Algorithms and Learning for Protein Science

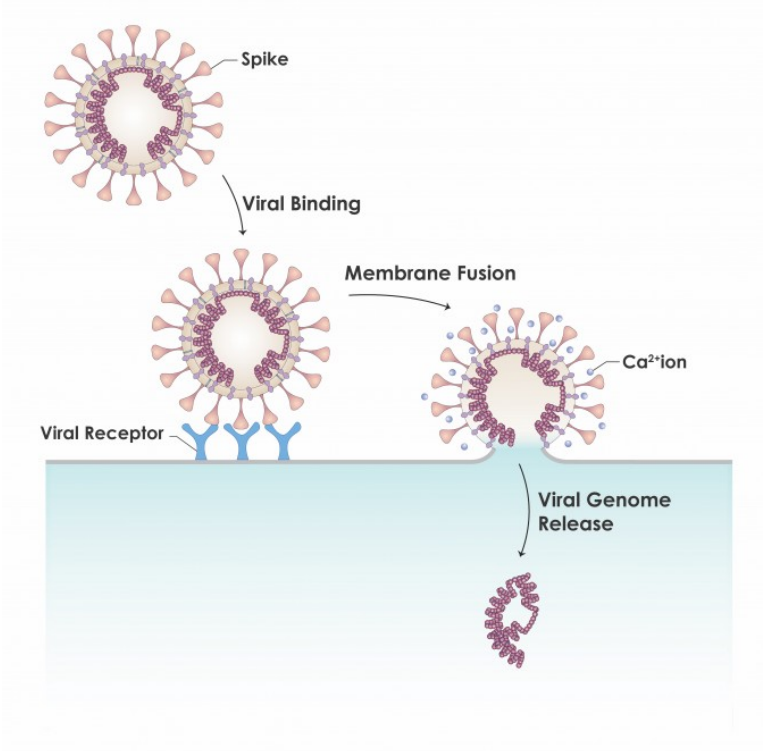
Frederic.Cazals@inria.fr

- Official: `https://www.master-mva.com/cours/algorithms-and-learning-for-protein-science/`
- Revised+slides: `http://www-sop.inria.fr/abs/teaching/ALPS/`

Algorithms and Learning for Protein Science

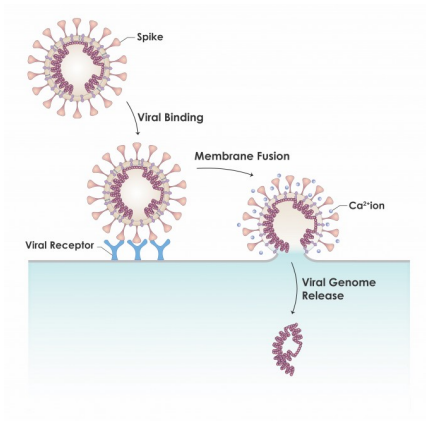
PART 1: Introduction
PART 2: Protein structure and resolution
PART 3: Physical chemistry & dynamics
PART 4: Biomolecular interactions and binding affinity
PART 5: Outlook

Protein functions, example: viral infection



https://youtu.be/e2Qi-hAXdJo

Protein functions, example: viral infection



https://youtu.be/e2Qi-hAXdJo

Computational Structural Biology

- Goals: unveil the *structure-dynamics-function* conundrum for biomolecules (proteins and nucleic acids)
- Methods: biophysics (crystallography, NMR, cryo-microscopy) + modeling
- Nobel prizes related to molecular/structural biology – as of 01/2025: 82 [1]
  - Chemistry or Physiology-medicine: structures and mechanisms
  - Chemistry or Physics: methods
  - Chemistry: modeling
    - 2013: Levitt, Karplus, Warshel for *the development of multiscale models for complex chemical systems*
    - 2024: D. Baker: *For computational protein design*; D. Hassabie and J. Jumper: *For protein structure prediction*
- An extraordinary field
  - Technology driven: novel biophysical experiments,
  - Raises open mathematical / computational questions,
  - Reveals the molecular foundations of biology and medicine.

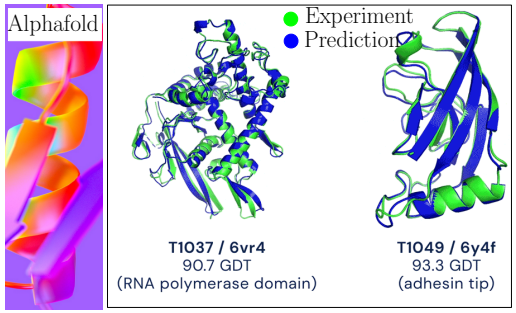
[1] https://pdb101.rcsb.org/learn/flyers-posters-and-other-resources/other-resource/structural-biology-and-nobel-prizes

Challenge *Structure of proteins*: specification

- Input: sequences from genome sequencing projects

```
P69905 (HBB_HUMAN) MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GS 53
P68871 (HBB_HUMAN) MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN 58
P02144 (MYG_HUMAN) -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKAS 59
                    : *:  :   *   ****         * * *::   * *   *  *  *
```

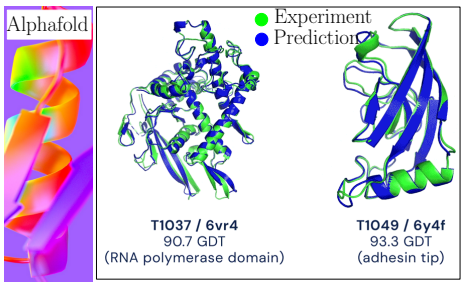
- Output: plausible structures i.e. atomic coordinates $\{(x_i, y_i, z_i)\}$



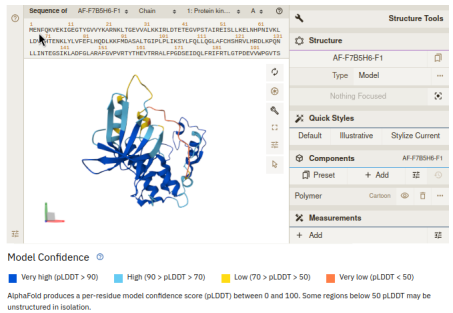
- Protein sequences versus structures: numbers
  - Num. sequences: `UniProtKB/TrEMBL`: $\sim 2.5 \times 10^8$; `UniProtKB/Swiss-Prot`: $\sim 6 \times 10^5$
  - Num. structures in the Protein Data Bank: $\sim 2.3 \times 10^5$ structures
- Recent & notable: the Deepmind combined approach (DL, optimization)
  - Bias towards well folded structure – no disorder (IDP)
  - Structure only – neither thermodynamics nor kinetics
  - Predicting is not explaining

Alphafold by Deepmind

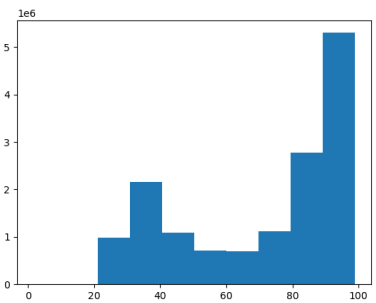
▷ Successes

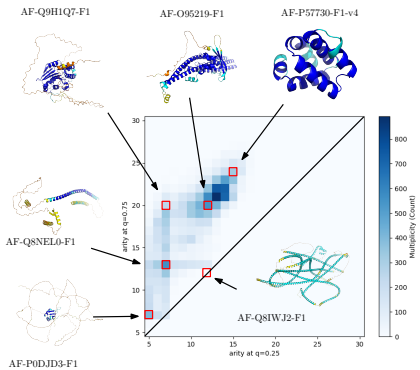


▷ . . . and failures



▷ Distribution of confidence value (pLDDT $\in [0, 1]$) per entire genome





- Structure only, no dynamics
- Biases towards well folded structures ... !!! flexible/disordered regions!!!
- Predicting is not explaining

▷Ref: Jumper et al, Nature, 2021

▷Ref: (Cazals and Sarti, 2025)

Challenge *Dynamics of proteins*: specification

Youtube

- Input: structure(s) of biomolecules + potential energy model
- Output
  - Thermodynamics: meta-stable states and observables
  - Kinetics: transition rates, Markov state models
- Time-scales
  - Biological time-scale $>$ millisecond
  - Integration time step in molecular dynamics: $\Delta t \sim 10^{-15} s$

- 162 amino acids, $>$ 2000 atoms
- 5.058ms of simulation time
- $\sim$ 230 GPU years on NVIDIA GeForce GTX 980 processor

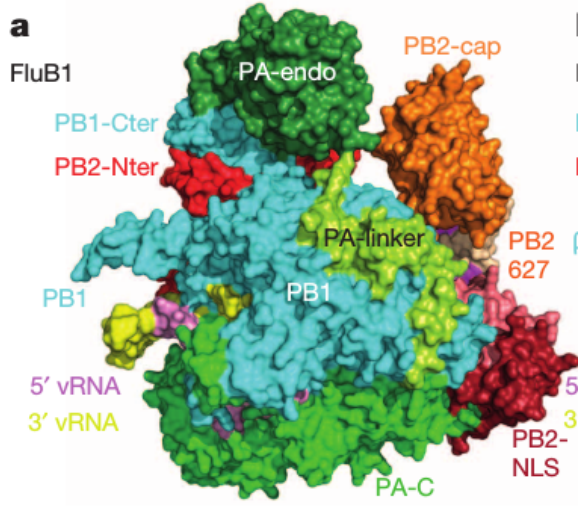
Ref: Chodera et al, eLife, 2019

Challenge *Molecular machines–structure and dynamics*: specification

- ▷ Molecular machines: assemblies with tens / hundreds of subunits
- ▷ Input
  - cryo-electron microscopy (cryo-EM) maps of whole assemblies
  - crystal structures of subunits
  - other data: native mass spectrometry data, . . .
- ▷ Output: structure(s) + mechanism(s)

▷ Polymerase of E. coli: structure+dynamics

▷ Polymerase of influenza: structure



▷Ref: Scheres et al, Elife, 2015; ▷Ref: Cusak et al, Nature, 2015

Cryo-electron microscopy: the microscope

- Titan Krios: boosting the resolution revolution
  - Space: near atomic resolution for most samples
  - Time: from images to thermodynamics and dynamics
- Approximate running costs:
  - Cost: 5 M€
  - Data acquisition: 2.5k€/day
  - One dedicated engineer

- A problematic situation in France (numbers to be confirmed):
  - France: 1 in Strasbourg, 1 in Grenoble, 1 at Soleil
  - UK, Germany: > 25 each
  - Cryo-microscopy lab of NY city: > 6

Hall of fame: more than 20 structural biology-related Nobel Prizes in 50 years

- (1962, chemistry) J. Kendrew and M. Perutz, *for their studies of the structures of globular proteins*
- (1962, medicine) F. Crick, J. Watson and M. Wilkins, *for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material*
- (1972, chemistry) Anfinsen, *for his work on ribonuclease, especially concerning the connection between the amino acid sequence and the biologically active conformation*
- (2002, chemistry) K. Wutricht and J. Fenn, *for the development of methods for identification and structure analyses of biological macromolecules*
- (2006, chemistry) R. Kornberg, *for his studies of the molecular basis of eukaryotic transcription*
- (2009, chemistry) V. Ramakrishnan, T. Steitz, A. Yonath, *for studies of the structure and function of the ribosome*
- (2013, chemistry) M. Karplus, M. Levitt, A. Warshell, *for the development of multiscale models for complex chemical systems*
- (2017, chemistry) F. Dubochet, J. Frank and R. Henderson, *for developing cryo-electron microscopy to identify high-resolution structure of biomolecules in solution*
- (2024, chemistry) D. Baker for *For computational protein design*, and D. Hassabie and J. Jumper for *For protein structure prediction*

Methods: molecular simulation

The Nobel Prize in Chemistry 2013
Martin Karplus, Michael Levitt, Arieh Warshel

The Nobel Prize in Chemistry 2013

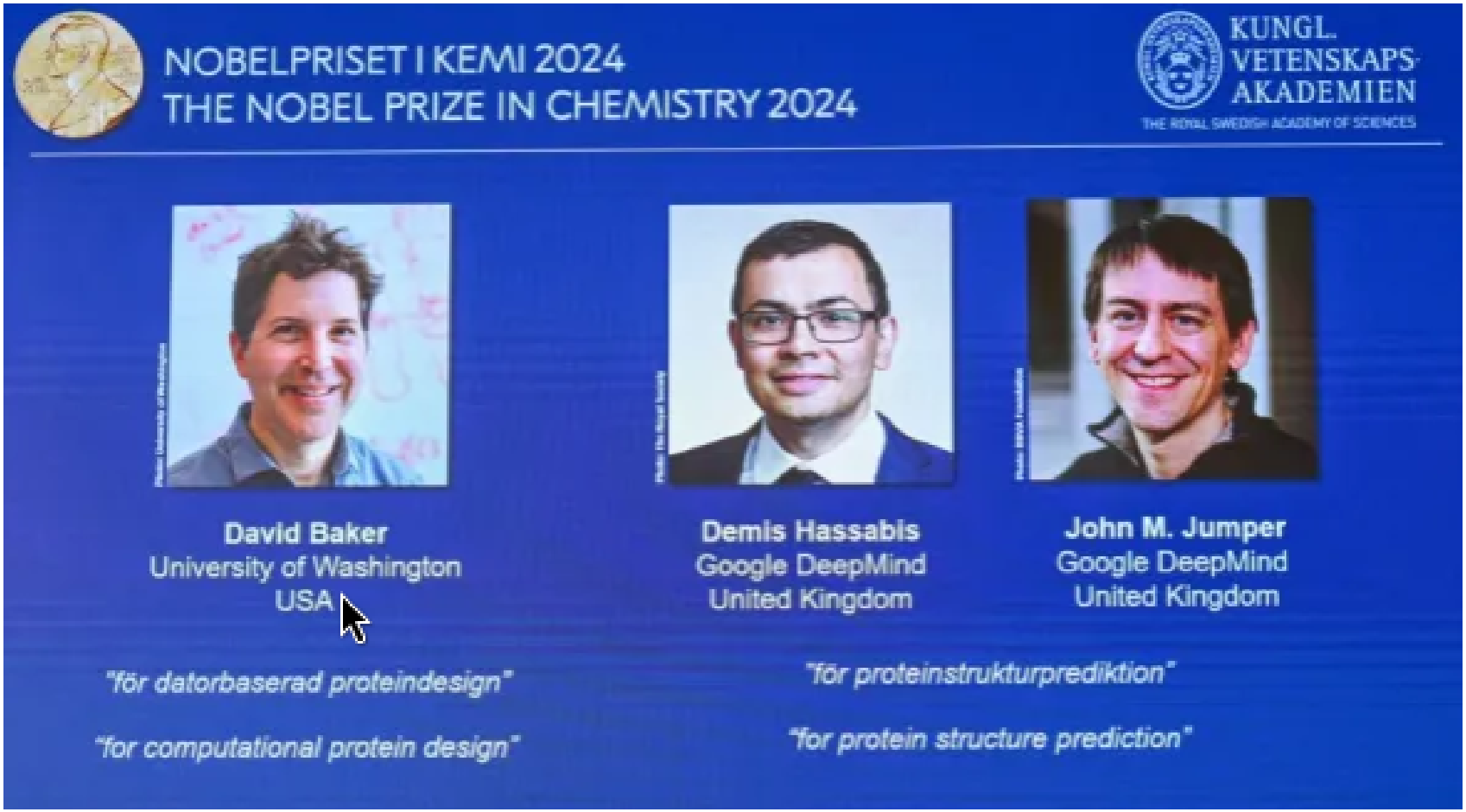
© Harvard University
Martin Karplus

Photo: © S. Fisch
Michael Levitt

Photo: Wikimedia Commons
Arieh Warshel

The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel *"for the development of multiscale models for complex chemical systems"*.

2024 Nobel prize in Chemistry



- D. Baker: *For computational protein design*
- D. Hassabie and J. Jumper: *For protein structure prediction*

Algorithms and Learning for Protein Science

PART 1: Introduction
PART 2: Protein structure and resolution
PART 3: Physical chemistry & dynamics
PART 4: Biomolecular interactions and binding affinity
PART 5: Outlook

Algorithms and Learning for Protein Science

Experiments for protein structure resolution

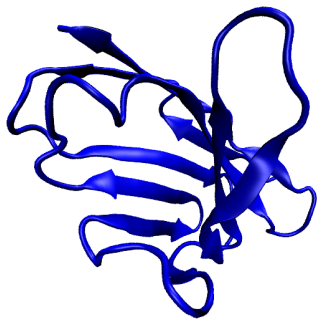
Experiments for protein structure resolution

What is a protein?

- Primary structure: sequence of amino acids

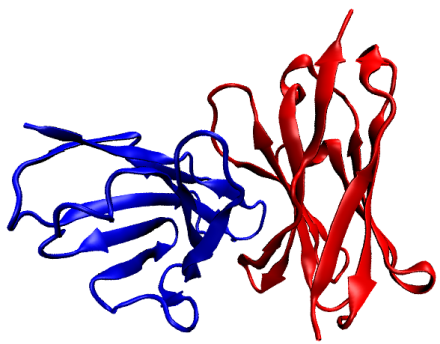
```
P69905 (HBB_HUMAN) MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GS 53
P68871 (HBB_HUMAN) MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN 58
P02144 (MYG_HUMAN) -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKAS 59
                    : *:  :   *   ****         * * *::   * *   *  *  *
```

- Polypeptide chain



- Protein - protein complex

- Heterodimeric protein



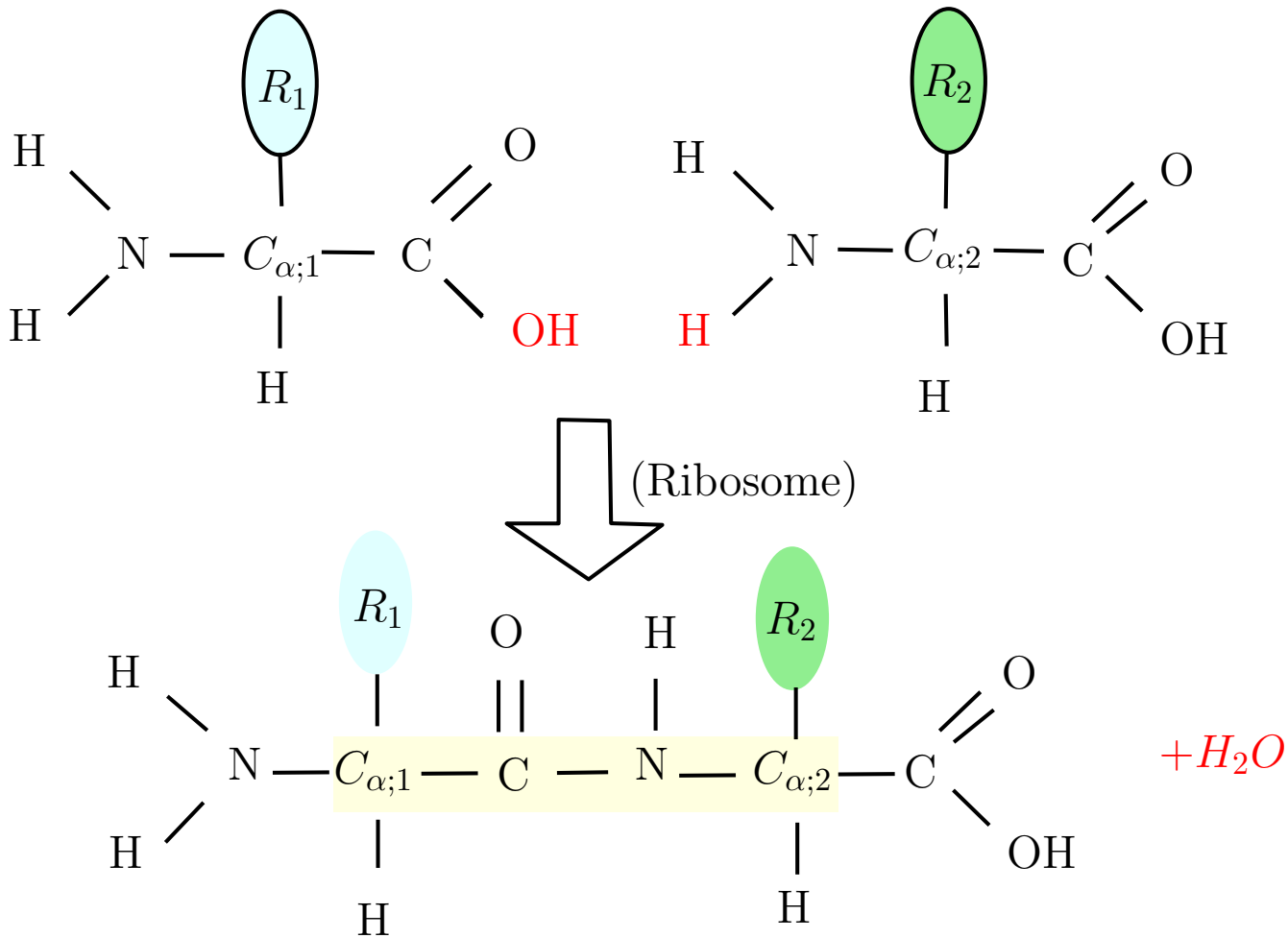
- Nb: median number of a.a. in a chain: ~ 450

Amino acids and the peptide bond

- ▷ Natural amino acids and their side chains

Nb: 0 to 10 heavy atoms per side chain

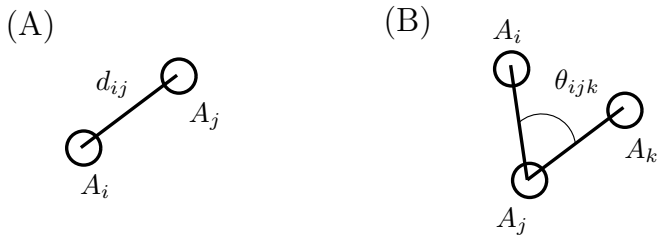
- ▷ Peptide bond synthesis:



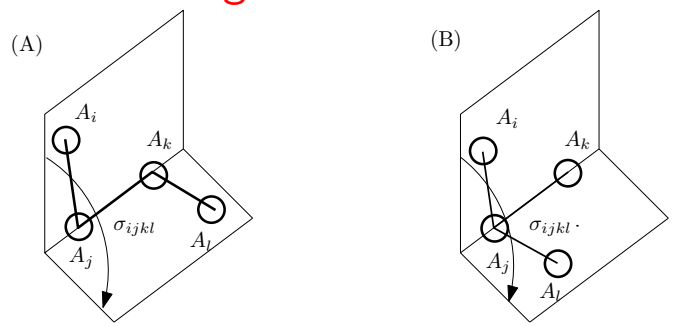
Geometric models: Cartesian and internal coordinates

▷ Cartesian versus internal coordinates: $\{x_i y_i z_i\}_i$ versus $\{d_{ij}, \theta_{ijk}, \sigma_{ijkl}\}$

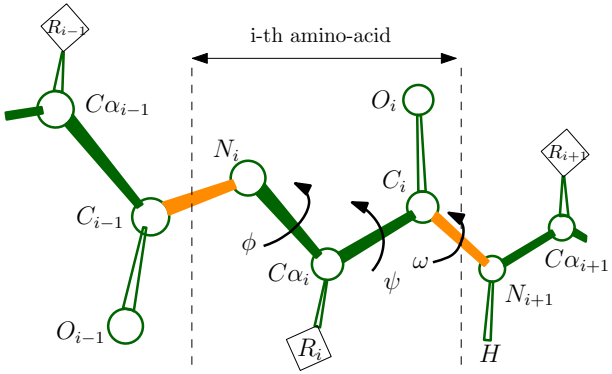
▷ Bond length and valence angle



▷ Dihedral angles



▷ Protein backbone

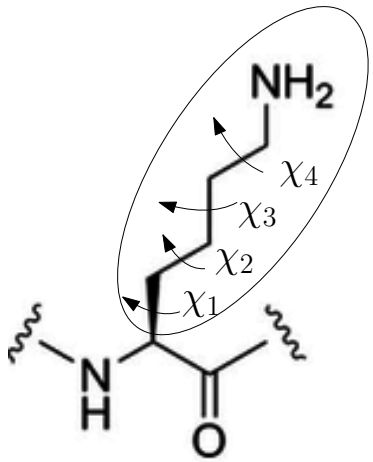


Ramachandran diagram, per a.a. type:

- bivariate distribution for (ϕ, ψ)

▷ Side chain: 20 natural amino acids

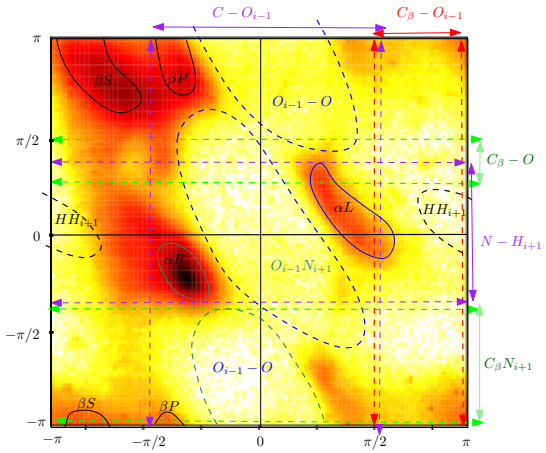
Exple: Lysine, 4 dihedral angles



LYS

The Ramachandran diagrams

▷ Ramachandran diagrams and populated regions

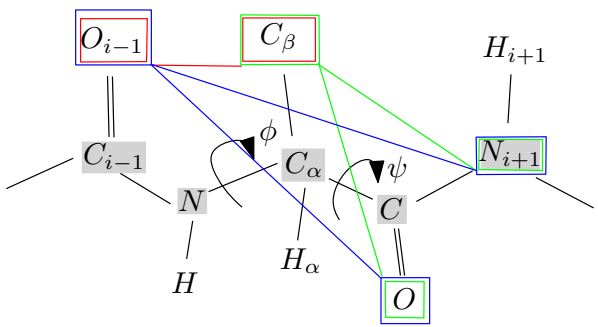


- Main regions: $\alpha L, \alpha R, \beta S, \beta P$
- Three prototypical diagrams
  - Glycine – no side chain/chiral C_α
  - Proline – side chain cycles on N
  - Others – with C_β and chiral C_α

▷ Distance constraints and the Ramachandran tetrahedron

$C1 : C_\beta - O_{i-1}$ $C2 : C_\beta - O + C_\beta N_{i+1}$

$C3 : O_{i-1} - O + O_{i-1} N_{i+1}$



▷Ref: Stereochemistry of polypeptide chain configurations, JMB, 1963; Ramachandran et al

▷Ref: Revisiting the Ramachandran plot, Protein Science, 2003; Ho et al

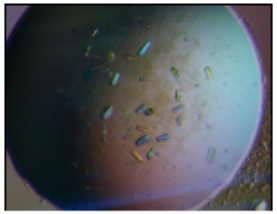
Algorithms and Learning for Protein Science

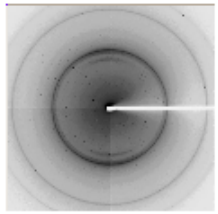
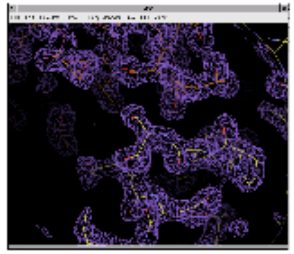
Experiments for protein structure resolution

Experiments for protein structure resolution

Structure resolution:
X ray crystallography, NMR, cryo-electron microscopy

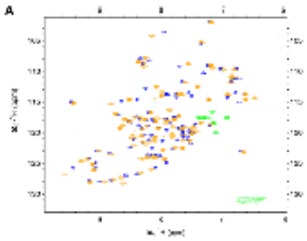
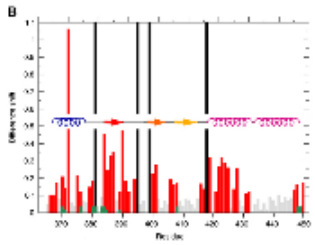
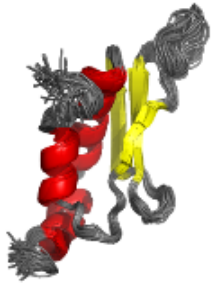
Crystallography



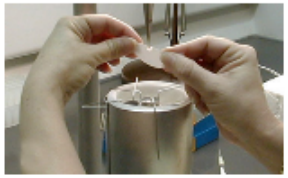



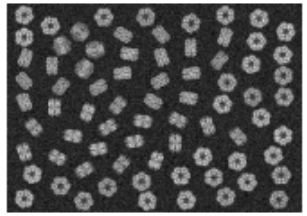
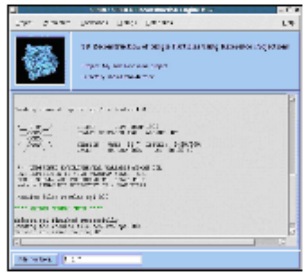
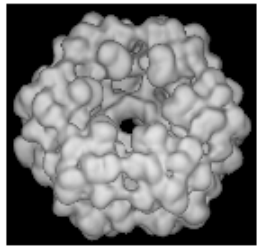

NMR



Cryo electron microscopy



Note: resolutions between 1 and 15 Å

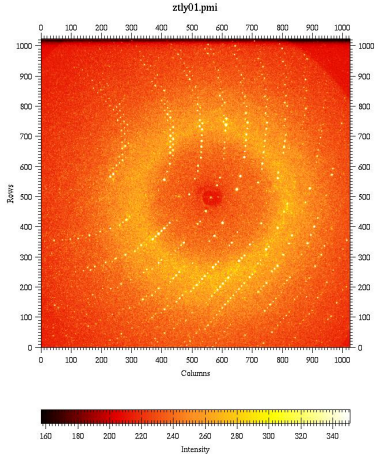
X ray crystallography

- (Selenium) crystals

- Protein crystals

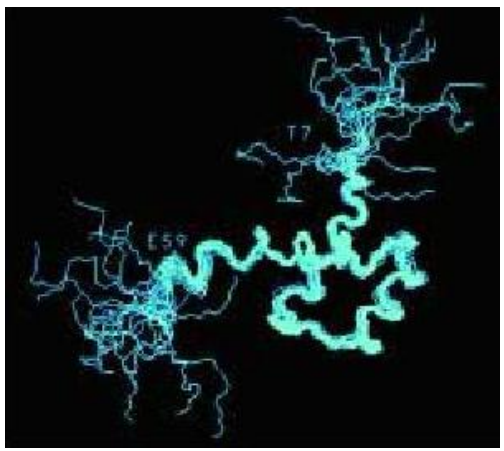
- X ray diffraction

- Diffraction pattern



Proteins and NMR

- One typically finds several types of regions within proteins:
  - well structured regions,
  - unstructured regions – see Fig.
  - regions undergoing conformational changes under selected conditions.

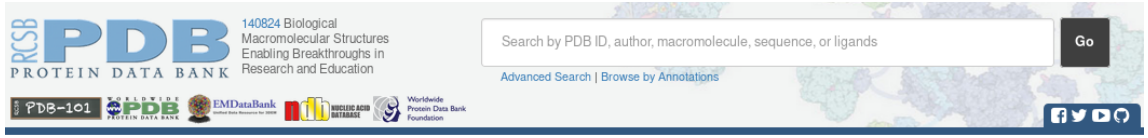


Flexibility of biomolecules: illustration. Structure of the Antennapedia homeodomain solved by NMR. Superimpositon of 20 conformations of the backbone. The tight packing of the regions 7-59 indicates that this region is stable, while the two ends are disordered. From the Nobel lecture of K. Wütricht.

- The dynamical properties are typically related to the functions of proteins.
- They depend on the conditions: pH, ionic strength.

The Protein Data Bank

- Structures in the PDB: origin and molecular type



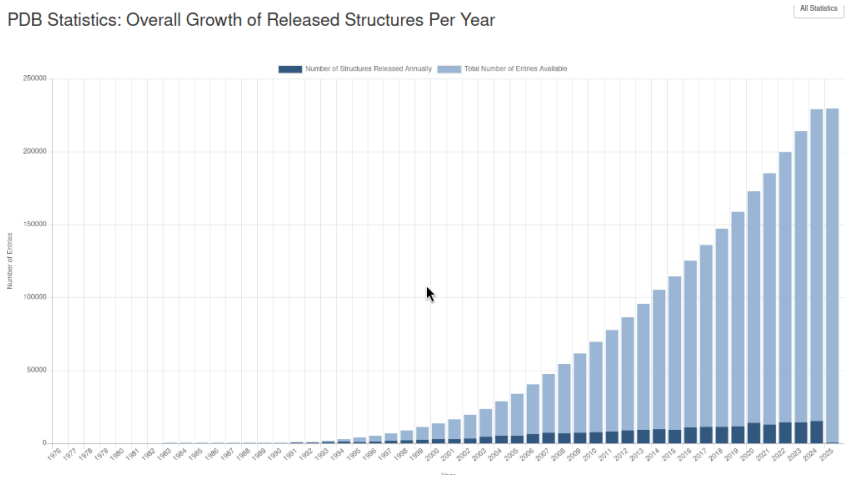
PDB Data Distribution by Experimental Method and Molecular Type

Other Statistics

Copy CSV

Experimental Method	Proteins	Nucleic Acids	Protein/NA Complex	Other	Total
X-Ray	118103	1931	6027	10	126071
NMR	10723	1243	249	8	12223
Electron Microscopy	1569	31	543	0	2143
Other	241	4	6	13	264
Multi Method	116	4	2	1	123
Total	130752	3213	6827	32	140824

- Growth of the PDB



- To learn more: PDB 101 `https://pdb101.rcsb.org/`

A typical PDB file

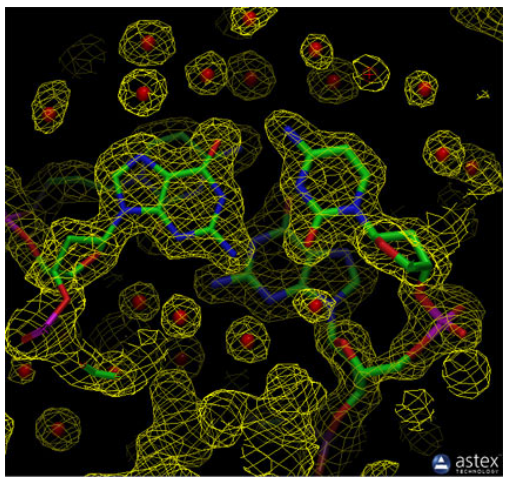
▷ Geometry information: n atoms yield $3n$ Cartesian coordinates . . . and $3n - 6$ degrees of freedom

```
ATOM      1  N   ASP A   1      23.963  -0.947  -1.031  1.00 37.52           N
ATOM      2  CA  ASP A   1      25.119  -0.797  -1.881  1.00 32.56           C
ATOM      3  C   ASP A   1      25.715   0.493  -1.356  1.00 29.72           C
ATOM      4  O   ASP A   1      24.964   1.396  -0.971  1.00 28.87           O
ATOM      5  CB  ASP A   1      24.721  -0.606  -3.341  1.00 34.71           C
ATOM      6  CG  ASP A   1      24.061  -1.777  -4.067  1.00 35.11           C
ATOM      7  OD1 ASP A   1      23.841  -2.849  -3.496  1.00 35.99           O
ATOM      8  OD2 ASP A   1      23.798  -1.612  -5.255  1.00 38.08           O
ATOM      9  H1  ASP A   1      23.429  -0.061  -1.100  1.00 20.00           H
ATOM     10  H2  ASP A   1      23.417  -1.821  -1.194  1.00 20.00           H
ATOM     11  H3  ASP A   1      24.348  -0.968  -0.067  1.00 20.00           H
ATOM     12  N   ILE A   2      27.025   0.577  -1.277  1.00 26.56           N
ATOM     13  CA  ILE A   2      27.669   1.808  -0.873  1.00 25.29           C
ATOM     14  C   ILE A   2      27.740   2.665  -2.147  1.00 26.50           C
ATOM     15  O   ILE A   2      28.123   2.164  -3.216  1.00 26.25           O
```

▷ Other pieces of information: organism, molecules / sequences (and their engineering), crystal resolution and symmetry group, secondary structures, disulfide bonds.

PDB files: pitfalls

- Focus on files from X ray crystallography:
  - Crystal structures: a confined environment
  - Asymetric unit versus biological unit
  - Extra atoms/molecules: water, chemical, co-factors, etc
  - Missing atoms: H systematically, heavy atoms . . . often
  - Alternate locations – if several conformations
  - Atoms retain dynamics encoded in B factors
  - Resolution and precision on coordinates – a complex problem



- To learn more:
  `https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/methods-for-determining-structure`

Visualization systems

- Main systems
  - Visual Molecular Dynamics
  - Pymol
  - Chimera
  - ...
- Demo

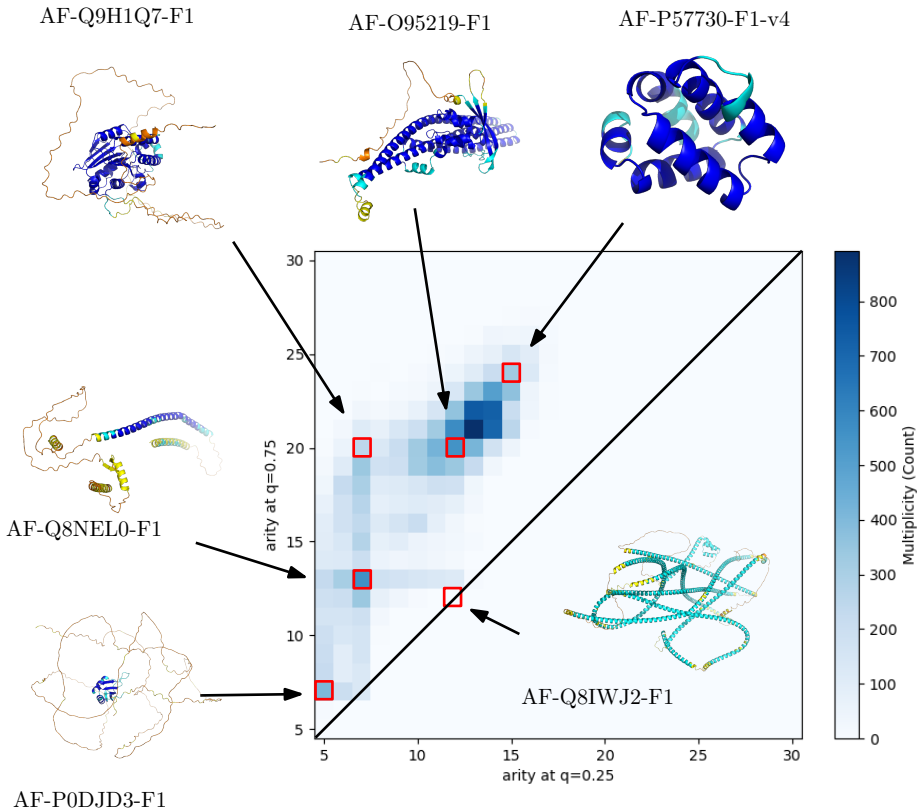
Databases of protein sequences

- (Reviewed) UniProtKB/Swiss-Prot – since 1986:
  - Refs: `https://www.uniprot.org/`, `https://www.uniprot.org/help/manual_curation`
  - High quality manually annotated and non-redundant protein sequence database,
  - Contains: experimental results, computed features and scientific conclusions
  - Size – Dec. 2024: $\sim 600k$ sequences
- (Unreviewed) UniProtKB/TrEMBL – since 1996:
  - Enriched from genome sequencing projects
  - Computationally analyzed records + automatic annotation and classification
  - Size – Dec. 2024: $\sim 250M$ sequences
- Removing redundancy: the UniRef databases UniRef100, UniRef90, UniRef50
  - `https://www.uniprot.org/help/uniref`
  - Clustering *homologous* sequences across organisms
- Gene Ontology:
  - `https://geneontology.org/`
  - Annotations on the function of the genes and gene products

Databases of protein models: AlphaFold-DB https://alphafold.ebi.ac.uk/

▷ Goal: provided AlphaFold predictions for all known sequences

▷ Ambition: to be expanded to cover most of the (over 100 million) representative sequences from the UniRef90 data set.



▷Ref: Varadi et al, NAR, 2021

▷Ref: Cazals and Sarti, 2025

Algorithms and Learning for Protein Science

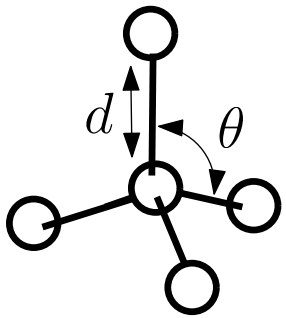
PART 1: Introduction
PART 2: Protein structure and resolution
PART 3: Physical chemistry & dynamics
PART 4: Biomolecular interactions and binding affinity
PART 5: Outlook

Algorithms and Learning for Protein Science

- Force fields and landscapes
- PEL and selected properties
- Thermodynamics vs kinetics
- Dynamics matter
- Protein dynamics: movies
- Dynamics and timescales

The potential energy of (bio-)molecules: force fields

- The $3n-6$ degrees of freedom of a molecule:



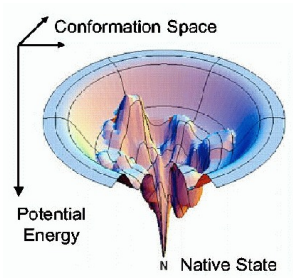
  - types for atoms (element, bonds)
  - covalent: bond lengths, angles
  - non covalent: pairwise distances
  - solvent model

- Potential energy: non linear function

$$V_{\text{total}} = V_{\text{bond}} + V_{\text{angle}} + (V_{\text{proper}} + V_{\text{improper}}) + (V_{\text{vdw}} + V_{\text{electro}}) \qquad (1)$$

V_{bond}: bonds
V_{angle}: covalent angles
V_{proper}: proper dihedrals
V_{improper}: improper dihedrals
V_{vdw}: van der Walls
V_{electro}: electrostatics

- Examples:
  - AMBER: $S_u = (73, 133, 112, 3, 14, 758)$
    1093 unique parameters
  - CHARMM: $S_u = (85, 152, 209, 13, 33, 1)$
    493 unique parameters
  - MARTINI: $S_u = (16, 4, 0, 2, 21, 3)$
    46 unique parameters



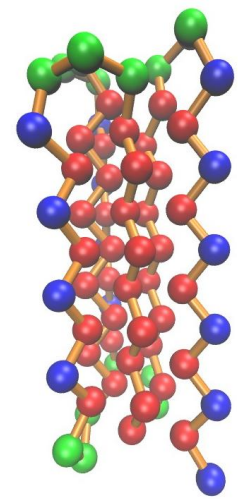
Force field from molecular mechanics: example

▷ BLN69 model protein: three types of Beads: hydrophobic(B), hydrophylic(L) and neutral(N).

▷ Dimension of conformation space: 3x69=207 cartesian coordinates

▷ Force field / potential energy:

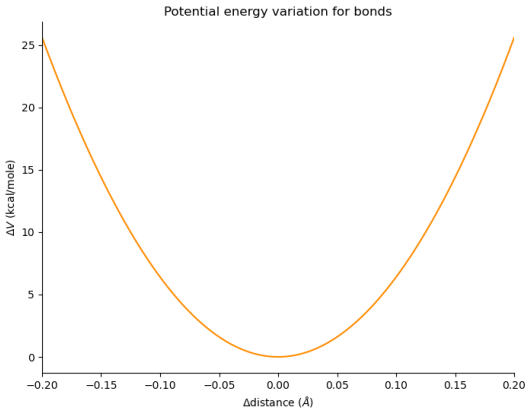
$$V_{BLN} = \frac{1}{2} \cdot K_r \sum_{i=1}^{N-1} (R_{i,i+1} - R_e)^2 + \frac{1}{2} K_0 \sum_{i=1}^{N-2} (\theta_i - \theta_e)^2$$
$$+ \epsilon \cdot \sum_{i=1}^{N-3} [A_i(1 + \cos\phi_i) + B_i(1 + 3\cos\phi_i)]$$
$$+ 4\epsilon \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \cdot C_{ij}[(\frac{\sigma}{R_{i,j}})^{12} - D_{ij}(\frac{\sigma}{R_{i,j}})^6]$$



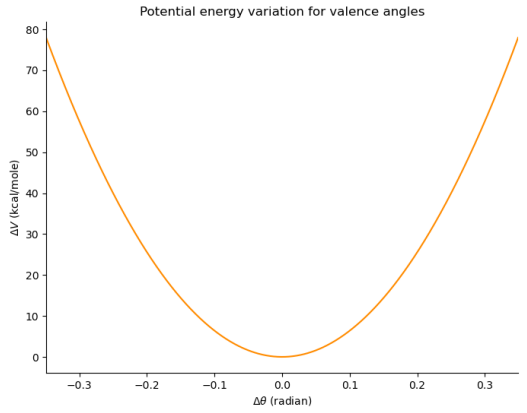
▷ Rmk. Model has been studied in detail, and 1/2 million of local minima have been reported.

▷Ref: Oaklet and Wales 2011; Cazals et al 2016

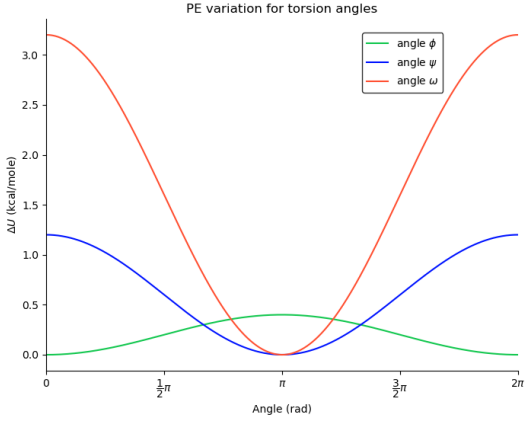
Softness of Internal coordinates –force constants from CHARMM 36



Bonds: $\delta d_{ij} \sim .2\text{Å} : \Delta V \sim 20 kcal/mol$



Valence angles: $\delta\theta_{ij} \sim 10^\circ : \Delta V \sim 20 kcal/mol$



Torsion angles:$\Delta V \sim 3 - 4 kcal/mol$

Dihedral angles:

- are indeed *soft* coordinates, but. . .
- long range steric clashes,
- yield complicated inverse problems. for loop closure

Potential of Mean Force: potential energy vs free energy

▷ Rationale: decouple the slow and fast dof of a system. Example: solvated protein:

- slow dof: protein
- fast dof: solvent molecules

▷ How to: replace the overall potential energy by an average, computed over the fast dof

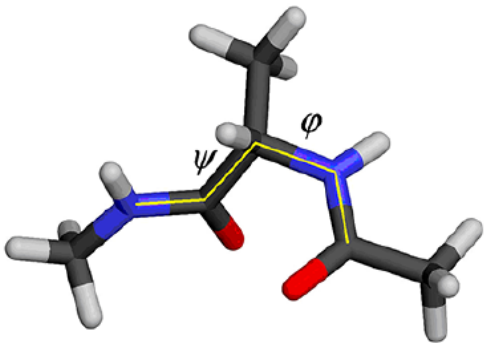
▷ PMF definition:

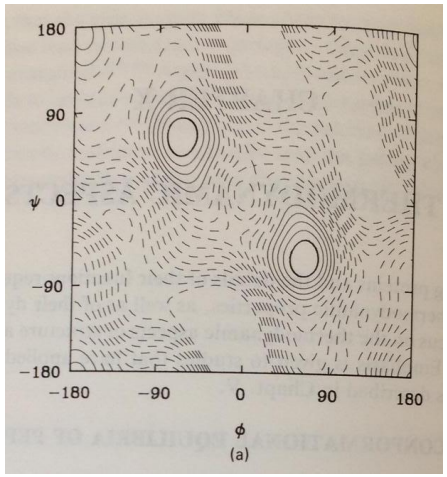
$$\exp(-\beta PMF(x_1,\ldots,x_n)) \propto \frac{\int \exp(-\beta V(x_1,\ldots,x_d))dx^{n+1,\ldots,d}}{\rho_{\text{unif.}(x_1,\ldots,x_n)}} \qquad (2)$$

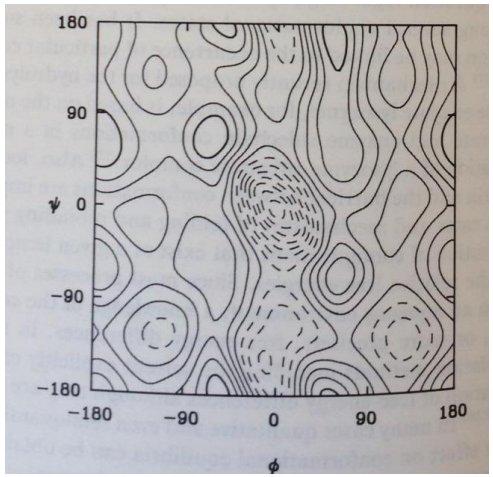
Nb: in this equation, $\rho_{\text{unif.}}$ stand for the uniform distribution on the slow dof, which naturally depends on the nature of these parameters – cartesian or internal coordinates.

Potential energy landscapes: illustration

- Potential energy map: vacuum (PE) versus solvated (PMF):

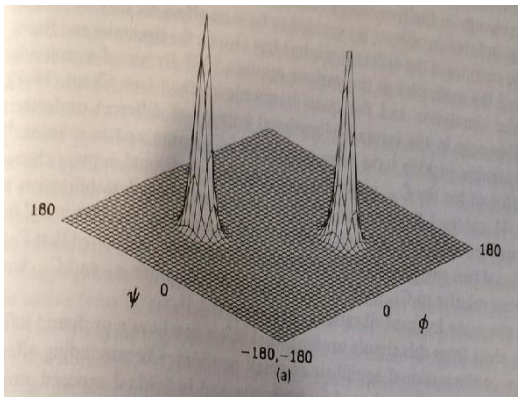


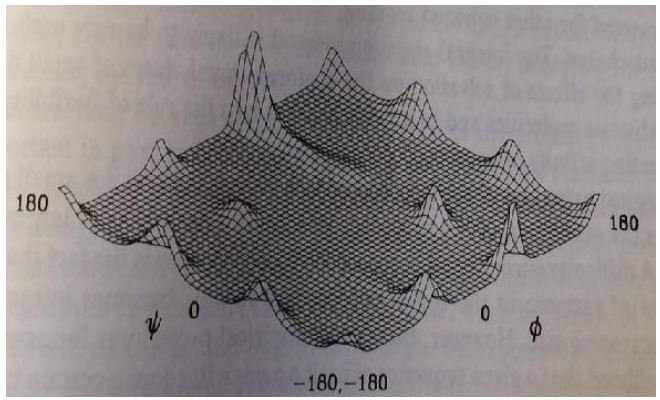




- Corresponding Boltzmann-weighted probability maps:

Solvent stabilizes many more conformers–hydrogen bonding.





- Ref: Petitt, Karplus, Chem. Phys. Lett., 121, 1985

Designing force fields

▷ A regression problem: predicting a response variable (potential energy) using descriptors of the system.

▷ Number of parameters to be fitted: 500-1000.

▷ Strategy: optimizing the various parameters to replicate physical properties of small organic molecules

- heat capacity, density, viscosity, T for state changes, surface tension, etc.

▷ Avoiding over-fitting: classical machine learning techniques are used, including cross-validation, Bayesian models, etc

Table 1. Organic Compounds Considered in the Calibration and Validation of 2016H66 along with Experimental Values of Their Target Properties[a]

chemical function	code	name	calibration	ε	T_{l+g} (K)	ρ_{liq} (kg m^{-3})	ΔH_{vap} (kJ mol^{-1})	ΔG_{wat} (kJ mol^{-1})	ΔG_{che} (kJ mol^{-1})
alcohols	MTL	methanol		33.50[b]		784.00[c]	37.43[c]	−21.37[d]	−5.40[d]
	ETL	ethanol	×	24.00[b]		784.93[c]	42.31[c]	−20.95[d]	−10.84[d]
	1PL	propanol	×	20.00[b]		799.60[c]	47.49[c]	−20.58[d]	−11.42[d]
	BTL	butanol		17.70[b]		805.75[c]	52.34[c]	−19.75[d]	−14.73[d]
	PTL	pentanol		15.10[b]		810.80[c]	56.94[c]	−18.70[d]	−15.10[d]
	HXL	hexanol		13.00[b]		815.34[c]	61.85[c]	−18.24[d]	−22.22[d]
	HPL	heptanol		11.50[b]		820.00[e]	66.81[f]	−17.70[d]	−25.18[d]
	OTL	octanol		10.10[b]		821.57[c]	70.98[c]	−17.10[g]	
	2PL	propan-2-ol		19.10[b]		780.00[c]	45.52[c]	−19.33[d]	−9.92[d]
	2BTL	butan-2-ol		16.70[b]		802.41[c]	49.66[c]	−19.16[g]	
	2PTL	pentan-2-ol		13.80[b]		805.40[c]	53.10[c]	−18.36[g]	
	3PTL	pentan-3-ol		13.40[b]		816.00[c]	53.10[c]	−18.20[g]	
	CHXL	cyclohexanol		16.40[b]		968.40[c]	62.01[c]	−22.90[g]	
	2M2P	2-methylpropan-2-ol		11.50[b]		781.20[c]	46.82[c]	−18.87[d]	−12.26[d]
	2M2B	2-methylbutan-2-ol		5.70[b]		805.00[c]	50.20[c]	−18.53[g]	
ethers	DME	methoxymethane	×	6.20[b]	254	722.00[h]	21.24[f,i]	−8.03[d]	
	DEE	ethoxyethane	×	4.20[b]		707.82[c]	27.10[c]	−7.36[d]	−12.68[d]
	MPH	1-methoxypropane		3.70[j]		735.60[f]	27.60[f]	−6.95[d]	
	DXE	1,2-dimethoxyethane	×	7.30[b]		863.70[c]	36.39[c]	−20.25[d]	
aldehydes	EAL	acetaldehyde	×	21.10[b]		778.00[c,k]	26.11[c]	−14.60[g]	
	PAL	propionaldehyde		18.40[b]		791.20[c]	29.63[c]	−14.40[g]	
	BAL	butyraldehyde		13.40[b]		796.40[c]	33.68[c]	−13.30[g]	

▷Ref: Horta et al, JCTC, 2016

Algorithms and Learning for Protein Science

Force fields and landscapes

PEL and selected properties

Thermodynamics vs kinetics

Dynamics matter

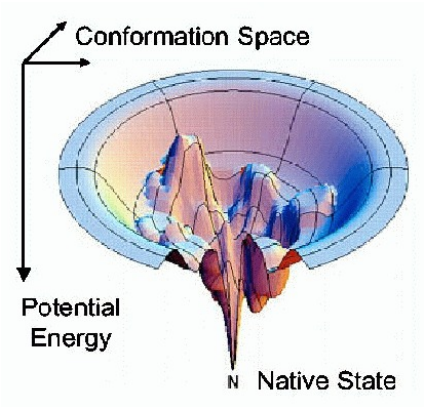
Protein dynamics: movies

Dynamics and timescales

PEL and Levinthal's paradox

▷ The Levinthal paradox: the exponential (infinite) number of conformations a protein can adopt is incompatible with its exhaustive exploration of the conformational space. Indeed, proteins work on observable time scales?

▷ Anfinsen's dogma: the *global* (free) energy minimum is encoded in the sequence, and is kinetically accessible.
Anfinsen: 1972, Nobel prize in chemistry.

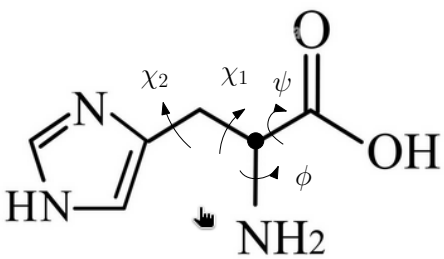


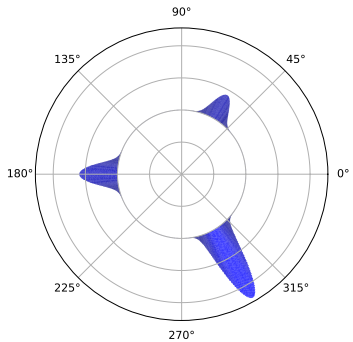
Funneled PEL which makes it easy to find the global minimum

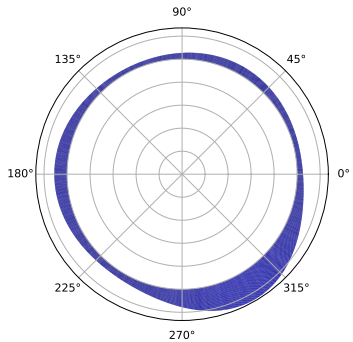
The Folding Problem: intrinsic difficulty

- C. Anfinsen's experiment (Nobel 1972): A.A. sequence $\Rightarrow$ structure
  - Identification of the native state —minimum of free energy
  - Determination of the folding pathways

- Torsion angles and their rotameric (discrete) structure







- Levinthal's paradox: yet, an exponential number of conformations
  - Counting: 4 torsion angles per amino acid, each 3 options: $3^4 = 81$
  - Counting: 500 amino acids per protein: $81^{500} \sim 1.7 \times 10^{954}$

PEL: a complex multi-scale structure

- Structure: conformations consisting of
  - local minima
  - saddle points connecting them
- Disconnectivity tree/graph:
  - Nodes: local minima
  - Edges: saddle points connecting local minima

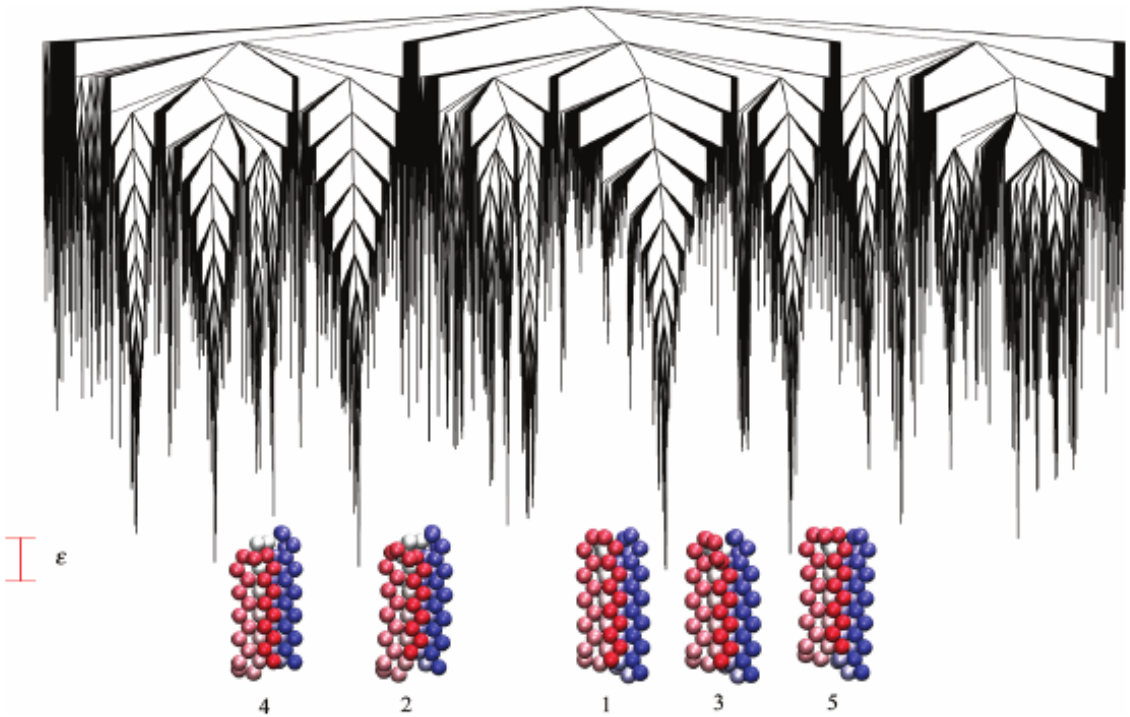


Figure: **Simplified view of the potential energy landscape of BLN69, encoded in its disconnectivity tree.** About 1/2 million of local minima are known. Overall, 8 structures which stand out in terms of minima.

Critical points of the potential energy

▷ Critical point of the potential energy V: point where the gradient vanishes
Diagonalizing the Hessian + performing a linear change of variables yields

$$V(u) = \sum_{i=1,\ldots,k} \lambda_i u_i^2 - \sum_{i=k+i1,\ldots,d} \lambda_i u_i^2. \tag{3}$$

▷ Def.: the number of negative eigenvalues is called the *index* of the critical points.

▷ Two types of special points are of interest:

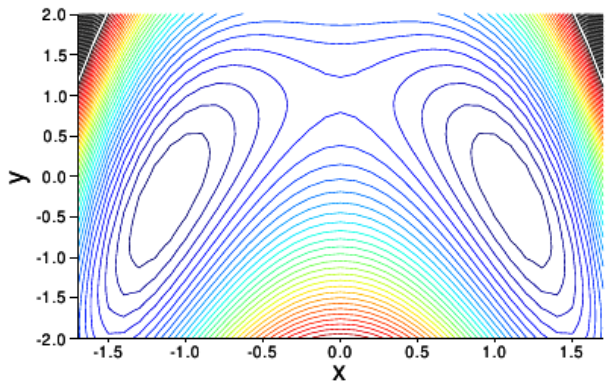
- local minima: zero negative eigenvalues.
- index one saddle points: one negative eigenvalue.

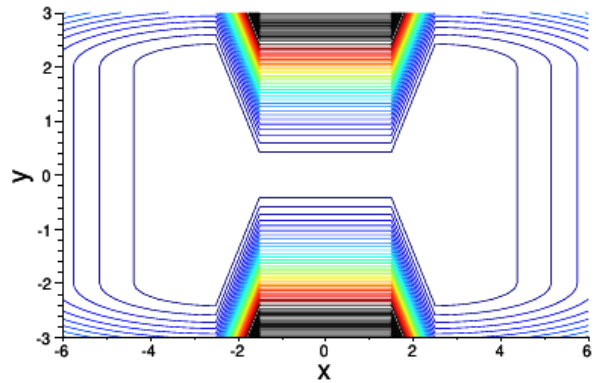
▷ Difficulties faced in exploring a PEL:

- the number local minima and saddle points generally grows exponentially with the dimension.
- the presence of the solvent complicate matters, and makes the surface very rugged.

Barriers on a PEL: enthalpic and entropic

- Two types of barriers:
  - enthalpic barrier. a region of the PEL requiring to gain significant elevation to visit a neighboring basin / local minimum.
  - entropic barrier. an almost flat region of the PEL, whose exhaustive exploration is likely to occur to find the exit point(s). NB: the foggy plateau metaphor.

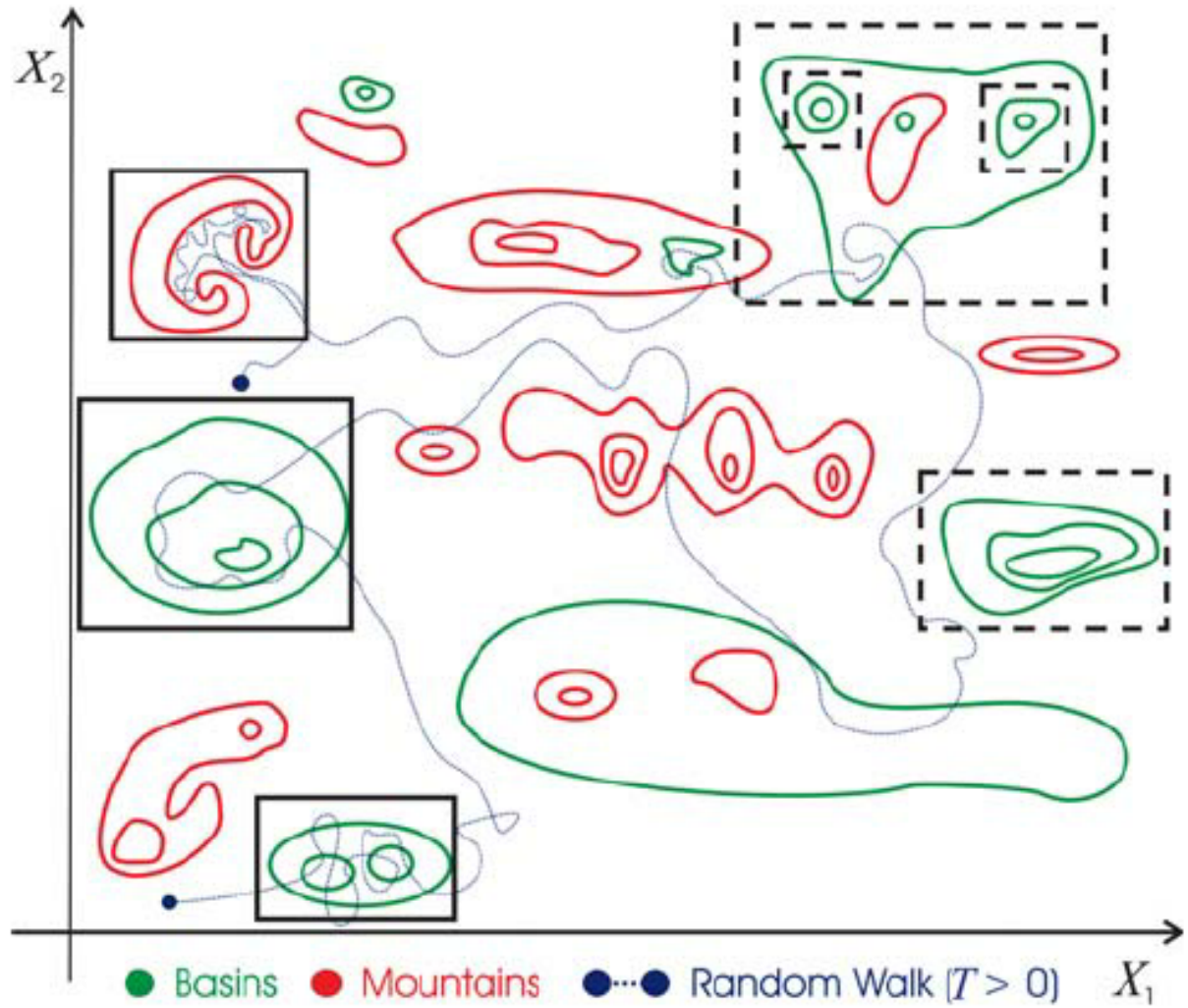




- Ref: `Free energy computations, Lelièvre, Stoltz, Rousset, 2010`

PEL: a typical exploration

- Typical exploration of a PEL energy landscape



- Ref: Schoen and Jansen, Int. J. Mat. Res., 2009

Classifying PEL

- Bad news:
  - Very large number of critical points.
  - Enthalpic - entropic barriers.
- Good news:
  - strong coupling / coherence between dof, making the effective dimensionality small.
  - few significant (large or deep) basins.
- Prototypical landscapes

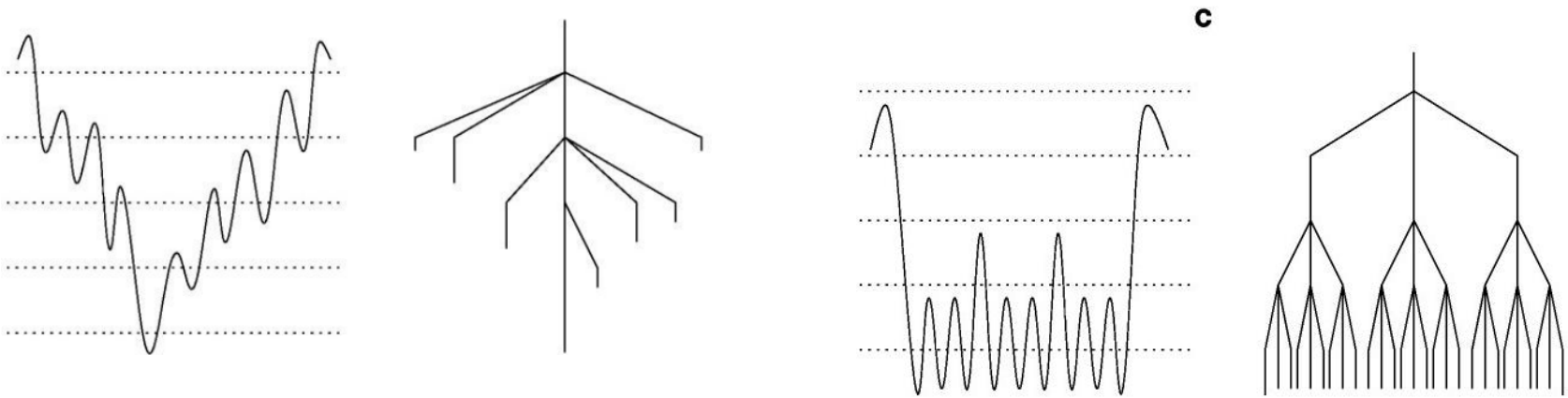


Figure: **Classification of PEL. (Left)** Protein **(Right)** Glasses

- Ref: Wales, 2003

Multiscale analysis of PEL: what is a deep basin ?

- Analogous problem for mountains: defining a peak is a matter of scales:
  - prominence: closest distance to the nearest local maximum with higher elevation
  - **culminance**: elevation drop to the saddle leading to a higher local maximum

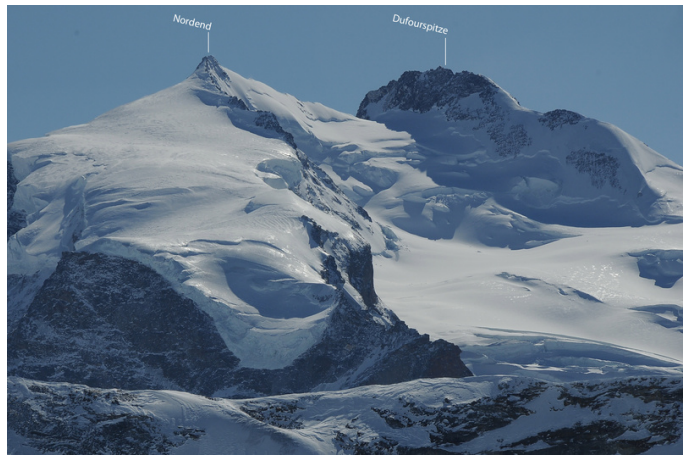


Figure: **Mountain topography: what is a peak?** The analysis requires defining local minima as opposed to local maxima.

- The Norden peak does not qualify:
  - fourth highest peak of the Mont Rose massif, 4609 meters
  - prominence: 575 meters; culminance: 94 meters

Algorithms and Learning for Protein Science

- Force fields and landscapes
- PEL and selected properties
- Thermodynamics vs kinetics
- Dynamics matter
- Protein dynamics: movies
- Dynamics and timescales

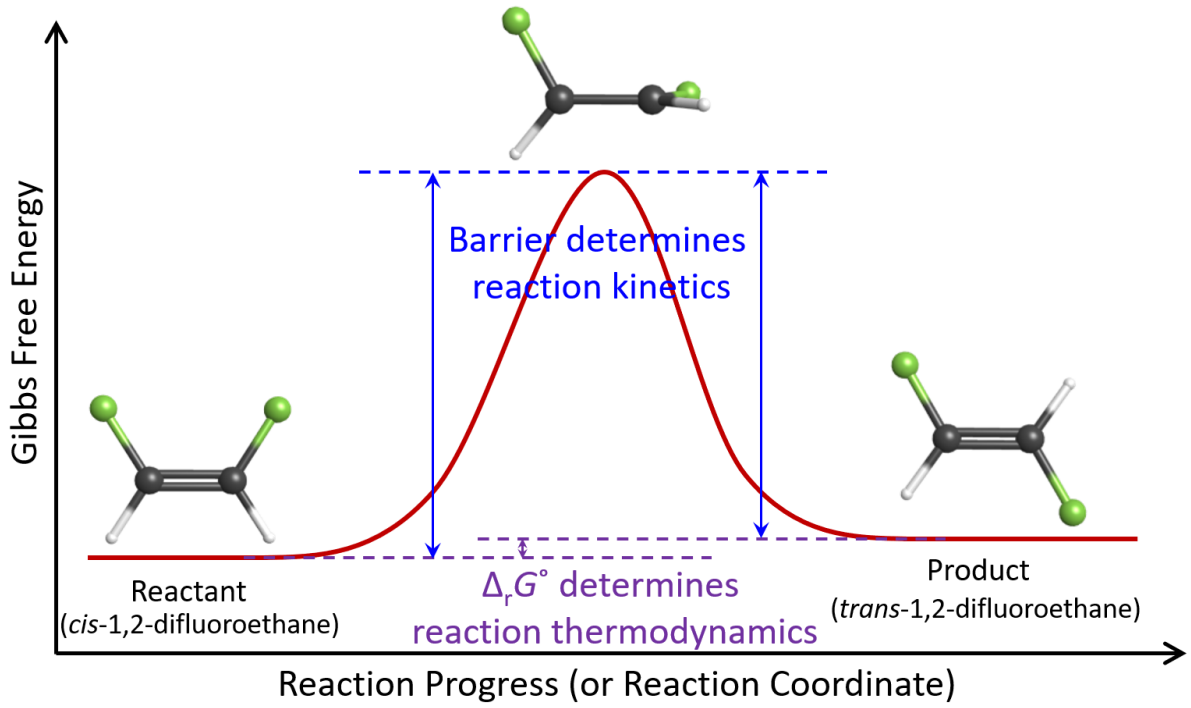
Thermodynamic ensembles

- Conformation versus state:
  - Conformation: microscopic state (think: coordinates)
  - Thermodynamic/meta-stable state: conformations *easily interconvertible* into one-another
- Stability and barriers:
  - Stability is a function of temperature and observation time
  - Two types of barriers: (potential) energy, entropy
- Ensemble: collection of all conformations belonging to the same thermodynamic state
- Thermodynamic ensembles: canonical (NVT), isobaric-isothermal (NPT), microcanonical (NVE)
- Application of ensembles: computing averages values of observables – next slide

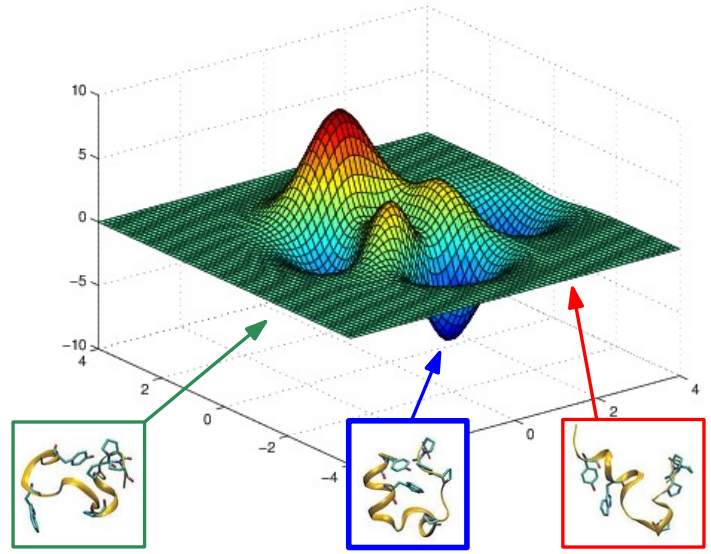
Thermodynamics and observables

- Quantities defined for a conformation x:
  - Potential energy: $V(x)$
  - Kinetic energy: $K(x)$
  - Total energy: $E(x) = V(x) + K(x)$
  - Boltzmann's distribution: $P^{\text{eq}}(x) = e^{-\beta E(x)}/Z, Z = \sum_{\text{Conformation}x} P^{\text{eq}}(x)$
- Quantities defined for ensembles:
  - Average of observable $\mathcal{O}$ wrt an ensemble: $< \mathcal{O} > \equiv \sum_{\text{Conformation}x} \mathcal{O}(x) P^{\text{eq}}(x)$
  - Exple: average total energy $U =< E >$
  - NVT: Helmholtz free energy $A = U - TS = k_B T \ln Z$
  - NPT: Gibbs free energy $G = U + PV - TS = H - TS$

Emergence of macromolecular function(s) from Structure – Thermodynamics – Kinetics



Emergence of macromolecular function(s) from Structure – Thermodynamics – Kinetics

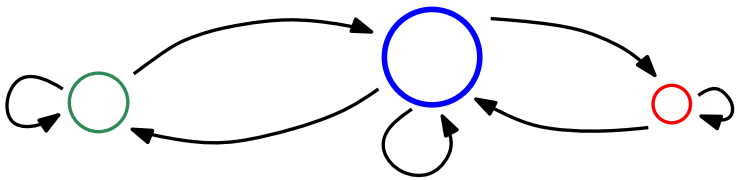


Potential Energy Landscape

- large number of local minima
- enthalpic barriers
- entropic barriers

Structure: stable conformations i.e. local minima of the PEL

Thermodynamics: meta-stable conformations i.e. ensemble of conformations easily inter-convertible into one - another.



Kinetics: transitions between meta-stable conformations e.g. Markov state model

Kinetics: Markov State Models (MSM)

- A MSM is characterized by:
  - the MSM is described by a graph whose nodes are the meta stable states, and edges transitions between them.
  - the system does not have memory: the current states determines the next one.
  - transitions between two states obeys transition probabilities.
- Example: Markov state model for the protein methyltransferase SETD8

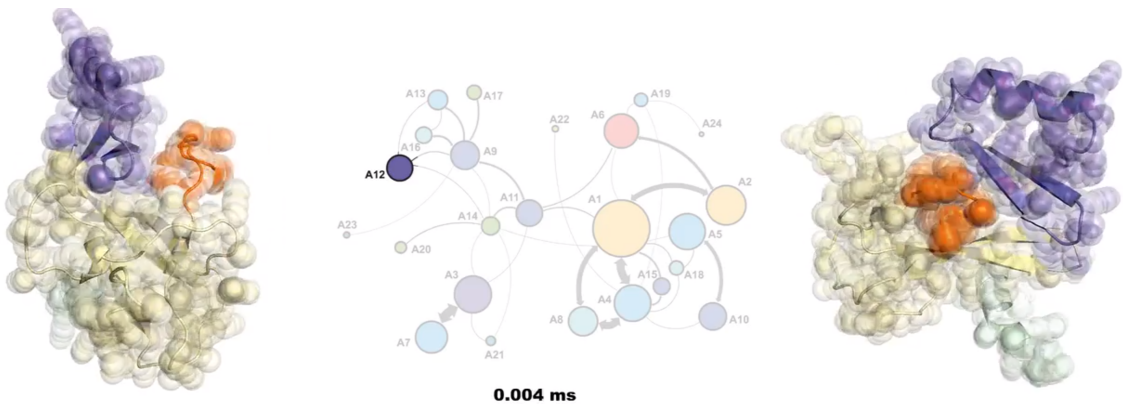


Figure: From Chodera et al, eLife, 2019

- Movie: `https://www.youtube.com/watch?v=IDLEi-M8Aow`

Kinetics: the Master Equation

▷ Notations:

- $\boldsymbol{P}(t)$: the occupancy probabilities of individual minima
- $\boldsymbol{W}$: matrix coding the gains/losses per basin associated with transition rates across saddles, *i.e.*

▷ Master equation for one species:

$$\frac{dP_i}{dt} = \sum_{j \neq i} k_{ij} P_j(t) - k_{ji} P_i(t). \tag{4}$$

The dynamics of the system are defined by the following ordinary differential equation [**?**]:

$$d\boldsymbol{P}(t)/dt = \boldsymbol{W}\boldsymbol{P}(t) \tag{5}$$

▷ Rmk. Bootstrapping from vibrational free energies –cf the harmonic oscillator.

▷`Ref: Van Kampen, Stochastic processes in physics and chemistry, 1992`

Algorithms and Learning for Protein Science

Force fields and landscapes

PEL and selected properties

Thermodynamics vs kinetics

Dynamics matter

Protein dynamics: movies

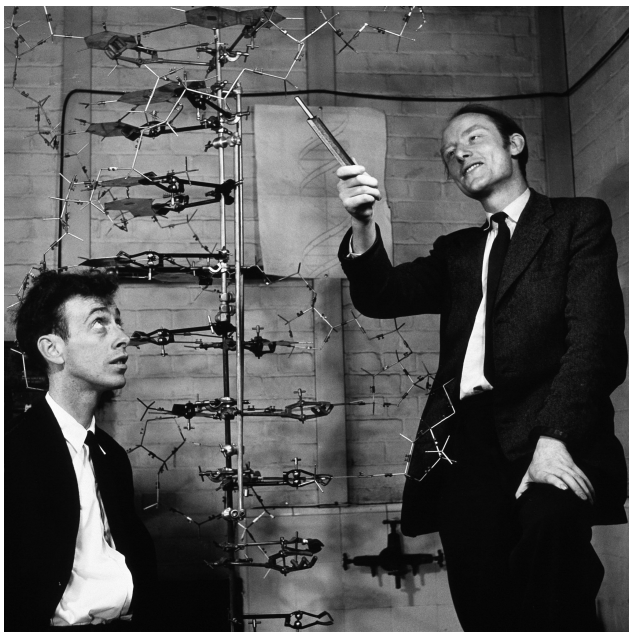
Dynamics and timescales

Molecular dynamics and protein functions: examples

- Molecular flexibility and the functions of proteins:
  - Protein folding: the adoption of a tertiary structure, possibly with the help of chaperonin proteins.
  - Enzymatic activity: the transformation of reactant into products, by a protein or a ribozyme (cf the ribosome)
  - Non covalent molecular recognition: the formation of a non covalent complex.
  - Allostery: activity regulation thanks to the binding of an effector protein.
  - Ion and small molecular transport: e.g. ion transport in membrane channels.
  - Active transport of large molecules: e.g. transport into the nucleus by the nuclear pore complex.
  - Molecular motions (in muscle e.g.): actin-myosin movements.

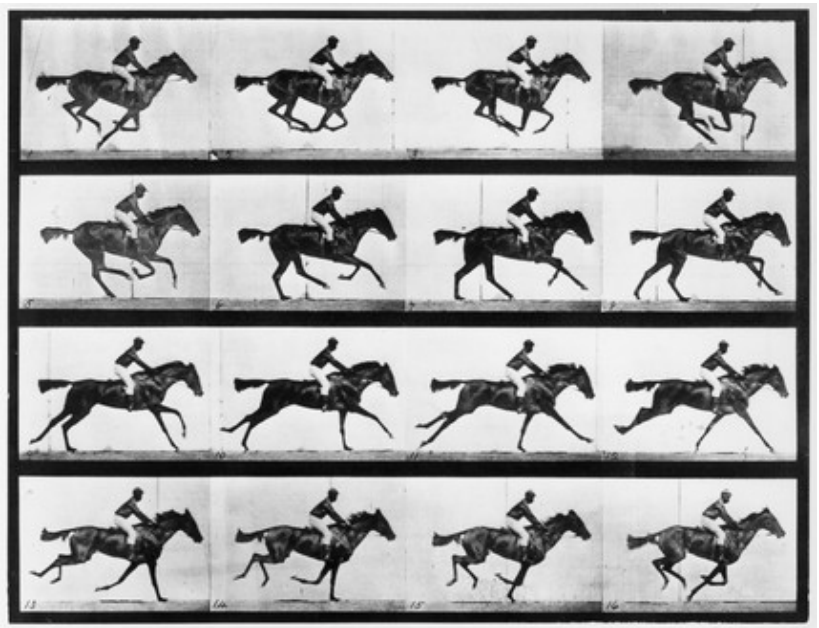
Two schools: static versus dynamic studies

▷ Balls and sticks



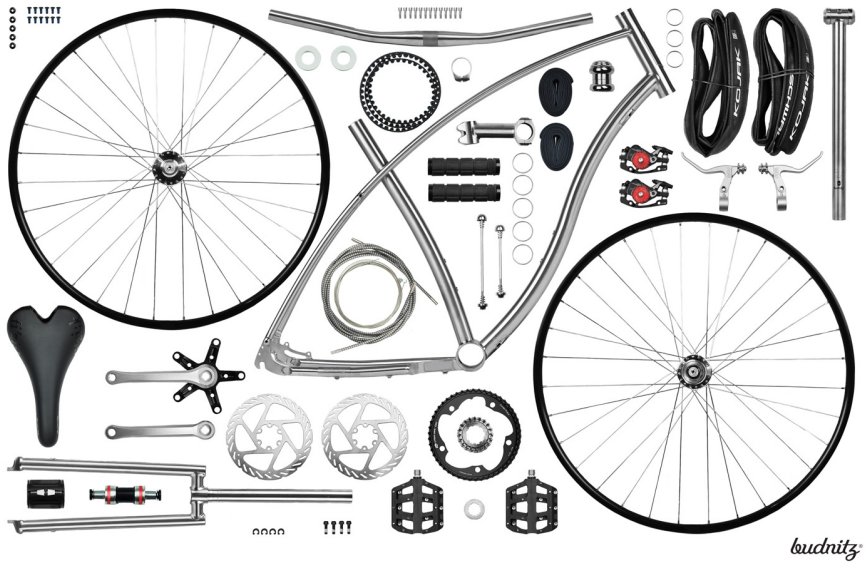
(Watson and Crick, DNA model)

▷ The Ballet & time lapse

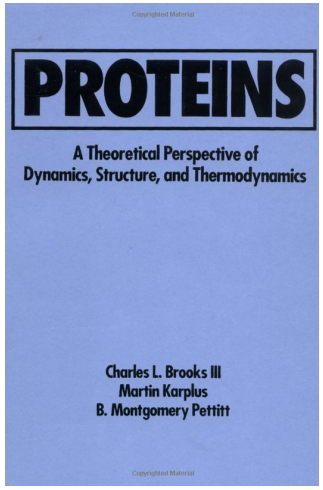


- Static analysis using crystal structure from the Protein Data Bank
  http://rcsb.org
- Dynamical analysis using molecular mechanics

Statics vs dynamics



Dynamics: *alea jacta est* in the mid eighties



CONTENTS

▷Ref: Brooks, Karplus, Montgomery Pettitt; Advances in Chemical Physics, Proteins; Wiley, 1988

Algorithms and Learning for Protein Science

Force fields and landscapes

PEL and selected properties

Thermodynamics vs kinetics

Dynamics matter

Protein dynamics: movies

Dynamics and timescales

Dynamics of biomolecules:
first simulation of a protein, and a recent one

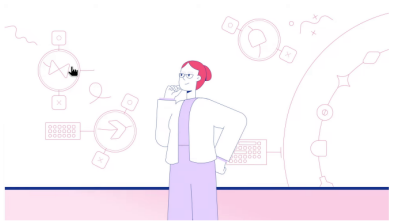
Molecular dynamics and protein functions: movies

- Selected (great) movies:
  - Proteins according to AlphaFold https://youtu.be/KpedmJdrTpY
  - Protein synthesis by the ribosome: https://www.youtube.com/watch?v=TfYf_rPWUdY
  - Membrane fusion-entry and infection by SARS-Cov-2: https://youtu.be/e2Qi-hAXdJo
  - Molecular motors: https://www.youtube.com/watch?v=X_tYrnv_o6A
- Other videos of interest:
  - Various phenomena in this movie: https://www.youtube.com/watch?v=wJyUtbn005Y
  - More XVivo movies at https://www.youtube.com/channel/UCAUL7Wl_lydKXI8q0oi4CUw

- Rmk. Remarkable illustration of the aforementioned mechanisms can be found in the book [?]; see also the gallery on the PDB portal, at https://pdb101.rcsb.org/sci-art/goodsell-gallery.

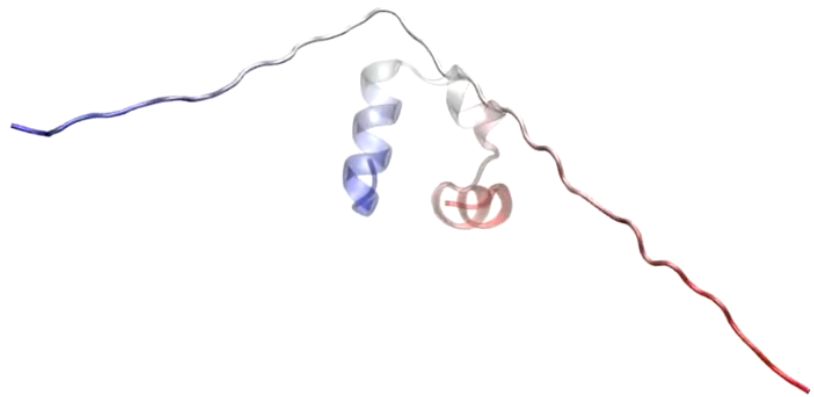
Proteins and protein folding

Movie by Deepmind/ The AlphaFold team



https://youtu.be/KpedmJdrTpY

Dynamics of biomolecules: protein folding
A molecular dynamics simulation



▷ More videos from the movie gallery Illinois:
`http://www.ks.uiuc.edu/Gallery/Movies/ProteinFoldingStretching/`

Algorithms and Learning for Protein Science

- Force fields and landscapes
- PEL and selected properties
- Thermodynamics vs kinetics
- Dynamics matter
- Protein dynamics: movies
- Dynamics and timescales

Protein dynamics: time scales

- Relevant quantities to quality dynamics and their time scales:
  - Spatial extent: the size of the region undergoing the change,
  - Amplitude: the displacement undergone,
  - Times scale: the duration required for the conformational change to occur.

Table 1. Characteristic Time Scales for Protein Motions

event	spatial extent (nm)	amplitude (nm)	time (s)	appropriate simulations
bond-length vibration	0.2−0.5	0.001−0.01	10^{-14}−10^{-13}	QM methods
elastic vibration of globular domain	1.0−2.0	0.005−0.05	10^{-12}−10^{-11}	conventional MD
rotation of solvent-exposed side chains	0.5−1.0	0.5−1.0	10^{-11}−10^{-10}	conventional MD
torsional libration of buried groups	0.5−1.0	0.05	10^{-11}−10^{-9}	conventional MD
hinge bending (relative motion of globular domains)	1.0−2.0	0.1−0.5	10^{-11}−10^{-7}	Langevin dynamics, enhanced sampling MD methods?
rotation of buried side chains	0.5	0.5	10^{-4}−1	enhanced sampling MD methods?
allosteric transitions	0.5−4.0	0.1−0.5	10^{-5}−1	enhanced sampling MD methods?
local denaturation	0.5−1.0	0.5−1.0	10^{-5}−10^{1}	enhanced sampling MD methods?
loop motions	1.0−5.0	1.0−5.0	10^{-9}−10^{-5}	Brownian dynamics?
rigid-body (helix) motions		1.0−5.0	10^{-9}−10^{-6}	enhanced sampling MD methods?
helix−coil transitions		>5.0	10^{-7}−10^{4}	enhanced sampling MD methods?
protein association	≫1.0			Brownian dynamics

▷Ref: Adcock and McCammon, Chem. Reviews, 2006

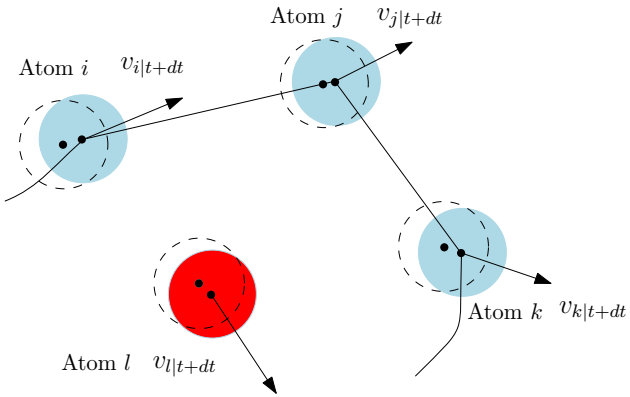
Molecular simulation: four schools

- Various classes of methods:
  - Molecular dynamics
  - Monte Carlo based methods
  - Energy landscapes methods
  - AI based methods: tokenization, diffusion, denoising diffusion maps

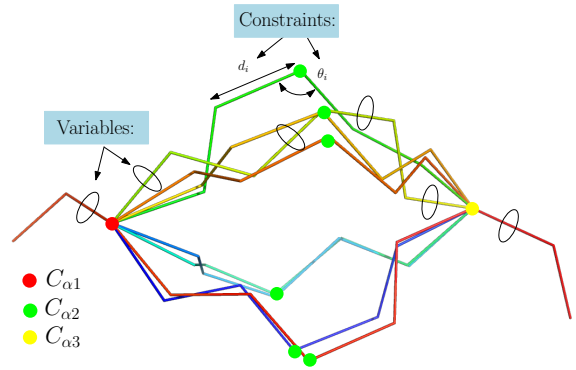
▷Ref: Field, A practical introduction to the simulation of molecular systems, 1999
▷Ref: Frenkel and Smit, Understanding molecular simulations, 2002
▷Ref: Wales, Energy landscapes, 2003
▷Ref: Stoltz et al, Free energy calculations, 2010
▷Ref: Jing et al, Berger et al, etc - posterior to 2023

Dynamics: *California dreamin'...*

- Direct problems / molecular dynamics versus inverse problems



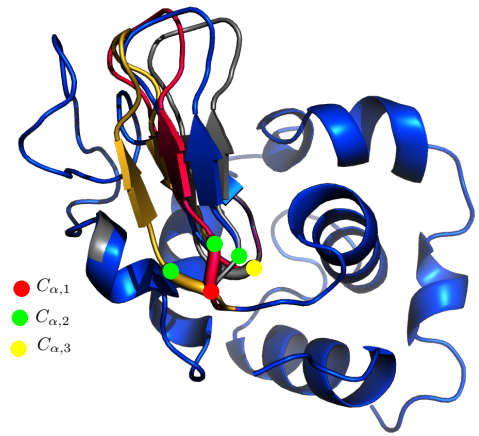
Molecular dynamics, time-steps of 10^{-15}s:
$\|\Delta x_i\| \sim 1/100$Å



Inverse problems, typical changes:
$\|\Delta x_i\| \sim 1 - 10$Å

- Next major scientific goal: a metaphor

Paris / San Francisco / Stanford:
30' + 30' minutes



Biomolecules: identifying stable states and their probabilities

Algorithms and Learning for Protein Science

PART 1: Introduction
PART 2: Protein structure and resolution
PART 3: Physical chemistry & dynamics
PART 4: Biomolecular interactions and binding affinity
PART 5: Outlook

Algorithms and Learning for Protein Science

Biomolecular recognition: proteins and binding affinity

The time dimension: $K_d, K_{\text{on}}, K_{\text{off}}$ and $1/K_{\text{off}}$

Binding affinity estimation: hardness

Application: influenza - antibody complexes

Main points

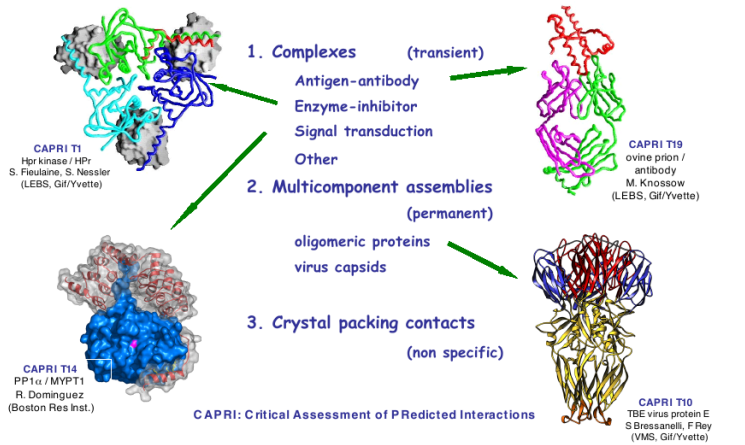
Main points:

- Proteins and binding affinity
- Enthalpy - entropy compensation
- The time dimension $1/K_{\text{off}}$
- Application: antibodies binding viruses

Biological complexes: structural diversity

▷ Biology rests on interactions biomolecules make with one another. A remarkable variety of such complexes exist, both in size and time scales spanned – see Janin et al.
▷ Size-wise, complexes span a range from $O(100\ kDa)$ up to 120 MDa (mammalian NPC). Note that the nuclear pore complex is the largest assembly known (to date) in eukaryotic cells, as it involves circa 500 polypeptide chains.

▷ Biological complexes: diversity



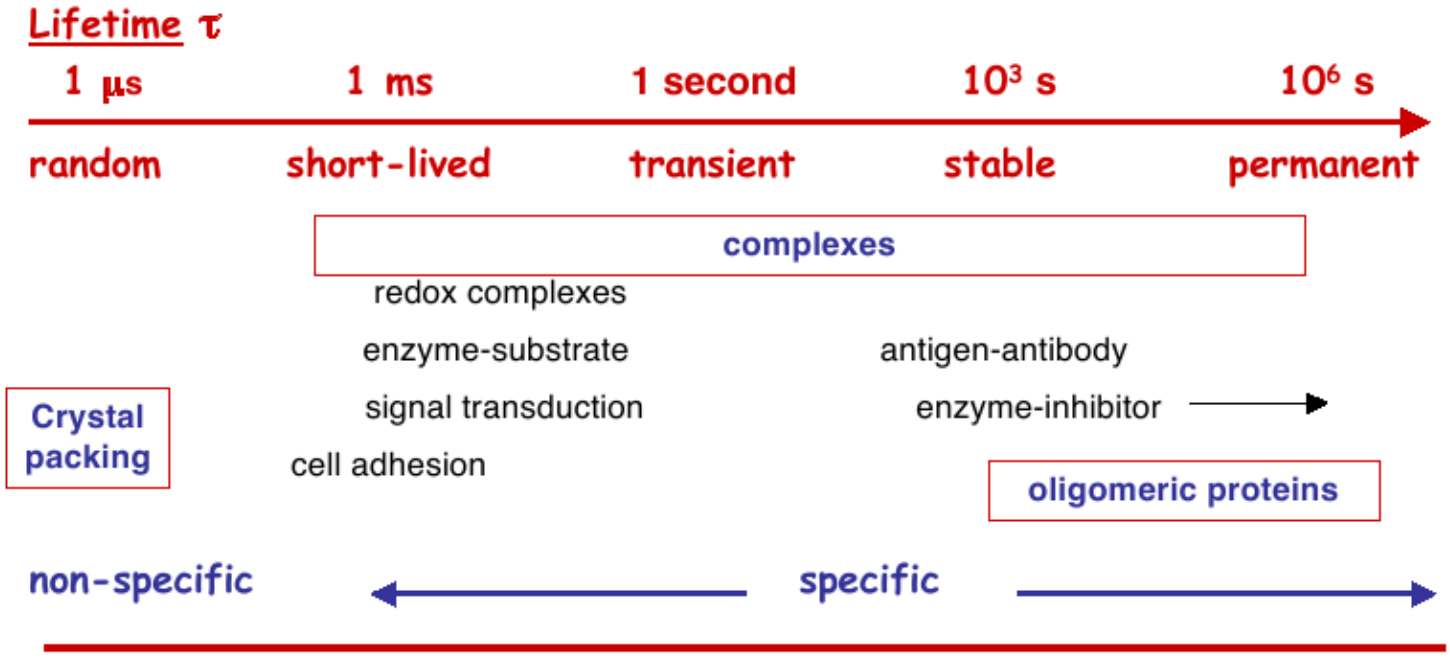
[J. Janin]

▷Ref: Janin et al, Quaterly reviews of biophysics, 2008

Biological complexes: time-wise

▷ Time-wise, biological complexes also span several orders of magnitude, say from the millisecond to years for permanent ones (Fig. **??**).

▷ Biological complexes: time scales



Short-lived complexes (τ<1 second) are relevant to many important biologically processes.
Only a few examples of these are present in the PDB (Nooren & Thornton, 2003).
These systems may resemble **crystal packing** more than permanent assemblies.

[J. Janin]

▷Ref: Janin et al, Quaterly reviews of biophysics, 2008

Docking models

- Over the years, several docking models have been proposed (Fig. 6):
  - Lock-and-key Fisher, 1894. In this model, the two partners associate as rigid bodies.
  - Induced fit: Koshland, 1958. While getting close, the partners *shape* one-another, resulting in the conformations found in the complex.
  - Conformer selection, Monod-Wyman-Changeux, 1965. In solution or in the cell, each molecule exists in a variety of conformations. In the course of their diffusion, *compatible* conformations stumble onto one-another, and the complex gets formed.

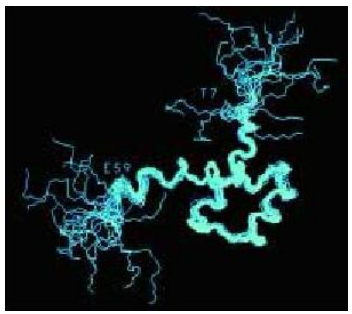


Figure: **Flexibility of biomolecules: illustration.** From the Nobel lecture of K. Wütricht.

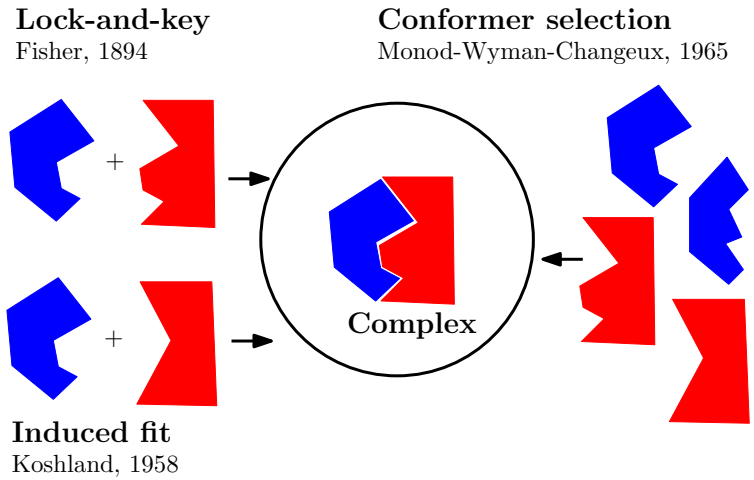
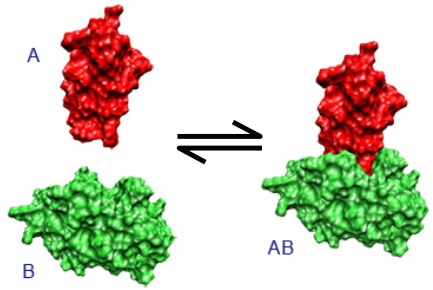


Figure: **Docking models.**

Binding affinity: dissociation free energy

- Protein complexes rock back and forth



- Dissociation constant / free energy as a function of concentrations:

$$K_d = [A][B]/[AB]$$
$$\Delta G_d = -RT \ln K_d/c^\circ = \Delta H - T\Delta S.$$

- Binding affinities (thermodynamics):
  - random complex: $K_d \sim 10^{-6}$
  - high: $K_d \sim 10^{-9}$
  - very high: $K_d \sim 10^{-12}$
  - extreme: $K_d \sim 10^{-15}$

- Time scales (kinetics):
  - short-lived complexes: $10^{-6}s$ (e.g. enzyme-substrate)
  - stable complexes: $10^{3}s$ (e.g. antibody-antigen)
  - permanent complexes: $10^{6}s$ (aggregates)

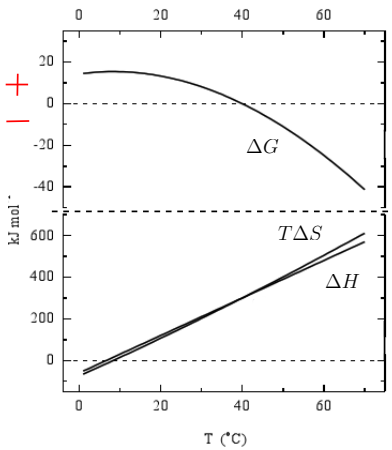
Binding affinity: thermodynamics

- Dissociation constant k_D for $C \leftrightharpoons A + B$:

$$K_d = \frac{[A][B]}{[C]}; \Delta G_d = -RT \ln K_d/c^\circ = \Delta H - T\Delta S. \tag{6}$$

- The enthalpy - entropy compensation:
  - enhanced packing of interface atoms due to attractive forces: $\Delta H < 0$
  - higher packing, restricted atomic motions: $T\Delta S < 0$

- Marginal stability of proteins and complexes:



  - Large ΔH and $T\Delta S$ compensate
  - Crossing of curves difficult to predict
  - Marginals stability is key to regulation

Pict. courtesy of Alan Cooper (Thermodynamics of unfolding)

Enthalpy - entropy compensation – thermodynamics

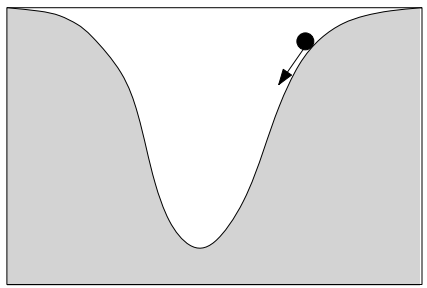
▷ Energy: loose some, gain some...

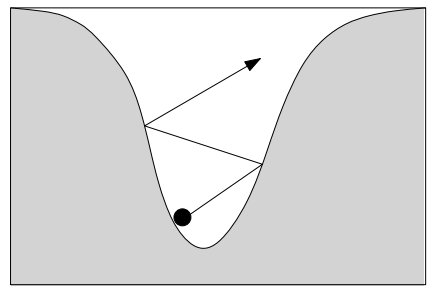
Energy minimization:

- $\Rightarrow \Delta H$ wants to be negative...

At the moleular level:

- thermal/Brownian motion tend to make things the other way around
- $\Rightarrow T\Delta S$ wants to be positive
- $\Rightarrow$ Variations counted as $-T\Delta S$





▷ Balance between these two tendencies: Gibbs free energy

▷Ref: Cooper, Biophysical Chemistry, Royal Society of Chemistry, text #24, Cambridge, 2001

Binding affinity: spectrum

- Typical binding affinity values are presented in Table 7.

Type of Interaction	K_D (molar)	ΔG^0_{bind} (at 300K) kJ mol^{-1}
Enzyme:ATP	~1×10^{-3} to ~1×10^{-6} (millimolar to micromolar)	–17 to –35
signaling protein binding to a target	~1×10^{-6} (micromolar)	-35
Sequence-specific recognition of DNA by a transcription factor	~1×10^{-9} (nanomolar)	-52
small molecule inhibitors of proteins (drugs)	~1×10^{-9} to ~1×10^{-12} (nanomolar to picomolar)	–52 to –69
biotin binding to avidin protein (strongest known non-covalent interaction)	~1×10^{-15} (femtomolar)	-86

Figure: **binding affinity: typical examples.** Table from Kuriyan et al.

- Ref: Kuriyan et al, The molecules of life, 2012

Binding affinity predictions: setting a reasonable goal

- Binding affinity measurements, experiments:
  ITC, SPR, and titration by fluorescence, with
  typical error range 0.1 - 0.25 kcal/mol

- Change in ΔG_d of 1.4, 2.8 and 4.2 kcal/mol:
  Change in K_d of 10x, 100x, 1000x respectively

- Binding affinity is a thermodynamic quantity, depending on:
  concentration, temperature, ionic strength, pH

- Changing the pH in the range 5.5 – 8.5: can change ΔG_d by 1.4-2.3 kcal/mol

- Methods not taking into account the solvent/dynamics:
  cannot claim an accuracy well beyond $\sim$ 1.4 kcal/mol

- `Ref: Kastritis et al; Protein Science (20), 2011`

Algorithms and Learning for Protein Science

Biomolecular recognition: proteins and binding affinity

The time dimension: $K_d, K_{\text{on}}, K_{\text{off}}$ and $1/K_{\text{off}}$

Binding affinity estimation: hardness

Application: influenza - antibody complexes

Chemical equilibrium

▷ Setup: we consider a protein P and a ligand L which interact in a non-covalent fashion. This means that no chemical bonds get created or removed. We further assume that these two species for a chemical equilibrium:

$$P + L \rightleftharpoons PL, K_{eq} = \frac{[PL]_{\text{Eq.}}}{[P]_{\text{Eq.}}[L]_{\text{Eq.}}} \tag{7}$$

▷ The notion of *equilibrium* is central here, and owes to competing effects:

- Due to attraction forces, P and L get closer to one another.
- Due in particular to thermal fluctuations, they get away.

▷ In the medium considered (test tube, cell): three chemical species: P, L, and the complex PL.

▷ In the sequel, we consider the standard setup:

- We start from std concentrations of the individual species, say 1 Molar
- We consider the equilibrium concentrations

Equilibrium constants K_a, K_d

- Consider the non-covalent interaction $P + L \rightleftharpoons PL$
- The law of mass action yields the association and dissociation constants:

$$\begin{cases} \text{Association constant} : K_a = \dfrac{[PL]_{\mathbf{Eq.}}}{[P]_{\mathbf{Eq.}}.[L]_{\mathbf{Eq.}}} \\ \text{Dissociation constant} : K_d = \dfrac{[P]_{\mathbf{Eq.}}.[L]_{\mathbf{Eq.}}}{[PL]_{\mathbf{Eq.}}} \end{cases} \tag{8}$$

Using std units, K_a is expressed in moles^{-1}, and K_d is in moles.

- Determine the concentration of the molecular species, here P, L, and PL, when the binding reaction reaches an equilibrium.
- The relationship between K_a and the variation of free energy satisfies:

$$\Delta G_a^0 = -RT \log c^0 K_a = RT \log \frac{K_d}{c^0}. \tag{9}$$

- Rmk. In Eq. 9, c^0 is meant to obtain a unit-less number: if K_a is expressed in moles^{-1} ,then C^0 is equal to 1 molar.

Fractional saturation

- The fraction of proteins with bound ligand satisfies:

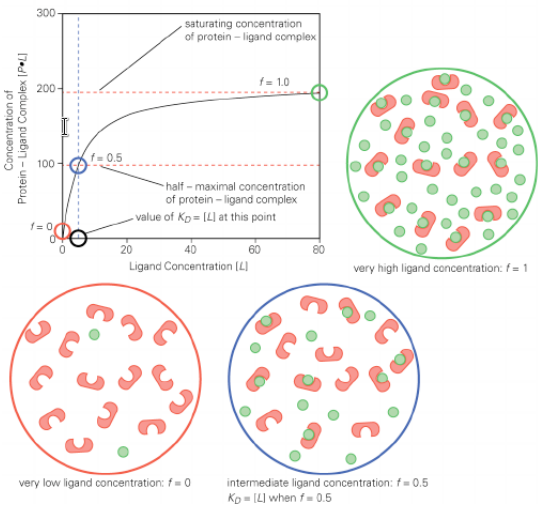
$$f = \frac{\#\text{num proteins with bound ligand}}{\text{total } \#\text{ proteins}} \tag{10}$$

$$= \frac{[PL]_{\text{Eq.}}}{[P]_{\text{Eq.}} + [PL]_{\text{Eq.}}} \tag{11}$$

$$= \frac{[P]_{\text{Eq.}}[L]_{\text{Eq.}}}{K_d([P]_{\text{Eq.}} + \frac{[P]_{\mathbf{Eq.}}[L]_{\mathbf{Eq.}}}{K_d})} \tag{12}$$

$$= \frac{1}{K_d(\frac{1}{K_d} + \frac{1}{[L]_{\mathbf{Eq.}}})} = \frac{[L]_{\text{Eq.}}}{[L]_{\text{Eq.}} + K_d} \tag{13}$$

- Varying the concentration of the ligand, one gets from Eq. 10:
- Observation: K_d is the concentration of the ligand such that the fraction of bound equals 1/2.



Equilibrium constants and reaction rates

- To account for kinetics, one resorts to the reaction rates

$$P + L \underset{K_{\mathbf{off}}}{\overset{K_{\mathbf{on}}}{\rightleftharpoons}} PL. \tag{14}$$

Note that $K_{\mathbf{on}}$ is expressed $mol^{-1}s^{-1}$ while $K_{\mathbf{off}}$ is expressed in s^{-1}. These rates account for the fact that in order to assemble, the molecules must first meet/collide.

- The relationship with dissociation is as follows:

$$K_a = \frac{K_{\mathbf{on}}}{K_{\mathbf{off}}}. \tag{15}$$

Residence times

Binding affinity is a thermodynamic quantity. On the other hand, time is clearly involved in biomolecular interactions – Chapter **??**.

- Mean life of the complex, $1/K_{off}$: average life span of the PL complex.
- Half-time of the complex, $\log 2/K_{off}$: the time required for half of a population of complexes to unbind.

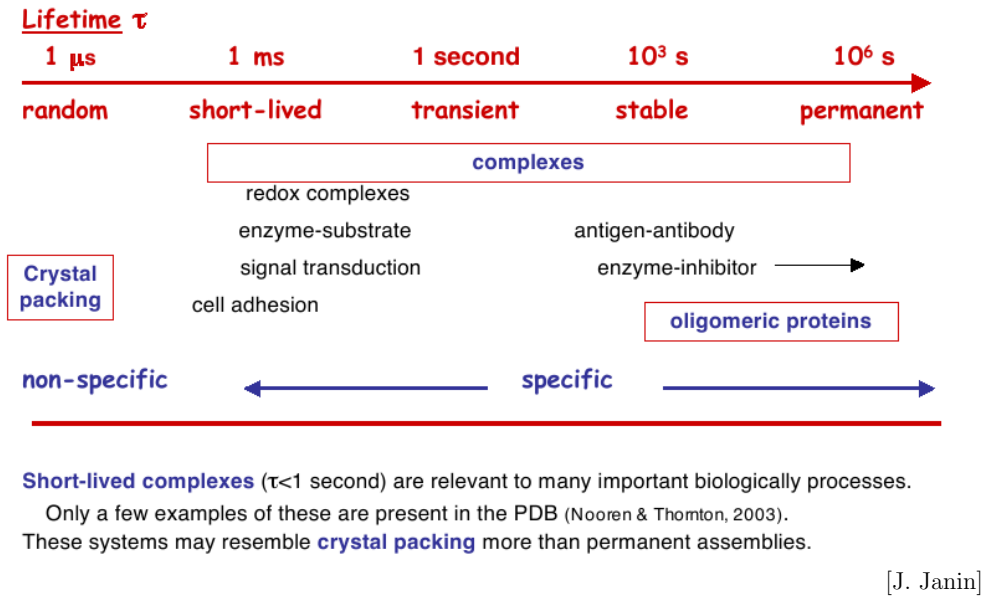


Figure: **Biological complexes: time scales.** From Janin et al.

▷Ref: Janin et al, Quaterly reviews of biophysics, 2008

Algorithms and Learning for Protein Science

Biomolecular recognition: proteins and binding affinity

The time dimension: $K_d, K_{\text{on}}, K_{\text{off}}$ and $1/K_{\text{off}}$

Binding affinity estimation: hardness

Application: influenza - antibody complexes

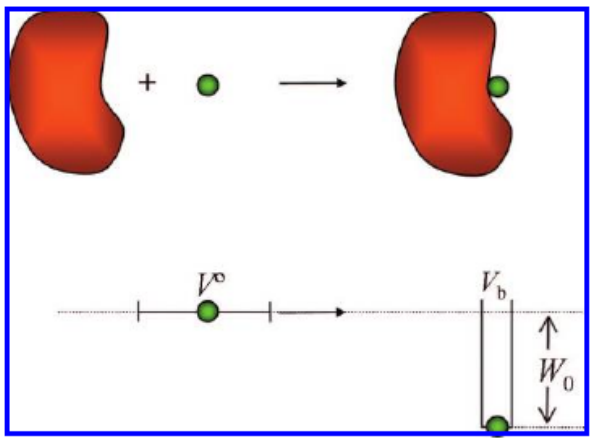
Enthalpy - entropy compensation - I

▷ To understand the components of binding, let us recall:

$$\Delta G_a^0 = -RT \log c^0 K_a = RT \log \frac{K_d}{c^0} = \Delta H - T\Delta S. \tag{16}$$

▷ To understand the relative variations of ΔH and $T\Delta S$, we need to discuss several components in turn:

- (1) System protein + ligand, enthalpy
- (2) Mixing: Two versus three species
- (3) Ligand and its translational / rotational entropy
- (4) System protein + ligand, conformational + vibrational entropy
- (5) Solvent and its entropy



Binding affinity: enthalpy-entropy competition illustrated along the binding process. The volume accessible to the ligand decreases, whence $T\Delta S < 0$ and $-T\Delta S > 0$. On the other hand, the interaction energy (enthalpy) decreases by W_0. From [**?**].

Enthalpy - entropy compensation - II

1. **System protein + ligand, enthalpy:**
   - Energy minimization when P and L get closer. (Exple: strong electrostatic interactions.)
2. **Mixing: Two versus three species:**
   - Three species (P, L, PL) have more entropy than two.
3. **Ligand and its translational / rotational entropy:**
   - Assuming P fixed: 6 dof of the ligand get constrained. Translation/rotational entropy decreases.
4. **System protein + ligand, conformational + vibrational entropy:**
   - In PL, conformational changes hindered + coupled harmonic oscillators: conformational and vibrational entropy decrease.
5. **Solvent and its entropy:**
   - Buried surface area at the interface $\Rightarrow$ the solvent S increases.

Summary:

- During association, grossly speaking: ΔH is negative, and $-T\Delta S$ is positive.
- Variation of enthalpy and entropy are very subtle, and the balance depends in general on the temperature.
- For biological systems: this subtlety is key to **regulation**. By slightly changing the conditions (temperature, pH, ionic strength which alter the electrostatic interactions), the behavior changes.

Enthalpy-entropy competition: illustration on protein unfolding

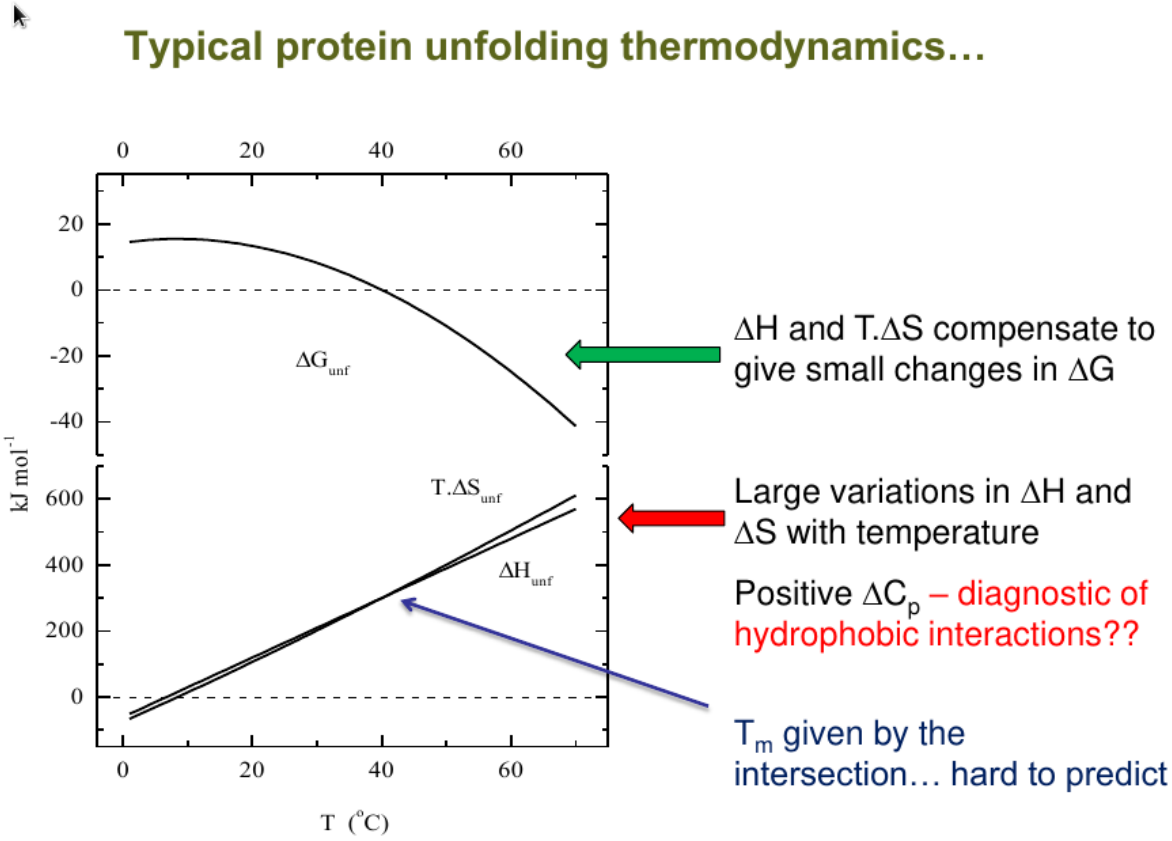
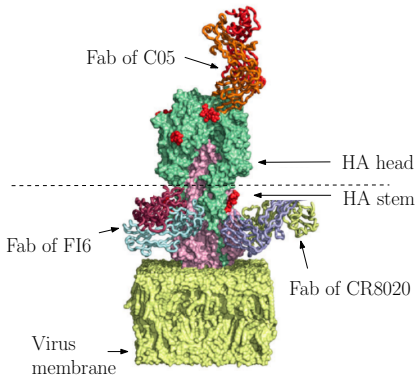


Figure: **Protein unfolding: illustration of the enthalpy-entropy competition.** Courtesy of Alan Cooper.

Partition functions determine macroscopic properties

Example: binding affinity as ratio of Zs

- A standard antibody-antigen complex:



- Model without solvent:
  - FAB of antibody $\sim 3k$ atoms
  - Hemaglutinin $\sim 14k$ atoms
  - One conformation: 1 point in $\mathbb{R}^{51,000}$

- ΔG_d as a multidimensional integral with V the PE and W the solvent PMF:

$$\Delta G_d = -\frac{1}{\beta} \ln \frac{Z_{IG} Z_{Ag}}{Z_{complex}} = -\frac{1}{\beta} \ln \left(\frac{8\pi^2}{c^\circ} \frac{\int e^{-\beta(V(r_A)+W(r_A))} dr_A \times \int e^{-\beta(V(r_B)+W(r_B))} dr_B}{\int e^{-\beta(V(r_C)+W(r_C))} dr_C} \right)$$

▷Ref: Woo ad Roux, PNAS 102 (19), 2005

▷Ref: Gilson and Zhou, Ann. Rev. Biophys. Biomol. Struct., 36, 2007

Algorithms and Learning for Protein Science

Biomolecular recognition: proteins and binding affinity

The time dimension: $K_d, K_{\text{on}}, K_{\text{off}}$ and $1/K_{\text{off}}$

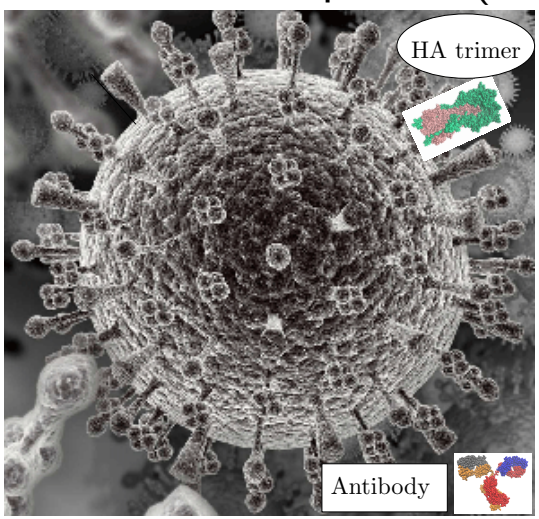
Binding affinity estimation: hardness

Application: influenza - antibody complexes

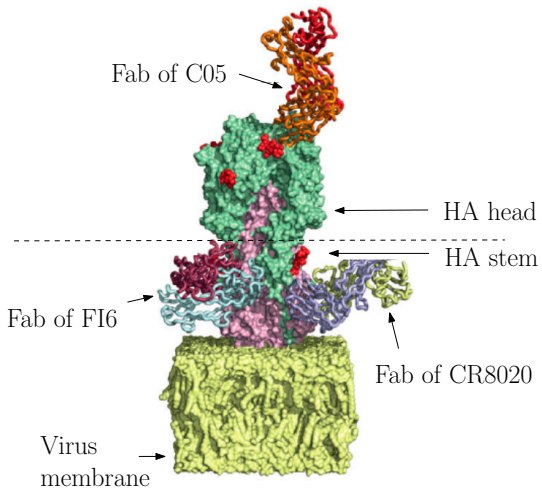
Virus neutralization by antibodies: the problem

- Enveloped viruses: the case of influenza
- Broadly neutralizing antibodies targeting the fusion protein of influenza:
  - Ig on top: prevent the virus attachment
  - Ig on stem: preventing the conformational changes required for envelope-membrane fusion

- The influenza virus. Drawn to scale a trimer of the fusion protein (HA)

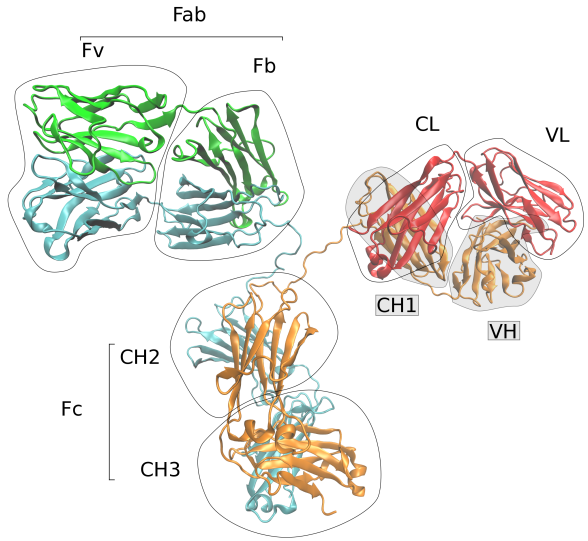


- Broadly neutralizing antibodies : hemaglutinin (HA) of influenza is depicted in green

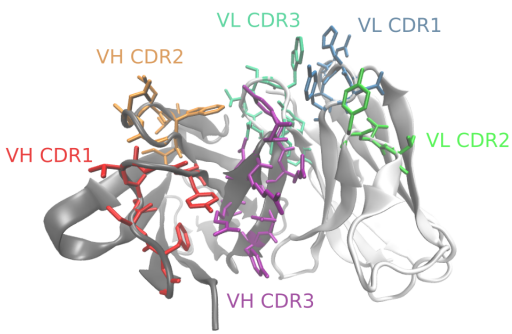


The structure of antibodies – IgG immunoglobulins

- Overall structure



- FABs and CDRs



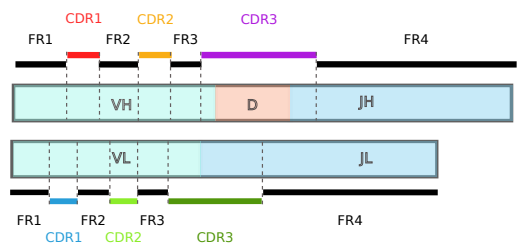


Figure: **(A) Antigen-binding fragment (FAB) and Complementarity Determining Regions (CDRs) (B) Encoding of CDRs and Frs by the V, D and J genes**

Affinity maturation: process

- Affinity maturation: secretion of more potent antibodies
- IgG lineage

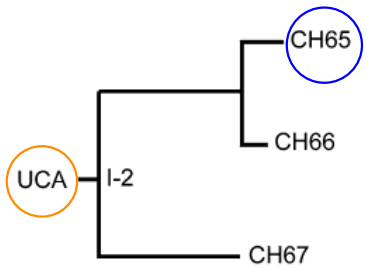


Figure: **Lineage of IgG observed during an immune response against influenza.**

- Evolution of the affinity

Fab	$K_d(\mu M)$
UCA	118 ± 14
I-2	142 ± 15
CH65	$0.49 \pm .10$
CH67	0.36 ± 0.04

Table: **Binding affinities: K_d analysis by SPR** NB: CH65 $\sim$ CH67; wrt UCA: $\Rightarrow \sim$ 200-fold improvement

Affinity enhancement: origin

▷ Ancestor and matured IgG have similar binding modes

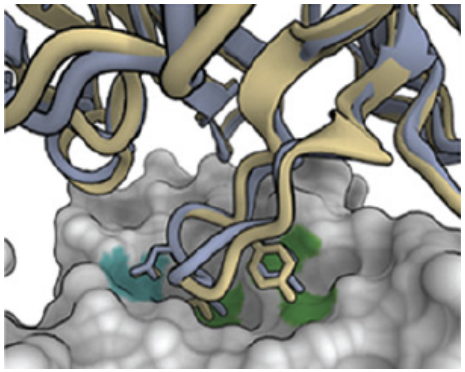


Figure: **But UCA and CH65 have similar binding modes.** Displayed: backbone traces of the CDR3. From Schmidt et al.

▷ But matured IgG have a pre-formed binding site:

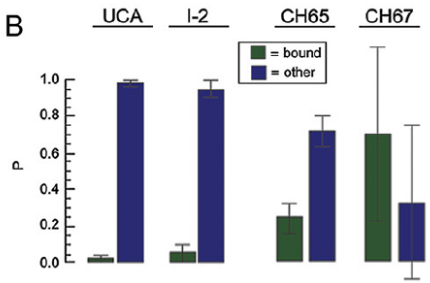


Figure: **CDR3: time spent in bound and unbound conformations.** Maturated IG (CH65, CH67): more time in the bound conformation. From Schmidt et al.

▷ Origin of the affinity enhancement: lesser entropic penalty.
"*In both branches (CH65, CH67)), increased conformational restriction of CDR H3 has been the principle consequence of affinity maturation.* "

▷Ref: Schmidt et al., PNAS, 2013

Binding affinity and specificity

- The two critical notions for protein interactions are
  - Binding affinity: the *strength* of the interactions.
  - Binding specificity: the variety of partners a molecules binds sufficiently strongly with.

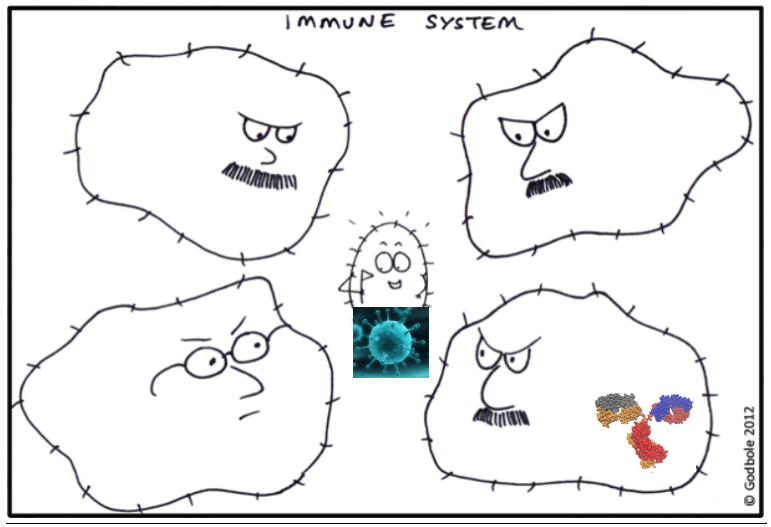


Figure: **Binding affinity and specificity: how to for the immune system.** The molecules secreted should bind strongly enough the pathogens; but they should also be quite specific.

Algorithms and Learning for Protein Science

PART 1: Introduction
PART 2: Protein structure and resolution
PART 3: Physical chemistry & dynamics
PART 4: Biomolecular interactions and binding affinity
PART 5: Outlook