Why biophysics and biochemistry

Systems under study

Physics

Math

Biophysics

Biology

Biochemistry

Physical Biochemistry

Chemistry

Physical Chemistry

Principles
models
methods

Baumketner, BioSim, Lviv 2019
A bit of biology

Ознаки “життя”:

отримання і перетворення енергії з середовища - метаболізм

самоорганізація через використання енергії – синтез, утворення макромолекулярних комплексів

здатність запам’ятовувати свою будову – генетичний код

здатність давати потомство - розмноження

Baumketner, BioSim, Lviv 2019
Diversity of life forms

- Classification based on similarities and differences in ribosomal RNA seq.:
  
  **Prokaryotes** = unicellular organisms that lack a membrane-bound nucleus, mitochondria, or any other membrane-bound organelle

  **Eukaryotes** = cells with a nucleus that stores genetic information.

  **Bacteria**
  - definition: a cell that contains its DNA genome within a membrane-bound nucleus
  - Some aspects of archaeal biochemistry are unique, such as their reliance on ether lipids in their cell membranes

  **Archaea**
  - <3μm
  - diverse habitats

  **Eukarya**
  - animals, plants, fungi (e.g., mushrooms, mold)

Baumketner, BioSim, Lviv 2019

*Haloquadratum walsbyi*

Geothermal vent on the Atlantic ocean floor

Permafrost in Antarctica
Structural hierarchy in eukaryotes

Клітина – цеглина для всього живого

組織 → 器官 → 軟組織 → 基本組織 → 細胞

Illustration by Cell Imaging Core of the Center for Reproductive Sciences.

Baumketner, BioSim, Lviv 2019
Great diversity of cells

1 μm

shape

size

structure

Baumketner, BioSim, Lviv 2019
Cell structure

Кишкова паличка

E. Coli (model prokaryotic cell)
Cell structure

Фібропласт – клітина сполучної тканини (колаген для загоювання ран)

Fibroblast (model higher eukaryotic cell)
Biological length scales

http://learn.genetics.utah.edu/content/begin/cells/scale/

Baumketner, BioSim, Lviv 2019
Hierarchical organization

Animals and Plants
→ Specialized Organs
→ Specialized Tissues
→ Cells
→ Organelles
→ Macromolecular Assemblies
→ Macromolecules
→ Simplest Molecular Building Blocks

Зростання рівня організації
Cell atoms

Organic Atoms (H, C, N, O)  99% of cells
Ions (Na, K, Mg, Ca, P, S, Cl)  0.9%

Baumketner, BioSim, Lviv 2019
Chemical bonds in biology

Two atoms are close enough for their atomic orbitals to mix.

The electronegativity values for the two atoms are...

- Similar
  - The atoms are classified as...
    - Metals
      - Metallic Bonding (marked as incorrect)
    - Nonmetals
      - Covalent Bonding
        - The electronegativity values are...
          - Very close
            - Nonpolar Covalent
          - Different
            - Polar Covalent

- Very different
  - Ionic Bonding
Covalent bonds

Polar Bond
Creates permanent dipoles

Nonpolar Bond

When sodium (Na) and chlorine (Cl) are combined, the sodium atoms each lose an electron, forming cations (Na\(^+\)), and the chlorine atoms each gain an electron to form anions (Cl\(^-\)). These ions are then attracted to each other in a 1:1 ratio to form sodium chloride (NaCl).

\[ \text{Na} + \text{Cl} \rightarrow \text{Na}^+ + \text{Cl}^- \rightarrow \text{NaCl} \]

Baumketner, BioSim, Lviv 2019
Noncovalent bonds

Hydrogen Bond

Electropositive hydrogen atom is shared by two electronegative atoms. Covalent bond is partially distorted. Interaction is weak, last a short period of time due to thermal motion.

Molecules that contain polar bonds and that can form H-bonds in water dissolve easily in water (hydrophilic). Nonpolar molecules do not dissolve in water (hydrophobic)

Van der Waals Interaction

The electron cloud of an atom fluctuates, producing a flickering dipole. Such dipoles induce oppositely flickering dipoles in a nearby atom, generating a weak interaction.

Baumketner, BioSim, Lviv 2019
Hydrophobic interactions

Гідрофоб – речовина яка не змішується з водою, олії, нафта та інші вуглеводні

Явище виштовхування

Гідрофобні частинки злипаються – ефективне притягання

Baumketner, BioSim, Lviv 2019
- Molecules do not fall apart by thermal agitation.
- The energy of noncovalent interactions are in the range of thermal noise in the environment.
- ATP hydrolysis energy exceeds noncovalent interactions and thermal motions
- Covalent bond energy can be used to synthesize multiple ATPs
1) Processes are driven by free energy not internal energy

- **Protein folding**
- **Formation of hydrogen molecule**

20–40 kJ/mol

**One order of magnitude difference!**

- a) Strong influence of thermal fluctuations and entropy
- b) Important role of non-covalent and solvent-mediated interactions
- c) Very little chemistry other than in enzymes
Biological molecules

Proteins

Lipids

Nucleic acids

Sugars

Baumketner, BioSim, Lviv 2019
Sugars

Different monosaccharides

RNA

1. Aldoses
   - D-Ribose (Rib)
   - D-Xylose (Xyl)
   - L-Arabinose (Ara)

   Pentoses

   - D-Glucose (Glc)
   - D-Mannose (Man)
   - D-Galactose (Gal)

   Hexoses

2. Ketoses
   - D-Ribulose (Rub)
   - D-Ribose (Rib)
   - D-Fructose (Fru)

3. Deoxyaldoses
   - 2-Deoxy-D-ribose (dRib)
   - L-Fucose (Fuc)

4. Acetylated amino sugars
   - N-Acetyl-d-glucosamine (GlcNAc)
   - N-Acetyl-d-galactosamine (GalNAc)

Carbohydrates with hydroxyl groups

Different isomers

Energy source and storage.
Cell wall (mechanical support)
Glycoproteins, glycolipids (surface adhesion, extracellular signaling, cell-cell interactions)

Baumketner, BioSim, Lviv 2019
Sugars

Polymerization

B. Disaccharides

1. Maltose
   $\alpha$-D-Glucopyranosyl-(1$\rightarrow$4)-D-glucopyranose

2. Lactose
   $\beta$-D-Galactopyranosyl-(1$\rightarrow$4)-D-glucopyranose

3. Sucrose
   $\alpha$-D-Glucopyranosyl-(1$\rightarrow$2)-$\beta$-D-fructofuranoside

Baumketner, BioSim, Lviv 2019
## Sugars

### Important polysaccharides

<table>
<thead>
<tr>
<th>Polysaccharide</th>
<th>Monosaccharide 1</th>
<th>Monosaccharide 2</th>
<th>Linkage</th>
<th>Branching</th>
<th>Occurrence</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bacteria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Murein Dextran</td>
<td>D-GlcNAc</td>
<td>D-MurNAc</td>
<td>$\beta 1 \to 4$</td>
<td>$\alpha 1 \to 6$</td>
<td>Cell wall</td>
<td>SC, WB</td>
</tr>
<tr>
<td></td>
<td>D-Glc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Plants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agarose Carrageenan</td>
<td>D-Gal</td>
<td>L-aGal</td>
<td>$\beta 1 \to 3$</td>
<td>$\alpha 1 \to 3$</td>
<td>Red algae (agar)</td>
<td>WB, WB</td>
</tr>
<tr>
<td>Cellulose Xyloglucan</td>
<td>D-Glc</td>
<td>D-Xyl (D-Gal, L-Fuc)</td>
<td>$\beta 1 \to 4$</td>
<td>$\beta 1 \to 4$</td>
<td>Cell wall</td>
<td>SC, SC, SC</td>
</tr>
<tr>
<td>Arabinan Amylose Amylopectin Inulin</td>
<td>L-Ara</td>
<td>D-Glc</td>
<td>$\alpha 1 \to 5$</td>
<td>$\beta 1 \to 2$</td>
<td>Amyloplasts</td>
<td>RC, RC, RC</td>
</tr>
<tr>
<td></td>
<td>D-Glc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Animals</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chitin Glycogen Hyaluronic acid</td>
<td>D-GlcNAc</td>
<td>D-GlcNAc</td>
<td>$\beta 1 \to 4$</td>
<td>$\alpha 1 \to 3$</td>
<td>Insects, crabs</td>
<td>SK, RK, SK, WB</td>
</tr>
<tr>
<td></td>
<td>D-GluC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*SC, WB = growth for submerged cultures, WB = growth for solid media.

[http://www.irmed.ir](http://www.irmed.ir)
Sugars attached to proteins = glyco proteins
Sugars

C. Glycoproteins: forms

O-linked

N-linked

Mannose-rich type

Complex type

Baumketner, BioSim, Lviv 2019
Lipids

B. Biological roles

1. Fuel
   - Fat
   - Glycerol
   - ADP + P_i
   - ATP
   - O_2
   - CO_2
   - H_2O
   - Mitochondrion

2. Building block
   - Membrane
   - Cytoplasm
   - Phospholipid
   - Lipid bilayer

3. Thermal insulator
   - Cell
   - 37 °C
   - 0 °C

4. Special tasks
   - Signaling
   - Anchor
   - CoQ
   - Cofactor
   - Visual pigment

Baumketner, BioSim, Lviv 2019
### Lipids

**A. Carboxylic acids**

<table>
<thead>
<tr>
<th>Name</th>
<th>Number of carbons</th>
<th>Number of double bonds</th>
<th>Position of double bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formic acid</td>
<td>1 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acetic acid</td>
<td>2 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propionic acid</td>
<td>3 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Butyric acid</td>
<td>4 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valerianic acid</td>
<td>5 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caproic acid</td>
<td>6 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caprylic acid</td>
<td>8 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capric acid</td>
<td>10 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lauric acid</td>
<td>12 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myristic acid</td>
<td>14 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palmitic acid</td>
<td>16 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stearic acid</td>
<td>18 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oleic acid</td>
<td>18 : 1; 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linoleic acid</td>
<td>18 : 2; 9,12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linolenic acid</td>
<td>18 : 3; 9,12,15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arachidic acid</td>
<td>20 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Arachidonic acid</strong></td>
<td>20 : 4; 5,8,11,14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behenic acid</td>
<td>22 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erucic acid</td>
<td>22 : 1; 13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lignoceratic acid</td>
<td>24 : 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nervonic acid</td>
<td>24 : 1; 15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Essential in human nutrition*
Proteins

Amino acid

Zwitterion
Proteins

No rotation around double bonds!
Proteins

No rotation around C-N or O-C bonds. The peptide bond is planar and rigid!

Phi/Psi angles are the only real degrees of freedom of the protein backbone.

Cost of rotation is 90kJ/mol. Compare to 2.5kJ/mol of thermal energy at 300K.
Ramachandran plot

Baumketner, BioSim, Lviv 2019
Primary structure

...-ASP-ALA-VAL-ILE-ASP-SER-GLU-PRO-THR-...

...DAVIDSEPT...
Secondary structure

H-bonding between N-H and C=O groups without involving side chains.

**Alpha Helix**

- C=O of one residue bond to N-H of the fourth residue
- 3.6 amino acid residues per turn.
- Helical pitch is 0.54 nm.

**Beta Sheet**

- C=O of one residue bond to N-H of a residue on another strand
- 0.48 nm between strands
- 0.35 nm per residue

Baumketner, BioSim, Lviv 2019
Tertiary structure

- Helices and sheets often combine in various ways.
- Certain combinations of α and β repeat over and over, called MOTIFS

Four Helix Bundle  Beta Barrel  Coiled Coil

Baumketner, BioSim, Lviv 2019
Protein folding

Driven by noncovalent bond formation and hydrophobic effect

Folded state is the energetically stable state, spontaneously occurring in water.

3D shape of a protein is determined by its amino acid sequence.

Baumketner, BioSim, Lviv 2019
Interactions between proteins

(A) free subunits → assembled structures
- binding site → dimer

(B) free subunits → assembled structures
- binding sites → helix
- binding sites → ring

(C) actin molecule
- minus end
- 37 nm
- plus end

Baumketner, BioSim, Lviv 2019
Enzymatic proteins
Function: Selective acceleration of chemical reactions
Example: Digestive enzymes catalyze the hydrolysis of bonds in food molecules.

- **Enzymes** are a type of protein that acts as a **catalyst** to speed up chemical reactions.
- Enzymes can perform their functions repeatedly, functioning as workhorses that carry out the processes of life.
- [http://www.biotopics.co.uk/other/morinf.html](http://www.biotopics.co.uk/other/morinf.html)
Function of proteins

Storage proteins

Function: Storage of amino acids
Examples: Casein, the protein of milk, is the major source of amino acids for baby mammals. Plants have storage proteins in their seeds. Ovalbumin is the protein of egg white, used as an amino acid source for the developing embryo.
Function of proteins

Hormonal proteins

Function: Coordination of an organism's activities
Example: Insulin, a hormone secreted by the pancreas, causes other tissues to take up glucose, thus regulating blood sugar concentration

Insulin protein entry:
Contractile and motor proteins

Function: Movement

Examples: Motor proteins are responsible for the undulations of cilia and flagella. Actin and myosin proteins are responsible for the contraction of muscles.
Function of proteins

**Defensive proteins**

Function: Protection against disease
Example: Antibodies inactivate and help destroy viruses and bacteria.

![Antibodies bind to virus](image1)

**Transport proteins**

Function: Transport of substances
Examples: Hemoglobin, the iron-containing protein of vertebrate blood, transports oxygen from the lungs to other parts of the body. Other proteins transport molecules across cell membranes.

![Transport protein in cell membrane](image2)

Recognizes pathogen (antigen) via fragment antigen binding motif

Baumketner, BioSim, Lviv 2019
Nucleotides

Phosphate

Nucleoside

Sugar ring (ribose or deoxyribose)

Nucleobase
Nucleic acids = poly-nucleotides

Uracil

Oligonucleotides (DNA, RNA)
Cellular energy (ATP)

ACTGU – primary bases

ACGT - DNA
ACGU - RNA

guanine
adenine
thymine
cytosine

Baumketner, BioSim, Lviv 2019
ATP as energy source

Baumketner, BioSim, Lviv 2019
DNA structure

AT and GC pairings are possible due to HB geometry

- Forms a double helix.
- Each turn is made of 10 nucleotide pairs.
- 3.4 nm between adjacent nucleotide

Baumketner, BioSim, Lviv 2019
Genetic code

Multiple codons for the same AA

Some codons are silent, or are they?

Figure 1.4 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Baumketner, BioSim, Lviv 2019
How is genetic information stored?

Ген – код для одного або декількох білків.

46 total chromosomes in each normal cell, grouped into 23 pairs, referred to by number
Corresponding sets of maternal and paternal genes in each pair of chromosomes
A specialized pair of chromosomes that determines a person's sex: females have two X chromosomes and males have one X and one Y.

Baumketner, BioSim, Lviv 2019
Transcription + translation

1) DNA synthesis (replication) → DNA

2) RNA synthesis (transcription) → RNA

3) Protein synthesis (translation) → PROTEIN

Proteins

CENTRAL DOGMA

DNA transcription and mRNA translation
http://www.youtube.com/watch?v=41_Ne5mS2ls

Baumketner, BioSim, Lviv 2019
DNA replication

Unzipping

Loosening the supercoil

Catalyzing the elongation by one unit
DNA replication

Show DNA polymerase advanced
http://www.youtube.com/watch?v=I9ArIJWTZHI&feature=related
1. RNA polymerase, together with one or more general transcription factors, binds to promoter DNA.
2. RNA polymerase creates a transcription bubble, which separates the two strands of the DNA helix. This is done by breaking the hydrogen bonds between complementary DNA nucleotides.
3. RNA polymerase adds RNA nucleotides (which are complementary to the nucleotides of one DNA strand).
4. RNA sugar-phosphate backbone forms with assistance from RNA polymerase to form an RNA strand.
5. Hydrogen bonds of the RNA–DNA helix break, freeing the newly synthesized RNA strand.

Speed of 10-100 nts/sec
Protein synthesis = translation

Transfer RNA (abbreviated tRNA) is an adaptor molecule composed of RNA, typically 73 to 94 nucleotides in length, that serves as the physical link between the nucleotide sequence of nucleic acids (DNA and RNA) and the amino acid sequence of proteins.

Baumketner, BioSim, Lviv 2019
Proteins modifications

**Modification**
- Phosphorylation
- Acetylation
- N-linked glycosylation
- Amidation
- Hydroxylation
- Methylation
- O-linked glycosylation
- Ubiquitylation
- Pyrrolidone carboxylic acid
- Sulfation

**Diagram**

- **D:** Pyroglutamyl-Acetyl-Formyl-Myristoyl-
- **B:** Oligosaccharide (O-glycosylation)
- **B:** Oligosaccharide (N-glycosylation)
- **D:** Phospho-Methyl-γ-Carboxy-(Glu)
- **D:** Acetyl-Methyl-γ-Hydroxy-(Glu)
- **B:** Pyridoxal-Liponat Biotin Retinal Ubiquitin

- **Modification**
  - Ser, Thr
  - Asn, Gln
  - Asp, Glu
  - Lys
  - Pro

- **Modification**
  - Tyr
  - Phe
  - His
  - Cys

- **Modification**
  - D: Amidated -(CONH₂)
  - D: Phospho-iodo-Sulfato-Adenyl-
  - D: 4-Hydroxy-(Tyrosine)
  - D: Phospho-Methyl-B: Flavin
  - D: Disulfide Prenyl-B: Heme Flavin

By Baumketner, BioSim, Lviv 2019
Абетки життя

NUCLEIC ACIDS

G  T
C  A
nucleotides

codon

PROTEINS

A  V  F  P  D
L  I  M  E  K
T  Y  R  C  N
Q  H  W  G
amino acids

α-helix

β-strand

protein

Figure 1.2 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Baumketner, BioSim, Lviv 2019
Metabolism

Процес перетворення енергії в клітинах

Катаболізм

Анаболізм
Приклад катаболізму: Гліколіз

Процес перетравлювання цукру (глюкози) досить складний. В результаті утворюються молекули ATP та піруват. Споживається кисень для реакції оксидації та виділяється дво-окис вуглецю.
Very little chemistry happens during the majority of all biological processes so the appropriate level of description is classical. This entails:

1) Adiabatic approximation. Nuclei are moving in the field created by the electrons.

2) Relaxation processes taking place on picosecond timescale and slower.

When are QM effects important?

For harmonic oscillator, for instance:

\[ \Delta E = \hbar \nu \]

at T=300K


Baumketner, BioSim, Lviv 2019
Perturbation theory doesn't work. Valence-bond theory, LDA, Hartree-Fock don't work.

Electrons shared by two nuclei

Perturbation theory doesn't work. Valence-bond theory, LDA, Hartree-Fock don't work.

This part of the curve needs to be handled differently from this part!

Baumketner, BioSim, Lviv 2019
Non-bonded energy

Multipole expansion for point charges

Potential energy of the set of charges $q_i$ interacting with charge $q$ at the origin

$$U = \sum_{i}^{N} \frac{qq_i}{|\vec{R} + \vec{r}_i|} = q \sum_{i}^{N} q_i e^{\frac{\vec{r}_i \cdot \vec{v}}{R}} \frac{1}{R}$$

where $e^{\frac{\vec{r}_i \cdot \vec{v}}{R}} = 1 + \vec{r}_i \cdot \vec{v} + \frac{1}{2} (\vec{r}_i \cdot \vec{v}) (\vec{r}_i \cdot \vec{v}) + ...$ - the translation operator

In a more compact form

$$U = q \Phi(\vec{R}) = q \hat{Q} \frac{1}{R}, \quad \Phi(\vec{R}) \quad \text{is the potential created by the charges at the origin}$$

where

$$\hat{Q} = \sum_{i} q_i e^{\frac{\vec{r}_i \cdot \vec{v}}{R}} = \sum_{i} q_i (1 + \vec{r}_i \cdot \vec{v} + \frac{1}{2} (\vec{r}_i \cdot \vec{v}) (\vec{r}_i \cdot \vec{v}) + ... ) = Q + \vec{d} \cdot \vec{v} + \frac{1}{2} \sum_{\alpha \beta} Q_{\alpha \beta} \frac{\partial}{\partial X_{\alpha}} \frac{\partial}{\partial X_{\beta}} + ...$$

$$Q = \sum_{i}^{N} q_i \quad \text{total charge} \quad \vec{d} = \sum_{i}^{N} q_i \vec{r}_i \quad \text{dipole moment} \quad Q_{\alpha \beta} = \sum_{i}^{N} q_i x_{\alpha}^i x_{\beta}^i \quad \text{quadrupole moment}$$

Baumketner, BioSim, Lviv 2019
Potential energy then can be written as series:

\[ U = q \Phi(R) = q \varphi_m(R) + q \varphi_d(R) + q \varphi_Q(R) + \ldots \]

\[ \varphi_m = \frac{Q}{R} \quad \text{potential created by the monopole=total charge} \sim \frac{1}{R} \]

\[ \varphi_d = -\frac{d \vec{R}}{R^3} \quad \text{potential created by point dipole (at vector } -\vec{R} \text{) } \sim \frac{1}{R^2} \]

\[ \varphi_Q = \frac{1}{2} \sum_{\alpha\beta} Q_{\alpha\beta} \left( \frac{3X_{\alpha}X_{\beta}}{R^2} - \delta_{\alpha\beta} \right) \frac{1}{R^3} \quad \text{potential created by point quadrupole } \sim \frac{1}{R^3} \]

Another way to look the interaction energy is to compute interaction of point multipoles with the field created by the charge at the origin:

\[ U = Q \frac{q}{R} \quad \text{potential created at the location of the charge distribution by point charge at the origin} \]

\[ -\hat{d} \hat{E} \quad \text{field created by the point charge at the location of the dipole} \]

\[ \hat{E} = -q \hat{\nabla} \frac{1}{R} = \frac{q \hat{R}}{R^3} \quad \text{derivatives of the electric field created by the point charge} \]

Monopole “interacts” with the potential

Dipole “interacts” with the field

Quadrupole “interacts” with the field derivative

Baumketner, BioSim, Lviv 2019
Apply these formulas to two distributions of charges

To simplify things use the following abbreviations:

- **charge**
- **dipole**
- **quadrupole**
- **octopole**

Potential at site A:

Electric field at site A:

Field derivative at site A:

\[
U = \frac{Q_A Q_B}{R} - Q_A \frac{d_B \vec{R}}{R^3} \sim \frac{1}{R^3} + \frac{Q_A \frac{d_{AB}}{R^3}}{R^3} \sim \frac{1}{R^4} + \ldots
\]

Multipole expansion

Interaction energy as a series in powers of \( \frac{1}{R} \)

\[ \sim \frac{1}{R^n} \]

Baumketner, BioSim, Lviv 2019
A table of the order of different interactions

<table>
<thead>
<tr>
<th></th>
<th>●</th>
<th>▲</th>
<th>□</th>
<th>●</th>
</tr>
</thead>
<tbody>
<tr>
<td>●</td>
<td>$\sim \frac{1}{R}$</td>
<td>$\sim \frac{1}{R^{2}}$</td>
<td>$\sim \frac{1}{R^{3}}$</td>
<td>$\sim \frac{1}{R^{4}}$</td>
</tr>
<tr>
<td>▲</td>
<td></td>
<td>$\sim \frac{1}{R^{3}}$</td>
<td>$\sim \frac{1}{R^{4}}$</td>
<td>$\sim \frac{1}{R^{5}}$</td>
</tr>
<tr>
<td>□</td>
<td></td>
<td></td>
<td>$\sim \frac{1}{R^{5}}$</td>
<td>$\sim \frac{1}{R^{6}}$</td>
</tr>
<tr>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td>$\sim \frac{1}{R^{7}}$</td>
</tr>
</tbody>
</table>
Interaction energy between two atoms

**Full Hamiltonian:**

\[ \hat{H} = \hat{H}_A + \hat{H}_B + \hat{U} \]

\[ \hat{U} = \frac{Z_A Z_B}{|\vec{R}_A - \vec{R}_B|} + \sum_{i=1}^{Z_A} \sum_{j=1}^{Z_B} \frac{1}{|\vec{R}_A + \vec{r}_{Ai} - \vec{R}_B - \vec{r}_{Bj}|} \]

\[ - \sum_{i=1}^{Z_A} \frac{Z_B}{|\vec{R}_A + \vec{r}_{Ai} - \vec{R}_B|} - \sum_{j=1}^{Z_B} \frac{Z_A}{|\vec{R}_A - \vec{R}_B - \vec{r}_{Bj}|} \]

Compute the total energy by the perturbation theory:

\[ E = E^{(0)} + E^{(1)} + E^{(2)} + \cdots \]

\[ E^{(0)} = E_0^A + E_0^B \]

*ground state of individual atoms*

**Each atom is electrically neutral**

\[ Z_A \quad Z_B \quad \# \text{ of valence electrons and nuclei’s charges} \]

\[ |n_A > \quad |n_B > \quad \text{eigenfunctions} \]

\[ E_n^A \quad E_n^B \quad \text{eigenvalues of energy operators} \]

**zero-order Hamiltonian**

\[ \hat{H}_0 = \hat{H}_A + \hat{H}_B \]

**perturbation term**

\[ \hat{U} \]

**first-order term**

\[ E^{(0)} = < 0_A 0_B | \hat{H}_0 | 0_A 0_B > \]

Baumketner, BioSim, Lviv 2019
Using the translation vector formula:

$$e^{r_{Ai} \cdot \nabla} e^{-r_{Bj} \cdot \nabla} \frac{1}{R} = \frac{1}{|R + r_{Ai} - r_{Bj}|},$$

$$\to \hat{U} = \sum_{i=1}^{Z_A} \sum_{j=1}^{Z_B} \left( e^{r_{Ai} \cdot \nabla} - 1 \right) \left( e^{-r_{Bj} \cdot \nabla} - 1 \right) \frac{1}{R}$$

Upon introducing charge operators:

$$\hat{Q}_A = \sum_{i=1}^{Z_A} \left( e^{r_{Ai} \cdot \nabla} - 1 \right)$$

Perturbation term can be written as:

$$\hat{U} = \hat{Q}_A \hat{Q}_B^+ \frac{1}{R}$$

First-order term:

$$E^{(1)} = <0 | \hat{U} | 0> = <0_A 0_B | \hat{Q}_A \hat{Q}_B^+ \frac{1}{R} | 0_A 0_B> = <0_A | \hat{Q}_A | 0_A> <0_B | \hat{Q}_B^+ | 0_B> \frac{1}{R}$$

Let see the first few terms explicitly:

$$\hat{Q}_A = \sum_i \left( 1 + \left( \mathbf{r}_{i}^{A} \cdot \nabla \right) + \frac{1}{2} \left( \mathbf{r}_{i}^{A} \cdot \nabla \right) \left( \mathbf{r}_{i}^{A} \cdot \nabla \right) + \cdots - 1 \right) = \mathbf{d}^{A} \cdot \nabla + \frac{1}{2} \sum_{\alpha \beta} Q_{A}^{\alpha \beta} \frac{\partial}{\partial X_{\alpha}} \frac{\partial}{\partial X_{\beta}} + \cdots$$

$$\hat{Q}_B^+ = \sum_i \left( 1 - \left( \mathbf{r}_{i}^{B} \cdot \nabla \right) + \frac{1}{2} \left( \mathbf{r}_{i}^{B} \cdot \nabla \right) \left( \mathbf{r}_{i}^{B} \cdot \nabla \right) + \cdots - 1 \right) = -\mathbf{d}^{B} \cdot \nabla + \frac{1}{2} \sum_{\alpha \beta} Q_{B}^{\alpha \beta} \frac{\partial}{\partial X_{\alpha}} \frac{\partial}{\partial X_{\beta}} + \cdots$$

If atoms have non-zero charge:  

$$\hat{Q}_A = Q_A + \left( \mathbf{d}^{A} \cdot \nabla \right) + \frac{1}{2} \sum_{\alpha \beta} Q_{A}^{\alpha \beta} \frac{\partial}{\partial X_{\alpha}} \frac{\partial}{\partial X_{\beta}} + \cdots$$

Baumketner, BioSim, Lviv 2019
Taking the expectation value:

\[
< 0_A | \hat{Q}_A | 0_A > = < \vec{d}^A > \vec{v} + \frac{1}{2} \sum_{\alpha \beta} < Q^A_{\alpha \beta} > \frac{\partial}{\partial x_\alpha} \frac{\partial}{\partial x_\beta} + ...
\]

\[
< 0_B | \hat{Q}^+_{-B} | 0_B > = - < \vec{d}^B > \vec{v} + \frac{1}{2} \sum_{\gamma \delta} < Q^B_{\gamma \delta} > \frac{\partial}{\partial x_\gamma} \frac{\partial}{\partial x_\delta} + ...
\]

\[
E^{(1)} = - \left( < \vec{d}_A > \vec{v} \right) \left( < \vec{d}_B > \vec{v} \right) \frac{1}{R} + \sum_{\alpha \beta} < d^A_\alpha > < d^B_\alpha > \left( \frac{3X_\alpha X_\beta}{R^5} - \frac{\delta_{\alpha \beta}}{R^3} \right) \sim \frac{1}{R^3}
\]

if dipole moments are non-zero. true for molecules but not for atoms

\[
+ \frac{1}{2} \sum_{\gamma \delta} \left\{ ( < \vec{d}_A > \vec{v} ) < Q^B_{\gamma \delta} > - ( < \vec{d}_B > \vec{v} ) < Q^A_{\gamma \delta} > \right\} \frac{\partial}{\partial x_\gamma} \frac{\partial}{\partial x_\delta} \frac{1}{R} \sim \frac{1}{R^4}
\]

\[
+ \frac{1}{4} \sum_{\alpha \beta \gamma \delta} < Q^A_{\alpha \beta} > < Q^B_{\gamma \delta} > \frac{\partial}{\partial x_\alpha} \frac{\partial}{\partial x_\beta} \frac{\partial}{\partial x_\gamma} \frac{\partial}{\partial x_\delta} \frac{1}{R} + ... \sim \frac{1}{R^5}
\]

\[
= \frac{C^{(3)}}{R^3} + \frac{C^{(4)}}{R^4} + \frac{C^{(5)}}{R^5} + ... = \frac{Q_A Q_B}{R} + \frac{C^{(2)}}{R^2} + \frac{C^{(3)}}{R^3} + \frac{C^{(4)}}{R^4} + \frac{C^{(5)}}{R^5} + ...
\]

"electronic interaction energy" due to permanent multipole moments

multipole expansion expansion if atoms have uncompensated charge

Baumketner, BioSim, Lviv 2019
Second-order term:

\[ E_n^{(2)} = \sum_{m(n \neq m)} \frac{|U_{mn}|^2}{E_n^{(0)} - E_m^{(0)}} \]

Correction to the ground-state energy relies on the matrix entry:

\[ U_{m0} = \langle m|\hat{U}|0\rangle = \langle m_A|\hat{Q}_A|0_A\rangle \langle 0_B|\hat{Q}^+_B|0_B\rangle \frac{1}{R}, \]

which has to be substituted into the energy formula:

\[ E_0^{(2)} = \sum_{m_A, m_B} \frac{|\langle m_A|\hat{Q}_A|0_A\rangle \langle m_B|\hat{Q}^+_B|0_B\rangle \frac{1}{R}|^2}{E_0^A - E_0^{m_A} + E_0^B - E_0^{m_B}} = E_{0,ind}^{(2)} + E_{0,disp}^{(2)} \]

Induction energy

Summation is performed while keeping one of the atoms in the ground state

\[ E_{0,ind}^{(2)} = \sum_{m_B} \frac{|\langle 0_A|\hat{Q}_A|0_A\rangle \langle 0_B|\hat{Q}^+_B|0_B\rangle \frac{1}{R}|^2}{E_0^B - E_0^{m_B}} + \sum_{m_A} \frac{|\langle m_A|\hat{Q}_A|0_A\rangle \langle 0_B|\hat{Q}^+_B|0_B\rangle \frac{1}{R}|^2}{E_0^A - E_0^{m_A}} \]

non-zero starting from non-zero multipole.

note that since the denominator is always negative and nominator – positive the correction is negative and corresponds to attraction

Baumketner, BioSim, Lviv 2019
The induction energy is non-zero starting from non-zero multipole terms.

Examples:
1) Atoms have non-zero dipole in the ground state

\[
E_{\text{ind}}(A) \sim \left| < \hat{d}^A > \hat{\nabla} < m_B | \hat{Q}^+_B | 0_B > \frac{1}{R} \right|^2 = \frac{I^{(6)}}{R^6} + \ldots
\]

is the lowest term in expansion which has to squared

2) Atoms have non-zero charge and may or may not have dipole moment

\[
E_{\text{ind}}(A) \sim \left| < 0_A | \hat{Q}_A | 0_A > < m_B | \hat{Q}^+_B | 0_B > \frac{1}{R} \right|^2 = \left| Q_A < m_B | \hat{Q}^+_B | 0_B > \frac{1}{R} \right|^2 = \frac{I^{(4)}}{R^4} + \ldots
\]

A convenient way to view induction energy is by introducing the concept of polarization.
Assume that the dipole moment induced by external field can be written as:

\[
\hat{\mu} = \alpha \hat{E}
\]

where \( \alpha \) is the polarizability constant (tensor in general)

The interaction energy of that moment with the field then is:

\[
E_{\text{ind}} = -\hat{\mu} \hat{E} = -\alpha \hat{E}^2
\]

If polarization is caused by point charge, \( E \sim \frac{1}{R^2} \) and so
\( E_{\text{ind}} \sim \frac{1}{R^4} \)

interaction of charge with induced dipole

If polarization is caused by a point dipole \( E \sim \frac{1}{R^3} \) and so
\( E_{\text{ind}} \sim \frac{1}{R^6} \)

interaction of dipole with induced dipole

Baumketner, BioSim, Lviv 2019
A formal expression for the induction energy:

\[
E_{\text{ind}}(A) = -\frac{1}{2} \alpha_{\alpha \beta}(A) E_{\alpha}(A) E_{\beta}(A) - \frac{1}{3} A_{\alpha, \beta \gamma}(A) E_{\alpha}(A) E_{\beta \gamma}(A) - \frac{1}{6} C_{\alpha \beta, \gamma \delta}(A) E_{\alpha \beta}(A) E_{\gamma \delta}(A) + \ldots
\]

\[
E_{\alpha}(A) = -\frac{X_{\alpha}}{R^3} Q_B + \sum_{\beta} \left( \frac{3X_{\alpha} X_{\beta}}{R^5} - \frac{\delta_{\alpha \beta}}{R^3} \right) d_{\alpha}(B) + \ldots
\]

\[
E_{\alpha \beta}(A) = -\left( \frac{3X_{\alpha} X_{\beta}}{R^5} - \frac{\delta_{\alpha \beta}}{R^3} \right) Q_B + \ldots
\]

**Dispersion energy**

Second-order correction when the summation is performed over excited states of both atoms:

\[
E_{0,\text{disp}}^{(2)} = \sum_{m_A, m_B \neq 0} \left| \langle m_A | \hat{Q}_A | 0_A \rangle \langle 0_B | \hat{Q}^+_B | 0_B \rangle \right|^2 \frac{1}{R^2} \frac{1}{E_0^A - E_{m_A}^A + E_0^B - E_{m_B}^B}
\]

 Baumketner, BioSim, Lviv 2019
Keeping only the lowest-order dipolar term: \( \hat{Q}_A = \overrightarrow{d^A} \vec{V} \quad \hat{Q}^+_B = -\overrightarrow{d^B} \vec{V} \)

\[
E_{0,\text{disp}}^{(2)} = \sum_{m_A,m_B \neq 0} \frac{|(\langle m_A | \overrightarrow{d^A} | 0_A > \vec{V} | (\langle m_B | \overrightarrow{d^B} | 0_B > \vec{V}) \frac{1}{R} \rangle|^2}{E_0^A - E_{m_A}^A + E_0^B - E_{m_B}^B} = \frac{D^{(6)}}{R^6}
\]

For charged systems

\[
\hat{Q}_A = Q_A + \overrightarrow{d^A} \vec{V}
\]

0 because of orthogonality condition

\[
\langle m_A | \hat{Q}_A | 0_A > = Q_A \langle m_A | 0_A > + \langle m_A | \overrightarrow{d^A} \vec{V} | 0_A > = \langle m_A | \overrightarrow{d^A} \vec{V} | 0_A >
\]

so \( \frac{1}{R^6} \) is genuinely the lowest order term in the dispersion interaction

Some general properties of dispersion interactions

- attractive regardless of molecule orientation
- weaker than normal covalent and ionic bonds
- strength is proportional to the polarizability of the atom
- additive and cannot be saturated
- short-range forces and hence only interactions between the nearest particles need to be considered

Dispersion, London, van der Waals interaction


Baumketner, BioSim, Lviv 2019
Extension to molecules

Nuclear interaction

\[ \frac{Z_1^A Z_1^B}{|\vec{R} - \vec{R}_1^B + \vec{R}_1^A|} + \frac{Z_1^A Z_2^B}{|\vec{R} - \vec{R}_2^B + \vec{R}_2^A|} + \frac{Z_2^A Z_2^B}{|\vec{R} - \vec{R}_1^B + \vec{R}_2^A|} + \frac{Z_2^A Z_2^B}{|\vec{R} - \vec{R}_2^B + \vec{R}_2^A|} = (\vec{Q}_A + Z_A)(\vec{Q}_B + Z_B) \frac{1}{R} \]

Electrons with nuclei

\[ -1 \sum_{i=1,Z_A} \left\{ \frac{Z_1^B}{|\vec{R} - \vec{R}_1^B + \vec{r}_i^A|} + \frac{Z_2^B}{|\vec{R} - \vec{R}_2^B + \vec{r}_i^A|} \right\} + \]

\[ -1 \sum_{j=1,Z_B} \left\{ \frac{Z_1^A}{|\vec{R} + \vec{R}_1^A - \vec{r}_j^B|} + \frac{Z_2^A}{|\vec{R} + \vec{R}_2^A - \vec{r}_j^B|} \right\} = -\{(\vec{Q}_A + Z_A)(\vec{q}_B + Z_B) + (\vec{Q}_B + Z_B)(\vec{q}_A + Z_A)\} \frac{1}{R} \]

Electrons with electrons

\[ \sum_{i=1,Z_A} \sum_{j=1,Z_B} \frac{1}{|\vec{R} - \vec{r}_j^B + \vec{r}_i^A|} = (\vec{q}_A \vec{q}_B^* + Z_A \vec{q}_B^* + Z_B \vec{q}_A + Z_A Z_B) \frac{1}{R} \]

Baumketner, BioSim, Lviv 2019
Full perturbation potential

\[ E_{0,n}^{(1)} \neq 0, E_{0,n}^{(2)} = 0 \]

\[ \hat{U} = \{ \widehat{Q}_A \widehat{Q}_B^+ + \widehat{q}_A \widehat{q}_B^+ - \widehat{Q}_A \widehat{q}_B^+ - \widehat{q}_A \widehat{Q}_B^+ \} \frac{1}{R} \]

Additional terms:

\[ E_{0,n}^{(1)} = \langle 0 | \widehat{Q}_A \widehat{Q}_B^+ | 0 \rangle > \frac{1}{R} = \frac{N^{(3)}}{R^3} + \ldots \]

\[ E_{0,n}^{(2)} \sim < m_A | \widehat{Q}_A \widehat{Q}_B^+ | 0_A > = \widehat{Q}_A \widehat{Q}_B^+ < m_A | 0_A > = 0 \]

because of the orthogonality of the excited states to the ground-state wave function

\[ E_{0,c}^{(1)} = \langle 0 | \widehat{Q}_A \widehat{q}_B^+ | 0 \rangle > \frac{1}{R} = \widehat{Q}_A < 0 | \widehat{q}_B^+ | 0 > \frac{1}{R} = \frac{CN^{(3)}}{R^3} + \ldots \]

\[ E_{0,c}^{(2)} \sim \sum_{m \neq 0} | < 0 | \widehat{Q}_A \widehat{q}_B^+ | m > \frac{1}{R} |^2 = \sum_{m \neq 0} | \widehat{Q}_A < 0 | \widehat{q}_B^+ | m > \frac{1}{R} |^2 = \frac{CN^{(6)}}{R^6} + \ldots \]

Nuclear degrees of freedom contribute additional terms to the multipole expansion corresponding to the permanent moments

Baumketner, BioSim, Lviv 2019
Putting all summands together one arrives at the most general representation of the interaction energy between two molecules:

\[
U = \frac{Q_A Q_B}{R} + \frac{C^{(2)}}{R^2} + \frac{C^{(3)}}{R^3} + \frac{C^{(4)}}{R^4} + \frac{C^{(5)}}{R^5} + \frac{C^{(6)}}{R^6} + \cdots
\]

Permanent multipoles.

Electronic and nuclear systems.

\[
\begin{align*}
+ \frac{I^{(4)}}{R^4} & + \frac{I^{(6)}}{R^6} + \cdots & \text{induction energy} \\
+ \frac{D^{(6)}}{R^6} & + \cdots & \text{dispersion energy}
\end{align*}
\]

Constants

\[C^{(i)}, I^{(i)}, D^{(i)}\]

a) Can be computed in QM studies. Difficult for large systems.

Almost impossible for dispersion force.

b) Obtained by fitting. Empirical parameters.
Atom-pair potential approximation

Proteins are modeled at the classical level. 1) Not much chemistry happens but 2) QM calculations are way too expensive.

Interaction energy $U(R)$ can be used to model the dynamics of the nuclei in the Born-Oppenheimer approximation. Adiabatic approximation.

Two options of how to proceed:

Option #1:

Assign proper (valence) charges to each nucleus and a certain number of permanent moments. These can be taken from QM calculations or some other source (from experiment in case of dipole moment for instance). The moments have to be attached to the local geometry of the molecules. As the local reference frame moves (rotates) the moments have to be recomputed.

Drawbacks:

1) Algorithms are not always straightforward to implement. It may be difficult to compute forces acting on each nucleus, especially for higher moments. See Stockmayer fluid for example. Torques are not always computed in a pairwise summation ($i$ acting on $j$ is not the opposite of $j$ acting on $i$).

2) The procedure is tedious and expensive. Each molecule has to have its moments recomputed at each step as it rotates in the course of the simulation.

3) A large amount of data needs to be stored.

4) A large number of parameters – dipole, quadrupole and higher moments, makes parametrization very challenging.

Baumketner, BioSim, Lviv 2019
Option #2:

1) Assign partial charges to each nucleus. These will generate multipole series of permanent dipoles.

\[
\sum_i \frac{q_i}{|\vec{R} - \vec{r}_i|} = \frac{C^{(1)}}{R} + \frac{C^{(2)}}{R^2} + \frac{C^{(3)}}{R^3} + \frac{C^{(4)}}{R^4} + \frac{C^{(5)}}{R^5} + \frac{C^{(6)}}{R^6} + \cdots
\]

Important: the series will contain all powers of 1/R, not just ones specific for a particular dipole.

2) Add polarization and dispersion interactions to each nucleus.

\[
\sum_{i,j} \frac{D_{ij}^{(4)}}{r_{ij}^4} + \frac{D_{ij}^{(6)}}{r_{ij}^6} + \frac{D_{ij}^{(7)}}{r_{ij}^7} + \frac{D_{ij}^{(8)}}{r_{ij}^8} + \cdots + \frac{D_{ij}^{(12)}}{r_{ij}^{12}}
\]

Parameters: \( q_1 \ldots q_N \)

\( D_{ij}^{(n)}, n = 4,6,7,8 \ldots 12 \) for each pair \( q_i, q_j \)

Baumketner, BioSim, Lviv 2019
Can the atom-pair approximation work?

The method of Clementi

1) Approximation: cut all terms with powers $1/R^7$ and higher. This will lower the number of parameters

2) Compute potential energy for a two-molecule system in QM calculations.

- $A =$ amino acids and some other systems. 25 in total
- $B =$ water molecule

$\Delta E_{AB}(R, \theta, \varphi)$ interaction energy as a function of mutual distance and orientations. 10,000 different values

The interaction potential

$$\Delta E = \sum_{a(X)} \sum_{b(W)} \left\{ c_{ab}^{(1)}/R_{ab} + c_{ab}^{(4)}/R_{ab}^4 + c_{ab}^{(6)}/R_{ab}^6 + c_{ab}^{(12)}/R_{ab}^{12}\right\}$$

Results:

1) Way too many parameters to perform a fit for all atoms. Introduce atom types. Typical types:

- sp3, sp3 hybridized carbon, carbon in aromatic residues etc. ~30 different classes

2) Electrostatic contribution can be well approximated by charges:

$$c_{ab}^{(1)} = q_a q_b$$

Baumketner, BioSim, Lviv 2019
3) Keeping the $1/R^4$ term doesn't improve the quality of the fit

4) Certain rules for cross terms seem to work well. For instance:

$$c_{a(X)H(W)}^{(12)} = c_{a(X)}^{(12)} c_{H(W)}^{(12)}$$

$$c_{a(X)O(W)}^{(12)} = c_{a(X)}^{(12)} c_{O(W)}^{(12)}$$

5) Coefficients $c_{ab}^{(6)}$ are too small and can't be determined reliably. This is the consequence of the dispersion interactions not being well described by the QM approximation.

Conclusions

The following model of potential energy will work well for proteins:

$$U = \sum_{i,j} \left\{ \frac{q_i q_j}{r_{ij}} + \frac{c_{ij}^{(12)}}{r_{ij}^{12}} - \frac{c_{ij}^{(6)}}{r_{ij}^6} \right\}$$

Adjustable parameters

$q_1 \ldots q_N$ partial charges

$c_{ij}^{(12)}, c_{ij}^{(6)} > 0$

Baumketner, BioSim, Lviv 2019
Examples of when this approximation will fail

**oxygen, nitrogen ...**
- Charge is zero at each nucleus because a) the molecule is neutral and b) charges are equivalent.
- \( q = 0 \)
- \( q \neq 0 \) (for instance, metal ion)

**Actual energy**
- \( I^{(4)} \frac{1}{R^4} + \ldots \)
- **polarization term.**

**Atom-pair model**
- \( D^{(6)} \frac{1}{R^6} \)
- **dispersion term.**

**benzene**

**Molecules that have zero charge, zero dipole moment but non-zero quadrupole moment**
- \( q = 0 \)

**Actual energy**
- \( C^{(5)} \frac{1}{R^5} + \ldots \)

**Atom-pair model**
- \( D^{(6)} \frac{1}{R^6} \)
- **dispersion term.**

Baumketner, BioSim, Lviv 2019
QM calculations predict certain geometry for the studied molecule. Distortions from that geometry are described by a number of potential energy terms that collectively are known as “bonded energy”.

Bond-stretching potential

Morse potential (some basis in QM calculations)

\[ v(l) = D_e \{1 - \exp[-a(l - l_0)]\}^2 \]

Three parameters. Not convenient

Harmonic approximation (typically used)

\[ v(l) = \frac{k}{2} (l - l_0)^2 \]

Force constant. Source: normal mode analysis of QM, vibrational spectra

Reference bond length. Source: crystal structures, QM calculations

<table>
<thead>
<tr>
<th>Bond</th>
<th>( l_0 ) (Å)</th>
<th>( k ) (kcal mol(^{-1}) Å(^{-2}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Csp(^3)–Csp(^3)</td>
<td>1.523</td>
<td>317</td>
</tr>
<tr>
<td>Csp(^3)–Csp(^2)</td>
<td>1.497</td>
<td>317</td>
</tr>
<tr>
<td>Csp(^2)=Csp(^2)</td>
<td>1.337</td>
<td>690</td>
</tr>
<tr>
<td>Csp(^2)=O</td>
<td>1.208</td>
<td>777</td>
</tr>
<tr>
<td>Csp(^3)=Nsp(^3)</td>
<td>1.438</td>
<td>367</td>
</tr>
<tr>
<td>C=N (amide)</td>
<td>1.345</td>
<td>719</td>
</tr>
</tbody>
</table>

Table 4.1 Force constants and reference bond lengths for selected bonds [Allinger 1977].

Baumketner, BioSim, Lviv 2019
Angle bending potential

*Harmonic approximation*

\[ v(\theta) = \frac{k}{2} (\theta - \theta_0)^2 \]

*Force constant. Source: normal mode analysis of QM, vibrational spectra*

*Reference bond angle. Source: crystal structures, QM calculations*

<table>
<thead>
<tr>
<th>Angle</th>
<th>(\theta_0)</th>
<th>(k) (kcal mol(^{-1})deg(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Csp(^3)–Csp(^3)–Csp(^3)</td>
<td>109.47</td>
<td>0.0099</td>
</tr>
<tr>
<td>Csp(^3)–Csp(^3)–H</td>
<td>109.47</td>
<td>0.0079</td>
</tr>
<tr>
<td>H–Csp(^3)–H</td>
<td>109.47</td>
<td>0.0070</td>
</tr>
<tr>
<td>Csp(^3)–Csp(^2)–Csp(^3)</td>
<td>117.2</td>
<td>0.0099</td>
</tr>
<tr>
<td>Csp(^2)–Csp(^2)=Csp(^2)</td>
<td>121.4</td>
<td>0.0121</td>
</tr>
<tr>
<td>Csp(^3)–Csp(^2)=O</td>
<td>122.5</td>
<td>0.0101</td>
</tr>
</tbody>
</table>

*Table 4.2 Force constants and reference angles for selected angles [Allinger 1977].

*Higher-order approximations approximation:*

\[ v(\theta) = \frac{k}{2} (\theta - \theta_0)^2 [1 - k'(\theta - \theta_0) - k''(\theta - \theta_0)^2 - k'''(\theta - \theta_0)^3 \ldots] \]
Dihedral angle potential

Several functional forms are in use

$$\nu(\Phi) = \sum_{n=0}^{N} \frac{V_n}{2} [1 + \cos(n\Phi - \gamma)]$$

Barrier height. Provides an idea on qualitative level about the barrier to rotation around particular bond.

Multiplicity. Determines how many minima the potential has. Depends on the chemistry of the central two atoms. For sp3 atoms, n=3, giving 3 minima. For sp2 atoms, n=2, leading to 2 minima.

The phase factor, Determines where the potential passes through a minimum.

Butane

Baumketner, BioSim, Lviv 2019
Improper dihedrals

Chemical compounds in which four non-consecutive atoms have to lie in-plane. This can’t be achieved with the help bond, angle and dihedral energy alone.

Improper dihedral energy:

\[ v(\theta) = \frac{k}{2} \theta^2; \]

Conformation favored by bond-angle terms

Experimentally observed conformation. Four atoms (1)(2)(3)(4) lie in the same plane.

Bonded cross-terms

Apply in Class II force-fields as opposed to Class I force-fields relying on fixed-charge model

Typically used in highly specialized force-fields such as MM2/MM3

Baumketner, BioSim, Lviv 2019
Parametrization

Partial charges

1) First principle approaches:

a) Partial charge is not an experimentally observable quantity. Can’t be determined directly

a) In QM calculations partial charges can’t be determined unambiguously. Many schemes exist. Mulliken charges are meant for intra-molecular interactions. They depend on the molecule chemistry, number of nuclei etc. Can’t be used to describe interactions between molecules.

2) Fitting:

a) Partial charges are fitted to reproduce certain thermodynamic properties of the studied system. See for instance OPLS/AA.

b) Partial charges are adjusted to reproduce electrostatic potential created around the molecule of interest. The latter are obtained in QM calculations. See AMBER and CHARMM.

\[
R = \sum_{i=1}^{N_{\text{points}}} w_i (\phi_i^0 - \phi_i^{\text{calc}})^2
\]

\[
\phi_i^{\text{calc}} = \sum_{j=1}^{N} \frac{q_j}{4\pi\varepsilon_0 r_{ij}} + \frac{Z - \sum_{j=1}^{N-1} q_j}{4\pi\varepsilon_0 r_{iN}}
\]
In principle there is no guarantee that such fit should be successful. Much depends on how fitting is performed. In all cases errors will be present.

Potential fitted to reproduce the long-range region. The fit will reproduce the dipole moment and will do a bad job for the intermediate distances because the partial-charge model and actual potential have different functional forms.

True potential measured for certain orientation of molecules A and B will tend to increase the dipole moment of this dipole

Electric field created by this dipole will tend to increase the dipole moment of this dipole

A low-energy configuration of two dipoles in a medium

Potential fitted to reproduce the entire curve. The long-range potential is not right but at a better agreement is seen at intermediate distances. The dipole moment produced by such fit will be larger than the anticipated. This is equivalent to the effect of polarization on the molecule. So the fit includes, in a way, the effect of polarization. Although not in a controlled manner.
c) Charges that are buried are statistically underdetermined. Difficult to obtain meaningful values.

The problem can be addressed by introducing weight factors for the charges. Example RESP charges of AMBER

d) The same set of charges cannot describe the potential equally well for different configurations of the target molecules.

Some force-fields consider multiple configurations so that the fitted charge produces the best agreement for the entire ensemble of structures. See AMBER.

e) The best performing fixed-charge model produce 5-15\% relative error in electrostatic potential with respect to QM results. For comparison, polarizable force-fields can achieve less than 1\% accuracy.
Van der Waals parameters

In almost all force-fields the vdW parameters to reproduce

a) Constants in molecular crystals
b) Heat of vaporization
c) Liquid densities

Bonded potentials

Bond-stretching, angle bending – normal modes, vibrational spectra. Most transferable part of force-field

Torsion potential.

QM calculations of potential energy as a function of the particular dihedral angle.

General scheme

Typically, parametrization of a force-field proceeds in three steps:

1) Bond-stretching and angle-bending parameters are set. Perhaps by borrowing values from AMBER.
2) Charges are fitted
3) Vdw parameters are fitted. The rule for 1-4 interactions is set.
4) Torsion potentials are fitted on QM simulations of dipeptides.

These are coupled. Don’t use torsion potentials obtained in one force-field in a different force-field!

$V_{LJ}(r_{ij}) = 4\epsilon_{ij} \left( \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right)$

Combination rules

$\sigma_{ij} = \sqrt{\sigma_i \sigma_j}$

$\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$

$\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j)$

Lorentz rule
### Force-fields in general use

**AMBER**

**CHARMM**

**OPLS**

<table>
<thead>
<tr>
<th>Force field</th>
<th>Potential type</th>
<th>Key references</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUFF</td>
<td>All Atom</td>
<td>Carlson, 2000</td>
</tr>
<tr>
<td>CEDAR</td>
<td>All Atom</td>
<td>Hermans et al., 1984; Hu et al., 2003</td>
</tr>
<tr>
<td>CVFF</td>
<td>All Atom</td>
<td>Kitson and Hagler, 1988</td>
</tr>
<tr>
<td>DISCOVER</td>
<td>All Atom</td>
<td>Maple et al., 1998</td>
</tr>
<tr>
<td>ECEPP/3</td>
<td>All Atom, Torsional</td>
<td>Némethy et al., 1993</td>
</tr>
<tr>
<td>ENCAD</td>
<td>All Atom</td>
<td>Daggett and Levitt, 1993; Levitt et al., 1995</td>
</tr>
<tr>
<td>GROMOS87</td>
<td>United Atom</td>
<td>van Gunsteren and Berendsen, 1987</td>
</tr>
<tr>
<td>GROMOS96</td>
<td>United Atom</td>
<td>Scott et al., 1999</td>
</tr>
<tr>
<td>MM2</td>
<td>All Atom</td>
<td>Lii et al., 1989</td>
</tr>
<tr>
<td>MM3</td>
<td>All Atom</td>
<td>Lii et al., 1991</td>
</tr>
<tr>
<td>MM4</td>
<td>All Atom</td>
<td>Langley and Allinger, 2002</td>
</tr>
<tr>
<td>MMFF</td>
<td>All Atom</td>
<td>Halgren, 1996a,b,c,d</td>
</tr>
<tr>
<td>NEMO</td>
<td>Polarizable</td>
<td>Hermida-Ramón et al., 2003</td>
</tr>
<tr>
<td>PROSA</td>
<td>Polarizable</td>
<td>Stern et al., 1999</td>
</tr>
<tr>
<td>SCHRODINGER</td>
<td>Polarizable</td>
<td>Kaminski et al., 2002</td>
</tr>
<tr>
<td>SDFF</td>
<td>Polarizable</td>
<td>Palmo et al., 2003</td>
</tr>
<tr>
<td>SIBFA</td>
<td>Polarizable</td>
<td>Gersh, 1997; Guo et al., 2000</td>
</tr>
<tr>
<td>SPASIBA</td>
<td>All Atom</td>
<td>Derreumaux and Vergoten, 1995</td>
</tr>
<tr>
<td>TRIPOS</td>
<td>All Atom</td>
<td>Clark et al., 1989</td>
</tr>
<tr>
<td>UCSD-WILSON</td>
<td>All Atom</td>
<td>Mackay et al., 1984</td>
</tr>
<tr>
<td>UFF</td>
<td>All Atom</td>
<td>Rappé et al., 1992</td>
</tr>
<tr>
<td>UPJOHN</td>
<td>All Atom</td>
<td>Oie et al., 1981</td>
</tr>
<tr>
<td>YETI</td>
<td>United, Torsional</td>
<td>Vedani, 1988</td>
</tr>
</tbody>
</table>
AMBER

History

ff84  united atom
ff94, ff96, ff99  all-atom
ff02  polarizable

Charges

Fritted to reproduce electrostatic potential of model peptides, ESP and then RESP charges.

In ff99 refitted using higher-order QM energies

vdW parameters

Combination rules: \( \epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j} \) geometric mean

\( \sigma_{ij} = \frac{1}{2} (\sigma_i + \sigma_j) \) arithmetic mean

Density and enthalpy of vaporization of CH4, C2H6, C3H8 and C4H10 liquids \( \rightarrow \) sp3 carbon and aliphatic hydrogen. sp2 carbon and aromatic H from liquid benzene. All others borrowed from OPLS/UA

1-4 interactions

Scaling factor of 0.5 in ff84 and 0.83 in all-atom force fields

Bonded

QM data and vibrational spectra

Energy function

\[
E_{\text{total}} = \sum_{\text{bonds}} K_i (r - r_{eq})^2 + \sum_{\text{angles}} K_d (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[ \frac{A_{ij}}{R_{ij}^6} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]
\]

RESP with the neutrality of AA enforced.

QM HF-6-31G* set. Multiple conformations.

The side-chain is varied. QM-optimized structures for GLY and ALA

Alanyl and glycyl dipeptide

Torsions

QM on GLY and ALA dipeptides as a function of rotation angles

Baumketner, BioSim, Lviv 2019
CHARMM

History
charmm19  united atom
charmm22, charmm27  all-atom
charmm36  polarizable

Charges

Supramolecular approach. QM energies are computed for AA-Water complexes. For neutral systems the energy is divided by 1.16. Charges are fitted to reproduce AA-Water interactions. TIP3P with non-zero vdW on H is used for water.

vdW parameters combination rules

Density and heat of vaporization.

\[
\begin{align*}
\epsilon_{ij} &= \sqrt{\epsilon_i \epsilon_j} \quad \text{geometric mean} \\
\sigma_{ij} &= \frac{1}{2}(\sigma_i + \sigma_j) \quad \text{arithmetic mean}
\end{align*}
\]

Bonded

QM data and vibrational spectra

1-4 interactions

No scaling. Scaling factor of 1.0

---

NMA water complexes

Backbone: QM yields geometry + force constants for bonds, angles. Water is important for geometry. Charges + vdW parameters

Alanyl dipeptide with water

+ simulations of proteins in gas and crystal phases

\[ R_c = \begin{cases} 
7.5-8.5\text{A W-W} \\
8.5-9.5\text{A S-W} 
\end{cases} \quad \text{switching} \]

Torsions

\[ \phi, \psi \]

QM on dipeptides. Matching of energy of different minima, C7, aR etc.
OPLS

**History**

- opls-UA: united atom
- opls-AA: all-atom

**Charges**

Fitted to reproduce interaction energy of model compounds with water estimated in QM simulations. Dipole moments are set about 15% larger than in gas phase to take polarization into account. TIP4P water is used in MM part, but TIP3P and SPC are also suitable. Concept of neutral groups is introduced, which reduces the number of requisite charges.

1-4 interactions

Scaling factor of 0.83 in OPLS-UA and 0.5 in OPLS-AA

vdW parameters

Experimental density and enthalpy of vaporization in liquid state are reproduced in MC simulations of model compounds that correspond to the peptide bond and side chains.

\[ \epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j} \]

\[ \sigma_{ij} = \sqrt{\sigma_i \sigma_j} \]

geometric mean

**Bonded**

- Borrowed from AMBER94 force-field

**Torsions**

Adopted from AMBER94 in OPLS-UA.

Fitted to QM energy functions computed for AA dipeptides for OPLS-AA/L

Parameters of CH3(C-O) are taken from hydrocarbons. After charges are fitted, the number of unknown parameters is equal the number of experimental measurements.

Neutral block

NMA is used as the model of peptide bond. Geometry from crystal structure. Charges from solute-water interactions.

[Diagram showing a neutral block with molecular structure and parameters]

Baumketner, BioSim, Lviv 2019
Convergence of force-fields

*United-atom force-fields show large variation in charges:*

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Amber 84</th>
<th>Amber 94/99</th>
<th>CHARMM19</th>
<th>CHARMM22/27</th>
<th>OPLS-UA</th>
<th>OPLS-AA</th>
<th>GROMOS96</th>
<th>BUFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>-0.463</td>
<td>-0.4157</td>
<td>-0.35</td>
<td>-0.47</td>
<td>-0.57</td>
<td>-0.50</td>
<td>-0.28</td>
<td>-0.749</td>
</tr>
<tr>
<td>HN</td>
<td>0.252</td>
<td>0.2719</td>
<td>0.25</td>
<td>0.31</td>
<td>0.37</td>
<td>0.30</td>
<td>0.28</td>
<td>0.328</td>
</tr>
<tr>
<td>CA</td>
<td>0.035</td>
<td>-0.0249</td>
<td>0.10</td>
<td>0.07</td>
<td>0.20</td>
<td>0.14</td>
<td>0.00</td>
<td>0.189</td>
</tr>
<tr>
<td>HA</td>
<td>0.048</td>
<td>0.0843</td>
<td>0.09</td>
<td>0.09</td>
<td>0.06</td>
<td>0.06</td>
<td>0.048</td>
<td>0.048</td>
</tr>
<tr>
<td>C</td>
<td>0.616</td>
<td>0.5973</td>
<td>0.55</td>
<td>0.51</td>
<td>0.50</td>
<td>0.50</td>
<td>0.38</td>
<td>0.828</td>
</tr>
<tr>
<td>O</td>
<td>-0.504</td>
<td>-0.5679</td>
<td>-0.55</td>
<td>-0.51</td>
<td>-0.50</td>
<td>-0.50</td>
<td>-0.38</td>
<td>-0.679</td>
</tr>
<tr>
<td>CB</td>
<td>0.018</td>
<td>0.2117</td>
<td>0.25</td>
<td>0.05</td>
<td>0.265</td>
<td>0.145</td>
<td>0.15</td>
<td>0.296</td>
</tr>
<tr>
<td>HB</td>
<td>0.119</td>
<td>0.0352</td>
<td>0.09</td>
<td>0.09</td>
<td>0.06</td>
<td>0.06</td>
<td>0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>OG</td>
<td>-0.55</td>
<td>-0.6546</td>
<td>-0.65</td>
<td>-0.66</td>
<td>-0.70</td>
<td>-0.683</td>
<td>-0.548</td>
<td>-0.764</td>
</tr>
<tr>
<td>HO</td>
<td>0.31</td>
<td>0.4275</td>
<td>0.40</td>
<td>0.43</td>
<td>0.435</td>
<td>0.418</td>
<td>0.398</td>
<td>0.491</td>
</tr>
</tbody>
</table>

*All-atom force-fields appear to converge*
Problems, ways to improve

**Intrinsic deficiencies**

The ansatz of partial charges placed at positions of nuclei not always is appropriate. It may not yield proper permanent dipoles. This can be fixed by adding more charges. Also atom-based multipole moments.

Example: nitrogen molecule

Has no dipole moment but has quadrupole moment.

Potential energy

\[ V \sim \frac{1}{R^5} \]

\[ V \sim \frac{1}{R^6} \]

the fixed-charge model has no moments so the first term is dispersion energy

\[ q = 0 \]

\[ q = 0 \]
Problems, ways to improve

Polarization

Molecules in condensed-phase environment acquire additional moments. This is a very strong effect that is seen even in molecular geometry, Polarization causes N-C distance in peptide bond to shorten while that of C-O bond to lengthen.

Polarization is taken into account implicitly by:

1) Errors in QM theory
2) Adding water molecules to the model compounds in QM calculations
3) Taking molecular geometry from crystal structures
4) Increasing the dipole moment of studied compounds by about 15%
5) Optimizing dihedral angles against NMR data in liquid state or proteins in solution

Still fixed-charge force-fields are only about 5-15% accurate. The chemical accuracy of 1kCal/mol is out of reach.

Baumketner, BioSim, Lviv 2019
Polarization in local environment

Better parametrization is unlikely to improve fixed-charge models by much. The fundamental problem is that they include polarization in an average sense.

Molecule in gas. Small dipole

Same molecule in polar medium, perhaps its own liquid. Increased dipole.

What happens during protein folding

The environment of target molecule changes. Could be transferred from polar medium where it’s polarized to non-polar medium where it’s dipole moment is small.

Polarization has to be included explicitly in order to make progress

Baumketner, BioSim, Lviv 2019
Polarizable force-fields

Three basic methods:

1) Fluctuating charge model
2) Drude oscillator
3) Induced dipole models

Polarization seems to get the ordering of different structures right

<table>
<thead>
<tr>
<th>Energy Model</th>
<th>cis-NMA</th>
<th>β-sheet</th>
<th>ΔE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPLS-AA</td>
<td>−11.5</td>
<td>−16.9</td>
<td>−5.4</td>
</tr>
<tr>
<td>CHARMM27</td>
<td>−11.6</td>
<td>−16.9</td>
<td>−5.3</td>
</tr>
<tr>
<td>AMBER ff94</td>
<td>−11.3</td>
<td>−14.8</td>
<td>−3.5</td>
</tr>
<tr>
<td>AMBER ff02</td>
<td>−13.5</td>
<td>−14.8</td>
<td>−1.3</td>
</tr>
<tr>
<td>AMOEBA</td>
<td>−18.5</td>
<td>−12.6</td>
<td>+5.9</td>
</tr>
<tr>
<td>SIBFA</td>
<td>−18.7</td>
<td>−17.1</td>
<td>+1.6</td>
</tr>
<tr>
<td>MP2/(CEP)4-31G+(2d)</td>
<td>−20.5</td>
<td>−17.5</td>
<td>+3.0</td>
</tr>
<tr>
<td>BP/DZVP (BSSE corrected)</td>
<td>−16.2</td>
<td>−8.4</td>
<td>+7.8</td>
</tr>
</tbody>
</table>

Baumketner, BioSim, Lviv 2019
Two different types of physical processes, deterministic and stochastic

Deterministic behavior:
The outcome of an experiment can be predicted exactly. Applies to many laws of physics: Newton’s laws, Maxwell equations etc.

Stochastic behavior:
The outcome of an experiment cannot be predicted exactly. This could be an intrinsic property of the physical object – quantum mechanics. Or, the lack of knowledge about the object= statistical mechanics. Fundamentally, all processes in nature are stochastic.

Example: Dice

1) The outcome of rolling dice “experiment” consists of 6 different realizations. It can be fully described by a discrete variable $g$ that takes on 6 values:

$$g_1, g_2, g_3, g_4, g_5, g_6$$

2) Although the laws of solid body mechanics are known, there is no way of predicting exactly the outcome of any experiment. Too many unknowns are involved: asymmetry in the mass distribution in the dice, temperature/pressure fluctuations, convection etc

For quantitative description of stochastic processes one needs the concept of distribution.

Baumketner, BioSim, Lviv 2019
Distributions:

Are easiest to introduce for discrete variables.

\[ \begin{align*}
&g_1 - \cdots - g_M \\
\text{possible realizations} & \text{of quantity } g
\end{align*} \]

If after \( M \) experiments value \( g_i \) is seen \( M p(g_i) \) times then

\[ p(g_i) = \text{probability distribution function} \]

Important properties:

\[ p(g_i) \geq 0, \text{ always positive or zero} \]

\[ \sum_{i=1}^{M} p(g_i) = 1 \]

has to be normalized, the sum is finite

Most generally:

\[ P(g_i) \rightarrow \frac{P(g_i)}{\sum_{i=1}^{N} P(g_i)} \]

For continuous variables sum are replaced with integrals:

\[ x \in [a, b] \quad \int_{a}^{b} P(x) \, dx \]

is the probability of seeing \( x \) in the interval \([a, b]\)

\[ P(x) \]

= probability distribution function

Normalization condition:

\[ \int_{a}^{b} P(x) \, dx = 1 \]

Averages:

By definition:

\[ \langle g \rangle = \frac{1}{M} \sum_{k=1}^{M} g_k = \frac{1}{M} (g_1 + g_2 + g_2 + g_2 + g_3 + \cdots) = \]

\[ k= \text{sum over different experiments} \]

\[ = \frac{1}{M} \sum_{i=1}^{N} M p(g_i) g_i = \sum_{i=1}^{N} p(g_i) g_i \]

\[ i= \text{sum over different realizations of variable } g \]

For any function of \( g \) and normalized dist.:

\[ \langle f \rangle = \sum_{i=1}^{N} f(g_i) \cdot P(g_i) \]

For any distr. funct.:

\[ \langle f \rangle = \frac{\sum_{i=1}^{N} f(g_i) P(g_i)}{\sum_{i=1}^{N} P(g_i)} \]

Baumketner, BioSim, Lviv 2019
Some basic definitions for distributions

\[ \mu = \int x \cdot \varphi(x) \, dx \quad \text{mean value} \]

\[ \sigma^2 = \int (x - \mu)^2 \cdot \varphi(x) \, dx \quad \text{standard deviation} \]

\[ \mu_n = \int (x - \mu)^n \cdot \varphi(x) \, dx \quad \text{n-order moment} \]

\( (n \text{-order moment may or may not exist}) \)

\[ \begin{bmatrix} \mu, \sigma \end{bmatrix} \] can be estimated from sampling

Geometrical interpretation

Say we have a sequence of measurements:

\[ x_1, \ldots, x_i, \ldots, x_n \]

Average over the sample will approximate the mean value

\[ \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \rightarrow \mu \]

Square deviation from the average will approximate the standard deviation:

\[ s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2} \rightarrow \sigma \]

\( (n-1 \text{ comes from Bessel correction for finite } n) \)

Baumketner, BioSim, Lviv 2019
Error estimate

Independent measurements

\[ SE = \frac{S}{\sqrt{n}} \]

\( S \) is sample error

\[ SD = \frac{S}{\sqrt{n}} \]

\( S \) is standard error

Correlated measurements

\[ SE = \frac{S}{\sqrt{n}} \sqrt{1 + (n-1) \rho^2 \frac{1}{1 - \rho}} \]

\( \rho = \langle \xi, \eta \rangle \) is correlation coefficient

\( \rho = 0 \) is for independent events

Examples:

\[ P(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x - \mu)^2}{2 \sigma^2}} \]

\( P(x) \) is normal distribution

Relative error:

\[ \text{Relative Error} = \frac{SE}{\mu} \rightarrow \frac{S}{\mu \sqrt{n}} \]

declines as inverse square root of the number of measurements. \( n \) must be large to achieve good accuracy

Accuracy of SE

Wide distributions require larger number of steps to converge

It’s safe to use SE to estimate the error in the measurement for \( n > 10 \)

Baumketner, BioSim, Lviv 2019
Transforming distribution functions

\( X \) - is a stochastic variable characterized by \( p(x) \)

\[ y = f(x) \quad Q: \text{what is distribution} \quad p(y) = ? \]

From the definition of the probability distribution:

\[ p(x) \, dx = p(y) \, dy \]

\[ p(y) = \frac{p(x(y))}{\left| \frac{dx}{dy}(x(y)) \right|} \]

(probability density transformation theorem)

This can be written in a more convenient form:

\[ \int \delta(y - f(x)) \, p(x) \, dx = \int \delta(y - f(x)) \, dx \]

Change of variables rule for delta function

\[ y' = f(x), \quad dx = \left| \frac{df}{dx}(x(y')) \right| \frac{dy'}{dx} \]

Baumketner, BioSim, Lviv 2019
Examples of probability transformations

1) Normal distribution with zero mean and $\sigma=1$

$$p_n(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

Q: what is the distribution of the linear transformation of variable $x$?

$$y = \mu + \sigma \cdot x$$

The shape of the distribution doesn’t change. But now it is characterized by new mean and variance

Normal distribution with zero mean and unit standard deviation can be used to generate Gaussians with arbitrary mean and variance through linear transformation of the variable!

Baumketner, BioSim, Lviv 2019
2) Uniform distribution:

\[ P_i(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1 \\ 0 & \text{otherwise} \end{cases} \]

\[ P(y) = \frac{d}{dy} P_i(x) = \frac{d}{dx} \left[ \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt \right] \quad \text{since} \quad P_i = \text{const} \]

3) Exponential distribution:

\[ P(y) = e^{-y} \quad \text{arbitrary constant} \]

\[ \frac{d}{dy} x = e^{-y} \rightarrow x = e^{-y} + C \rightarrow y = -\log(x - C) \]

Set \( C = 0 \) to get \( x(0) = 1, x(+\infty) = 0 \) \( \rightarrow y = -\log(x) \)

4) Gaussian:

\[ P(y) \sim e^{-y^2} \]

\[ \frac{d}{dy} x \sim e^{-y^2} \rightarrow x(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-t^2} dt = \text{Erf}(y), \quad y \geq 0 \]

\[ x(0) = 0, \quad x(+\infty) = 1 \]

\[ y = \text{Erf}^{-1}(x) \]

will generate normally distributed positive numbers.

inversion of the error function can be costly numerically

For negative numbers, use the property:

\[ P(y) = P(-y) \]

Two random numbers \( x \in [0,1] \)

\[ -y \]

Baumketner, BioSim, Lviv 2019
Multiple events can be observed simultaneously. For two variables $x$ and $y$ one introduces:

Multivariate distributions

$P(y), y \in [y_1, y_2]$ target distribution. Define: $F(y) = \int_{y_1}^{y} P(x')dx'$ cumulative dist.

$z = F^{-1}(x)$ what’s dist. for this variable $P_n(Z)$?

 indeed, the desired distribution

$P_n(z)dz = 1dx \rightarrow P_n(z) = \frac{d}{dz}x(z), x(z) = F(z), \frac{dx}{dz} = \frac{dF(z)}{dz} = P(z), P_n(z) = P(z)$

5) For distributions of arbitrary shapes: cumulative distribution transformation theorem

$$P(y), y \in [y_1, y_2]$$ target distribution. Define: $F(y) = \int_{y_1}^{y} P(x')dx'$ cumulative dist.

$z = F^{-1}(x)$ what’s dist. for this variable $P_n(Z)$?

 indeed, the desired distribution

$P_n(z)dz = 1dx \rightarrow P_n(z) = \frac{d}{dz}x(z), x(z) = F(z), \frac{dx}{dz} = \frac{dF(z)}{dz} = P(z), P_n(z) = P(z)$

Multivariate distributions

Multiple events can be observed simultaneously. For two variables $x$ and $y$ one introduces:

$\mathbf{r} = (x, y)$ to denote a joint event

Joint distribution function $P(x, y)$ is introduced so that

is the probability of seeing $x \in [x - \Delta x, x + \Delta x]$ and $y \in [y, y + \Delta y]$.

Distributions for individual variables:

$P(x) = \int P(x, y)dy, P(y) = \int P(x, y)dx$ Normalization condition:

$\int P(x, y)dx dy = 1$

Baumketner, BioSim, Lviv 2019
Conditional probabilities

Assume 2-D for simplicity

Event $A: [x, x + 8 \times]

Event $B: [y, y + 6 \times]

$N = \# \text{ of 2D events total}

$N_A = \# \text{ of } A

$N_{AB} = \# \text{ of both } A \text{ and } B

$N_B = \# \text{ of } B

\text{ of both occurring at the same time}

Define conditional probability:

$P(A|B) = \frac{\text{prob. of } A \text{ once } B \text{ occurred}}{P(B)} = \frac{N_{AB}}{N_B} = \frac{N - N_{AB}}{N} \cdot \frac{P(A|B)}{P(B)}$

$P(B|A) = \frac{N_{AB}}{N_A} = \frac{P(A,B)}{P(A)}$

$P(A|B) = P(B|A) \frac{P(A)}{P(B)}$

Bayes’ theorem

Baumketner, BioSim, Lviv 2019
Concept of independent events

Assume that event A is not conditioned upon event B. The conditional probability then is equal to the simple probability of event A:

\[ P(A|B) = P(A) \]

The joint probability then becomes:

\[ P(A, B) = P(A) \cdot P(B) \]

If this condition is met the events are known as independent. The distribution can be used to judge the degree of independence or correlation.

Quantitatively this can be measured by correlation coefficient:

\[ \rho = \frac{\langle \Delta x \cdot \Delta y \rangle}{\sqrt{\Delta x^2 \cdot \Delta y^2}} \]

where \( x \) and \( y \) are two stochastic variables and

\[ \Delta x = x - \langle x \rangle, \Delta y = y - \langle y \rangle \]

Case 1: \( x \) and \( y \) are independent:

\[ \langle \Delta x \cdot \Delta y \rangle = \int \int (x - \langle x \rangle) \cdot (y - \langle y \rangle) \cdot p(x, y) \, dx \, dy = \]

\[ \int x \cdot y \cdot p(x) \cdot p(y) \, dx \, dy = \int x \cdot \int y \cdot p(x) \cdot p(y) \, dx \, dy = \]

\[ = \langle x \rangle \langle y \rangle - \langle x \rangle \langle y \rangle = 0 \]

Baumketner, BioSim, Lviv 2019
Case 2: $x$ and $y$ are fully correlated $x= \alpha y$, $\alpha > 0$

$$\langle \Delta x \cdot \Delta y \rangle = \frac{\partial \langle \Delta y^2 \rangle}{\partial \langle \Delta x^2 \rangle} = \frac{\partial}{\partial \langle \Delta x^2 \rangle} \sqrt{\frac{\langle \Delta y^2 \rangle}{\langle \Delta x^2 \rangle}} = 1$$

Case 3: $x$ and $y$ are fully anti-correlated $x=-\alpha y$, $\alpha > 0$

$$\langle \Delta x \cdot \Delta y \rangle = -\frac{\partial \langle \Delta y^2 \rangle}{\partial \langle \Delta x^2 \rangle} = -\frac{\partial}{\partial \langle \Delta x^2 \rangle} \sqrt{\frac{\langle \Delta y^2 \rangle}{\langle \Delta x^2 \rangle}} = -1$$

What multivariate distributions can be used for

Generation of normal distributions. Let $x$ and $y$ be random variables uniformly distributed between 0 and 1. Introduce new variables:

$$x' = \sqrt{-2 \log(x)} \cos(\sqrt{2 \log(1-x^2)})$$
$$y' = \sqrt{-2 \log(x)} \sin(\sqrt{2 \log(1-x^2)})$$

$$x'^2 + y'^2 = -2 \log(1-x) \Rightarrow x = e^{-\frac{1}{2}(x'^2 + y'^2)}$$

Baumketner, BioSim, Lviv 2019
Transformation of the joint distribution function:

\[
P(x) = P(x) \quad P(y) = P(y)
\]

\[
P(x', y') \, dx' \, dy' = P(x) \, P(y) \, dx \, dy
\]

\[
P(x', y') \left\lvert \frac{\partial (x', y')}{\partial (x, y)} \right\rvert \, dx \, dy = P(x) \, P(y)
\]

\[
\begin{bmatrix}
\frac{\partial x'}{\partial x} &= -\frac{1}{\sqrt{2 \log(x)}} \times \cos(2\pi y) \\
\frac{\partial y'}{\partial x} &= -\frac{1}{\sqrt{2 \log(x)}} \times \sin(2\pi y)
\end{bmatrix}
\]

\[
\begin{bmatrix}
\frac{\partial x'}{\partial y} &= \frac{1}{2 \log(x)} \times \cos(2\pi y) \\
\frac{\partial y'}{\partial y} &= \frac{1}{2 \log(x)} \times \sin(2\pi y)
\end{bmatrix}
\]

Jacobian of the transformation

\[
P(x', y') = \frac{1 \times 1}{e^{\pi}}
\]

\[
P(x', y') = \frac{1}{2\pi} \, e^{-\frac{1}{2}(x'^2 + y'^2)} = \frac{1}{\sqrt{\pi a}} \, e^{-\frac{1}{2} x'^2} \cdot \frac{1}{\sqrt{\pi a}} \, e^{-\frac{1}{2} y'^2} = P_n(x') \cdot P_n(y')
\]

two uniformly distributed numbers \( x \) and \( y \)

two normally distributed numbers \( x' \) and \( y' \)

Baumketner, BioSim, Lviv 2019
Statistical mechanics = theory to extract **macroscopic** properties from **microscopic** variables

Microscopic description:

2N-D phase space:

Coordinates + momenta fully define the state of a system with N degrees of freedom

Measuring property $A(\Gamma)$ always yields time average (take pressure in tires for example):

$$A_{\text{obs}} = \langle A \rangle_{\text{time}} = \langle A(r(t)) \rangle_{\text{time}} = \lim_{t_{\text{obs}} \to \infty} \frac{1}{t_{\text{obs}}} \int_{t_{\text{obs}}-1}^{t_{\text{obs}}} A(r(t)) \, dt$$

*real or virtual experiment*

Observation time $t_{\text{obs}}$ is always finite. Furthermore, in practice it is always discretized so that the integral can be carried out.

$$t_{\text{obs}} = \delta t \cdot t_{\text{obs}} - t_{\text{obs}} = \#$$

must be large enough to eliminate dependence on the initial conditions

Baumketner, BioSim, Lviv 2019
The expression for the observable now reads:

\[
\langle x \rangle \quad A_{\text{obs}} = \langle A \rangle_{\text{time}} = \frac{1}{\tau_{\text{obs}}} \sum_{\tau=1}^{\tau_{\text{obs}}} A(\Gamma(\tau))
\]

\(\tau\) is now just a blind index that enumerates all measurements.

Recall how we computed averages for random variables:

\[
< g > = \frac{1}{M} \sum_{k} g_k = \frac{1}{M} (g_1 + g_2 + g_2 + g_2 + g_3 + \ldots)
\]

it’s the same formula.

On one hand we have **time evolution** but on the other – different realizations of some random variables that can be described by certain **distribution**. Both descriptions lead to the same average. The one based on distributions is the subject of statistical mechanics.

**The concept of ensembles**

**Ensemble = Multiple copies of the system at time t=0**

*Time evolution of one system*

Points in the phase space are distributed according to certain function

\[
P_{\text{ens}}(r) = \frac{1}{\sum P_{\text{ens}}(r)}
\]

Baumketner, BioSim, Lviv 2019
At first glance the connection between time average and ensemble average appears to be straightforward. But there are important subtleties.

Let us consider $N$ members of the ensemble, each corresponding to a gamma point $\Gamma_i$. Consider that in general the distribution function may have explicit dependence on time. See what happens in a certain volume $\delta \Gamma$ when the time changes from $t$ to $t+dt$.

\[ N_{\text{points}} = NP_{\text{ens}}(\Gamma, t)\delta \Gamma \]

1) Some points will enter the volume
2) Some will leave it
3) Some will remain within the volume
4) None will be created or destroyed

The balance equation for the number of points:

\[ N \delta \Gamma \frac{\partial}{\partial t} P_{\text{ens}}(\Gamma, t) = N(\text{enter}) - N(\text{leave}) + F(\Gamma, t) \]

\[ = 0, \text{ not present} \]

Baumketner, BioSim, Lviv 2019
Let’s forget about momenta in $\Gamma$ for the moment and focus on Cartesian coordinates only.

$v_s$ component of the velocity along $\vec{n}$

$\rho = NP_{ens}$ density of points

all those contained in this parallelepiped

$$N(\text{enter}) = \frac{\rho}{dS} \frac{dl}{dt} = \frac{\rho}{dS} v_s \frac{dt}{dt} = J \vec{n} dS$$

$N(\text{leave}) = \int J \vec{n}' dS$

surface integral of flux

divergence theorem

$N(\text{enter}) - N(\text{leave}) = -\int \vec{J} \cdot d\vec{S} = -\int (\nabla \vec{J}) dV \approx -N(v \nabla P_{ens}) \delta \Gamma$

$J = \rho \vec{v}$ flux = the number of points passing through the boundary per unit surface area per unit time

Baumketner, BioSim, Lviv 2019
Putting the estimate into the balance equation:

\[ N\delta\Gamma \frac{\partial}{\partial t} P_{ens}(\Gamma, t) = -N(\ddot{\nu}\vec{\nu}P_{ens})\delta\Gamma \]

\[ \frac{\partial}{\partial t} P_{ens}(\Gamma, t) + (\ddot{\nu}\vec{\nu}P_{ens}) = 0 \]

\[ \frac{\partial}{\partial t} P_{ens}(\Gamma, t) + \dot{\vec{r}} \frac{\partial}{\partial \vec{r}} P_{ens} = 0 \]

If we add momenta back to the equation we will get (by analogy):

\[ \frac{\partial}{\partial t} P_{ens}(\Gamma, t) + \dot{\vec{r}} \frac{\partial}{\partial \vec{r}} P_{ens} + \dot{\vec{p}} \frac{\partial}{\partial \vec{p}} P_{ens} = 0 \]

\[ \left( \frac{\partial}{\partial t} + \dot{\vec{r}} \frac{\partial}{\partial \vec{r}} + \dot{\vec{p}} \frac{\partial}{\partial \vec{p}} \right) P_{ens}(\Gamma, t) = 0 \]

The probability distribution is constant along any trajectory

\[ \frac{dP_{ens}(\Gamma, t)}{dt} = 0 \]

\[ \frac{d\rho_{ens}(\Gamma, t)}{dt} = 0. \]

Baumketner, BioSim, Lviv 2019
In equilibrium the probability distribution can not depend on time. Because otherwise the averages would depend on time as well. That contradicts the definition of “equilibrium”. Therefore:

\[ P_{ens}(\Gamma, t) = P_{ens}(\Gamma) \quad \rightarrow \quad \frac{\partial}{\partial t} P_{ens}(\Gamma, t) = 0 \]

\[ N(\text{enter}) - N(\text{leave}) = 0 \]

The system is evolving in is such a way that

\[ P_{ens}(\Gamma) = \text{const} \quad \text{at each point} \]

As one point exits certain cell in the phase space, another point immediately enters it. As a result, all points are moving in concert in what resembles a Conga line.

The line snakes around the phase space as time passes by. How this happens has important consequences.

**Option 1.** The snakes passes through all points available in the phase space. The entire phase space is accessible. **Ergodic** behavior.

**Option 2.** There are regions in the phase space from which the snake cannot break out. It moves in a circular manner. **Non-ergodic** behavior.

Baumketner, BioSim, Lviv 2019
In ergodic systems, all phase points are visited multiple times during a simulation. Only in this case is the time average equal to ensemble average (with particular distribution function)

\[
\hat{A}_{\text{obs}} = \langle A \rangle_{\text{time}} = \langle A \rangle_{\text{ens}} = \frac{\sum_{r} A(r) \cdot p_{\text{ens}}(r)}{\sum_{r} p_{\text{ens}}(r)}
\]

1) Determining whether or not a system is ergodic is not a trivial task. Rigorous proof exists only for a few model systems, such as coupled harmonic oscillators.

2) There are different reasons for non-ergodicity
   a) Frustration – multiplicity of potential energy minima of the same depth prevents their thorough exploration
   b) Low temperature. Creates very high barriers in the free energy landscape that can be overcome. The system becomes locked up in certain parts of the phase space. Glass transition is one example.

3) Certain models are known to be non-ergodic. For instance, certain lattice models of proteins

4) It’s easy to design a non-ergodic system. All it takes is to arrange a circular movement in the phase space. Can be achieved through specific Monte Carlo moves.

Baumketner, BioSim, Lviv 2019
Stat mech. vs. thermodynamics

Normalization constant of the probability distribution is not needed to compute averages. Consider
\[ \rho_{\text{ens}}(\Gamma) \] and another distribution proportional to it
\[ \rho'_{\text{ens}}(\Gamma) = \alpha \rho_{\text{ens}}(\Gamma) \]

According to the definition:
\[
< A >' = \sum_{\Gamma} A(\Gamma) \rho'_{\text{ens}}(\Gamma) / \sum_{\Gamma} \rho'_{\text{ens}}(\Gamma) = \sum_{\Gamma} A(\Gamma) \alpha \rho_{\text{ens}}(\Gamma) / \sum_{\Gamma} \alpha \rho_{\text{ens}}(\Gamma) =
\]
\[
= \sum_{\Gamma} A(\Gamma) \rho_{\text{ens}}(\Gamma) / \sum_{\Gamma} \rho_{\text{ens}}(\Gamma) = < A >
\]

But this quantity is central to establishing link between \textit{microscopic} description in terms of coordinates/momenta and \textit{macroscopic} description in terms of thermodynamic functions

\[ Q_{\text{ens}} = \sum_{\Gamma} \rho_{\text{ens}}(\Gamma) \quad \text{\textit{partition function=}} \]
the sum of \( \rho_{\text{ens}}(\Gamma) \) over all possible states

\[ < A >_{\text{ens}} = \sum_{\Gamma} A(\Gamma) \rho_{\text{ens}}(\Gamma) / Q_{\text{ens}} \]

\[ \Psi_{\text{ens}} = -\log(Q_{\text{ens}}) \]

\text{The function that reaches minimum in equilibrium.}

Baumketner, BioSim, Lviv 2019
Microcanonical ensemble (NVE)

**Distribution function:**

\[ \rho_{\text{ens}} \sim \delta(E - H(\Gamma)) \]

**Hamiltonian:**

\[ H(\Gamma) = E_k + E_p = \sum_i \frac{\vec{p}_i^2}{2m_i} + U(\vec{q}_1, \ldots, \vec{q}_N) \]

**Partition function:**

\[ Q_{NVE} = \sum_{\Gamma} \delta(E - H(\Gamma)) \]

The volume of the hypersurface that corresponds to energy \( E \)

**Thermodynamic potential:**

\[ \Psi_{NVE} = -k\log(Q_{NVE}) = -S(N, V, E) \]

**Physical equivalent:** an isolated system

Boltzmann's constant

Boltzmann's constant first introduced on the grounds of dimensionalities. Then recognized as the Planck constant when QM came about.

Boltzmann's constant

Boltzmann's constant

Boltzmann's constant

Baumketner, BioSim, Lviv 2019
**Canonical ensemble (NVT)**

**Distribution function:**

\[ \rho_{NVT}(\Gamma) \sim e^{-\beta H(\Gamma)}, \beta = \frac{1}{kT} \]

external parameter that is associated with temperature

**Partition function:**

\[ Q_{NVT} = \sum_{\Gamma} e^{-\beta H(\Gamma)} = \sum_{E} n(E)e^{-\beta E} \]

separation of the partition function

\[ Q_{NVT} = \frac{1}{h^{3N}N!} \int d\vec{p}d\vec{q}e^{-\beta H(\Gamma)} = \frac{1}{h^{3N}N!} \int d\vec{p}e^{-\beta E_K} \int d\vec{q}e^{-\beta U(\vec{q})} = Q_{NVT}^{id} \times Q_{NVT}^{ex} \]

\[ Q_{NVT}^{id} = \frac{V^N}{N! \lambda^{3N}}, \lambda = \sqrt{\frac{\hbar^2}{2\pi mkT}} \]

**Thermodynamic potential:**

\[ \Psi_{NVT} = F(N, V, T) = -kT\log(Q_{NVT}) = F_{id}(N, V, T) + F_{ex}(N, V, T) \]

Helmholtz free energy ideal gas part excess part (due to interactions)

Baumketner, BioSim, Lviv 2019
Isothermal-isobaric ensemble (NPT)

Distribution function:

\[ \rho_{NPT}(\Gamma) \sim e^{-\beta(H(\Gamma,V)+PV)} \]

Partition function:

\[ Q_{NPT} = \sum_V \sum_{\Gamma} e^{-\beta(H(\Gamma,V)+PV)} = \sum_V Q_{NVT} e^{-\beta PV} \]

Thermodynamic potential:

\[ \Psi_{NPT} = G(N,P,T) = -kT \log(Q_{NPT}) = G_{id}(N,P,T) + G_{ex}(N,P,T) \]

Physical equivalent - system under a piston

Baumketner, BioSim, Lviv 2019
Grand canonical ensemble ($\mu VT$)

**Distribution function:**

$$\rho_{\mu VT}(\Gamma) \sim e^{-\beta(H(\Gamma,V) - \mu N)}$$

**Partition function:**

$$Q_{\mu VT} = \sum_N \sum_{\Gamma} e^{-\beta(H(\Gamma,V) - \mu N)} = \sum_N Q_{NVT} e^{\beta \mu N}$$

no separation but the kinetic part can be integrated explicitly

$$Q_{\mu VT} = \sum_N e^{\beta \mu N} Q_{NVT}^{id} Q_{NVT}^{ex} = \sum_N e^{\beta \mu N} \frac{1}{N! \lambda^{3N}} Z(N,V,T)$$

**Thermodynamic potential:**

$$\Psi_{\mu VT} = \Phi_G(\mu, V, T) = -kT \log(Q_{\mu VT}) = F - \mu N = -PV$$

Physical equivalent=

system exchanging heat and particles with the environment

energy and number of particles are allowed to change

key property to be evaluated

Baumketner, BioSim, Lviv 2019
Summary on ensembles/thermodynamic functions

Microcanonical (primary)

\[
\frac{S(N,V,E)}{k} = \log(Q_{NVE})
\]

entropy

Canonical

\[
F(NTV) = -kT\log(Q(NTV)) = E - TS
\]

Helmholz free energy

Isobaric-isothermic

\[
G(NPT) = -kT\log(Q(NPT)) = F + PV
\]

Gibbs free energy

Grand canonical

\[
\Phi(\mu TV) = -kT\log(Q(\mu TV)) = F - \mu N = -PV
\]

Grand canonical or Hill free energy

Link to thermodynamics

Fundamental law:

\[
TdS = dU + pdV - \mu dN
\]

\[
dF = -pdV - sdT + \mu dN \rightarrow \frac{\partial F}{\partial N} \bigg|_{TV} = \mu
\]

\[
dG = Vdp - sdT + \mu dN \rightarrow \frac{\partial G}{\partial N} \bigg|_{TP} = \mu
\]

\[
\Phi = F - \mu N = G - PV - \mu N = -PV
\]

\[\mu_{TP}\] is not an ensemble. it contains only intensive variables some of which are related

Baumketner, BioSim, Lviv 2019
Monte Carlo – a class of algorithms that rely on random sampling to obtain numerical results

a) Developed by Ulam, von Nuemann and Metropolis in the 40's to model diffusion of neutrons in fissile material.

b) The term is the codename coined after Monte Carlo casino where Ulam's uncle used to borrow money to gamble.

c) Many flavors exist designed to address specific problems

d) Use in math: applied statistics – the inference problem, integration, optimization, inverse problems etc.

e) Under the name of Markov Chain Monte Carlo (MCMC), used widely in physics, chemistry, biology, finance, quantitative linguistics etc

The main idea: use **stochastic** methods (random sampling) to solve **deterministic equations**.

How is that possible? Why is that needed?

**Example: Buffon’s needle experiment**

A needle of length \( l \) is thrown onto a striped field. What is the probability that the needle will cross the middle line?

Two outcomes of the experiment:

a) Needle crosses the middle line

b) Needle doesn’t cross the line

Baumketner, BioSim, Lviv 2019
The state of the needle is defined by two variables:

a) $x$ – the center of the needle

b) $\varphi$ – the angle it makes with the vertical axis

$x$ and $\varphi$ are random variables, independently distributed.

$P(x), P(\varphi), P(x, \varphi)$ are all uniform distributions – no preferential $x$ or $\varphi$.

\[
P(x, \varphi) = \frac{1}{\pi} \cdot \frac{1}{2t}, \quad \int P(x, \varphi)dx\,d\varphi = 1
\]

The needle crosses the line when $x$ coordinate of points 1 and 2 satisfies the following conditions:

\[
x + \frac{l}{2} \cdot \sin(\varphi) \geq t
\]

\[
x - \frac{l}{2} \cdot \sin(\varphi) \leq t
\]

\[
x = t - \frac{l}{2} \sin(\varphi)
\]

\[
x = t + \frac{l}{2} \sin(\varphi)
\]

$\Omega$ is the surface area

\[
P_{1/2} = \int_0^\pi dx\,d\varphi P(x, \varphi) = \frac{1}{\pi 2t} S_c(\Omega) = \frac{1}{\pi 2t} \int_0^\pi \sin(x) dx = \frac{2l}{\pi 2t} = \frac{l}{\pi t}
\]

Baumketner, BioSim, Lviv 2019
Full probability:

\[ P_{\text{crossing}} = 2P_{1/2} = \frac{2l}{\pi t} \]

\[ \pi = \frac{2l}{tP_{\text{crossing}}} \]

Deterministic result = exact number

Lazzarini's experiment: 3408 trials

\[ \pi = \frac{355}{113} = 3.1415929 \quad 10^{-7} \text{ accuracy} \]

**Example: hit and miss integration**

Throw particles (generate pairs of random numbers) on this square and count how many fall within the circle \( r = 1 \). The goal is to compute \( \pi \).

If particles are distributed uniformly then:

\[ P_A - \text{probability to hit circle} \]
\[ P_B - \text{probability to square} \]

\[ \begin{align*}
P_A &\approx S_A = \frac{\pi}{4} \\
\frac{P_A}{P_B} &\approx \frac{S_A}{S_B} = 1 \\
\frac{S_A}{S_B} &\approx \frac{\pi}{4} \\
S_A &\approx \frac{\pi 2^2}{4} = \frac{\pi}{2} \\
S_B &\approx \frac{\pi}{4} 
\end{align*} \]

Baumketner, BioSim, Lviv 2019
If the number of particles hitting the circle is \( n_A \) and those hitting the square is \( n_B \):

\[
\pi = \frac{4n_A}{n_B}
\]

**Accuracy estimates:**

\[ 10^{-4} \text{ accuracy} \]

**Hit & miss:** 3.14173 after 10^7 shots

\[ 10^{-3} \text{ accuracy} \]

**Needles:** 3.140472 after 10^7 shots

**Conclusion:** Lazzarini was able to compute \( \pi \) with 10^{-7} accuracy after \( \sim 10^4 \) shots so he:

a) cheated

b) had a VERY lucky afternoon with numbers!
Hit & miss experiment is an example of the “sample mean” integration method.

Goal: compute \( F = \int_{x_1}^{x_2} f(x)dx \)

Rewrite \( F = \int_{x_1}^{x_2} f(x)dx = \int_{x_1}^{x_2} \frac{f(x)}{\rho(x)} \cdot \rho(x)dx = \begin{cases} \rho(x) \text{-- arbitrary distribution} \\ f(x) \text{ function} \end{cases} \)

If \( N \) trials are performed to sample random variable \( \xi \), distributed according to \( \rho(\xi) \), then

\[
F = < \frac{f(\xi)}{\rho(\xi)} >_{\text{trials}} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(\xi_i)}{\rho(\xi_i)}
\]

the average is taken over different trials

Example:

uniform distribution \( \rho(x) = \begin{cases} \frac{1}{x_2-x_1} & x_2 \leq x \leq x_1 \\ 0 & \text{otherwise} \end{cases} \)

\[
F = \frac{x_2 - x_1}{N} \sum_{i=1}^{N} f(\xi_i)
\]

Baumketner, BioSim, Lviv 2019
Compute \( \pi \) number by the sample mean method:

\[
f(x) = \sqrt{1 - x^2}, x \in [0, 1], \int_0^1 \sqrt{1 - x^2} \, dx = \frac{\pi}{4}
\]

\[
\pi \approx 3.1416 \text{ for } N = 10^7
\]

\( 10^{-4} \) accuracy

Simpson’s integration rule:

\[
\pi = 3.141593 \text{ after } 10^4 \text{ steps!}
\]

\( 10^{-6} \) accuracy

The cost of Simpson’s rule (or similar quadrature method based on interpolation) is prohibitive for multi-dimensional integrals!

\[
D = n \rightarrow \int f(\vec{x}) \, dx_1 \ldots dx_n \quad n_s = \text{number of sample points per dimension} \quad n_s^n = \text{number of function evaluations}
\]

\[
\begin{cases}
  n = 300 \\
  n_s = 10
\end{cases} \quad \text{for an ensemble of 100 particles}
\]

\( 10^{300} \) function evaluations. That’s an astronomical number that no computer can handle!

Baumketner, BioSim, Lviv 2019
Two steps involved in the integration using uniform $\rho$:

(in math this method is known as random Monte Carlo integration)

1) Pick a point in the configuration space $(\vec{r}_1 ... \vec{r}_N)$ by generating $3N$ random numbers uniformly

2) Compute the potential energy $U(\vec{r}_1 ... \vec{r}_N)$ and the integrand for select $\beta$

Repeat these steps $N_s$ times and compute the configuration integral as:

$$Z(N, V, T) = \frac{1}{N_s} \sum_{i=1}^{N_s} e^{-\beta U(\vec{r}_1(i) ... \vec{r}_N(i))}$$

Problems with the uniform $\rho$:

1) $N_s$ has to be VERY large. For most systems of practical interest in physics convergence is not attainable.

2) Ensemble averages $< A > = \frac{\sum A_i e^{-\beta U_i}}{\sum e^{-\beta U_i}}$ are even less accurate and in most cases meaningless

the integral has to converge with $N_s$

for many evaluation points $e^{-\beta U} \sim 0$

Baumketner, BioSim, Lviv 2019
To reduce the number of zeroes choose a distribution function $\rho(\Gamma)$ that has a strong overlap with the integrand.

For the canonical ensemble:

$$< A >_{NVT} = \frac{\int A(\Gamma)\rho_{NVT}(\Gamma)d\Gamma}{\int \rho_{NVT}(\Gamma)d\Gamma} = \frac{< A \rho_{NVT}/\rho >_{\text{trials}}}{< \rho_{NVT}/\rho >_{\text{trials}}}$$

rho - sampling distribution

Specific choice $\rho(\Gamma) = \rho_{NVT}(\Gamma)$ is known as importance sampling:

$$< A >_{NVT} = < A >_{\text{trials}}$$

How does one generate a sequence of configurations $\Gamma_1 \ldots \Gamma_N$ that satisfy the given distribution $\rho(\Gamma)$?

Answer: design a Markov chain of states whose limiting distribution is $\rho(\Gamma)$!

(the corresponding method is known as Markov Chain Monte Carlo (MCMC))

Baumketner, BioSim, Lviv 2019
What does it mean “Markov chain”? Stochastic process in which:

a) there is a finite (countable) set of configurations in the phase space \{ \Gamma_1 \ldots \Gamma_N \}.

b) transition from state \( i \) to state \( j \) does not depend on the history prior to state \( i \). There is no memory. Correlation only between neighboring sites.

Transition probabilities among states make a matrix \( \pi \):

\[
\begin{align*}
\pi_{mn} & \quad \text{probability of making a transition from state } m \text{ to state } n. \\
\sum_{n \neq m} \pi_{mn} & \quad \text{probability of transitioning to any state} \\
\pi_{mm} & \quad \text{probability of remaining in state } m. \\
\sum_{n} \pi_{mn} = 1 & \quad \text{consequence of the phase space finiteness}
\end{align*}
\]

Baumketner, BioSim, Lviv 2019
How are $\pi_{mn}$ and $\rho(\Gamma)$ related?

**Example:** prediction of computer’s up time

Computer can be either up or down. So the phase space consists of two states $|\uparrow> \text{ and } |\downarrow>

Computer has:
- $60\%$ chance of being up today if it was up the day before
- $70\%$ chance of being down today if it was down the day before

Transition matrix from day 1 to day 2

$$\hat{\pi} = \begin{pmatrix} \pi(\uparrow\uparrow) & \pi(\uparrow\downarrow) \\ \pi(\downarrow\uparrow) & \pi(\downarrow\downarrow) \end{pmatrix} = \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix}$$

*Day 1:* computer is up, $\rho(\uparrow) = 1, \rho(\downarrow) = 0$

$$\rho(\downarrow) = 0.6 \times 0.6 + 0.4 \times 0.3 = 0.48$$

$$\rho(\uparrow) = 0.4 \times 0.7 + 0.6 \times 0.4 = 0.52$$

*Day 2:* $\rho(\uparrow) = 0.6, \rho(\downarrow) = 0.4$

$$\hat{\rho}(2) = (\rho(\uparrow)\rho(\downarrow)) = (0.6 \ 0.4) = (1 \ 0)\begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix}$$

*Day 3:

$$\hat{\rho}(3) = \hat{\rho}(2) \cdot \hat{\pi} = \hat{\rho}(1) \cdot \hat{\pi} \cdot \hat{\pi} = \hat{\rho}(1) \cdot \hat{\pi}^2$$

*Day N:

$$\hat{\rho}(N) = \hat{\rho}(1) \cdot \hat{\pi}^{N-1}$$

Probability on day $N$ depends on probability on day 1

Baumketner, BioSim, Lviv 2019
Is there a limiting probability for large $N$ in which the dependence on the initial state disappears?

$$\lim_{N \to \infty} \hat{\rho}(N) = \hat{\rho} \quad \text{does this limit exist?}$$

If it does, then the probability becomes independent of time so it is equal for day $N+1$ and day $N$:

$$\hat{\rho}(N + 1) = \hat{\rho}(N) = \hat{\rho}$$

However,

$$\hat{\rho}(N + 1) = \hat{\rho}(N) \cdot \hat{\pi} \quad \rightarrow \quad \bar{\rho} = \hat{\rho} \cdot \hat{\pi} \quad \text{an equation for the limiting distribution}$$

Eigenvector eigenvector eigenvalue

Eigenvector problem

$$\bar{\rho} = \hat{\rho} \cdot \hat{\pi}$$

$$\rho_n = \sum_m \rho_m \pi_{mn}$$

$\hat{\pi}$ is a stochastic matrix – its rows sum up to 1. We will also assume that it is an irreducible matrix which means that all states are accessible leading to ergodic behavior.

Perron-Frobenius theorem for stochastic irreducible matrices: The maximum eigenvalue is $\lambda(1) = 1$. It’s simple (non-degenerate). Its eigenvector is real. No other real eigenvector exists.

Consequences:

a) limiting distribution $\bar{\rho}$ exists

b) eigenvalues $\lambda(n), n > 1$ control the convergence rate to $\bar{\rho}$

Baumketner, BioSim, Lviv 2019
Application to the canonical configuration integral

The transition matrix $\hat{\pi}$ is unknown. But we know the limiting distribution $\rho_m = \rho_{NVT}(\Gamma_m)$

Some rules for constructing $\hat{\pi}$

a) avoid the need to compute $Q_{NVT} = \sum_m \rho_m$

$\rho_n \pi_{nm} = \rho_m \pi_{mn}$

Take the sum over $m$:

\[
\sum_m \rho_n \pi_{nm} = \rho_n \sum_m \pi_{nm} = \rho_n = \sum_m \rho_m \pi_{mn}
\]

not needed for averages

may be impossible to compute if the size of the phase space is large

$\sum_m \rho_n \pi_{nm} = \rho_n \sum_m \pi_{nm} = \rho_n = \sum_m \rho_m \pi_{mn}$

b) Detailed balance :

$\rho_n \pi_{nm} = \rho_m \pi_{mn}$

# transitioning from $n$ to $m$

# transitioning from $m$ to $n$

$\rho_n$ is the desired eigenvector

A number of methods are available to build $\hat{\pi}$ that satisfies detailed balance

Metropolis-Hastings   Wood/Glauber/Barker   Kawasaki

Baumketner, BioSim, Lviv 2019
Metropolis-Hastings method: asymmetrical solution

\[
\begin{align*}
\pi_{mn} &= \alpha_{mn} \quad \rho_n \geq \rho_m \quad m \neq n \\
\pi_{mn} &= \alpha_{mn} \cdot \frac{\rho_n}{\rho_m} \quad \rho_n < \rho_m \quad m \neq n \\
\pi_{mm} &= 1 - \sum_{m \neq n} \pi_{mn} 
\end{align*}
\]

\[\sum_n \pi_{mn} = 1\]

\[\alpha_{mn} = \alpha_{nm} \quad \text{a symmetric stochastic matrix underlying Markov chain}\]

Proof that the solution satisfies the detailed balance:

\[\rho_n \leq \rho_m: \quad \rho_n \pi_{nm} = \rho_n \alpha_{nm} = \boxed{\rho_m \alpha_{mn}} \frac{\rho_n}{\rho_m} = \rho_m \alpha_{mn} \frac{\rho_n}{\rho_m} = \rho_m \pi_{mn}\]

\[\rho_n > \rho_m: \quad \rho_n \pi_{nm} = \rho_n \alpha_{nm} \frac{\rho_m}{\rho_n} = \alpha_{nm} \rho_m = \alpha_{mn} \rho_m = \rho_m \pi_{mn}\]

Important point: \(\pi_{nm}\) depends on the ratio \(\rho_n/\rho_m\) but not on these quantities individually.

One needs to know \(\rho_n\) up to a multiplicative constant to arrange a Markov chain. The normalizing factor \(Q_{NVT} = \sum_n \rho_n\) is not required.
Symmetric solution

\[
\begin{align*}
\pi_{mn} &= \alpha_{mn} \frac{\rho_n}{\rho_n + \rho_m} \quad m \neq n \\
\pi_{mm} &= 1 - \sum_{n \neq m} \pi_{mn} \\
\alpha_{mn} &= \alpha_{nm} \quad \text{a symmetrical stochastic matrix}
\end{align*}
\]

Proof that the solution satisfies the detailed balance:

\[
\rho_n \pi_{nm} = \rho_n \alpha_{nm} \frac{\rho_m}{\rho_n + \rho_m} = \rho_m \alpha_{mn} \frac{\rho_n}{\rho_n + \rho_m} = \rho_m \pi_{mn}
\]

Which solution is better?

Statistical inefficiency to measure the rate of convergence to the limiting distribution: low inefficiency = fast convergence

\[
s = \frac{\tau_{run} \sigma^2(<A>_{run})}{\sigma^2(A)}
\]

\[
[\hat{\pi}_1]_{nm} > [\hat{\pi}_2]_{nm} \quad m \neq n
\]

\[
s(\pi_1) < s(\pi_2)
\]

Metropolis algorithm has faster convergence rate

---

Baumketner, BioSim, Lviv 2019
Let’s see how Monte Carlo can be applied to simulate liquids, in particular Lennard-Jones liquid.

**Configuration space:**

\[ \Gamma = (\vec{r}_1, ..., \vec{r}_N) \]

\[ U(\Gamma) = \frac{1}{2} \sum_{i \neq j}^N U_{LJ}(r_{ij}) \]

Periodic boundary conditions (PBC) are applied to remove the surface artifacts

**Probability distribution:**

\[ \rho_{NVT}(\Gamma) \sim e^{-\beta H(\Gamma)}, \beta = \frac{1}{kT} \]

\[ U_{LJ}(r) = 4\epsilon \left( \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right) \]
How to design MC moves

In order to run MC one needs to select the symmetric matrix $\alpha$.

For simplicity the matrix $\alpha_{nm} = \alpha_{mn} = \alpha$ is assumed to be a constant

One possible choice for this constant is related to how new configurational states $r_n$ are generated.

Assume new trial states are generated by random displacement of atom $i$ by vector: $\delta \vec{r} = (\xi_1 \delta x, \xi_2 \delta x, \xi_3 \delta x)$ where $\delta x$ is the maximum allowed displacement and $\xi_i \in [0,1]$ are random numbers

If the initial state is $n$, then the final states $m$ will make a cube with side $\delta x$. This cube will contain a large but finite (on computers) number of points $N_k$.

Any one of these points will have an $\frac{1}{N_k}$ probability of being occupied upon transition.

Therefore random displacements occur with transition probability $\frac{1}{N_k} = \pi_{nm} = \alpha_{nm}$

A natural choice:

$$\alpha = \frac{1}{N_k} \quad N_k \sim \delta x^3$$

so $\alpha$ will be set by the magnitude of $\delta x$

Baumketner, BioSim, Lviv 2019
Simulation scheme

The system is in an initial state $n$. Potential energy is available.

→ **Step 1:** pick a random displacement vector. This will happen with probability $\alpha$.

**Step 2:** If the resulting $\rho_m > \rho_n$, accept the move as this means $\pi_{nm} = \alpha$ transition probability.

**Step 3:** If the resulting $\rho_m < \rho_n$, accept the move with $\frac{\rho_m}{\rho_n}$ probability. This means $\pi_{nm} = \alpha \frac{\rho_m}{\rho_n}$ transition probability

For $\rho_n \sim e^{-\beta E_n}$ and $\Delta E = E_m - E_n$, the algorithm can be written as follows:

Accept $n \rightarrow m$ move with $\min\{1, e^{-\beta \Delta E}\}$ probability

How to decide whether a given move should be accepted or rejected based on its desired probability $P$?

The outcome is stochastic so it has to rely on a stochastic/random process. The simplest method is to flip a coin or roll a dice.

Baumketner, BioSim, Lviv 2019
Coin example:

If we don’t want/are unable to make a decision we leave it to chance.

If we flip a coin, the positive outcome of the decision will have \( \frac{1}{2} \) probability.

Decision is made with \( P = \frac{1}{2} \).

If we roll a dice, the positive outcome of the decision may have probability \( \frac{1}{6} \) or a number of other values.

\[
P = \frac{1}{6} \quad \text{or} \quad P = \frac{2}{6} = \frac{1}{3}
\]
In general, a dice with $N$ sides may encode $N - 1$ distinct probabilities:

$$P = \frac{1}{N}$$

or

$$P = \frac{N - 1}{N}$$

Generalization to continuous $P$:

$$\xi \in [0,1] \quad \text{a random number}$$

$$P_1(\xi) \quad \text{uniform distribution}$$

The event that $\xi$ is below $P$ will occur with $P$ probability.

Generate a random number $\xi$. If it is below $P$ - accept the move, otherwise – reject it.
**Step 1:** Generate new positions for particle $k$:

\[
\begin{align*}
rx_{\text{new}} &= rx(k) + (2 \times \text{rand}([0,1]) - 1.0) \times dx \\
ry_{\text{new}} &= ry(k) + (2 \times \text{rand}([0,1]) - 1.0) \times dx \\
rz_{\text{new}} &= rz(k) + (2 \times \text{rand}([0,1]) - 1.0) \times dx
\end{align*}
\]

Apply PBC

**Step 2:** Compute the resulting change in potential energy

\[\Delta E = E_m - E_n\]

\[2E = \sum_{i \neq j} U_{LJ}(r_{ij}) = \sum_{j \neq k} U_{LJ}(r_{jk}) + \sum_{k \neq j} U_{LJ}(r_{kj}) + \sum_{i \neq j, i \neq k, j \neq k} U_{LJ}(r_{ij})\]

\[2(\Delta E) = \sum_{j \neq k} U_{LJ}(r_{jk}^m) + \sum_{k \neq j} U_{LJ}(r_{kj}^m) - \sum_{j \neq k} U_{LJ}(r_{jk}^n) - \sum_{k \neq j} U_{LJ}(r_{kj}^n)\]

**Step 3:** If $\Delta E < 0$, accept the move. Otherwise, generate $\xi \in [0,1]$

- If $\xi \leq w = e^{-\beta \Delta E}$ accept the move
- If $\xi > w$ reject the move

Baumketner, BioSim, Lviv 2019
Implementation example in Fortran. One particle moves at a time

deltaE = E_{new} - E_{old}
deltab = \frac{\Delta E}{k \cdot T}

\textbf{if} (\text{deltab} \leq 0.75) \textbf{then}

\textbf{if} (\text{deltab} \leq 0.0) \textbf{then}
\begin{align*}
e &= e + \text{deltaE} \\
x_r(k) &= x_{rnew} \\
y_r(k) &= y_{rnew} \\
z_r(k) &= z_{rnew} \\
naccp &= naccp + 1
\end{align*}
\textbf{else}
\begin{align*}
e &= e + \text{deltaE} \\
x_r(k) &= x_{rnew} \\
y_r(k) &= y_{rnew} \\
z_r(k) &= z_{rnew} \\
naccp &= naccp + 1
\end{align*}
\textbf{endif}
\textbf{endif}

ntrial = ntrial + 1

guard against overflow
don't generate the random number if \( \Delta E < 0 \). Saves time

generate the random number
reassign coordinates
update the counter of accepted moves
update the counter of total trials.

\textbf{Extensions:}

a) multiple-particle moves
b) all-particle moves

Baumketner, BioSim, Lviv 2019
The goal is to reproduce the NTP ensemble. Volume now has to be treated as a dynamical variable.

**Phase space:**

\[
\begin{align*}
(r_1 \ldots r_N, V) & \quad \rightarrow \quad (r_1 \ldots r_N, V + \delta V) \\
(r_1 \ldots r_N, V) & \quad \rightarrow \quad (r_1 \ldots r_N, V - \delta V)
\end{align*}
\]

After relaxation, particles will fill up the box.

PBC will return the particles to the main simulation box but there will be many steric clashes.

Alternative approach: introduce scaled coordinates

**old variables:**

\[
\begin{align*}
(r_1 \ldots r_N, V)
\end{align*}
\]

**new variables:**

\[
\begin{align*}
(s_1 \ldots s_N, V)
\end{align*}
\]

\[
\hat{r} = V^{\frac{1}{3}} \cdot \hat{s}, \quad (s_x, s_y, s_z) \in [0, 1], \quad d\hat{r_i} = V \cdot ds_i
\]

The box will experience uniform expansion or contraction.

Baumketner, BioSim, Lviv 2019
How to compute NPT averages?

![Equation](image)

\[<A>_{NPT} = \frac{\int dVe^{-\beta PV} \int d\vec{r}e^{-\beta U(\vec{r})} A(\vec{r})}{\int dVe^{-\beta PV} \int d\vec{r}e^{-\beta U(\vec{r})}} = \frac{\int dVe^{-\beta PV} V^N \int d\vec{s}e^{-\beta U(\vec{s})} A(\vec{s})}{\int dVe^{-\beta PV} V^N \int d\vec{s}e^{-\beta U(\vec{s})}} = \frac{\int dV d\vec{s}e^{-\beta U(\vec{s})} e^{-\beta PV} e^{N \log(V)} A(\vec{s})}{\int dV d\vec{s}e^{-\beta U(\vec{s})} e^{-\beta PV} e^{N \log(V)}} = <A>_{sV} \]

Sample from NVT ensemble for the extended system defined by phase space coordinates \( \Gamma = (\vec{s}_1, \ldots, \vec{s}_N, V) \) with the limiting distribution function:

\[\rho(\Gamma) = e^{-\beta(U(\vec{s}) + PV - kTN \log(V))} \]

Algorithm:
Simulation is conducted as in NVT ensemble except that now we have two types of moves.

1) Coordinate moves:
\[\vec{s}_i^m = \vec{s}_i^n + \delta x(2 \cdot \xi - 1) \quad \xi \in [0,1] \]

2) Volume changes:
\[V^m = V^n + \delta V(2 \cdot \xi - 1) \]

Moves accepted with the probability:
\[P = \min\{1, e^{-\beta \Delta E}\}, \Delta E = E_m - E_n + P(V_m - V_n) - kT \log\left(\frac{V_m}{V_n}\right)\]

Baumketner, BioSim, Lviv 2019
Both coordinates and the number of particles are allowed to change. It's more convenient to introduce dimensionless coordinates as this will make the distribution function dimensionless as well. This is important when comparing systems with different numbers of particles, \( N \) and \( N + \mu_{\text{reservoir}} \). For instance, as they have different volumes of the phase space.

Parameters of the ensemble: \( \mu, V, T \)

Phase-space variables: \((\mathbf{r}_1, ..., \mathbf{r}_N, N)\)  

Scaled variables: \((\mathbf{s}_1, ..., \mathbf{s}_N, N)\)  

Ensemble averages:

\[
< A >_{\mu VT} = \frac{1}{N!} \int d\mathbf{s} e^{-\beta U(\mathbf{s})} A(\mathbf{s}) / Q_{\mu VT} =
\]

\[
= \frac{\sum_{N=1}^{\infty} \int d\mathbf{s} e^{-\beta(U(\mathbf{s})-\mu N-kTN \log(V)+kTN \log(\lambda^3)+kT \log(N!)} A(\mathbf{s}; N)}{\sum_{N=1}^{\infty} \int d\mathbf{s} e^{-\beta(U(\mathbf{s})-\mu N-kTN \log(V)+kTN \log(\lambda^3)+kT \log(N!)} =
\]

Limiting distribution in MC chain:

\[
\rho(\mathbf{s}_1, ..., \mathbf{s}_N, N) = e^{-\beta(U(\mathbf{s})-\mu N-kTN \log(V)+kTN \log(\lambda^3)+kT \log(N!)}
\]

There are a number of implementations of GCMC that differ in how particles are added to/removed from the system.

Baumketner, BioSim, Lviv 2019
Three types of moves:

1) Coordinate moves: \( \vec{s}_i^m = \vec{s}_i^n + \delta x(2 \cdot \xi - 1) \quad \xi \in [0,1] \)

\[ P = \min\{1, e^{-\beta \Delta U}\} \]

2) Particle creation:

Insertion at a random position. Difficult in dense fluids because of steric clashes.

\[ P \sim \frac{\rho_m(\vec{s}, N + 1)}{\rho_n(\vec{s}, N)} = \min \{1, e^{-\beta \Delta U + \log \frac{z V}{N+1}}\} \]

\[ \Delta U = U(\vec{s}, N + 1) - U(\vec{s}, N) \]

\[ z = e^{\beta \mu / \lambda^3} \quad \text{activity} \]

3) Particle destruction:

Deletion of a random particle. Difficult in dense fluids because the particle may experience strong attraction in the media.

\[ P \sim \frac{\rho_m(\vec{s}, N - 1)}{\rho_n(\vec{s}, N)} = \min \{1, e^{-\beta \Delta U + \log \frac{N}{z V}}\} \]

\[ \Delta U = U(\vec{s}, N - 1) - U(\vec{s}, N) \]

Direct computation of free energy in GCMC:

\[ A/N = \mu - \langle P \rangle_{\mu VT} V / \langle N \rangle_{\mu VT} \]

Baumketner, BioSim, Lviv 2019

For fastest convergence choose \( P_1 = P_2 = P_3 = \frac{1}{3} \)
Gibbs-ensemble simulations

This method is designed to simulate phase equilibria.

\[
\begin{align*}
\text{Liquid} & : & P_l &= P_g \\
& & T_l &= T_g \\
& & \mu_l &= \mu_g \\
\text{Gas} & & \\
\end{align*}
\]

The best choice for these experimental conditions is $\mu^{PT}$ "ensemble"

One of the state variables has to be extensive ($N$ or $V$)

Consider two coupled systems with the total $V$ and $N$ fixed:

\[
\begin{align*}
V_1 &= V - V_2 \\
N_1 &= N - N_2 \\
V_2 & & \\
N_2 & & \\
\end{align*}
\]

Exchange of particles and volume is allowed. This enables coexistence between two phases. The advantage is that there is no interface. Molecules in system 1 and 2 do not interact with one another.

Such ensemble does not exist!

If both $N$ and $V$ are allowed to change a simulation box can’t be defined uniquely

Baumketner, BioSim, Lviv 2019
The partition function:

\[ Q_G(N, V, T) = \sum_{n_1=0}^{N} \frac{1}{V^3 \lambda^{N-n_1} n_1! (N-n_1)!} \int_{V_1}^{V} dV_1 V_1^{n_1} (V - V_1)^{N-n_1} \times \]

\[ \times \int ds_1^{n_1} e^{-\beta U(s_1)} \int ds_2^{N-n_1} e^{-\beta U(s_2)} \]

Distribution function:

\[ \rho(n_1, V_1, s_1, s_2) = \frac{V_1^{n_1} (V - V_1)^{N-n_1}}{n_1! (N-n_1)!} e^{-\beta (U(s_1) + U(s_2))} \]

MC process that samples from that distribution:

1) Coordinate moves applied in both systems independently:

\[ \vec{s}_i^{\text{new}} = \vec{s}_i^{\text{old}} + \delta x (2 \cdot \xi - 1) \quad \xi \in [0,1] \]

\[ P = \min\{1, e^{-\beta \Delta U}\} \]

Variables:

\[ (\vec{s}_1 \ldots \vec{s}_{n_1}, n_1, V_1) \]

\[ (\vec{s}_1 \ldots \vec{s}_{N-n_1}) \]
2) Volume change:

\[ V_1^m = V_1^n + \delta V \]

\[ V_2^m = V_2^n - \delta V \]

Probability of \( n \) to \( m \) transition:

\[ P = \min\{1, \frac{(V_1^m)^{n_1} (V-V_1^m)^{N-n_1}}{(V_1^n)^{n_1} (V-V_1^n)^{N-n_1}} e^{-\beta(\Delta U(s_1)+\Delta U(s_2))} \} \]

3) Particle exchange:

\[ N_1^m = N_1^n - 1 \]

\[ N_2^m = N_2^n + 1 \]

Probability of \( n \) to \( m \) transition:

\[ P = \min\{1, \frac{n_1(V-V_1)}{(N-n_1+1)V_1} e^{-\beta(\Delta U(s_1)+\Delta U(s_2))} \} \]
**Typical results:**

After initial relaxation, densities in the two boxes will settle down to their equilibrium values.

In the coexistence region $T < T_c$ there will be two distinct densities corresponding to two distinct phases.

In the supercritical region $T > T_c$ there will be only one density. Boxes may have different sizes and numbers of particles.

Large finite size effect for gas-lattice models. Minimal effects for continuous models. Systems with <100 particles are OK for LJ model in both 2D and 3D.

Baumketner, BioSim, Lviv 2019
Electrostatics in biomolecular systems

Coulomb interaction between two charges \( U_c(r) = \frac{q_1 q_2}{r} \).

If \( r < r_0 \), \( U_c \) is attractive, as with LJ potential, the long-range correction is:

\[
\frac{-k_c e^2}{4\pi\varepsilon_0} \int \frac{e^2}{r^2} dr = \frac{k_c e^2}{4\pi\varepsilon_0} \text{ - integral diverges.}
\]

The long-range part of the potential energy is greater than the short-range part.

---

**Attraction forces between two positively charged ions**

1) Truncate the potential at \( r_0 \)

   - Bad approximations
   - Wrong results even at low \( r_0 \)

Example: effective interaction \( U_e(r) \) between two iron ions.

For large \( R \), \( U_2 \approx \frac{q_1 q_2}{R} \) and \( \varepsilon R_0 \)

Attractive forces are seen at \( R < R_0 \).
\[ \text{II) Reaction field corrections (RF method)} \]

Compute contributions of the truncated long-range part of the using continuum electrorheological models. For homogeneous media:

AtOMIC REprESENTATION \[ \xrightarrow{\text{}} \] CONTINUUM REprESENTATION

Electrostatics of charges in spherical cavities:

Poisson equation relating potential \( \Phi \) and charge density \( \rho \):

\[
\begin{cases}
\Delta \Phi(x) = -4\pi \rho(x) & \text{in } \Omega \\
\Delta \Phi(x) = 0 & \text{in } \Omega^c
\end{cases}
\]

Boundary conditions:

\[
\begin{cases}
\Phi(x) = \Phi(x) & \text{on boundary} \\
\frac{\partial \Phi(x)}{\partial n} = \varepsilon \frac{\partial \Phi(x)}{\partial n} & \text{normal derivative, or displacement } \vec{u} = \varepsilon \vec{E}, \text{ \( \varepsilon \) continuous}
\end{cases}
\]

Baumketner, BioSim, Lviv 2019
For the potential inside the cavity:

\[ \Phi_i(\vec{r}) = \frac{q_i}{|\vec{r} - \vec{r}_i|} + \Phi_e(\vec{r}) \]

*Induced charge density that creates reaction potential

\[ \Phi_e(\vec{r}, \Theta) = \sum_{n=0}^{\infty} B_n \cdot \hat{n} \cdot P_n(\cos \Theta) \]

\[ B_n = \frac{2^{n+1}}{n+1} \gamma \left( \frac{1+\gamma}{1-\gamma+2n}, \frac{\gamma-1}{\gamma+1} \right) \]

In the limit of small \( n \):

\[ \Phi_e(\vec{r}) = -\frac{q_i (t-1)}{a \cdot \epsilon} + \frac{\gamma}{q_a^2} \frac{2(t-1)}{2t+1} \frac{2t}{t^2+1} \]

Two important results that follow from this expansion:

1) Born solution energy

Assume that the charge is placed at the center \( \vec{r} = 0 \). The reaction potential it creates at the center \( \vec{r} = -\frac{q_i (t-1)}{a \cdot \epsilon} \)
It will have \( dW = \Phi_k(0) \, dy \) work against that potential to add by change at the center. The total work associated with creating charge \( \gamma \) - changing free energy, is then:

\[
\Delta \Gamma_{\text{Born}} (\gamma = 1) = 0 - \text{no changing energy in vacuum} \ (\gamma = \gamma_{\text{ext}})
\]

2) Oxygen solvation at point dipoles

Place two charges \( \gamma \) and \(-\gamma\) at positions \( z_s \) and \(-\tilde{z}_s \) respectively.

Dipole moment

\[
\mathbf{M} = \gamma \mathbf{z}_s + \gamma \mathbf{z}_s = 2 \gamma \mathbf{z}_s
\]

Point dipole \((2\mathbf{z}_s \to 0, \mathbf{M} = \text{const})\)
1) change particle +z. Work has to be done against its own reaction field:

\[ \Delta f_2 = \int_0^y \left( -\frac{y}{a \cdot t} + \frac{y}{a^3} \frac{2(1-t)}{2t+1} \frac{z^2}{2z^2} \right) dy = -\frac{1}{a} \frac{y^2 (t+1)}{a \cdot t} + \frac{1}{2} \frac{y^2}{a^3} \frac{2(1-t)}{2t+1} z^2 = \Delta G_1 \]

and against the reaction field created by the first particle:

\[ \Delta f_2 = \int_0^y \Phi(-z) \, dy \]

\[ = \int_0^y \left( -\frac{y}{a \cdot t} - \frac{y}{a^3} \frac{2(1-t)}{2t+1} z^2 \right) \, dy = \frac{y^2 (t-1)}{a \cdot t} + \frac{y^2}{a^3} \frac{2(1-t)}{2t+1} z^2 \]

Putting these terms together:

\[ \Delta G = \Delta G_1 + \Delta G_2 + \Delta G_3 = \frac{2y^2}{a} \frac{2(1-t)}{2t+1} z^2 \]

In the limit of point dipole, \( z \to 0 \), \( m = \text{const} \):

\[ \Delta G_0 = \frac{1}{a^2} \frac{1 - \delta}{2t+1} M^2 \]

- higher-order terms in \( z \) disappear. \( \Delta G_0 \) is full solvation energy
- \( \Delta G \) is dipolar approximation of the changing free energy

It needs to be corrected for non-point dipoles.

As with the Born energy for Born:

\[ \Delta G_0 (\epsilon = 1) = 0 \] for dipoles in a changing energy of vacuum!

Baumketner, BioSim, Lviv 2019
The changing free energy can also be interpreted as interaction of the dipole with its own reaction field. By Laplace field created at the center by a charge $q_i$ at position $\vec{r}_i$:

$$E_i(\vec{r}) = -\frac{\partial \Phi(\vec{r})}{\partial r} = -\frac{q_i}{4\pi \varepsilon_0} \frac{2(1-\ell)}{2\ell+1} \frac{1}{r^5} + O(1/r^7)$$

If there is more than one charge, $q_i, \vec{r}_i$:

$$E(\vec{r}) = -\frac{1}{4\pi \varepsilon_0} \sum_{i} \frac{2(1-\ell)}{2\ell+1} q_i \delta(\vec{r} - \vec{r}_i) = -\frac{1}{4\pi \varepsilon_0} \frac{2(1-\ell)}{2\ell+1} \frac{1}{r^5}$$

Let's assume now that the field is created by a pair of charges $-q_i, \vec{r}_i$ and $q_i, \vec{r}_i$ so that $\vec{E} = 2q_i \vec{r}_i$. Adding $-d\vec{y}$ and $d\vec{y}$ amounts of charge to these particles, will cost $d\Psi = -d\vec{r} \cdot \vec{E} = -2\vec{r}_i \cdot d\vec{y}$. This is the electric energy. Creating these charges from zero will require work:

$$\left\{ d\Psi = \frac{q_i}{4\pi \varepsilon_0} \frac{2(1-\ell)}{2\ell+1} \frac{1}{r^5} \delta(\vec{r} - \vec{r}_i) = \frac{1}{4\pi \varepsilon_0} \frac{1-\ell}{2\ell+1} q_i^2 r_i^5. \right\}$$

Outsider solvation energy

Baumketner, BioSim, Lviv 2019
Why is $\Delta C$ called "solution" energy?

Assume $X$ denotes configuration of our molecule & solvent. Ei energy in vacuum is $U_\text{vac}(x)$. Probability of seeing $x$

$P(x) \sim e^{-\beta U(x)}$

If the molecule is placed in a medium, a solvent:

$P(x) = \int e^{-\beta U(x) - \beta E(x,y)} dy / Q = \frac{1}{Q} \int e^{-\beta U_\text{med}(x)} dy$

$Q = \text{partition function}$

$U_\text{med}(x) - \text{potential energy of the molecule in the medium}$

Introduce $U_\text{eff}(x) = -kT \log \int e^{-\beta E(x,y)} dy$ - effective potential due to the medium

$U_\text{med}(x) = U_\text{vac}(x) + U_\text{eff}(x)$ - total potential is vacuum + $U_\text{eff}(x)$

What's the physical meaning of $U_\text{eff}(x)$?

Let's introduce a scaling variable $\lambda$ that will "turn on" the medium:

$Q(\lambda) = \int e^{-\beta U_\text{vac}(x) - \beta \lambda E(x,y)} dy$ - partition function

$G(\lambda) = -kT \log Q(\lambda)$ - free energy

$G_\text{vac}(\lambda) = G(\lambda=0) = U_\text{vac}(x) - \text{vacuum}$

$G_\text{med}(\lambda) = G(\lambda=1) = U_\text{med}(x) - \text{medium}$

Free energy of turning on the medium is the effective potential:

$U_\text{eff}(x) = G_\text{med}(\lambda) - G_\text{vac}(\lambda)$

$U_\text{med}(x) = U_\text{vac}(x) + U_\text{eff}(x)$

Baumketner, BioSim, Lviv 2019
Effect of dielectric continuum

potential energy in vacuum

\[ U_m(x) = U_v(x) + U_s(x) \]

potential energy & free energy

in solvent 

solute

Design a changing/dechanging path: a) dechanging in vacuum
b) placing neutral particles into a cavity, c) changing the particles in the medium

\[ \Delta G = \Delta G_0 + \Delta G_s + \Delta G_c = \Delta G_v(x) - U_s(x) \]

Total potential work against the potential

inside the cavity created by other particles work against the reaction field

\[ \Delta G_v(x) = \int dV \Phi = \int dy \Phi_v + \int dy \Phi_p \]

\[ = U_v + \int dy \Phi_p \]

\[ \Delta G = \int dy \Phi_p \]

Solvation energy = the energy that needs to be added to the potential in order to mimic the presence of solvent

Baumketner, BioSim, Lviv 2019
Moving boundary reaction field method

Use the Kirkwood expansion to compute electric fields created at the center by a source charge $q_i$:

$$
\vec{E}_k(\vec{r} = 0) = -\frac{2}{a^2} \Phi_k(\vec{r}) \bigg|_{\vec{r} = 0} = \frac{q_i}{a^3} \frac{2(c-1)}{2c+1} \vec{z}
$$

Force experienced by the probe charge $q_j$ at the center due to polarization is:

$$\vec{F} = q_j \vec{E}_k(\vec{r} = 0) = \frac{\mu \nu}{\ell} \frac{2(c-1)}{2c+1} \vec{z}$$

(Force from its own polarization or zero)

This force can be modeled by introducing an effective potential energy between particles $i$ and $j$:

$$U_{\text{eff}}(\vec{r}) = \frac{q_i q_j}{a^3} \frac{c-1}{2c+1} \vec{r}^2
$$

$$
\vec{F}_i = -\vec{F}_j = \frac{2}{a^3} \frac{\mu}{\ell} U_{\text{eff}}(\vec{r}) = \frac{q_i q_j}{a^3} \frac{c-1}{2c+1} \vec{r}
$$

The total force acting between charges $q_i$ and $q_j$ if $n < a$ is:

$$U_{\text{tot}}(\vec{r}) = U_c(\vec{r}) + U_{\text{eff}}(\vec{r}) = \frac{q_i q_j}{\ell} + \frac{q_i q_j}{a^3} \frac{c-1}{2c+1} \vec{r}^2 = \frac{q_i q_j}{\ell} \left( 1 + \frac{c-1}{2c+1} \left( \frac{r}{a} \right)^3 \right)
$$
The reaction field results in a modified Coulomb potential of $a = \varepsilon_0$.

In all other aspects simulation are the same.

A method is applicable to homogeneous systems only.

The area outside cut-off sphere? This approximation has to be homogeneous continuum.

Applied to some solute atoms.

Doesn't apply to water solvent atoms!

6) The modified potential is still long-ranged.

$$ U_{fp}(r = \varepsilon_0) = \frac{9 \varepsilon_0 q_s^2}{\varepsilon_0} \frac{3 \pi}{2\varepsilon_0^2} \approx \frac{9 \varepsilon_0 q_s^2}{2 \varepsilon_0^2} \frac{3}{2} U_C(\varepsilon_0) $$

The potential is still significant at the cut-off distance. The force is discontinuous at $\varepsilon_0$ or worse. This requires large $\varepsilon_0$ to minimize the contribution at the truncated part.
1) Resonance-field correction removes a big difference. Solution of a \( ^{16}O^{16} \) ions

resonance fields removes medium-range repulsion

Resonance fields remove long-range repulsion

Fig. 1. Potential of mean force obtained in this work for NaCl using group-based long-range cutoff method. Two values of the cutoff distance were considered, \( r_c = 0.5 \) nm and 1.5 nm. The maximum of the correlation function is smaller at large cutoff radii.

Fig. 4. Potential of mean force obtained in this work for NaCl using group-based reaction-field correction method. Two values of the cutoff distance were considered, \( r_c = 0.5 \) nm and 1.5 nm. Compared to Fig. 3, the unphysical minima resulting from the potential at \( r < r_c \) is smaller. Significant residual errors are observed in the association energy of two ions.

d) In molecular systems, the way truncation is implemented matters a big difference

Group-based truncation

Atom-based truncation

Group-based truncation

Large error that depends on the scale of the simulation box: lattice-sum method

Good agreement with theory

Edited changes
Ewald summation

To avoid summing the potential, consider the infinite lattice in which particles in the central simulation box interact with all their periodic images:

$$E_{\text{es}} = \sum_{i,j} \frac{q_i q_j}{|R_i - R_j + R_z'|} \quad \text{plane wave}$$

This seems a conditionally convergent, meaning the result now depend on how summation is performed. A physically appealing way is to systematically grow the lattice from all directions into infinity.

The summation can be performed directly without convergence to each atomic configuration, but this method is extremely inefficient.

Baumketner, BioSim, Lviv 2019
The difficulty is that for each pair \( i \) and \( j \) alone, an infinite (very large) number of terms have to be considered. The situation could be simplified if some or these terms were small, or were made small. This can be achieved if instead of Coulomb, some short-range potential were considered.

Each change in the electrostatic potential with the infinite lattice:

\[ \phi_c \propto \text{some short-range potential} \]

\[ \phi_c = \frac{q_i}{r_{ij}} \]

Only a limited number of \( \tilde{r}_n \) vectors will contribute to screened potential energy.

Assume that some compensating charge contribution \( l_6(\tilde{r}) \) can be added to each \( y_i \) so that the potential becomes short-range \( \phi_{sc}(\tilde{r}) \). The potential created by that change is:

\[ 1 \leq i \leq n \]

\[ \frac{1}{2} \sum_{i,j} y_i \phi_{sc}(\tilde{r}_{ij} + \tilde{r}_n) = \frac{1}{2} \sum_{i,j} y_i \phi_{sc}(\tilde{r}_{ij} + \tilde{r}_n) - \sum_{i,j} y_i \phi_{sc}(\tilde{r}_{ij} + \tilde{r}_n) \]

\[ \tilde{r}_n \]

\[ T \]

Ward energy \[ \rightarrow \]

Short-range interactions \[ \rightarrow \]

Long-range interactions

easy to compute

Baumketner, BioSim, Lviv 2019
The sum on the long-range part can be supplemented by the \( i = j, \bar{r}_i = \bar{r}_j \) term

\[
\frac{1}{2} \sum_{i,j,\bar{r}_i,\bar{r}_j} \psi_i \psi_j (\bar{r}_{ij} + \bar{r}_n) - \frac{1}{2} \sum_{i,j,\bar{r}_i,\bar{r}_j} \psi_i \psi_j (\bar{r}_{ij} + \bar{r}_n) - \frac{1}{2} \sum_{i,j,\bar{r}_i,\bar{r}_j} \psi_i \psi_j (\bar{r}_{ij} + \bar{r}_n) - \frac{1}{2} \sum_{i,j,\bar{r}_i,\bar{r}_j} \psi_i \psi_j (\bar{r}_{ij} + \bar{r}_n)
\]

The total energy then is easily to compute

\[
\mathcal{E}_B = \sum_{i,j,\bar{r}_i,\bar{r}_j} \psi_i \psi_j (\bar{r}_{ij} + \bar{r}_n) - \sum_{i,j,\bar{r}_i,\bar{r}_j} \psi_i \psi_j (\bar{r}_{ij} + \bar{r}_n) - \sum_{i,j,\bar{r}_i,\bar{r}_j} \psi_i \psi_j (\bar{r}_{ij} + \bar{r}_n)
\]

\( \mathcal{E}_B \) is background energy.

Change density of the central cell \( l_1(z) = \frac{1}{L_xL_yL_z} \psi_1 (\bar{z} - \bar{z}_1) \) background change \( l_1(z) \)

\[
\mathcal{V} = L_x \times L_y \times L_z
\]

\[
\mathcal{E}_B = \frac{1}{2} \sum_{i,j,\bar{r}_i,\bar{r}_j} \psi_i \psi_j (\bar{r}_{ij} + \bar{r}_n) = \frac{1}{2} \int \mathcal{S}_B \phi_1 (\bar{z}) \phi_1 (\bar{z}) \mathcal{B}_B = \frac{1}{2} \int \mathcal{S}_B \phi_1 (\bar{z}) \phi_1 (\bar{z}) \mathcal{B}_B
\]

\( \mathcal{E}_B \) and its accumulated change contribution \( l_1(z) \) are related by the Poisson eq.

\[
\Delta \psi_1 (\bar{z}) = - \Delta \mathcal{E}_B (\bar{z})
\]

Introduction.

\[
\psi_1 (\bar{z}) = \phi_1 (\bar{z}) e^{i \mathcal{B}_B \bar{z}}
\]

\[
\Delta \psi_1 (\bar{z}) = - \frac{i}{\mathcal{B}_B} \psi_1 (\bar{z}) e^{i \mathcal{B}_B \bar{z}}
\]

\[
\psi_1 (\bar{z}) = \frac{1}{\mathcal{B}_B} \phi_1 (\bar{z}) e^{i \mathcal{B}_B \bar{z}}
\]

Baumketner, BioSim, Lviv 2019
Poisson equation reads:

$$\kappa^{-2} \psi(x) = -4\pi \rho(x)$$

$$E_0 = \frac{1}{2} \int \psi_0(x) \psi_1(x) \delta(x) \, dx = \frac{1}{2} \int \delta(x-\vec{r}_i) \psi_0(x) \, dx$$

$$\psi_0(x) = \frac{\psi_1(x)}{\psi_1(\vec{r}_i)} e^{-i\frac{\vec{k}_i}{\kappa} \cdot \vec{x}}$$

$$\Phi_0(x) = -4\pi \rho_0(x) = -\Psi_0(x) = -\frac{q_i}{\kappa} \int_{0}^{\infty} e^{-x^2} \, dx$$

$$\Phi_{sc}(x) = \frac{q_i}{\kappa} - q_i \int_{0}^{\infty} \frac{\psi_0(x)}{\psi_1(\vec{r}_i)} e^{-x^2} \, dx$$

$$\Phi_{sc}(\vec{r}) \overset{\text{when}}{=} 0$$

$$\Phi_{sc}(\vec{r}) = -q_i \frac{2}{\sqrt{\pi}} \int_{0}^{\infty} e^{-x^2} \, dx$$

$$\Phi_{sc}(\vec{r}) = -q_i \frac{2}{\sqrt{\pi}}$$

\(\Phi_{sc}(\vec{r})\) to be used in \(E_0\)

If \(\Phi_{sc}(\vec{r})\) decays rapidly with \(
\vec{r}\), the sum may be easy to compute.

Assume that the background charge density \(\Phi_{sc}\) is Gaussian. For end charge \(q_i\):

$$\Phi_{sc}(\vec{r}) = -q_i \left( \frac{\vec{r}}{\kappa} \right)^{3/2} e^{-\frac{r^2}{\kappa^2}}$$

\(\kappa\) = adjustable parameter

Baumketner, BioSim, Lviv 2019
For the discrete lattice at background charge:

\[
l_b(x) = \sum_{i,j,k,l} l_b(i,j,k,l,e_u) = -\left(\frac{2\pi}{a}\right)^d \sum_{j,k,l} \alpha_j e^{-a(i-j)^2 + (k-l)^2 + e_u^2}
\]

\[
l_b(x) = -\frac{1}{\sqrt{2\pi}} \int \frac{2\pi}{a} \sum_{j,k,l} y_j e^{-a(i-j)^2 + (k-l)^2 + e_u^2} e^{-i \cdot x} d^2 x = 0
\]

\[
\int d^2 \bar{z} \leq F(\bar{z} + e_u) = \int d^2 \bar{z} F(\bar{z})
\]

all space

\[
\int d^2 \bar{z} = \int e^{-a(i-j)^2 + (k-l)^2 + e_u^2} e^{-i \cdot x} d^2 x = 0
\]

with \( \bar{z} = z - e_u \) substitution

\[
\int d^2 \bar{z} = \int e^{-a(i-j)^2 + (k-l)^2 + e_u^2} e^{-i \cdot x} d^2 x = 0
\]

For \( \bar{z} \):

\[
l_b(\bar{z}) = -e^{\frac{k^2}{a^2}} l_b(\bar{z})
\]

\[
l_b = -\frac{1}{2} \sum_{i,j} \frac{q_i q_j}{k^2} l_b(i,j) e^{-i \cdot z} = 2 \pi \sum_{k} \frac{1}{e^k} l_i(\bar{e}) l_i(\bar{e}) = 2 \pi \sum_{k} \frac{1}{e^k} e^{-\frac{k^2}{a^2}} |l_i(\bar{e})|^2
\]
Background charge contribution can be computed in the Fourier space:

\[ E_0 = \sum \frac{k^2}{2} e^{\frac{-k^2}{2}} |\phi_i(\vec{k})|^2 \]

If \( \phi_i(\vec{k}) \) are available, the sum will be rapidly converging with \( |\vec{k}| \) due to the exponential decay of \( e^{-\frac{k^2}{2}} \).

\( E_0 \) contains one term \( \vec{k} = 0 \) which needs separate treatment. If we introduce \( f(\vec{k}) = \frac{1}{\vec{k}^2} e^{\frac{-k^2}{2}} |\phi_i(\vec{k})|^2 \), then \( E_0 = \sum \phi_0 + \sum f(\vec{k}) \).

\[ f(\vec{k}) = \delta_{\vec{k}, \vec{0}} f(\vec{0}) \]

\[ E_0 = \sum \phi_i(\vec{0}) \]

- \( a) \) Limit does not exist. The formal sum is divergent and should not be used.
- \( b) \) Limit exists. \( E_0 \) can be determined uniquely.
- \( c) \) Limit is conditional. \( E_0 \) can be determined up to a constant.

Expand \( f(\vec{k}) \) in powers of \( |\vec{k}| \) around \( |\vec{k}| = 0 \). Choose vector \((\vec{k}_0, 0, 0)\) for simplicity:

\[ f(\vec{k}) = \frac{1}{\vec{k}^2} \sum_i \phi_i (1 + i\vec{k}_0 \cdot \vec{x}_i - \frac{1}{2} \vec{k}_0^2 \cdot \vec{x}_i^2 + \ldots) \]

\[ |f(\vec{k})|^2 = \frac{1}{\vec{k}^2} \sum_i \sum_j \phi_i \phi_j (1 + i\vec{k}_0 \cdot (\vec{x}_i - \frac{1}{2} \vec{k}_0 \cdot \vec{x}_i^2 + \ldots) \cdot \phi^*_j (1 + i\vec{k}_0 \cdot \vec{x}_j - \frac{1}{2} \vec{k}_0 \cdot \vec{x}_j^2 + \ldots) \]

Baumketner, BioSim, Lviv 2019
\[ \left| \mathbf{i} \cdot (\mathbf{E}) \right|^2 = \frac{1}{\nu^2} \left( (\mathbf{i} \cdot \mathbf{q}_i)^2 + \kappa \nu \left( \mathbf{i} \cdot \mathbf{q}_i \times \mathbf{q}_i \right)^2 - \kappa \nu^2 \mathbf{i} \cdot \mathbf{q}_i \times \mathbf{q}_i \times \mathbf{q}_i \right) + O(\kappa \nu^2) \]

\[ G(\mathbf{E}) = \frac{1}{\nu^2} e^{-\frac{\kappa \nu}{2}} \int \frac{d\mathbf{k}}{\kappa^2} \left( (\mathbf{k} \cdot \mathbf{q}_i)^2 + \kappa^2 (\mathbf{k} \cdot \mathbf{q}_i \times \mathbf{q}_i)^2 - \kappa^2 \mathbf{k} \cdot \mathbf{q}_i \times \mathbf{q}_i \times \mathbf{q}_i \right) + O(\kappa \nu^2) \]

\[ E \mathbf{i} \cdot \mathbf{q}_i = 0 \]

\( \text{(in } G(\mathbf{E}) \text{ does not exist and the Bould sum is divergent.)} \]

Assume change neutral simulation cell \( \mathbf{q}_i = 0 \).

\[ G(\mathbf{E}) = \frac{1}{\nu^2} e^{-\frac{\kappa \nu}{2}} \int \frac{d\mathbf{k}}{\kappa^2} (\mathbf{k} \cdot \mathbf{q}_i)^2 = \frac{1}{\nu^2} e^{-\frac{\kappa \nu}{2}} \int \frac{d\mathbf{k}}{\kappa^2} (\mathbf{k} \cdot \mathbf{M})^2, \quad \mathbf{M} = \mathbf{q}_i \cdot \mathbf{q}_i \]

The limit depends on direction from which \( \nu \) is approached!

\[ \omega \mathbf{u} \cdot G(\mathbf{E}) = \frac{1}{\nu^2} \mathbf{M} \mathbf{v}^2 \]

\[ \omega \mathbf{u} \cdot G(\mathbf{E}) = \frac{1}{\nu^2} \mathbf{M} \mathbf{v}^2 \]

\( \omega \mathbf{u} \cdot G(\mathbf{E}) \) is non-analytical at \( \nu = 0 \). This is the consequence of Bould summation being only conditionally convergent.

The type of summation we employed is when lattice grows in all 3 directions simultaneously. This implies that minimum \( \nu = \frac{\sqrt{3}}{L_3}, \quad \nu = \frac{\sqrt{3}}{L_2}, \quad \nu = \frac{\sqrt{3}}{L_1} \)

also have to change at the same time and at the same rate: \( \nu_1 - \nu_2 = \nu_2 \) (cube too)
In other words, it should change along the diagonal to be appropriate for the chosen summation method. But there's more than 1 diagonal in the $k$-space! 

Consider 1-D case:

$$A(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathcal{F}(k) e^{-i k x} \, dk = \frac{1}{2\pi} \left[ \int_{0}^{+\infty} \mathcal{F}(k) e^{i k x} \, dk + \int_{0}^{-\infty} \mathcal{F}(k) e^{-i k x} \, dk \right]$$

$$\lim_{k \to 0} \frac{\int_{-\infty}^{\infty} \mathcal{F}(k) e^{-i k x} \, dk}{\delta k = 0}$$

$$l(0) = \lim_{k \to 0} \frac{1}{\delta k} \left( A(0) + \frac{1}{2} \mathcal{F}(0) e^{-i k x} \right)$$

$$l(0) \text{ is analytical}$$

$$\lim_{k \to 0} \frac{1}{\delta k} \left( A(0) + \frac{1}{2} \mathcal{F}(0) e^{-i k x} \right)$$

The limit has to be taken on all equivalent directions and the result averaged.

Baumketner, BioSim, Lviv 2019
So there are 8 diagonals to reach $E \rightarrow 0$. Set $p$ a varying parameter

$$E = p / \sqrt{3} (1, 1, 1, 1)$$

would give all 8 diagonals.

For instance,

$$E_1 = \frac{p}{\sqrt{3}} (1, 1, 1), \quad v_1^2 = p^2, \quad \lim_{E \rightarrow 0} G(E_1) = \frac{1}{3} \sqrt{3} (m_1^2 + m_2^2 + m_3^2)$$

Taking all 8 limits,

$$G(0) = \lim_{E \rightarrow 0} \frac{1}{8} \sum_{i=1}^{8} G(E_i) = \frac{1}{3} \sqrt{3} (m_1^2 + m_2^2 + m_3^2)$$

This gives for the background energy:

$$E_4 = \frac{25}{2 \sqrt{3}} (\sum q_i E_i )^2 + \sum q_i \sum \frac{1}{2} \epsilon_{ij} \epsilon_{ik} e^{x \frac{-k^2}{2}} \left| l_i (E_i) \right|^2$$

Putting all terms together:

$$E_1 = \frac{1}{2} \sum_{i,j,k} \epsilon_{ij} \epsilon_{ik} \left| \sum_q \left( \frac{q_i + q_k}{2} \right) \right|^2$$

$$E_2 = \frac{1}{2} \sum_{i,j,k} \epsilon_{ij} \epsilon_{ik} \left| \sum_q \frac{q_i + q_k}{2} \right|^2$$

$$E_3 = \frac{1}{2} \sum_{i,j,k} \epsilon_{ij} \epsilon_{ik} \left| \sum_q \frac{q_i + q_k}{2} \right|^2$$

$$E_4 = \frac{25}{2 \sqrt{3}} (\sum q_i E_i )^2 + \sum q_i \sum \frac{1}{2} \epsilon_{ij} \epsilon_{ik} e^{x \frac{-k^2}{2}} \left| l_i (E_i) \right|^2$$

$$E_5 = \frac{25}{2 \sqrt{3}} \left( \sum q_i E_i \right)^2$$

$$E_0 = \frac{25}{2 \sqrt{3}} \left( \sum q_i E_i \right)^2$$

Baumketner, BioSim, Lviv 2019
assume that the stimulated system is considered in
a dispersive medium and not in vacuum.
Periodic images of the central cell will
gradually fill up space on approximately
spherical shapes. They will generate reaction
fields which will contribute to the solvation
tree energy of the central cell \( \Delta G \). Since on
the coarse limit the central cell will shrink
to a point, one can use dipole approximation:

\[
\Delta G = - \int \delta \mathbf{d} \cdot \mathbf{E}
\]

\[
\mathbf{E} = - \frac{1}{E^3} \frac{2(1-\epsilon)}{2\epsilon+1} \mathbf{M}_t
\]

\[
\mathbf{M}_t = \sum \mathbf{M}_q \ (r_q \cdot \mathbf{R}_n) = \mathbf{M} \cdot \mathbf{R}
\]

where \( \mathbf{M} \) is the dipole moment of the infinite lattice
encapsulates inside the sphere.

\[
\frac{4}{3} \pi R^3 = \text{volume of the sim. cell.}
\]

\[
\mathbf{E} = -\frac{4\pi}{3} \frac{1-\epsilon}{2\epsilon+1} \mathbf{R}
\]

\[
\Delta G = -\frac{4\pi}{3} \frac{1-\epsilon}{2\epsilon+1} \mathbf{R}^2
\]
The solvation energy can be combined with the $\tilde{\nu} = 0$ entry to give

$$\tilde{\nu} = \frac{1}{2} \sum_{i \neq j}^{N} \xi_{ij}^{SC} (\tilde{\nu}_{ij} + \tilde{\nu}_{ji}) + \sum_{i} \tilde{\nu}_{i} \sum_{k \neq i} \frac{1}{\epsilon_{k}} e^{-\frac{L_{i}^{2}}{\epsilon_{k}}} \left| L_{i} \right|^{2} - \sum_{i} \tilde{\nu}_{i} \frac{\epsilon_{i}}{\epsilon_{i}^{2}} \frac{\partial}{\partial \tilde{\nu}_{i}}$$

Boundary conditions:

1) "Adjusted" $\epsilon$ should be set equal to the detector constant of the simulated material

2) Vacuum $\epsilon = 1$

3) "Thin foil" or metallic $\epsilon = + \infty$. No solvation or $\tilde{\nu} = 0$ term

Most often used. Which boundary is better is still debated.
Accuracy and complexity

one adjustable parameter - L

The error is controlled by how many terms are retained in the real and inverse parts. Assume \( \rho_c \) cutoff is introduced for the real part, and \( \rho_c \) for the inverse part. \( \rho_c = 2 \pi / L \cdot n_c \). The error can then be estimated:

\[
\begin{align*}
\delta \rho_R & \approx Q \sqrt{\frac{\rho_c}{\beta L^2}} \left( \frac{1}{\rho_c} \right)^n e^{-\rho_c K^2} \\
\delta \rho_I & \approx Q \sqrt{\frac{\rho_c}{\beta L^2}} \left( \frac{1}{\rho_c} \right)^n e^{-\left( \frac{\pi \rho_c}{\beta} \cdot L \right)^2}
\end{align*}
\]

Both \( \delta \rho_R \) and \( \delta \rho_I \) are determined by the same function \( e^{-x^2} \). Let's impose an accuracy \( \varepsilon = \frac{e^{-s^2}}{s^2} \) on both terms.

\[
\begin{align*}
\rho_c = \frac{S}{\beta}, \quad n_c = \frac{s \cdot L \cdot \rho}{\pi} \rightarrow \delta \rho_R & \approx Q \sqrt{\frac{S}{\beta L^2}} e^{-S^2 / S^2}, \quad \delta \rho_I = Q \sqrt{\frac{S}{2 \beta L^2}} e^{-s^2 / s^2}
\end{align*}
\]

These two quantities need to be fixed areas of time.
Total computational expense:

\[ \tau = \frac{\tau_k \cdot \mu}{\tau_F \cdot L^6} \]

\( \mu_k \) - number of pair interactions that are evaluated within \( \tau_k \)

\[ \mu_k = \mu_c \cdot \mu = \frac{4}{3} \cdot \frac{\pi^3}{\beta} \cdot \frac{s^3 m^2}{\beta^3 L^3} \]

\( \mu_F \) - number of terms in the Fourier series with \( \mu_c \) cut-off

\[ \mu_F = \mu_c \cdot \mu = \frac{4}{3} \cdot \frac{\pi^3}{\beta} \cdot \frac{s^3 m^2}{\beta^3 L^3} \cdot \mu_f \]

Maximize \( \tau(\beta) \):

\[ \frac{\partial \tau}{\partial \beta} = 0 \]

\[ L = \left( \frac{\tau_k \cdot \mu}{\tau_F \cdot L^6} \right)^{1/6} \quad \tau = \frac{8 \sqrt{\tau_k \cdot \tau_\infty} \cdot m^2}{3 \sqrt{\pi}} = O \left( L^{3/2} \right) \]

\( \tau_k / \tau_F \) ratio can be determined in short simulations.
In canonical ensemble:

\[ \rho_{NVT}(\Gamma) \sim e^{-\beta H(\Gamma)}, \beta = \frac{1}{kT} \]

\[ Q_{NVT} = \frac{1}{h^{3N} N!} \int d\vec{p} d\vec{q} e^{-\beta H(\Gamma)} = Q_{NVT}^{id} \times Q_{NVT}^{ex} \]

\[ Q_{NVT}^{id} = \frac{V^N}{N! \lambda^{3N}}, \lambda = \sqrt{\frac{\hbar^2}{2\pi m kT}} \]

Partition function:

\[ Q_{NVT}^{ex} = \frac{Z_{NVT}}{V^N}, Z_{NVT} = \int d\vec{q} e^{-\beta U(q)} \]

Free energy splits into two parts as well:

\[ F = -kT \log Q_{NVT} = F^{id} + F^{ex} \]

Gibbs free energy:

\[ G = F + PV = F^{id} + F^{ex} + (P^{id} + P^{ex})V = G^{id} + G^{ex} \]

Chemical potential:

\[ \mu = \frac{G}{N} = \mu^{id} + \mu^{ex} \]

\[ \mu^{id} = kT \log(\lambda^3) + kT \log(\rho) = kT \log(\frac{\rho}{n_Q}) \]

\[ n_Q = \lambda^{-3} \]

\[ \mu^{ex} = U^{ex} - TS^{ex} + P^{ex}V \]

Baumketner, BioSim, Lviv 2019
Assume the following approximation for the ions solvated near a charged wall:

\[ \mu = U^e + \mu^d = q\varphi + kT \log \left( \frac{\rho}{n_Q} \right) \]

The total interaction energy is approximated by the electrostatic potential. The potential needs calibration.

Consider a system of ions confined between two surfaces. All properties depend on coordinate \( x \).

In the state of equilibrium, or more generally stationary state, the chemical potential should not depend on \( x \) to avoid exchange of particles between different parts of the system.

\[ \mu(x) = \text{const} \]

\[ q\varphi(x) + kT \log \left( \frac{\rho(x)}{n_Q} \right) = \text{const} \]

\[ \rho(x) = \rho^0 e^{-\beta q\varphi(x)} \]

Boltzmann distribution prescribing how density of ions will change depending on the potential

\[ \varphi(x = 0) = 0 \]
\[ \rho(x = 0) = \rho^0 \]

Calibration conditions

Baumketner, BioSim, Lviv 2019
Density and the potential are connected by laws of electrostatics:

\[ \vec{V} \overrightarrow{D}(x) = 4\pi \rho_c(x) \]

*displacement created by charge distribution*

*charge distribution*

In polarizable media:

\[ \overrightarrow{D}(x) = \varepsilon(x)\overrightarrow{E}(x) = -\varepsilon(x)\vec{V}\varphi(x) \]

For vacuum:

\[ -\vec{V}\overrightarrow{E}(x) = \Delta \varphi = -4\pi q \rho(x) \]

\[ \frac{d^2 \Delta \varphi}{dx^2} = -4\pi q \rho_0 e^{-\beta q \varphi(x)} \]

*Poisson-Boltzmann (PB) equation*

\[ \varphi(x = 0) = 0 \quad \sigma_s = \text{const} \]

*constant charge boundary condition*

\[ \sigma_s = f(\varphi) \]

*constant "potential" boundary condition*

PB gives ion density, potential and electric field at any point within the system

Baumketner, BioSim, Lviv 2019
How well does the PB model work?

Analytical solution:

\[ \rho(x) = \rho_0 e^{-q_\beta \varphi(x)} = \rho_0 / \cos^2(Kx) \]

\[ -\frac{2K}{\beta q} \tan\left(\frac{KD}{2}\right) = \sigma_s \text{ equation for } K \]

The agreement is remarkably good for concentrations in the range up to 16M!

Extension to mixtures

Assume that we have a mixture of ions with varying valency. The condition on the constancy of the chemical potential has to be satisfied for each component.

\[ \mu_i = q_i \varphi + kT \log\left(\frac{\rho_i}{nQ}\right) \rightarrow \rho_i(x) = \rho_i^0 e^{-\beta q_i \varphi(x)} \]
Poisson equation

\[ \Delta \varphi = -4\pi \sum_i q_i \rho_i(x) \quad \rightarrow \quad \Delta \varphi = -4\pi \sum_i q_i \rho_i^0 e^{-\beta q_i \varphi(x)} \]

PB equation

Linearize the RHS

\[ e^{-\beta q_i \varphi(x)} = 1 - \beta q_i \varphi(x) + \cdots \]

\[
4\pi \sum_i q_i \rho_i^0 e^{-\beta q_i \varphi(x)} = 4\pi \sum_i q_i \rho_i^0 (1 - \beta q_i \varphi(x) + \cdots) = 4\pi \sum_i q_i \rho_i^0 - 4\pi \sum_i q_i^2 \rho_i^0 \beta \varphi(x) + \cdots
\]

\[ \approx -\kappa^2 \varphi \]

\[ \kappa = \sqrt{\frac{4\pi \beta}{\sum_i q_i^2 \rho_i^0}} \]

inverse screening length

After putting everything together:

\[ \Delta \varphi - \kappa^2 \varphi = 0 \]

linear PB (LPB) for multicomponent systems

Zero for electrically neutral systems

Baumketner, BioSim, Lviv 2019
Limitations of the PB model

The key approximation: \[ \mu = U^{ex} + \mu^{id} = q\varphi + kT \log\left(\frac{\rho}{n_Q}\right) \]

1) The total energy is approximated by the electrostatic component only. What’s neglected:
   a) other energy contributions, excluded volume, vDW etc. \textbf{The size of the ions} is missing
   b) the electrostatic energy is included at the mean-field level. Approximated by the average value. \textbf{Ion-ion correlations} are missing

2) Full entropy is approximated by its ideal part. Effect of particle-particle interactions on the excess entropy is neglected. In particular:
   a) \textbf{steric effects} are missing

3) Discreetness of surface charge. May contribute additional attraction when discrete surface charges are mobile.
How can PB model be used for biomolecules?

Extension of the Poisson equation to multiple media with fixed and mobile charges:

\[ \nabla \cdot \mathbf{D} = 4\pi (\rho_f + \rho_m) \]

fixed charges

mobile ions

charge density

fixed charges

Electric field in the continuum approximation:

\[ \mathbf{D}(\mathbf{r}) = \varepsilon(\mathbf{r}) \mathbf{E}(\mathbf{r})\]

\[ \nabla (\varepsilon(\mathbf{r}) \mathbf{E}(\mathbf{r})) = -4\pi (\rho_f(\mathbf{r}) + \rho_m(\mathbf{r})) \]

\[ \nabla (\varepsilon(\mathbf{r}) \nabla \varphi(\mathbf{r})) = -4\pi (\rho_f(\mathbf{r}) + \sum_i q_i \rho_i^0 e^{-\beta q_i \varphi(\mathbf{r})}) \]

\[ \nabla (\varepsilon(\mathbf{r}) \nabla \varphi(\mathbf{r})) - \kappa^2 \varphi(\mathbf{r}) = -4\pi \rho_f(\mathbf{r}) \]

LPB equation

Baumketner, BioSim, Lviv 2019
LPB can be solved numerically subject to specific boundary conditions (constant charge):

\[
\begin{align*}
\varepsilon_I \Delta \varphi_I (\vec{r}) &= -4\pi \rho_f (\vec{r}) \\
\varepsilon_{II} \Delta \varphi_{II} (\vec{r}) - \kappa^2 \varphi_{II} (\vec{r}) &= 0
\end{align*}
\]

Continuity of potential

\[ \varphi_I \bigg|_{\Omega} = \varphi_{II} \bigg|_{\Omega} \]

Displacement

\[ \varepsilon_I \frac{\partial \varphi_I}{\partial n} \bigg|_{\Omega} = \varepsilon_{II} \frac{\partial \varphi_{II}}{\partial n} \bigg|_{\Omega} \]

Boundary between two media: \( \Omega \)

Computation of the charging free energy

Recall that the potential in solvent (continuum approximation) is:

\[ U_M (\Gamma) = U_V (\Gamma) + \Delta G (\Gamma) \]

where \( \Delta G (\Gamma) \) is the free energy associated with turning the solvent “on”.

\[ U_V (\Gamma) \] electrostatic energy in vacuum

Also recall that \( \Delta G (\Gamma) = \Delta G_{ch} (\Gamma) - U_V (\Gamma) \)

\[ \Delta G_{ch} (\Gamma) = \int dq \, \varphi \]

charging free energy. Work needed to create charge in a medium

Baumketner, BioSim, Lviv 2019
Since the PBL equation is linear:

\[ \Delta G_{ch}(\Gamma) = \frac{1}{2} \sum_i q_i \varphi_{sol}(i) \]

potential acting on charge qi obtained for particular solute in solvent with \( \epsilon = 80 \)

Potential energy in vacuum:

\[ U_V = \frac{1}{2} \sum_i q_i \varphi_{vac}(i) \]

potential acting on charge qi in vacuum or solvent with \( \epsilon = 1 \)

Combining the formulas:

\[ \Delta G(\Gamma) = \frac{1}{2} \sum_i q_i (\varphi_{sol}(i) - \varphi_{vac}(i)) = \Delta G_{PB} \]

this is the term that needs to be added to the potential energy in order to include the effect of the solvent.

The approach that combines certain force-field with PB equation for solvation energy is known as PB/MM model.
Limitations of the PB/MM model

1) All that apply to PB equation.

2) Error due to continuum approximation for water. Model breaks down on length scales comparable to the size of water molecule.

3) Missing non-polar solvation forces

4) High computational cost

Fine structure of the potential is due to the finite size of water molecules

All implicit solvent models are missing local minima, including GB which is parameterized against PB
Numerical solutions of PB for real molecular shapes are very costly. Much faster implicit solvation models are needed.

**Generalized Born (GB) model**

Draw an analogy with the Born solvation energy:

\[
\Delta G_{\text{Born}} = \frac{1}{2} \left( \frac{1}{\varepsilon} - 1 \right) \frac{q^2}{a}
\]

Approximate solvation energy for a molecule:

\[
\Delta G_{\text{GB}} = \frac{1}{2} \left( \frac{1}{\varepsilon} - 1 \right) \sum_{i,j} \frac{q_i q_j}{f_{ij}}
\]

Empirical function:

\[
f_{ij} = \sqrt{r_{ij}^2 + R_i R_j e^{-r_{ij}^2/4R_i R_j}}
\]

Still's formula

\[
R_i \quad \text{is the effective Born radius for atom } i
\]

If there's only one atom in the system:

\[
f_{ij} = R_i, \quad \Delta G_{\text{GB}} = \frac{1}{2} \left( \frac{1}{\varepsilon} - 1 \right) \frac{q_i^2}{R_i} = \Delta G_{\text{Born}}
\]

Born energy for particle with radius \( R_i \)

For two charges at a large separation: \( r_{12} \geq R_1, r_{12} \geq R_2, \quad f_{ij} \approx r_{ij} \)

\[
\Delta G_{\text{GB}} = \frac{1}{2} \left( \frac{1}{\varepsilon} - 1 \right) \frac{q_1^2}{R_1} + \frac{1}{2} \left( \frac{1}{\varepsilon} - 1 \right) \frac{q_2^2}{R_2} + \left( \frac{1}{\varepsilon} - 1 \right) \frac{q_1 q_2}{r_{12}}
\]

Baumketner, BioSim, Lviv 2019
Full electrostatic interaction then is:

\[ U_{ele}(r_{12}) = \frac{q_1 q_2}{r_{12}} + \Delta G_{GB} = \frac{1}{2} \left( \frac{1}{\epsilon} - 1 \right) \frac{q_1^2}{R_1} + \frac{1}{2} \left( \frac{1}{\epsilon} - 1 \right) \frac{q_2^2}{R_2} + \frac{q_1 q_2}{\epsilon r_{12}} \]

Self-energy of the two ions  
Coulomb interaction in a continuum

To retain their physical meaning, Born radii are introduced via charging free energy:

\[ R_i = \frac{2 \Delta G_{ch}}{q_i^2} \frac{\epsilon}{1 - \epsilon} \]

No atoms except number \( i \) have charge
Coulomb field approximation for the Born radii

Another way to compute electrostatic work is via scalar product of the displacement vector with the vector of electric field:

$$W = \frac{1}{8\pi} \int \mathbf{E} \cdot \mathbf{D} \, dV$$

Assume that the displacement vector created by charge $q_i$ retains its Coulomb form even outside of the solute molecule:

$$\mathbf{D}_i \approx \frac{q_i \mathbf{r}}{r^3}$$

Work needed to move charge $q_i$ from infinity to its proper location:

$$W = \frac{1}{8\pi} \int (\mathbf{D}/\varepsilon) \cdot \mathbf{D} \, dV \approx \frac{1}{8\pi} \int_{\text{in}} \frac{q_i^2}{r^4} \, dV + \frac{1}{8\pi} \int_{\text{out}} \frac{q_i^2}{r^4\varepsilon} \, dV$$

To compute electrostatic solvation energy one needs to subtract the same quantity evaluated at

$$\varepsilon = 1 \quad \rightarrow \quad \Delta G_{ch} = -\frac{1}{8\pi} \left( 1 - \frac{1}{\varepsilon} \right) \int_{\text{out}} \frac{q_i^2}{r^4} \, dV$$

One then arrives at the following formula:

$$R_i^{-1} = \frac{1}{4\pi} \int_{\text{out}} r^{-4} \, dV$$

Baumketner, BioSim, Lviv 2019
The Born radius in CF approximation can be evaluated exactly for the spherical geometry. The integral can be taken analytically to yield:

\[ R_{CF} = \frac{4\pi}{\int_{\text{out}} r^{-4} dV} = 2a \left( \frac{1}{1-p^2} + \frac{1}{2p} \ln \frac{1+p}{1-p} \right)^{-1}, \quad p = \frac{r_s}{a} \]

The spherical geometry has an exact solution. In particular, the reaction field is a sum of the field created by Kelvin image and a charged ray:

\[ \varphi_{RF}(\vec{r}) = \varphi_K(\vec{r}) + \varphi_{\text{line}}(\vec{r}) \]

Kelvin charge image. \( \vec{r}_K = \frac{q^2}{r_s} \quad q_K = \gamma \frac{a \cdot q}{r_s} \quad \gamma = \frac{1 - \epsilon}{1 + \epsilon} \)

Reaction filed at the location of the source charge:

\[ \varphi_{RF}(\vec{r}_s) = \frac{q_K}{r_K - r_s} = \frac{1 - \epsilon}{1 + \epsilon} aq \]

Charging free energy:

\[ \Delta G_{ch} = \int dq \varphi_{RF} = \frac{1}{2} \frac{1 - \epsilon}{1 + \epsilon} aq^2 \frac{1}{a^2 - r_s^2} = \frac{1}{2} \left( \frac{1}{\epsilon} - 1 \right) \frac{q^2}{R_K} \]

Baumketner, BioSim, Lviv 2019
GB radius in the Kelvin approximation then is:

\[ R_K = \frac{1 + \varepsilon}{\varepsilon} a(1 - p^2) \approx a(1 - p^2) \]

Good approximation for water with \( \varepsilon = 80 \)

Exact radius for arbitrary dielectric constant:

\[ \frac{1}{R_i} = \frac{1 - \gamma}{2} \left( \frac{1}{a(1 - p^2)} + \frac{1}{a} \sum_{n=0}^{\infty} \frac{(1+\gamma)p^{2n}}{1 - \gamma + 2n} \right) \]

Kelvin solution also leads to a better expression for the Born energy of multi-particle systems:

\[ f_{ij} = \sqrt{r_{ij}^2 + R_i R_j} \quad \text{Grycuk's formula} \]

\[ f_{ij} = \sqrt{r_{ij}^2 + R_i R_j e^{-r_{ij}^2/4R_i R_j}} \quad \text{Still's formula} \]

For atoms near the molecules boundary, the CF approximation overestimates the Born radius 2 times.
To evaluate Born radius for non-spherical objects, an empirical formula is proposed:

\[
\frac{1}{R} = \left( \frac{3}{4\pi} \int_{\text{excl}} \frac{1}{r^6} \, dV \right)^{1/3}, \quad \text{this leads to} \quad \frac{1}{a(1 - p^2)} \quad \text{for a spherical molecule}
\]

This is still an approximation. Works exactly only for a sphere in high-dielectric constant medium. Perform tests for everything else!

Uniformly good performance for any location of the source charge. Deterioration of quality for charges buried inside the molecule.

Numerical PB \(1/R6 GB\)

J. Chem. Phys., Vol. 119, No. 9, 1 September 2003

Baumketner, BioSim, Lviv 2019
Comparison for varying dielectric constant in the context of a protein

1) CFA is not accurate for any eps

2) R6 is accurate for interior dielectric of 1 but much worse for dielectric=20

3) Small shift of inverse $R$ can improve the agreement with PB a lot. Reason – unknown

$$\frac{1}{R_i} = \frac{1}{R_i} + 0.028 \, \text{Å}^{-1}$$
Non-polar molecules tend to minimize exposure to solvent. As a result, configurations with minimal surface area are assumed. Solvation free energy is assumed to be proportional to the surface area exposed to solvent:

\[ \Delta G_{np} = \gamma (SA) + b \]

- \( \gamma \) is an empirical parameter. May depend on the curvature of the solvent molecule. Typical value \( \gamma = 72 \text{ cal mol}^{-1} \text{Å}^{-1} \)

1) Parametrizations by many groups (Sitkoff, Sheraga etc.)

2) Only part or the whole molecule can be treated as non-polar. Separate parameters can be introduced for polar and non-polar groups.

3) Several definitions of “accessible surface” are in use.

\[ \Delta G = \Delta G_{PB/GB} + \gamma \Delta S \]

common models PB/SA or GB/SA

SASA constructed as the surface area of a figure swept by the center of a probe molecule rolled on the surface of the solute molecule

\[ R_p = 1.4 \text{Å} \]

typically used for water

\[ \text{vdW surface} = \text{SASA} \]

with the probe radius set to zero

Molecular surface is made by the points of closest approach of the probe to the solute
Weighted histogram analysis method (WHAM)

Assume that we have a number $K$ of different simulations, each performed at a separate temperature $\beta_k, k = 1, K$

The number of sampled conformations for each temperature is $N$

Each simulation produces an energy histogram $h_k(E)$

where the bin size is $\Delta E$

$$\sum_E h_k(E) = N$$

The energy can be obtained at a discrete set of temperatures $< E > \beta_k$

Q: How do we compute energy (and other functions) at intermediate temperatures? Is there new information hidden in these data?
Sampled histograms can be used to estimate density of states:

\[ h_k(E') = N n_k(E') e^{-\beta_k(E - f_k)} \]

\[ n_k(E) = \frac{h_k(E)}{N} e^{\beta_k(E - f_k)} \]

density of states.  
Most accurate for energy levels close to \(< E >_{\beta_k}

\[ f_k \] is the free energy at temperature \( \beta_k \)

Let us combine histograms at all temperatures in order to obtain a more accurate estimate of the density of states:

\[ n(E) = \sum_{i=1}^{K} w_i n_i(E) \]

some weight coefficients for each temperature

\[ \sum_{i=1}^{K} w_i = 1 \]

normalization condition that the coefficients needs to satisfy

If the density of states is known, energy distribution at any temperature can be computed as follows

\[ P_\beta(E) = \frac{n(E)e^{-\beta E}}{\sum_E n(E)e^{-\beta E}} \]

Baumketner, BioSim, Lviv 2019
How to compute the weight coefficients?

Let us estimate the error in the density of states and try to minimize it. This will produce an optimal set of weights.

If we perform multiple experiments = sets of measurements, the resulting energy histograms will fluctuate. This will lead to fluctuations in the estimate of the density of states.

\[
n_k(E) = \frac{h_k(E)}{N} e^{\beta_k(E-f_k)} \rightarrow \delta n_k(E) = \frac{\delta h_k(E)}{N} e^{\beta_k(E-f_k)}
\]

The average of fluctuations across many experiments is zero

\[
< \delta n_k(E) >_{\text{exp}} = 0
\]

Fluctuation of the weighted density of states:

\[
\delta n(E) = \sum_{i=1}^{K} w_i \delta n_i(E) \rightarrow < \delta n(E) > = \sum_{i=1}^{K} w_i < \delta n_i(E) > = 0
\]
The error can be estimated from fluctuation squared:

\[
\left( \delta n(E) \right)^2 = \sum_{i=1,j=1}^{K} w_i w_j \delta n_i(E) \delta n_j(E) \quad \rightarrow \quad < \left( \delta n(E) \right)^2 > = \sum_{i=1,j=1}^{K} w_i w_j < \delta n_i(E) \delta n_j(E) > =
\]

\[
\sum_{i=1}^{K} w_i^2 < \left( \delta n_i(E) \right)^2 >
\]

since measurements at different temperatures are uncorrelated

Average fluctuations in the density of states at fixed temperatures can be estimated as follows:

\[
\left( \delta n_k(E) \right)^2 = \frac{\left( \delta h_k(E) \right)^2}{N^2} e^{2\beta_k(E-f_k)} \quad \rightarrow \quad < \left( \delta n_k(E) \right)^2 > = \frac{< \left( \delta h_k(E) \right)^2 >}{N^2} e^{2\beta_k(E-f_k)}
\]

Let us rewrite the squared fluctuation of the histogram in explicit form:

\[
< \left( \delta h_k(E) \right)^2 > = < h_k^2(E) > - < h_k(E) >^2
\]

By definition:

\[
h_k(E) = \sum_{i=1}^{N} \delta_{E,E_i} \quad \text{where} \quad \delta_{E,E_i} = \frac{1}{2} \left( \Theta(E_i - E) + \Theta(E + \Delta E - E_i) \right) = \begin{cases} 1 & \text{if } E \leq E_i \leq E + \Delta E \\ 0 & \text{otherwise} \end{cases}
\]

Baumketner, BioSim, Lviv 2019
The square then is

\[
< (h_k(E))^2 > = \sum_{i=1, j=1}^{N} \delta_{E_i} \delta_{E_j} = \sum_{i=1}^{N} \delta_{E_i}^2 + \sum_{i \neq j} \delta_{E_i} \delta_{E_j}
\]

under the assumption that conformations \(i\) and \(j\) are independent

\[
< h_k(E) > = Ng
\]

For canonical distribution, quantity \(g\) can be evaluated directly:

\[
g = \frac{\int_{E}^{E+\Delta E} \delta_{E_i} n(E) e^{-\beta E} dE}{\int n(E) e^{-\beta E}} \approx n(E) e^{-\beta (E-F(\beta))} \Delta E
\]

so

\[
\lim_{\Delta E \to 0} g = 0
\]

Number that shows how likely a random conformation at temperature \(\beta\) is to have energy \(E \in [E + \Delta E]\)

Average square:

\[
< h_k(E)^2 > = \sum_{i=1}^{N} \delta_{E_i} < \sum_{j=1}^{N} \delta_{E, j} > = N^2 g^2
\]

Baumketner, BioSim, Lviv 2019
The difference then is

\[
< (\delta h_k(E))^2 > = < h_k^2(E) > - < h_k(E) >^2 = N g + N(N - 1)g^2 - N^2 g^2 = N g - N g^2
\]

\[\approx N g = < h_k(E) > \quad \text{if} \quad g^2 \ll g \quad \text{which can always be achieved by making an appropriate choice for } \Delta E
\]

Going back to the estimate of error in the density of states:

\[
< (\delta n_k(E))^2 >= \frac{< (\delta h_k(E))^2 >}{N^2} e^{2\beta_k(\textbf{E} - f_k)} = \frac{< h_k(E) >}{N^2} e^{2\beta_k(\textbf{E} - f_k)}
\]

The best estimate of the average histogram can be obtained from the improved density of states:

\[
<h_k(E)> = N n(E) e^{-\beta_k(\textbf{E} - f_k)} = Ne^{-\beta_k(\textbf{E} - f_k)} \sum_i^N w_i n_i(\textbf{E})
\]

\[
< (\delta n_k(E))^2 >= \frac{n(E)}{N} e^{\beta_k(\textbf{E} - f_k)}
\]

The error in the improved density of states finally:

\[
< (\delta n(E))^2 > = \sum_{i=1}^K w_i^2 < (\delta n_i(\textbf{E}))^2 > = \sum_{i=1}^K w_i^2 \frac{n(E)}{N} e^{\beta_i(\textbf{E} - f_i)}
\]

Baumketner, BioSim, Lviv 2019
Let us consider the relative error in the density of states estimate:

\[ F(w_1, \ldots, w_K) = \frac{<(\delta n_k(E))^2>}{n(E)} + \lambda \sum_{i=1}^{K} w_i = \sum_{i=1}^{K} w_i^2 \frac{1}{N} e^{\beta_i(E-f_i)} + \lambda \sum_{i=1}^{K} w_i \]

Minimize the cost function:

\[ \frac{\partial F(w_1, \ldots, w_K)}{\partial w_l} = \frac{2w_l}{N} e^{\beta_l(E-f_l)} + \lambda = 0 \quad l = 1, K \]

\[ w_l = -\frac{\lambda N}{2} e^{-\beta_l(E-f_l)} \]

Use the constraint to determine \( \lambda \)

\[ \sum_{l=1}^{K} w_l = -\frac{\lambda N}{2} \sum_{l=1}^{K} e^{-\beta_l(E-f_l)} = 1 \quad \rightarrow \quad -\frac{\lambda N}{2} = \frac{1}{\sum_{l=1}^{K} e^{-\beta_l(E-f_l)}} \quad \rightarrow \quad w_l = \frac{e^{-\beta_l(E-f_l)}}{\sum_{l=1}^{K} e^{-\beta_l(E-f_l)}} \]
The best estimate of the density of states:

\[ n(E) = \sum_{i=1}^{K} w_i n_i(E) = \frac{\sum_{i=1}^{K} n_i(E) e^{-\beta_i(E-f_i)}}{\sum_{i=1}^{K} e^{-\beta_i(E-f_i)}} = \frac{\sum_{i=1}^{K} n_i(E) e^{-\beta_i(E-f_i)}}{\sum_{i=1}^{K} e^{-\beta_i(E-f_i)}} = \frac{\sum_{i=1}^{K} p_i(E)}{\sum_{i=1}^{K} e^{-\beta_i(E-f_i)}} \]

Density of states can be evaluated from energy histograms at all temperatures and free energies.

By definition:

\[ e^{-\beta f_i} = \sum_E n(E) e^{-\beta_i E} \]

Then one obtains a set of coupled equations:

\[ n(E) = \frac{\sum_{i=1}^{K} p_i(E)}{\sum_{i=1}^{K} e^{-\beta_i(E-f_i)}} \quad (1) \]

\[ e^{-\beta f_i} = \sum_E n(E) e^{-\beta_i E} \quad (2) \]

WHAM equations can be solved by iterations:

Step 1: Adopt some values for the free energies \( f_1 \ldots f_K \)

Step 2: Compute the density of states using eq. (1)

Step 3: Obtain more accurate free energy from eq. (2)

Step 4: Go back to step 1. Continue until convergence.

Outcome:

1) Relative free energy for a set of temperatures. \( f_1 \) has to be fixed.

2) Density of states so energy dist. for any temperature

Pitfalls:

1) Temperatures have to be narrowly spaced for energy distributions to overlap

2) Energy distributions have to be converged. Problems at low temperature may arise in some systems

Reweighting:

\[ P_\beta(E) = \frac{n(E) e^{-\beta E}}{\sum_E n(E) e^{-\beta E}} \]

Baumketner, BioSim, Lviv 2019
Two-dimensional WHAM

Imagine that it’s some structural parameter \( X \) for which we seek temperature dependence.

To obtain reweighted function one needs to consider joint distribution of that parameter with \( E \).

\[
P_i(E, X) = \frac{n_i(E, X)e^{-\beta_i E}}{\sum_{E, X} n_i(E, X)e^{-\beta_i E}} = n_i(E, X)e^{-\beta_i(E-f_i)} \quad \rightarrow \quad n_i(E, X) = P_i(E, X)e^{\beta_i(E-f_i)}
\]

\[
\sum_{E, X} P_i(E, X) = 1
\]

2D density of states

Free energy

\[
P_i(X) = \sum_E P_i(E, X)
\]

Following the WHAM procedure, build a better estimate of the density of states:

\[
n(E, X) = \sum_i w_i n_i(E, X) \quad \rightarrow \quad w_l = \frac{e^{-\beta_l(E-f_l)}}{\sum_{k=1}^K e^{-\beta_k(E-f_k)}}
\]

2D WHAM equations:

\[
n(E, X) = \frac{\sum_i^K P_i(E, X)}{\sum_i^K e^{-\beta_i(E-f_i)}}
\]

\[
e^{-\beta_i f_i} = \sum_{E, X} n(E, X)e^{-\beta_i E}
\]

Distribution function at any temperature by reweighting:

\[
P_\beta(X) = \frac{\sum_E n(E, X)e^{-\beta E}}{\sum_{E, X} n(E, X)e^{-\beta E}}
\]

\[
<X>_{\beta} = \sum_X XP_\beta(X), \quad <X^2>_{\beta} = \sum_X X^2 P_\beta(X)
\]

<transition point>

Baumketner, BioSim, Lviv 2019
Umbrella sampling

The idea of combining multiple trajectories can be used to obtain distributions along selected degrees of freedom with the help of biased simulations.

Recall that for a degree of freedom $X$:

$$P(X) = \sum_E P(E, X) = \frac{\sum_E n(E, X) e^{-\beta E}}{\sum_{E, X} n(E, X) e^{-\beta E}} = \frac{P_0(X)}{\sum_X P_0(X)} = P_0(X) e^{\beta F(\beta)}$$

Not normalized distribution function

$$P_0(X) = \sum_E n(E, X) e^{-\beta E}$$

$$\sum_X P_0(X) = e^{-\beta F}$$

Imagine that we apply external potential to bias the value of coordinate $X$ sampled in simulations

$$E \rightarrow E + \frac{\alpha}{2} (X - X_i)^2$$

biasing "umbrella" potential

Distribution in the umbrella simulation:

$$P_i(X) = \sum_E n(E, X) e^{-\beta E} e^{-\beta \frac{\alpha}{2} (X - X_i)^2} e^{f_i} = P_0(X) e^{-\beta \frac{\alpha}{2} (X - X_i)^2} e^{f_i}$$

$$\sum_X P_i(X) = 1$$

some normalization constant specific to $X_i$

Baumketner, BioSim, Lviv 2019
Unbiased distribution can be recovered as follows:

\[ P^i_0 (X) = P_i (X) e^{\frac{\alpha}{2} (X - x_i)^2} e^{-f_i} \rightarrow \sum X P^i_0 (X) e^{-\frac{\beta}{2} (X - x_i)^2} = e^{-f_i} \]

this is normalized histogram from biased simulation

Let us now conduct multiple biasing simulations to obtain improved distribution:

\[ P_0 (X) = \sum_i w_i P^i_0 (X) \quad \sum_i w_i = 1 \]

Follow the WHAM procedure to estimate the error and then minimize it:

\[ (\delta P^i_0 (X))^2 = (\delta P_i (X))^2 e^{2 \frac{\alpha}{2} (X - x_i)^2} e^{-2f_i} = P_0 (X) e^{\frac{\alpha}{2} (X - x_i)^2} e^{-f_i} \]

\[ (\delta P_0 (X))^2 = \sum_i w_i^2 (\delta P^i_0 (X))^2 = \sum_i w_i^2 e^{\frac{\alpha}{2} (X - x_i)^2} e^{-f_i} P_0 (X) \]

The function to be minimalized:

\[ F = \sum_i w_i^2 e^{\frac{\alpha}{2} (X - x_i)^2} e^{-f_i} + \lambda \sum_i w_i \rightarrow w_l = \frac{e^{-\frac{\alpha}{2} (X - x_l)^2 + f_l}}{\sum e^{-\frac{\alpha}{2} (X - x_i)^2 + f_i}} \]

Baumketner, BioSim, Lviv 2019
WHAM equations:

\[ P_0(X) = \frac{\Sigma_i P_i(X)}{\Sigma_i e^{-\beta \frac{\alpha}{2}(x-x_i)^2 + f_i}} \]
\[ e^{-f_i} = \sum_X P_0(X) e^{-\beta \frac{\alpha}{2}(x-x_i)^2} \]

**Reweighting:**

\[ P(X) = \frac{P_0(X)}{\Sigma_X P_0(X)} \]

**Potential of mean force (PMF):**

\[ F(X) = -kT \log(P(X)) \]

\[ Z(X) = P(X)Z \]

Let \( X \) be some generalized coordinate. By definition:

\[ P(X) = \int \delta(X - X(\Gamma)) e^{-\beta U} d\Gamma Z^{-1} = \frac{\int e^{-\beta U} d\Gamma_X}{\int e^{-\beta U} d\Gamma} = \frac{Z(X)}{Z} \]

\[ d\Gamma = dXd\Gamma_X \]

\[ d\Gamma_X \quad \text{differential over variables other than } X \text{ (orthogonal variables)} \]

The average force that corresponds to the coordinate:

\[ <F_X>_X = \frac{-\int \frac{\partial U}{\partial X} e^{-\beta U} d\Gamma_X}{\int e^{-\beta U} d\Gamma_X} \]

\[ = \frac{1}{\beta} \int e^{-\beta U} d\Gamma_X \frac{\partial}{\partial X} \int e^{-\beta U} d\Gamma_X \]

\[ = \frac{1}{\beta Z(X)} \frac{\partial}{\partial X} Z(X) = -\frac{\partial}{\partial X} - kT \log(Z(X)) \]

\[ = -\frac{\partial}{\partial X} \{-kT \log(P(X)) - kT \log(Z)\} \]

\[ = -\frac{\partial}{\partial X} F(X) \]

function that can be used as generalized potential along degree of freedom \( X \)

free energy profile along one variable

"potential" that generates it
Free energy estimation

Key thermodynamic function that helps to describe stability of condensed matter systems

1) Phase equilibria, transitions

2) Binding strength for macro complexes …

By definition:

\[
F(N, V, T) = -kT\log(Q_{NVT}) = F_{\text{id}}(N, V, T) + F_{\text{ex}}(N, V, T)
\]

Helmholtz free energy

\[
Q_{NVT} = Q_{NVT}^{\text{id}} \times Q_{NVT}^{\text{ex}}, \quad Q_{NVT}^{\text{id}} = \frac{V^N}{N! \lambda^3 N}, \quad \lambda = \sqrt{\frac{\hbar^2}{2\pi m kT}}
\]

\[
F_{\text{id}}(N, V, T) = \frac{N}{\beta} \log(\rho) + \frac{N}{\beta} \log\left(\frac{1}{\lambda^3}\right) - \frac{N}{\beta}, \quad \rho = \frac{N}{V}
\]

\[
F_{\text{ex}}(N, V, T) = -kT\log(Q_{NVT}^{\text{ex}}), \quad Q_{NVT}^{\text{ex}} = \frac{Z_{NVT}}{V^N}, \quad Z_{NVT} = \int d\Gamma e^{-\beta U(\Gamma)}
\]

Not formulated as an average over ensemble so difficult to estimate in simulations
1) Hit and miss method

\[ Z_{NVT} \] is an integral in multidimensional space. The most efficient integration method is sample mean.

\[ Z_{NVT} = \int e^{-\beta U(\Gamma)} d\Gamma = \langle \frac{e^{-\beta U(\Gamma)}}{\rho(\Gamma)} \rangle \rho \int \rho(\Gamma) d\Gamma \]

sampling from the given distribution

The scheme can have multiple realizations:

1) Uniform distribution in the phase space \( \rho = \frac{1}{V^N} \), \( \int d\Gamma \rho(\Gamma) = 1 \)

\[ Z_{NVT} = V^N \langle e^{-\beta U(\Gamma)} \rangle \]

Configurations are generated by randomly displacing particles anywhere in the available volume \( V \)

Due to overlaps between particles, very few entries in this sum will be non-zero. The sum will never converge

Baumketner, BioSim, Lviv 2019
2) Non-uniform distribution in the phase space

\[ p = e^{-\beta U(\Gamma)} \]

\[
1 = \frac{\int e^{-\beta U(\Gamma)} e^{\beta U(\Gamma)} d\Gamma}{V^N} = \frac{\int e^{-\beta U(\Gamma)} e^{\beta U(\Gamma)} d\Gamma}{V^N} \frac{\int e^{-\beta U(\Gamma)} d\Gamma}{V^N} = \frac{\langle e^{\beta U(\Gamma)} \rangle}{Z_{NVT}}
\]

The weight function is non-zero where \( U(\Gamma) \) is large and negative. But \( \exp(\beta U(\Gamma)) \) is zero precisely at those points!

No overlap between the weight function and the integrand

The sum will never converge!
2) Thermodynamic integration

1) Integration over density

\[ dF = -PdV - SdT \quad \rightarrow \quad P = -\left( \frac{\partial F}{\partial V} \right)_T \]

\[ F(V) - F(V_0) = - \int_{V_0}^{V} P(V) dV = N \int_{\rho_0}^{\rho} d\rho \frac{P(\rho)}{\rho^2} \]

Free energy has to be known exactly for this density

Certain density for which free energy is supposed to be known

Let's pick density sufficiently low so that the system can be approximated by ideal gas

\[ F(\rho_0) = F_{id}(\rho_0) = \frac{N}{\beta} \log(\rho_0) + \frac{N}{\beta} \log \left( \frac{1}{\lambda^3} \right) - \frac{N}{\beta} \]

Baumketner, BioSim, Lviv 2019
Then the ideal part at low density can be written in terms of the ideal part at any density:

\[
F_{id}(\rho_0) = \frac{N}{\beta} \log(\rho) - \frac{N}{\beta} \log(\rho) + \frac{N}{\beta} \log(\rho_0) + \frac{N}{\beta} \log \left( \frac{1}{\lambda^3} \right) - \frac{N}{\beta} = F_{id}(\rho) + \frac{N}{\beta} \log(\rho_0) - \frac{N}{\beta} \log(\rho)
\]

\[
F_{id}(\rho_0) = F_{id}(\rho) - \frac{N}{\beta} \int_{\rho_0}^{\rho} d\rho \frac{1}{\rho}
\]

\[
F(\rho) = F(\rho_0) + N \int_{\rho_0}^{\rho} d\rho \frac{P(\rho)}{\rho^2} = F_{id}(\rho) - \frac{N}{\beta} \int_{\rho_0}^{\rho} d\rho \frac{1}{\rho} + N \int_{\rho_0}^{\rho} d\rho \frac{P(\rho)}{\rho^2} = F_{id}(\rho)
\]

\[
+ N \int_{\rho_0}^{\rho} d\rho \frac{\beta P - \rho}{\beta \rho^2} = F_{id}(\rho) + N \int_{\rho_0}^{\rho} d\rho \frac{P^{ex}(\rho)}{\rho^2} = F_{id}(\rho) + N \int_{0}^{\rho} d\rho \frac{P^{ex}(\rho)}{\rho^2}
\]

\[p^{ex} = P - \frac{\rho}{\beta}\]

the integral converges in the limit of low density

\[
F_{ex}(\rho) = N \int_{0}^{\rho} d\rho \frac{P^{ex}(\rho)}{\rho^2}
\]

Baumketner, BioSim, Lviv 2019
2) Integration over temperature

\[
Q_{NVT}^{ex} = \frac{1}{VN} \int d\Gamma e^{-\beta U(\Gamma)} \quad \rightarrow \quad \frac{\partial Q_{NVT}^{ex}}{\partial \beta} = -\frac{1}{VN} \int d\Gamma U(\Gamma)e^{-\beta U(\Gamma)} = -<U> Q_{NVT}^{ex}
\]

\[
-\beta F(\beta)
\]

\[
\frac{\partial \log(Q_{NVT}^{ex})}{\partial \beta} = -<U>(\beta) \quad \rightarrow \quad \log(Q_{NVT}^{ex}(\beta)) = \log(Q_{NVT}^{ex}(\beta_0)) - \int_{\beta_0}^{\beta} d\beta <U>
\]

\[
\beta F(\beta) = \beta_0 F(\beta_0) + \int_{\beta_0}^{\beta} d\beta <U>(\beta)
\]

a) Integration from high temperature

\[
\lim_{\beta_0 \to 0} \beta_0 F(\beta_0) = 0 \quad F(\beta) = \frac{1}{\beta} \int_{0}^{\beta} d\beta <U>(\beta)
\]

For potentials that diverge at zero, the average energy will diverge at high temperature/low beta.

For potentials bound at the origin, the average energy will also be bound.
b) Integration from low temperature

\[ F(\beta) = \frac{\beta_{\text{max}}}{\beta} F_0(\beta_{\text{max}}) - \frac{1}{\beta} \int_{\beta}^{\beta_{\text{max}}} d\beta < U > (\beta) \]

Free energy in harmonic approximation. Involves Hessian matrix. Easy to compute for a single structure, for instance in crystals. For liquids, an ensemble of local minima has to be considered.

\[ U_h = U_0 + \frac{N_f}{2} kT \]

number of the degrees of freedom

Thermodynamic integration: integration paths should not cross phase boundaries!
3) Thermodynamic perturbation

Assume that we want to measure free energy difference between two different systems described by Hamiltonian $A$ and $B$.

$$Z_B = \int d\Gamma e^{-\beta U_B(\Gamma)} = \int d\Gamma e^{-\beta U_A(\Gamma)} e^{-\beta (U_B(\Gamma) - U_A(\Gamma))} = < e^{-\beta \Delta U} >_A Z_A$$

$$\Delta U = U_B - U_A$$

$$-\beta F(B) = -\beta F(A) + \log < e^{-\beta \Delta U} >_A$$

$$F(B) = F(A) - \frac{1}{\beta} \log < e^{-\beta \Delta U} >_A$$

1) Only trajectory for one system $A$ is required to compute free energy difference.

2) Energy difference has to be small in order for the average to converge.

3) A path between two states $A$ and $B$ can be constructed that contains intermediate states with mutual overlap.

Free energy of system $B$ is expressed in terms of free energy of system $A$ and some average obtained in ensemble $A$. 

Baumketner, BioSim, Lviv 2019
4) “Artificial” thermodynamic integration

Let’s consider the same problem with two different systems A and B.

Introduce a variable that transforms one system into another

\[ U(\lambda) = U_A + \lambda(U_B - U_A) = U_A + \lambda \Delta U \]

\[ U(\lambda = 0) = U_A, \quad U(\lambda = 1) = U_B \]

Partition function that depends on \( \lambda \)

\[
Z(\lambda) = \int d\Gamma e^{-\beta (U_A + \lambda \Delta U)} \quad \rightarrow \quad \frac{\partial Z(\lambda)}{\partial \lambda} = \int d\Gamma -\beta \Delta U e^{-\beta (U_A + \lambda \Delta U)} = -\beta < \Delta U >_\lambda Z(\lambda)
\]

\[ d\log(Z(\lambda)) = -\beta < \Delta U >_\lambda d\lambda \quad \rightarrow \quad \log(Z(\lambda)) = \log(Z(0)) - \beta \int_0^\lambda < \Delta U >_\lambda d\lambda \]

The most reliable and widely used method

1) Integral has to be evaluated for a number of intermediate points. Each point has to be computed in a separate simulation

2) May have issues with integrand not being smooth enough

3) Applied to a large variety of tasks, for instance mutations
Perturbation and artificial integration are equivalent when the difference between the systems is small.

Free energy for a small $\lambda^*$

\[
F(\lambda^*) = F(A) + \int_0^{\lambda^*} <\Delta U > d\lambda = \underbrace{F(A) + <\Delta U > A}_{\text{if } \lambda^* \sim 0} \lambda^*
\]

In perturbation theory:

\[
U_{B'} = U_A + \lambda^* (U_B - U_A) \quad \rightarrow \quad \Delta U' = U_{B'} - U_A = \lambda^* (U_B - U_A) = \lambda^* \Delta U
\]

\[
F(B') = F(A) - \frac{1}{\beta} log <e^{-\beta \Delta U'} >_A = F(A) - \frac{1}{\beta} log <e^{-\beta \lambda^* \Delta U} >_A \approx F(A) + \lambda^* <\Delta U >_A
\]

\[
\log(1 - \beta \lambda^* <\Delta U >_A + \cdots) = \beta \lambda^* <\Delta U >_A
\]
5) Method of Einstein crystal

Can be used to compute absolute free energy of crystals

Introduce an artificial Hamiltonian that will drive the system into a state with known free energy

\[ U(\lambda) = \lambda U + (1 - \lambda) U_H \]

\[ U(\lambda = 0) = U_H \] system where particles are held at their positions by harmonic potentials

\[ U(\lambda = 1) = U \] actual system

\[ F(\lambda) = F(0) + \int_{0}^{\lambda} < U - U_H > \lambda \ d\lambda \]

free energy of harmonic oscillator

The integrand is well behaved if particles occupy the same position at the start and the end of the integration (in both Hamiltonians)

Baumketner, BioSim, Lviv 2019
6) Widom’s particle insertion method

Free energy can be computed from the chemical potential: \[ F(N_{VT}) = G - PV = N\mu - PV \]

According to the definition:
\[ G(N, TP) = N\mu \quad \rightarrow \quad \mu = G(N + 1, PT) - G(N, PT) \]
\[ G(N, PT) = -kT \log(Q(N_{PT})) \quad \rightarrow \quad \mu = -kT \log\left(\frac{Q(N + 1, PT)}{Q(N, PT)}\right) \]

Recall the definition of the partition function in NPT ensemble:
\[ Q(N, PT) = \frac{1}{N!\lambda^{3N}V_0} Z(N, PT), \quad Z(N, PT) = \int dV e^{-\beta PV} Z(N_{VT}) \]

Evaluate the N+1 term first:
\[ Q(N + 1, PT) = \frac{1}{(N + 1)!\lambda^{3(N+1)}V_0} Z(N + 1, PT) = \frac{1}{(N + 1)!\lambda^{3(N+1)}V_0} \int dV e^{-\beta PV} \int dq_1 dq_2 ... dq_{N+1} e^{-\beta U_{1}} = \]
\[ \frac{1}{(N + 1)!\lambda^{3(N+1)}V_0} \int dV e^{-\beta PV} \int dq_1 e^{-\beta U_{1}} dq_{2} ... dq_{N+1} e^{-\beta U_{2 ... N}} = \frac{1}{(N + 1)!\lambda^{3(N+1)}V_0} \int dV e^{-\beta PV} \int dq_1 e^{-\beta U_{1}} < Z(N_{VT}) = \]
\[ \int dq_1 < e^{-\beta U_{1}} > Z(N_{VT}) \]

Free energy can be computed from the chemical potential: \[ F(N_{VT}) = G - PV = N\mu - PV \]

Volume distribution function:
\[ P(V) = \frac{e^{-\beta PV} Z(N_{VT})}{\int dV e^{-\beta PV} Z(N_{VT})} \]

Baumketner, BioSim, Lviv 2019
Similar expression for the N-particle term:

\[ Q(N, PT) = \frac{1}{N! \lambda^N V_0^3} \int dV e^{-\beta PV} Z(NVT) \]

Combining them one obtains:

\[ \mu = -kT \log\left( \frac{N! \lambda^N}{(N+1)! \lambda^{N+1}} \int dV e^{-\beta PV} < e^{-\beta U_1} > Z(NVT) \right) = -kT \log\left( \frac{1}{(N+1) \lambda^3} \right) < V e^{-\beta U_1} > = \]

\[ -kT \log\left( \frac{V}{(N+1) \lambda^3} \right) \]

\[ \mu^{id} = -kT \log\left( \frac{1}{\rho \lambda^3} \right) \]

Chemical potential can be evaluated from simulations in canonical ensemble:

\[ \mu^{ex} = G^{ex} (N + 1, PT) - G^{ex} (N, PT) = F^{ex} (N + 1, TV) - F^{ex} (N, TV) + (P^{ex} (N + 1) - P^{ex} (N))V \]

\[ F^{ex} (NVT) = -kT \log(Q(NVT)) \quad \rightarrow \quad F^{ex} (N + 1, VT) - F^{ex} (N, VT) = -kT \log\left( \frac{Q(N + 1, VT)}{Q(N, VT)} \right) \]

Configuration integral explicitly:

\[ Q(N + 1, VT) = \frac{1}{V^{N+1}} Z(N + 1, VT) = \frac{1}{V^{N+1}} \int dq_1 e^{-\beta U_1} dq_2 ... dq_{N+1} e^{-\beta U (q_2 ... q_{N+1})} = \frac{1}{V^N} < e^{-\beta U_1} > Z(NVT) = < e^{-\beta U_1} > Q(NVT) \]

Baumketner, BioSim, Lviv 2019
Pressure contribution:

\[
P^{\text{ex}}(N+1) = P^{\text{ex}}(\rho^{\cdot} = \frac{N+1}{V}) \approx P^{\text{ex}}(\rho = \frac{N}{V}) + \frac{dP^{\text{ex}}(\rho)}{d\rho}(\rho^{\cdot} - \rho) = P^{\text{ex}}(N) + \frac{K^{\text{ex}}}{\rho} \frac{1}{V}
\]

bulk modulus

\[
K^{\text{ex}} = \rho \frac{dP^{\text{ex}}(\rho)}{d\rho} = a\rho^2 + b\rho^3 + ...
\]

from virial expansion

Two terms together:

\[
\mu^{\text{ex}} = -kT \log(< e^{-\beta U_1} >) + \frac{K^{\text{ex}}}{\rho} \approx a\rho + b\rho^2 + ...
\]

this summand vanishes at low densities but doesn't disappear when N tends to infinity

At low densities:

\[
\mu^{\text{ex}} \approx -kT \log(< e^{-\beta U_1} >) = F^{\text{ex}}(N+1,VT) - F^{\text{ex}}(N,VT) = \frac{F^{\text{ex}}(N+1,VT) - F^{\text{ex}}(N,VT)}{N+1-N} \approx \frac{\partial F^{\text{ex}}(NVT)}{\partial N} \bigg|_{VT} = \mu^{\text{ex}}
\]

the original Widom's formula. Due to large fluctuations it applies only at low densities. So it's safe to use it there. At high densities, large variations in \( U_1 \) resulting from particle overlaps hinder convergence.

Baumketner, BioSim, Lviv 2019
7) Chemical potential from the Gibbs ensemble simulations

Two boxes in GEMC simulations are maintained at constant temperature and pressure. The Widom formula for the NPT ensemble can be used to estimate chemical potential during the transfer move.

\[
\mu = -kT \log \left( \frac{1}{(N+1)\lambda^3} < Ve^{-\beta U_1} > \right)
\]

when a particle is added to box 1 a presumed change of the box’s energy is computed

\[
U_1 = U(n_1 + 1) - U(n_1)
\]

The volume and the number of particles change so the average includes both of them:

\[
\mu = -kT \log \left( \frac{1}{\lambda^3} < V_1 e^{-\beta U_1} > \right)
\]

NPT

Gibbs ensemble

1) The identity of box 1 can’t change during chemical potential evaluation. If it’s vapor it has to remain vapor to the end of the simulation. Similarly for liquid. This is easy to achieve unless the system is near critical point.

2) If box 1 contains all particles of the system (the other box is empty) one should still attempt to add a particle to it to evaluate \( U_1 \). This step is not executed in normal GEMC.
Imagine that we have a number of trajectories simulated at a set of temperatures. WHAM can be used to combine the energy histograms and obtain a) density of states and b) relative free energies. The relevant equations are:

\[ n(E) = \frac{\sum_i^K p_i(E)}{\sum_i^K e^{-\beta_i(E-f_i)}} \]

\[ e^{-\beta_i f_i} = \sum_n n(E)e^{-\beta_i E} \]

Histograms are built assuming certain size of the bins \( \Delta E \). The bin size has to be a) small so that the density of states is constant within \([E, E + \Delta E]\) and b) large so that reasonable statistics of \( n(E) \) is obtained. The finite size of \( \Delta E \) introduces errors. If only the free energies are of interest, the error can be minimized by taking the limit \( \lim \Delta E \to 0 \). The WHAM equations then can be reduced to a simpler form.

Let us introduce an indicator function which is unity if energy of a given configuration \( k \) is within \( \Delta E \) of \( E \) and zero otherwise:

\[ \delta_{E,E_k} = \frac{1}{2} \left( \Theta(E_k - E) + \Theta(E + \Delta E - E_k) \right) = \begin{cases} 1 & \text{if } E \leq E_k \leq E + \Delta E \\ 0 & \text{otherwise} \end{cases} \]
The energy histogram for temperature \( i \) is

\[ p_i(E) = \sum_k \delta_{E,E_k} \]

summation runs over all sampled conformations

The density of states then can be re-written:

\[ n(E) = \frac{\sum_{i}^{K} p_i(E)}{\sum_{i}^{K} e^{-\beta_i(E-f_i)}} = \frac{\sum_{i}^{K} \sum_{j} \delta_{E,E_j}}{\sum_{i}^{K} e^{-\beta_i(E-f_i)}} \]

for sufficiently small \( \Delta E \) the indicator function will kill the summation over \( E \)

Let us substitute this expression into the second WHAM equation:

\[ e^{-\beta_s f_s} = \sum_{E} \frac{\sum_{i}^{K} \sum_{j} \delta_{E,E_j}}{\sum_{i}^{K} e^{-\beta_i(E-f_i)}} e^{-\beta_s E} = \sum_{i}^{K} \sum_{j} \delta_{E,E_j} e^{-\beta_s E} \]  

\[ = \sum_{i}^{K} \sum_{j} \frac{e^{-\beta_s E_j}}{\sum_{i}^{K} e^{-\beta_i(E_j-f_i)}} \]

summation over all trajectories  

summation over all conformations in trajectory \( i \).

MBAR equations

\[ \beta_s f_s = -\log\{ \sum_{i}^{K} \sum_{j} \frac{e^{-\beta_s E_j}}{\sum_{i}^{K} e^{-\beta_i(E_j-f_i)}} \} \]

1) Non-linear set of equations for \( f_i \)  
2) Can be solved by iterations  
3) No binning is required  
4) Solution is not unique. Only relative free energies are obtained

General solution

\[ \beta_s f_s + A \]
Other methods:

9) Grand canonical ensemble (covered)

10) Methods for approximate evaluation of the free energy: chemical Monte Carlo, lambda dynamics, linear response theory, ....
MC simulations in generalized (non-Boltzmann) ensembles
Outline

- Introduction
- Generalized ensembles:
  - multicanonical ensemble
  - Tsallis statistics
  - Wang-Landau method
  - $1/k$ ensemble
  - $J$-walking algorithm
  - expanded-ensemble method
- Replica-exchange (REX) approach
- Applications:
  - replica-exchange simulations of peptide aggregation
  - folding of a short $\beta$-peptide in explicit water

Baumketner, BioSim, Lviv 2019
Why do we need *generalized* ensembles?

Two ingredients of a successful simulation:

(I) *accurate representation of the system and solvent*

(II) *adequate sampling of the conformational space*

Simulation times must be at least **10 times** longer than the relevant relaxation time!
Time scales

- Side-chain rotations
- Loop closure
- Helix formation
- Folding of \( \beta \)-hairpins
- Protein folding
- Protein aggregation

Baumketner, BioSim, Lviv 2019
Brute force approach


- **Villin headpiece** subdomain, 256 CPUs of Cray T3E.

![Diagram](image)

- RMSD~10A
- RMSD~4A
- RMSD<1A
Why are the relaxation times so long?

- Each minimum acts as a kinetic trap. The relaxation time is determined by the escape time from the minima.

\[ \tau = \tau_0 \exp(\Delta G / k_B T) \]

- Number of local potential energy minima grows exponentially with \( N \)

\( \Delta G \)

- \( \Delta G \): ~1 ps
- ~1.5 ns
- ~1 ms and longer

Baumketner, BioSim, Lviv 2019
Generalized ensembles

Canonical (Bolzmann) distribution is narrow!

Broad distributions facilitate escapes from minima!

\[ P(E) \sim n(E) e^{-\beta E} \]

\[ \Delta E > \delta E \]

\[ \Delta E < \delta E \]
Multicanonical ensemble

- sampling on a “deformed” potential energy surface $W(E)$:

$$P^*(E) \sim n(E) e^{-\beta W(E)}$$

- flat energy distribution for multicanonical ensemble:

$$P^*(E) = P_\mu(E) \sim \text{const}$$

$$W_\mu(E) = ?$$

- equation (1) can be rewritten as a non-linear equation in $W_\mu(E)$ (valid up to a constant which drops during normalization):

$$W_\mu(E) = \frac{1}{\beta} [\ln n(E) - \ln P_\mu(E)] = \frac{1}{\beta} \ln n(E)$$

- density of states is estimated from a simulation. It depends (as a functional) on $W_\mu(E)$ and simulation parameters: $n(E)[W_\mu; \text{param}]$. Simplest solution to equation (2) is given by successive iterations:

$$W_\mu^{n+1}(E) = \frac{1}{\beta} \ln n^n(E)$$

Baumketner, BioSim, Lviv 2019
Multicanonical ensemble

- final result:  
  \[ W_\mu^{n+1}(E) = W_\mu^0(E) + \frac{1}{\beta} \ln[P_\mu^0(E) \cdots P_\mu^n(E)] \]  
  (4)

- an example:  
  \[ W_\mu^0(E) = E \], zero energy distribution \( P_\mu^0(E) \) is canonical and equation (4) is the standard **multicanonical recursion**:

  \[ W_\mu^{n+1}(E) = E + \frac{1}{\beta} \ln[P_\beta(E) \cdots P_\mu^n(E)] \]  
  (5)

- canonical distribution can be recovered from \( P_\mu(E) \) through a reweighting procedure:

  \[ P_\beta(E) \sim P_\mu(E) e^{\beta(W_\mu(E) - E)} \]  
  (6)

- three steps of a multicanonical simulation:
  
  (i) generating \( W_\mu(E) \) in successive iterations

  (ii) equilibrium sampling

  (iii) recovering canonical expectations for various observables

Baumketner, BioSim, Lviv 2019
Monte Carlo studies of spin glasses


- 2D 10-state Potts model

- No exponential increase in the tunneling time between two free energy minima

- Speedup up to 2 orders of magnitude compared to standard simulations

**FIG. 2.** Multicanonical action density distribution $\mathcal{P}_{\tau_0}(s)$ together with its reweighted distribution $\mathcal{P}_{\tau_0}(s)$. 

Baumketner, BioSim, Lviv 2019
Molecular dynamics simulations


- Equations of motion:

\[
\dot{q}_i = \frac{dq_i}{dt} = \frac{p_i}{m_i} \quad \text{“multicanonical” force}
\]

\[
\dot{p}_i = f_i^\mu - \xi p_i \quad \text{factor to keep kinetic energy constant}
\]

\[
f_i^\mu = -\frac{\partial W_\mu(E)}{\partial q_i} = -\frac{dW_\mu(E)}{dE} \frac{\partial E}{\partial q_i} = -\frac{dW_\mu(E)}{dE} f_i
\]

\[
\xi = \frac{\sum f_i^\mu \cdot \dot{q}_i}{2 \sum p_i^2 / 2m_i}
\]

- The only modification is in how *forces* are calculated!

Baumketner, BioSim, Lviv 2019
The amino acid sequence of met-enkephalin is **TYR-GLY-GLY-PHE-MET**. Enkephalins belong to the family of endorphins that are expressed in reproductive organs. Their precise function is not well understood.
**Summary**

- **Advantages:**
  
  (i) multicanonical simulations do not get trapped in local minima  
  (ii) various thermodynamic quantities are obtained as a function of temperature from a single run. No need to run multiple simulations at different temperatures

- **Disadvantages:**
  
  (i) the energy transformation function $W_\mu(E)$ is non-analytical. Several preliminary simulations are needed to generate it  
  (ii) these simulations can not be run in parallel  
  (iii) convergence is sensitive to the details of numerical implementation. Can be quite poor if $P_\mu^0(E)$ is not accurately determined

Baumketner, BioSim, Lviv 2019
Tsallis energy transformation


- analytical form for the multicanonical weight factor. Non-exponential falloff at large $E$:

$$w_T(E) = [1 + x\beta(E - E_0)]^{-\frac{1}{x}} \xrightarrow[x\to0]{} e^{-\beta(E-E_0)}$$

- energy transformation:

$$W_T(E) = \frac{1}{x\beta} \ln[1 + x\beta(E - E_0)]$$

- at low temperature the density of states can be calculated in harmonic approximation:

$$n(E) \sim (E - E_0)^{\frac{N_F}{2}}$$

- if low-energy states are to be populated:

$$x < x_c = \frac{2}{N_F}$$

- optimal value for $x$:

$$x_0 = 0.5 x_c$$

Baumketner, BioSim, Lviv 2019
Applications

- progressive **broadening** of the potential energy distribution of a met-enkephalin model as $x$ is reduced

- same $x_0$ does not work for all systems!

Wang-Landau method

- Acceptance probability of a Monte Carlo move $i \rightarrow j : \min \{ n(E_i) / n(E_j), 1 \}$

**Algorithm:**
0) $n(E)=1$ for all $E$, $f=3$
1) every time an energy level $E$ is visited, $n(E)=f n(E)$
2) simulation is continued until energy histogram $H(E)$ is “flat”. $H(E)$ for all $E$ is not less than 80% of $<H(E)>$
3) change the factor $f_{i+1} = \sqrt{f_i}$
4) if $f_{i+1} > f_c (\sim 1.001)$ set $H(E)=0$ and return to step 1

- Detailed balance is satisfied at $f=1$
- Applicable to **large** systems. The desired energy interval can be broken into smaller pieces which are simulated in **parallel**
- **Caveat:** Make energy intervals large enough to avoid trapping!

- Application to proteins: [A. Cavalli et al., *Biophys J.* 88, (2005),3158]
1/k ensemble


- Acceptance probability of a Monte Carlo move \( i \rightarrow j : \min \{ w_j / w_i, 1 \} \)
  - *Metropolis (canonical distribution)*
  \[ w_i = e^{-\beta E_i} \]
  - *Multicanonical, Wang-Landau*
  \[ w_i = 1 / k_i, \quad k_i = n(E_i) \]

- Definitions of entropy:
  \[ S(E) = k \log( n(E) ) \quad S^*(E) = k \log( \int_{E'} dE' n(E') ) \]
  are equivalent in the therm. limit:
  \[ S^*(E) = S(E) + O(\log(N)) \quad N \rightarrow \infty \]

Differ for finite \( N \). Energy distribution is not flat.

Baumketner, BioSim, Lviv 2019
J-walking algorithm

- D. D. Frantz, D. L. Freeman and J. D. Doll, JCP 93, (1990), 2769, “Reducing quasi-ergodic behavior in Monte Carlo simulations by J-walking: Applications to atomic clusters”

- It’s easier to overcome potential energy barriers at \textit{high} T (low \(\beta\))!

- two types of Monte Carlo moves: \textit{local} to sample free energy minima and \textit{global} to transition among minima

- \(J\)-walks (global moves) are generated at high \(T\)

- too high \(T\) = low acceptance rate
- too low \(T\) = ergodicity problems

Low-\(\beta\) distribution provides an “intrinsic” size of a global move!

Double-well potential

Local moves get trapped!

“Global” moves would help. But how to pick them?

Baumketner, BioSim, Lviv 2019
Expanded-ensemble concept


- Temperature $\beta$ is treated as a **dynamical variable**. Canonical ensembles for each $\beta$ are treated as sub-ensembles of a larger, **expanded** ensemble. Distribution function in this larger ensemble is defined through a supplementary function $\alpha(\beta)$:

$$P^*(\beta, \Gamma) = \frac{\exp(\beta \alpha(\beta) - \beta H(\Gamma))}{Z^*}, \quad Z^* = \sum_{\beta} \exp(\beta \alpha(\beta))Z(\beta), \quad Z(\beta) = 1/N! \int d\Gamma \exp(-\beta H(\Gamma))$$

- Probability to occupy states with temperature $\beta$:

$$P^*(\beta) = \frac{\exp(\beta \alpha(\beta))Z(\beta)}{Z^*}$$

- Free energy difference for two temperatures: available from simulations

$$\beta_2 F(\beta_2) - \beta_1 F(\beta_1) = \beta_1 \alpha(\beta_1) - \beta_2 \alpha(\beta_2) - \log \{ P^*(\beta_1)/P^*(\beta_2) \}$$

  given by the model

Baumketner, BioSim, Lviv 2019
MC in expanded ensembles

- MC algorithm:

(I) Standard Metropolis moves for fixed $\beta$.

\[ p = \min \{ \exp(-\beta \Delta E), 1 \} \]

(II) Temperature swaps: $\beta_1 \rightarrow \beta_2$ with probability:

\[ p = \min \{ \exp((\beta_1 - \beta_2)E + \beta_2 \alpha(\beta_2) - \beta_1 \alpha(\beta_1)), 1 \} \]

(III) Total time spent at each temperature is accumulated in histograms and used for estimating $P^*(\beta)$

- Random walk in temperature is realized for:

\[ \alpha(\beta) = F(\beta) = 1/\beta \log(Z(\beta)) \]

- $\alpha(\beta)$ are determined in successive iterations

- For each $\beta$ canonical distributions are recovered!

- The algorithm is also known as *simulated tempering*

---

Random temperature walk!

Baumketner, BioSim, Lviv 2019
Replica-exchange algorithm


*Replica-exchange (REX) = parallel tempering = multiple Markov chain method*
Replica-exchange algorithm

- the method is based on the **expanded-ensemble** idea \( P^\ast(\beta, \Gamma) \sim \exp(\beta \alpha(\beta) - \beta H(\Gamma)) \)
- \( N \) independent replicas are considered (**parallel tempering**)
- double-jumps are attempted

\[
p = \begin{cases} 
  e^\Delta, & \Delta < 0 \\
  1, & \Delta < 0
\end{cases}
\]

\[
\Delta = (\beta_m - \beta_n)(E_j - E_i)
\]

\( \alpha(\beta) \) drop in double jumps

- **uniform distribution** over sampled temperatures (one replica per temperature condition)
- **canonical distribution** for each considered temperature \( \beta \)

Baumketner, BioSim, Lviv 2019
**Practical points**

- Four parameters need to be set: $T_{\text{max}}$, $T_{\text{min}}$, $N$, $\tau_0$

  - $T_{\text{max}}$ —— the higher the better (~500-700K). Typical relaxation time at this temperature should be ~1-100ps
  - $T_{\text{min}}$ —— the temperature of your interest (300K?)
  - $N$ —— large enough to ensure 10-50% acceptance probability for swaps between replicas
  - $\tau_0$ —— the longer the better. Typically 100-1000 simulation time steps

- What to look out for:
  (i) replica-exchange acceptance ratio is more than 10%
  (ii) each replica visits $T_{\text{min}}$ and $T_{\text{max}}$ at least several times
  (iii) all **relevant** order parameters undergo sufficient relaxation

---

Nadler [PCL B 112 (2008) 10386]

$N \approx 1 + 0.594 \sqrt{C} \ln T_{\text{max}} / T_{\text{min}}$

---

Baumketner, BioSim, Lviv 2019
Kinetic data from REX simulations

- There is no physical kinetics in the REX simulations. A number of approaches to extract kinetic information:

  
  Andrec [PNAS 102 (2005) 6801]
  van der Spoel [PRL 96 (2006) 238102]
  Yang [JMB 372 (2007) 756]
  Buchete [JPC B 112 (2008) 6057]
  Muff [JPC B 113 (2009) 3218]
  Chodera [JCP 134 (2011) 244107]

- Most approaches rely on the following ingredients:

  a) discretization of the available configuration space

     \{ Very difficult to get a representative ensemble \}

  b) obtaining rates of transition between the identified states

     \{ Most often in straight MD simulations. Not reliable \}

  c) solving master equation to generate reaction time

     \{ Relies on assumption on how transition rates depend on temperature \}

**Example:** ETNA of Muff and Caflisch

*Folding time for a β-sheet peptide predicted for varying temperature over a range that spans an order of magnitude*

Baumketner, BioSim, Lviv 2019
**Replica-exchange flavors**

- **$\textit{REX}$** coupled with Tsallis energy deformation function:
  - [Jang et al., *PRL*, 91 (2003) 058305](#)

- Muticanonical $\textit{REX}$ and $\textit{REX}$ multicanonical:

- $\textit{REX}$ in constant pressure (CPT) ensemble:

- Mutidimensional $\textit{REX}$:

- *Ab initio* Monte Carlo $\textit{REX}$:

- Hamiltonian $\textit{REX}$:

Baumketner, BioSim, Lviv 2019
Replica-exchange flavors

- **REX** coupled with RISM:  

- Local **REX**:  

- Non-equilibrium switches **REX**:  
  Ballard and Jarzynski, *PNAS*, 106 (2009) 12224

- Further reading:
Multiple-histogram reweighting technique


- $i=1,m$ temperatures, $N_i$ samples, $H_i(E)$ energy histograms
  \[ P_i(E) = H_i(E) / N_i = n(E) \exp(-\beta_i (E - f_i)) \]

- by definition, free energy: \( \exp(-\beta_i f_i) = \sum_{E_k} n(E_k) \exp(-\beta_i E_k) \) \( P(E) \)

- inaccurate estimate of the density of states from simulation at temperature $i$:
  \[ n_i(E) = P_i(E) \exp(\beta_i (E - f_i)) \]

- improved density of states:
  \[ n(E) = \frac{\sum_i g_i n_i(E)}{\sum_i g_i} \]

- $\delta n(E)^2 / n(E)$ is minimized with respect to weighting $g_i$ coeff.

- set of non-linear equations in $f_i$: \( n(E) = \frac{\sum_i P_i(E)}{\sum_i \exp(-\beta_i (E - f_i))} \)

Baumketner, BioSim, Lviv 2019
Why is replica exchange the method of choice in biomolecular simulations?

- Parallelism
- No empirical parameters or fitting involved
- Access to low free energy minima through accelerated relaxation
- Availability of all thermodynamical properties as a function of temperature through histogram reweighting techniques