Algorithms and Learning for Protein Science

Motions and energies

Frederic.Cazals@inria.fr

Overview

- Algorithms
  - Independent Component Analysis (ICA)
  - time lagged ICA (tICA)
- Theoretical biophysics
  - Harmonic oscillators in several guises
  - Notions of statistical physics
- Molecular science
  - Formation of clusters of hard spheres
  - Slow modes in protein motions

Structural alignments and analysis

Normal modes

- Intro
- Vibrational normal modes
- Elastic Network Models: pre-requisites
- Gaussian Network Model
- Anisotropic Network Model

A primer in statistical physics

- Notations
- The (canonical) Boltzmann distribution
- The classical 1D harmonic oscillator
- Polyatomic molecules: the rigid-rotor harmonic oscillator approximation
- Application: free energy landscape of hard spheres

Independent component analysis

tICA

- Method
- Application: one example

Normal modes: movie

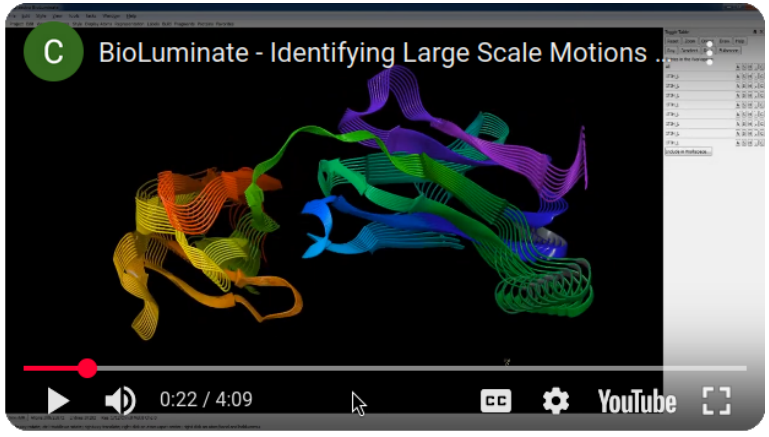


Figure: **NMA in** `Schrödinger`

Normal modes: overview

- Pros.
  - Based on physics, but coarse grain – much faster than molecular dynamics
  - Theory of normal modes: physics of vibrations, which exhibit *correlated motions*
  - At low temperature ($< 250K$) do vibrate like solids
  - At room temperature, a number of biological mechanisms exploit low frequency correlated motions
    - opening closing of enzymes
    - functioning of the ribosome
    - functioning of hemoglobin, etc
- Cons.
  - Harmonic models are local
  - No anharmonic effects
  - On the other hand: mechanisms can exploit harmonic effects at early stages, and then proceed with the energy provided/released

Normal modes and correlations

- Goal: study covariances between atomic displacements

$$\begin{cases} C(i,j) & =< \langle \Delta X_i, \Delta X_j \rangle >, \\ \text{corr(i,j)} & = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}}. \end{cases} \tag{1}$$

- Two types of models
  - Vibrational normal modes
  - Energetic normal modes / elastic network models:
    - (GNM) Gaussian network model: correlation using whole fluctuations vector ΔX_i
    - (ANM) Anisotropic network model: understanding directional preferences along the three coordinate axis, via the decomposition $\Delta X_i = \Delta x_i + \Delta y_i + \Delta z_i$

Vibrational normal modes: generalized eigenproblem

▷ **Kinetic energy:** With M the diagonal matrix of individual masses

$$K = \frac{1}{2}\dot{u}^{\mathbf{T}} M \dot{u} \tag{2}$$

▷ **Potential energy:** locally given by a quadratic form

$$V = \frac{1}{2}u^{\mathbf{T}} H u, \text{t with } H_{ij} = \frac{\partial^{\mathbf{2}} V}{\partial u_i \partial u_j}_{|\{u_i\}=\mathbf{0},\{\dot{u}_i\}=\mathbf{0}} \tag{3}$$

▷ **The Lagrangian of the system:**

$$L = \frac{1}{2}\dot{u}^{\mathbf{T}} M \dot{u} - \frac{1}{2}u^{\mathbf{T}} H u. \tag{4}$$

The Euler-Lagrange equations gives

$$M\ddot{u} + Hu = 0. \tag{5}$$

With $A = (a_{\mathbf{1}}, \ldots, a_n)^{\mathbf{T}}$ – the a_i are amplitudes, we seek a solution of the form $u(t) = A\exp(i\omega t)$:

$$\begin{cases} u(t) = A\exp(i\omega t), \\ \dot{u}(t) = i\omega u(t), \\ \ddot{u}(t) = -\omega^{\mathbf{2}} u(t). \end{cases} \tag{6}$$

Plugging into Eq. 5 yields the following **generalized eigenvalue problem:**

$$HA = \omega^{\mathbf{2}} MA, \tag{7}$$

$\Rightarrow$ the vector A is solution of a generalized eigenproblem.

Solving the generalized eigenproblem

The previous equation can be rewritten as

$$HA - \omega^2 MA = (H - \omega^2 M)A = 0. \tag{8}$$

This is a system of n homogeneous linear equations, so that we aim at solving

$$\det(H - \omega^2 M) = 0. \tag{9}$$

Expanding yields a n-th order polynomial in ω^2.

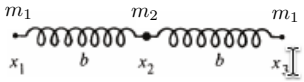
▷ Remark. Eq. 5: Newton's equations of motion in the form

$$M\ddot{q} = F = -\nabla V. \tag{10}$$

Assume the potential energy is invariant by rigid motion. By Newton's equations – Eq. 10, no forces are applied to the system: there are actually exactly six null frequencies / eigenvalues.

Normal modes, example: the triatomic molecule

▷ Model



▷ Potential

$$V = \frac{k}{2}(x_2 - x_1 - b)^2 + \frac{k}{2}(x_3 - x_2 - b)^2. \quad (11)$$

We introduce coordinates wrt equilibrium positions. With $x_{3,0} - x_{2,0} = x_{2,0} - x_{1,0} = b$:

$$\nu_i = x_i - x_{i,0}, \quad (12)$$

Then

$$V = \begin{pmatrix} k & -k & 0 \\ -k & 2k & -k \\ 0 & -k & k \end{pmatrix}, M = \begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_1 \end{pmatrix} \quad (13)$$

$$\det(V - \omega^2 M) = \det(\begin{bmatrix} k - \omega^2 m_1 & -k & 0 \\ -k & 2k - \omega^2 m_2 & -k \\ 0 & -k & k - \omega^2 m_1 \end{bmatrix}) \quad (14)$$

Solving

$$\omega^2(k - \omega^2 m_1)(k(m_2 + 2m_1) - \omega^2 m_1 m_2) = 0 \quad (15)$$

with the following solutions

$$\omega_1 = 0, \omega_2 = \sqrt{\frac{k}{m_1}}, \omega_3 = \sqrt{\frac{k}{m_1}(1 + \frac{2m_1}{m_2})}. \quad (16)$$

NB: $\omega_1 = 0$ is expected, see previous Remark on null eigenvalues.

Mass weighted coordinates (I)

- Eigenproblem revisited: $A \to e_i$, and $\omega_i^2 \to \lambda_i$:

$$He_i = \lambda_i M e_i. \tag{17}$$

Proposition. 1. The normal mode vectors solutions of the generalized eigenproblem are M-orthogonal:

$$e_i{}^{\mathsf{T}} M e_j = \delta_{ij}. \tag{18}$$

Hint: from $He_i = \lambda_i M e_i$, transposing and multiplying by e_j ...

Proposition. 2. Using normal mode coordinates $\{q_i\}$, the potential satisfies

$$H = \frac{1}{2} \sum_i \lambda_i q_i^2. \tag{19}$$

Hint: manipulate $H = \frac{1}{2} u^{\mathsf{T}} H u$ using $u = \sum_i q_i e_i$

Mass weighted coordinates (II)

Normal mode vectors are also solutions of a std eigenproblem with modified vector/matrices, yielding *mass weighted coordinates*:

- Let $f_i = M^{1/2} e_i$

We have

$$
\begin{aligned}
& He_i = \lambda_i M e_i && (20)\\
\Leftrightarrow & HM^{-1/2} M^{1/2} e_i = \lambda_i M^{1/2} M^{1/2} e_i && (21)\\
\Leftrightarrow & HM^{-1/2} f_i = \lambda_i M^{1/2} f_i && (22)\\
\Leftrightarrow & M^{-1/2} HM^{-1/2} f_i = \lambda_i f_i && (23)\\
\Leftrightarrow & \tilde{H} f_i = \lambda_i f_i, \text{ with } \tilde{H} = M^{-1/2} HM^{-1/2}. && (24)
\end{aligned}
$$

Intermezzo: the equipartition theorem

- Canonical NVT ensemble: Boltzmann's distribution applies
  - NB: in the microcanonical NVE ensemble: the system has a fixed total energy, there is no Boltzmann's factor, and there is a uniform probability for conformations.

- Equipartition theorem: the average energy associated with each quadratic degree of freedom (dof) is

$$< E_i >= \frac{1}{2} k_B T$$

- For one dof of the the harmonic oscillator: in the NM basis:

$$E_i = \frac{1}{2} \lambda_i q_i^2, \tag{25}$$

and we get

$$< E_i > = \frac{1}{2} k_B T = \frac{1}{2} \lambda_i < q_i^2 >, \tag{26}$$

and therefore

$$< q_i^2 >= \frac{k_B T}{\lambda_i} \tag{27}$$

- NB: the stiffer the spring/larger λ_i, the smaller $< q_i^2 >$

Pairwise correlations

▷ Goal: compute C_{ab} using the NM basis $\{e_i\}$

Proposition. 3. The pairwise correlations are given by

$$C_{ab} =< u_a u_b >= k_B T \sum_i \frac{e_{i,a} e_{i,b}}{\lambda_i}. \tag{28}$$

Proof. Rewrite an original displacement vector in the basis of NM:

$$u = \sum_i q_i e_i. \tag{29}$$

For two original coordinates u_a and u_b: $u_a = \sum_i q_i e_{i,a}$, $u_b = \sum_j q_j e_{j,b}$. Whence

$$C_{ab} =< u_a u_b >= \sum_{ij} < q_i q_j > e_{i,a} e_{j,b}. \tag{30}$$

But since NM coordinates are uncorrelated, that is $< q_i q_j >= 0$ if $i \neq j$, we get

$$< u_a u_b >= \sum_i < q_i^2 > e_{i,a} e_{i,b}. \tag{31}$$

But each q_i is an independent harmonic oscillator, the equipartition theorem gives the following expectation

$$\mathbb{E}\left[q_i^2\right] =< q_i^2 >= \frac{k_B T}{\lambda_i}, \tag{32}$$

whence Eq. 3. □

Application: atomic fluctuations

- Atom i: three Cartesian coordinates, whence

$$\Delta X_i^2 = \Delta x_i^2 + \Delta y_i^2 + \Delta z_i^2. \tag{33}$$

Denote $J = \{j_1, j_2, j_3\}$ the indices of these coordinates, the linearity of expectation gives

$$< \Delta R_i^2 >= k_B T \sum_i \sum_{j \in J} \frac{e_{i,j} e_{i,j}}{\lambda_i}. \tag{34}$$

Should be compared to the experimentally measured B-factor, defined by

$$B_i = 8\pi^2 < u_i^2 > . \tag{35}$$

Elastic Network Models

- Goal: depart from full atom models and (complicated) atomic force fields

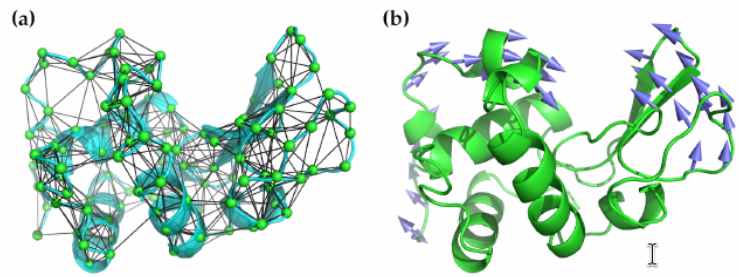


Figure: **Elastic network models GNM/ANM: example.** In both cases, one attaches a spring to residues within a distance threshold.

Graph Laplacian: pre-requisites (I)

Consider the standard Laplacian $L = D - W$ of the graph, with $w_{ij} = 1$ iff the nodes i and j are connected, and D the generalized degree defined by $d_i = \sum_j w_{ij}$.

Lemma 1. For any vector $F = (f_1 \dots f_n)^\mathsf{T} \in \mathbb{R}^d$, the Laplacian satisfies

$$F^\mathsf{T} L F = \frac{1}{2} \sum_{i,j} w_{ij} (f_i - f_j)^2. \tag{36}$$

Proof.

$$F^\mathsf{T}(D - W)F = \sum_i d_i f_i^2 - \sum_{ij} w_{ij} f_i f_j \tag{37}$$

$$= \frac{1}{2}\Big(\sum_i \big(\sum_j w_{ij}\big) f_i^2 - 2\sum_{ij} w_{ij} f_i f_j + \sum_j \big(\sum_i w_{ij}\big) f_j^2\Big) \tag{38}$$

$$= \frac{1}{2} \sum_{i,j} w_{ij} (f_i - f_j)^2. \tag{39}$$

□

▷ Question: positive definite or semi-definite ?

Graph Laplacian: pre-requisites (II)

Proposition. 4. Consider a connected graph. The eigenvalues of its Laplacian satisfy $\mu_1 \geq \cdots \geq \mu_{n-1} > \mu_n = 0$.
Moreover, the pseudo-inverse satisfies of Γ satisfies:

$$\Gamma^{\dagger} = \sum_{k=1}^{n-1} \frac{1}{\mu_k} u_k u_k^{\mathsf{T}}, \tag{40}$$

or

$$[\Gamma^{\dagger}]_{ij} = \sum_{k=1}^{n-1} \frac{1}{\mu_k} u_{ik} u_{jk}. \tag{41}$$

▷ NB: one null eigenvalue: the previous quadratic form is positive semi-definite.

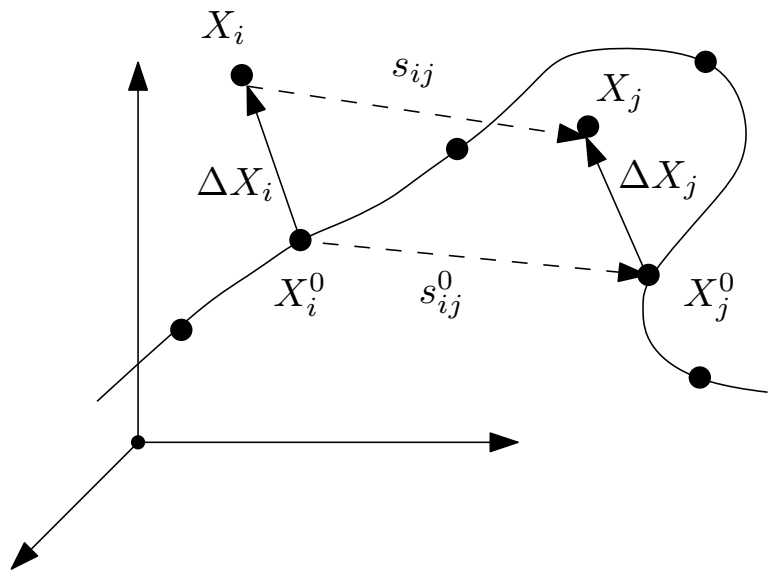
▷ Remark. The pseudo-inverse can also be obtained using the SVD of L, that is $L = USV^{\mathbf{T}}$. If the inverse exists, it satisfies $L^{-\mathbf{1}} = VS^{-\mathbf{1}}U^{\mathbf{T}}$. But since L is symmetric, we get $L^{-\mathbf{1}} = US^{-\mathbf{1}}U^{\mathbf{T}}$. Now, since the graph is connected, it has a single null eigenvalue, so that the pseudo-inverse is given by Eq. 40.

Gaussian Network Model: geometric model

▷ Focus: C_α carbons and their displacements ΔX_i with respect to the equilibrium positions

▷ Vector equality: for $C_{\alpha;i}$ and $C_{\alpha;j}$

$$\overset{\circ}{s}_{ij} + \Delta X_j - s_{ij} - \Delta X_i = 0 \Rightarrow s_{ij} - \overset{\circ}{s}_{ij} = \Delta X_j - \Delta X_i. \tag{42}$$



We also use the following notations for distances: $d_{ij} = \left\|s_{ij}\right\|$, $\overset{\circ}{d}_{ij} = \left\|\overset{\circ}{s}_{ij}\right\|$.

GNM: potential

- Pairwise potential:

$$V_{ij} = \frac{w_{ij}}{2} \left\| s_{ij} - \overset{\circ}{s}_{ij} \right\|^2 \tag{43}$$

$$= \frac{w_{ij}}{2} \left\| \Delta X_i - \Delta X_j \right\|^2 \tag{44}$$

- wlog, using $w_{ij} = \gamma$:

$$V_{GNM} = \frac{\gamma}{2} \sum_{ij} (\Delta X_j - \Delta X_i)^{\mathsf{T}} (\Delta X_j - \Delta X_i). \tag{45}$$

- Laplacian as a quadratic form:

$$V = \frac{\gamma}{2} \Delta X^{\mathsf{T}} \Gamma \Delta X. \tag{46}$$

Using this potential, one typically uses a unit matrix M for masses, to that the generalized eigenproblem of Eq. 7 become a standard eigenproblem:

$$HA = \lambda MA \quad \Rightarrow \quad \Gamma A = \lambda A. \tag{47}$$

That is, the eigenvectors are those of the Laplacian.

Fluctuations

Assume now that a given conformation is given by Boltzmann's distribution. We derive the fluctuations using the pseudo-inverse of the Laplacian:

$$\mathbb{P}[X] = \frac{1}{Z_n} \exp(-\frac{V_{GNM}(X)}{k_B T}) = \frac{1}{Z_n} \exp(-\frac{\gamma}{2k_B T} \Delta X^{\mathsf{T}} \Gamma \Delta X). \tag{48}$$

The atomic fluctuations, akin to covariances, are given by

$$\begin{aligned} Cij =< \langle \Delta X_i, \Delta X_j \rangle > &= \mathbb{E}\left[\langle \Delta X_i, \Delta X_j \rangle\right] && (49)\\ &= \frac{1}{Z_n} \int \langle \Delta X_i, \Delta X_j \rangle \exp(-\frac{V(X)}{k_B T}) d\Delta X && (50)\\ &= \frac{1}{Z_n} \int \langle \Delta X_i, \Delta X_j \rangle \exp(-\frac{\gamma}{2k_B T} \Delta X^{\mathsf{T}} \Gamma \Delta X) d\Delta X && (51)\\ &= [\Gamma^{\dagger}]_{ij} && (52)\\ &= \sum_{k=1}^{n-1} \frac{1}{\mu_k} u_{ik} u_{jk}. && (53) \end{aligned}$$

The penultimate line comes from the definition of the covariance for a multivariate Gaussian, and the last one is the expression of the pseudo-inverse of the Laplacian – Eq. 40.

Using the previous, we can define a theoretical B-factors as [**?**]:

$$B_i = 8\pi^2 \frac{< \Delta X_i^2 >}{3}. \tag{54}$$

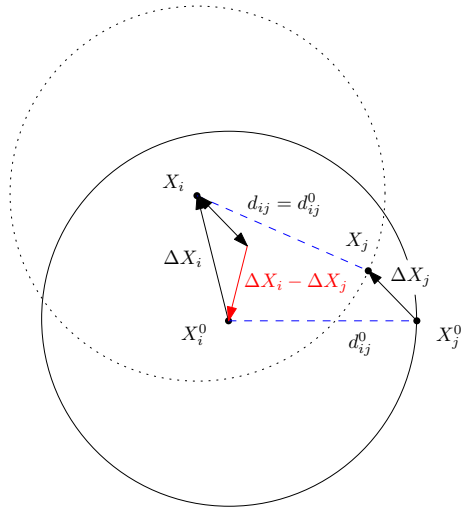
The Anisotropic Network Model

▷ Rationale: the GNM does not treat coordinates individually yielding isotropic motions ⇒ use a model coordinate-based

▷ Two potentials: GNM versus ANM

$$V_{GNM} = \frac{1}{2}\sum_{ij} k_{ij} \left\|\Delta X_j - \Delta X_i\right\|^2, \tag{55}$$

$$V_{ANM} = \frac{1}{2}\sum_{ij} k_{ij}(d_{ij} - \overset{\circ}{d}_{ij})^2. \tag{56}$$



Equal distances but different displacement vectors.

▷ Approximating V_{ANM}: Eq. 56 is not quadratic in the individual atomic coordinates. However, it can be checked that the initial positions $\{X_i = X_i^0\}$ define a local minimum, so that a second order Taylor expansion yields

$$V_{ANM} = \frac{1}{2}\sum_{ij} k_{ij}(d_{ij} - \overset{\circ}{d}_{ij})^2 \tag{57}$$

$$= \frac{1}{2}(X - X_0)^{\mathsf{T}} H_0 (X - X_0) + \text{ higher order terms.} \tag{58}$$

Hessian H_0: expression

- Goal: for N atoms, find out the $3N \times 3N$ Hessian matrix
- Overall:

$$H_0 = \begin{pmatrix} H_{11} & H_{12} \ldots & H_{1N} \\ H_{21} & & H_{2N} \\ \vdots & & \vdots \\ H_{N1} & & H_{NN}. \end{pmatrix} \tag{59}$$

- Off-diagonal blocks: The calculation yields:

$$H_{ij} = -\frac{k_{ij}}{d_{ij}^2} \begin{pmatrix} (x_j - x_i)^2 & (x_j - x_i)(y_j - y_i) & (x_j - x_i)(z_j - z_i) \\ (y_j - y_i)(x_j - x_i) & (y_j - y_i)^2 & (y_j - y_i)(z_j - z_i) \\ (z_j - z_i)(x_j - x_i) & (z_j - z_i)(y_j - y_i) & (z_j - z_i)^2 \end{pmatrix}_{X_i = X_i^0, X_j = X_j^0} \tag{60}$$

or, using the components of the vector $X_i X_j$

$$H_{ij} = -\frac{k_{ij}}{d_{ij}^2} \begin{pmatrix} X_{ij}X_{ij} & X_{ij}Y_{ij} & X_{ij}Z_{ij} \\ Y_{ij}X_{ij} & Y_{ij}Y_{ij} & Y_{ij}Z_{ij} \\ Z_{ij}X_{ij} & Z_{ij}Y_{ij} & Z_{ij}Z_{ij} \end{pmatrix}_{X_i = X_i^0, X_j = X_j^0} \tag{61}$$

- Off-diagonal blocks:

$$H_{ii} = -\sum_{j \neq i} H_{ij}. \tag{62}$$

Fluctuations

▷ Use V in a Gaussian model:

$$\mathbb{P}[X] = \frac{1}{Z_n} \exp(-\frac{V_{ANM}(X)}{k_B T}) = \frac{1}{Z_n} \exp(-\frac{1}{2k_B T} \Delta X^{\mathsf{T}} H_0 \Delta X). \tag{63}$$

▷ Compute a pseudo-inverse: using the $N-6$ eigenvalues/vectors $\{\lambda_i\}$ and $\{u_i\}$ of H_0

$$H_0{}^{\dagger} = \sum_{k=1}^{3N-6} \frac{1}{\lambda_i} u_i u_i{}^{\mathsf{T}}. \tag{64}$$

This pseudo-inverse is also organized in 3×3 blocks H_{ij}^{-1}, and the cross-correlation of Eq. 1 reads as:

$$C_{ij} =< \langle \Delta X_i, \Delta X_j \rangle >= \frac{\mathrm{Tr}(H_{ij}^{-1})}{\sqrt{\mathrm{Tr}(H_{ii}^{-1})\mathrm{Tr}(H_{jj}^{-1})}}. \tag{65}$$

This is an alternative to the expression of Eq. 53 obtained for the Gaussian network model.

Exercise: calculations

- Distance between atoms X_i and X_j:

$$d_{ij} = ((x_j - x_i)^{\mathbf{2}} + (y_j - y_i)^{\mathbf{2}} + (z_j - z_i)^{\mathbf{2}})^{\mathbf{1/2}}, \tag{66}$$

$$\overset{\circ}{d}_{ij} = d_{ij}|(X_i = X_i^{\mathbf{0}}, X_j = X_j^{\mathbf{0}}). \tag{67}$$

- Selected calculations to obtain $H_{\mathbf{0}}$:

$$\frac{\partial d_{ij}}{\partial x_i} = (x_i - x_j)\frac{\mathbf{1}}{d_{ij}} \tag{68}$$

$$\frac{\partial V_{ij}}{\partial x_i} = k_{ij}(x_i - x_j)(\mathbf{1} - \frac{d_{ij}}{\overset{\circ}{d}_{ij}}) \tag{69}$$

$$\frac{\partial^{\mathbf{2}} V_{ij}}{\partial x_i^{\mathbf{2}}} = k_{ij}(\mathbf{1} + (x_j - x_i)\overset{\circ}{d}_{ij}/d_{ij}^{\mathbf{3}} - \overset{\circ}{d}_{ij}/d_{ij}) \tag{70}$$

$$\ldots \tag{71}$$

Using normal modes

- Vibrational NM: fluctuations from eigen decomposition of the Hessian of the potential energy. Uses $3n - 6$ modes.
  - Depend on the force filed, require diagonalizing the Hessian
  - Tend to be too local
- Elastic network models: GNM / ANM
  - GNM: fluctuations from eigen decomposition of the Laplacian, which acts as a quadratic form / Hessian. Uses $n - 1$ modes.
  - GNM: fluctuations from the eigen decomposition of a Taylor expansion of the potential.
  - May significantly distort internal coordinates / create steric clashes

▷ NM: resources

Table 1. The most commonly used NMA web services.

Service	url	Reference
ElNémo	http://www.sciences.univ-nantes.fr/elnemo/	[20,68]
AD-ENM	https://enm.lobos.nih.gov/index.html	[104–106]
NOMAD-Ref	http://lorentz.immstr.pasteur.fr/nomad-ref.php	[107]
oGNM	https://dyn.life.nthu.edu.tw/oGNM/oGNM.php	[108]
iGNM	http://gnm.csb.pitt.edu/index.php	[109,110]
DynOmics	http://gnm.csb.pitt.edu/index.php	[111]
ANM 2.1	http://anm.csb.pitt.edu/	[112,113]
HingeProt	http://www.prc.boun.edu.tr/appserv/prc/hingeprot/hingeprot.html	[114]
MolMovDB	http://molmovdb.org/	[115,116]
iMODS	http://imods.chaconlab.org/	[117,118]
WEBnm@	http://apps.cbu.uib.no/webnma/home	[120]

▷Ref: Bauer et al, Normal Mode Analysis as a Routine Part of a Structural Investigation, Molecules, 24, 2019

Structural alignments and analysis

Main points

Main points:

- Canonical distribution in the NVT ensemble
- Associated average internal energy, specific heat, entropy, Helmoltz free energy
- Study of the classical 1D harmonic oscillator
- Application to the free energy landscape of hard spheres

Notations and goal

- We consider a physical system with state space Ω, and an energy $E(s)$ or E_S for $s \in \Omega$.
- For the sake of notational convenience, we use notations associated with a discrete system. For example, evaluating a function $f(\cdot)$ over all states yields $\sum_{s\in\Omega} f(s)$.
- The number of particles is denoted N. For example, if the system consists of N atoms moving independently, since each atom has three cartesian coordinates $\Omega = \mathbb{R}^{3N}$.
- Main goal: understand the importance of the partition function Z

Perfect gases

- Consider a perfect gas whose number of particles and moles satisfy $N = n\mathcal{N}$, with $\mathcal{N}$ the Avogadro number.
- The law of perfect gases satisfies

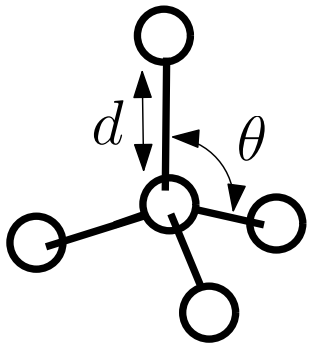
$$pV = nRT = Nk_B T, \text{ with } k_B = \frac{R}{\mathcal{N}}. \tag{72}$$

- We let $\beta = 1/(k_B T)$, with k_B the Boltzmann constant, and note that

$$\frac{\partial \beta}{\partial T} = -\frac{1}{k_B T^2}. \tag{73}$$

Molecular potential energy

- Molecule with N atoms: 3N Cartesian coordinates, and $3N - 6$ degrees of freedom



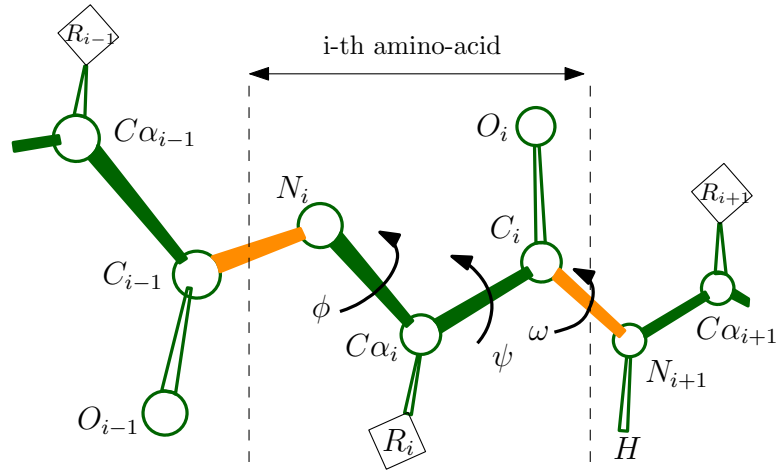


Figure: Internal coordinates

- The *potential energy* is a non linear function whose general equation satisfies

$$V_{\text{total}} = V_{\text{bond}} + V_{\text{angle}} + (V_{\text{proper}} + V_{\text{improper}}) + (V_{\text{vdw}} + V_{\text{electro}}) \qquad (74)$$

- Instantiating such an equation requires:
  - types for atoms and bonds
  - covalent: bond lengths, angles
  - non covalent: pairwise distances
  - solvent model
- Force field: hundreds/thousands of parameters
- : `AMBER`, `CHARMM`, `MARTINI` etc

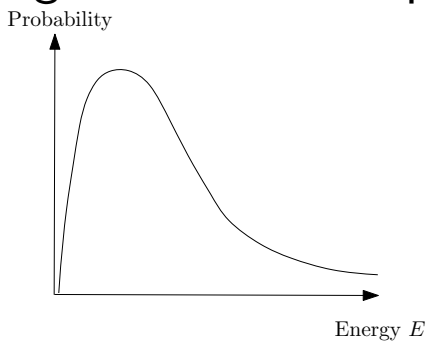
Boltzmann's distribution

- Internal energy in the sequel: $E = V$ (or $E = U + K$)
- Our system: exchanges energy with a *heat bath*; constant temperature T
- We take for granted the Boltzmann distribution, with assigns to each state i a probability

$$\mathbb{P}[i] = \frac{e^{-\beta E_i}}{Z}, \text{ with } Z = \sum_i e^{-\beta E_i}. \tag{75}$$

The normalization constant Z is called the *partition function*.

- Microscopic versus macroscopic
  - Eq. 75 refers to a microscopic configuration.
  - Averaging using Eq. 75 yields macroscopic quantities, also called observables.
- Low energy states are not in general the most populated



Average internal energy

Nb:

$$\frac{\partial Z}{\partial \beta} = -\sum_i E_i e^{-\beta E_i} \tag{76}$$

Averaging the energy E_i over all states yields

$$< E >= \sum_i E_i \mathbb{P}[i] = \frac{\sum_i E_i e^{-\beta E_i}}{Z} = -\frac{1}{Z}\frac{\partial Z}{\partial \beta} = -\frac{\partial \log Z}{\partial \beta}. \tag{77}$$

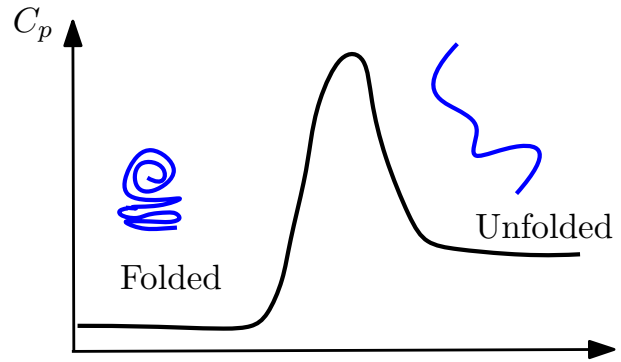
Specific heat per particle

- Heat capacity: energy needed to increase the temperature by one unit.
- For our system with N particles, using Eq. 77 for the last equality:

$$NC_v = \frac{\partial < E >}{\partial T} = \frac{\partial < E >}{\partial \beta}\frac{\partial \beta}{\partial T} = -\frac{1}{k_B T^2}\frac{\partial < E >}{\partial \beta} \tag{78}$$

$$= \frac{1}{k_B T^2}\frac{\partial^2 \log Z}{\partial \beta^2}. \tag{79}$$

- Intuition
  - Condensed phase: C_v related to the potential energy
  - Opposite: C_v related to the kinetic energy
  - In-between: C_v is higher



C_v as a function of T: intuition

Entropy

- Using the counting based formulae

$$S = -k_B \sum_i \mathbb{P}[i] \log \mathbb{P}[i] = -k_B \sum_i \frac{e^{-\beta E_i}}{Z} \log \frac{e^{-\beta E_i}}{Z} \tag{80}$$

$$= -k_B \sum_i \frac{e^{-\beta E_i}}{Z} (-\beta E_i - \log Z) \tag{81}$$

$$= k_B \beta < E > + k_B \log Z = \frac{< E >}{T} + k_B \log Z. \tag{82}$$

Or equivalently

$$S = \frac{< E >}{T} + k_B \log Z. \tag{83}$$

Helmoltz free energy – denoted A or F

- From Eq. 83, one defines the so-called Helmoltz free energy:

$$A(T, N, V) \stackrel{Def}{=} -k_B T \log Z =< E > -TS. \tag{84}$$

- Constant volume!

- From the previous equation, one also get the following interesting parallel:
  - Boltzman factor for a microscopic state

$$\exp(\frac{-E_i}{k_b T}) \tag{85}$$

  - Macroscopic analogous based on the Helmoltz free energy:

$$\exp(-\frac{A}{k_b T}) = Z(= \sum_i e^{-\beta E_i}) \tag{86}$$

Classical 1D harmonic oscillator

- 1D Harmonic oscillator: spring

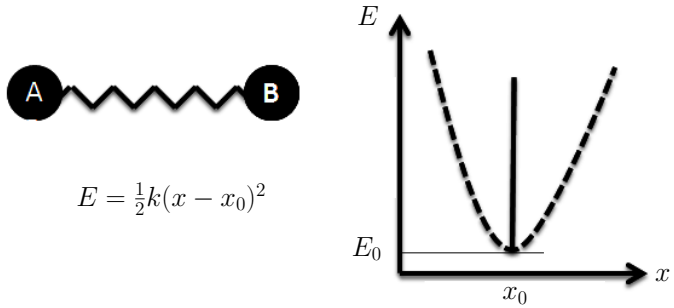


Figure: **1D harmonic oscillator**

- Associated potential energy

$$E(x) = E_0 + \frac{k}{2}(x - x_0)^2 \text{ or equivalently } E(u) = E_0 + \frac{1}{2}ku^2. \qquad (87)$$

- NB: k is the stiffness; geometrically, this the curvature of paraboloid representing the energy.

1D harmonic oscillator: partition function and free energy

- Partition function: using Gaussian integrals, we get

$$Z = \int \exp(-\beta E(u))du = \exp(-\beta E_0) \int \exp(-\beta \frac{ku^2}{2})du \tag{88}$$

$$= \exp(-\beta E_0)\sqrt{\frac{2\pi}{\beta k}}. \tag{89}$$

- Helmoltz free energy. Using the previously found partition function, we get:

$$F = -k_B T \log Z = E_0 - k_B T \log(\frac{2\pi}{\beta k})^{1/2}. \tag{90}$$

1D harmonic oscillator: average internal energy and entropy

- Average internal energy. One gets

$$< E >= \frac{1}{Z}\int (E_0 + \frac{ku^2}{2}) \exp(-\beta E_0 \frac{\beta k u^2}{2}) du = E_0 + \frac{k_B T}{2}. \quad (91)$$

- Entropy. One gets

$$S = \frac{< E >}{T} + k_B \log Z = \frac{E_0}{T} + \frac{k_B}{2} + k_B(-\beta E_0 + \log(\frac{2\pi}{\beta k})^{1/2} \quad (92)$$

$$= \frac{k_B}{2} + k_B \log \frac{2\pi}{\beta k}^{1/2}. \quad (93)$$

It should be noticed that when the curvature k decreases, the entropy increases: there is indeed more uncertainty on the position of the oscillator.

Rigid-rotor harmonic oscillator

- Assumptions: for an isolated rigid molecule, the energy can be decoupled as
  - translation
  - rotation
  - vibrational
  - electrostatic

$$\begin{cases} E = E_{trans} + E_{rot} + E_{vib} + E_{elec} \\ Z = Z_{trans} * Z_{rot} * Z_{vib} * Z_{elec} \\ S = S_{trans} + S_{rot} + S_{vib} + S_{elec} \end{cases} \tag{94}$$

- Caveat: does not apply to *floppy* (soft bending) molecules

Rotational degrees of freedom

- Mass and moment of inertia:
  - mass: determines the force needed to obtain a given acceleration
  - moment of inertia: determines the torque needed to produce an angular acceleration
    - torque/moment/force moment: vector representing the ability of a force to produce a rotational movement

- Tensor of inertia of a polyatomic molecule:

$$\mathsf{I} = \begin{pmatrix} \sum_i m_i(y_i^2 + z_i^2) & -\sum_i m_i x_i y_i & -\sum_i x_i z_i \\ -\sum_i m_i x_i y_i & \sum_i m_i(x_i^2 + z_i^2) & -\sum_i m_i y_i z_i \\ -\sum_i m_i x_i z_i & -\sum_i m_i y_i z_i & \sum_i m_i(x_i^2 + y_i^2) \end{pmatrix} \tag{95}$$

Eigenvalues/vectors: moments of inertia / principal axis of inertia

- Approximation of the partition function – with σ the rotational symmetry number:

$$Z_{rot} = \frac{\sqrt{\pi}}{\sigma}\left(\frac{8\pi^2 kT}{h^2}\right)^{3/2}\sqrt{I_1 I_2 I_3}. \tag{96}$$

- Ref: Jensen, Introduction to computational chemistry, Wiley
- Ref: Arnold, Math. methods of classical mechanics, Springer

Polyatomic molecule and vibrations: quantum partition function

- Generalization of the harmonic model for a diatomic molecule:

$$E_{vib} = \sum_{i=1,\ldots,3N-6} (n_i + 1/2) h\nu_i \tag{97}$$

$$Z_{vib} = \prod_{i=1,\ldots,3N-6} \frac{e^{-h\nu_i/2kT}}{1 - e^{-h\nu_i/2kT}}. \tag{98}$$

- Ref: Jensen, Introduction to computational chemistry, Wiley

Polyatomic molecules and vibrations: classical multi-dimensional harmonic oscillator

▷ Physical model. Consider a set of N atoms, whence $3N$ cartesian coordinates, whose potential energy is quadratic.

▷ The Hessian of the potential energy is the $3N \times 3N$ matrix defined by

$$\mathbf{H}_{ij} = \frac{\partial^2 E}{\partial x_i \partial x_j} \tag{99}$$

▷ Three types of models:

- Force field based: $3N - 6$ eigenvalues
- Gaussian Network Model: $N - 1$ eigenvalues
- Anisotropic Network Model: $3N - 6$ eigenvalues

▷ Vibrational partition function: from frequencies of normal modes

$$Z_{vib} \propto \prod_i \sqrt{\frac{2\pi}{\beta k_i}}. \tag{100}$$

Experiment and goal

- ▷ Setup: one places small polystyrene (PS) balls into a well.
- ▷ Question: most likely geometries for clusters of N balls?
- ▷ Experiment: how. The PS balls are placed in cylindrical micro-wells filled with an appropriate chemical (irrelevant here). The walls of the cylinders are treated to avoid interactions with the balls.
- ▷ Counting: using a microscope, one takes images, and consider the geometry of the clusters obtained.

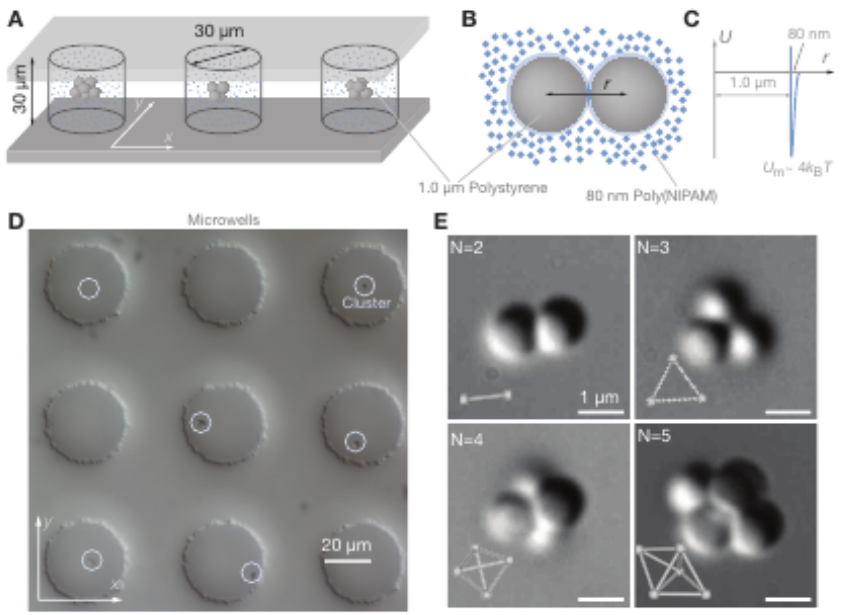


Figure: **Conformations of hard spheres: experiment.** From [?].

Physical model

▷ Model. The coating of the balls is such that interactions are very local, so that the potential energy of a collection of N balls is given by the number C of pairs in contact.

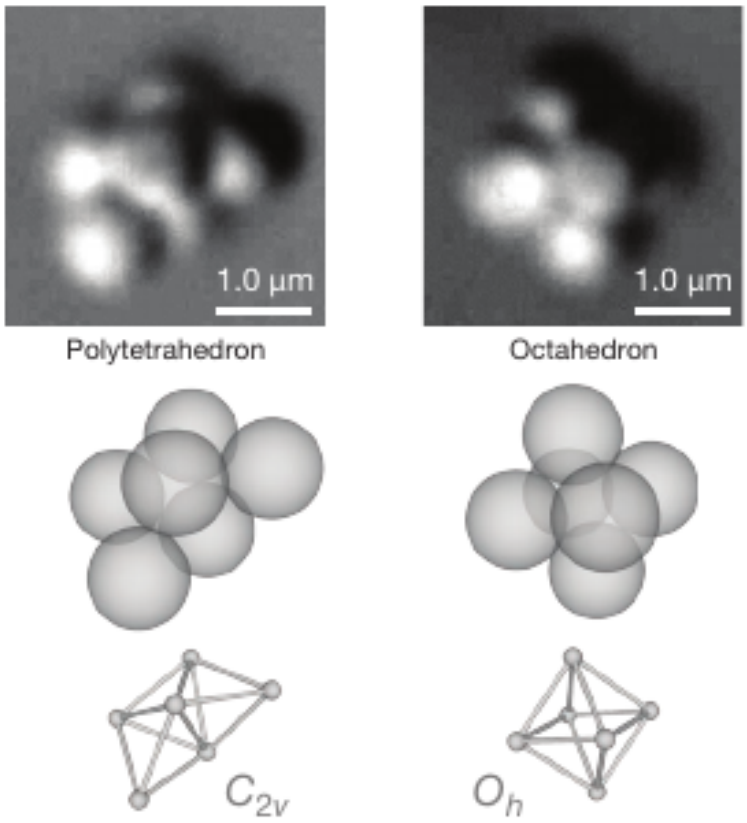
▷ Potential energy approximation

$$V = CV_m, \tag{101}$$

with V_m the depth of the interaction potential.

Experiments: observations for $N = 6$

- The observations are as follows:
  - Two clusters are observed: polytetra(hedron) and octahedron.
  - Their respective percentages are $\sim 96\%$ and $\sim 4\%$, whence the ratio $96/4 = 24$.
  - This is so, even though the two clusters exhibit the same number of contacts ($m = 12$), so that the potential energy is the same by Eq. 101



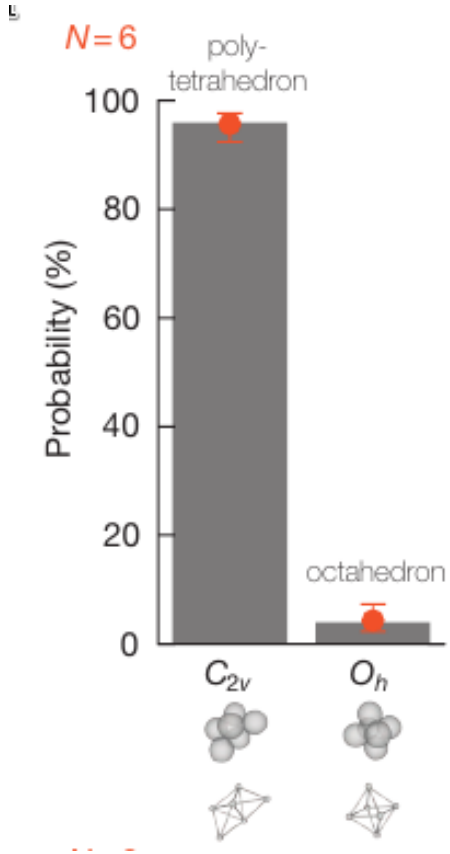


Figure: **Configurations for N=6 balls.** From [?]. **(Left)** Clusters **(B)** Population statistics. Note that the polytetrahedron is $\sim$ 24 times more

Statistical model: macroscopic states

▷ States. Let a cluster be called a state: a number of microscopic configurations populate this state. Using Eq. 84, using the partition function Z_s for this state to define the macroscopic analogous of Boltzmann's weight (Eq. 86), one has:

$$A_s = -k_B T \log Z_s \text{ or equivalently } \exp(-\frac{A_s}{k_b T}) = Z_s. \tag{102}$$

Therefore, the relative abundance of state s is given by

$$p_s = \frac{Z_s}{\sum Z_s}. \tag{103}$$

Statistical model: microscopic states

- The partition function Z_s of a state has three components:

$$Z_s = Z_{s,t} Z_{s,rot.} Z_{s,vib.} \tag{104}$$

with

- $Z_{s,t}$: translational. This term depends on the free volume within a cylindrical well. Assuming that the volume of all clusters (at fixed N) is the same, $Z_{s,t}$ is the same for all clusters.
- $Z_{s,rot.}$: the rotational partition function depends on the number of orientations and particle permutations in a cluster. With I the inertia tensor, and σ the rotational symmetry number (based on point group), it can be shown that

$$Z_{s,rot.} \propto \frac{\sqrt{\text{determinant}(I)}}{\sigma}. \tag{105}$$

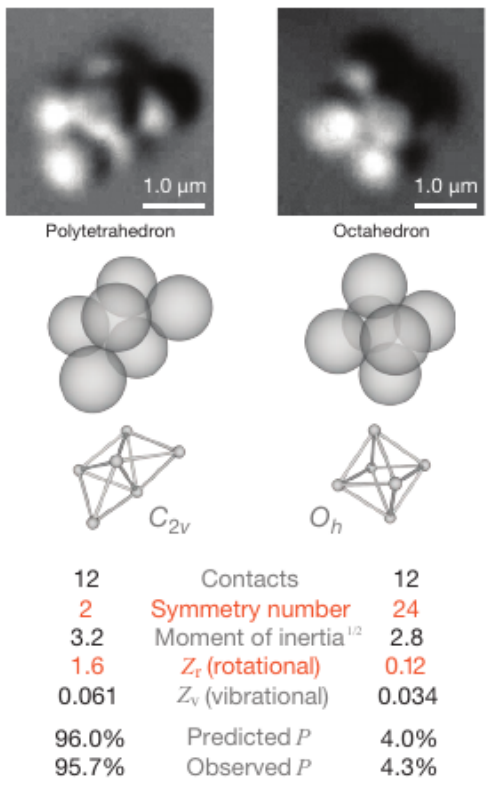
- $Z_{s,vib.}$: the vibrational partition function given by Eq. (100):

$$Z_{vib} \propto \prod_{i=1}^{3N-6} \sqrt{\frac{2\pi}{\beta k_i}}.$$

Calculations for the two configurations

- The calculation yields, see Fig. 48:
  - polytetra(hedron) : $Z_{s,rot.} Z_{s,vib.} \sim 1.6 * 0.061 = 0.0976$
  - octahedron: $Z_{s,rot.} Z_{s,vib.} \sim 0.12 * 0.034 = 0.00408$

Taking the ratio, we get $0.0976/0.00408 \sim 23.9$, which is highly consistent with the observed frequencies since $96/4 = 24$.



Conformations of hard spheres: why is the polytetrahedron more abundant?

- Conclusion: **the cluster geometry favored is that which is less symmetric, and the main source for this is not the interaction energy, but the rotational entropy.**

A didactical view

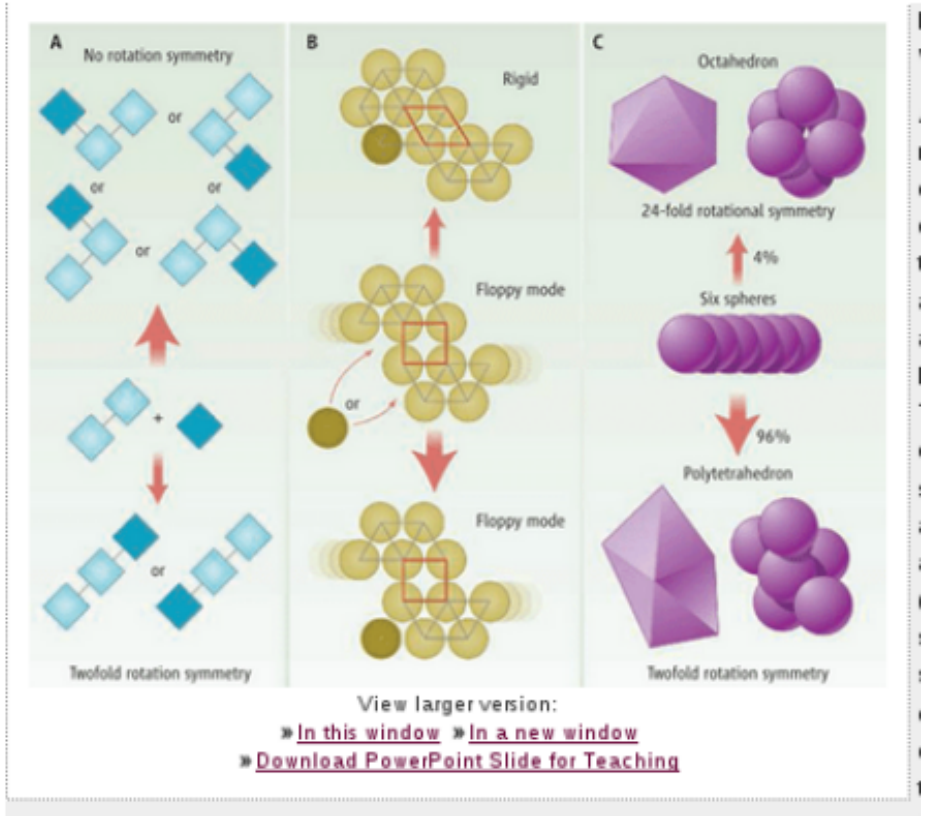
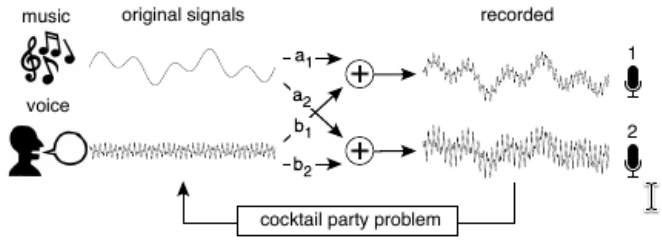


Figure: **Entropy driven association: didactical explanation.** From [?]

Structural alignments and analysis

Independent component analysis: the cocktail party problem

▷ The cocktail party problem: disentangling voice and music sources



▷ Model: recorded data x and the signals S yield

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} a_1 s_1 + a_2 s_2 \\ b_1 s_1 + b_2 s_2 \end{pmatrix} \tag{106}$$

In more compact form : $X = AS$

•*Assumption:* The sources are statistically independent in the following sense:

$$f(y) = \prod_i f_i(s_i). \tag{107}$$

Consequently, the covariance matrix reads as $\mathbb{E}\left[SS^{\mathsf{T}}\right] = \mathrm{I}_d$

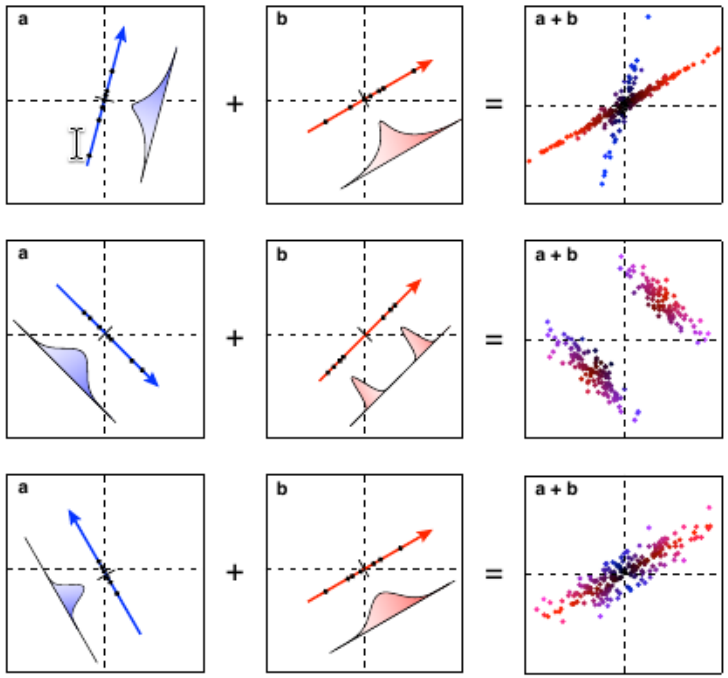
Problem 2. Under the previous assumption, and assuming that the linear operator A is invertible, find the inverse A^{-1} providing the signal from the mixed measurements:

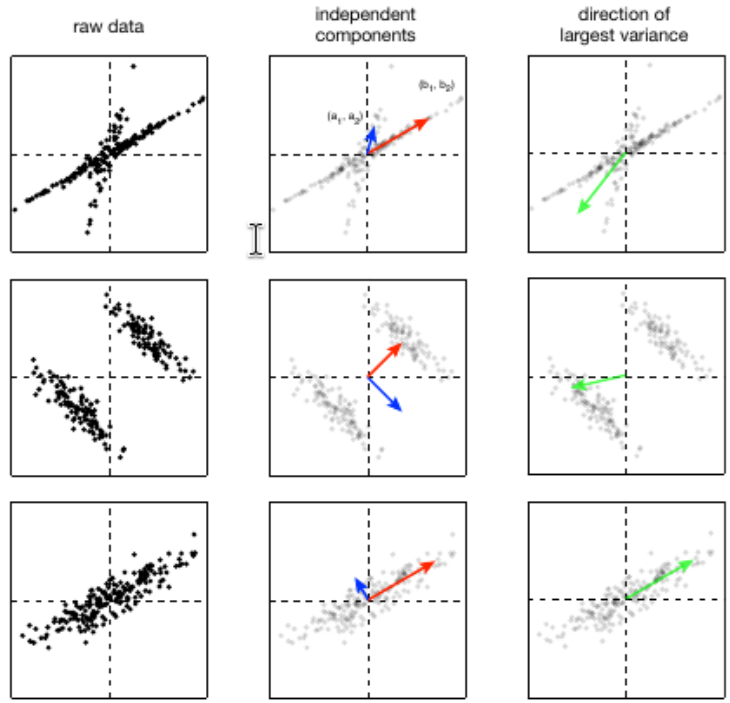
$$S = A^{-1}X. \tag{108}$$

NB: in the sequel: we assume centered data.

▷Ref: J. Shlens, A tutorial on ICA, preprint, 2014

Datasets with linear mixing: examples





- Individual models
- Merged dataset

- Blue, red: tICA vectors
- Green: PCA

▷Ref: J. Shlens, A tutorial on ICA, preprint, 2014

ICA: overview of the two steps

1. Step 1: whitening - via PCA
2. Step 2: find an orthogonal matrix V associated with a mutual information minimization problem

Step 1: whitening (I)

▷ Problem formulation using the SVD of A: with $A = U\Sigma V^{\mathbf{T}}$

$$S = A^{-\mathbf{1}}X = V\Sigma^{-\mathbf{1}}U^{\mathbf{T}}X. \tag{109}$$

▷ Correlation matrix C: direct calculation

$$C = \mathbb{E}\left[XX^{\mathbf{T}}\right]. \tag{110}$$

By the spectral decomposition theorem, we can decompose C as follows:

$$C = EDE^{\mathbf{T}}. \tag{111}$$

▷ Correlation matrix C: using the data model ie $X = AS$, and the SVD of A

$$\begin{aligned} C &= \mathbb{E}\left[(AS)AS^{\mathbf{T}}\right] && (112)\\ &= \mathbb{E}\left[U\Sigma V^{\mathbf{T}}SS^{\mathbf{T}}V\Sigma U^{\mathbf{T}}\right] && (113)\\ &= U\Sigma V^{\mathbf{T}}\mathbb{E}\left[SS^{\mathbf{T}}\right]V\Sigma U^{\mathbf{T}} && (114)\\ &= U\Sigma^{\mathbf{2}}U^{\mathbf{T}}. && (115) \end{aligned}$$

NB: we have used $\mathbb{E}\left[SS^{\mathbf{T}}\right] = \mathrm{I}_d$.

Comparing Eqs. 111 and 115 makes it possible to choose $U = E$ and $\Sigma = D^{\mathbf{1}/\mathbf{2}}$. With these, Eq. 109 rewrites as

$$S = A^{-\mathbf{1}}X = VD^{-\mathbf{1}/\mathbf{2}}E^{\mathbf{T}}X \tag{116}$$

The operator $D^{-\mathbf{1}/\mathbf{2}}E^{\mathbf{T}}$ is a so-called *whitening* or *sphereing* operation [?] – Fig. 8. Our problem rewrites as

$$S = VX_w, \text{ with } X_w = D^{-\mathbf{1}/\mathbf{2}}E^{\mathbf{T}}X. \tag{117}$$

We are left with the problem of identifying the optimal rotation V.

Step 1: whitening (II)

- Inversion problem re-written:

$$S = VX_w, \text{ with } X_w = D^{-1/2}E^{\mathsf{T}}X. \tag{118}$$

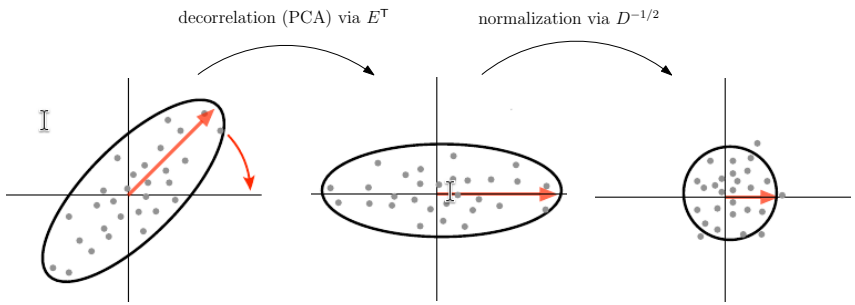


Figure: **ICA: whitening.** The input data are transformed by a rotation then a scaling, yielding the whitened data $X_w = D^{-1/2}E^{\mathsf{T}}X$.

$\Rightarrow$ We are left with the problem of identifying the optimal rotation V.

Step 2: finding the orthogonal matrix V

Problem 3. Find an orthogonal matrix V such that $S = VX_w$ and $\mathbb{E}\left[SS^\mathsf{T}\right] = \mathrm{I}_d$.

▷ Model: consider a random vector S with joint density $f(y) = \prod_i f_i(y_i)$ – cf our assumption

▷ Mutual information: Kullback-Leibler divergence between the joint density and the product of marginals

$$I(S) = \int f(y) \log \frac{f(y)}{\prod_i f_i(y_i)} dy. \tag{119}$$

Following the assumption $f(y) = \prod_i f_i(y_i)$, we aim at minimizing:

$$I(S) = \sum_i H(S_i) - H(S). \tag{120}$$

Assuming estimators from the data:

$$I(\hat{S}) = \sum_i H((VX_w)_i) - H(VX_w) \tag{121}$$

$$= \sum_i H((VX_w)_i) - (H(X_w) + \log_2(\det(V))) = \sum_i H((VX_w)_i) - H(X_w) \tag{122}$$

Since $H(X_w)$ does not depend on V, we end up with the minimization problem:

$$\arg\min_V \sum_i H((VX_w)_i) \tag{123}$$

Difficulties to solve this problem:

- Eq. 123 is non convex in general – local minima.
- Estimating the individual entropies is non trivial per se.

ICA and its limitations

- ICA is under-constrained: several degrees of freedom yield alternative solutions
  - Permutation of labels
  - Flip of independent components
  - Rescaling the ICA vectors. The rescaling can compensated in the assumption $\mathbb{E}\left[SS^{\mathsf{T}}\right] = \mathsf{I}_d$. See the derivation of Eq. 116.

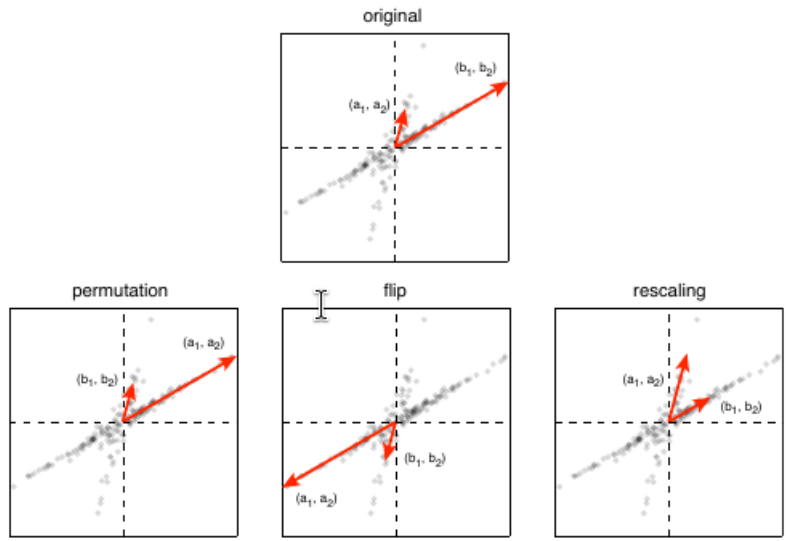


Figure: **ICA: ambiguities.** Left: permuation of the lables; Middle: flip of the ICA; Right: rescaling of the ICA. From [**?**].

- Ref: J. Shlens, A tutorial on ICA, preprint, 2014

Structural alignments and analysis

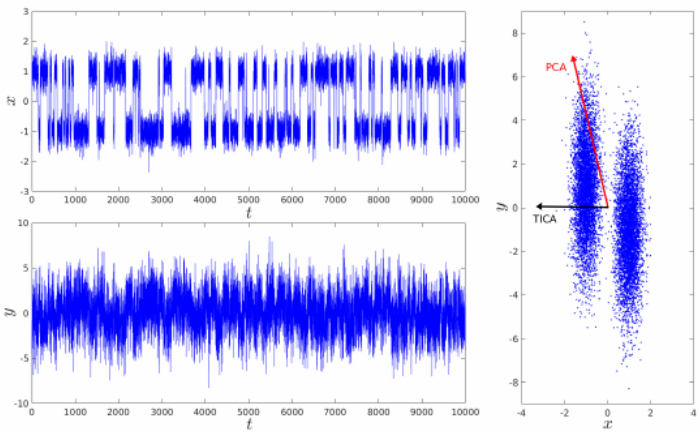
Time lagged ICA (tICA): goals

▷ Assumption: consider a dynamical system **at equilibrium**, given by a vector valued time series $X(t)$

▷ Two goals:

1. Project the dynamics onto a lower dimensional space whose coordinates are uncorrelated – that is cross-covariances are null auto-covariances are maximal.
2. Identify fast and slow motions.

NB: spirit analogous to ICA, but a single source, and exploitation of temporal coherence



System with

- two meta-stable states, with rare transitions.
- red vector: first principal component by PCA.
- black vector: first tICA component.

▷ NB: with a proper **lag time** τ, tICA identifies the *slow* directions, x here.

▷Ref: Noe et al, J. of Nonlinear science, 2018

tICA: the generalized eigenproblem

Definition 4. The *time lagged covariance matrix* at lag time τ is defined by:

$$C^r(\tau) = \left(c_{ij}^r(\tau)\right), \text{with } c_{ij}^r(\tau) = \mathbb{E}_t\left[X_{t,i} X_{t+\tau,j}\right]. \tag{124}$$

▷ Central def.: motivated by the theory of transfer operators

Definition 5. The *time-lagged independent components* are defined as the solution of the following eigenvalue problem:

$$C_\tau f_l = \lambda_l C_0 f_l, \tag{125}$$

or equivalently, using the pseudo-inverse ${C_0}^\dagger$ of C_0, of

$$M_{\text{TICA}} f_l = \lambda_l f_l, \text{ with } M_{\text{TICA}} = {C_0}^\dagger C_\tau. \tag{126}$$

The time scale of the mode l is the quantity

$$t_l = -\frac{\tau}{\log \lambda_l}. \tag{127}$$

tICA: time scales

Proposition. 5. The λ_i values of Eq. 126 are ≤ 1.

The proof is based on the Cauchy-Schwarz inequality for expected covariances in time series:

Proof. Multiplying the generalized eigen pb. equation $C_\tau f_i = \lambda_i C_0 f_i$ by ${f_i}^\mathsf{T}$ yields

$$\lambda_i = \frac{{f_i}^\mathsf{T} C_\tau f_i}{{f_i}^\mathsf{T} C_0 f_i}. \tag{128}$$

Covariance matrices satisfy the Cauchy-Schwarz inequality for expectations $C_\tau \preceq C_0$, that is, for any vector v, one has $v^\mathsf{T} C_\tau v \preceq v^\mathsf{T} C_0 v$. One conclude from the preceding equation. □

▷ Important comments:

- Correlations between states decay as a function of time. Thus, $\lambda_i (\leq 1)$ captures how much correlation remains at the time lag τ.
- Eigenvalues close to 1 indicate modes with long time persistence. These are the modes of interest.
- tICA modes are assessed by sorting the time scales of Eq. 127 by decreasing value.

tICA vectors: C_0 orthogonality

Proposition. 6. Consider the eigenvectors $\{f_i\}$ of the generalized eigenproblem of Eq. 126. These vectors are C_0 orthogonal, that is

$$f_i^{\mathsf{T}} C_0 f_j = \delta_{ij}. \tag{129}$$

Proof. Consider two eigenvector f_i and f_j, that is

$$\begin{cases} C_\tau f_i = \lambda_i C_0 f_i \\ C_\tau f_j = \lambda_i C_0 f_j \end{cases} \tag{130}$$

Taking the transpose of the first eq. and multiplying by f_j yields

$$f_i^{\mathsf{T}} C_\tau f_j = \lambda_i f_i^{\mathsf{T}} C_0 f_j. \tag{131}$$

Likewise, for the second equation, we get

$$f_j^{\mathsf{T}} C_\tau f_i = \lambda_j f_j^{\mathsf{T}} C_0 f_i. \tag{132}$$

But since C_τ is symmetric, we have

$$\lambda_i f_i^{\mathsf{T}} C_0 f_j = \lambda_j f_j^{\mathsf{T}} C_0 f_i \Rightarrow (\lambda_i - \lambda_j) f_i^{\mathsf{T}} C_0 f_j = 0. \tag{133}$$

Then $\lambda_i \neq \lambda_j$ implies $f_i^{\mathsf{T}} C_0 f_j = 0$, that is, the eigenvectors f_i and f_j are C_0-orthogonal. Finally, assuming that C_0 is positive definite, we can normalize the eigenvectors so that

$$f_i^{\mathsf{T}} C_0 f_i = \delta_{ij}. \tag{134}$$

□

Using tICA

▷ Using tICA:

procedure Enjoy-tICA(X, Y)
  Compute the covariance matrices C_0 and C_τ
  Solve the generalized eigenvalue problem *e.g.* using `AMUSE`
  Threshold the eigenvalues using the time scales – Eq. 127 ▷ Assume k get selected
  Assemble the $d \times k$ transformation matrix $W = [f_1, \ldots, f_k]$
  Project onto tICA coordinates with to obtain $Z(t) = W^\mathsf{T} X(t)$ ▷ Shape is $d \times 1 = (k \times d) \times (d \times 1)$
end procedure

▷ Covariance matrices: with datasets $X = [x_0 \ldots x_{m-1}]$, $Y = [x_\tau \ldots x_{\tau+m-1}]$, compute

$$C_0 = \frac{1}{m-1} X X^\mathsf{T}, C_\tau = \frac{1}{m-1} X Y^\mathsf{T}. \tag{135}$$

▷ Projection onto the tICA vectors selected:

$$Z(t) = W^\mathsf{T} X(t). \tag{136}$$

▷ NB: if we denote $g_i = C_0 f_i$, Because of the C_0 orthogonality between the f_is – Eq. 129, the projection onto the tICA vectors amounts to rewriting the time series in the $\{g_i\}$ basis, that is

$$Z(t) = \sum_i (f_i{}^\mathsf{T} X(t)) g_i. \tag{137}$$

Computing tICA with AMUSE

- Computing tICA when $n \gg d$:

Proposition. 7. Algorithm **??** computes the solutions of Eq. 126.

procedure AMUSE(X, Y)
  Compute a reduced SVD of X, *i.e.* $X = U\Sigma V^{\mathbf{T}}$
  Whiten the data as $\tilde{X} = \Sigma^{-\mathbf{1}} U^{\mathbf{T}} X$ and $\tilde{Y} = \Sigma^{-\mathbf{1}} U^{\mathbf{T}} Y$
  Compute $\bar{M}_{\mathbf{TICA}} = \tilde{X}\tilde{Y}^{\mathbf{T}}$
  Solve the eigenvalue problem $\bar{M}_{\mathbf{TICA}} w_l = \lambda_l w_l$
  Obtain the tICA coordinates as $f_l = U\Sigma^{-\mathbf{1}} w_l$
end procedure

Algorithm AMUSE: proof of correctness

Proof. We also introduce the following linear transformation:

$$f_l = U\Sigma^{-1} w_l. \tag{138}$$

Note the following expression of the pseudo-inverse of $(XX^{\mathsf{T}})^{\dagger}$ obtained from the SVD of X:

$$(XX^{\mathsf{T}})^{\dagger} = U\Sigma^{-2}U^{\mathsf{T}}. \tag{139}$$

Note also

$$\tilde{X}\tilde{Y}^{\mathsf{T}} = \Sigma^{-1}U^{\mathsf{T}}XYU\Sigma^{-1}. \tag{140}$$

Starting from Eq. 126, we make $\bar{M}_{\mathrm{TICA}}$ appear:

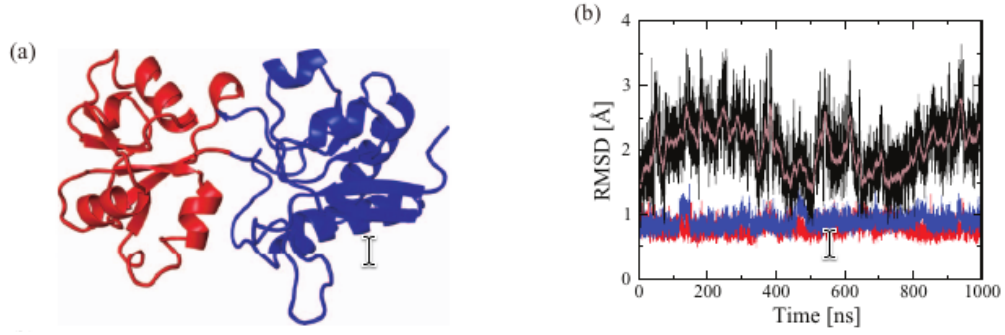
$$\begin{aligned}
M_{\mathrm{TICA}} f_l &= {C_0}^{\dagger} C_\tau f_l = (XX^{\mathsf{T}})^{\dagger} XY^{\mathsf{T}} f_l && (141)\\
&= U\Sigma^{-1}\Sigma^{-1}U^{\mathsf{T}}XY^{\mathsf{T}}U\Sigma^{-1}w_l && (142)\\
&= U\Sigma^{-1}\tilde{X}\tilde{Y}^{\mathsf{T}}w_l && (143)\\
&= \lambda_l U\Sigma^{-1}w_l - w_l \text{ eigenvector of } \bar{M}_{\mathrm{TICA}} && (144)\\
&= \lambda_l f_l. && (145)
\end{aligned}$$

□

Lysine Arginine Ornithine (LAO) protein

▷ LAO: member of the PBP family; transports low molecular weight ligands from the outer to the inner membrane in the ABC transport mechanism of Gram-negative bacteria

▷ LAO protein and its two domains:



- The LAO protein, 238 a.a., is composed of 2 domains, large (blue) and small (red). Molecular dynamics simulation of $1\mu s$ in the NVE ensemble.
- RMSD wrt the crystal structure, overall and per domain.
- MD simulation: $1\mu s$

▷ Setup: tICA analysis with time lag $\tau = 1ns$

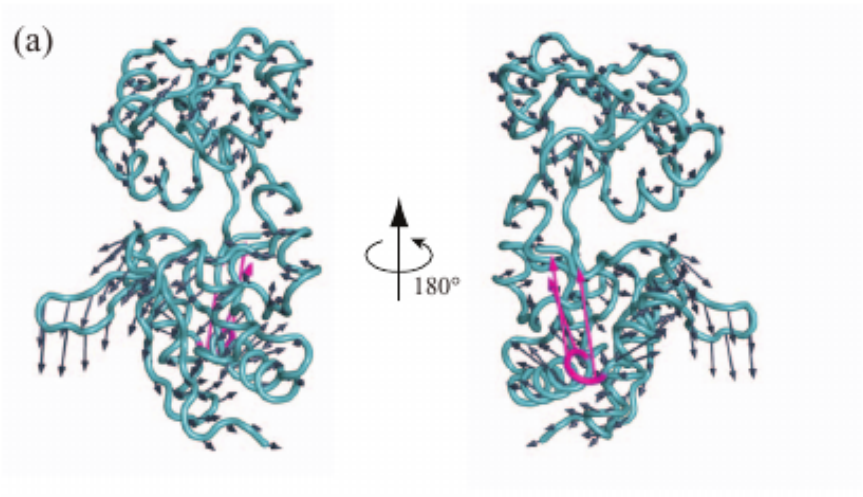
▷ Five modes selected: all with time scales $\gg \tau$

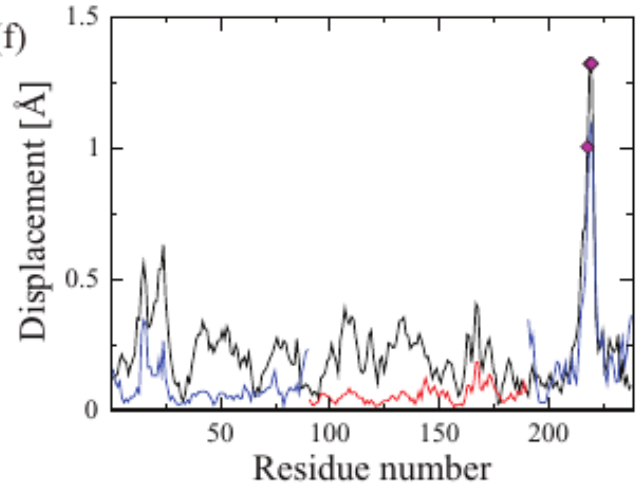
TABLE I. Summary of the properties of IC modes.

Properties	IC1	IC2	IC3	IC4	IC5
Time scale (ns)	28.0	13.4	10.7	6.6	4.5
Contribution ratio (%)	5.6	11.9	2.6	1.6	0.6
Remarkably mobile C_α atoms	R218, Q219, D220	—	A15, P16	G24	K186
Ratio of interdomain motion (%)	60.6	92.1	65.7	76.7	49.0

▷Ref: Naritomi et al, J. Chem. Physics, 2013

LAO: slowest mode IC1





- (a) Mode vectors of IC1 restricted to each C_α carbon, with remarkable vectors in purple.
- (b) Displacements of C_α accounted for by IC1. Remarkable C_α marked by magenta diamonds.

LAO: IC1 and local motion of the backbone

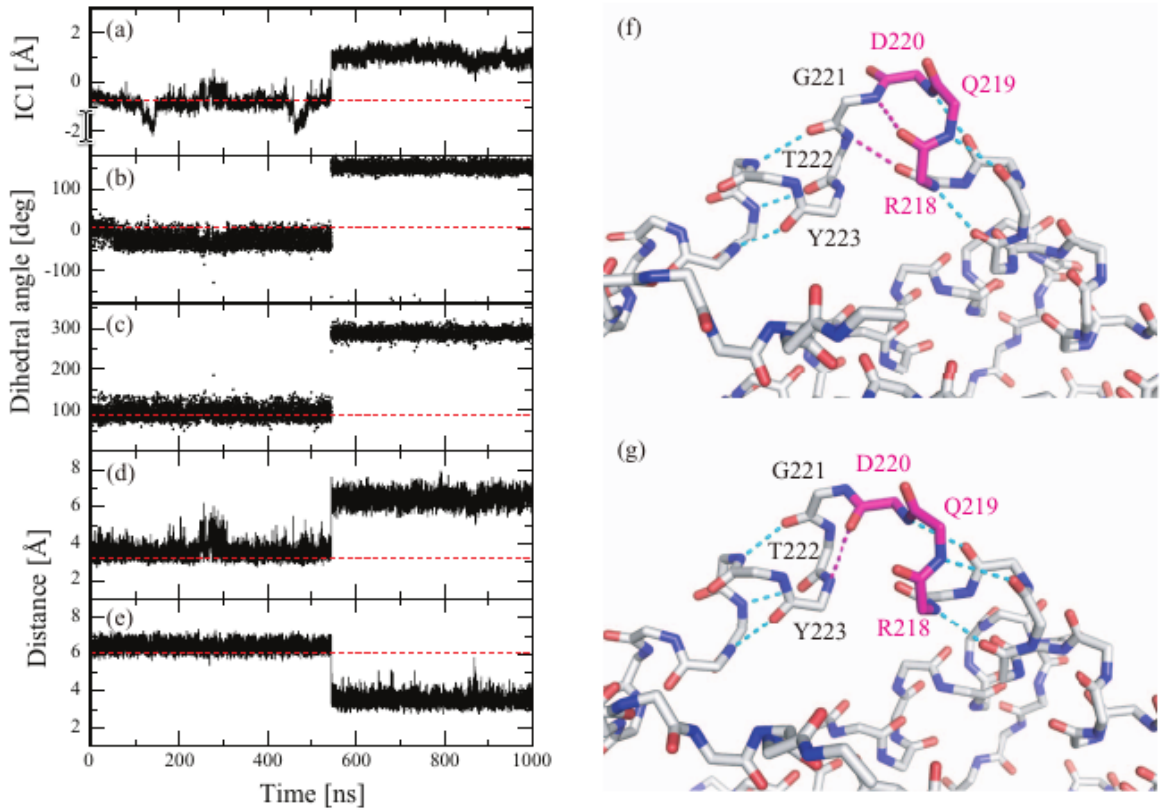
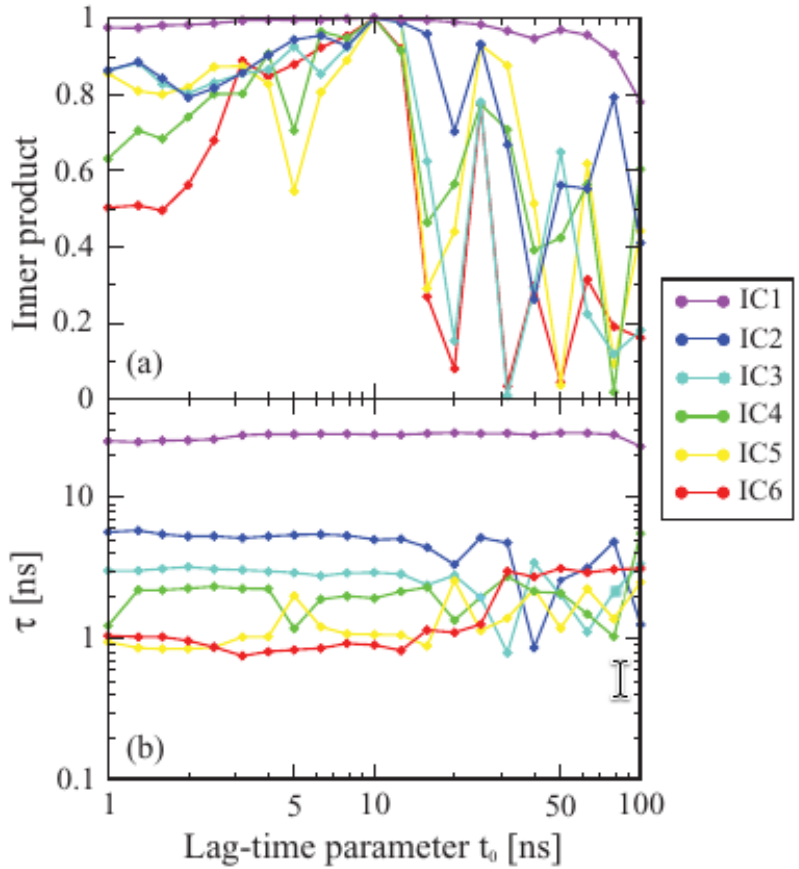


FIG. 4. Local motion of the backbone of LAO detected by IC1: Trajectories of (a) IC1, (b) D220ψ, (c) G221ϕ, (d) R218O–G221N, and (e) D220O–Y223N.

- (i) the trajectory along IC1 (panel (a)),
- (ii) two dihedral angles ψ, ϕ working in tandem (compensating one another) in a so-called crankshaft move (panels (b,c)),
- (iii) distance between 2 atoms: aspartic acid D220-O – tyrosine Y223-N.

LAO: stability of tICA vectors as a function of τ



NB: from a twin paper in which τ is renamed t_0