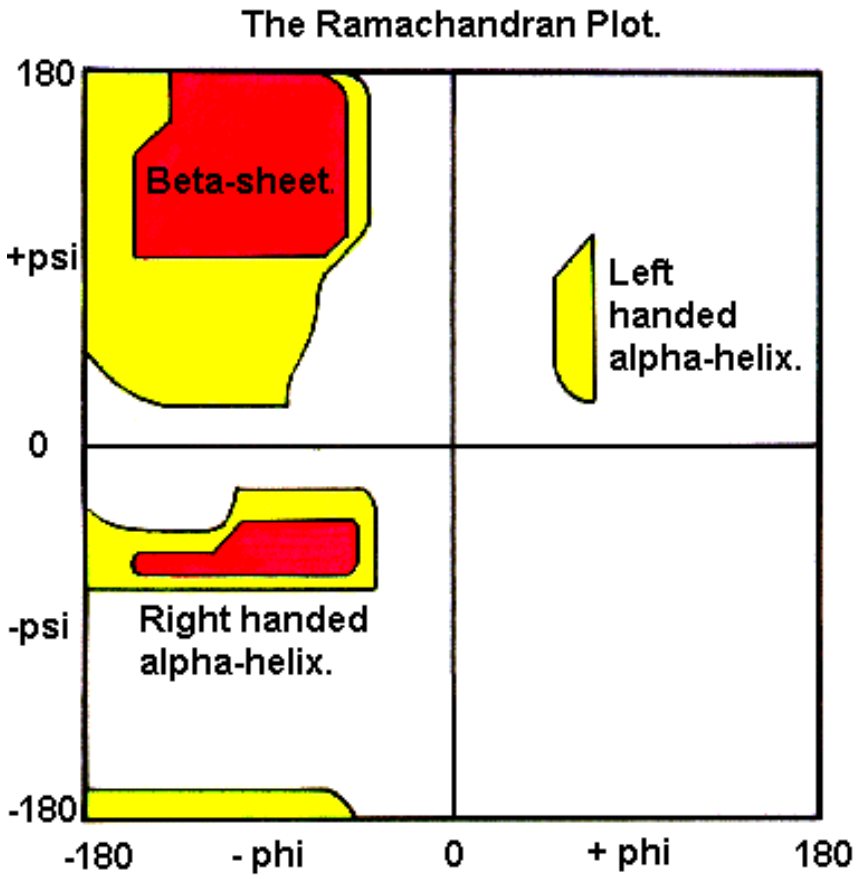
Experiments usually do not provide atomic level information (there are exceptions, such as NMR)

Topics

- **Introduction to protein structure and dynamics**
- **Characterization of protein structure**
  - **Prediction of secondary structure**
  - **Intrinsically disordered proteins**
  - **Tertiary structure**

Secondary structure





wikipedia

Prediction of secondary structure

Using only available structures **60 %**

Combining with sequence aligments **70-80 %**

Implementations:

- **neural networks,**
- **support vector machines,**
- **hidden Markov models, etc.**

Scoring each positions
GOR4, HNN, Prof, JPred/JNet

Intrinsically Disordered Proteins

- **25 % of proteins are predicted disordered**
- **Increased disorder with increased complexity**

  50 % of human proteins contains a disordered region with 30 a.a. or longer

- **Not fully random**
- **Structure flexibility**
- **No compact globular folding and residual structure**

The paradigm
protein function needs a well-defined 3D structure
has changed.

Intrinsically Disordered Proteins

Benefits

- Specificity and adaptation
- Reversible transition between ordered/disordered states
- Large binding surface
- Fast binding

Roles

- Entropic chain: inactivation of K^+ channels
- Effectors: peptide inhibitors
- Scavengers: casein
- Assembly: calmodesmon, F-actin
- Presentation: phosphorylation and cleavage sites

Intrinsically Disordered Proteins

K. Dunker, Péter Tompa, Zsuzsa Dosztányi

DisProt database: http://www.disprot.org

Disorder prediction

Learning algorithms
Based on disordered sequences
in the PDB database

Disopred2

Predicting interaction energies

IUPred.enzim.hu

IUPred, statistical potentials

For an existing 3D structure:

$$E_{calculated} = \sum_{i,j} M_{ij} C_{ij}$$

There is only protein sequence:

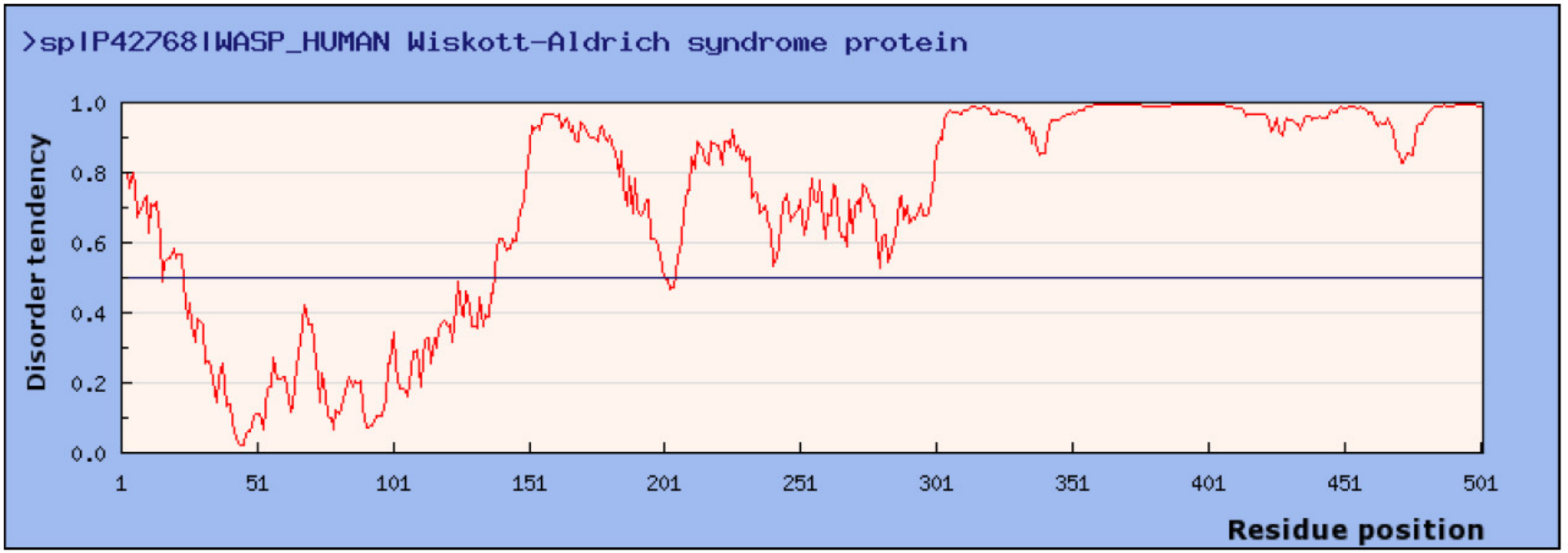
$$E_{estimated} = L \sum_{i,j} P_{ij} f_i f_j$$

Disorder level of an a.a.:

$$E_j^k = \sum_{i=1}^{20} P_{ij} f_i^k (w_0)$$

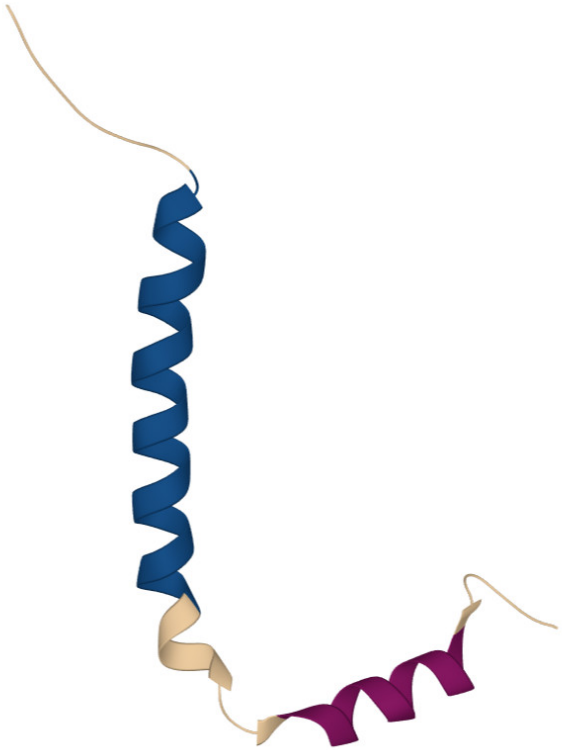
B Mészáros, G Erdős, Z Dosztányi
IUPred2A
Nucleic Acids Research 2018;46(W1):W329-W337.

An IUPred output



MemMoRF

Membrane Molecular Recognition Feature

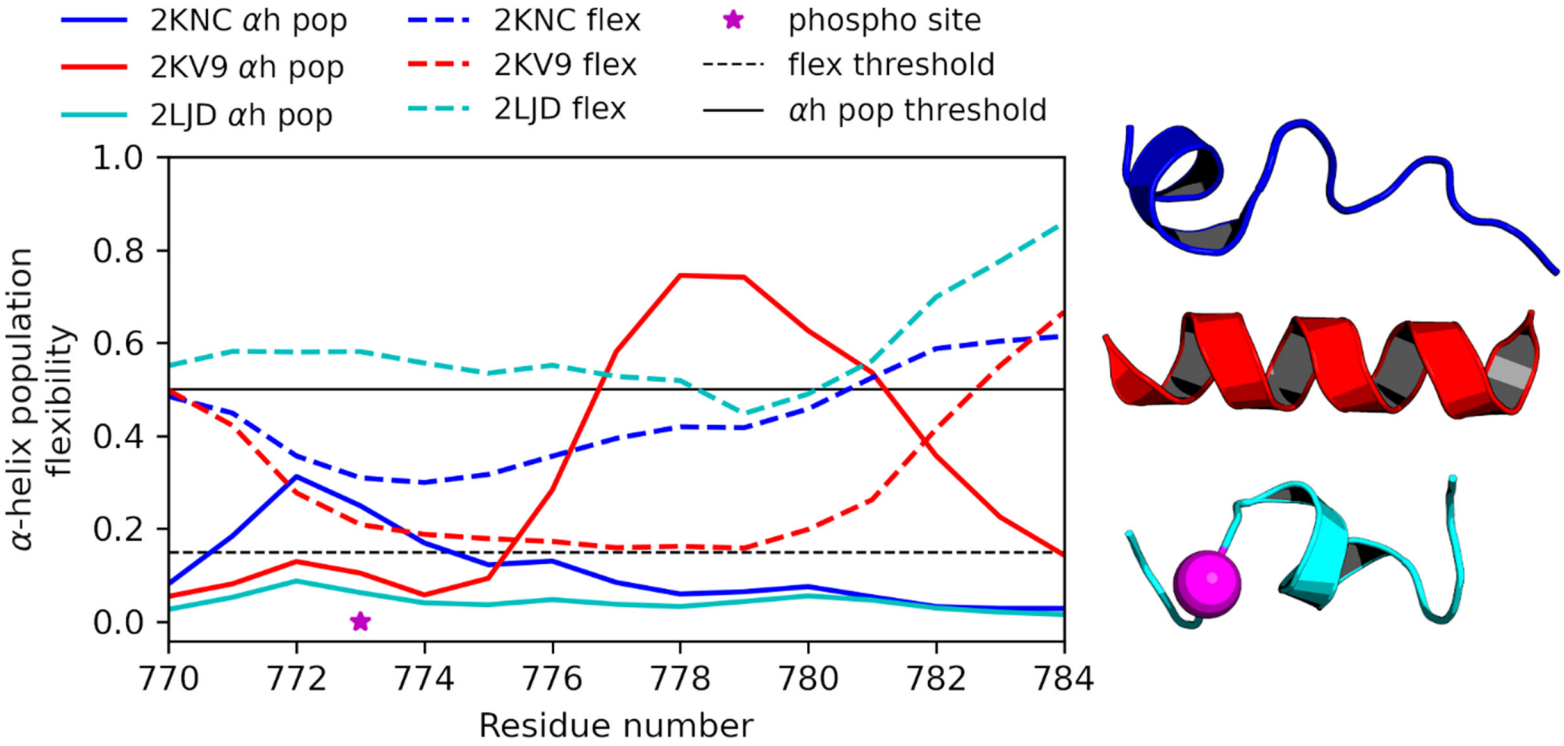


EGFR, PDBID: 2N5S

MemMoRF

Membrane Molecular Recognition Feature

Integrin beta-3



blue: in organic solvent
red: in DPC
cyan: phosphorylated in DPC

magenta: phosphorylation site; αh pop: α-helix population calculated by δ2D;
flex: 1-S2 calculated by RCI, αh pop threshold: 0.5, flex threshold: 0.15.

http://memmorf.hegelab.org

3D structure prediction

Ab initio folding

- CASP (Critical Assessment of Techniques for Protein Structure Prediction)
- constraints from experiments

Homology modelling

- conserved sequence == conserved structure
- > 30% similarity
- most important: the sequence alinement

Threading

- If similarity is low

Homology modelling

- Searching a template
- Sequence alignment
- Modelling
- Energy minimization

BLOSUM
(BLOcks of Amino Acid SUbstitution Matrix) matrix
is a substitution matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

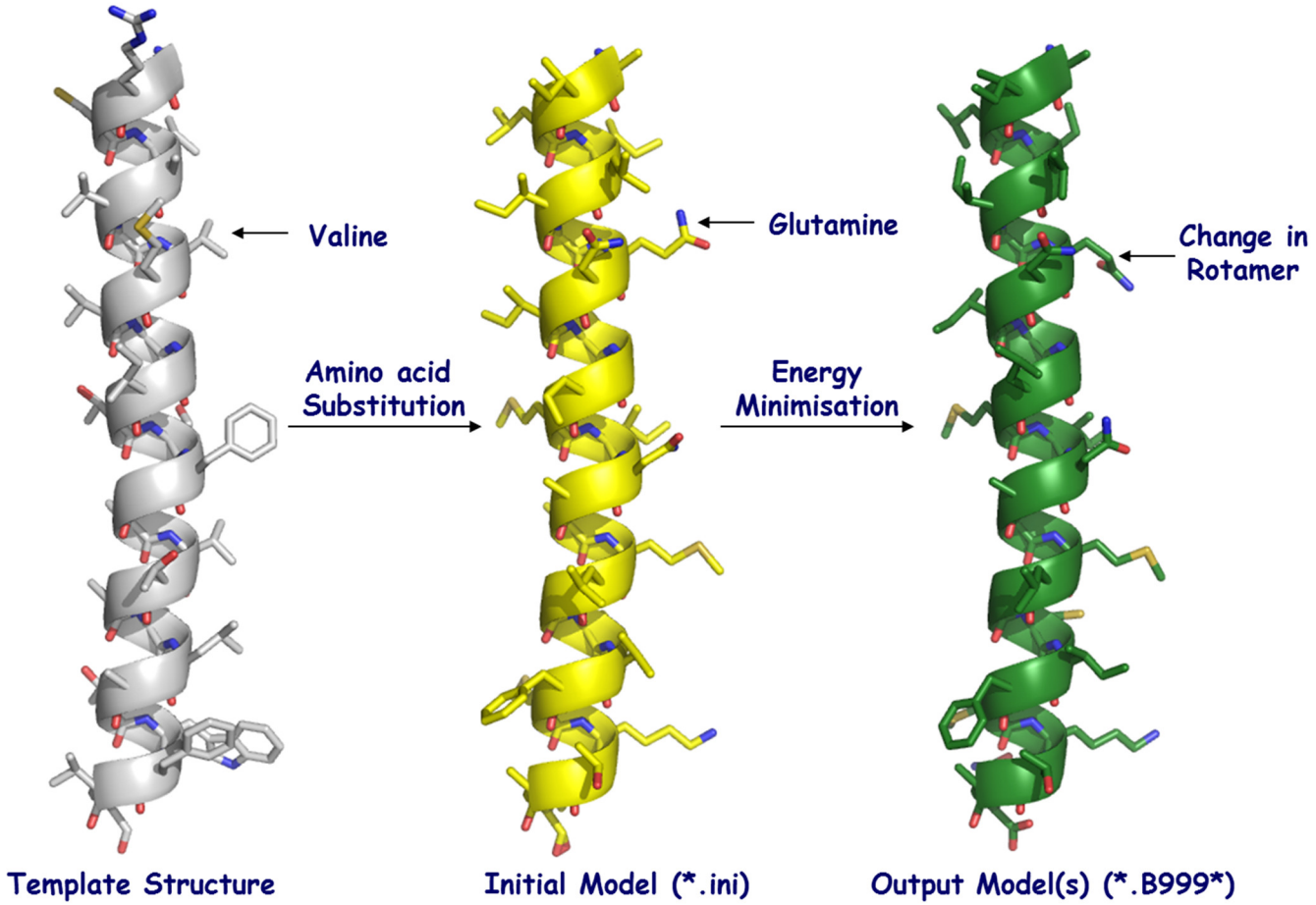
Basic Local Alignment Search Tool (BLAST)

Alignement – pl. ClustalW

```
CLUSTAL W (1.83) multiple sequence alignment


2HYD        -----------MIKRYLQFVK-----PYKYRIFATIIVGIIKFGIPMLIP
3B5X        ---------WQTFKRLWTYIR-----LYKAGLVVSTIALVINAAADTYMI
CFTR_HUMAN  MQRSPLEKASVVSKLFFSWTRPILRKGYRQRLELSDIYQIPSVDSADNLS
                              *    : :      *:  :  : *  : .      :
```

Homology modelling

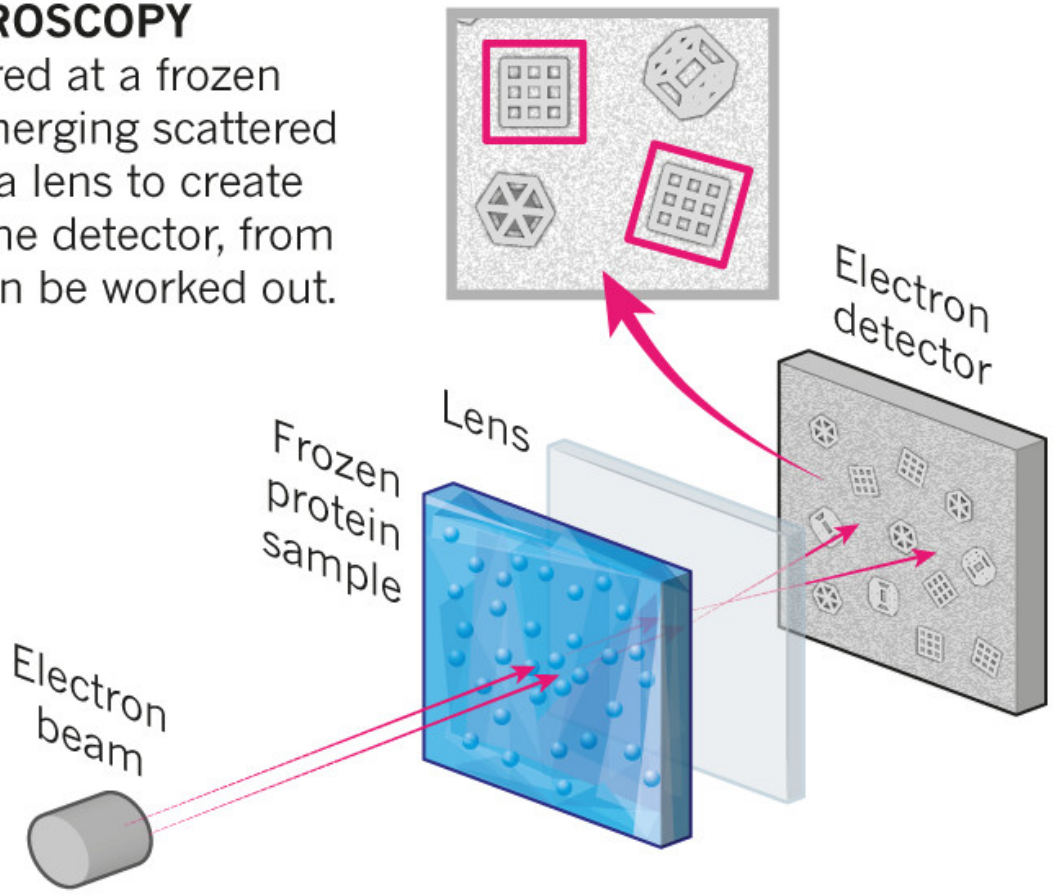


source: SBCB, Oxford, UK

Structure determination – „single particle”

Cryo-electron microscopy

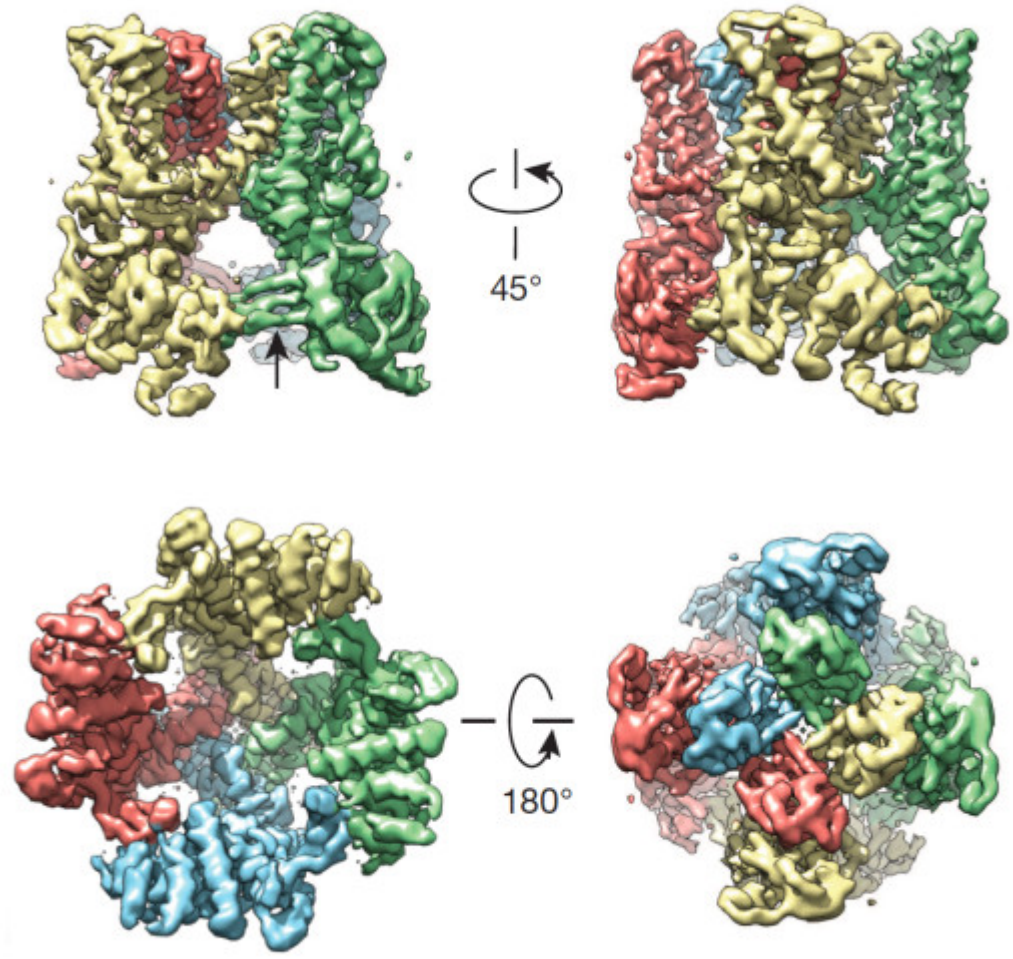
CRYO-ELECTRON MICROSCOPY
A beam of electron is fired at a frozen protein solution. The emerging scattered electrons pass through a lens to create a magnified image on the detector, from which their structure can be worked out.



Ewen Callaway, Nature | News Feature
The revolution will not be crystallized: a new method sweeps through structural biology, 09 September 2015

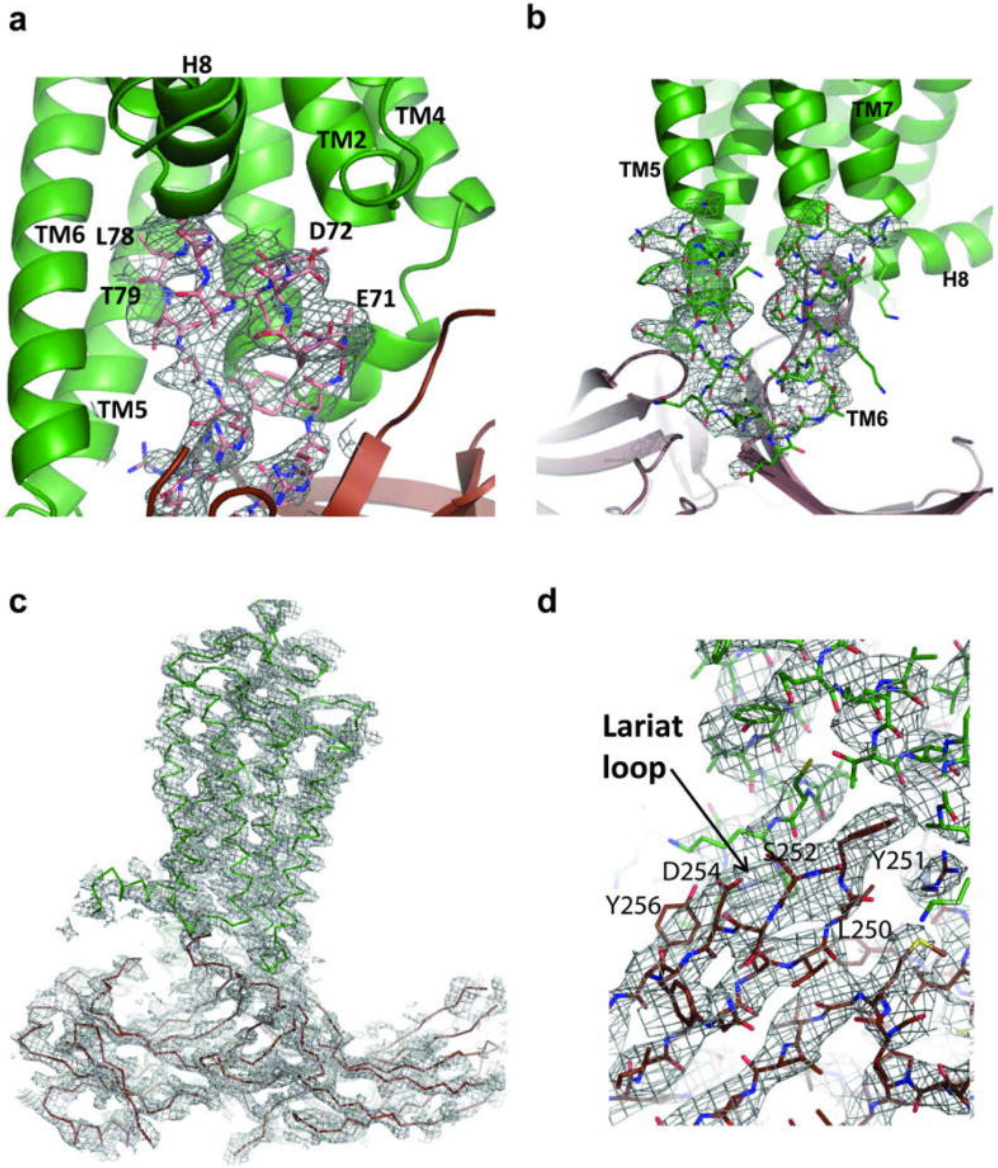
Structure determination – „single particle”

“The TRPV1 channel detects the burn of chilli peppers, and this 3.4-Å structure is considered super-hot in the structural-biology world.”



Structure determination – „single particle”

Free Electron Laser (FEL)

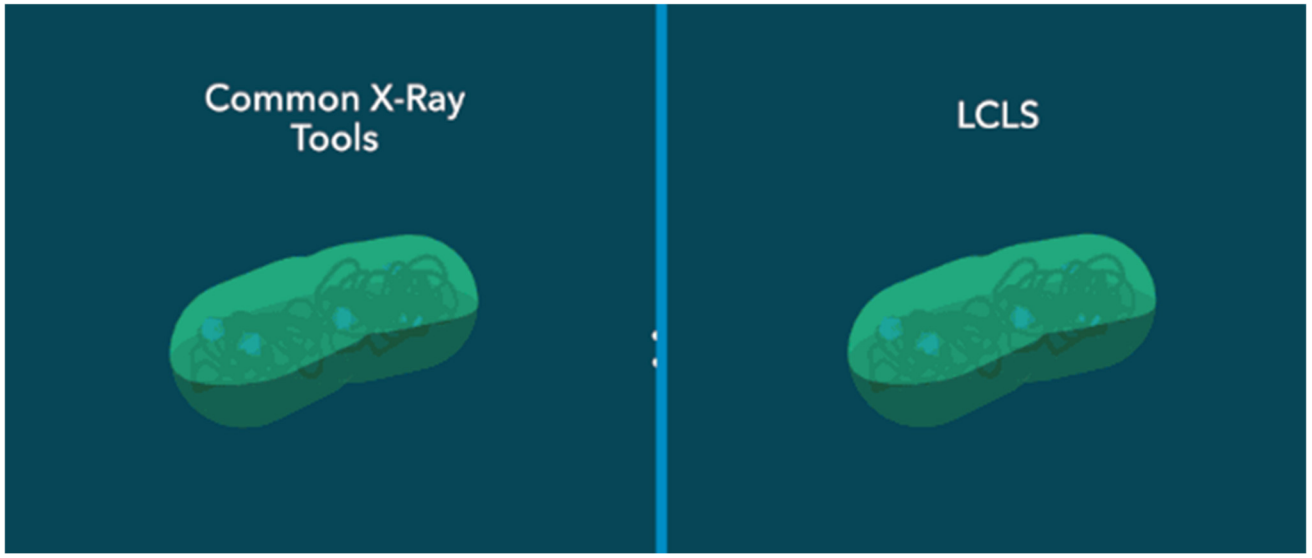


Nature. 2015 Jul 30;523(7562):561-7.

Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser.

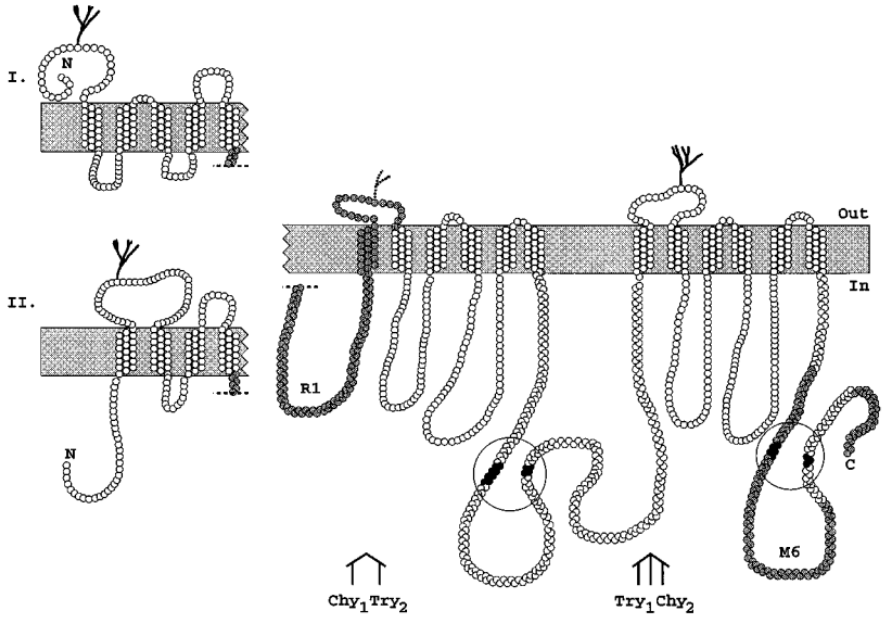
Structure determination – „single cell”

Free Electron Laser (FEL)



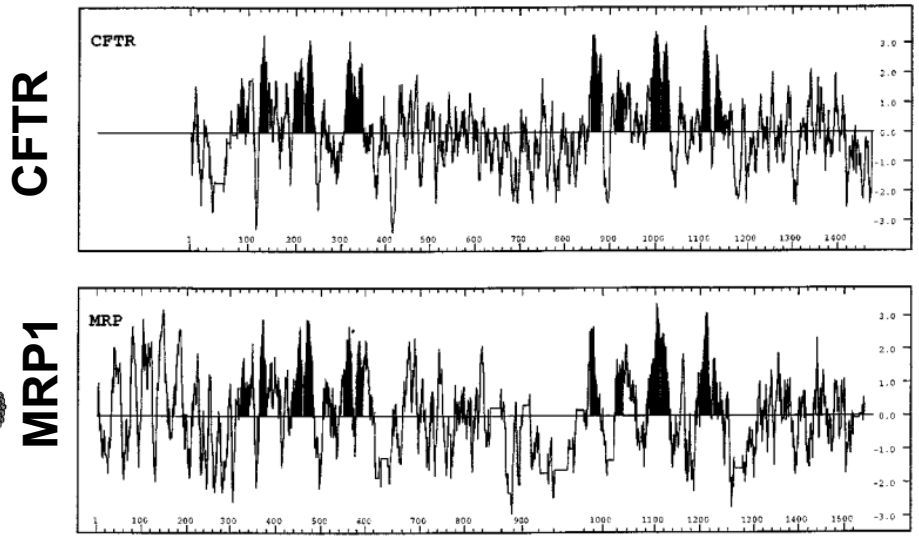
Membrane protein topology

MRP1



Bakos *et al.* JBC 1996

Hydrophobicity profiles:



CFTR topology is known from experiments

Chang et al. J Biol Chem. 1994 Jul 15;269(28):18572-5

Experimental methods for topology determination:

- *tag* recognition
- Cys accessibilitly

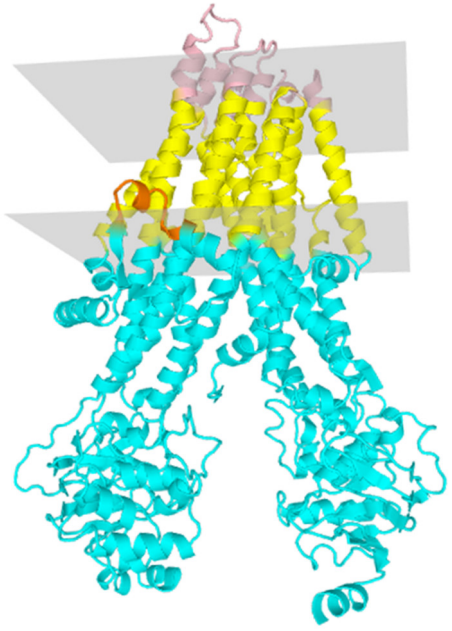
Prediction of membrane topology

- **Based on chemical properties of amino acids**
- **a.a. distribution in TM and soluble regions (statistics)**
- **Incorporation of experimental knowledge**
- **Integration of several predictors**

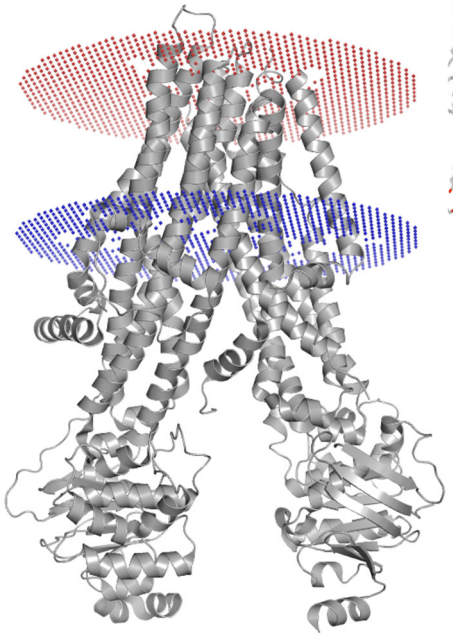
e.g. CCTOP.enzim.hu

Prediction of TM helices based on structure

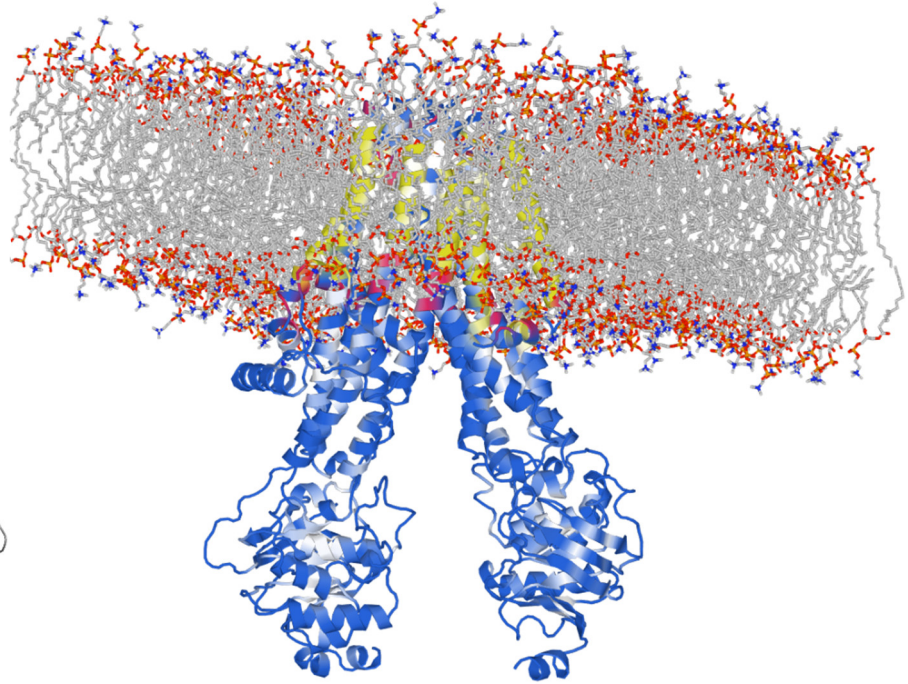
TMDET
PDBTM



PPM
OPM

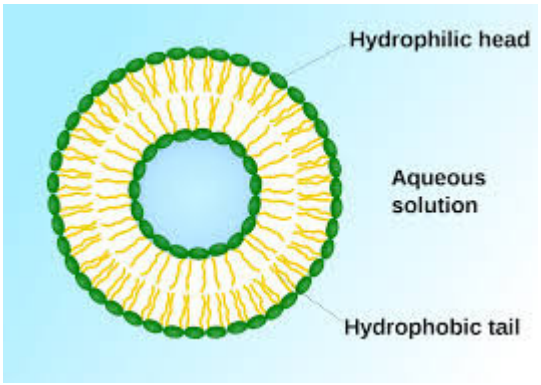


MemProtMD

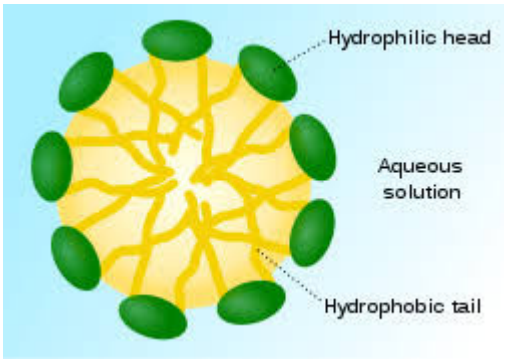


Membrane mimetics

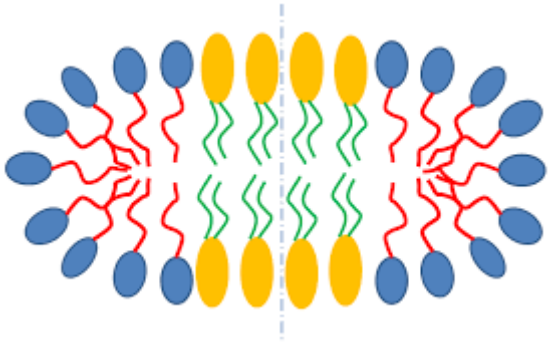
Vezikula



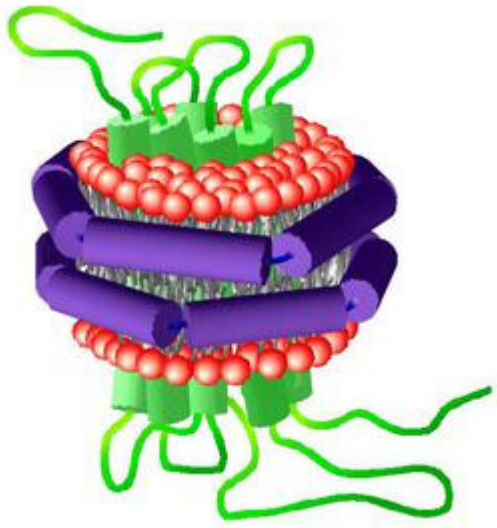
Micella



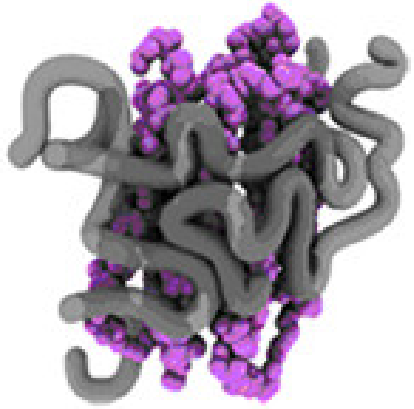
Bicella



Nanodisc

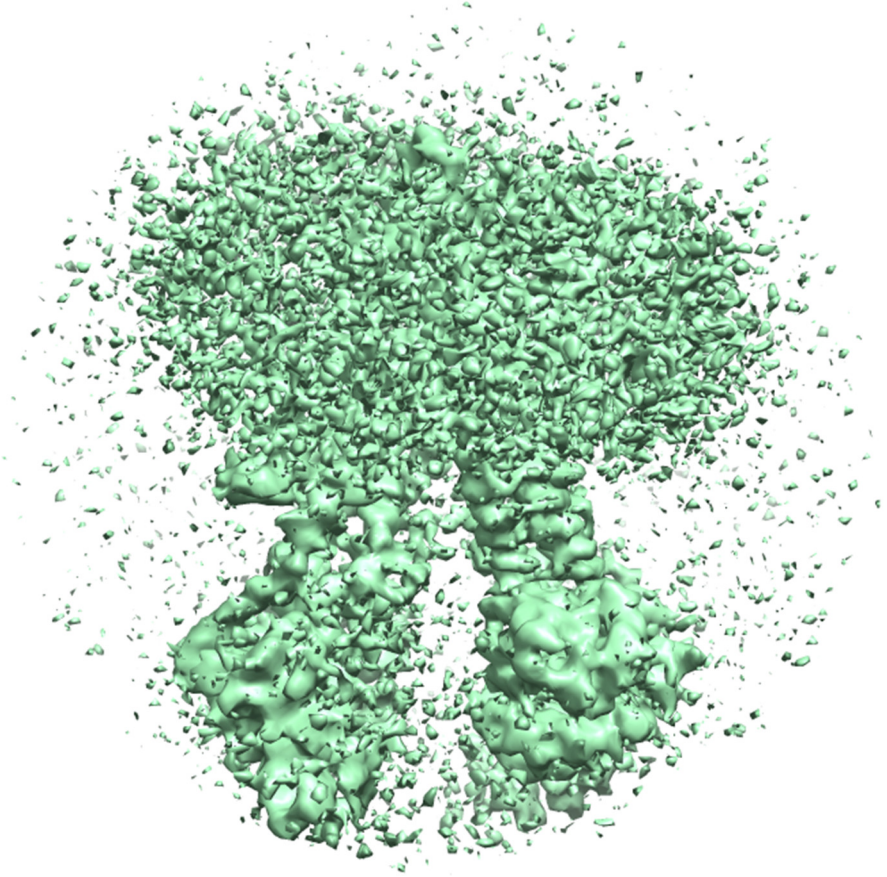


Amphipol

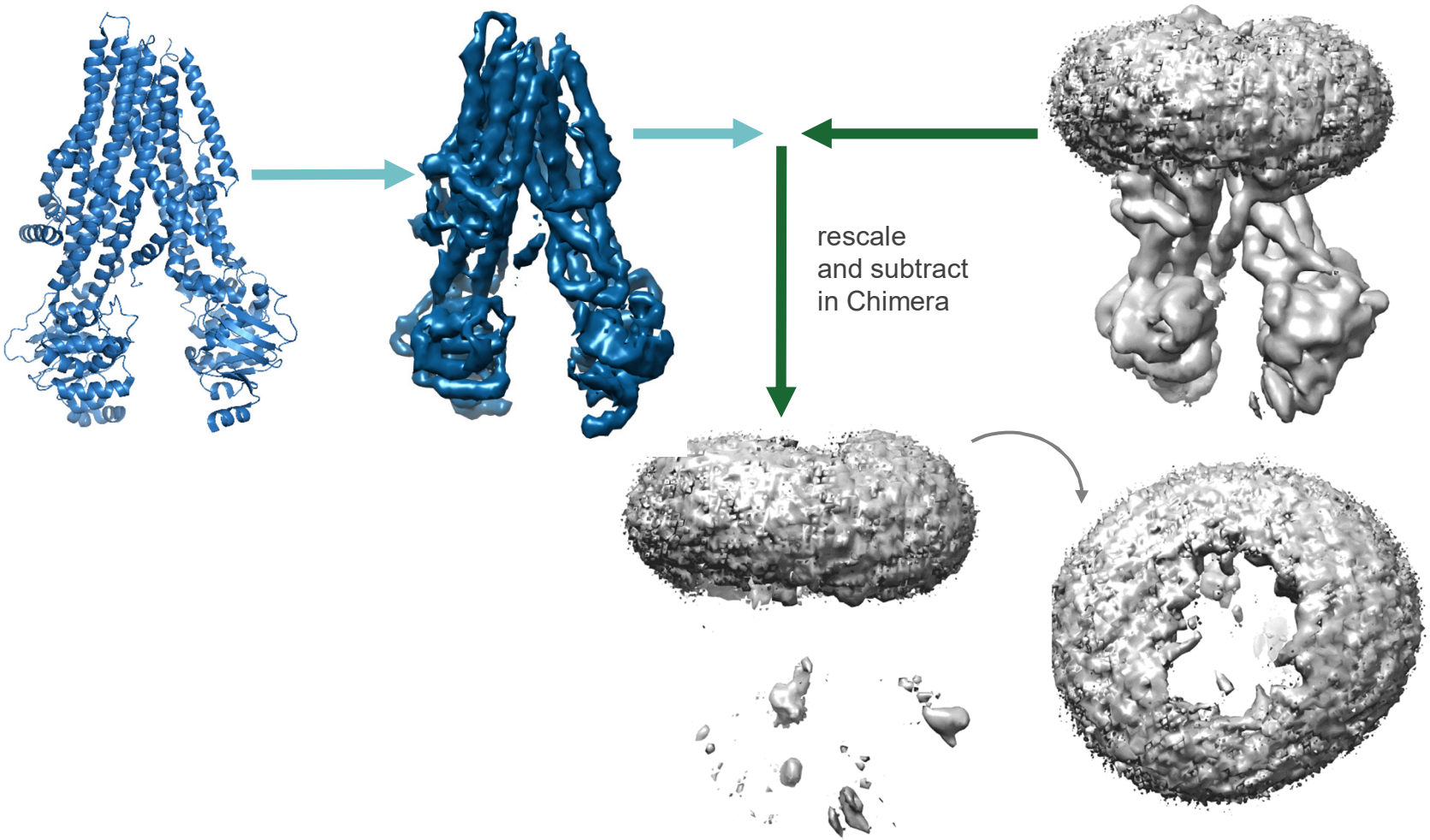


Membrane embedment data is in the electron density maps

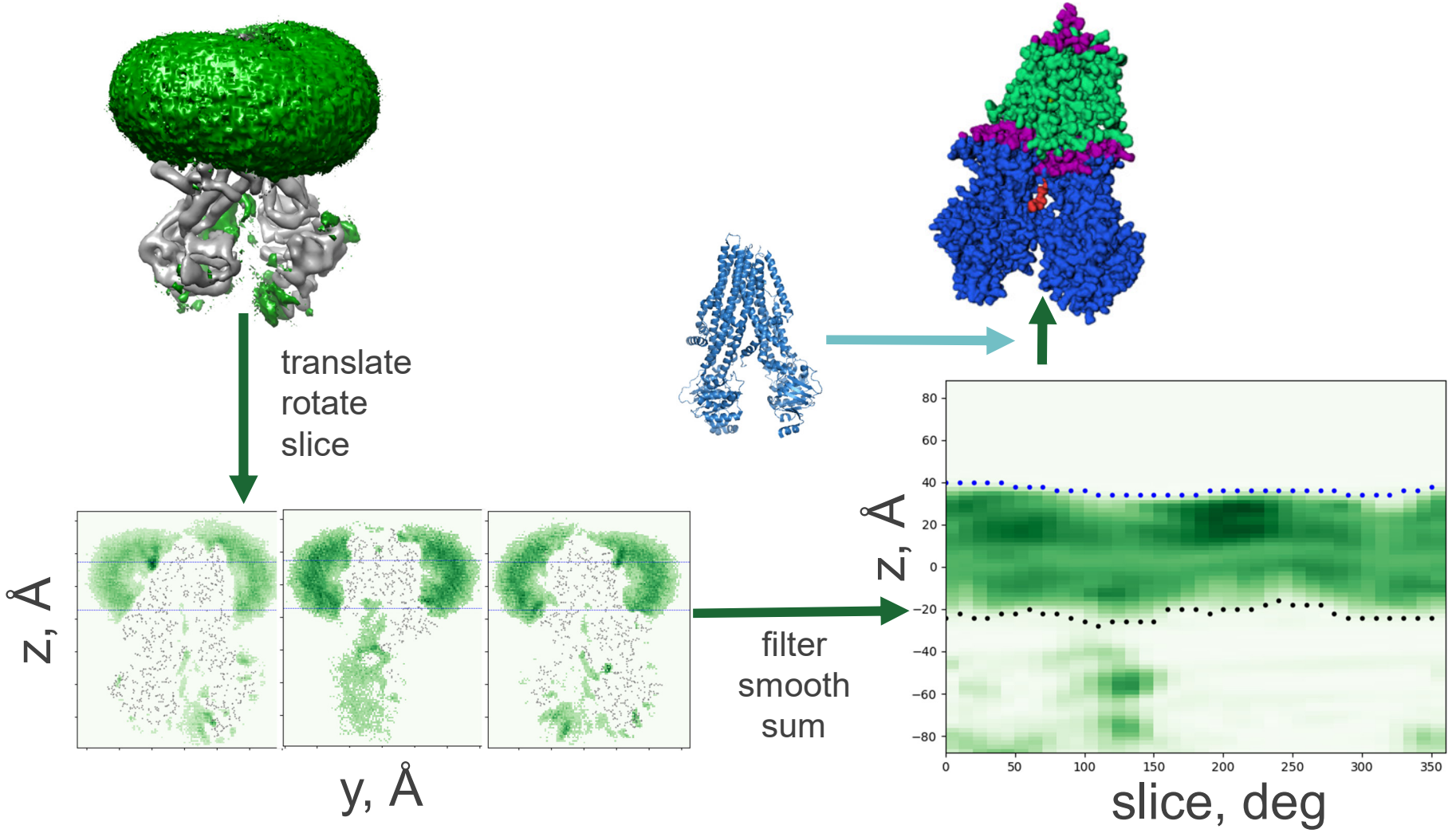
CFTR (PDBID: 5UAK) EMD



A membrane blob can be extracted from the electron density map



The MemBlob can be converted to membrane boundaries



Protein-protein interactions

Docking of proteins – challenging (surface shape, dynamics)

PISA - Protein Interfaces, Surfaces and Assemblies

Molecular Dinamics

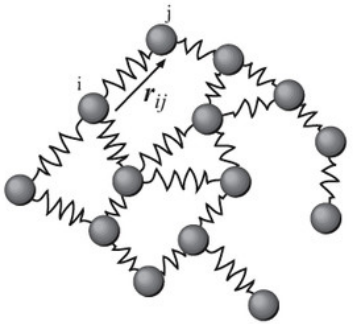
Topics

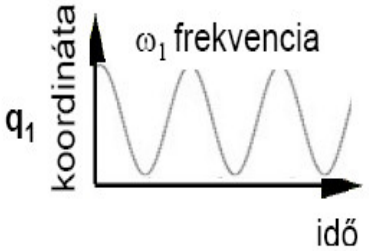
- **Protein structure**
- **Protein dynamics**
- **Protein folding**

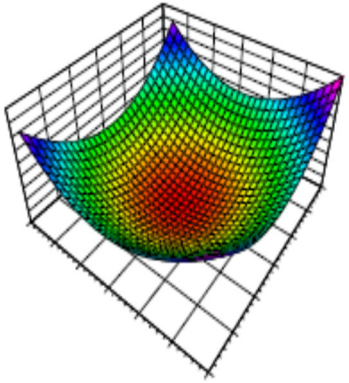
Methods for studying protein dynamics

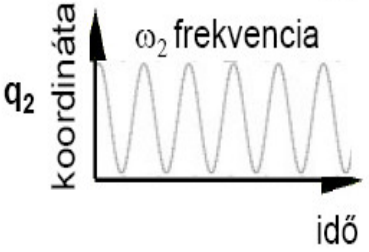
Normal mode analysis

- harmonic potential
- analytic equation of motions
- normal modes









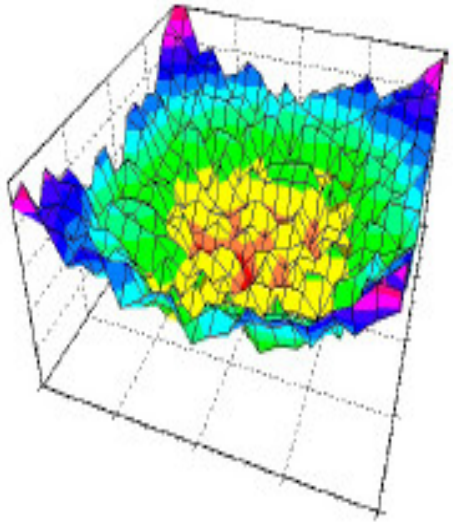
- Gaussian network model (GNM)
  - mean squared displacements
  - cross-correlations between fluctuations
- Anisotropic network model (ANM)
  - directionality by projection of motions to a mode space of N dimensions

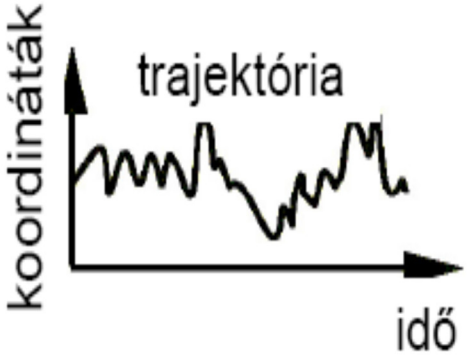
Tools: http://prody.csb.pitt.edu

Methods for studying protein dynamics

Molecular dynamics

- realistic potential surface
- numerical integration of Newton's equations
- a system of interacting particles
- forces between the particles and their potential energies are calculated by using interatomic potentials (molecular mechanics force fields)
- output: trajectory





The force field

$$E_{\text{prot}} = W_{\text{rot}} E_{\text{rot}} + W_{\text{atr}} E_{\text{atr}} + W_{\text{rep}} E_{\text{rep}} + W_{\text{solv}} E_{\text{solv}} + W_{\text{pair}} E_{\text{pair}}$$
$$+ W_{\text{mbenv}} E_{\text{mbenv}} + W_{\text{hbond}} E_{\text{hbond}} - E_{\text{ref}}$$

$$E_{\text{solv}} = -\sum_{i}^{\text{natom}} \sum_{j>i}^{\text{natom}} \left\{ \frac{2\Delta G_i^{\text{free}}}{4\pi\sqrt{\pi}\lambda_i r_{ij}^2} \exp\left(-d_{ij}^2\right) V_j + \frac{2\Delta G_j^{\text{free}}}{4\pi\sqrt{\pi}\lambda_j r_{ij}^2} \exp\left(-d_{ji}^2\right) V_i \right\}$$

Lazaridis (2003)

TABLE I. Solvation Parameters[†]

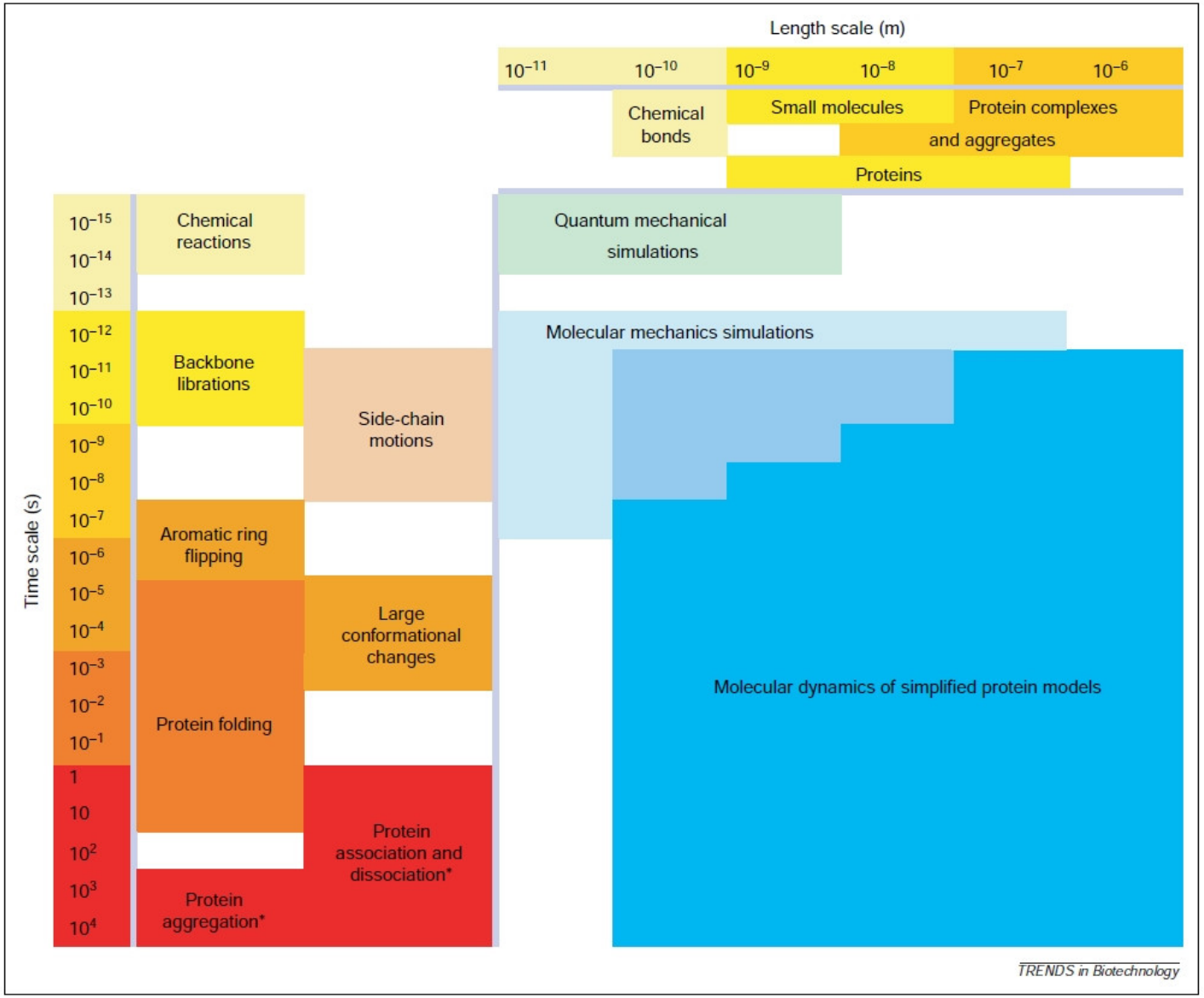
Atom types[a]	Volume	ΔG_i^{ref} [b]	ΔG_i^{free} [c]	ΔH_i^{ref} [b]	ΔCp_i^{ref} [d]
C	14.7	0.000	0.00	0.000	0.00
CR	8.3	−0.890	−1.40	2.220	6.90
CH1E	23.7	−0.187	−0.25	0.876	0.00
CH2E	22.4	0.372	0.52	−0.610	18.60
CH3E	30.0	1.089	1.50	−1.779	35.60
CR1E	18.4	0.057	0.08	−0.973	6.90
NH1	4.4	−5.950	−8.90	−9.059	−8.80
NR	4.4	−3.820	−4.00	−4.654	−8.80
NH2	11.2	−5.450	−7.80	−9.028	−7.00
NH3	11.2	−20.000	−20.00	−25.000	−18.00
NC2	11.2	−10.000	−10.00	−12.000	−7.00
N	0.0	−1.000	−1.55	−1.250	8.80
OH1	10.8	−5.920	−6.70	−9.264	−11.20
O	10.8	−5.330	−5.85	−5.787	−8.80
OC	10.8	−10.000	−10.00	−12.000	−9.40
S	14.7	−3.240	−4.10	−4.475	−39.90
SH1E	21.4	−2.050	−2.70	−4.475	−39.90

Lazaridis (1999)

The limitations of MD

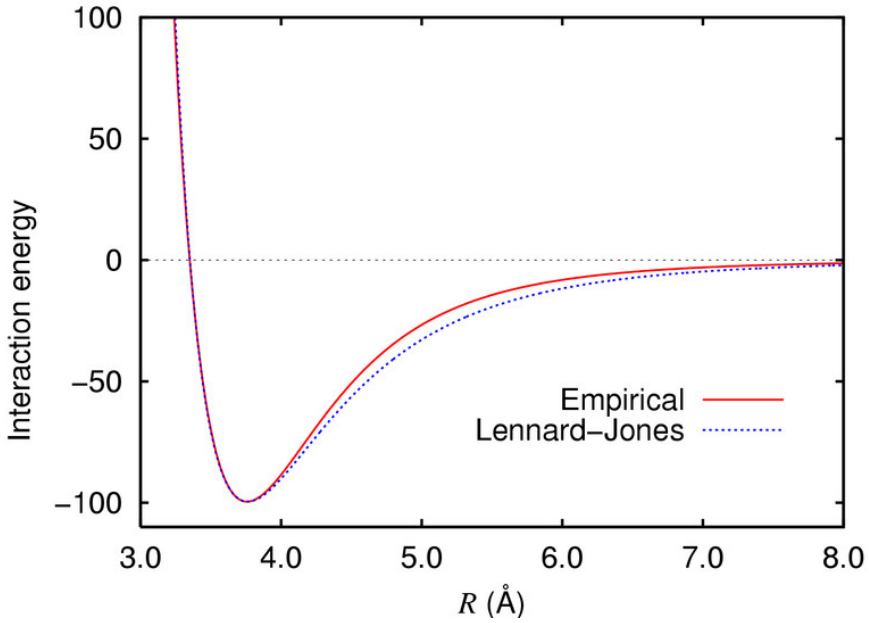
- time (computation time versus real time)
- calculation of the potential is the bottle-neck
- fs long integration steps
- „periodic boundary condition”
- solvent (explicit/implicit)

The time scale of various molecular events

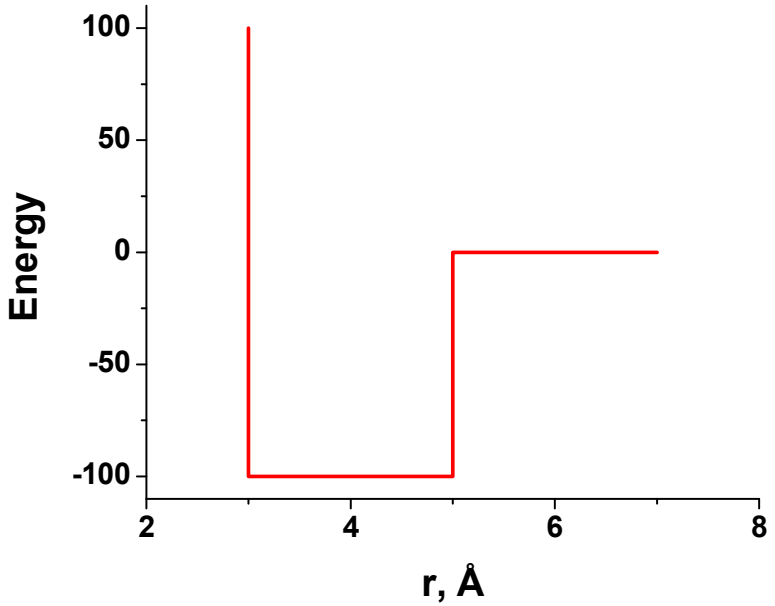


F. Ding and N.V. Dokholyan, TRENDS in Biotechnology, **23:**450 (2005)

Discrete Molecular Dynamics (DMD)



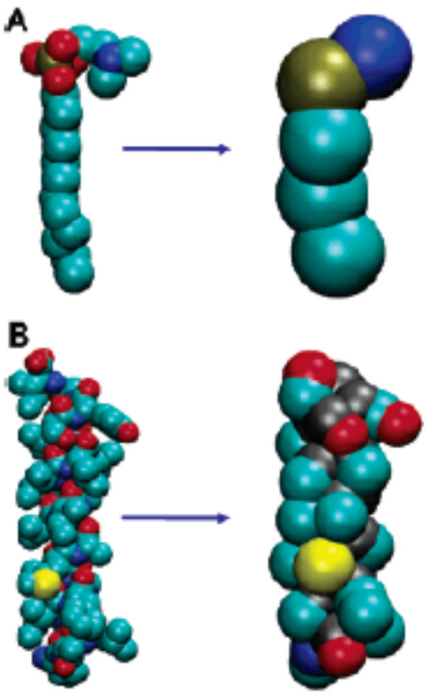
wikipedia

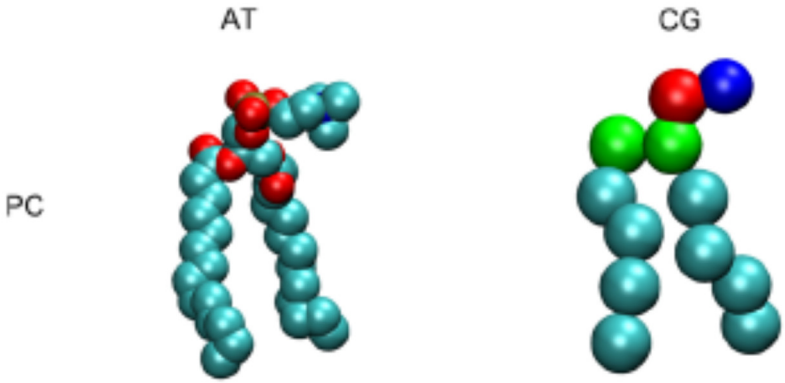


Ding, F., Dokholyan, N. V. PLoS Comput Biol 2:e85

$$\mathcal{V}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] = \varepsilon\left[\left(\frac{R_{min}}{r}\right)^{12} - 2\left(\frac{R_{min}}{r}\right)^{6}\right]$$

Simplified coarse-grained models

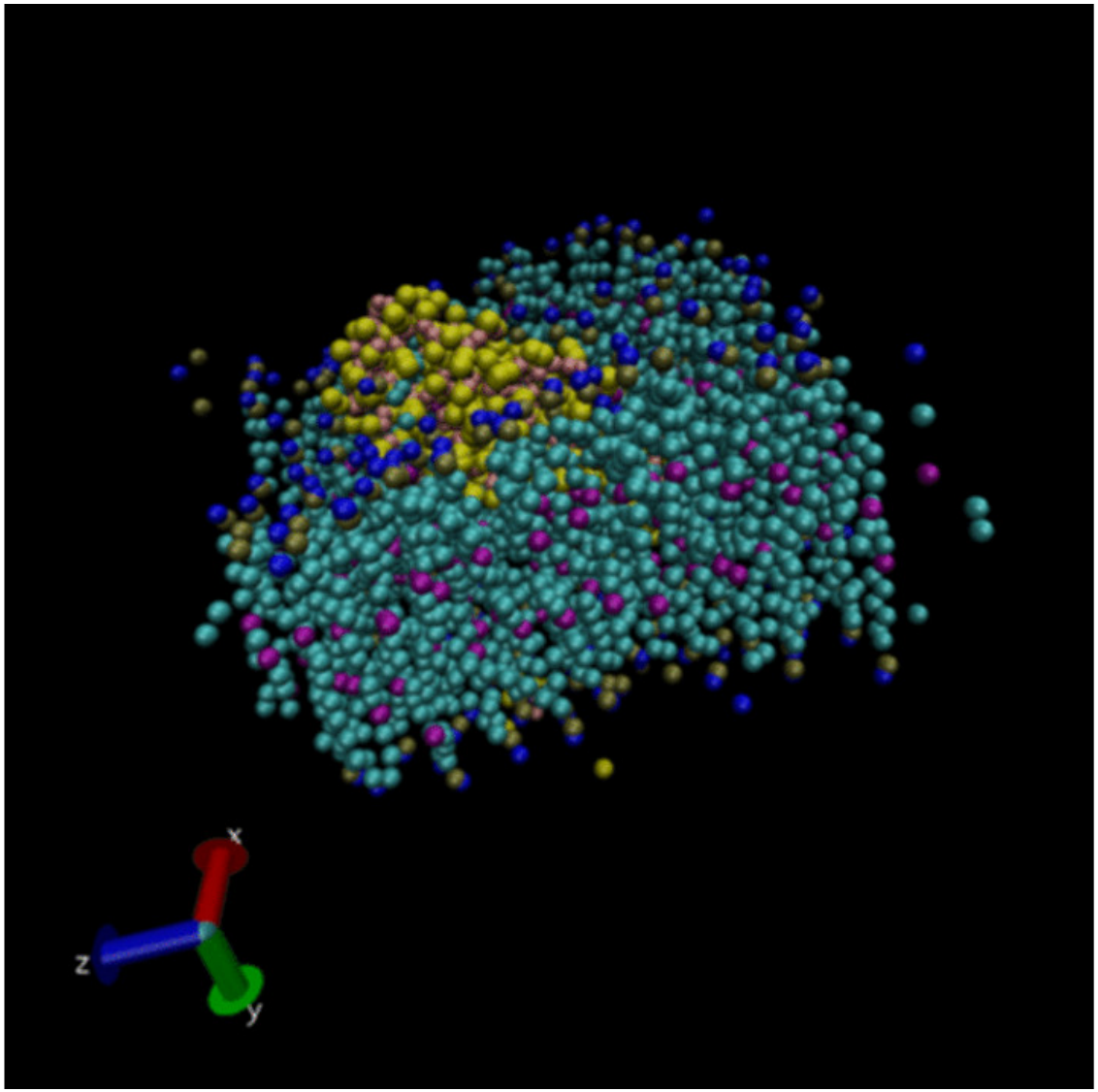




AT: atomistic, all-atom
CG: coarse grained

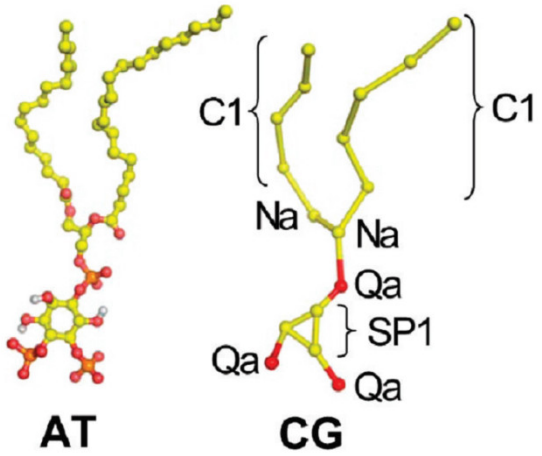
e.g. 2 bead or 4+ bead models for proteins
e.g. MARTINI CG force field

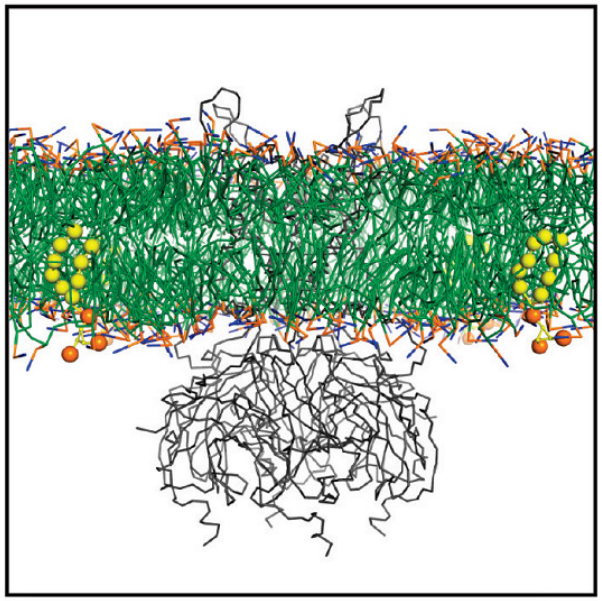
Membrane bilayer formation

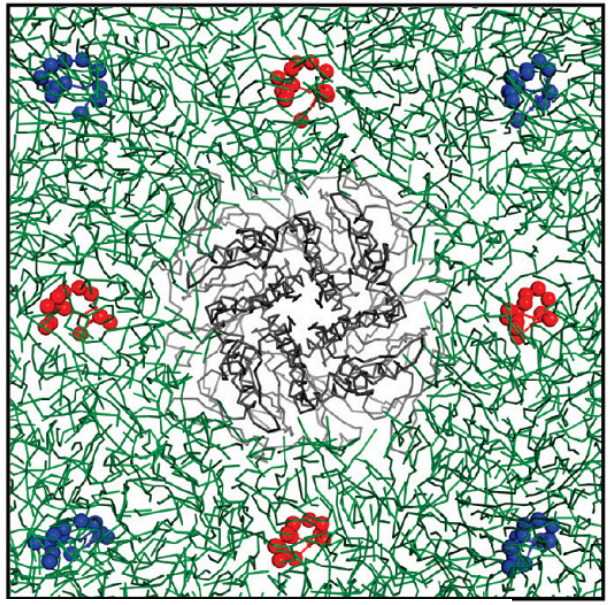


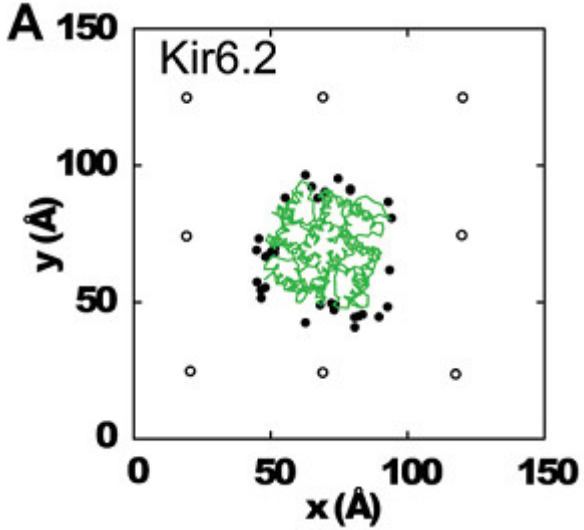
Binding of PIP2 to a Kir channel

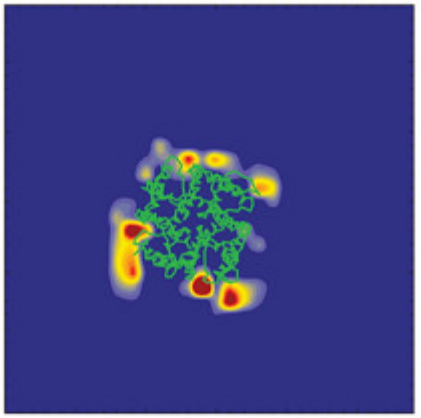
Biochemistry, Vol. 48, No. 46, 2009



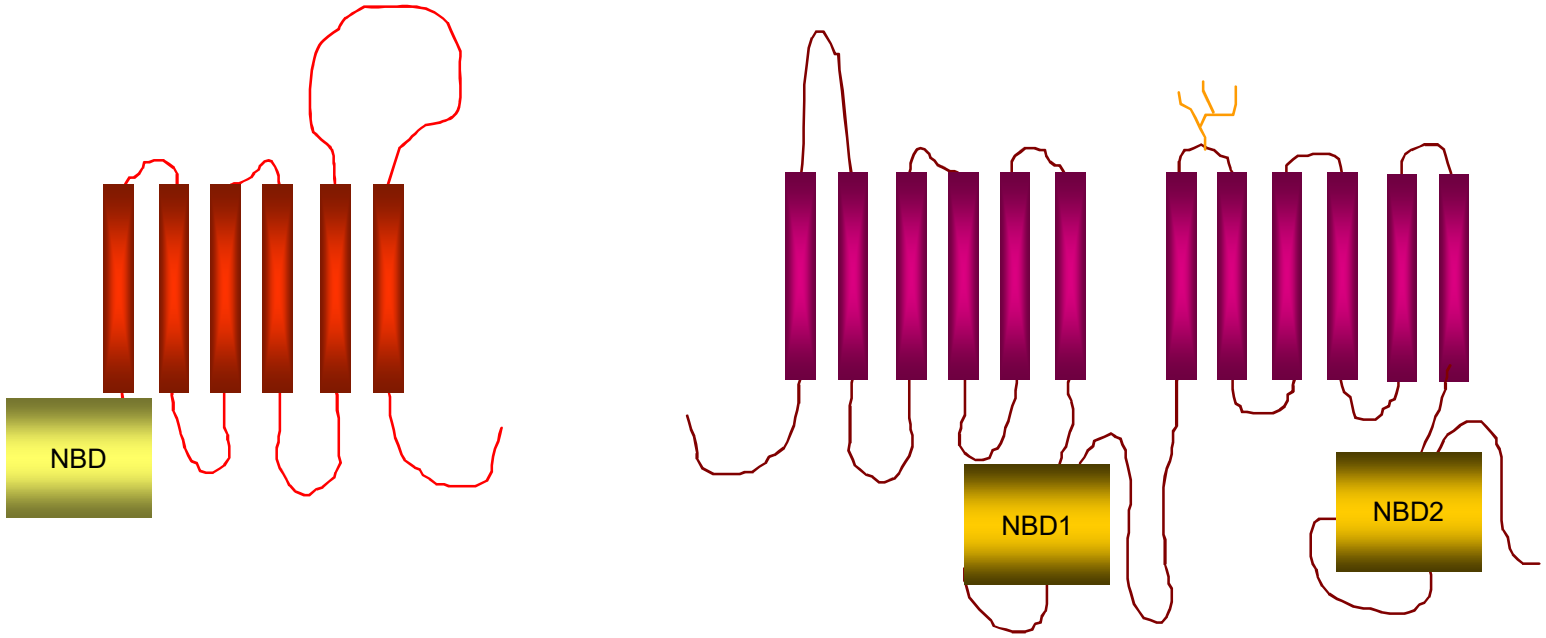




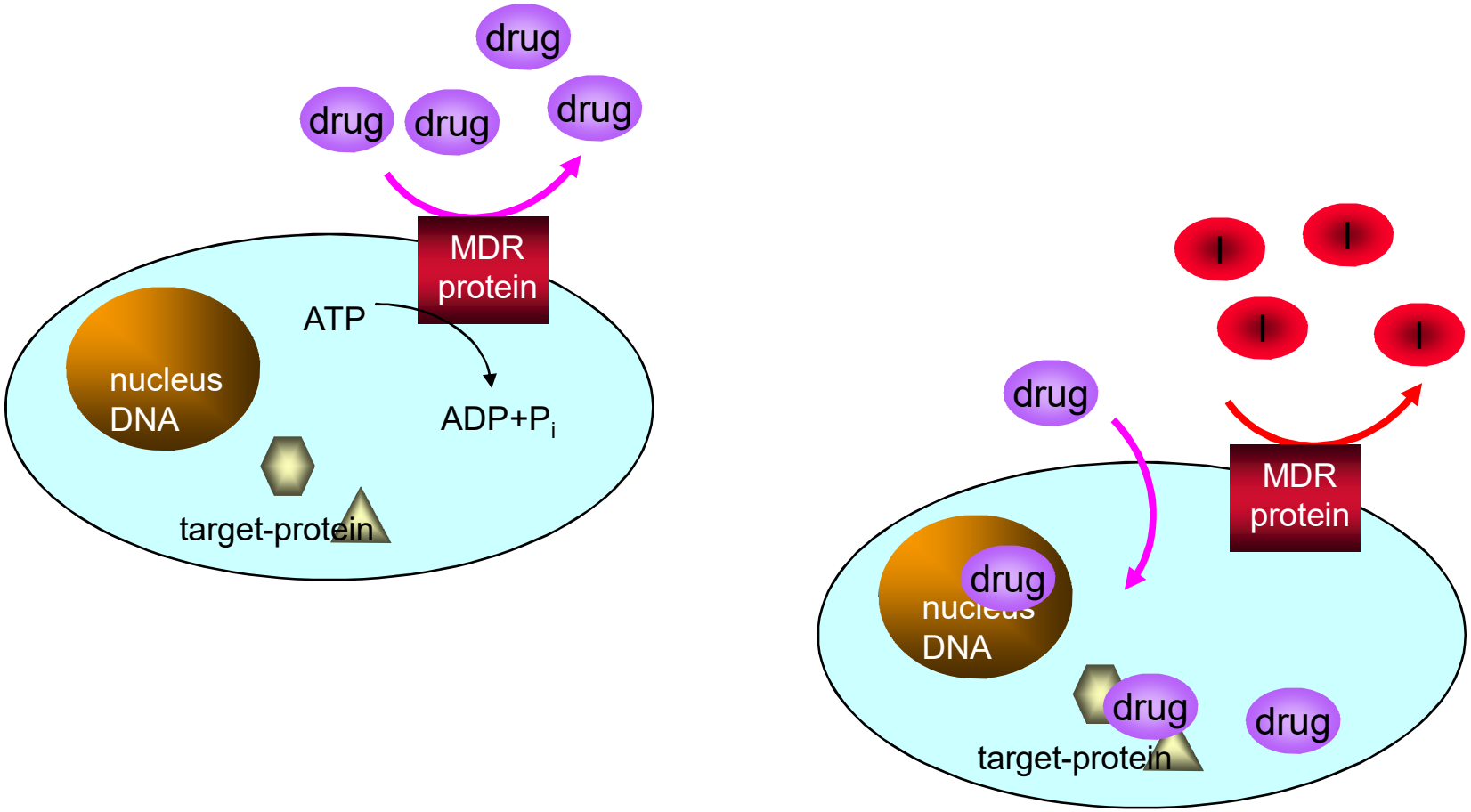




ATP Binding Cassette (ABC) proteins

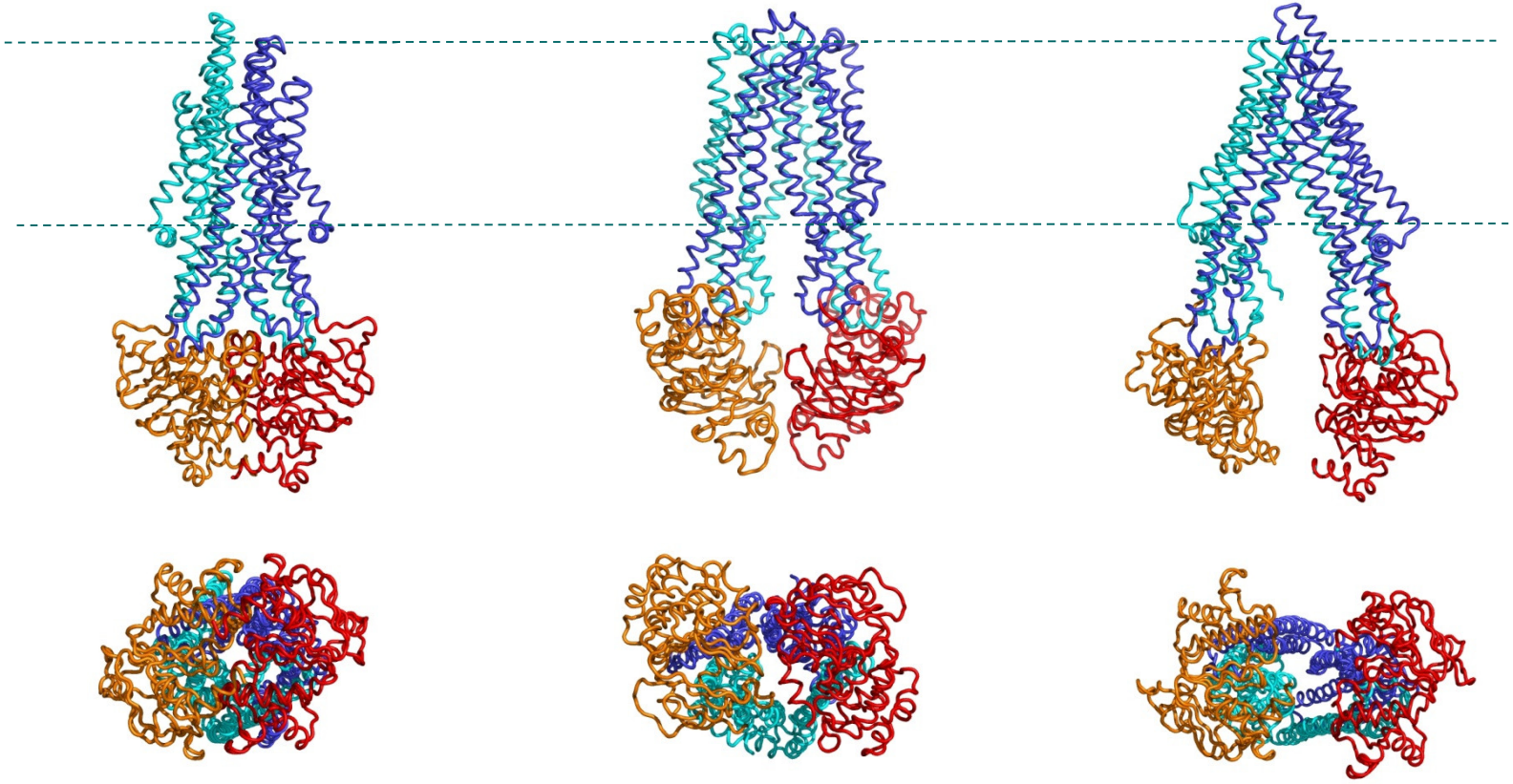


Multidrug resistance



I: inhibitor

Conformation of ABC proteins



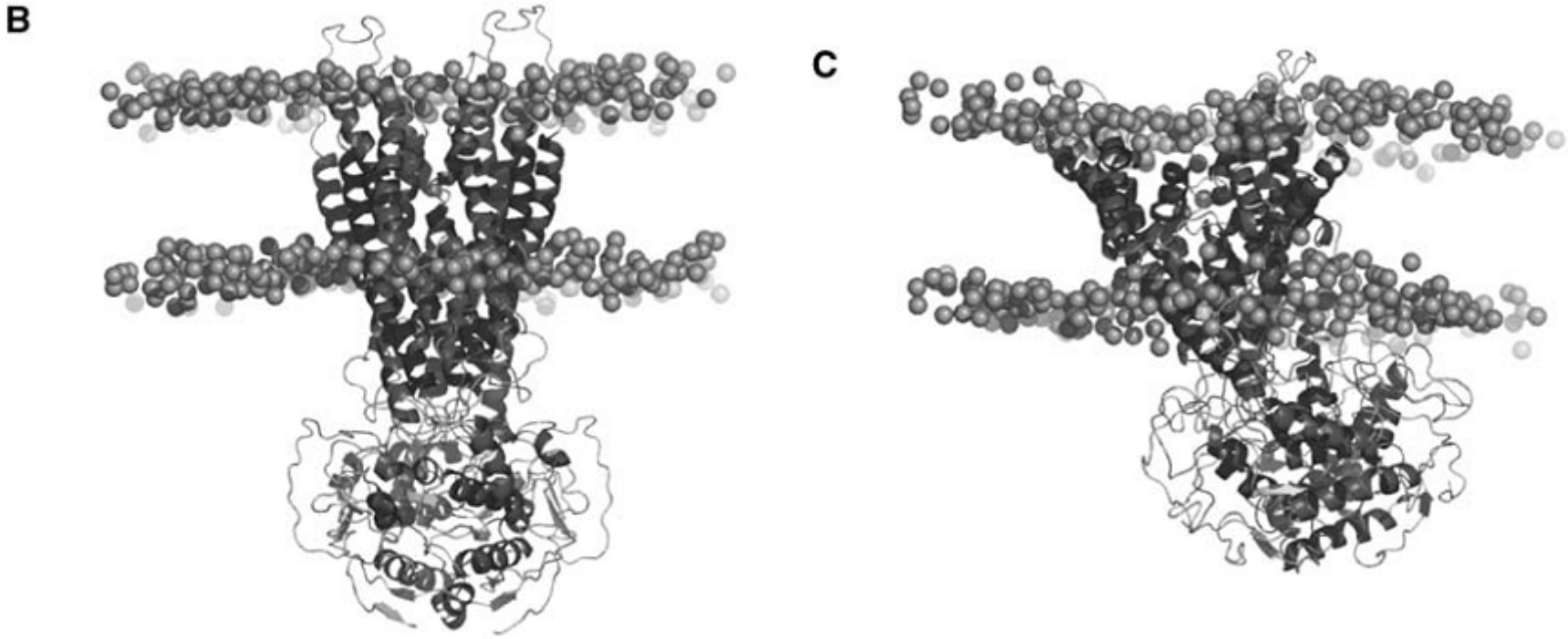
“bottom-closed” (+ATP)

“bottom-closed” apo (-ATP)

“bottom-open” apo (-ATP)

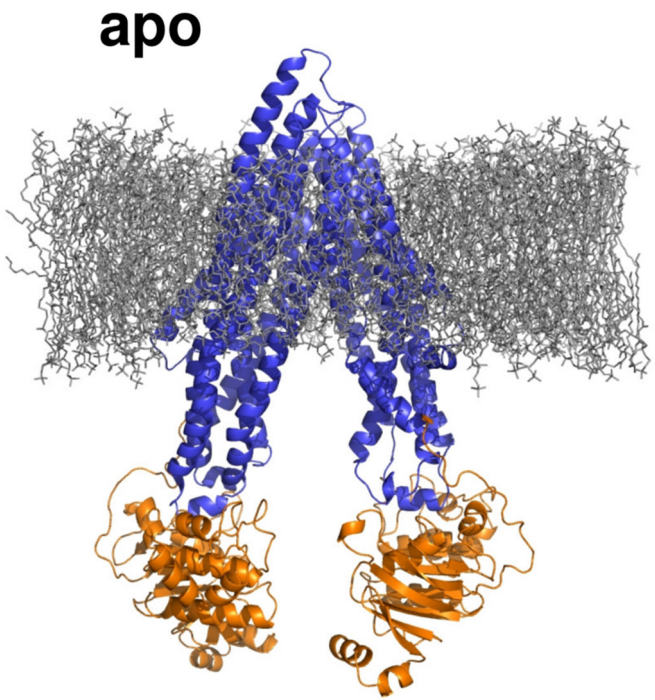
Stability of simulations

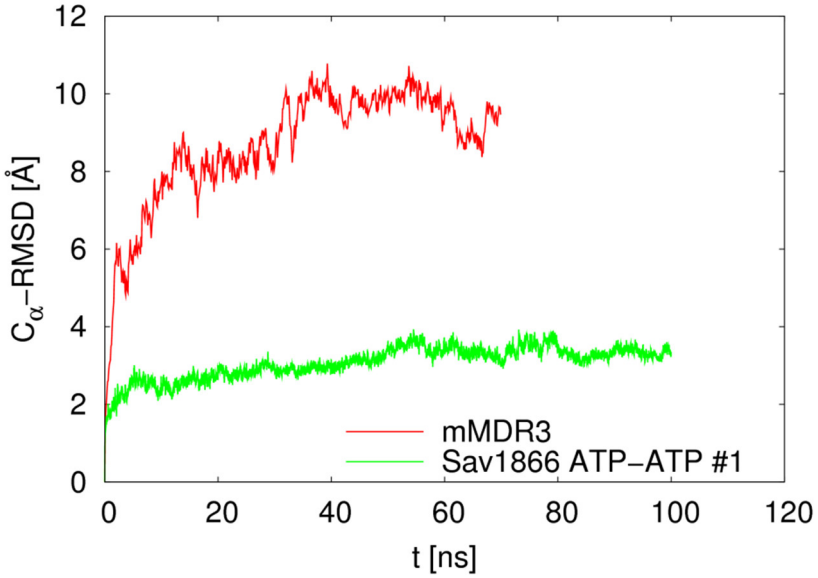
Eur Biophys J (2008) 37:403–409



0 ns **20 ns**

The bottom-open apo conformation is unstable

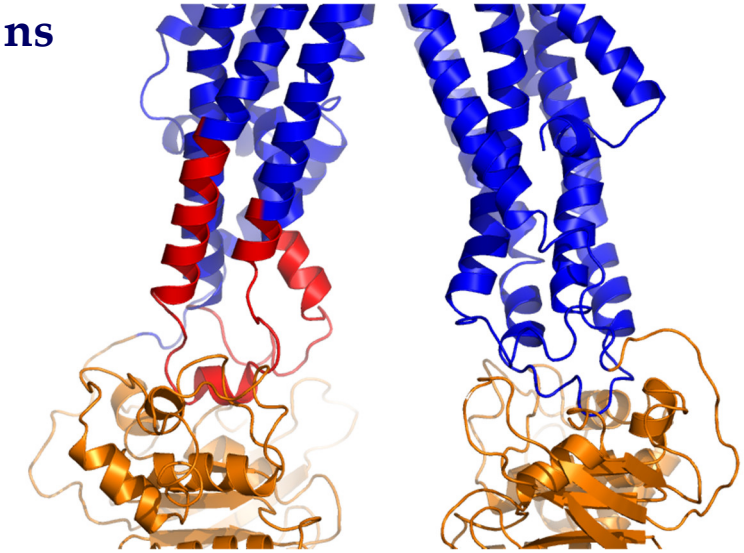




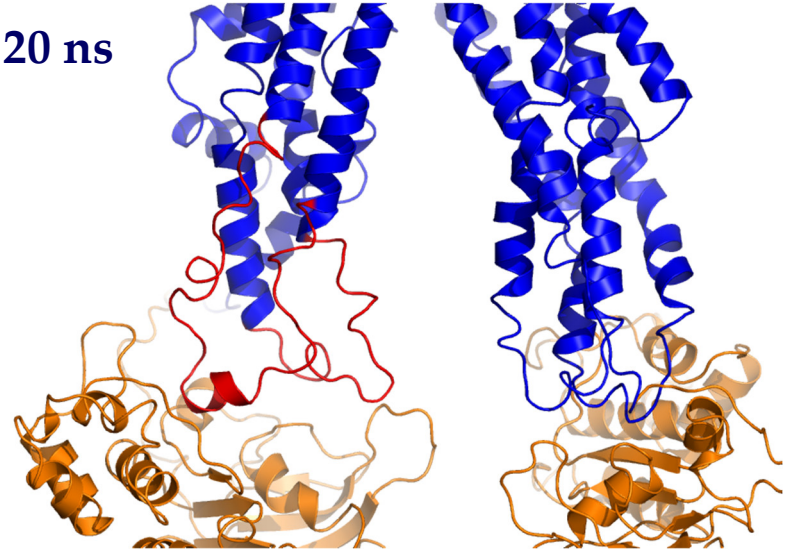
Gyimesi *et al.* BBA 2012

The bottom-open apo conformation is unstable

t = 0 ns

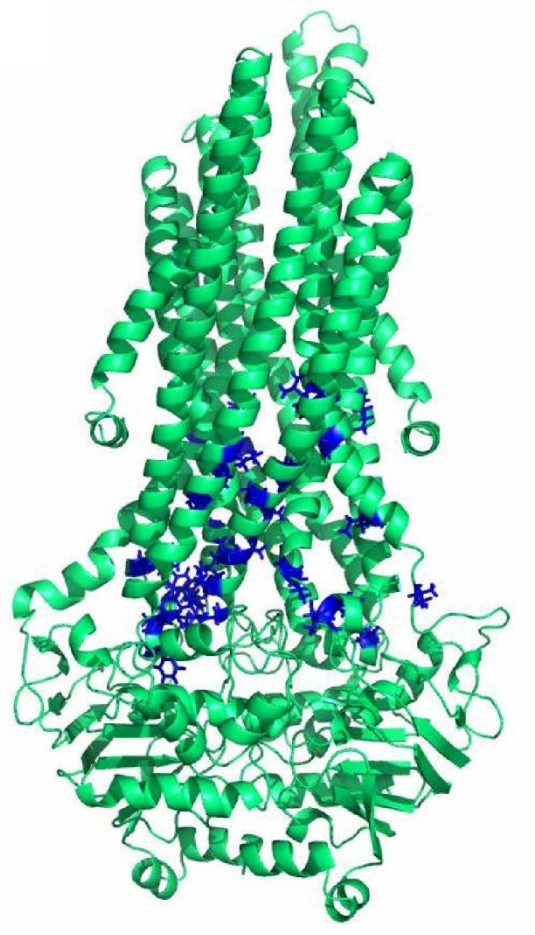


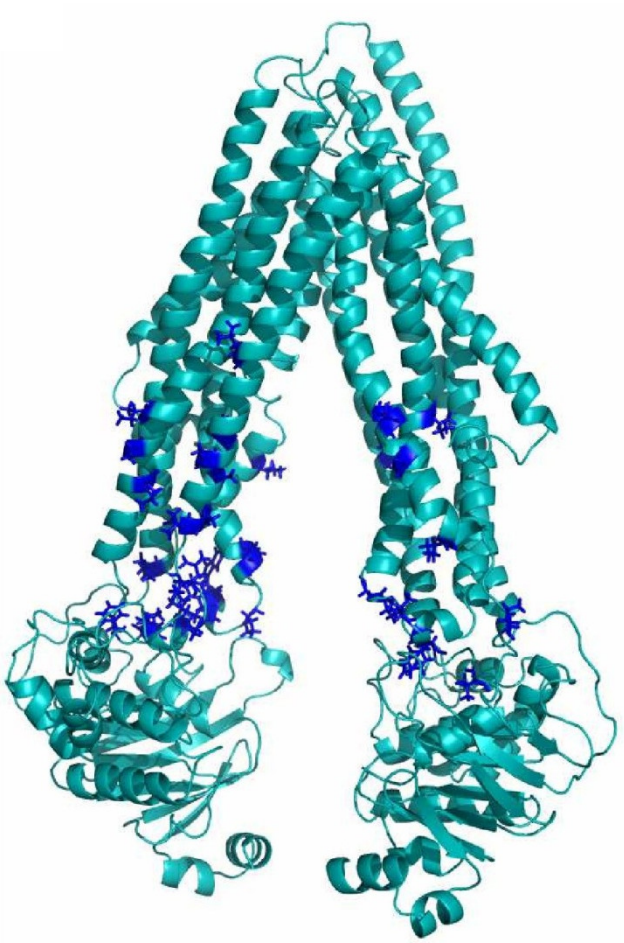
t = 20 ns



system	helical content
Sav1866 ATP/ATP #1	90.04%
hMDR1 holo	91.84%
hMDR1 apo	64.30%
mMDR3	63.13%

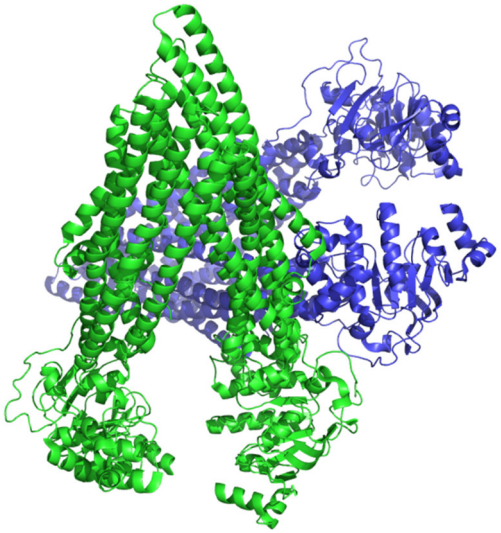
Hydrophobic amino acids are surface exposed



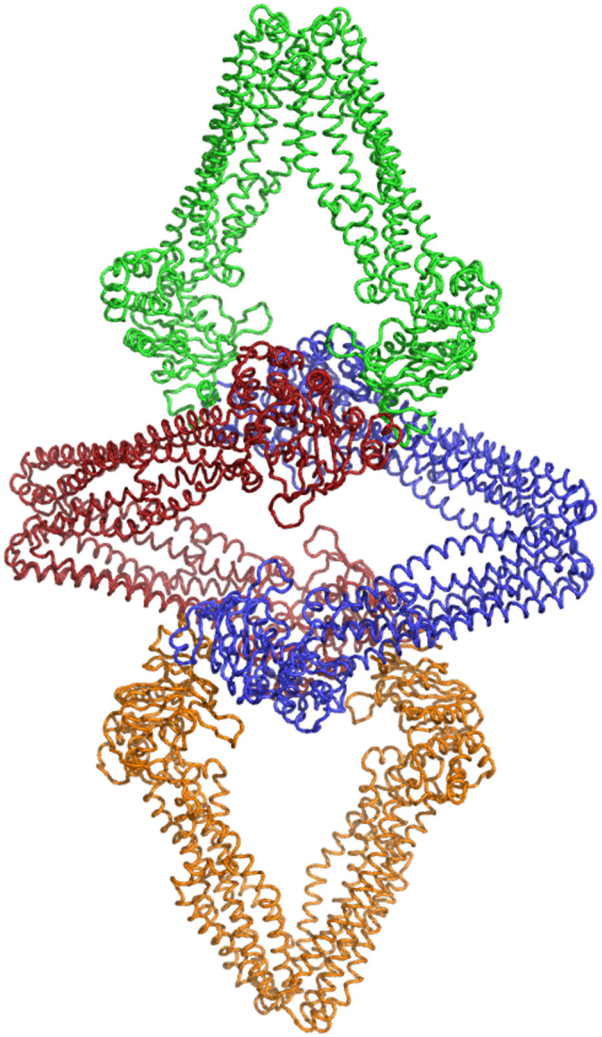


The content of the unit cell

mMDR3, PDBID:3G5U



MsbA, PDBID:3B5W



Simulating special and slow events

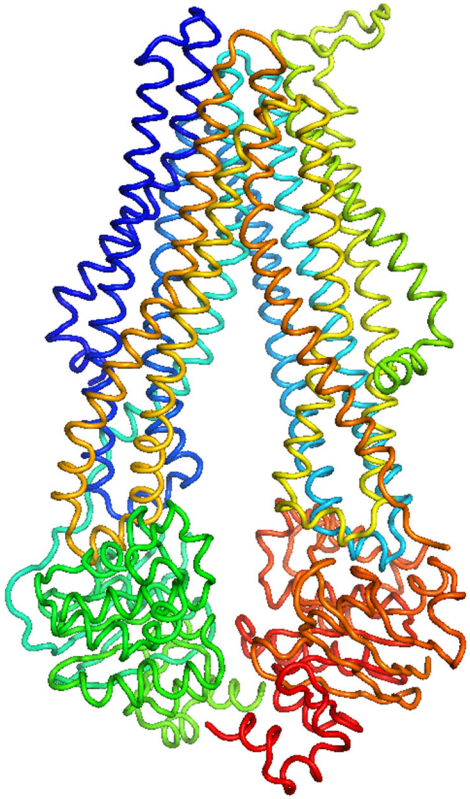
- How does ATP hydrolysis affect the protein dynamics?

  e.g. steered MD

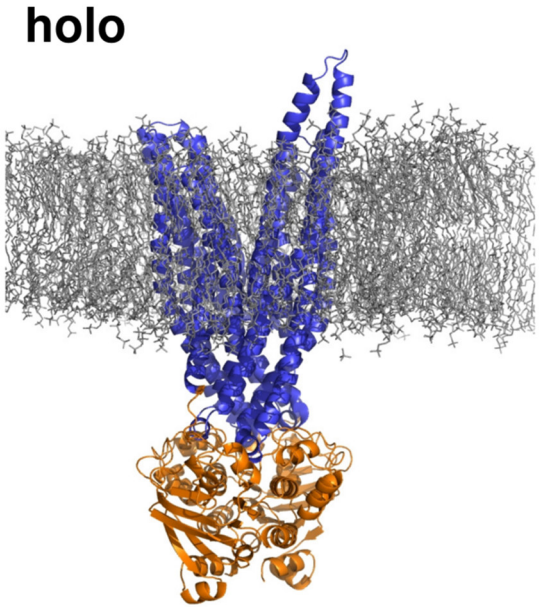
- What is the transition pathway between the bottom-open and bottom-closed conformation?

  e.g. targeted MD, Metadynamics

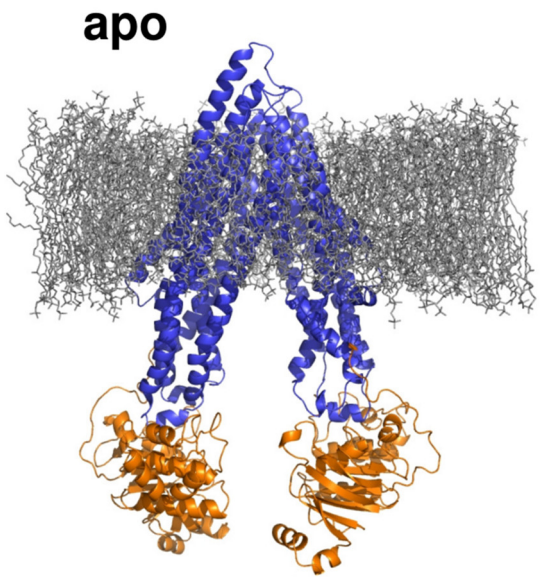
Targeted MD



Describing the transition using MD+ED



hMDR1 homology model
(3x100 ns)



hMDR1 homology model

MD trajectories

⇩

Essential Dynamics

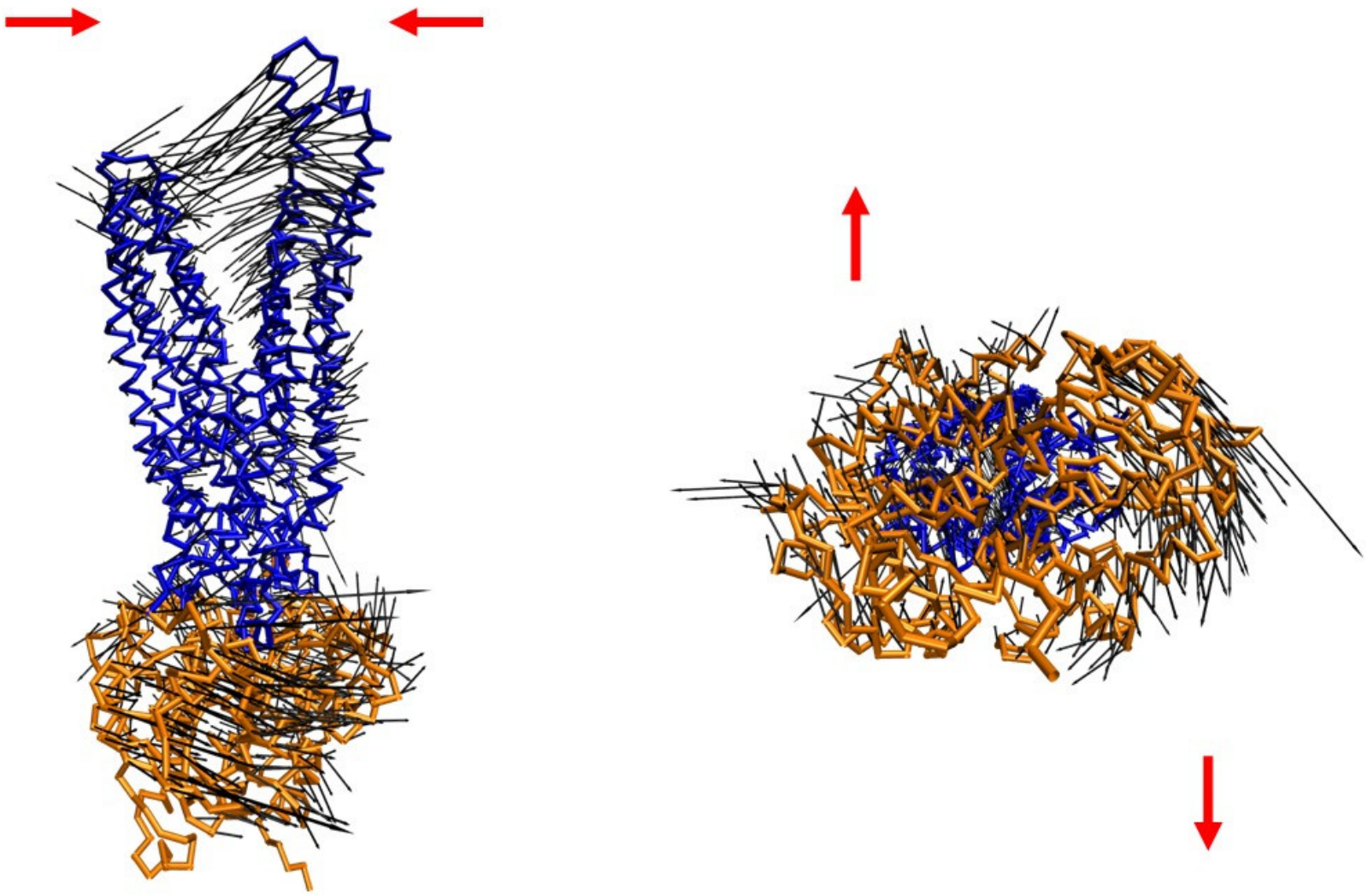
⇩

Selection of modes

⇩

Collective motions

Describing the transition using MD+ED

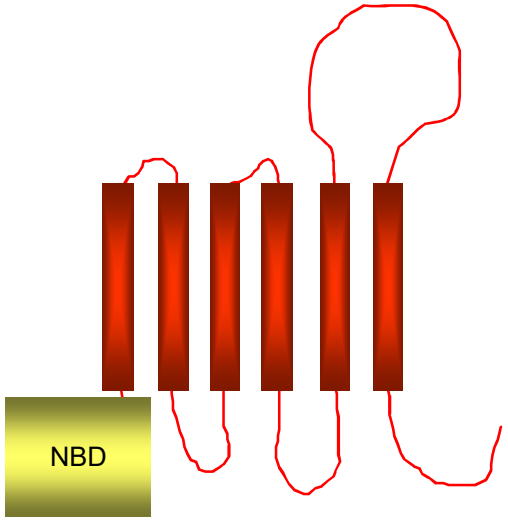


Calculating the correlation of motions

- Pearson correlation
- MI (mutual information)
- DiCC (distance correlation coefficient)

The complex example of ABCG2

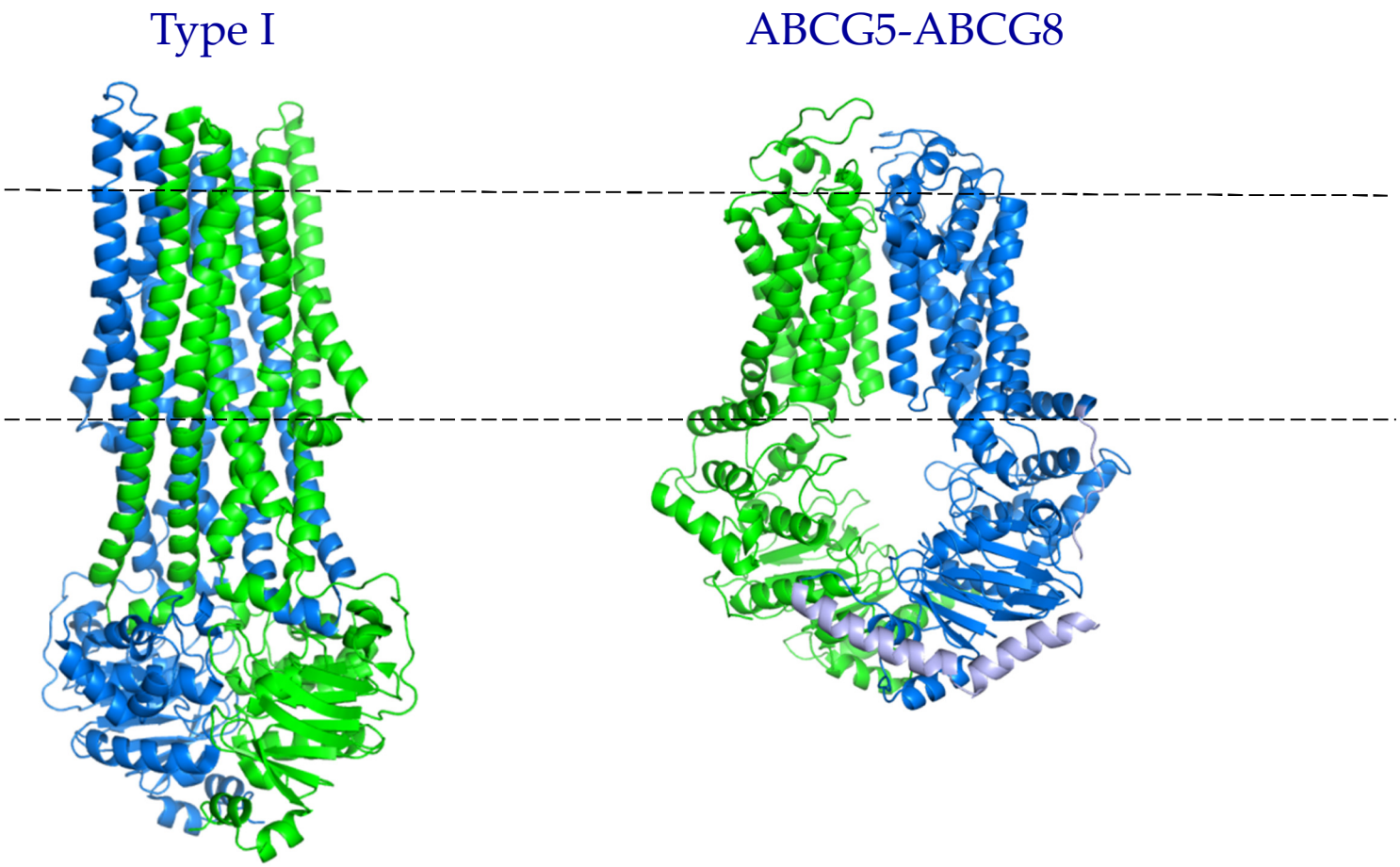
- Homology modelling of the structure (2016)
- Investigating the effect of mutations using MD
- Effect of cholesterol on function
- Identification of drug binding sites
- Describing the transport process by MD and METAD



Importance of ABCG2

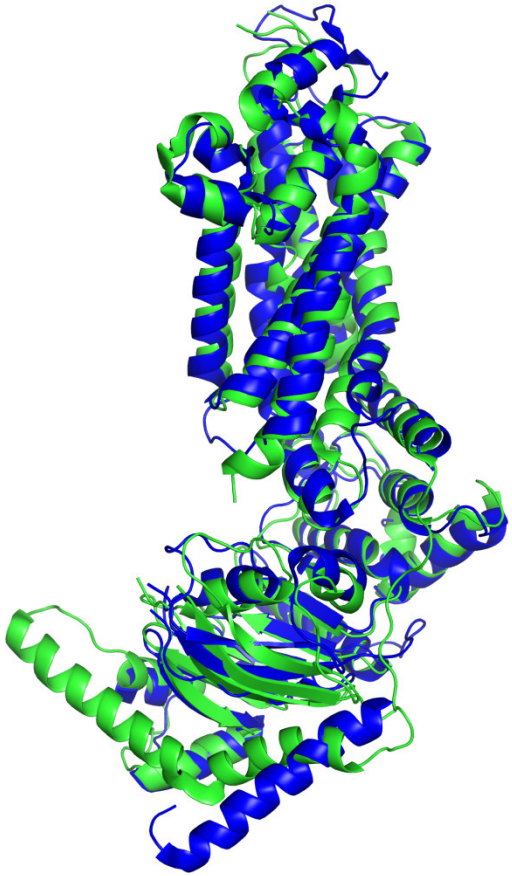
- Expressed in stem cells, tissue barriers, cancer cells
- Multidrug transporter of xenobiotics and endobiotics
  - antitumor agents
  - uric acid
- The Q141K variant exhibit decreased function and expression

The type II ABC exporter fold



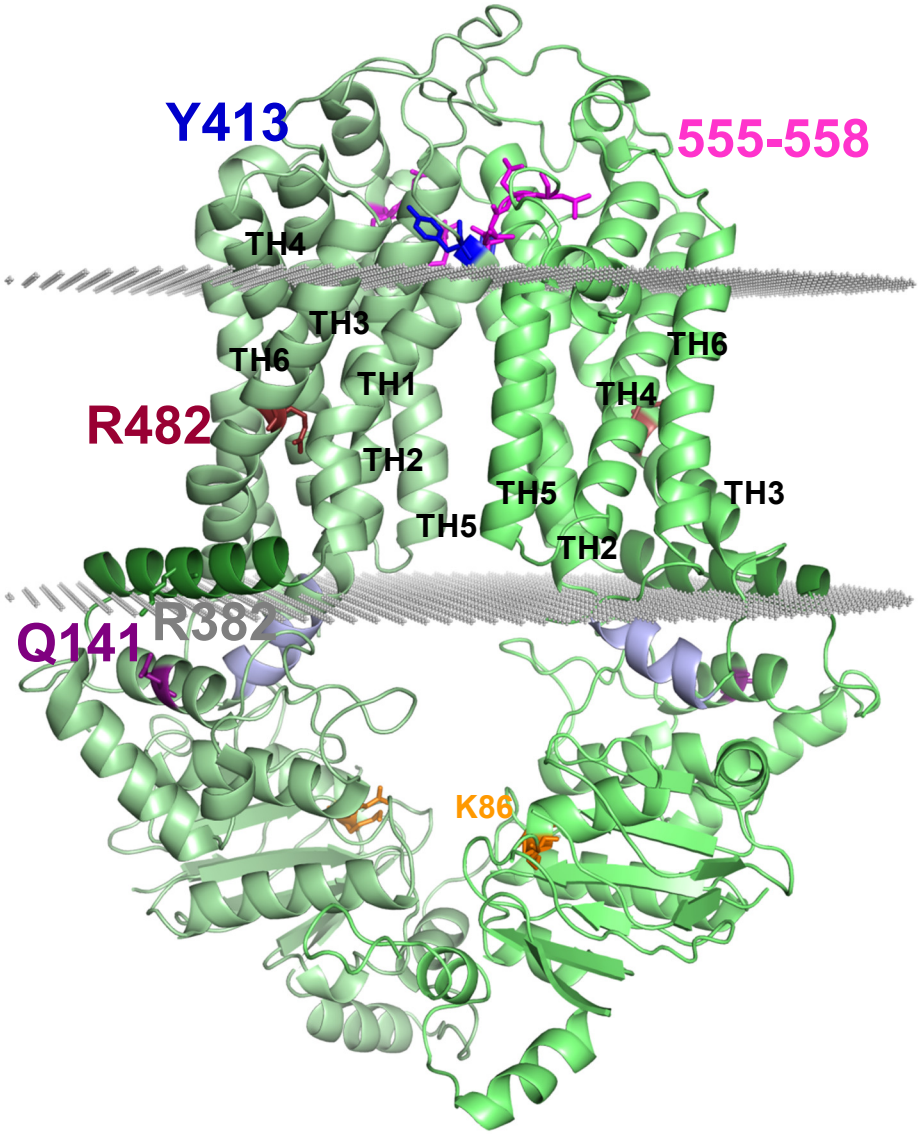
Homology modelling of ABCG2 based on ABCG5/G8

- Approx. 25% identity and 45% similarity
- Generation of a sequence alignment was ~trivial
- 100 models were built using Modeller
- The model with the best DOPE score was selected and used



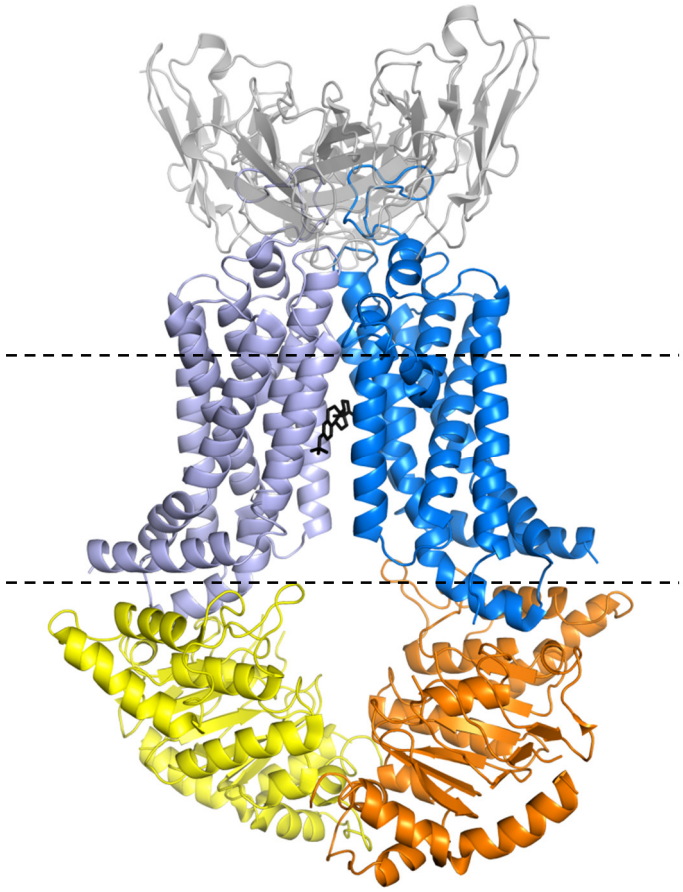
ABCG5-ABCG8
PDBID:5DO7

The ABCG2 model

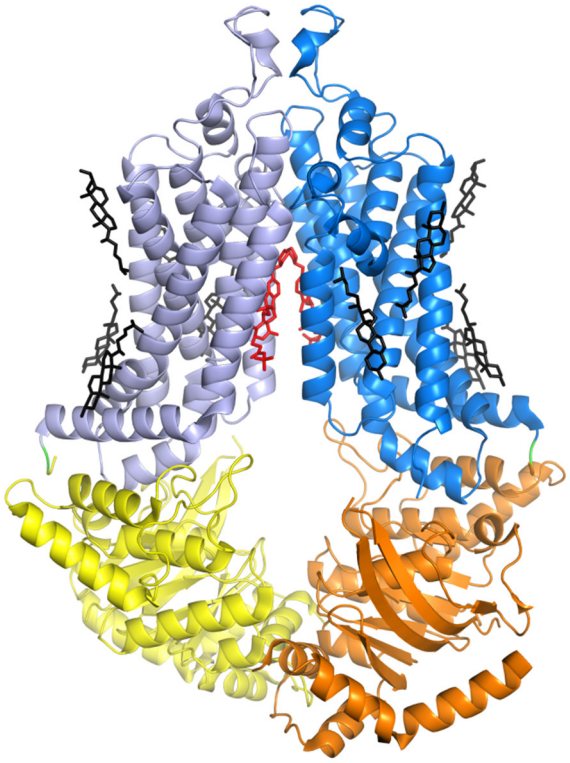


ABCG2 structures

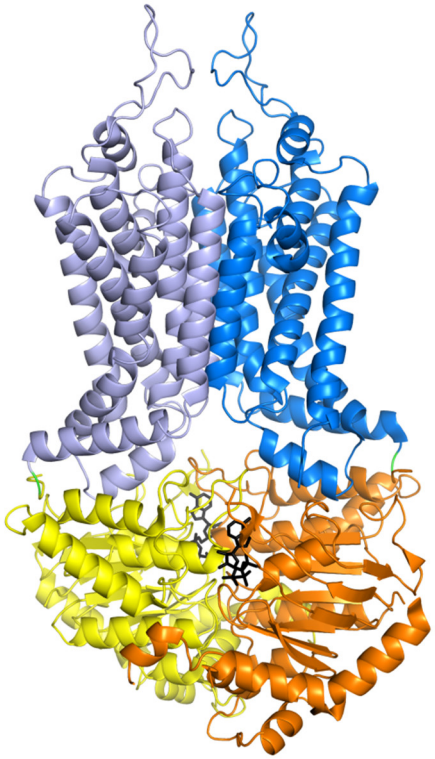
K. Locher, ETH, Zurich



6HCO

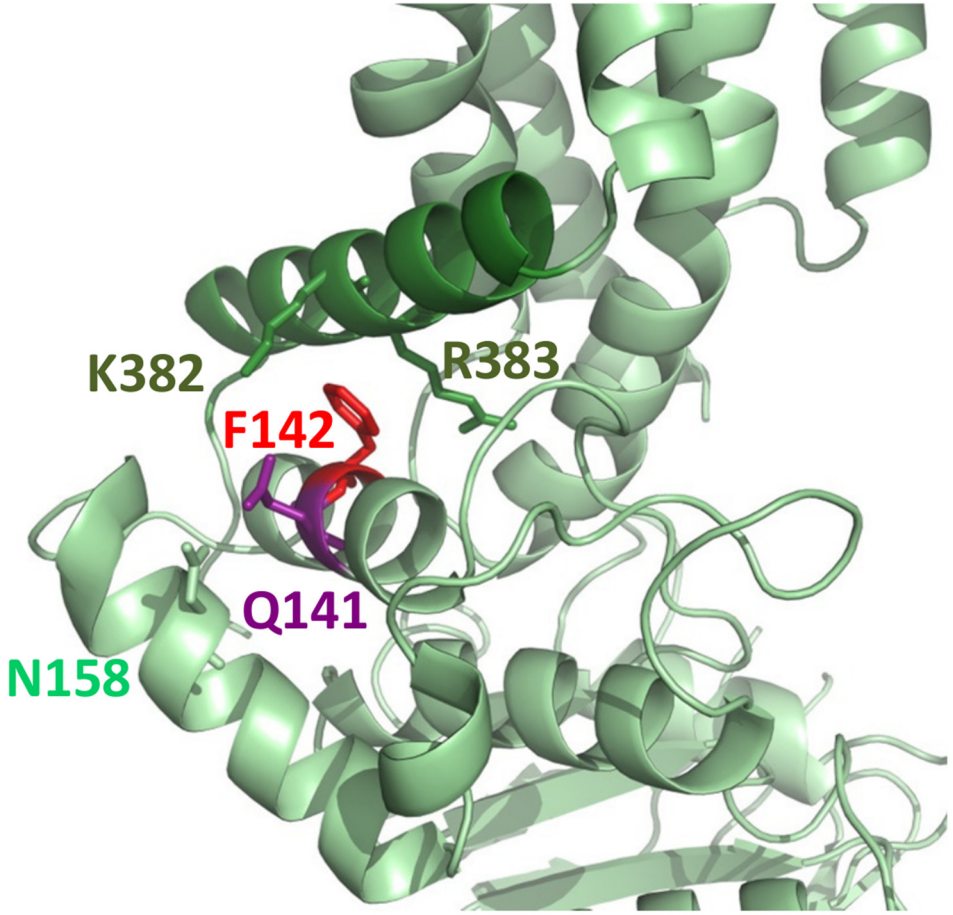


6HIJ



6HZM

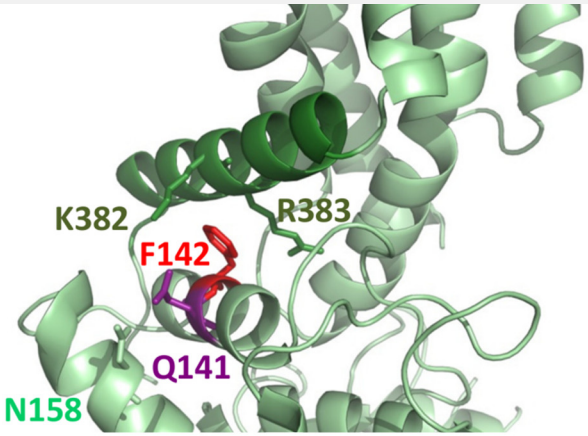
The Q141 position

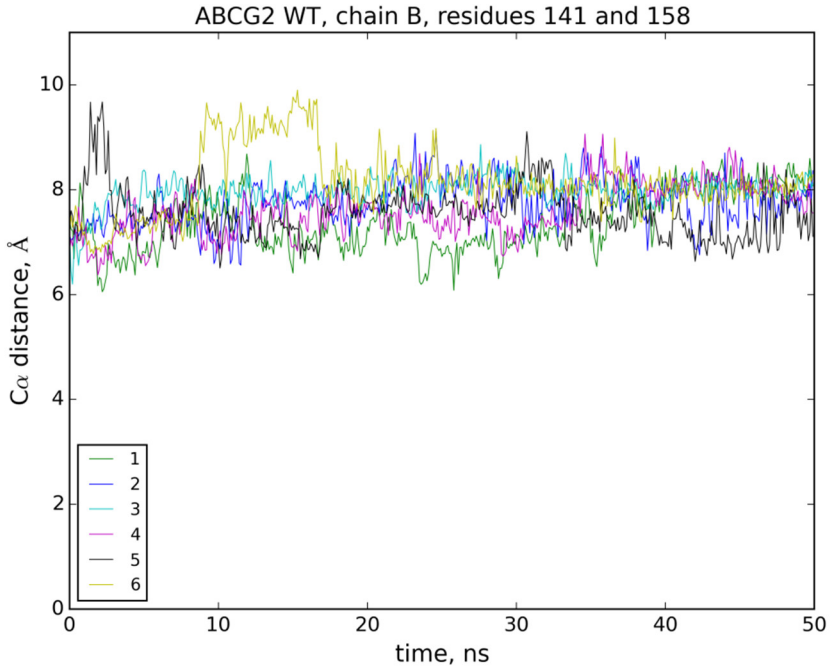


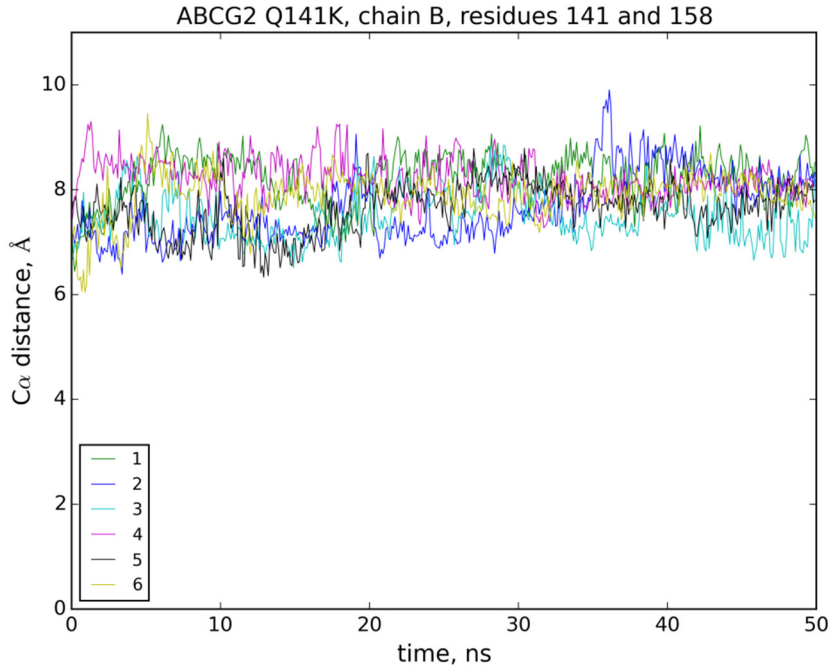
MD simulations

- The protein was embedded in POPC bilayer
- Optimizing the orientation of water, lipids, amino acid side chains:
  - energy minimization
  - equilibration
  - minimal backbone motions (position constrains)
- Production run
  - no constraints
  - 50 ns x 6 = 300 ns
- Comparing WT és mutants (e.g. Q141K, R482G)

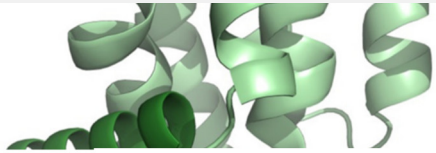
The effect of Q141K on protein dynamics

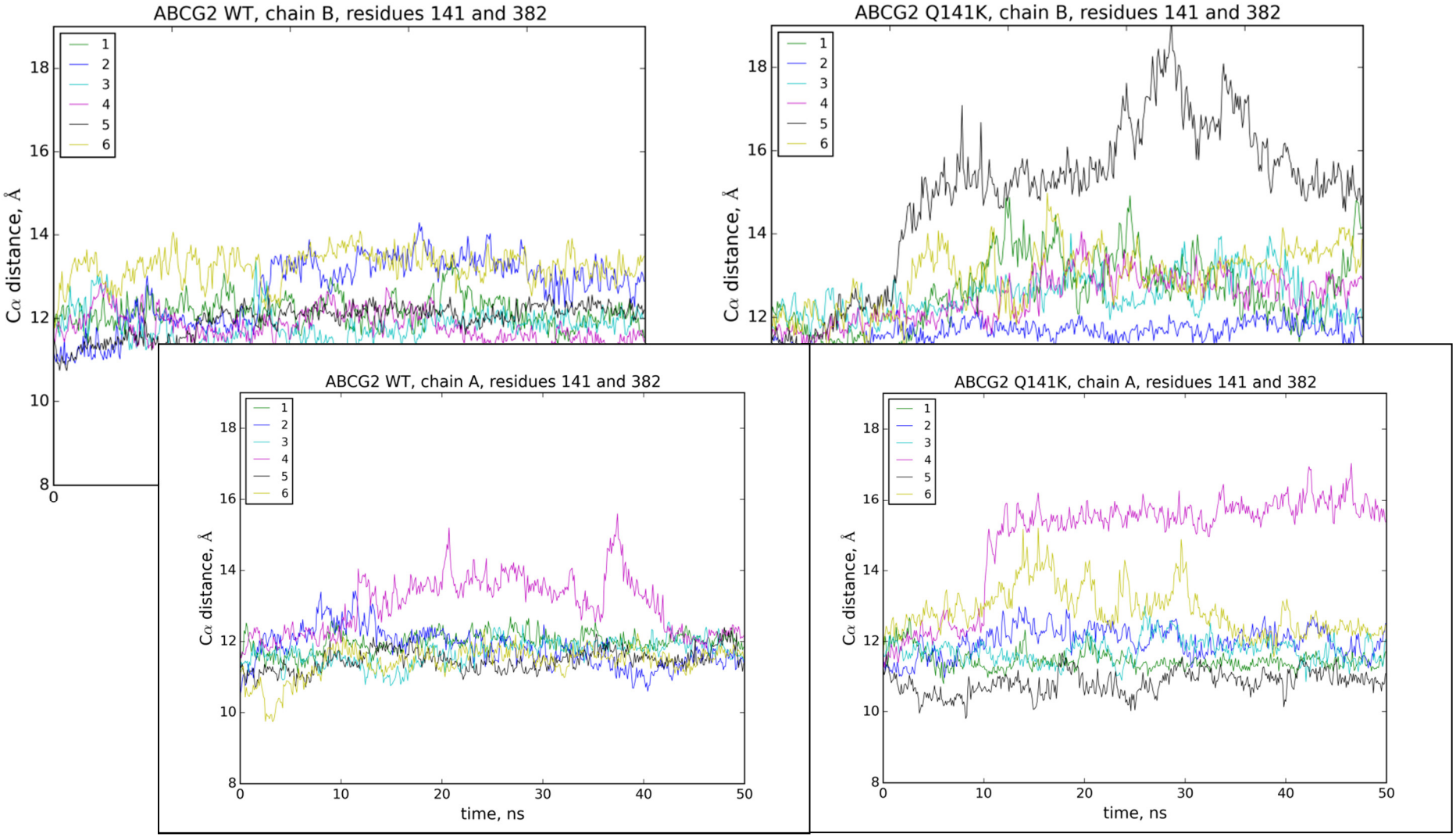






The effect of Q141K on protein dynamics

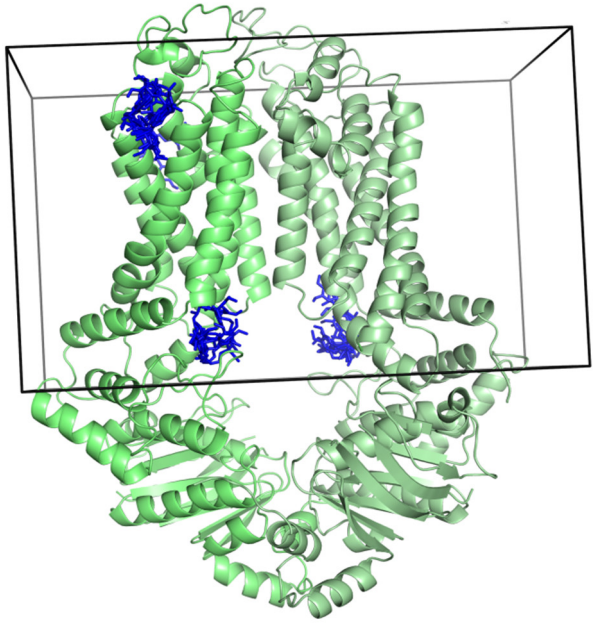




Identification of drug binding sites

in silico docking, AutoDock Vina

- Flexible ligand, non-flexible protein
- Several conformations from simulations
- Search space defined by a box

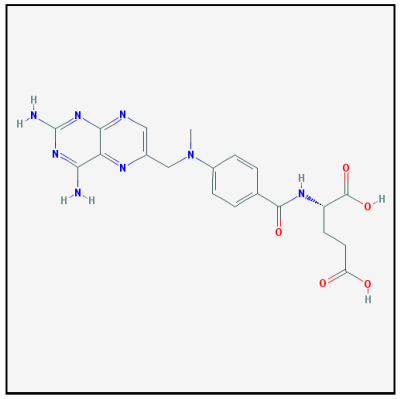


Identification of drug binding sites

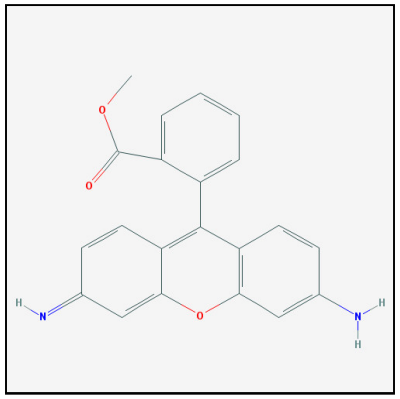
in silico docking, AutoDock Vina

- Flexible ligand, non-flexible protein
- Several conformations from simulations
- Search space defined by a box
- Everything gets docked; docking substrates and non-substrates

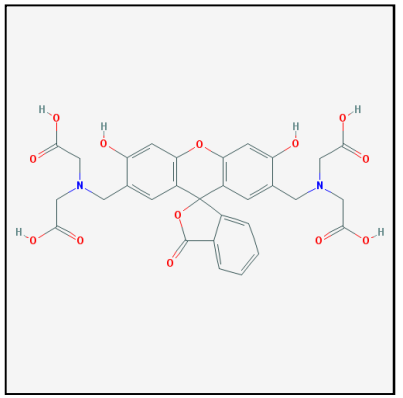
methotrexate



rhodamine123



calcein



Identification of drug binding sites

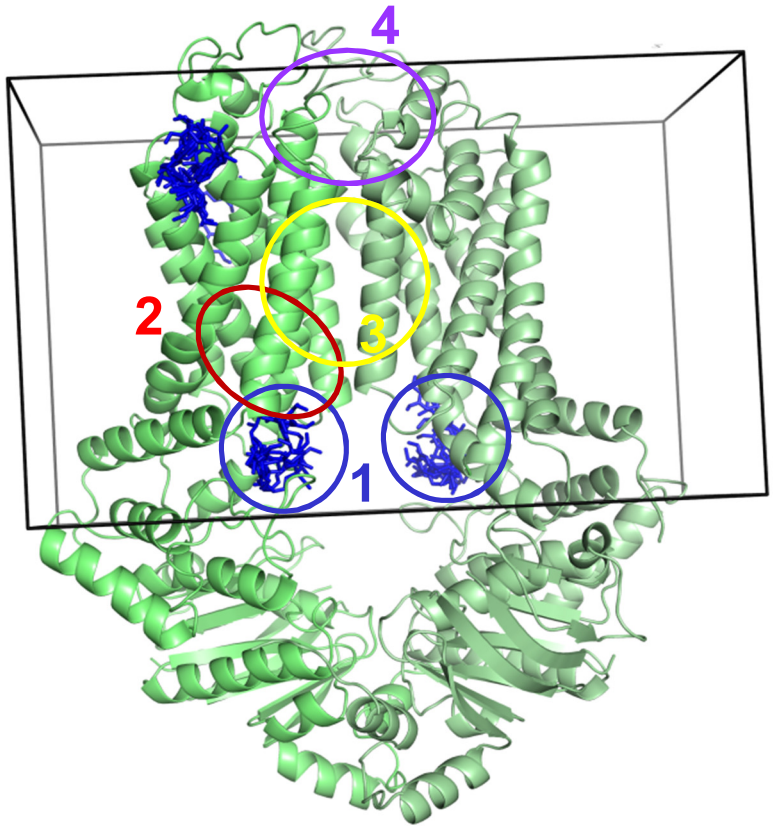
in silico docking, AutoDock Vina

- Flexible ligand, non-flexible protein
- Several conformations from simulations
- Search space defined by a box
- Everything gets docked; docking substrates and non-substrates
- (6 ABCG2 conformations) * (3 parallel dockings) * (20 poses) *
  (25 substrates + 14 non-substrates)
- Clustering poses

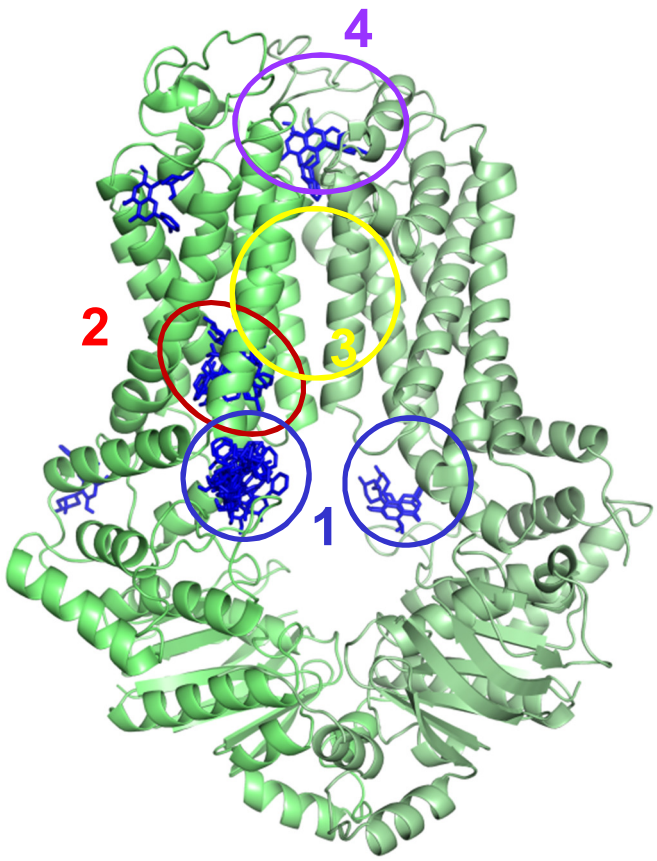
Identification of drug binding sites

in silico docking, AutoDock Vina

verapamil

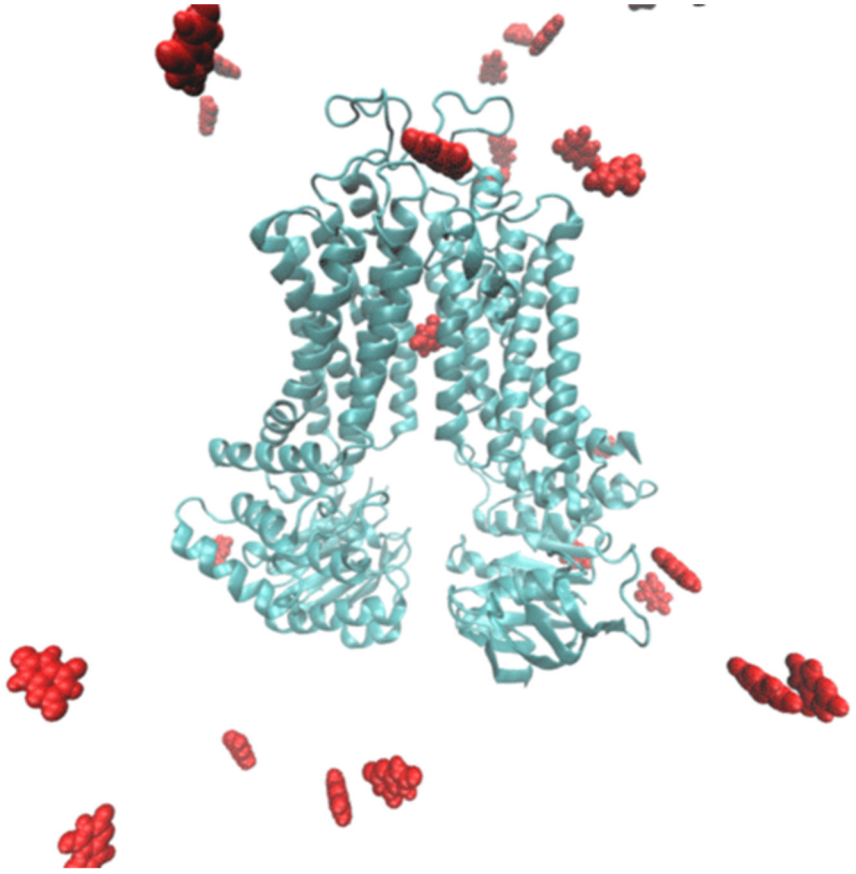


flavopiridol



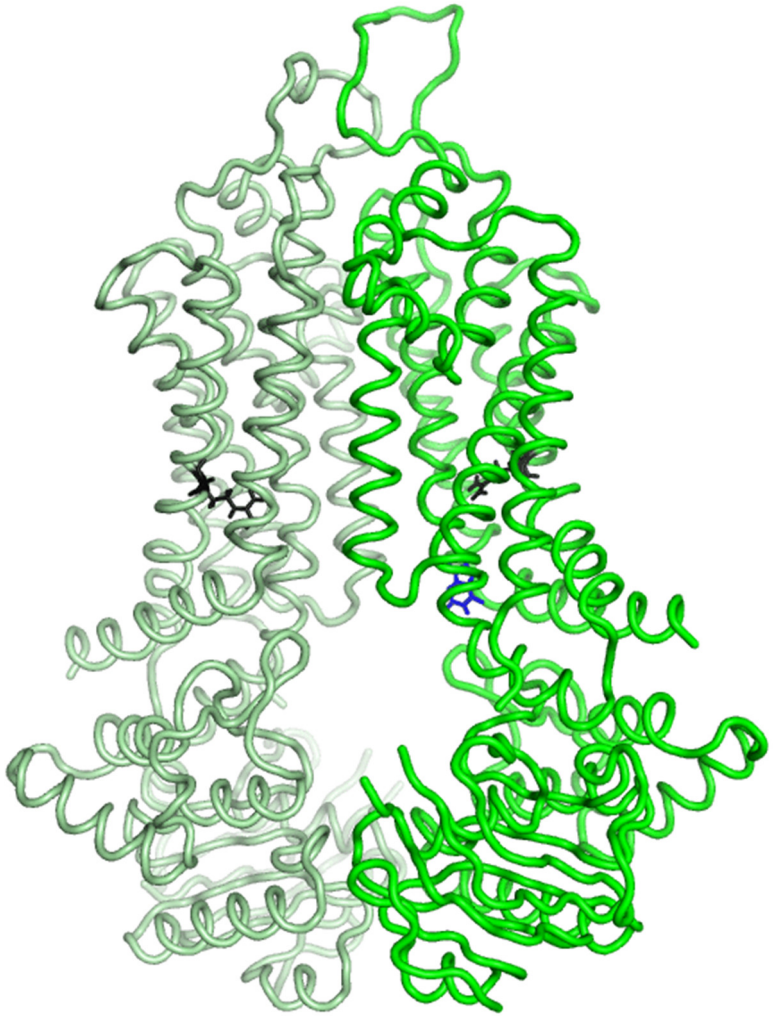
Describing the transport using MD

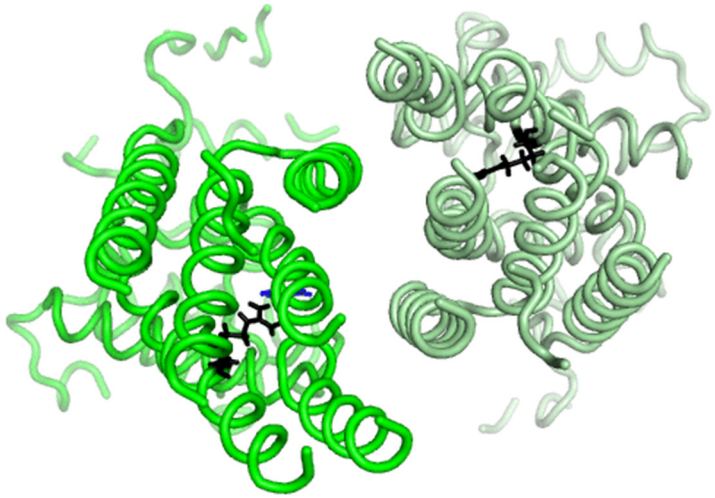
equilibrium simulations, uric acid molecules



Describing the transport using MD

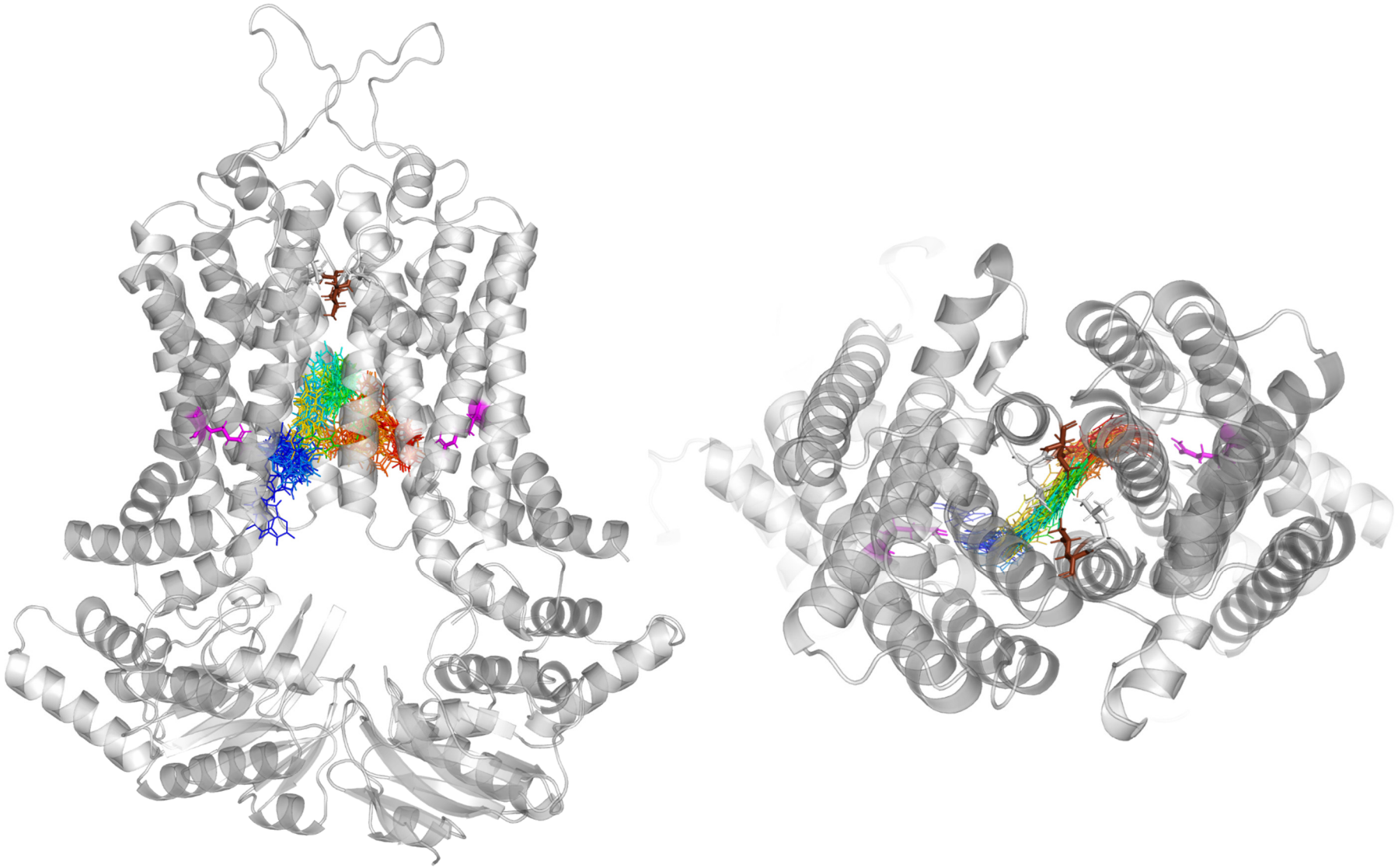
equilibrium simulations, uric acid molecules



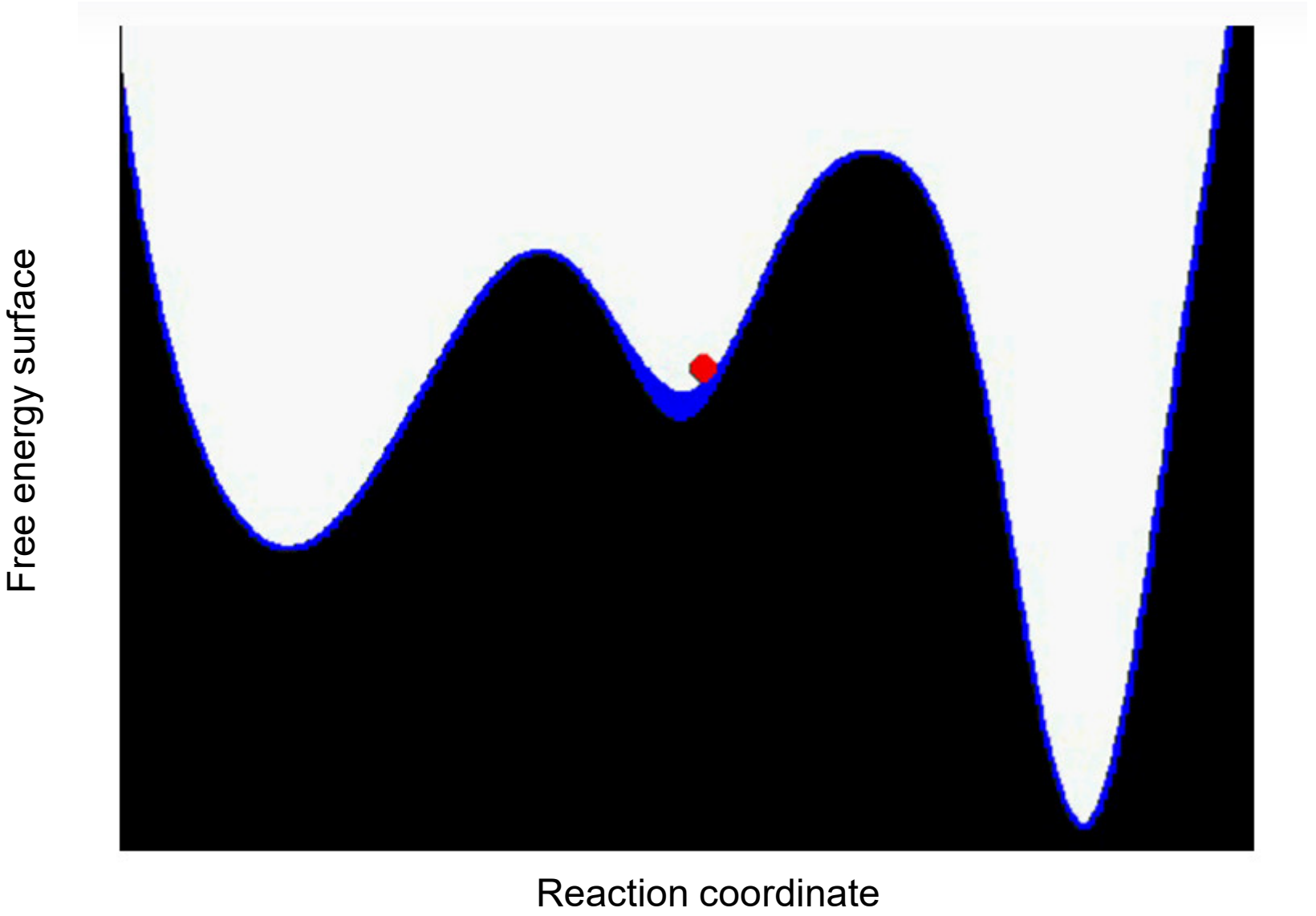


Describing the transport using MD

equilibrium simulations, uric acid molecules

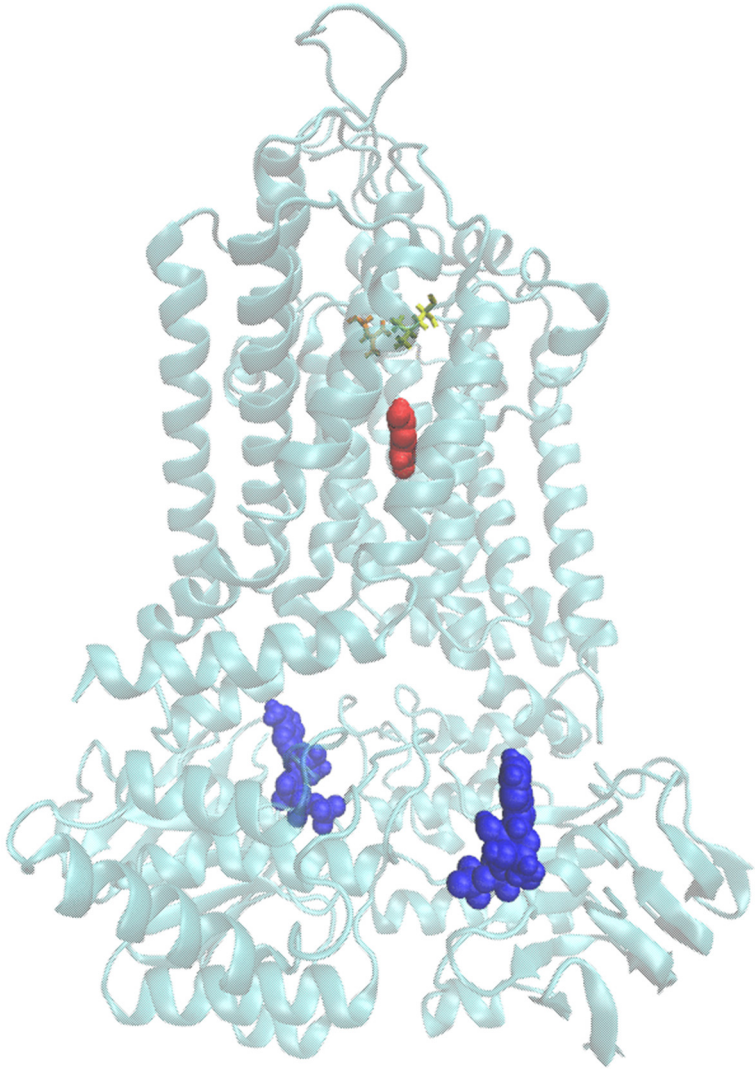


Exploring substrate transport by biased MD simulations



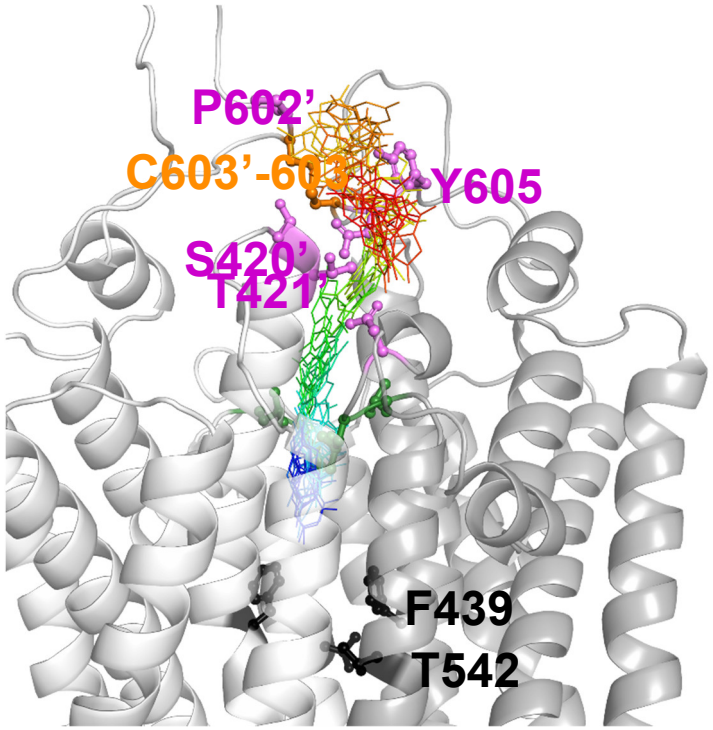
Describing the transport using MD

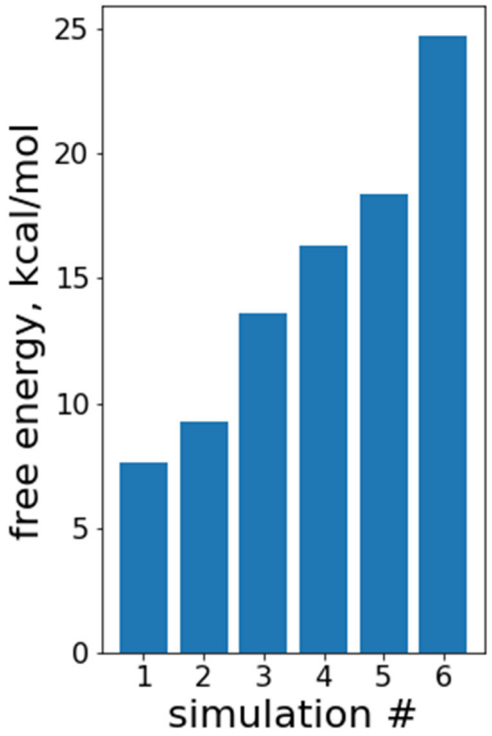
metadynamics simulations, uric acid molecule



Describing the transport using MD

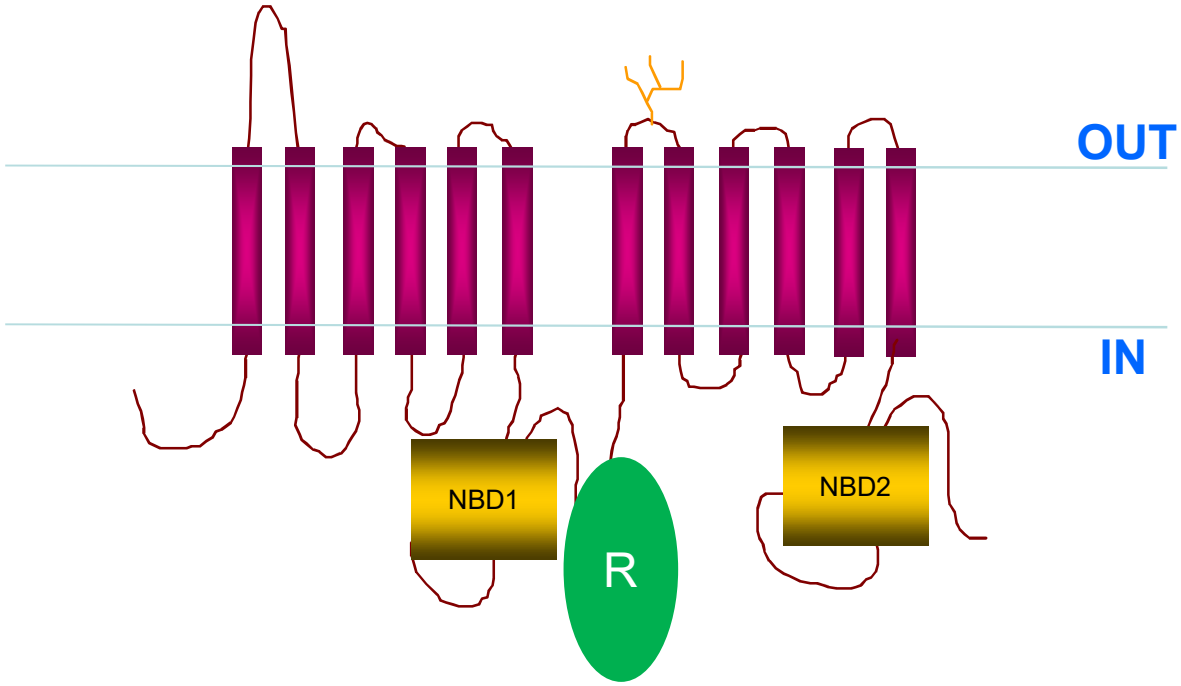
metadynamics simulations, uric acid molecule



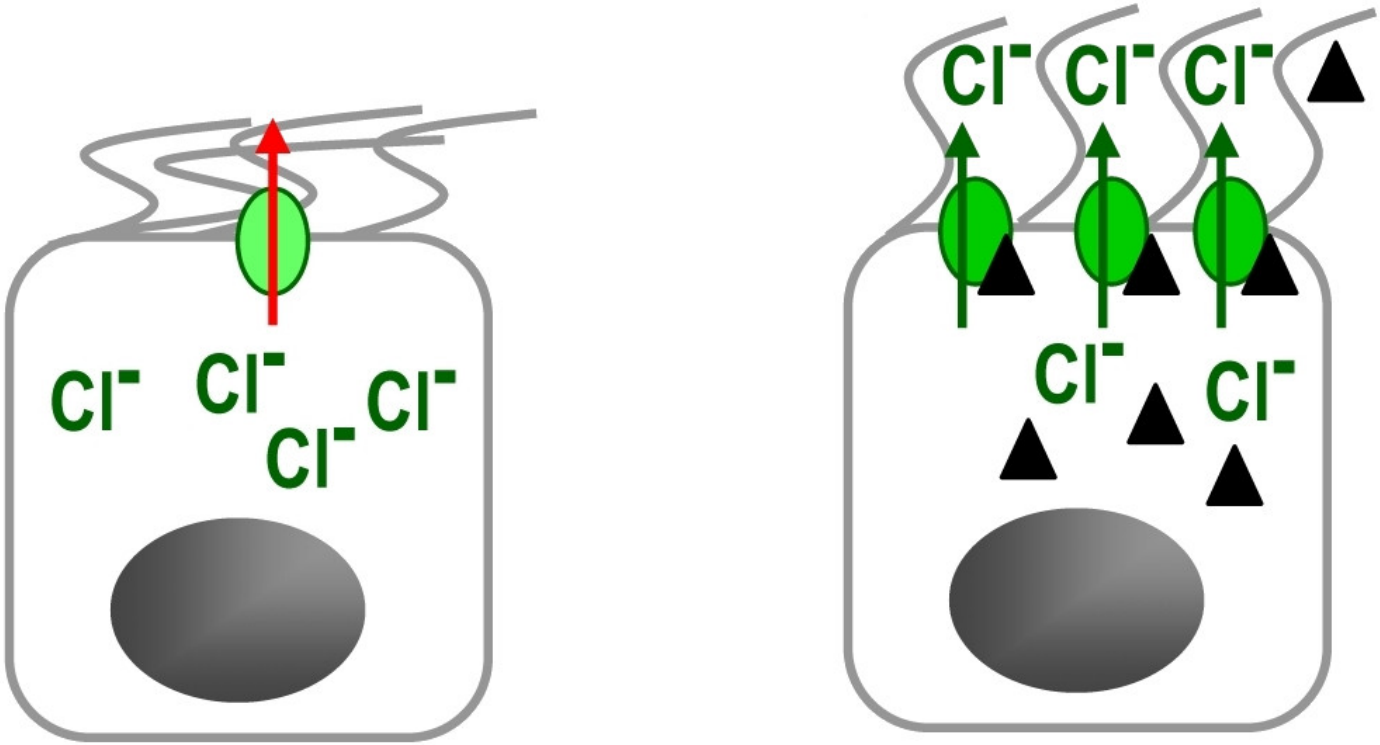


ABCC7/CFTR

Cystic Fibrosis Transmembrane Conductance Regulator

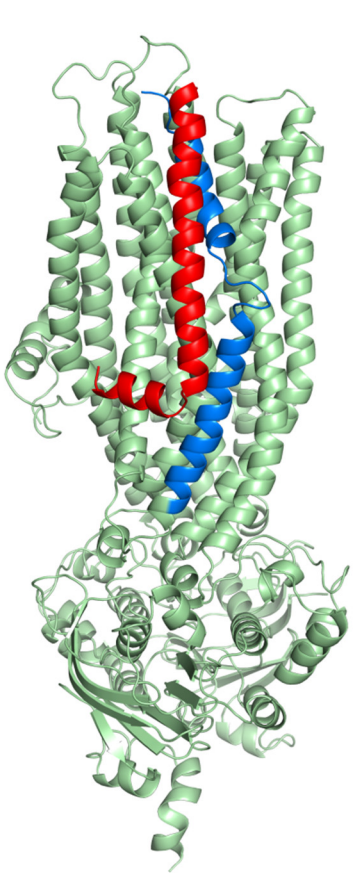


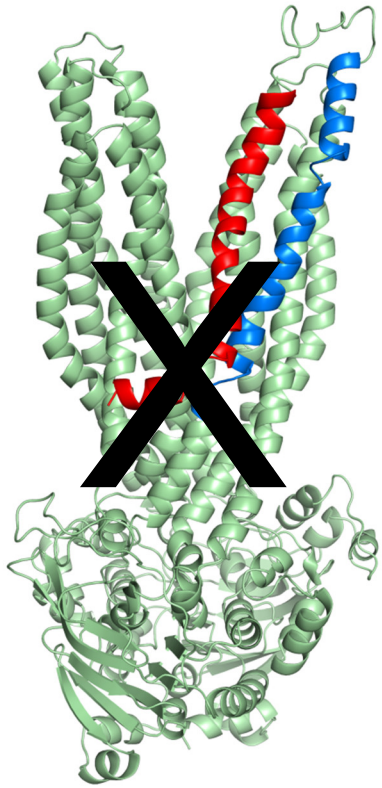
Cisztás fibrózis (CF)

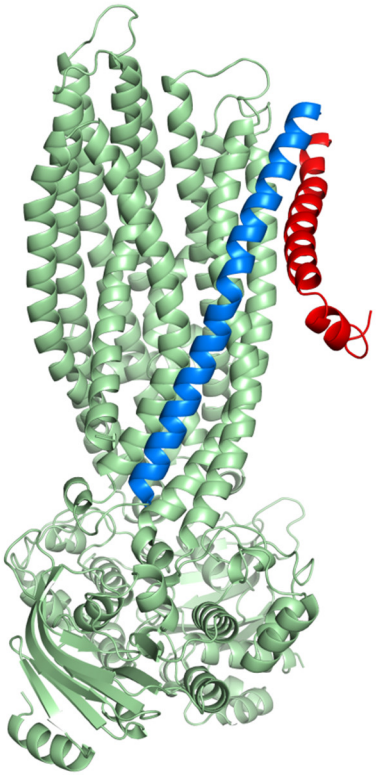


Full-length, ATP-bound CFTR structures

Cryo-EM revolution





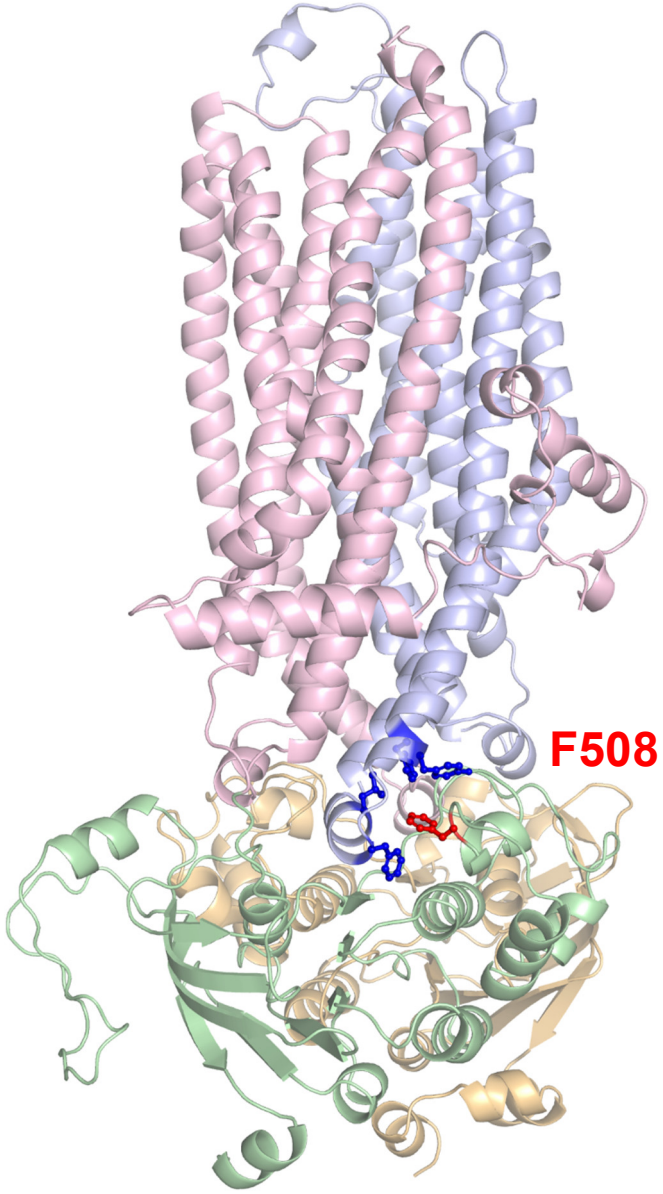


Zhang et al. (2017) Cell 170: 483-491.e8
PDBID:5W81

Bob Ford
University of Manchaster, UK

J. Fay, Jack Riordan
UNC, Chapel Hill, USA

ΔF508 mutation



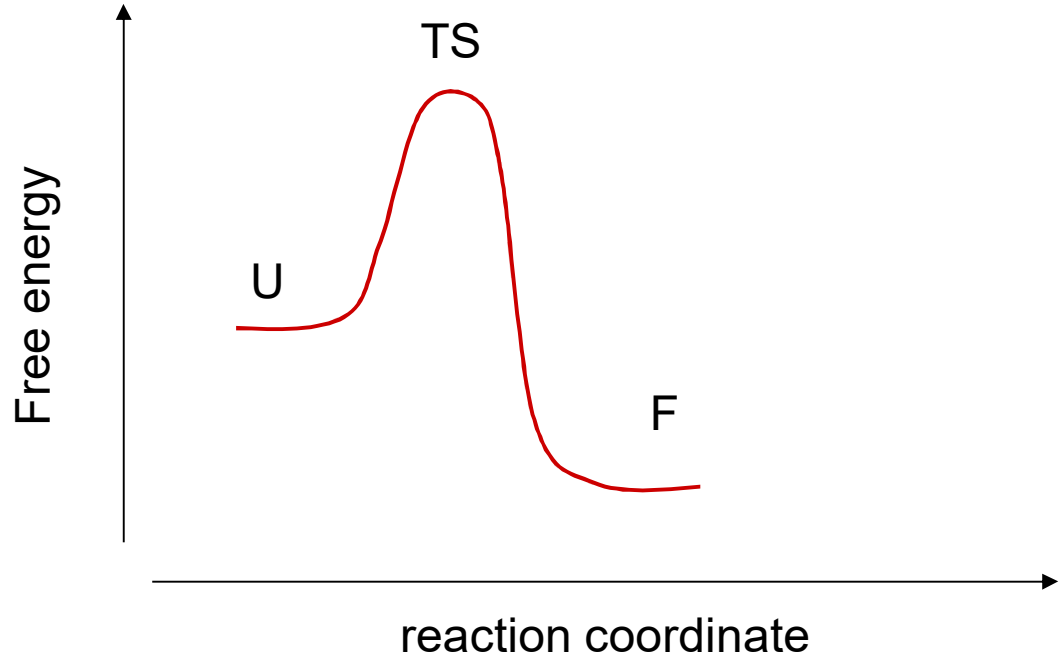
Many experimental and computational studies

Domain folding
Domain stability
Domain-domain assembly

Transmission of the consequence of a mutation; allosteric propagation of alterations in dynamics

Protein folding

Two-state model



nucleation

Levinthal's paradox

Simulation techniques for protein folding

all-atom force-field
(temperature) replica exchange MD

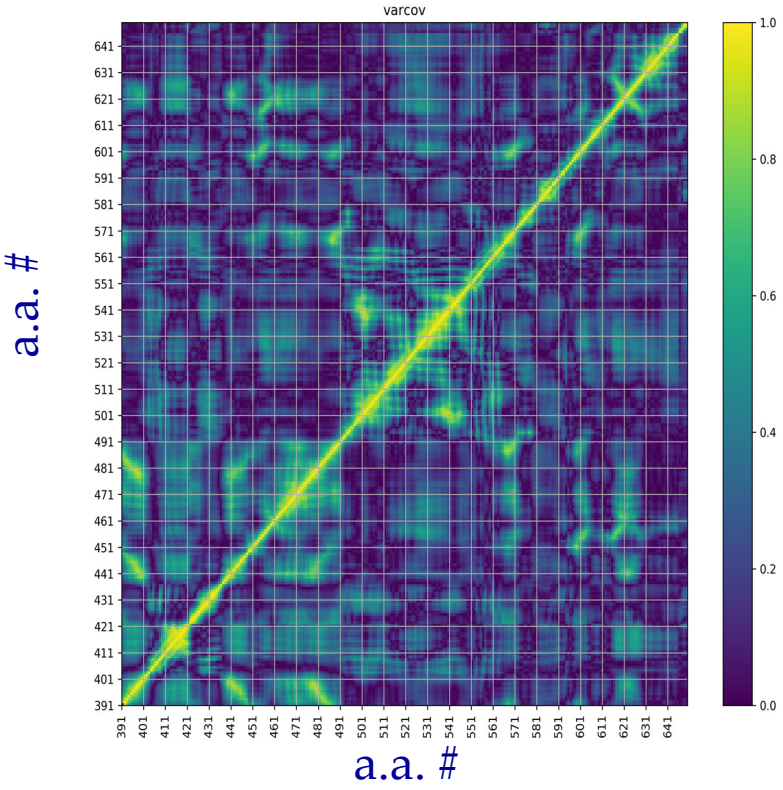
NBD1 allostericity

correlation in amino acid motions

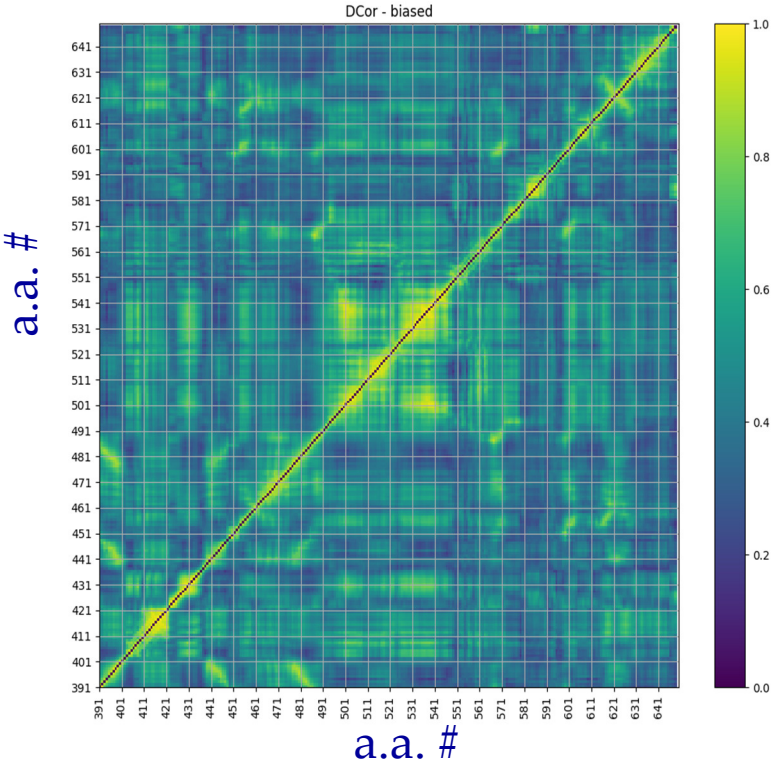
- Pearson correlation
- MI (mutual information)
- DiCC (distance correlation coefficient)

$$C_{ij} = \frac{\langle \Delta\vec{r}_i(t) \cdot \Delta\vec{r}_j(t) \rangle}{(\langle \Delta\vec{r}_i(t)^2 \rangle \langle \Delta\vec{r}_j(t)^2 \rangle)^{1/2}}$$

Pearson

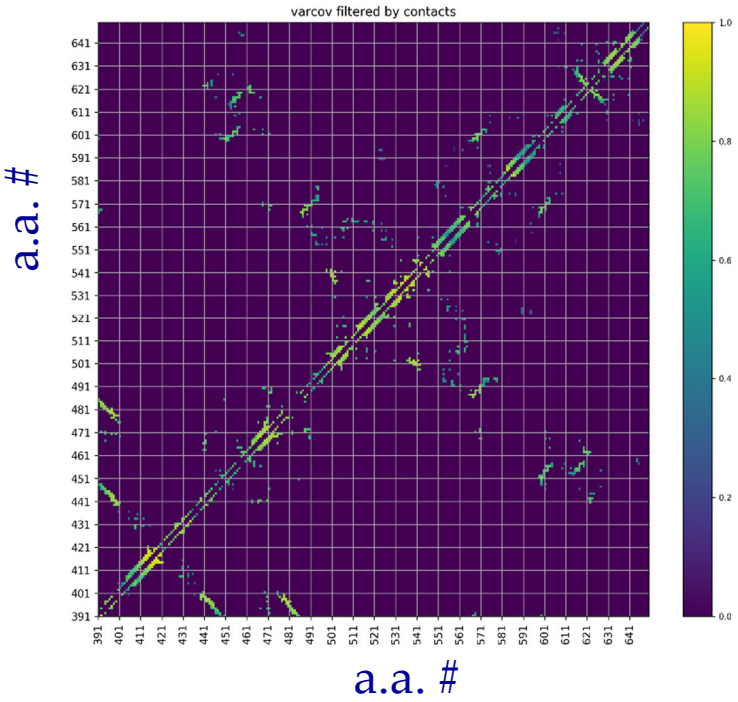


DiCC v1
Gábor Székely

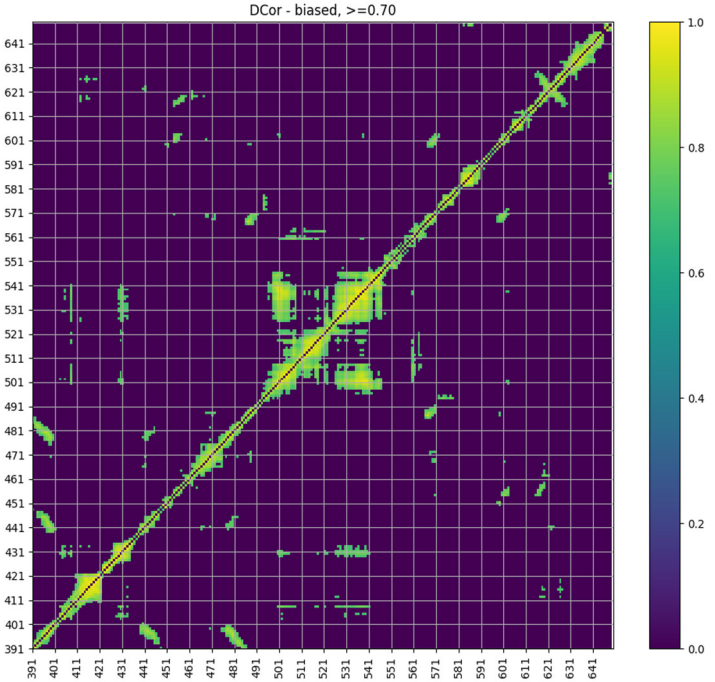


Correlation in amino acid motions

Pearson, filtered by contacts (75%)

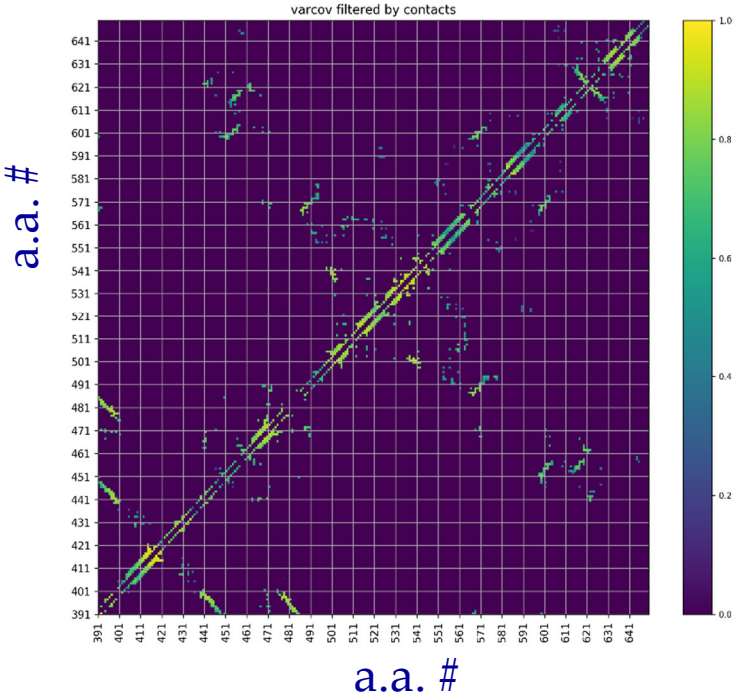


DiCC, cutoff = 0.7

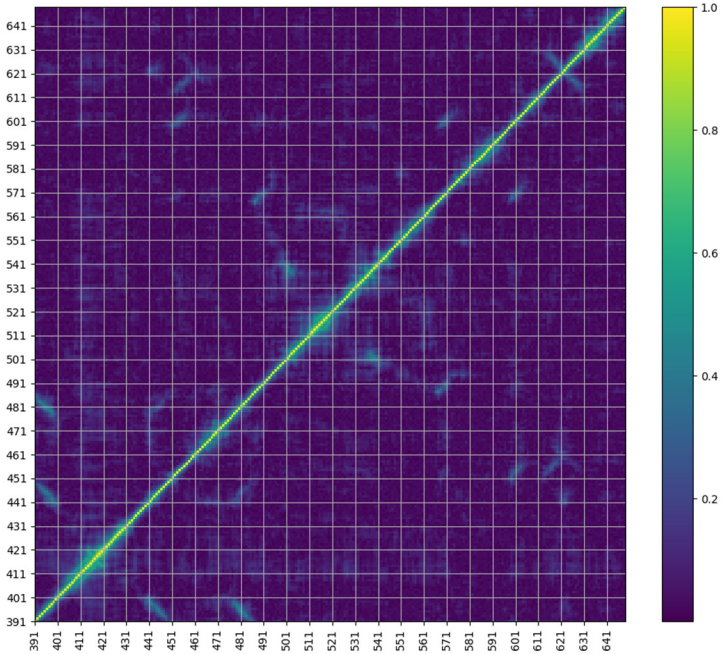


Correlation in amino acid motions

Pearson, filtered by contacts (75%)



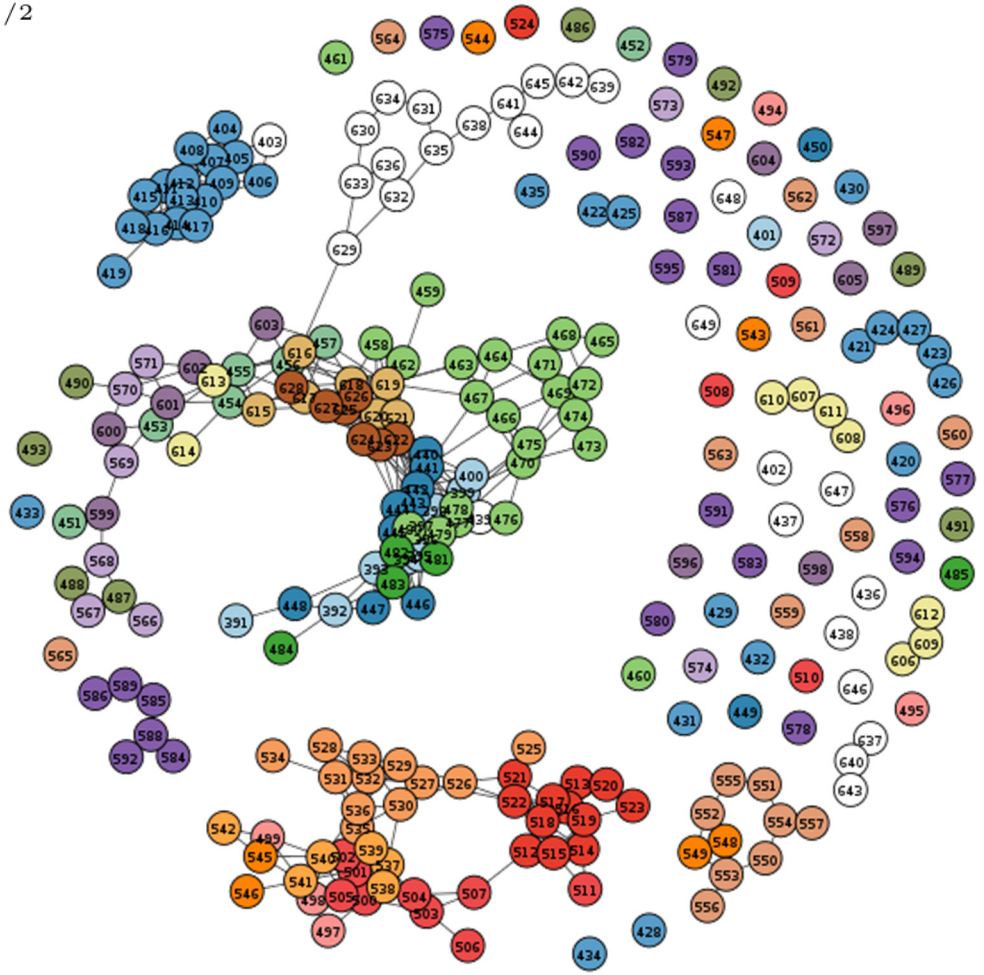
DiCC v2



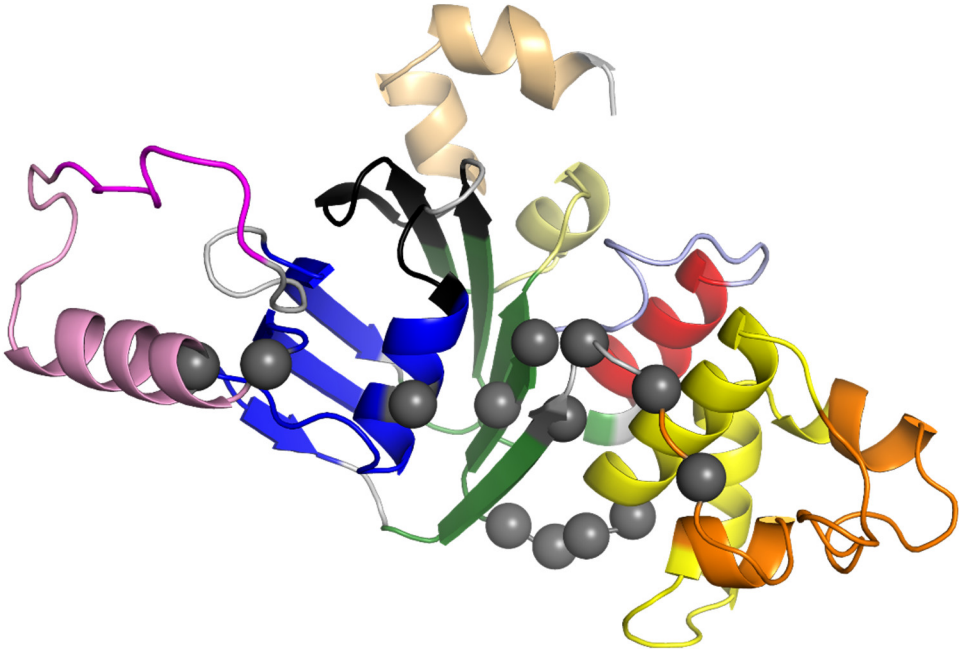
A network, based on motion correlation

$$C_{ij} = \frac{\langle \Delta\vec{r}_i(t) \cdot \Delta\vec{r}_j(t) \rangle}{(\langle \Delta\vec{r}_i(t)^2 \rangle \langle \Delta\vec{r}_j(t)^2 \rangle)^{1/2}}$$

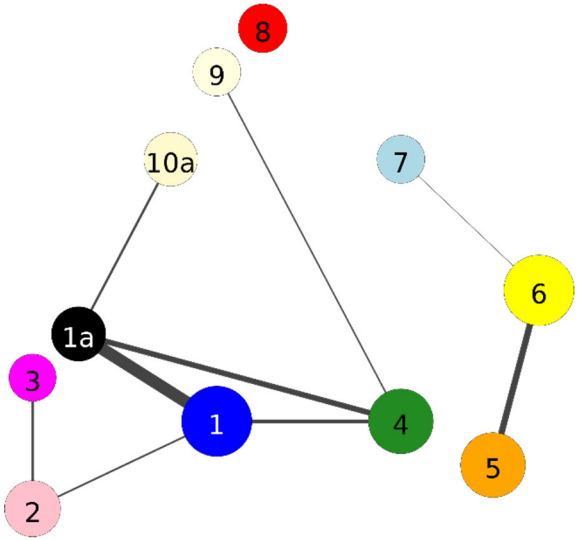
$$d_{ij} = -log(|C_{ij}|)$$



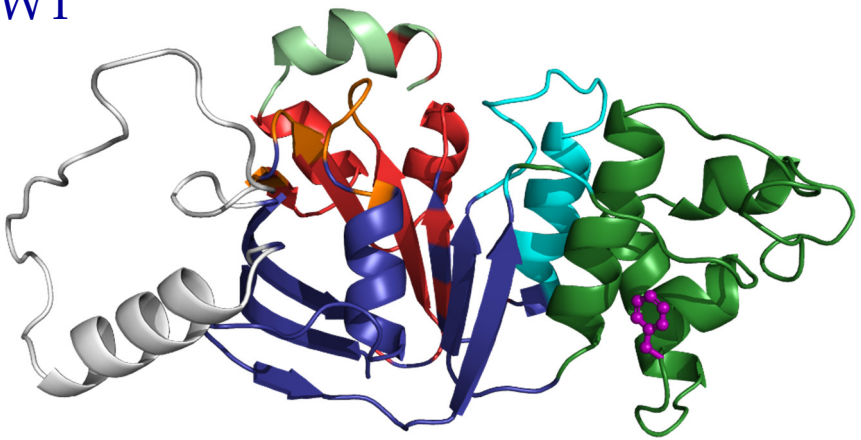
Betweenness centrality – critical residues



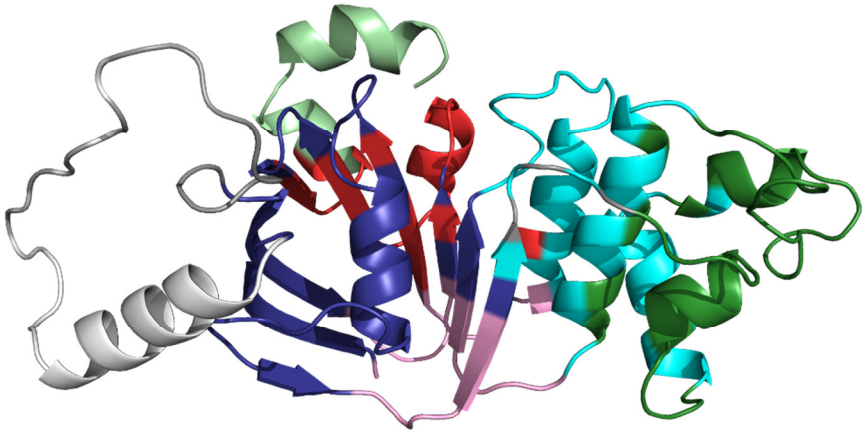
Community analysis



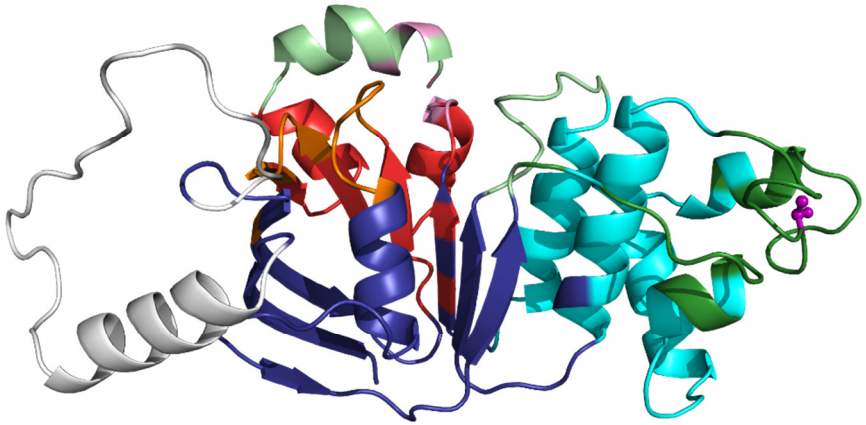
WT



ΔF508



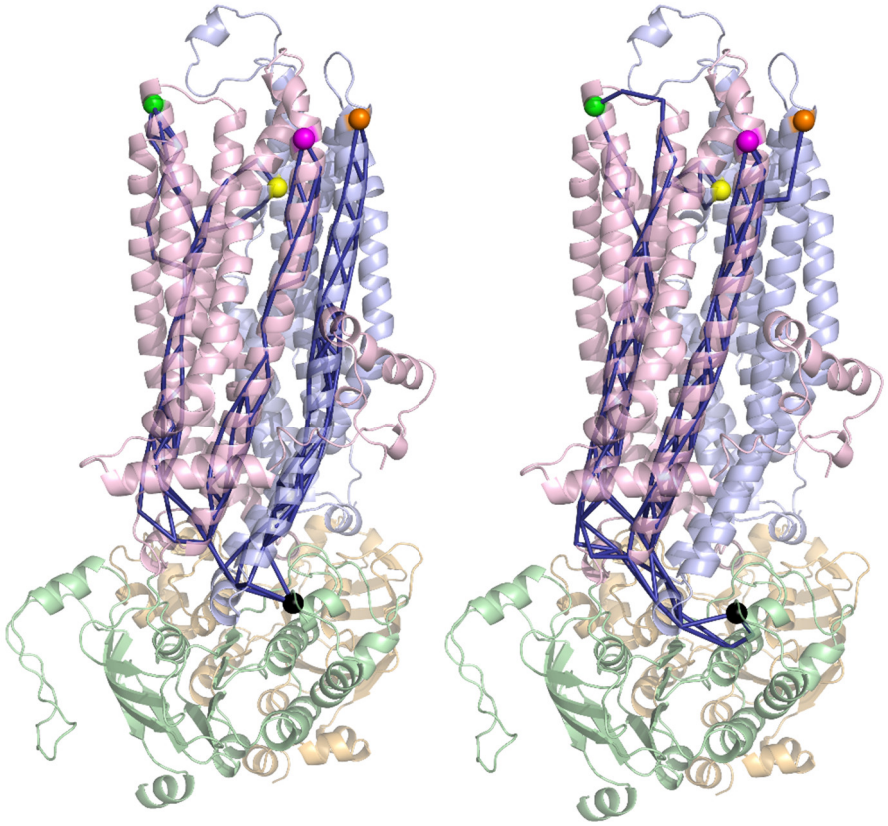
ΔF508 + rescue mutation



Optimal and suboptimal pathways

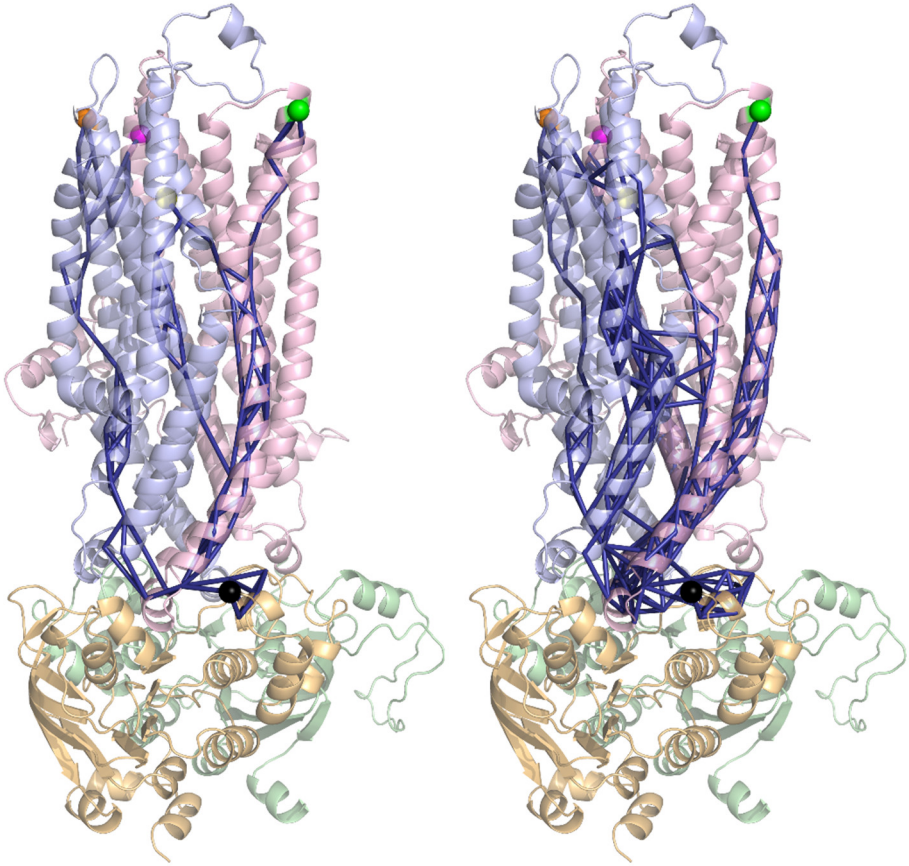
source #1
sinks #1-#4

WT mutant



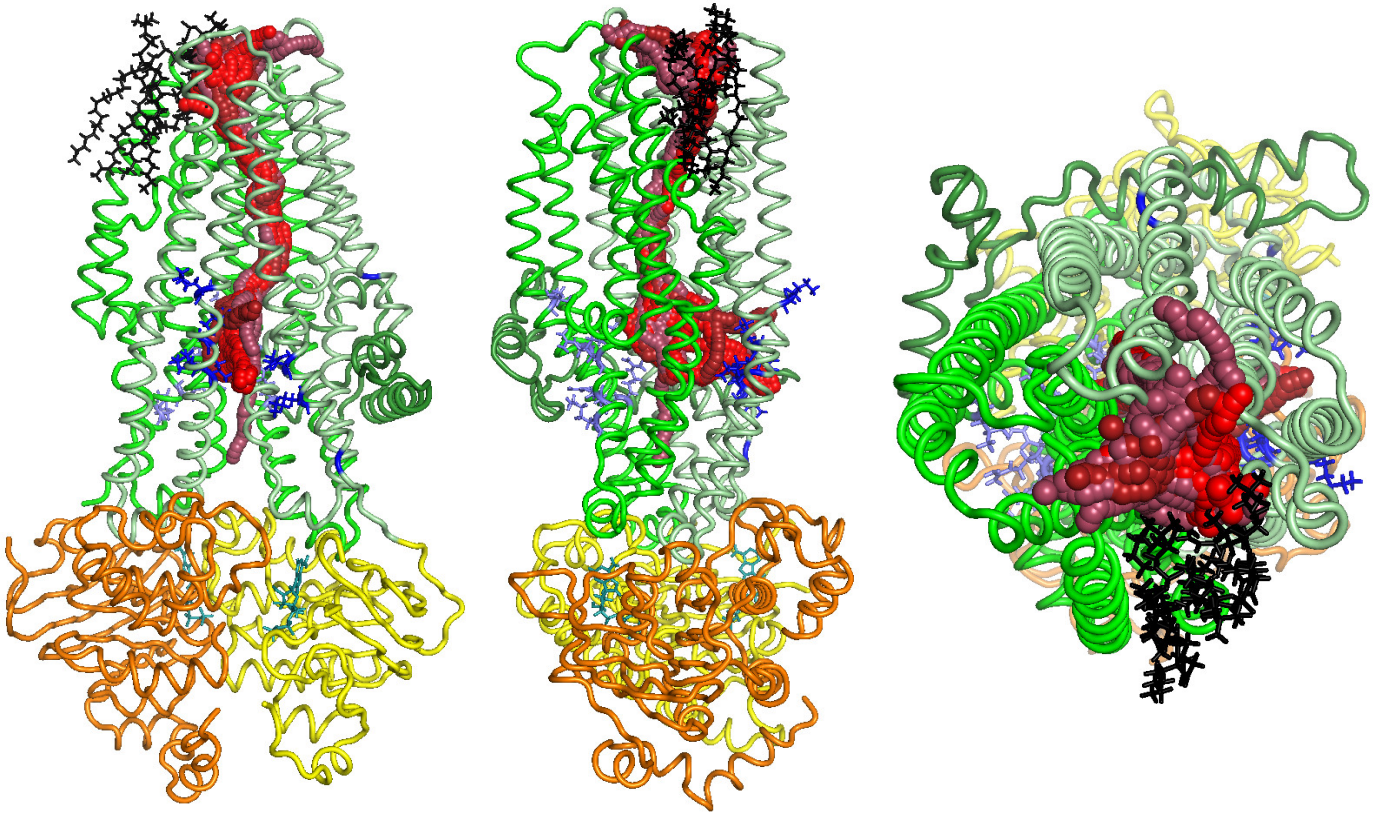
source #2
sinks #1-#4

WT mutant



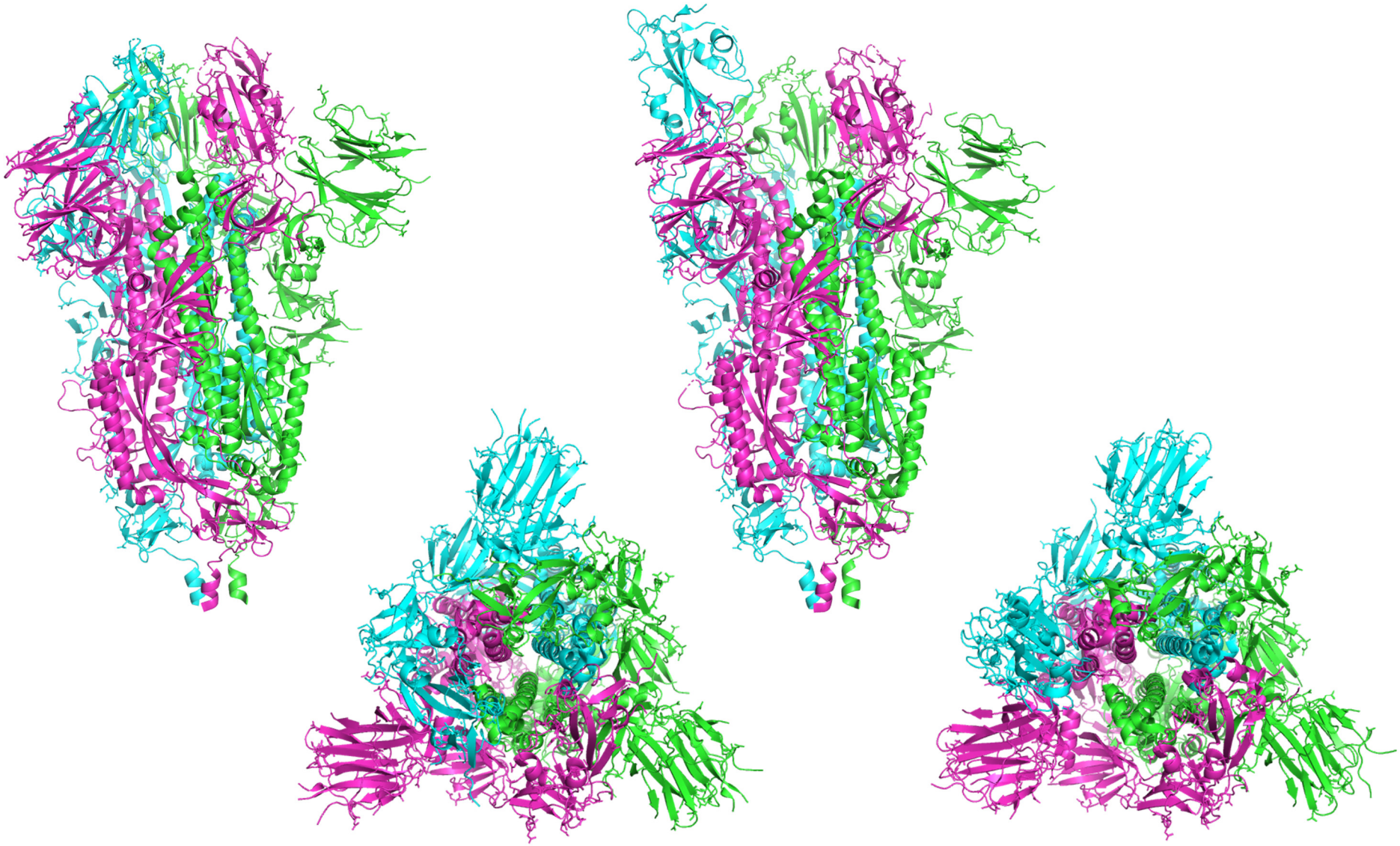
Identification of the chloride permeation pathway

Farkas *et al.* Cell Mol Life Sci. 2019 Jul 20. doi: 10.1007/s00018-019-03211-4



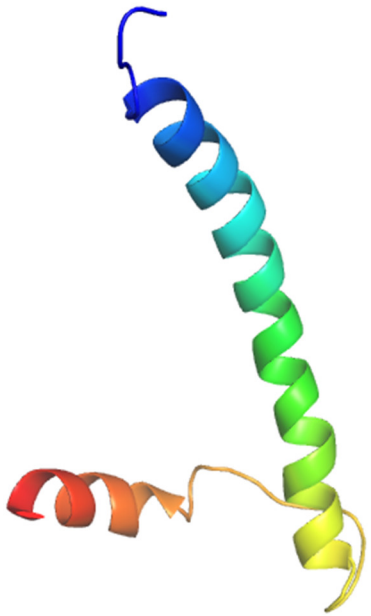
SARS CoV-2, Spike protein

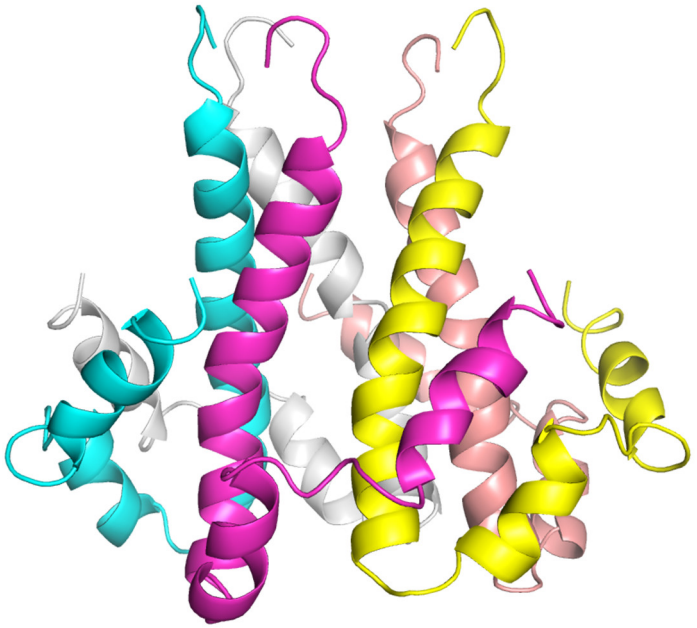
protein-protein interaction with human ACE2



SARS CoV-2, E protein

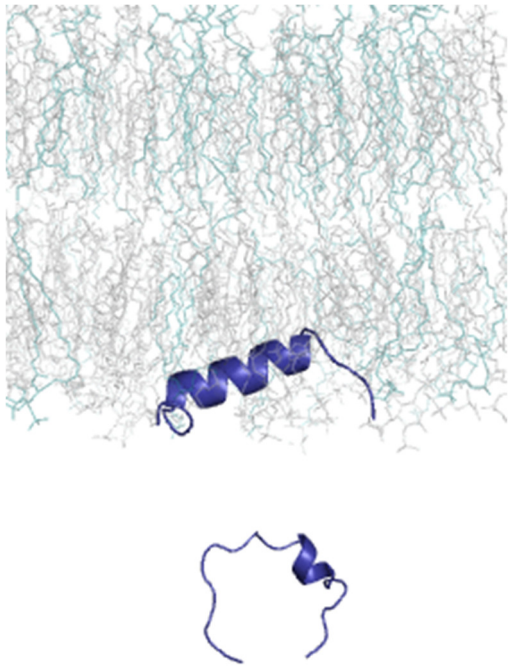
E protein





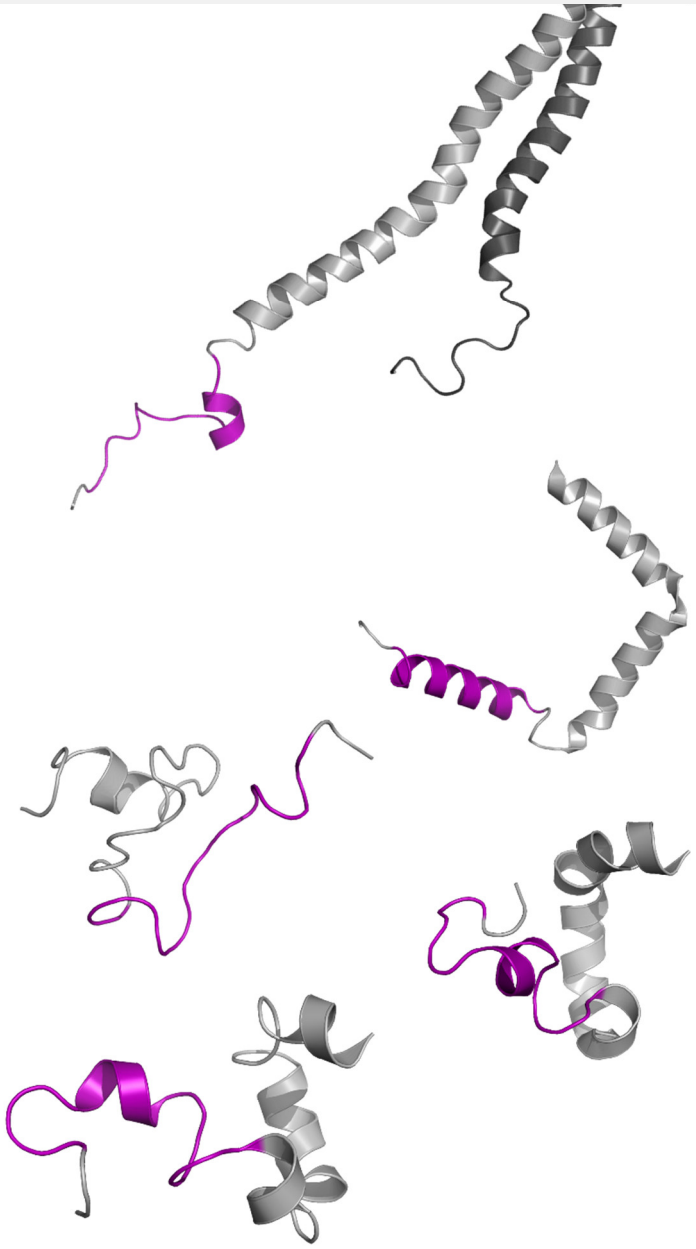
MemMoRFs

Membrane Molecular Recognition Features, https://memmorf.hegelab.org



Integrin β3

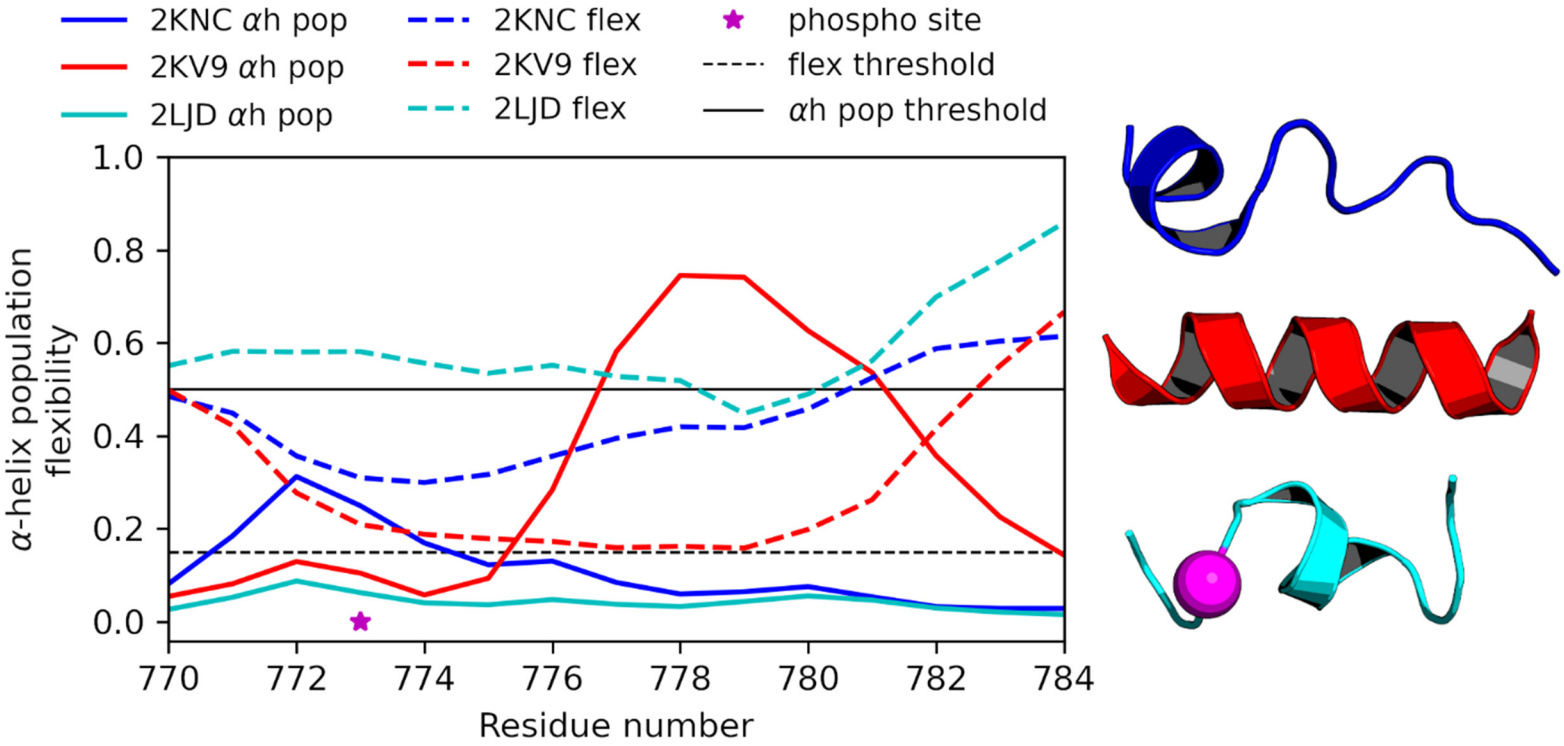
PDBID	Pozíció	Membrán-környezet	Egyéb
2KNC	TMD-CD	organic solvent	ALFAIIB-BETA3 hetero-complex
2KV9	CD	micelle	disulfide linked alphaIIb-beta3
2LJF	CD	aqueous conditions	TYR 773 PTR
2LJD	CD	micelle	TYR 773 PTR
2LJE	CD	micelle	TYR 773 PTR, TYR 785 PTR



MemMoRF

Membrane Molecular Recognition Feature

Integrin beta-3



blue: in organic solvent
red: in DPC
cyan: phosphorylated in DPC

magenta: phosphorylation site; αh pop: α-helix population calculated by δ2D;
flex: 1-S2 calculated by RCI, αh pop threshold: 0.5, flex threshold: 0.15.

http://memmorf.hegelab.org

Summary

- **Protein structure**

  prediction, homology modelling, disorder

- **Protein dynamics**
  - ensembles at 37°C
  - molecular dynamics, simplified models
  - describing effect of mutations: ABCG2 Q141K, CFTR ΔF508
  - describing function:
    - ABCG2 transport
    - CFTR channel
    - SARS Cov-2, S and E proteins
    - MemMoRF

Thanks for your attention!

hegedus.tamas@hegelab.org