3D-Bioinformatics
Protein structure and dynamics

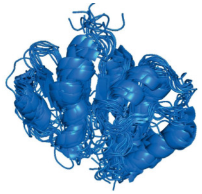
Tamás Hegedűs

hegedus.tamas@hegelab.org

Importance of protein structure and dynamics

Provides the atomic-level basis of diseases.

Helps to understand the shape of drug binding sites.

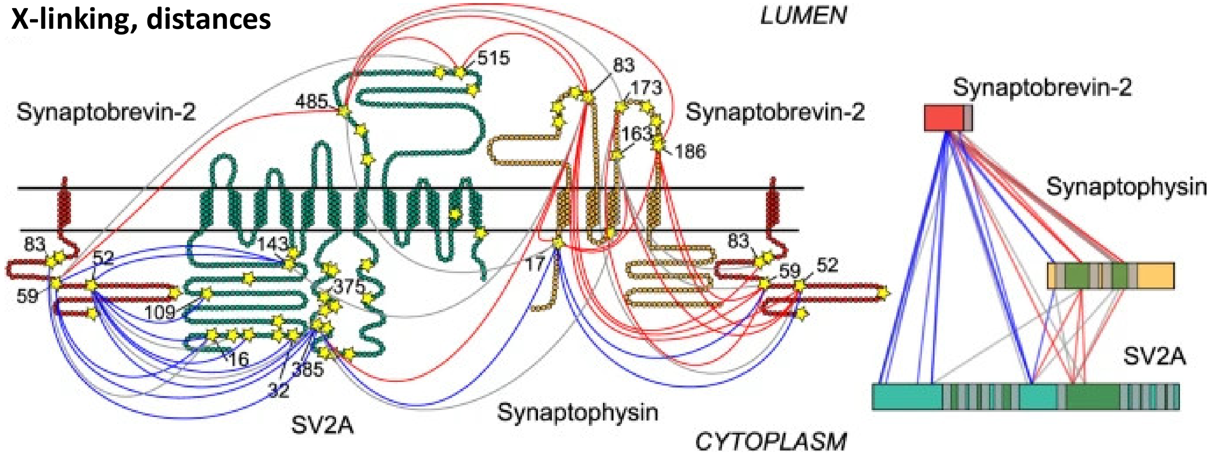


Proteins do not exist as a single structure, but as a **conformational ensemble** at 37°C.

Importance of Computational Modelling

- Offers atomic-level information on protein motions.
- Experiments typically do not provide this level of detail (with exceptions like NMR).

X-linking, distances



Wittig et al. Nat Comm 2021 12:858

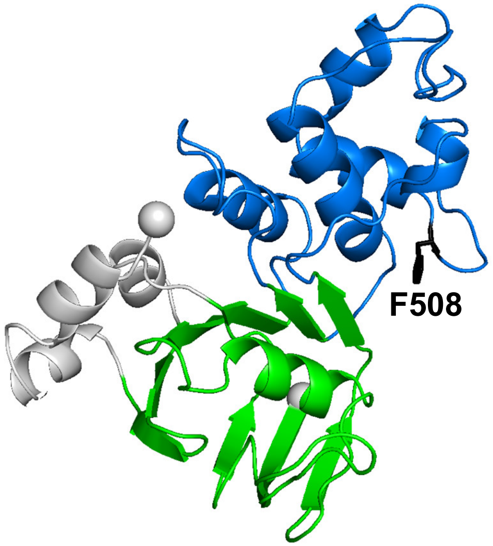
Topics

- ➢ Short introduction to protein structure
- ➢ Non-structured and dynamic disordered proteins
  - Membrane Molecular Recognition Features
  - Predictions
  - Protein Language Models
- ➢ Protein 3D structure
  - Cryo-EM
    TM topology
  - Prediction
    Homology modelling
    AlphaFold

Secondary structure

- Local folded structures
- Main types: alpha-helix, beta-sheet, turns and loops
- Help to define the overall 3D shape (tertiary structure)
- Contribute to the protein's stability and function
- Identified experimentally (e.g. X-ray crystallography and NMR) and computational prediction tools (e.g., PSIPRED, JPred)

CFTR NBD1



- **α-subdomain – blue**
- **β-subdomain – green**

Intrinsically Disordered Proteins (IDPs) – Overview

- 25% of proteins are predicted to be disordered.
- Disorder increases with the complexity of organisms.
- 50% of human proteins contain a disordered region that is 30 amino acids (a.a.) or longer.

- IDPs are not fully random but display structural flexibility.
- They lack compact globular folding and residual structure.

Benefits and roles of IDPs

- Specificity and adaptation to different environments.
- Reversible transitions between ordered and disordered states.
- Large binding surface, allowing for multiple interactions.
- Fast binding, providing efficiency in cellular processes.

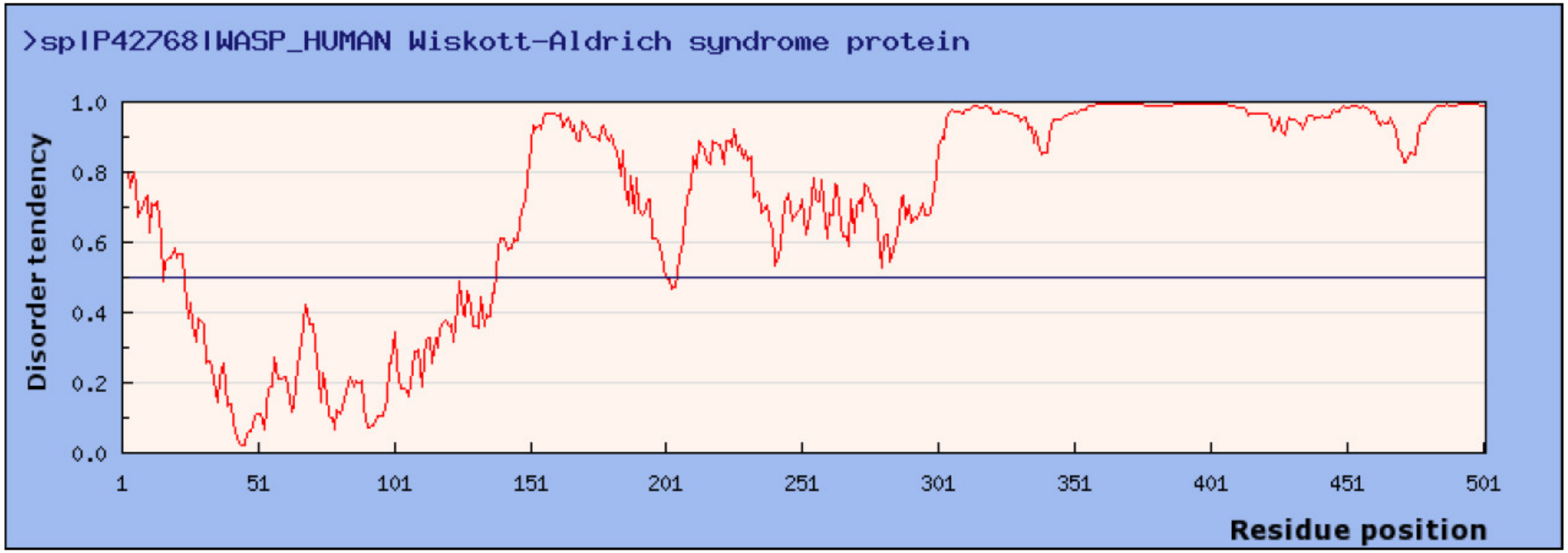
- **Entropic chain**: Inactivating **K+ channels**.
- **Effectors**: Acting as **peptide inhibitors**.
- **Scavengers**: Example: **casein**.
- **Assembly**: Role in forming structures, e.g., **calmodesmon** with **F-actin**.
- **Presentation**: Providing sites for **phosphorylation** and **cleavage**.

Computational Approaches for IDPs

- Learning algorithms trained on disordered sequences from the PDB database.
- Algorithms predict interaction energies of disordered regions.

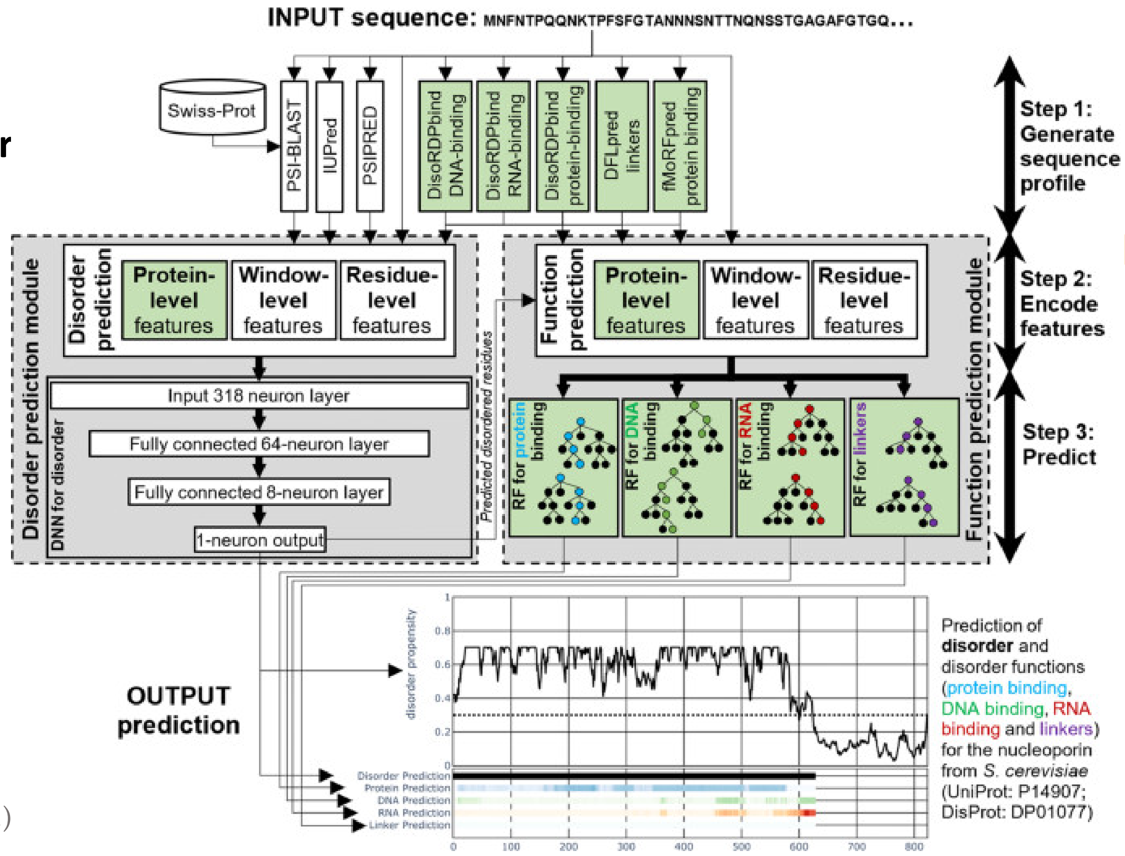
Disopred2
AlphaFold2

https://iupred3.elte.hu



Computational Approaches for IDPs

flDPnn: Winner of the **Critical Assessment of protein Intr Disorder (CAID)** competition.



Prediction of **disorder** and disorder functions (protein binding, DNA binding, RNA binding and linkers) for the nucleoporin from *S. cerevisiae* (UniProt: P14907; DisProt: DP01077)

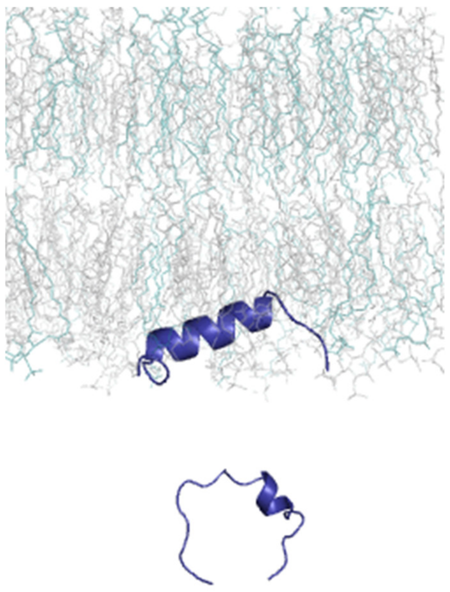
hu

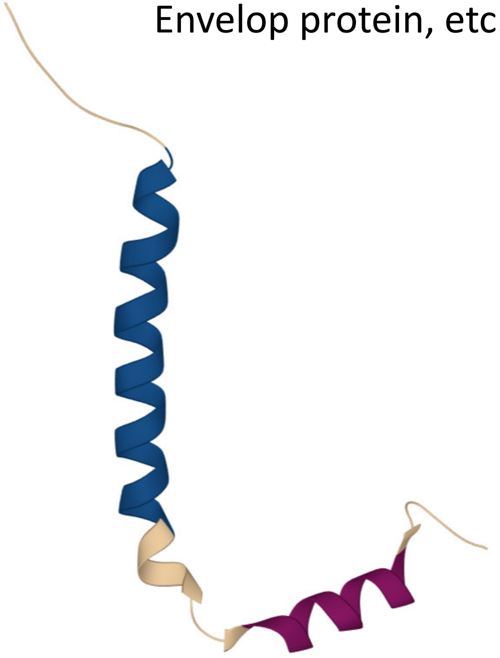
Hu *et al*. Nature Communications volume 12, Article number: 4438 (2021)

MemMoRFs

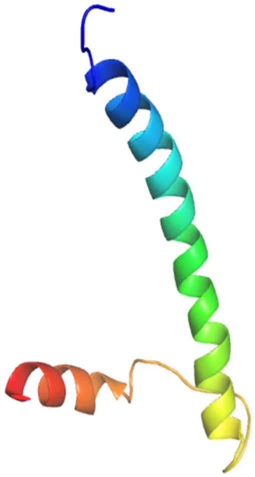
Membrane Molecular Recognition Features
https://memmorf.hegelab.org

KRAS, STX17 autophagy protein, SARS-Cov2
Envelop protein, etc.



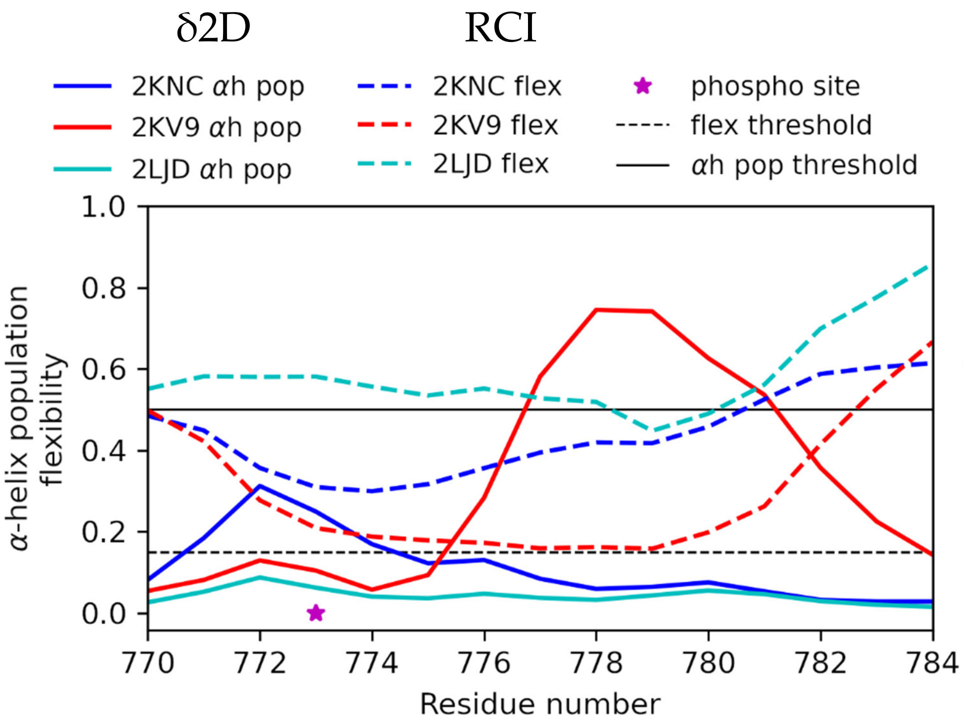


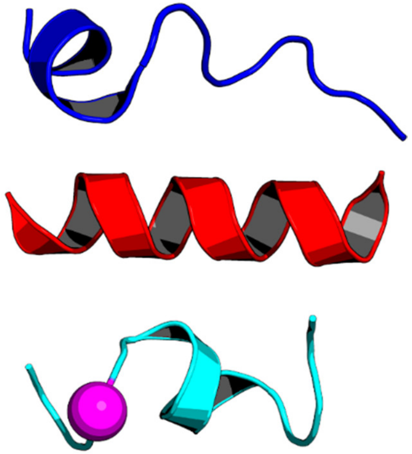
EGFR, PDBID: 2N5S



E protein

Integrin beta-3





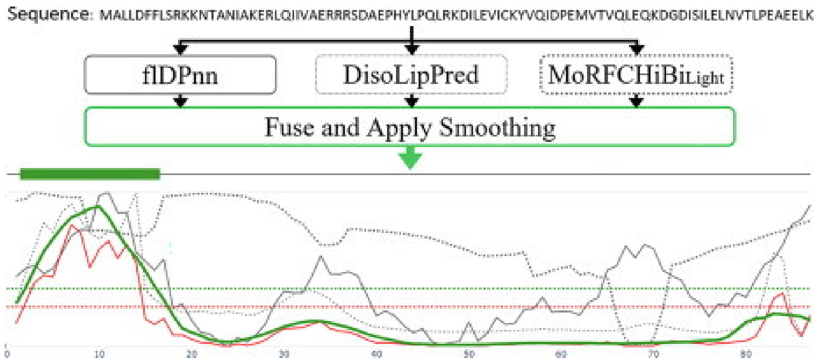
in organic solvent

in DPC

in DPC, phosphorylated

CoMemMoRFPred

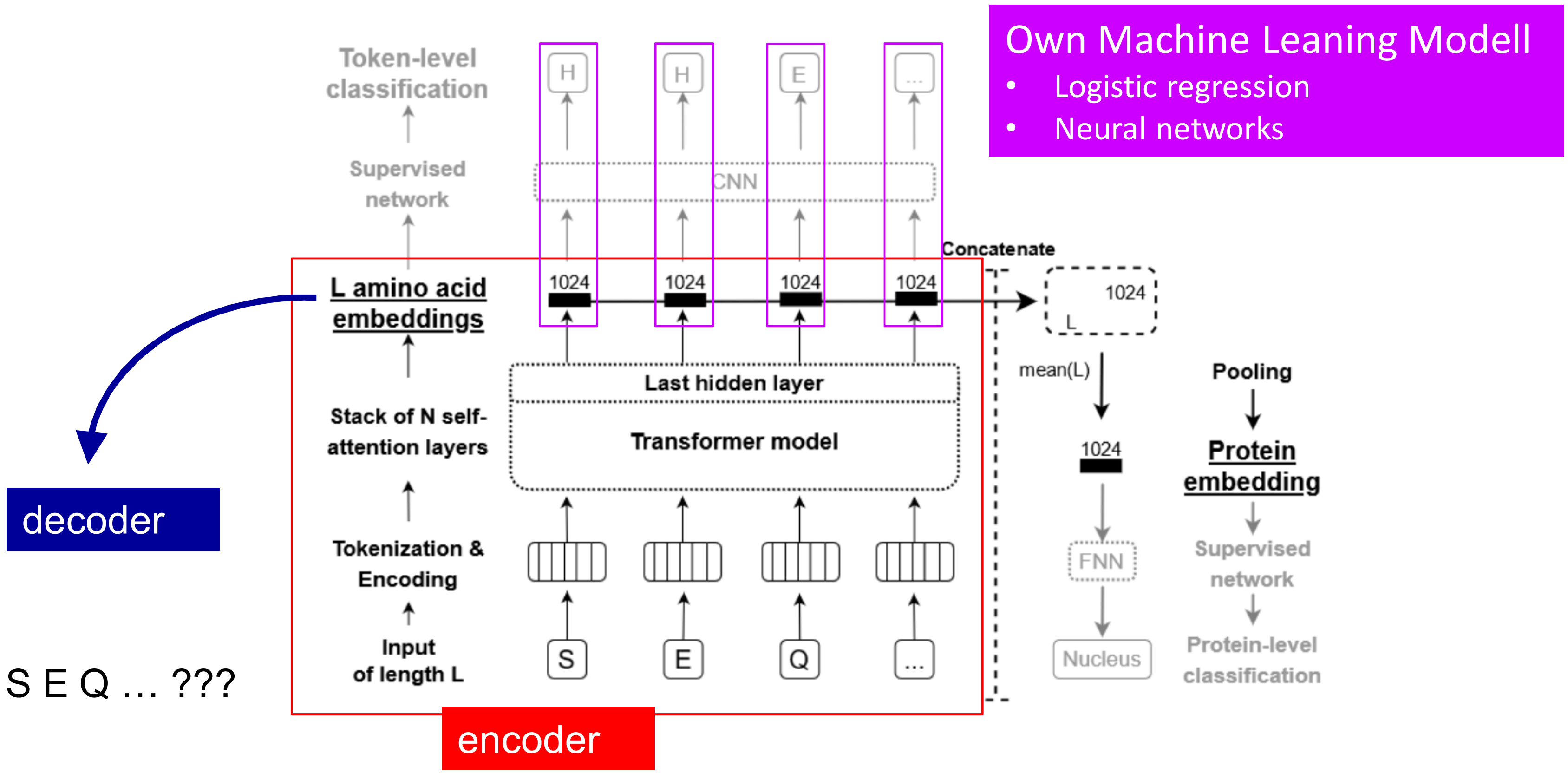
N(proteins)	67
N(proteins) w. 30% cutoff	41
N(MemMoRF-residues)	684
N(other residues)	20,275


Sequence: MALLDFFLSRKKNTANIAKERLQIIVAERRRSDAEPHYLPQLRKDILEVICKYVQIDPEMVTVQLEQKDGDISILELNVTLPEAEELK
fIDPnn
DisoLipPred
MoRFCHiBiLight
Fuse and Apply Smoothing
0
10
20
30
40
50
60
70
80

Basu, Hegedus, Kurgan
J Mol Biol. 2023 Nov 1;435(21):168272

AUC	0.765
rateAUC	5.7
F1_max	0.204

pLM (protein Language Model)



Own Machine Leaning Modell

- Logistic regression
- Neural networks

decoder

S E Q … ???

encoder

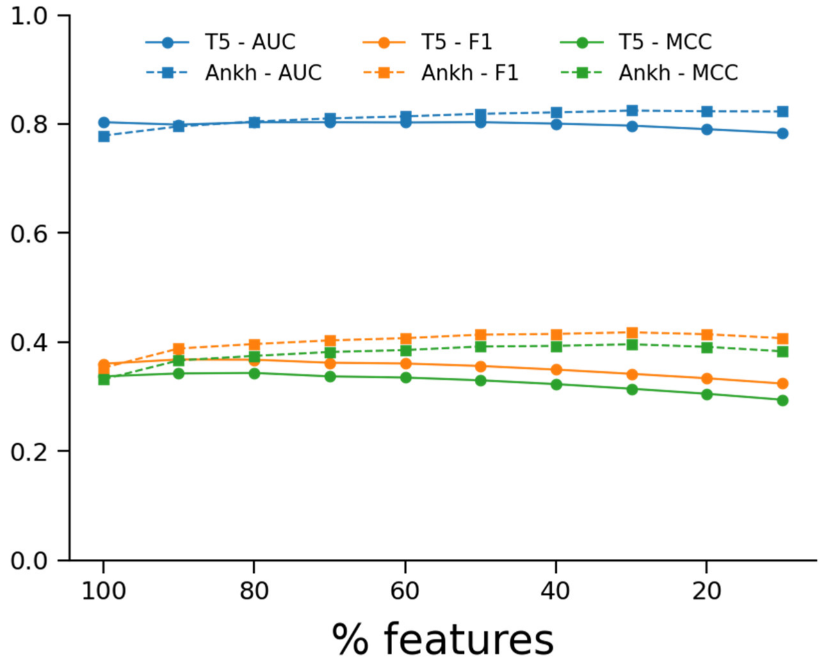
Elnaggar *et al.* 2022, https://rostlab.org
IEEE transactions on pattern analysis and machine intelligence

Logistic regression

pLM: protT5, Ankh
solver
penalty
regularization strength

	AUC	F1_max	MCC
T5	0.803	0.36	0.337
Ankh	0.779	0.352	0.329

Recursive feature elimination



Neural network

1 Linear layers
1 Dropout layer
1 Linear layer
1 Dropout layer
1 Linear layers
Sigmoid activation

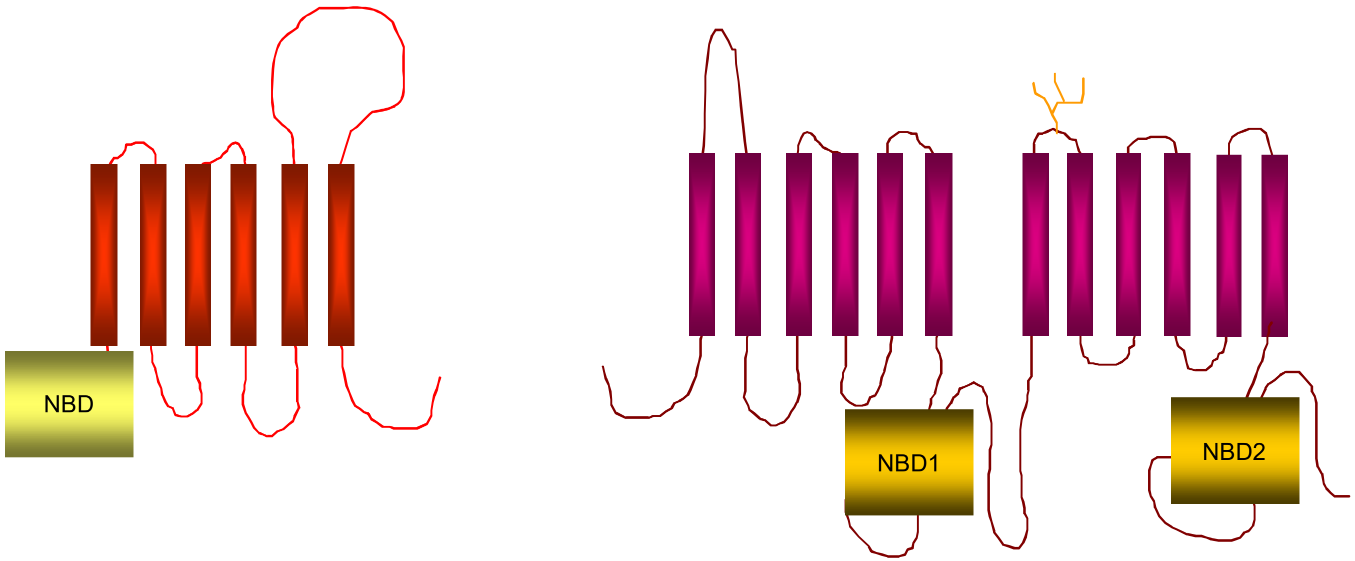
Training:

new non-MemMoRF set for each epoch

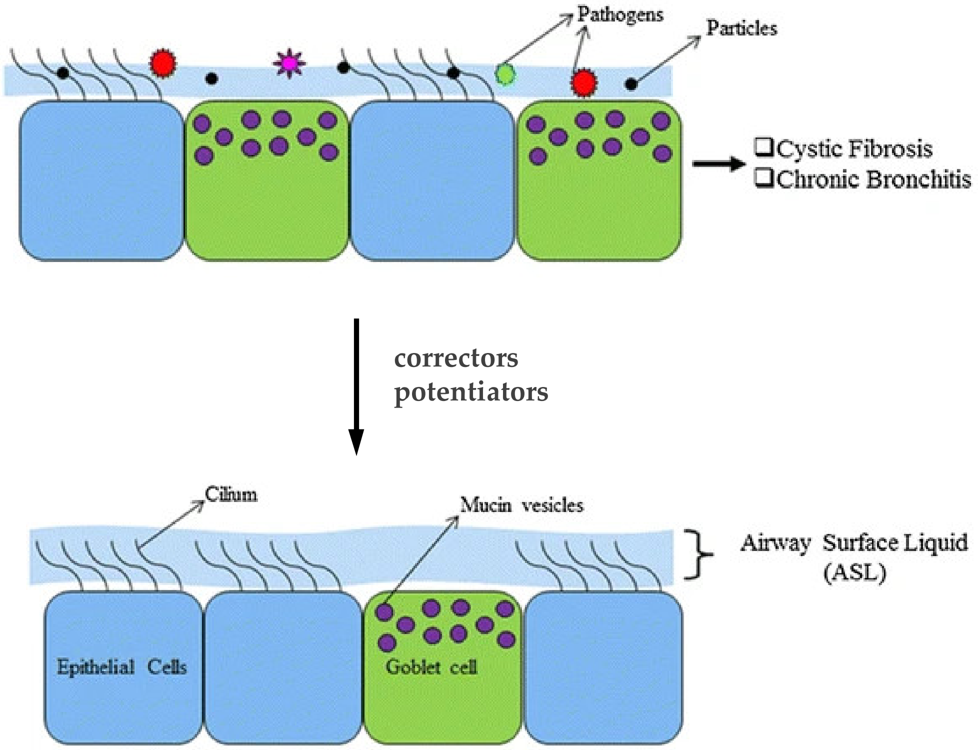
	AUC	F1_max	MCC
All features	0.790	0.757	0.468
30% of features	0.830	0.783	0.542
Best model	0.967	0.929	0.855
Best model, full length	0.944	0.664	0.630

True \ Predicted	P	N
P	75	35
N	41	1,081

ABC (ATP Binding Cassette) proteins



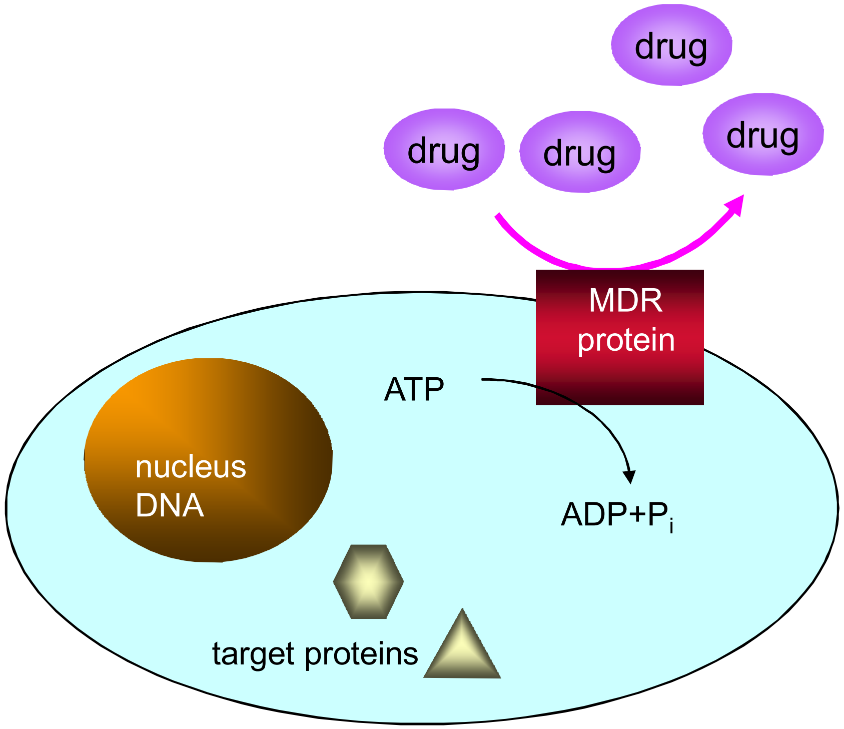
Cystic fibrosis – CFTR channel



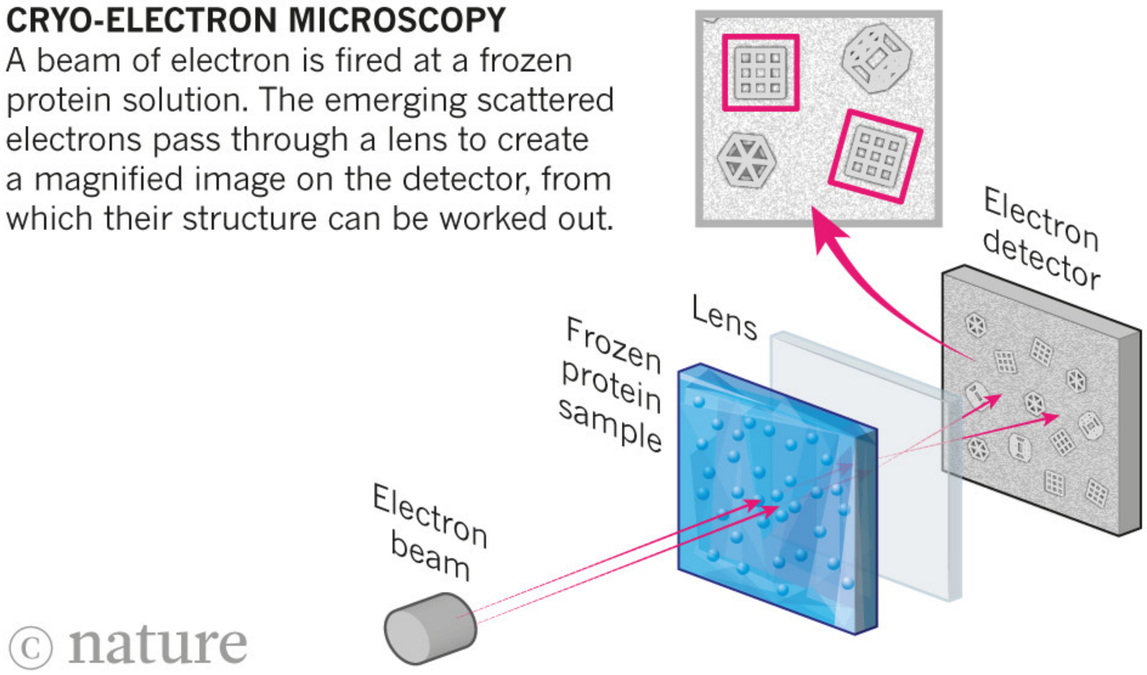
correctors
potentiators

Ghosh, Boucher, Tarran,
CMLS 2015

Multidrug transporters



Cryo-EM

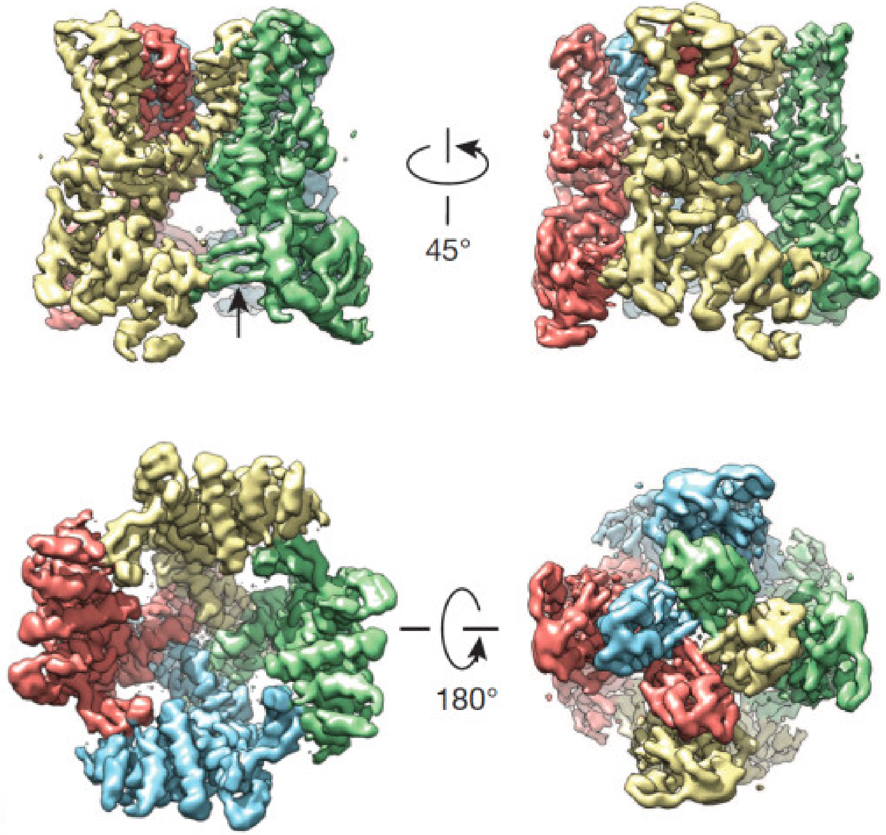


CRYO-ELECTRON MICROSCOPY

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

TRPV1 channel

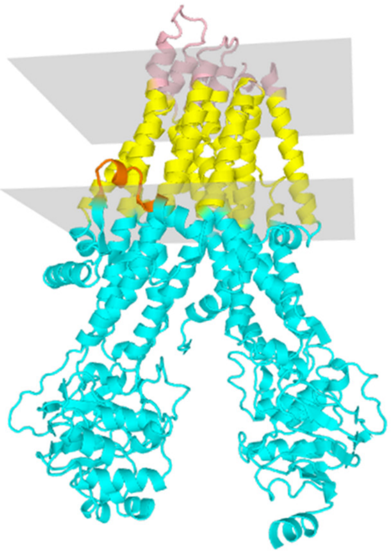


Prediction of TM topology

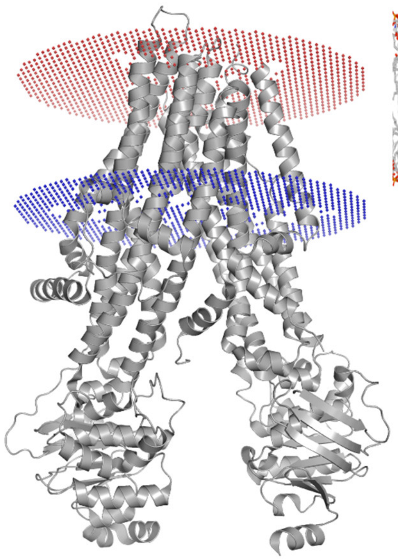
- Based on chemical properties of amino acids
- a.a. distribution in TM and soluble regions (statistics)
- Incorporation of experimental knowledge
- Integration of several predictors

UNIfied database
of TransMembrane Proteins
https://www.unitmp.org

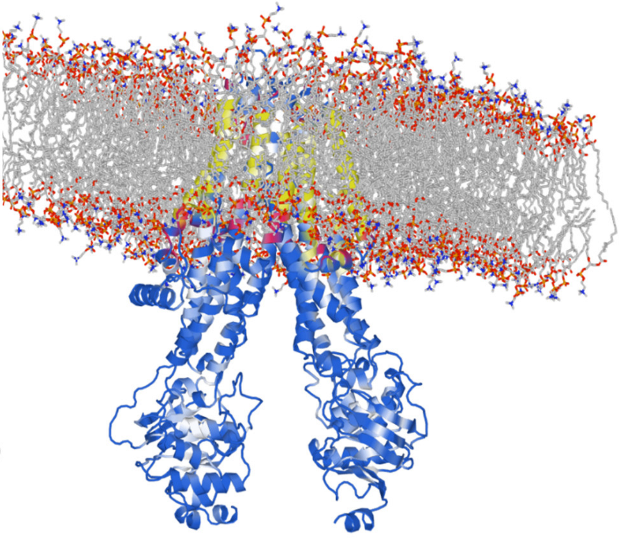
TMDET
PDBTM



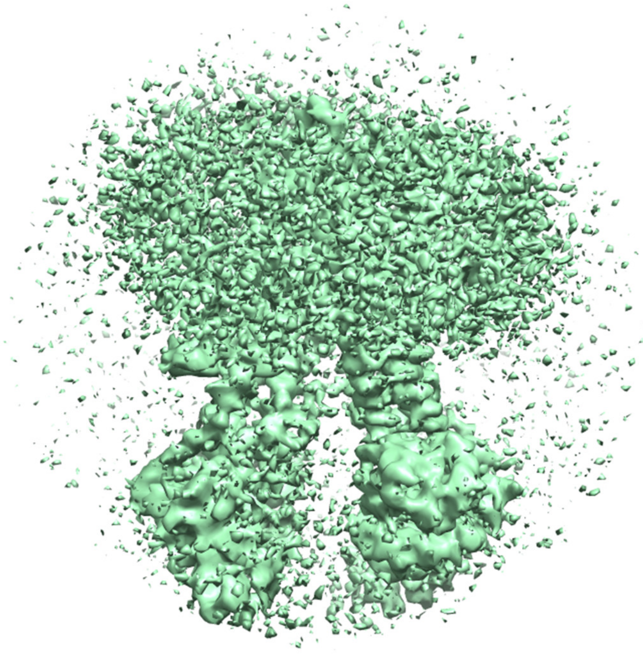
PPM
OPM



MemProtMD

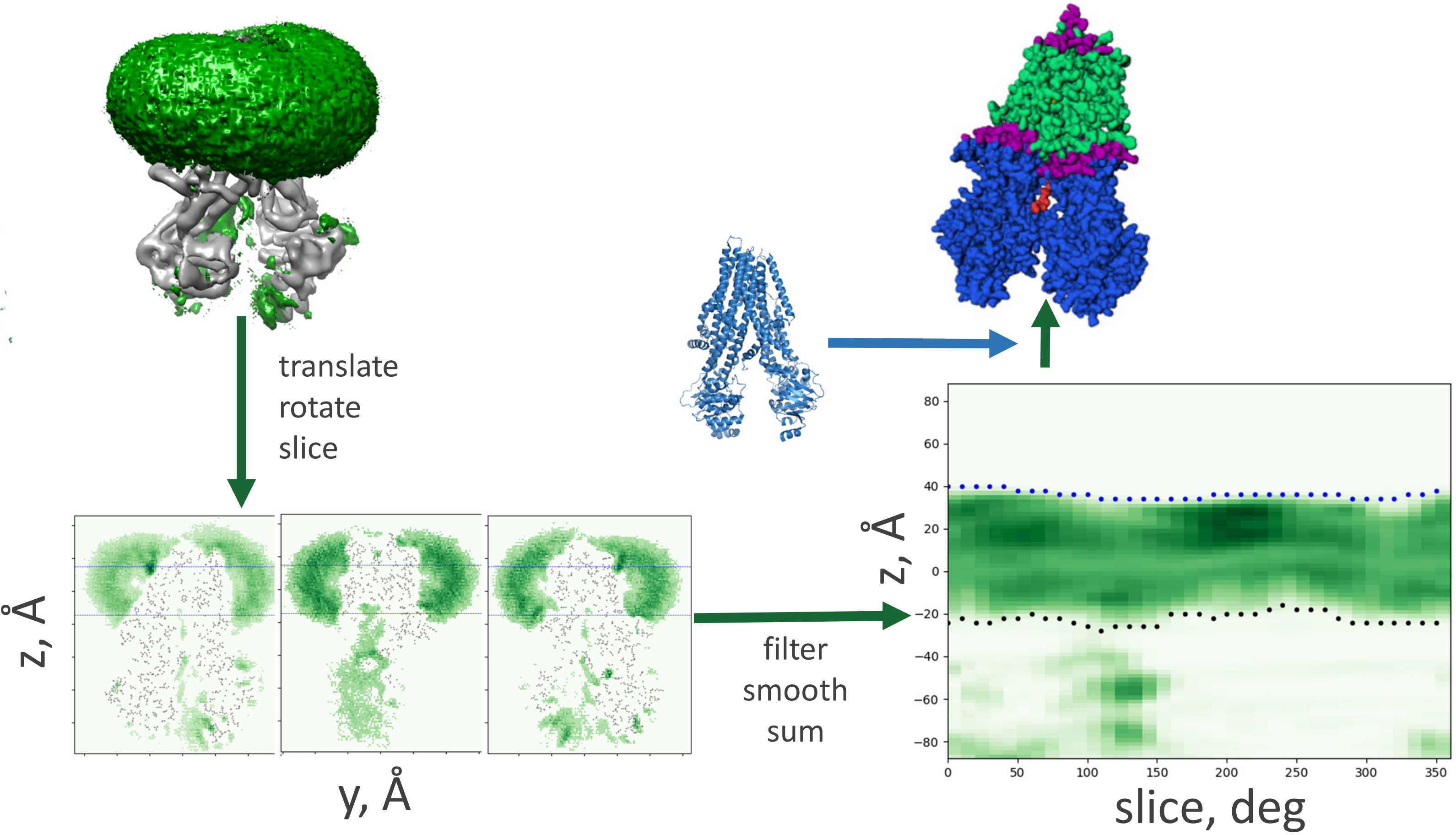


Membrane embedment in cryo-EM



CFTR (PDBID: 5UAK)

http://memblob.hegelab.org



3D structure prediction

Homology modelling

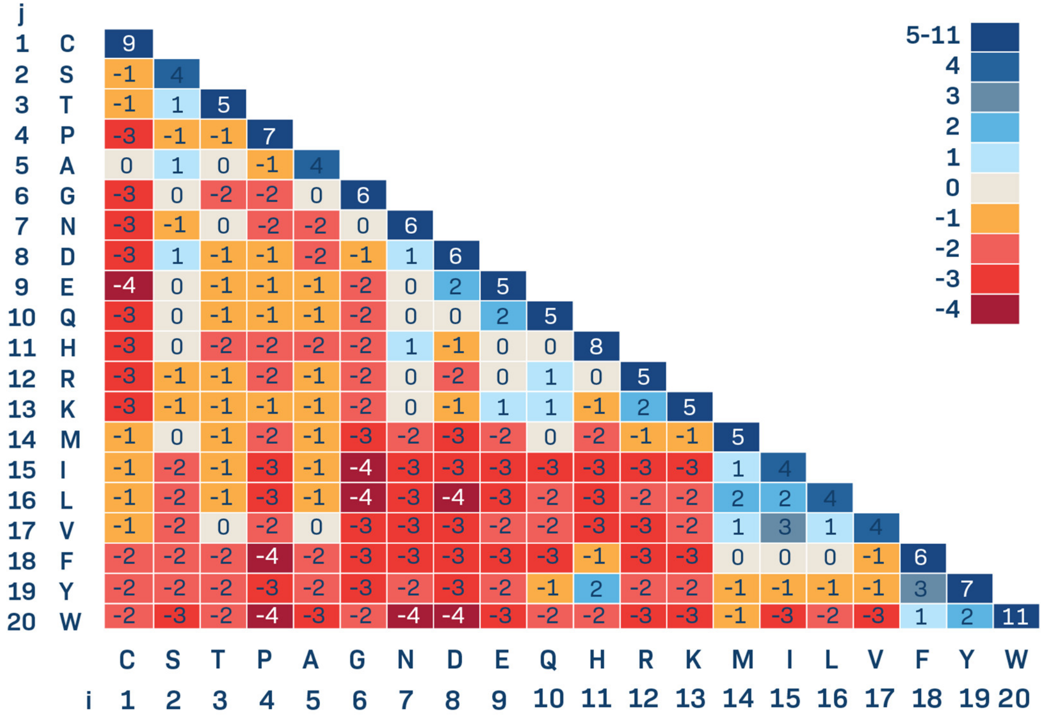
- conserved sequence == conserved structure
- > 30% similarity
- most important: the sequence alinement

„*Ab initio*" folding

- CASP (Critical Assessment of Techniques for Protein Structure Prediction)
- constraints from experiments
- deep learning (e.g. AlphaFold2, RoseTTAFold)

Homology modelling

1. Searching a template
2. Sequence alignment
3. Modelling
4. Energy minimization



j																					
1	C	9																			
2	S	-1	4																		
3	T	-1	1	5																	
4	P	-3	-1	-1	7																
5	A	0	1	0	-1	4															
6	G	-3	0	-2	-2	0	6														
7	N	-3	-1	0	-2	-2	0	6													
8	D	-3	1	-1	-1	-2	-1	1	6												
9	E	-4	0	-1	-1	-1	-2	0	2	5											
10	Q	-3	0	-1	-1	-1	-2	0	0	2	5										
11	H	-3	0	-2	-2	-2	-2	1	-1	0	0	8									
12	R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
13	K	-3	-1	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
14	M	-1	0	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
15	I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
16	L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
17	V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
18	F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
19	Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
20	W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
	i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

https://www.labxchange.org

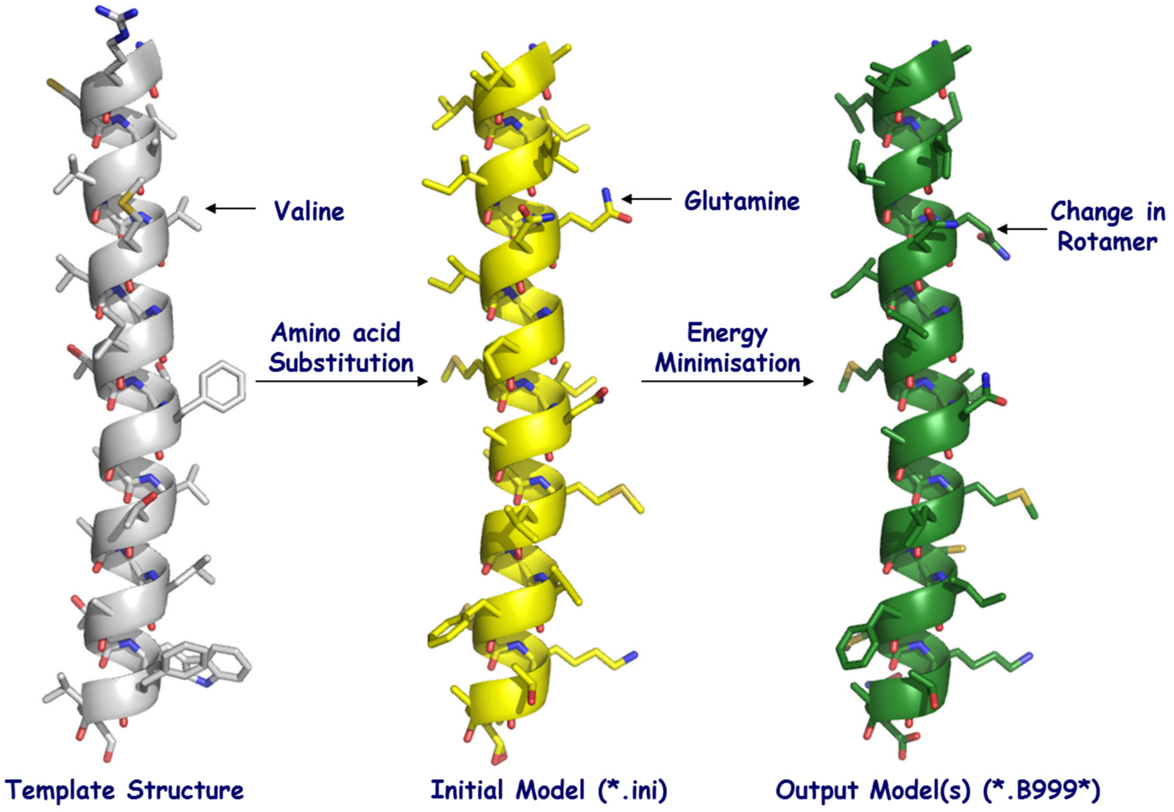
BLOSUM62
(BLOcks of Amino Acid SUbstitution Matrix)

Basic Local Alignment Search Tool (BLAST)

Alignement – pl. ClustalW

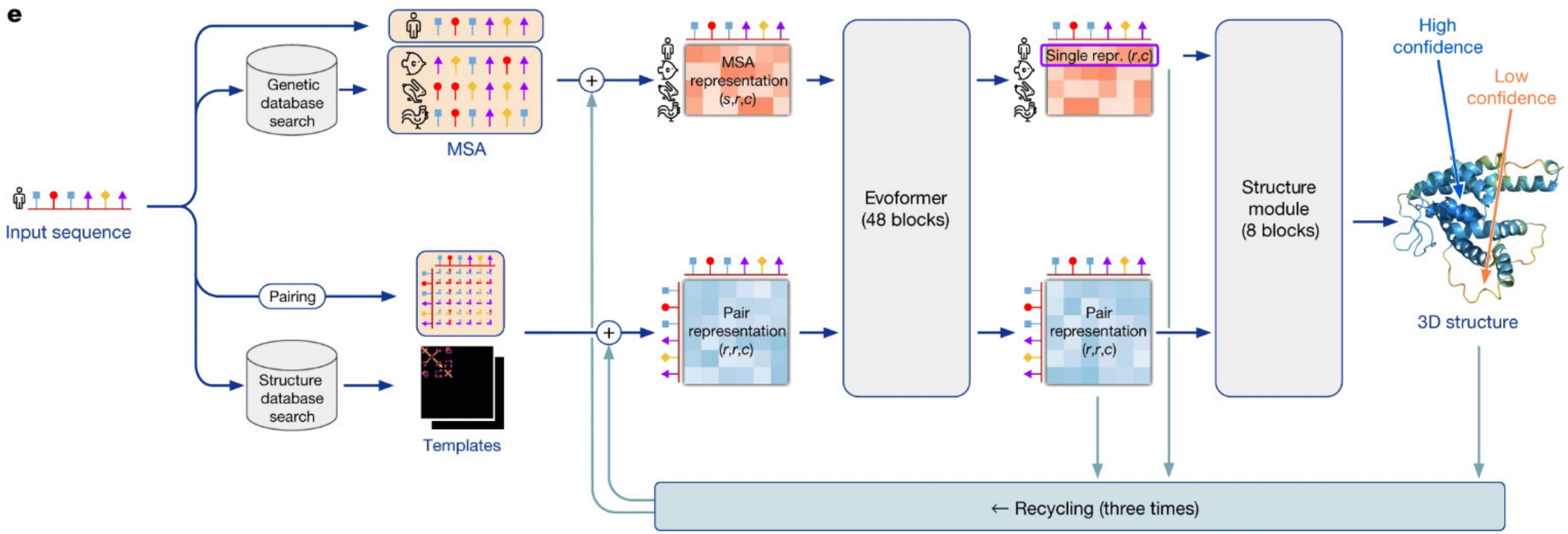
```
CLUSTAL W (1.83) multiple sequence alignment


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

Model building



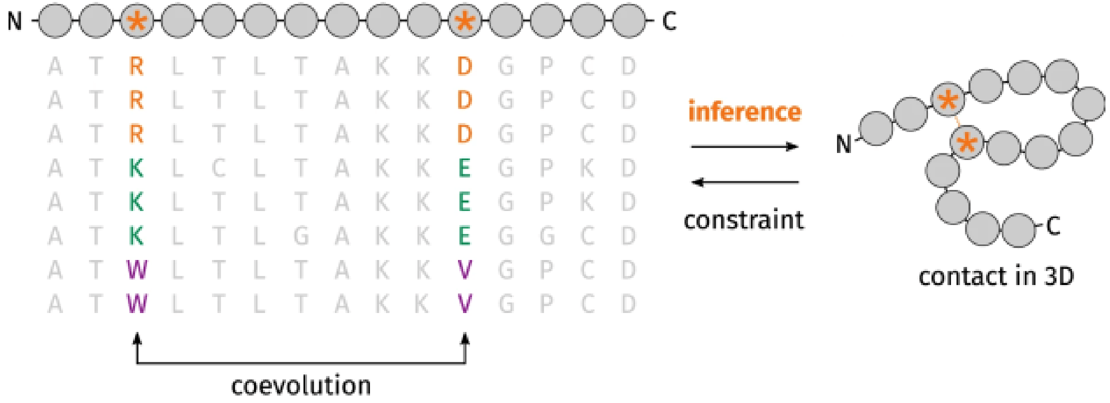
source: SBCB, Oxford, UK

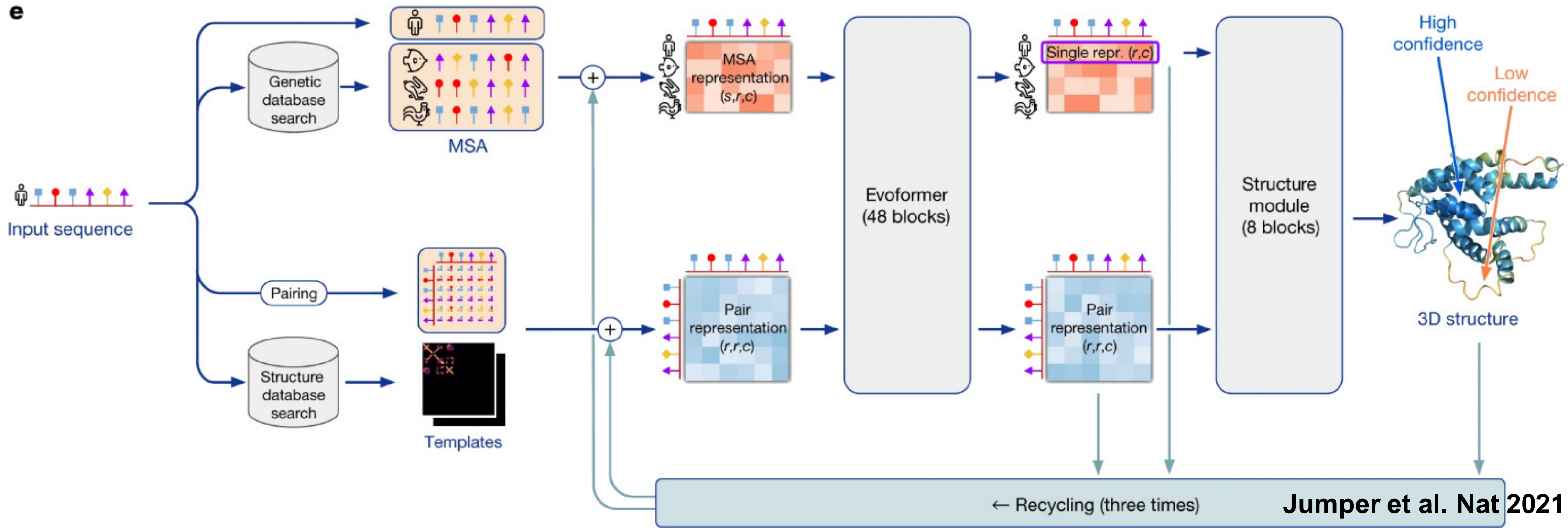
AlphaFold from DeepMind


e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Jumper et al. Nat 2021

AlphaFold from DeepMind

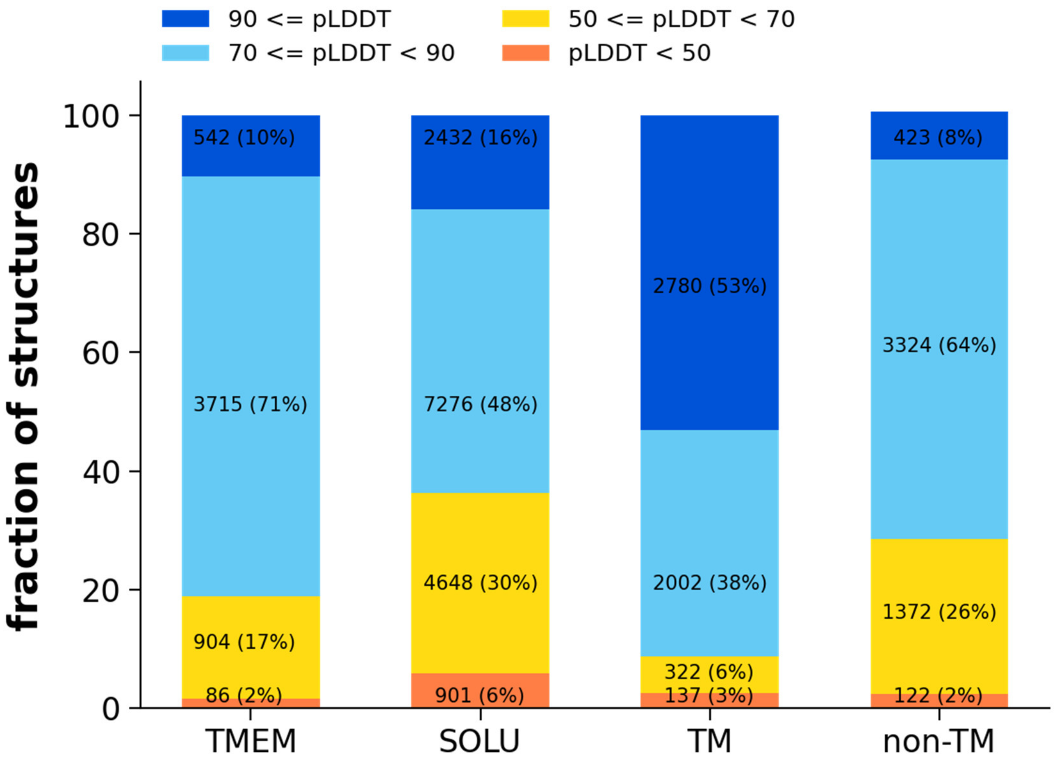

N
C
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
coevolution
inference
constraint
N
C
contact in 3D


e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

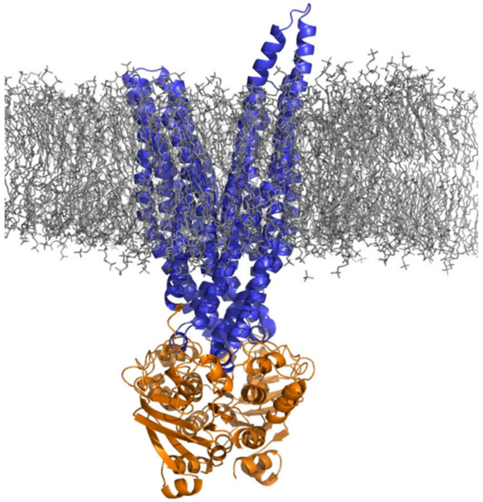
Jumper et al. Nat 2021

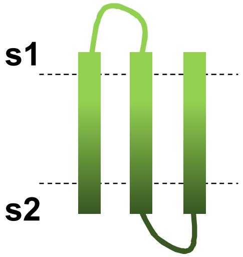
TM protein structure prediction by AF2

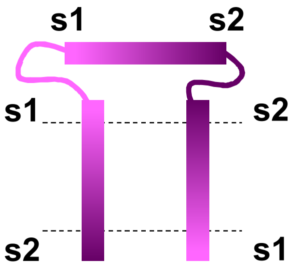
Hegedus *et al.* Cell Mol Life Sci. 2022

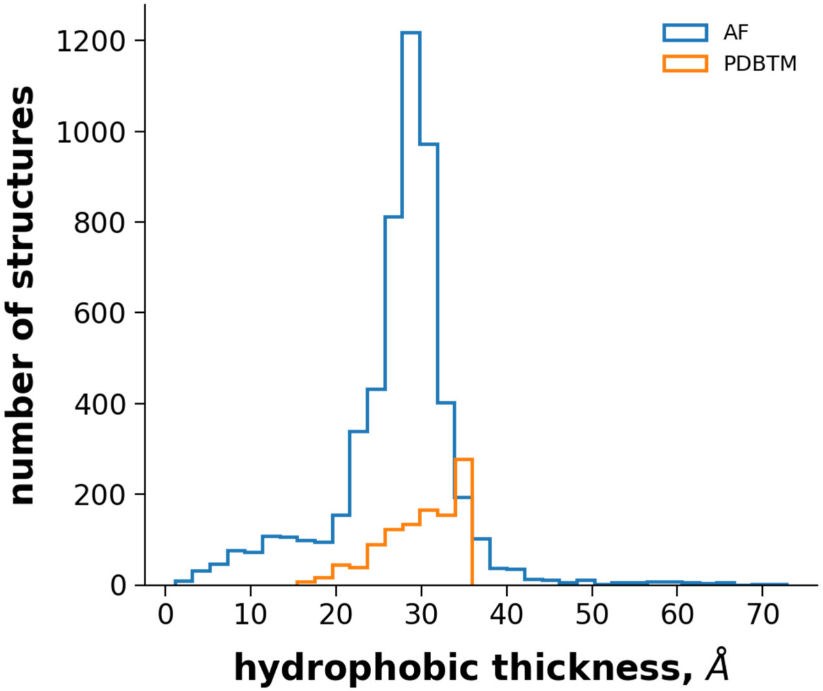


TM protein structure prediction by AF2

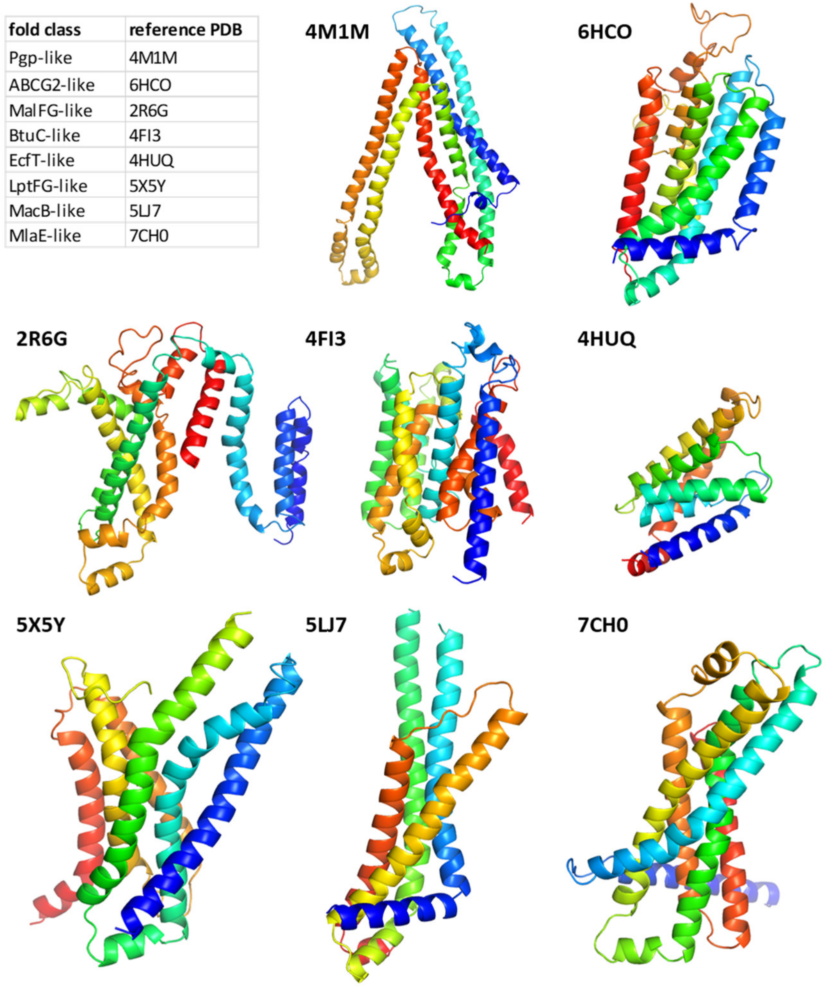




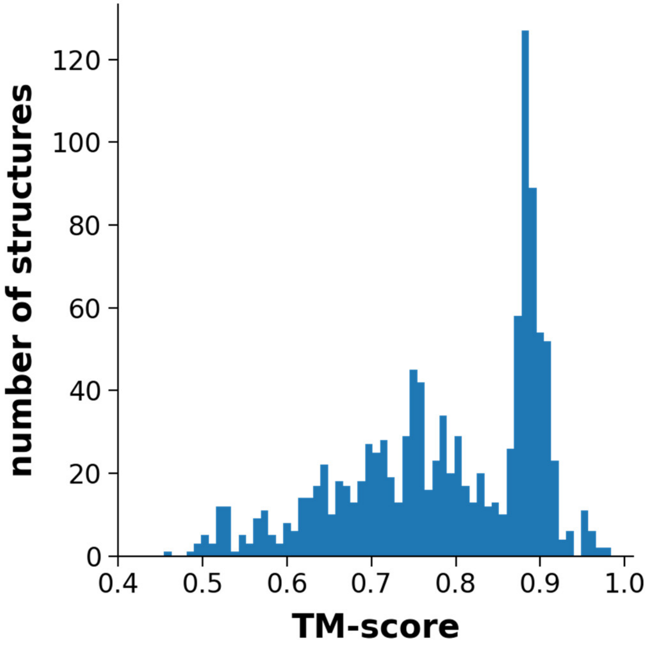




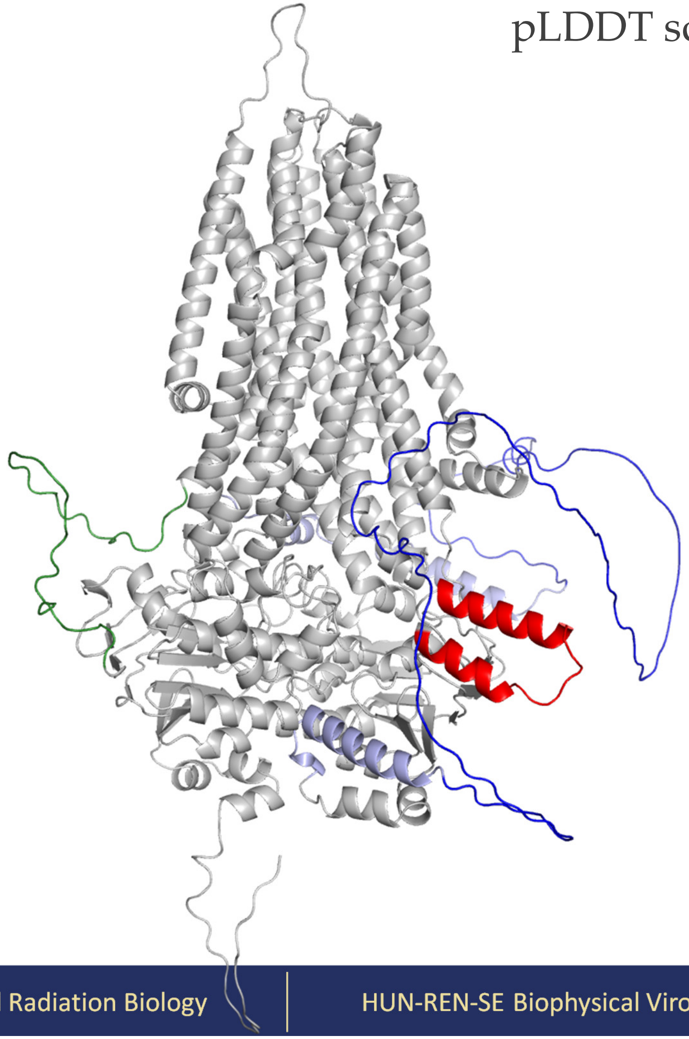
ABC protein folds



fold class	reference PDB
Pgp-like	4M1M
ABCG2-like	6HCO
MalFG-like	2R6G
BtuC-like	4FI3
EcfT-like	4HUQ
LptFG-like	5X5Y
MacB-like	5LJ7
MlaE-like	7CH0

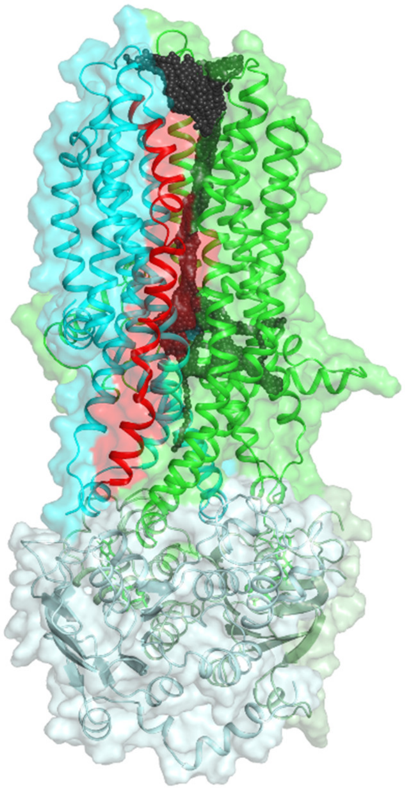


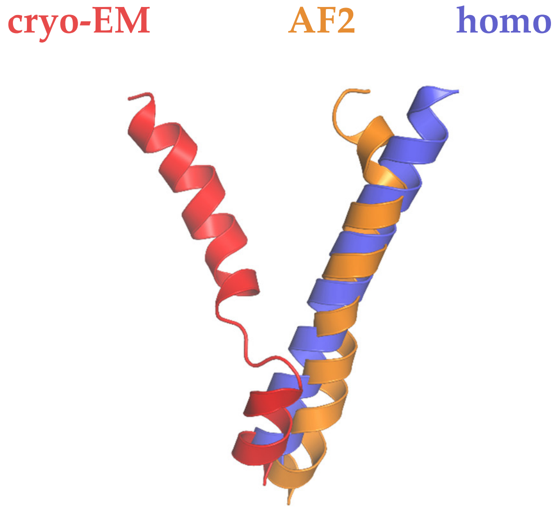
AlphaFold – TM – disorder

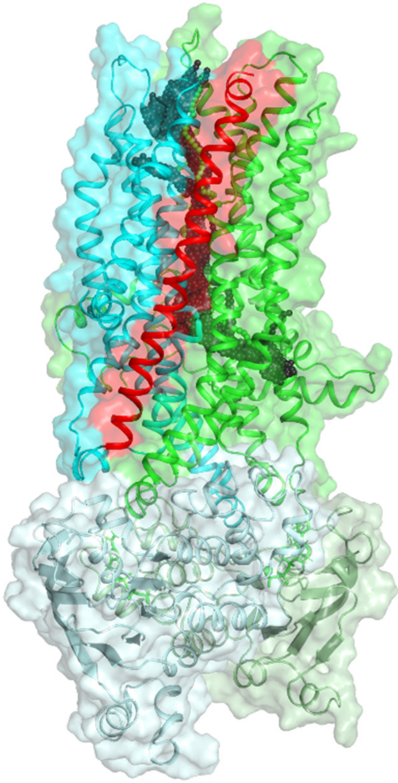


pLDDT score - IDR prediction

CFTR TM8



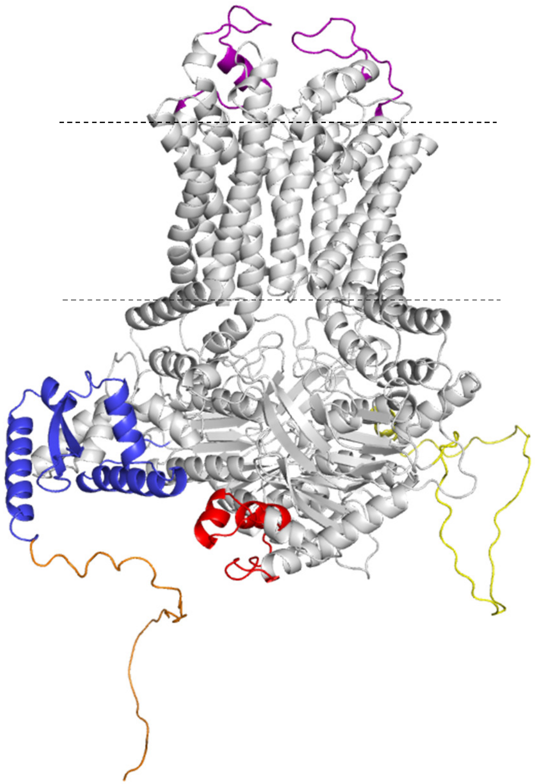


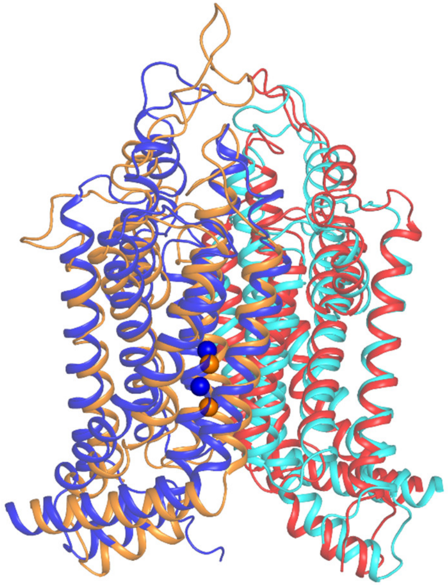


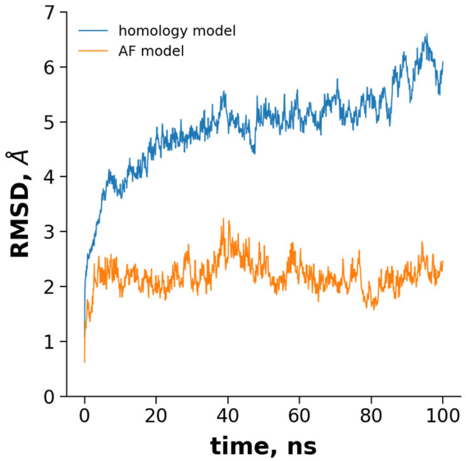
cryo-EM
1/22 simulations
0.36%

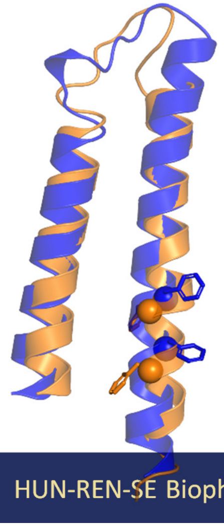
homo
5/6 simulations
3.74%

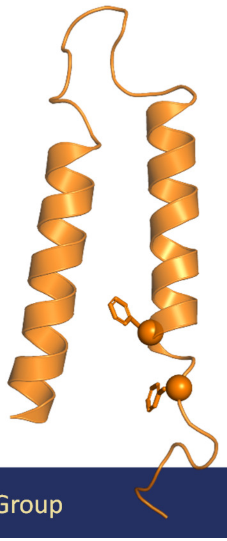
A plant transporter, AtABCG36



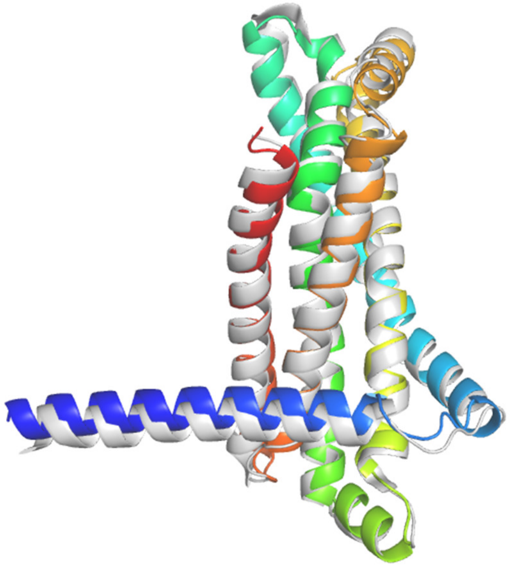






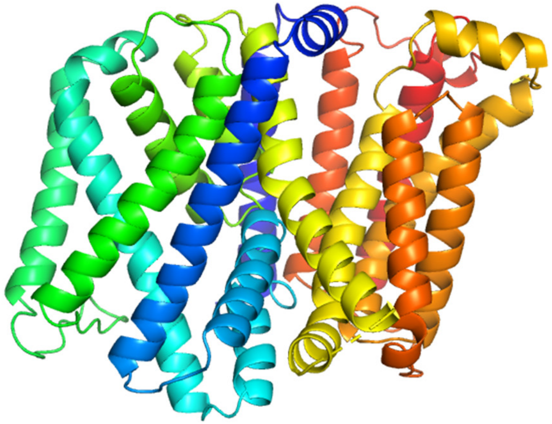


Prediction of new TM folds

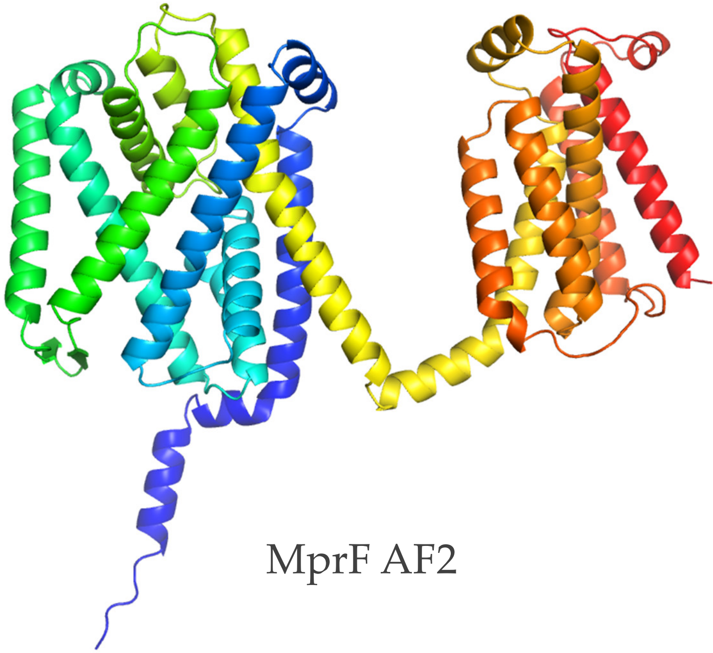


MlaE-like fold
PDBID: 7ch0
RMSD of 1.28 Å

ER membrane protein complex subunit 6
PDBID: 6ww7
RMSD of 0.96 Å

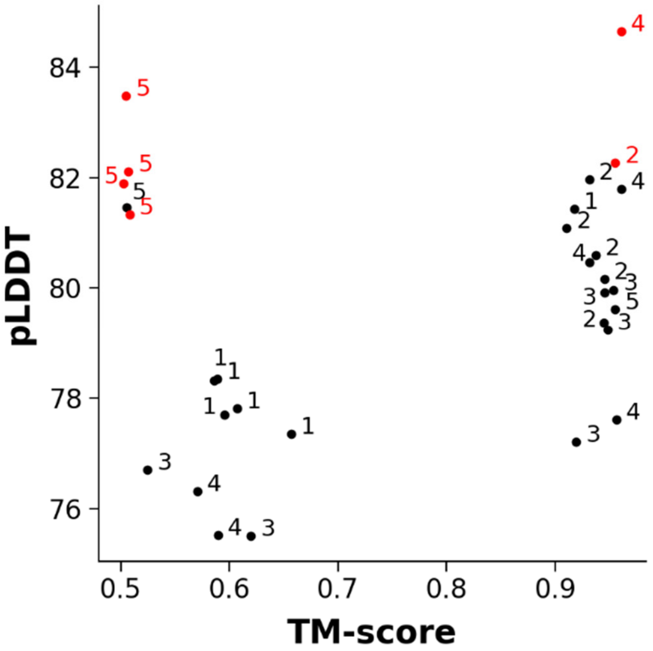


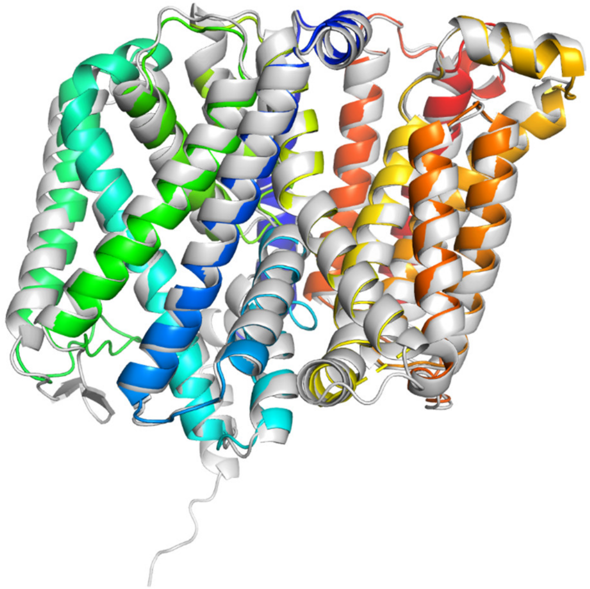
MprF (PDBID: 7DUW)



MprF AF2

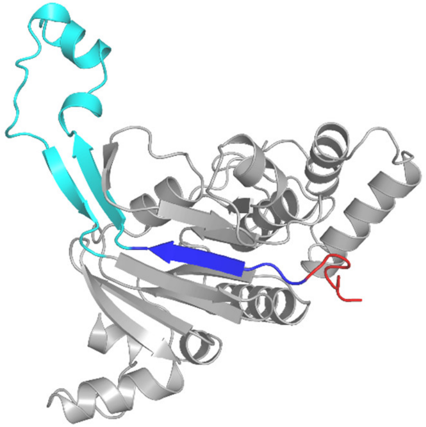
Prediction of MprF





AF2 corrects an experimental structure

```
G2 6HCO AVLSFHNICY ┐
G8 5DO7 NSLYFTYSGQ ┘ ✗
G2 6HCO AVLSFHNICY ┐
G8 seq  NTLEVRDLNY ┘ ✓
G2 6HCO AVLSFHNICY ┐
G8 AF   NTLEVRDLNY ┘ ✓
```



Protein-protein interactions

Docking of proteins – challenging (surface shape, dynamics)

PISA - Protein Interfaces, Surfaces and Assemblies

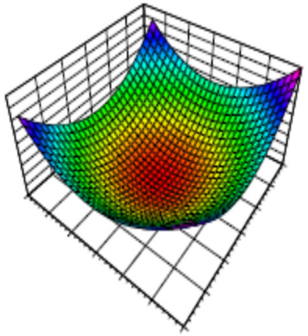
Molecular Dynamics

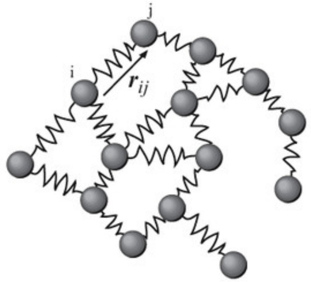
AlphaFold2-Multimer

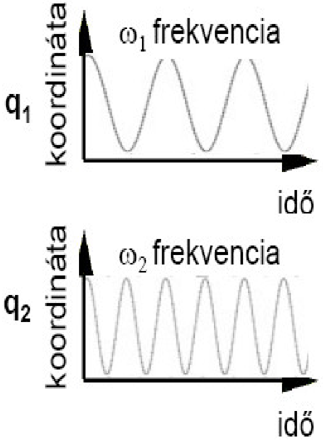
Methods for studying protein dynamics

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)

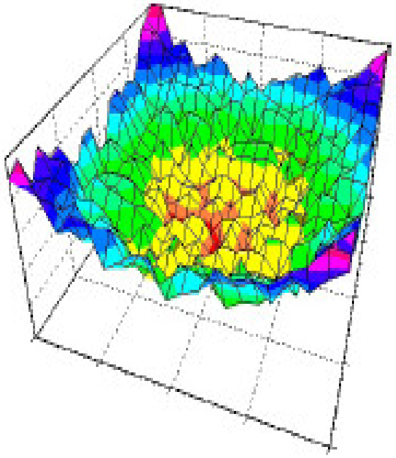
  directionality by projection of motions
  to a mode space of N dimensions

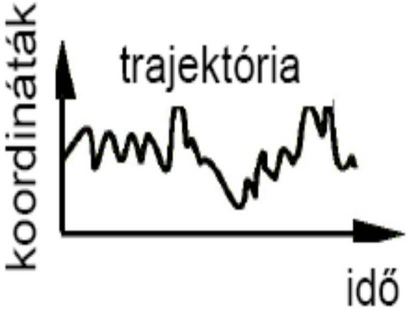
Tools: http://prody.csb.pitt.edu

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

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

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

Lazaridis (2003)

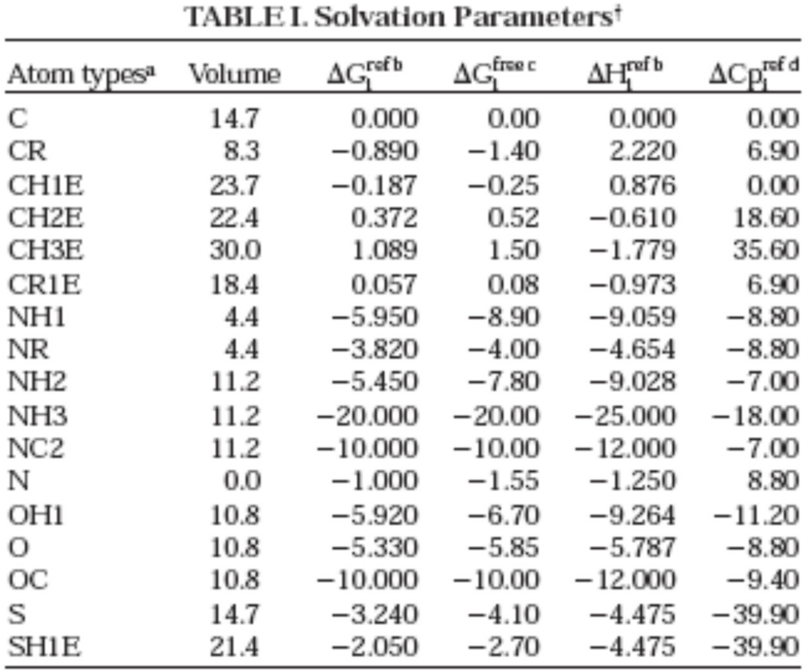


TABLE I. Solvation Parameters[†]

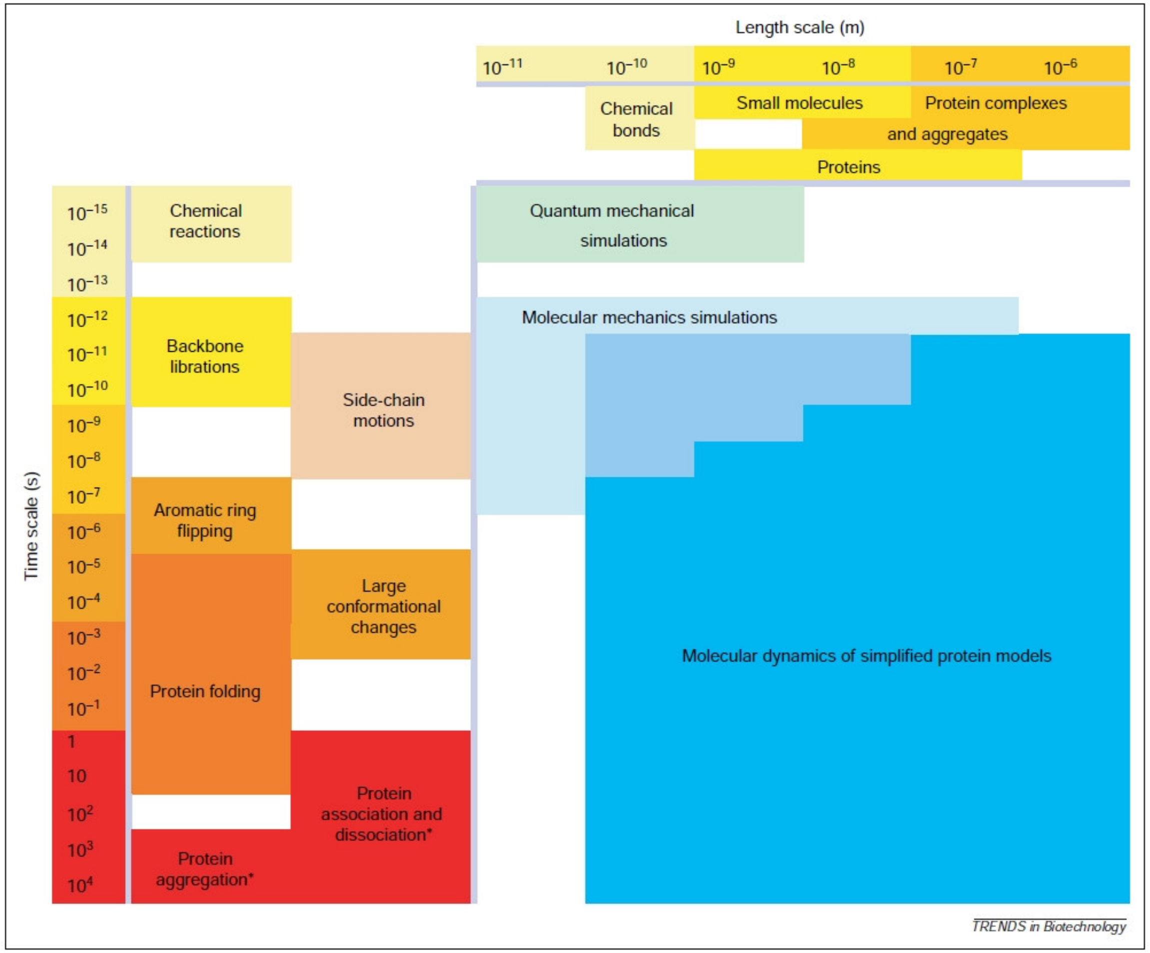
Atom types[a]	Volume	$\Delta G_i^{\mathrm{ref}\,b}$	$\Delta G_i^{\mathrm{free}\,c}$	$\Delta H_i^{\mathrm{ref}\,b}$	$\Delta Cp_i^{\mathrm{ref}\,d}$
C	14.7	0.000	0.00	0.000	0.00
CR	8.3	−0.890	−1.40	2.220	6.90
CH1E	23.7	−0.187	−0.25	0.876	0.00
CH2E	22.4	0.372	0.52	−0.610	18.60
CH3E	30.0	1.089	1.50	−1.779	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

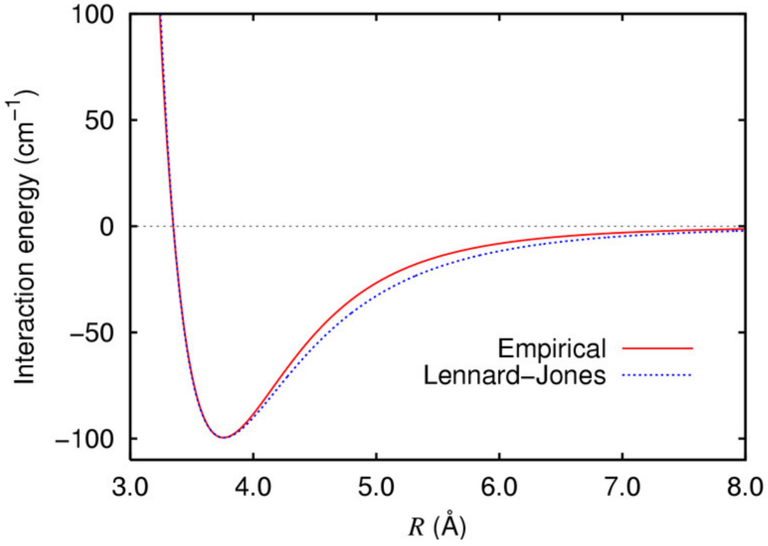
- time (computation time versus real time)
- calculation of the potential is the bottle-neck
- fs long integration steps
- „periodic boundary condition”
- solvent (explicit/implicit)

The time scale of various molecular events

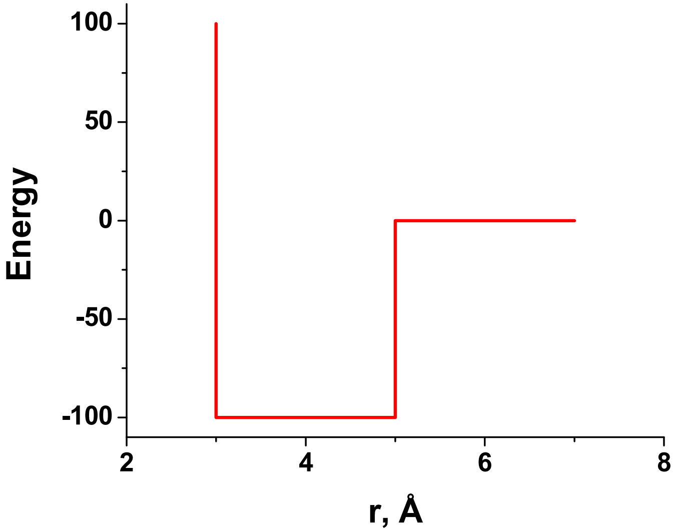


F. Ding and N.V. Dokholyan
TRENDS in Biotechnology, 2005

Discrete Molecular Dynamics (DMD)



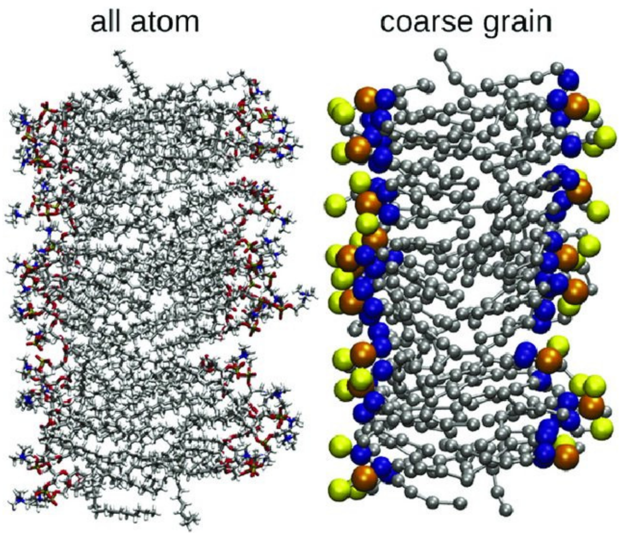
wikipedia



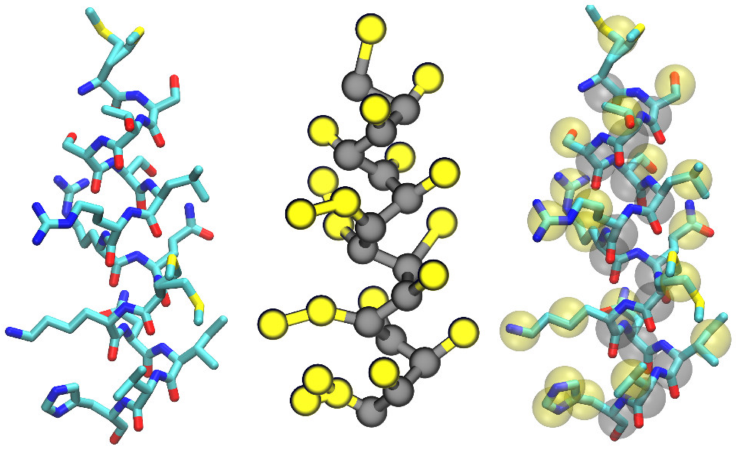
Ding, F., Dokholyan, N. V. PLoS Comput Biol 2:e85

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

Simplified coarse-grained models



Awoonor-Williams and Rowley BBA 2015

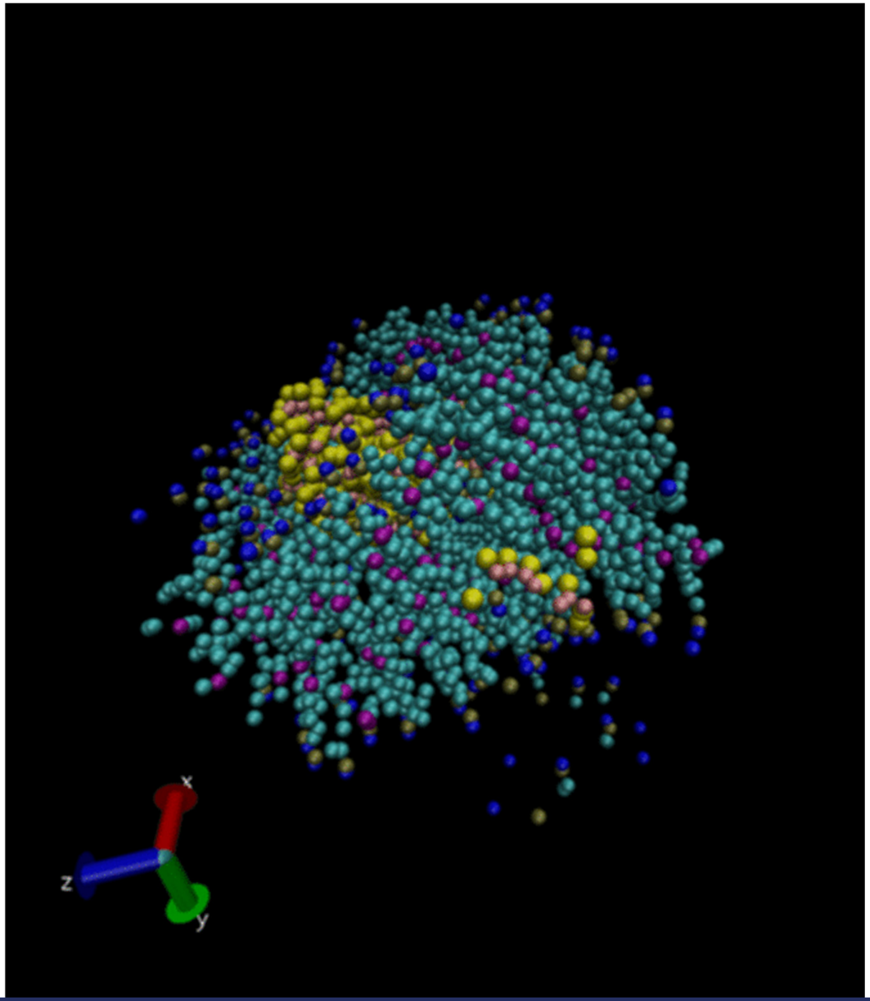


Bradley and Radhakrishnan Polymers 2013

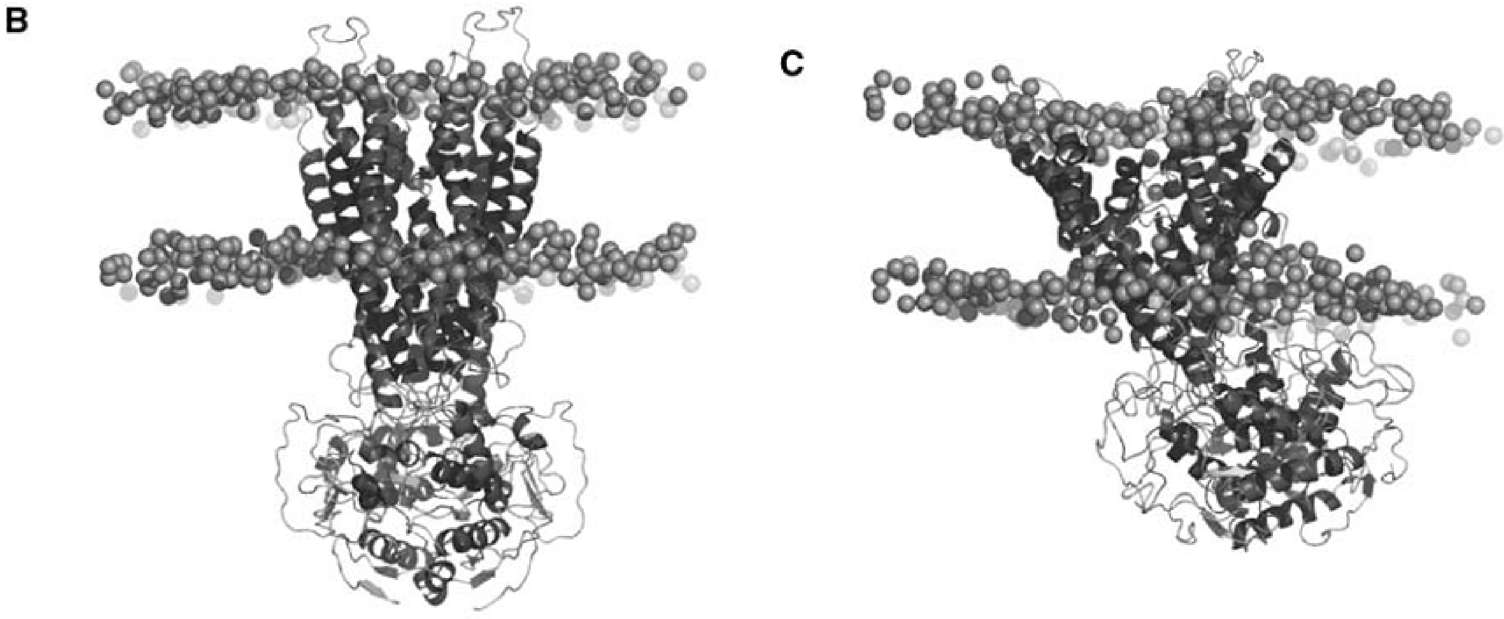
AT:	atomistic, all-atom
CG:	coarse grained

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

Membrane bilayer formation



Stability of simulations

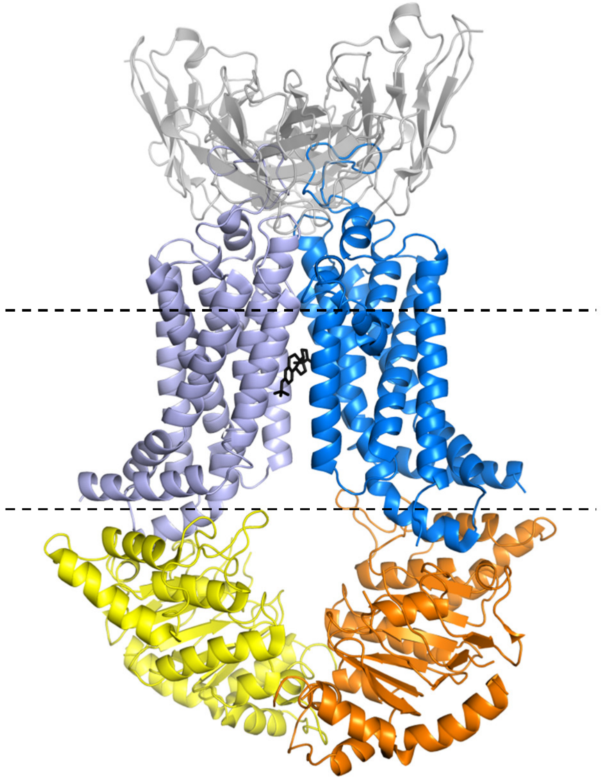


0 ns

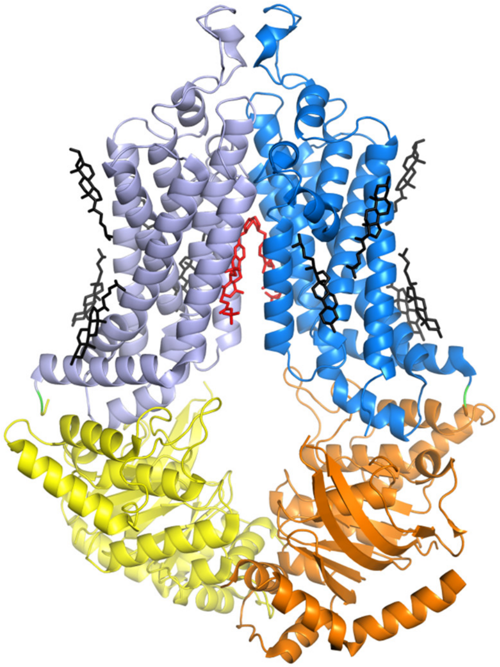
20 ns

Eur Biophys J (2008) 37:403–409

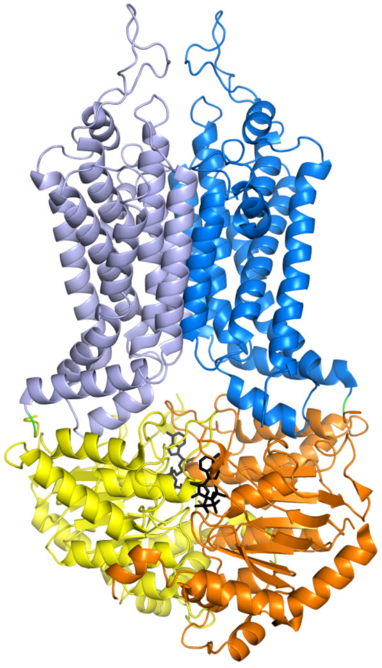
ABCG2 structures



6HCO

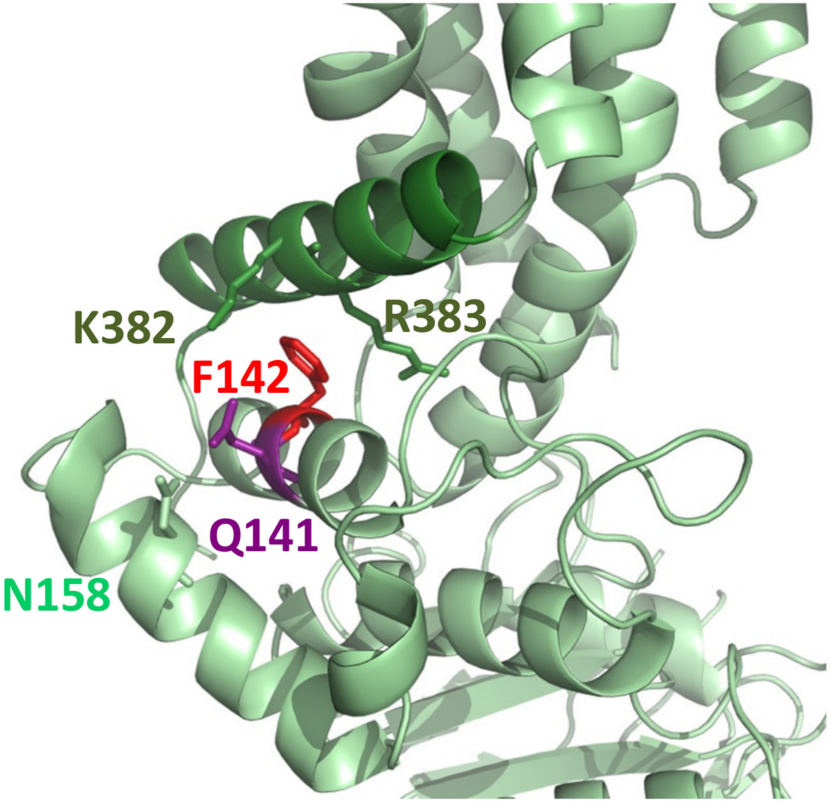


6HIJ



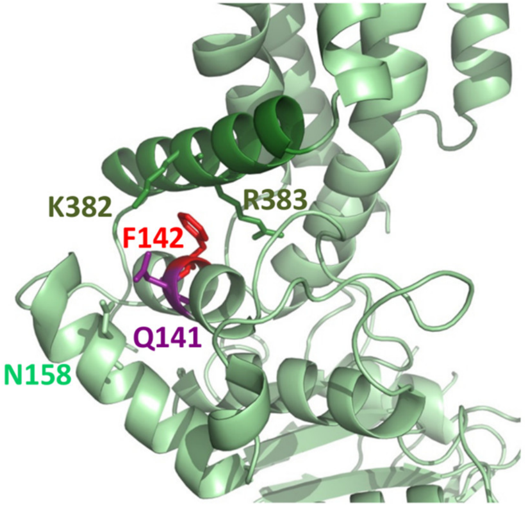
6HZM

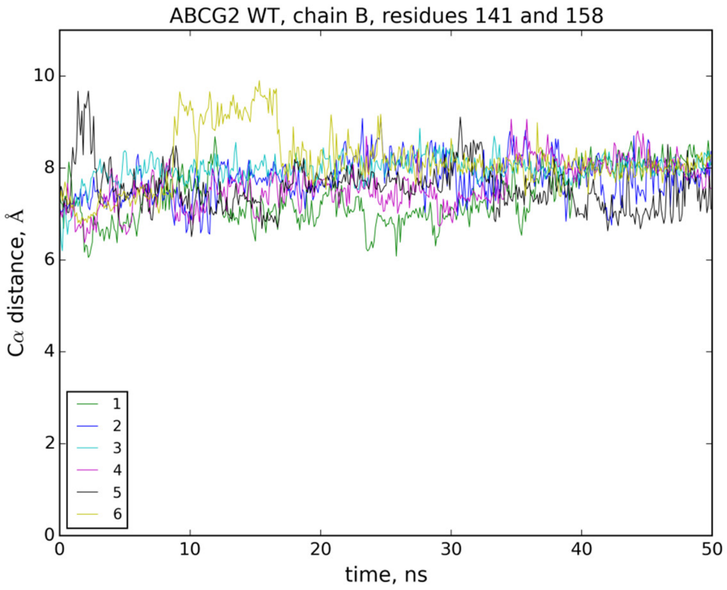
The Q141 position

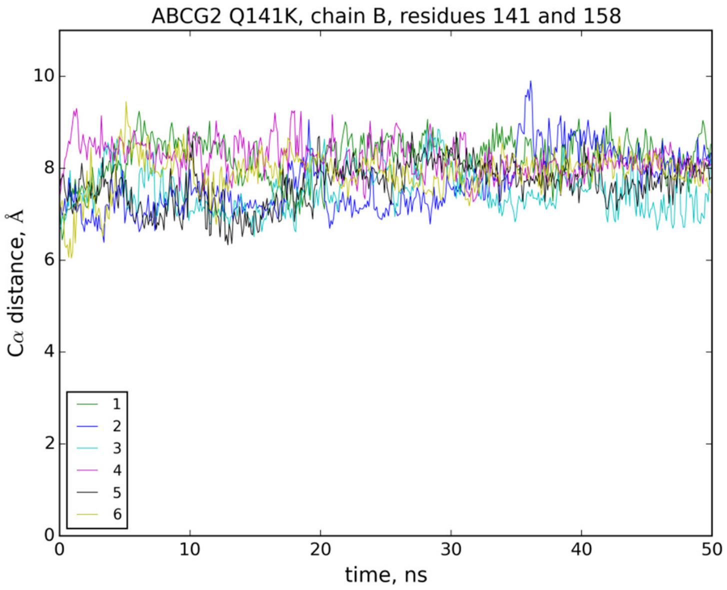


- The protein was embedded in POPC bilayer
- Optimizing the orientation of water, lipids, amino acid side chains:
  - energy minimization
  - equilibration
  - minimal backbone motions (position constrains)
- Production run
  - no constraints
  - 50 ns x 6 = 300 ns
- Comparing WT és mutants (e.g. Q141K,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

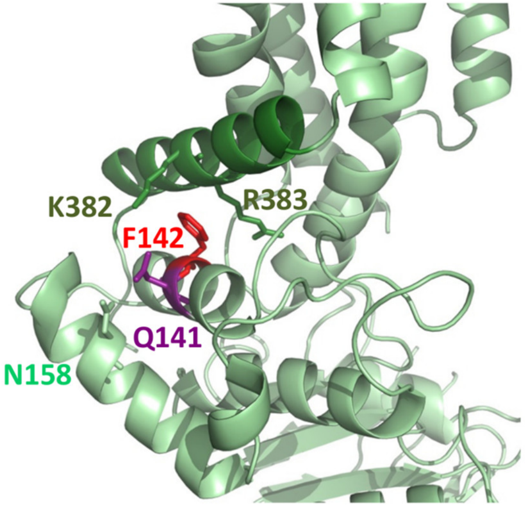
The effect of Q141K on protein dynamics

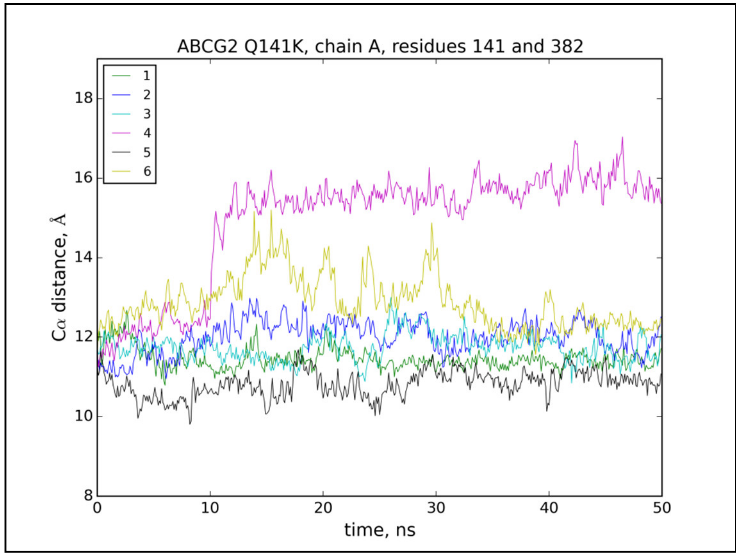


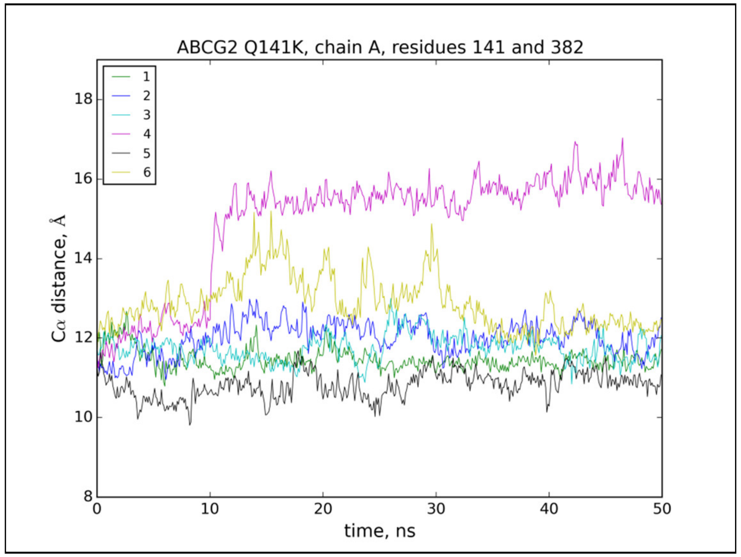




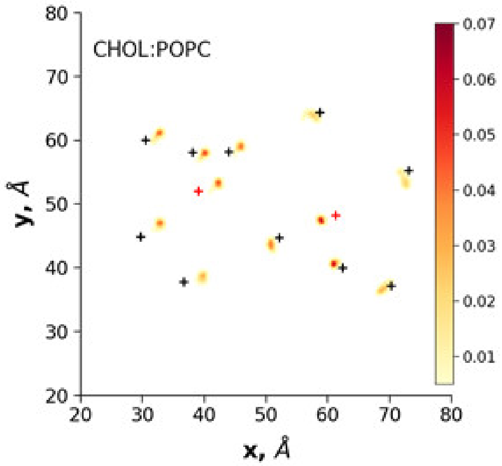
The effect of Q141K on protein dynamics

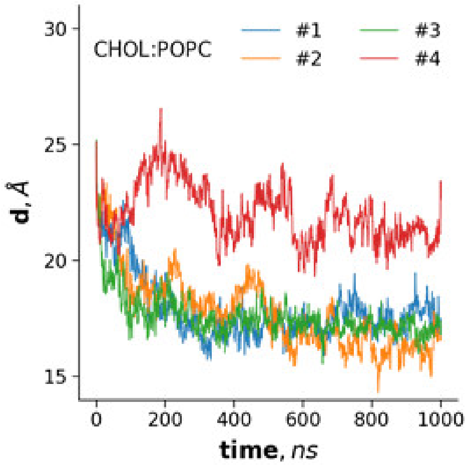


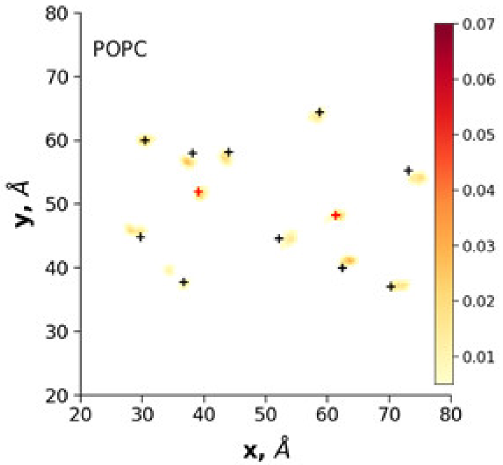


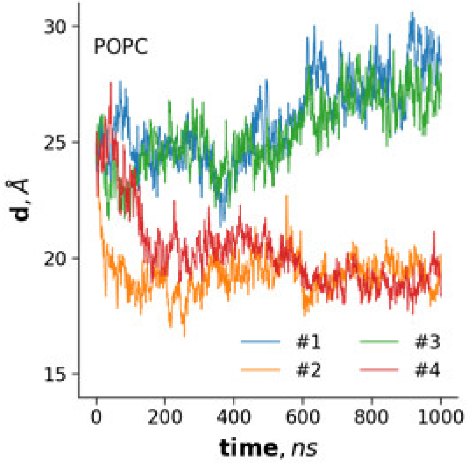


The effect of cholesterol on ABCG2





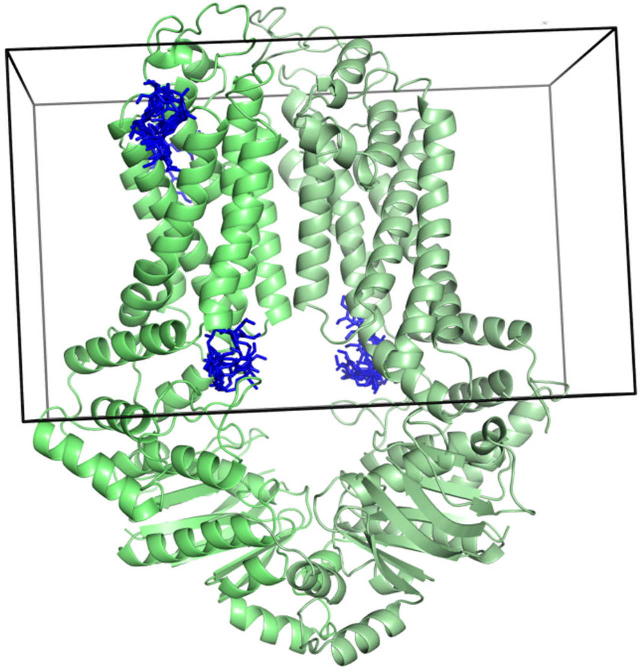




Identification of drug binding sites

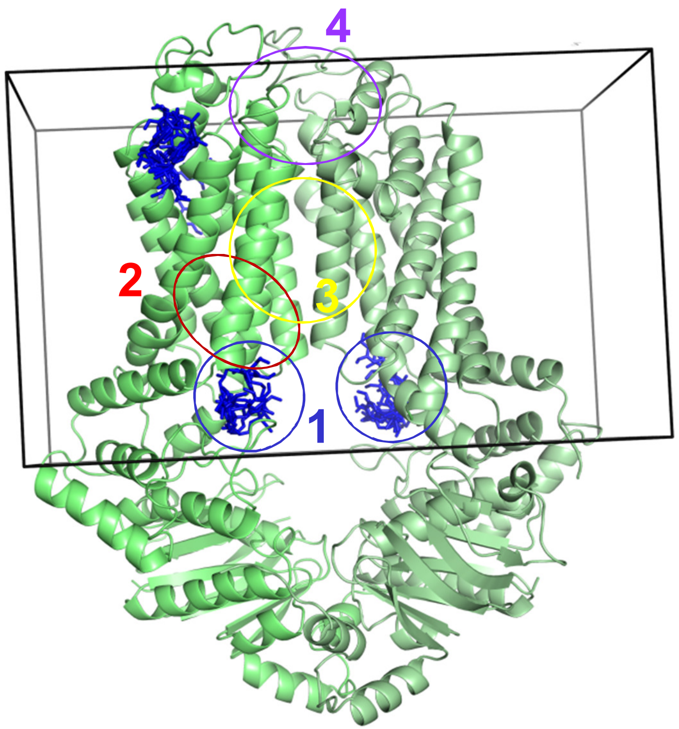
***in silico* docking, AutoDock Vina**

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

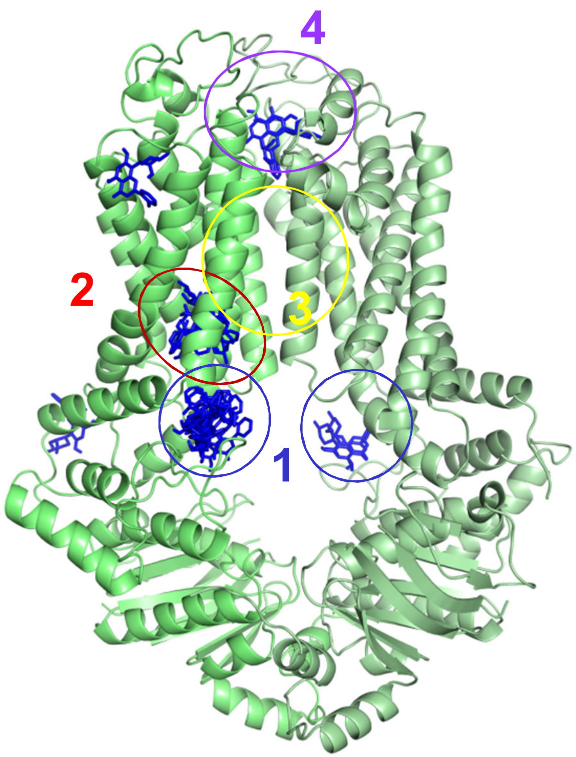


Identification of drug binding sites

verapamil

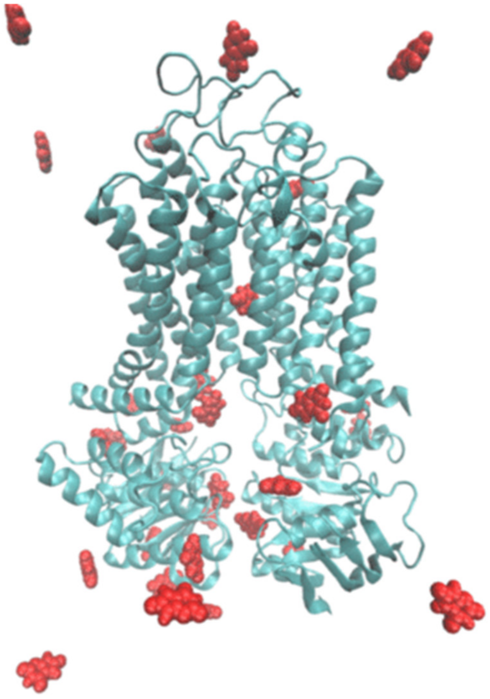


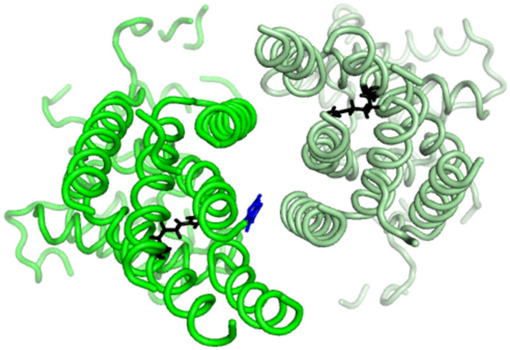
flavopiridol



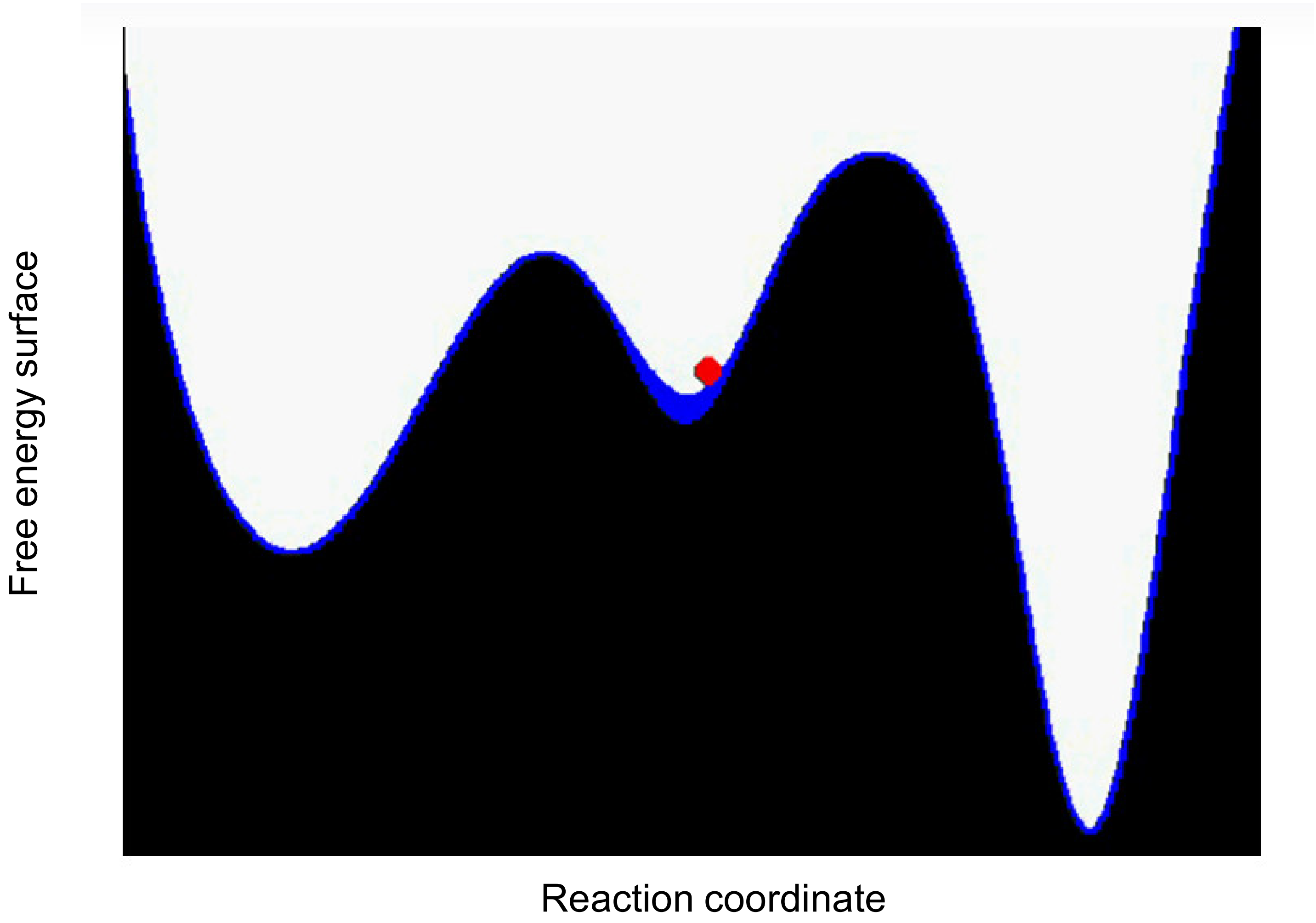
Describing the transport using MD

equilibrium simulations, uric acid molecules



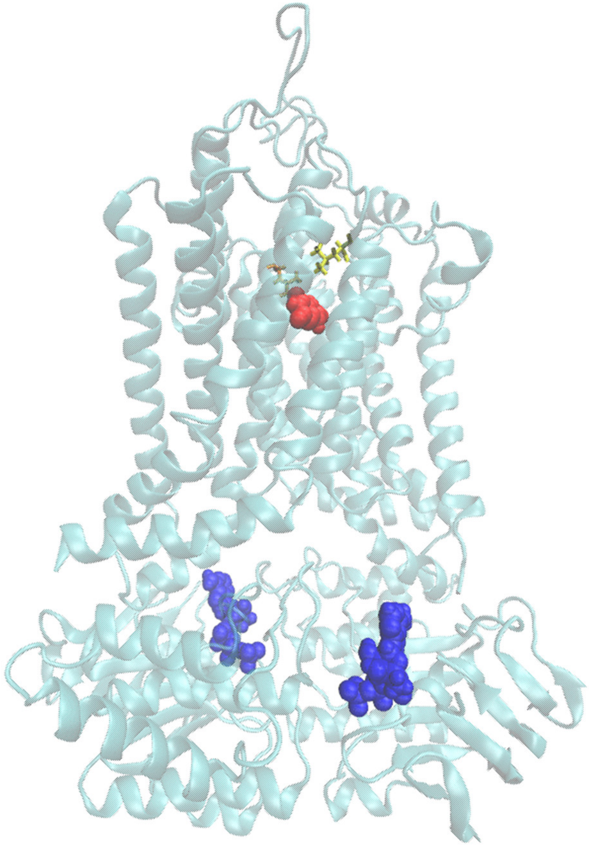


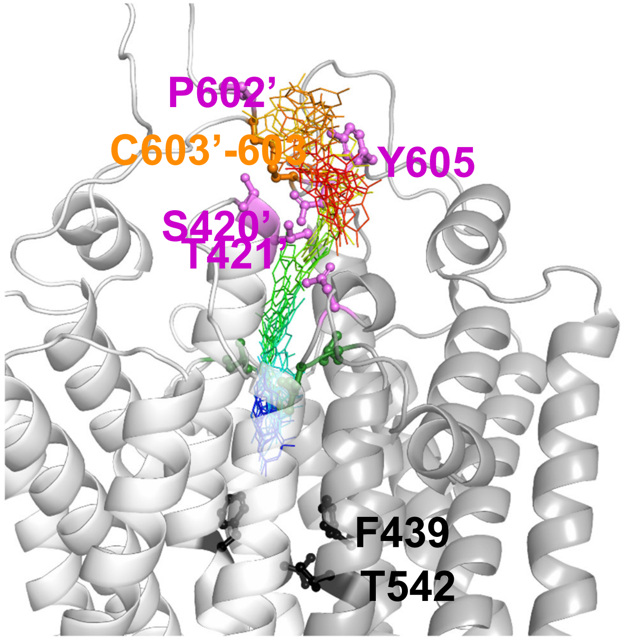
Exploring substrate transport by biased MD simulations

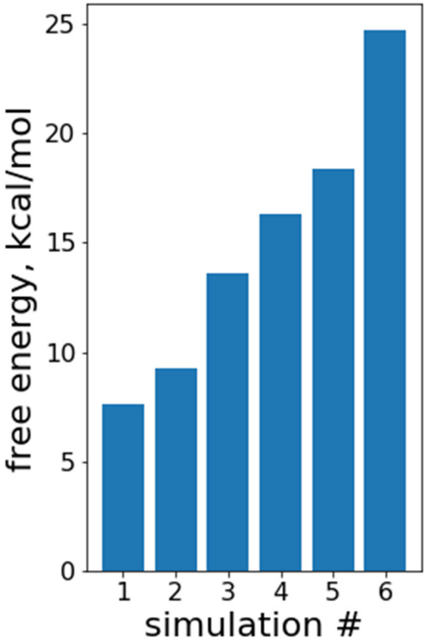


Simplified coarse-grained models

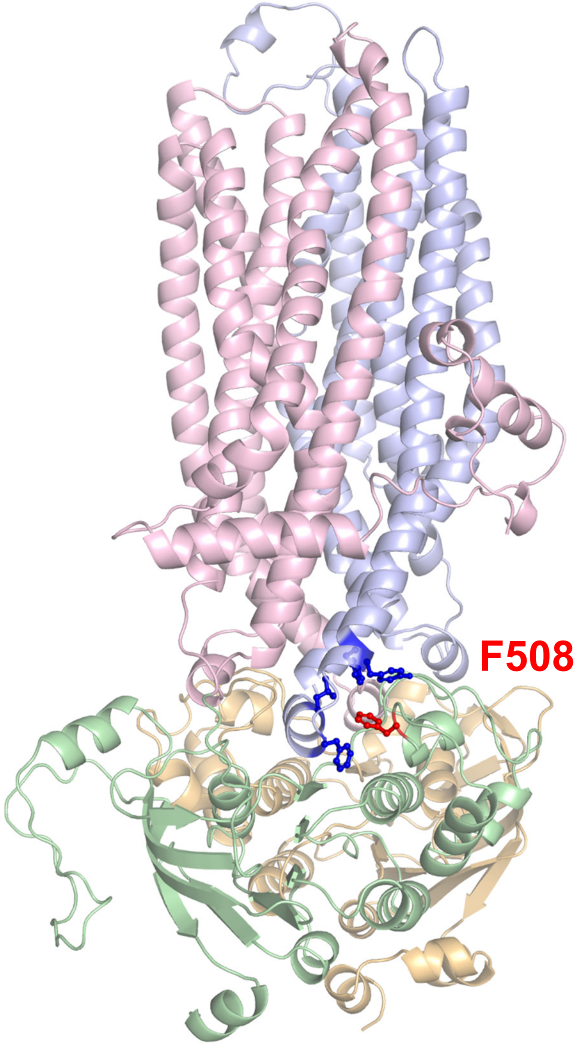
metadynamics simulations, uric acid molecule







CFTR / ΔF508 mutation



Many experimental and computational studies

Domain folding
Domain stability
Domain-domain assembly

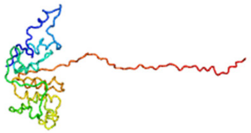
Transmission of the consequence of a mutation; allosteric propagation of alterations in dynamics

CFTR / NBD1 folding

Padanyi *et al.* Cell Mol Life Sci. 2022

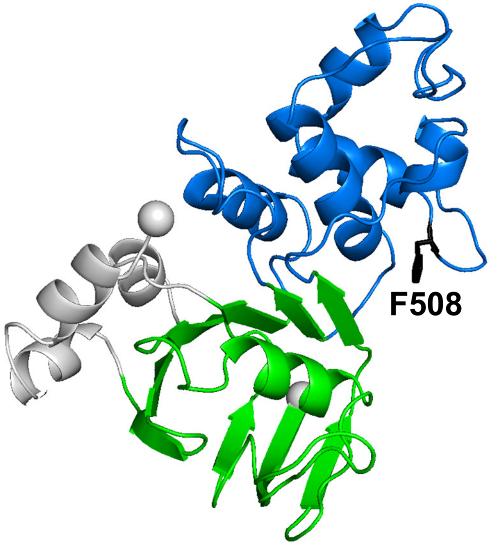
To learn folding
computationally
experimentally

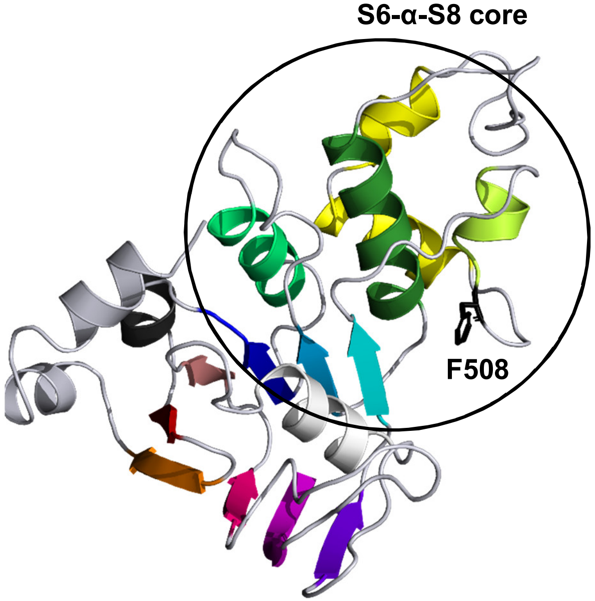
highly challenging

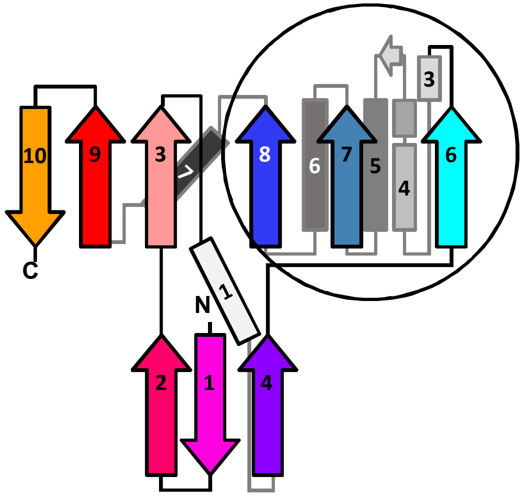


Unfolding
pulling molecular dynamics (MD) simulations
atomic force microscopy (AFM) experiments

NBD1 architecture

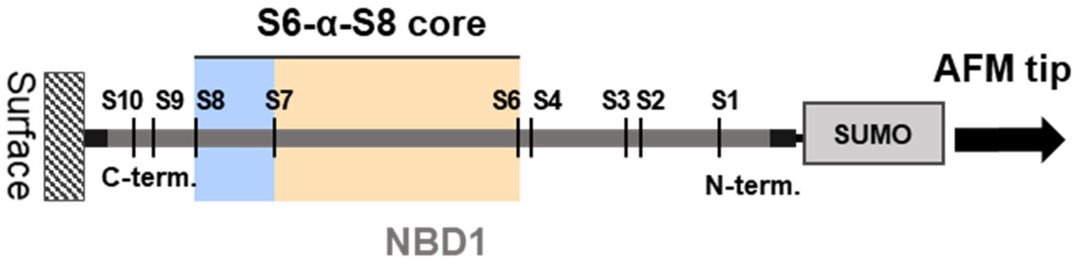


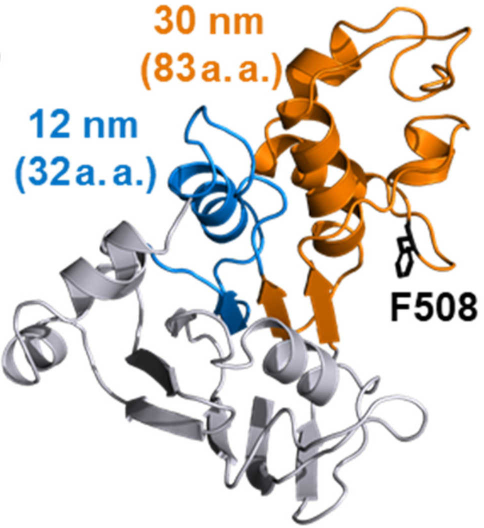


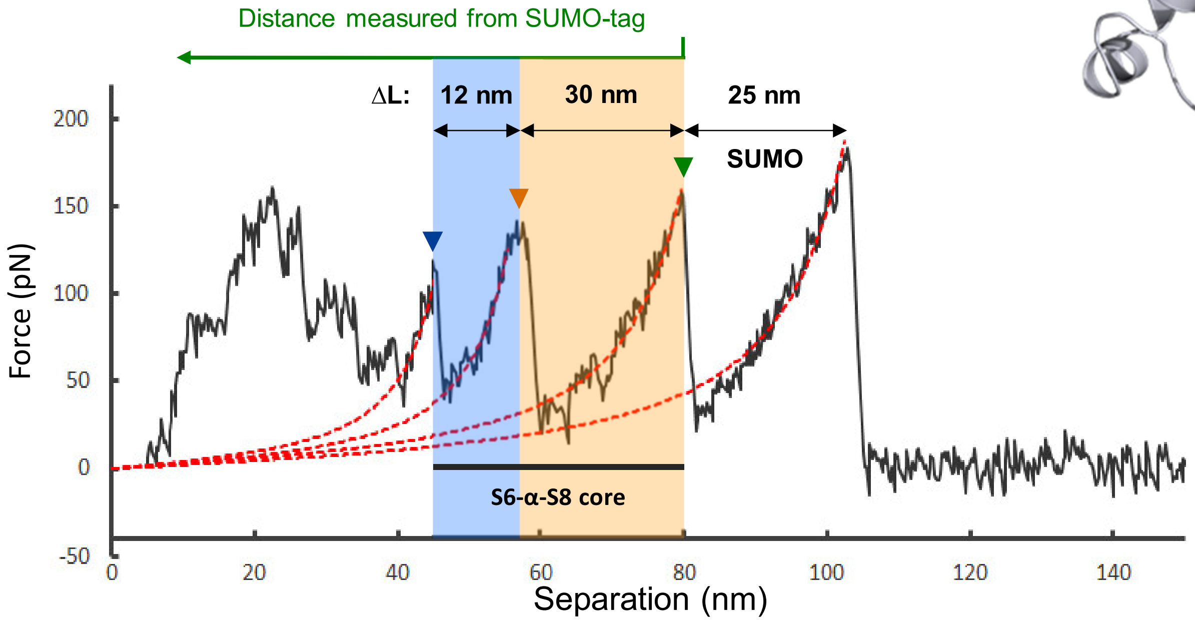


- α-subdomain – blue
- β-subdomain – green

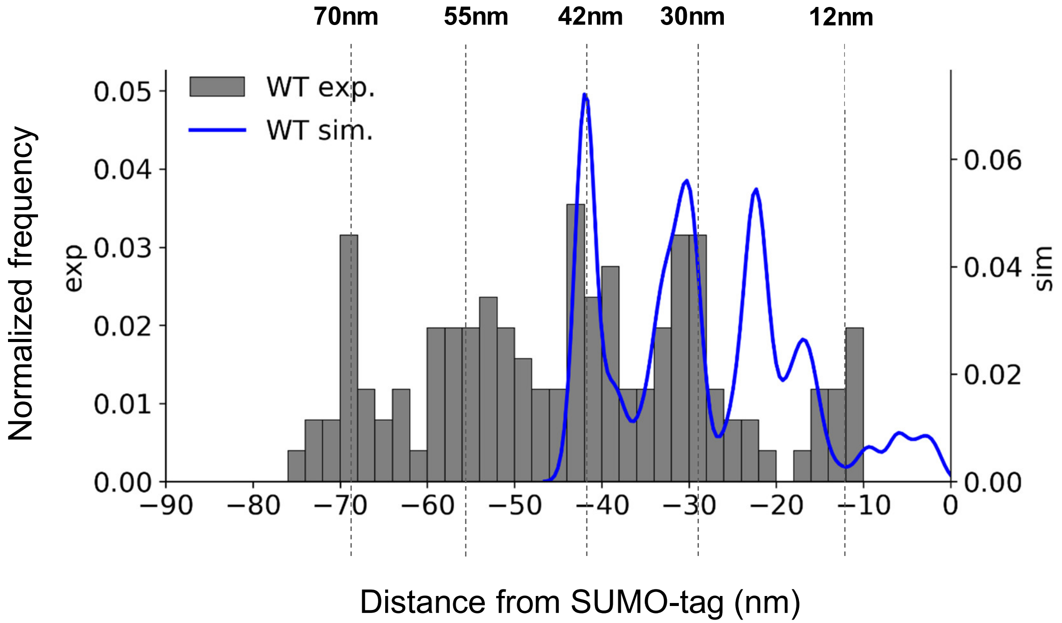
AFM experiments



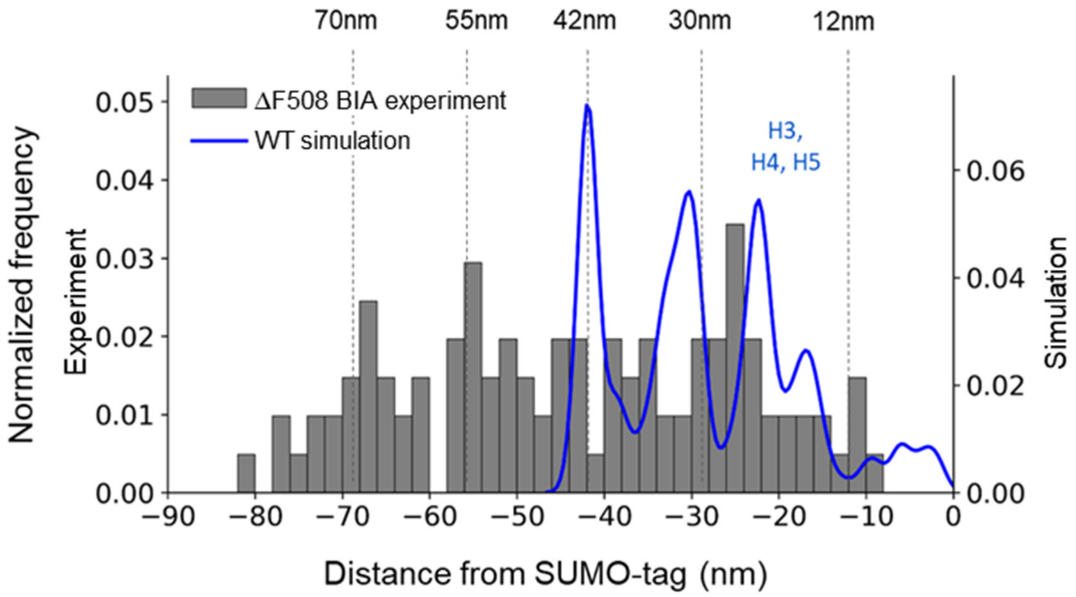


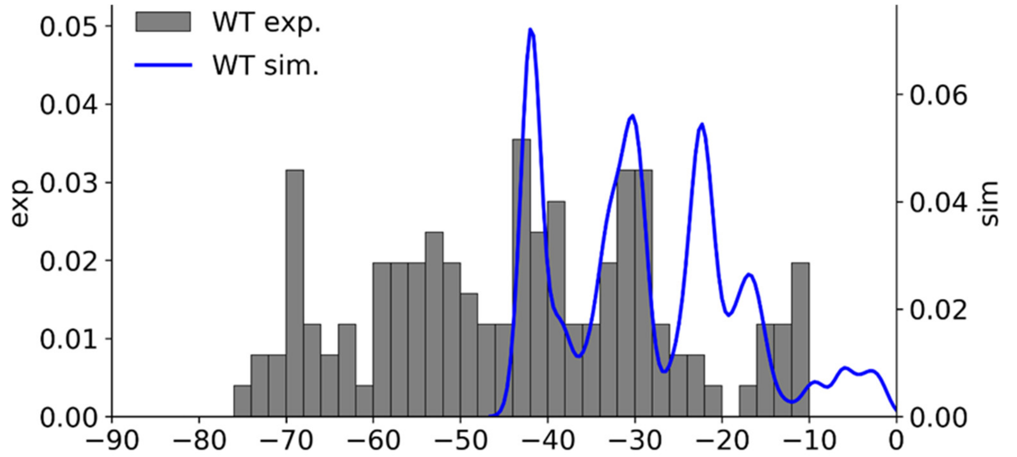


AFM experiments



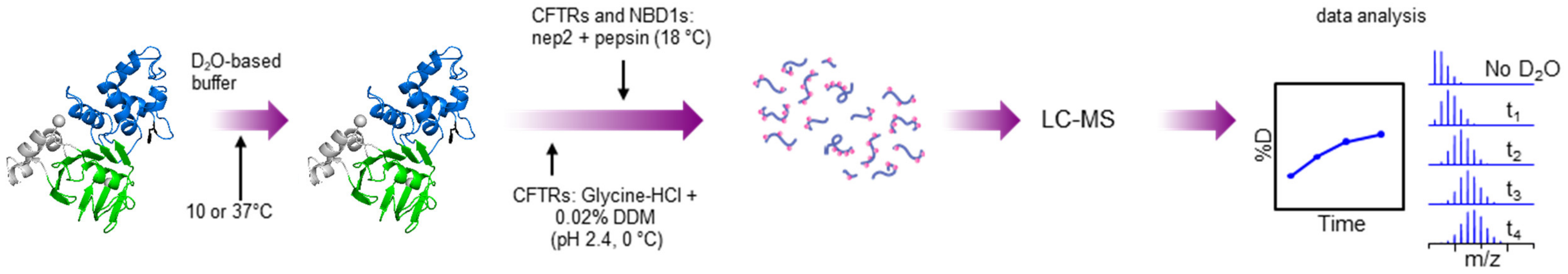
AFM experiments - ΔF508+BIA vs WT



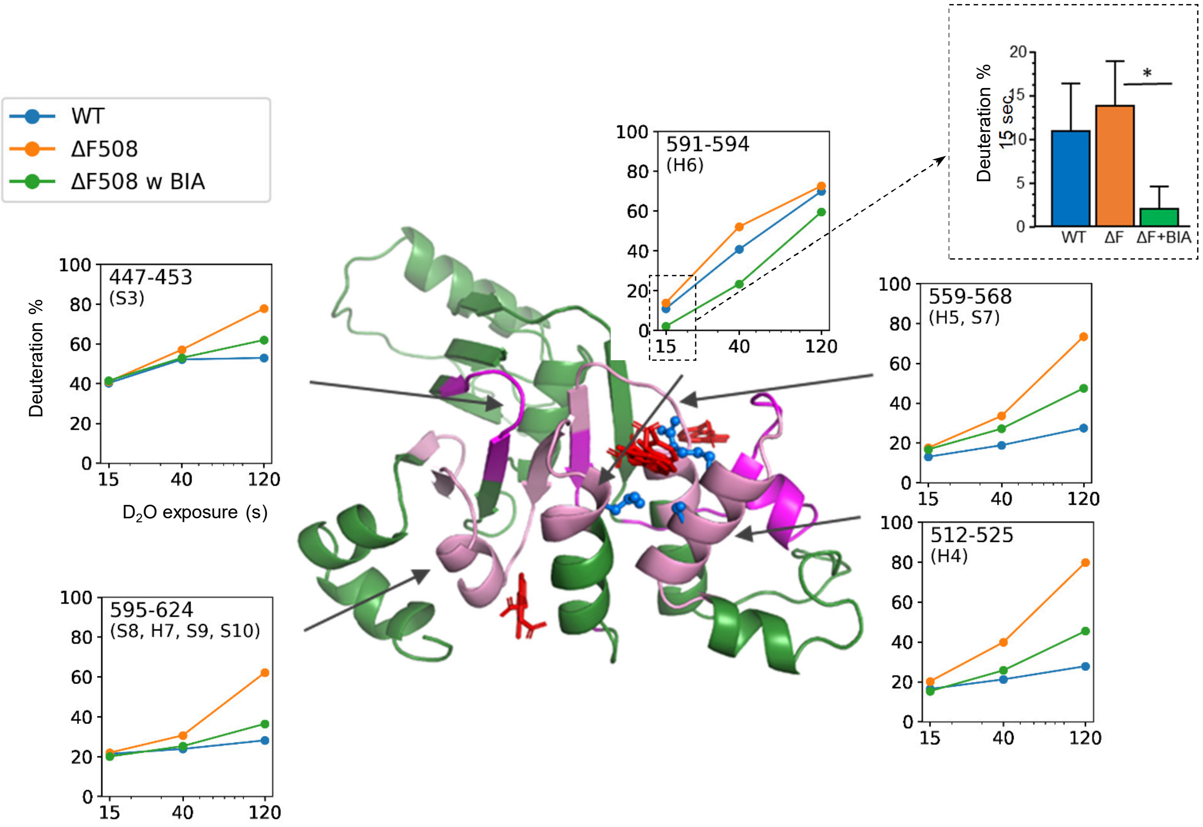


HDX experiments

Hydrogen-Deuterium Exchange
Gergely Lukács, McGill University, Montreal

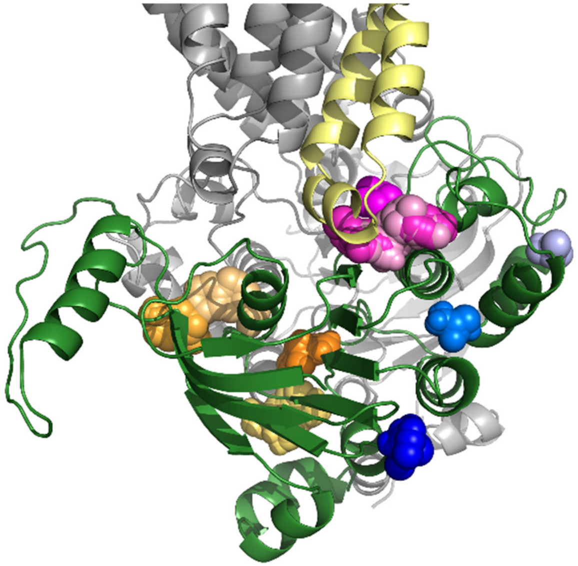


BIA binding site

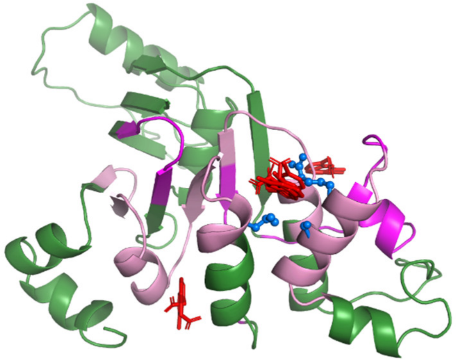


BIA binding site

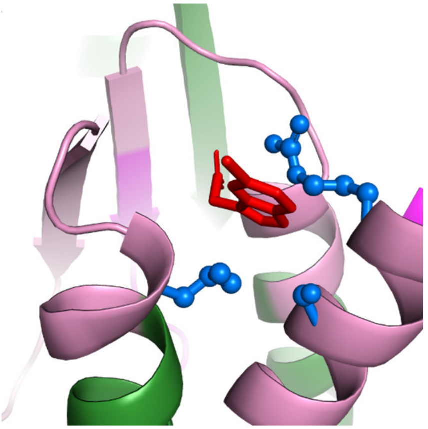
pocket detection
fpocket



docking
AutoDock Vina



concluded site

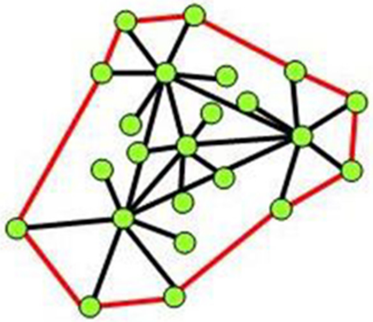


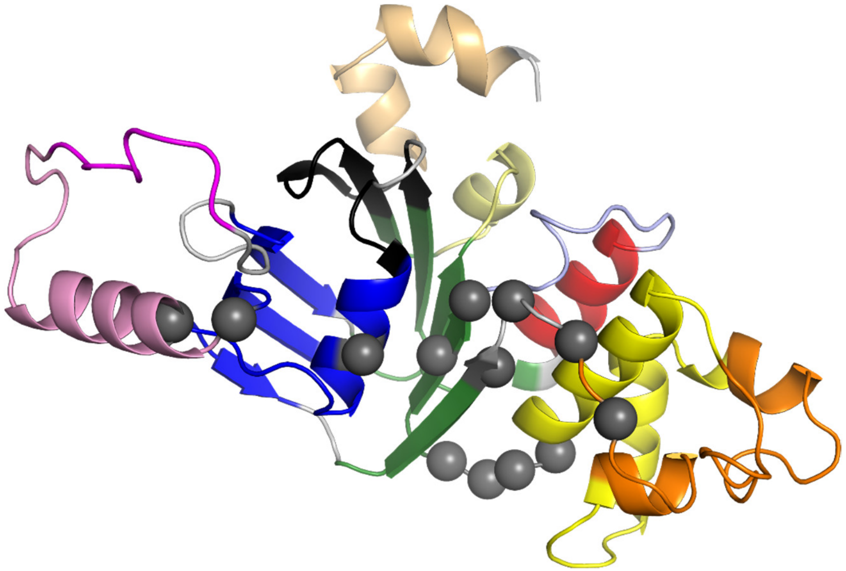
Allostery

- MD simulations
- pairwise correlated motion, c_{ij}
- network

  node – a.a.

  edge if $c_{ij} > 0$
- graph analysis

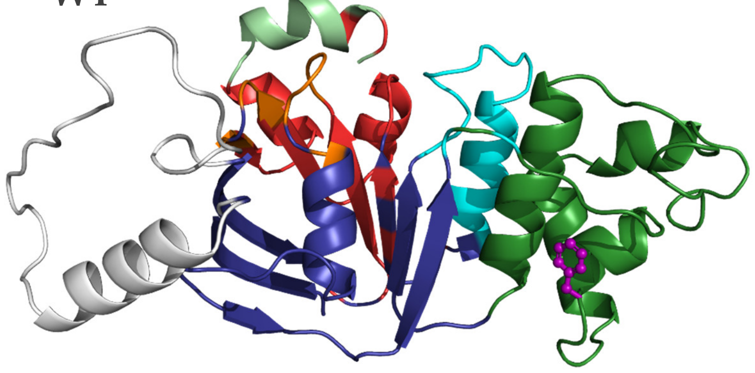




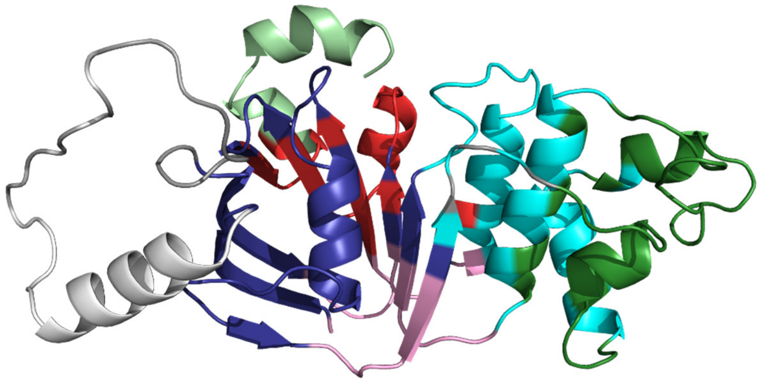
spheres: critical residues
betweenness centrality

Community analysis

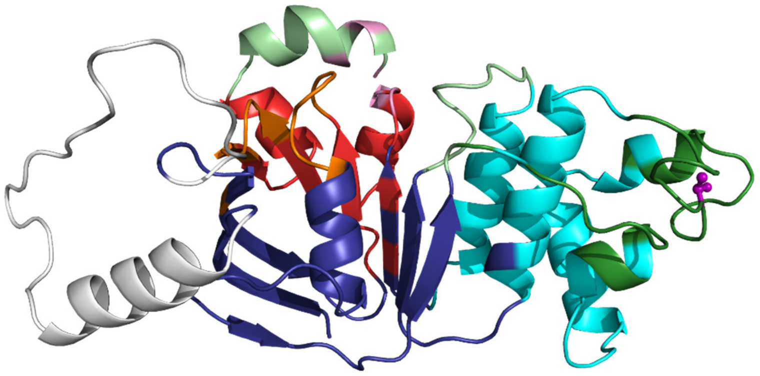
WT



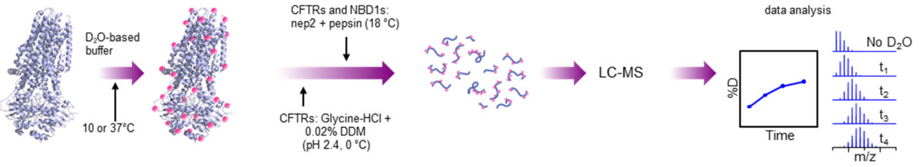
ΔF508



ΔF508 + rescue mutations



HDX experiments with full length CFTR



D$_2$O-based buffer

10 or 37°C

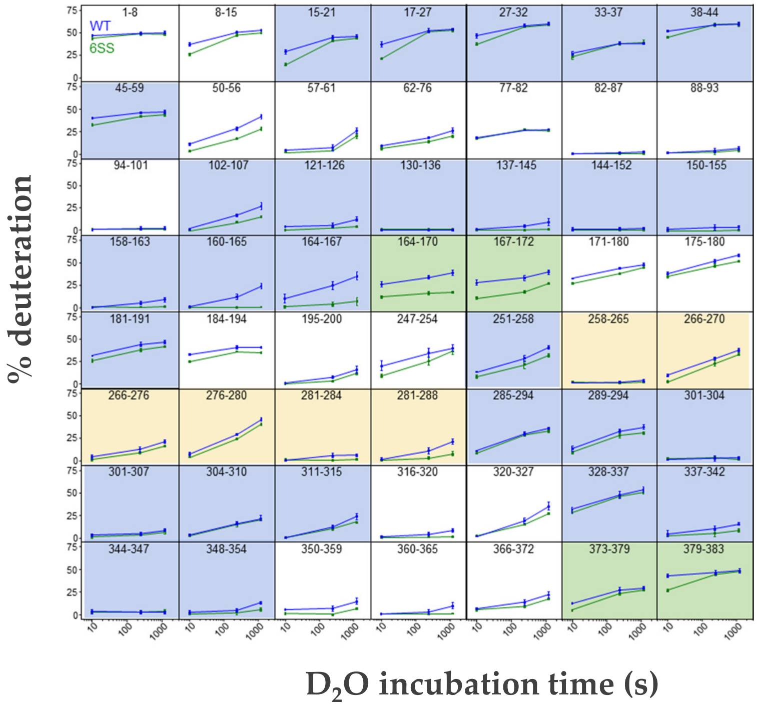
CFTRs and NBD1s: nep2 + pepsin (18 °C)

CFTRs: Glycine-HCl + 0.02% DDM (pH 2.4, 0 °C)

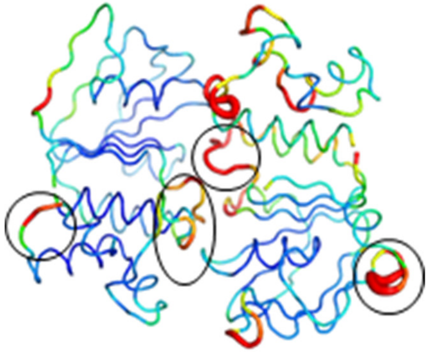
LC-MS

data analysis

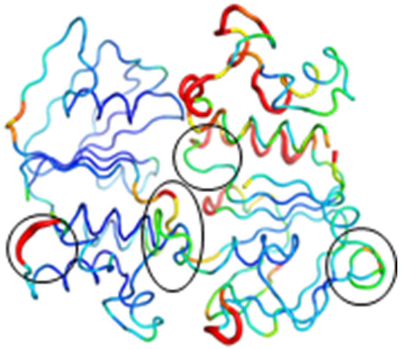
TMD1



F508G

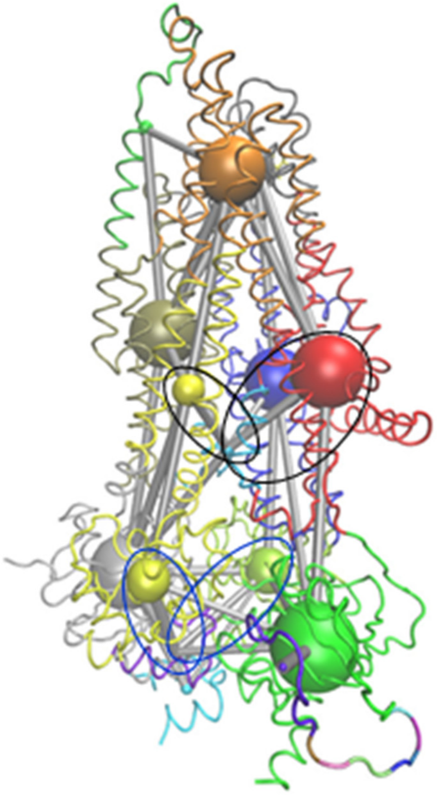


F508G-6SS

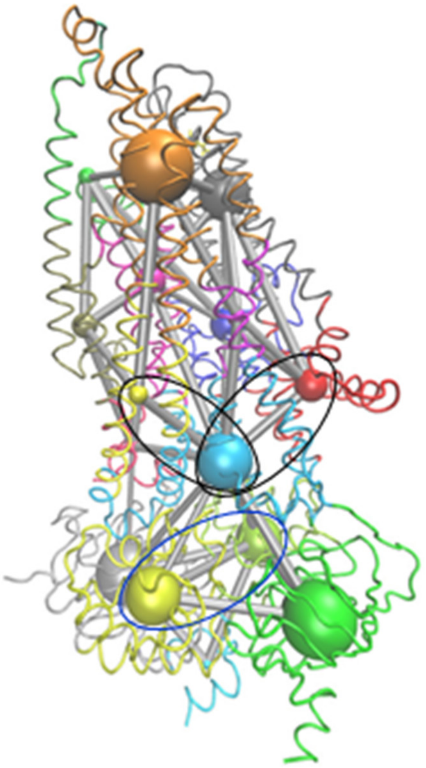


Allosteric stabilization of TM_{IC}

F508G



F508G-6SS



Thanks for your attention!

hegedus.tamas@hegelab.org