Energy functions and their relationship to molecular conformation

CS/CME/BioE/Biophys/BMI 279

Sept. 22 and 24, 2020

Ron Dror

Outline

- Energy functions for biomolecular systems
  - Definition and properties
  - Molecular mechanics force fields
- What does the energy function tell us about biomolecular conformation?
  - The Boltzmann distribution
  - Microstates and macrostates
  - Free energy

Energy functions for biomolecular systems

Energy functions for biomolecular systems

Definition and properties

Specifying atom positions

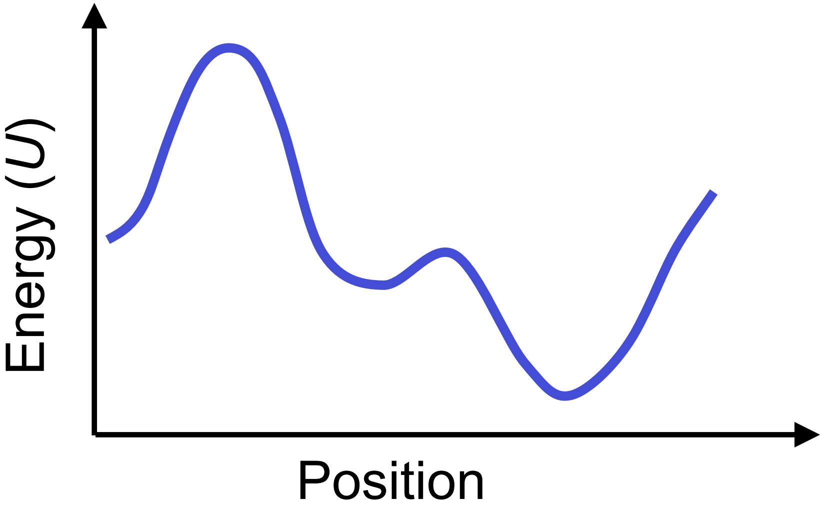
- For a system with *N* atoms, we can specify the position of all atoms by a single vector ***x*** of length 3*N*
  - This vector contains the *x*, *y*, and *z* coordinates of every atom

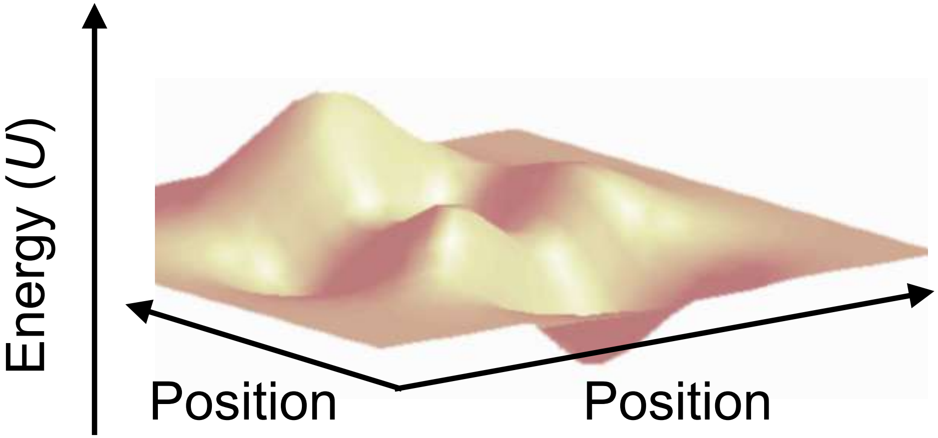
$$\mathbf{x} = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \\ x_2 \\ y_2 \\ z_2 \\ \vdots \\ x_N \\ y_N \\ z_N \end{pmatrix}$$

Energy function

High dimensional energy landscape

- A potential energy function $U(\boldsymbol{x})$ specifies the total potential energy of a system of atoms as a function of all their positions ($\boldsymbol{x}$)
  - In the general case, include not only atoms in the protein but also surrounding atoms (e.g., water)



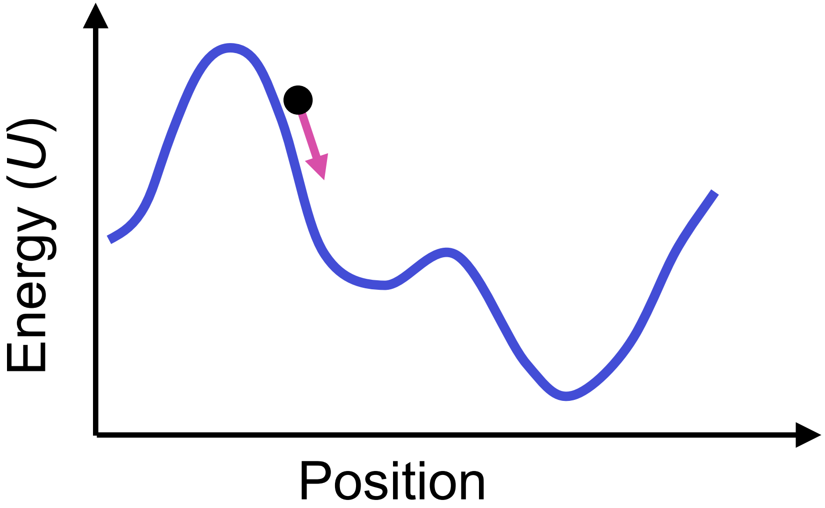


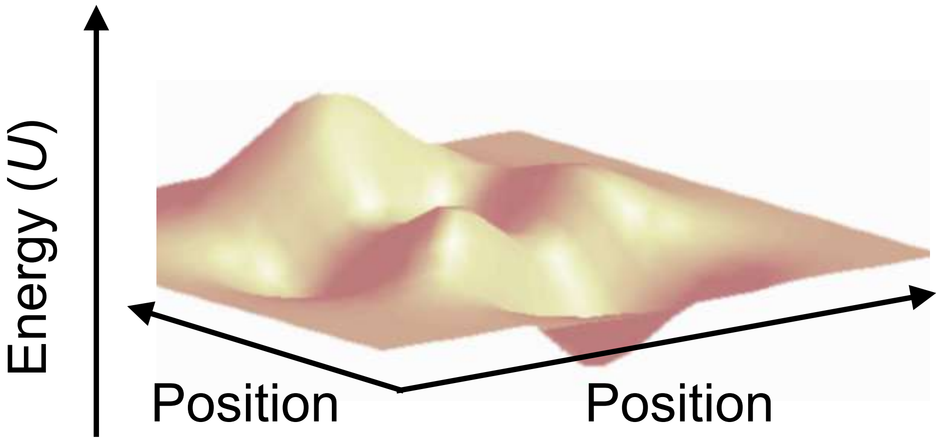
Relationship between energy and force

- Force on atom *i* is given by derivatives of *U* with respect to the atom's coordinates x_i, y_i, and z_i

$$F(\boldsymbol{x}) = -\nabla U(\boldsymbol{x})$$

- At local minima of the energy *U*, all forces are zero
- The potential energy function *U* is also called a *force field*





Types of force fields (energy functions)

- A wide variety of force fields are used in atomic-level modeling of macromolecules
- Physics-based vs. knowledge-based
  - Physics-based force fields attempt to model actual physical forces
  - Knowledge-based force fields are based on statistics about, for example, known protein structures
  - Most real force fields are somewhere in between
- Atoms represented
  - Most realistic choice is to model all atoms
  - Some force fields omit waters and other surrounding molecules. Some omit certain atoms within the protein.

Some commonly omit hydrogens in protein. Some even take large group of atoms and replace those with “pseudo-atom” that represents the whole group

Energy functions for biomolecular systems

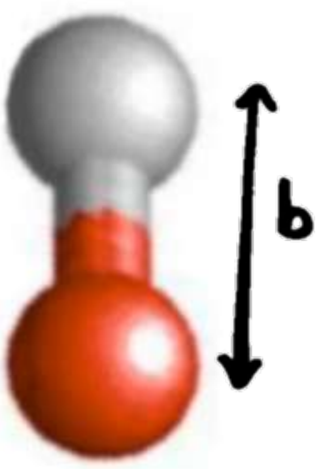
Molecular mechanics force fields

Molecular mechanics force fields

- Today, we'll focus on *molecular mechanics force fields*, which are often used for molecular simulations
- These are more toward the physics-based, all-atom end (i.e., the more "realistic" force fields)
  - Represent physical forces explicitly
  - Typically represent solvent molecules (e.g., water) explicitly
- We'll revisit the forces acting between atoms and write down the functional forms typically used to approximate them

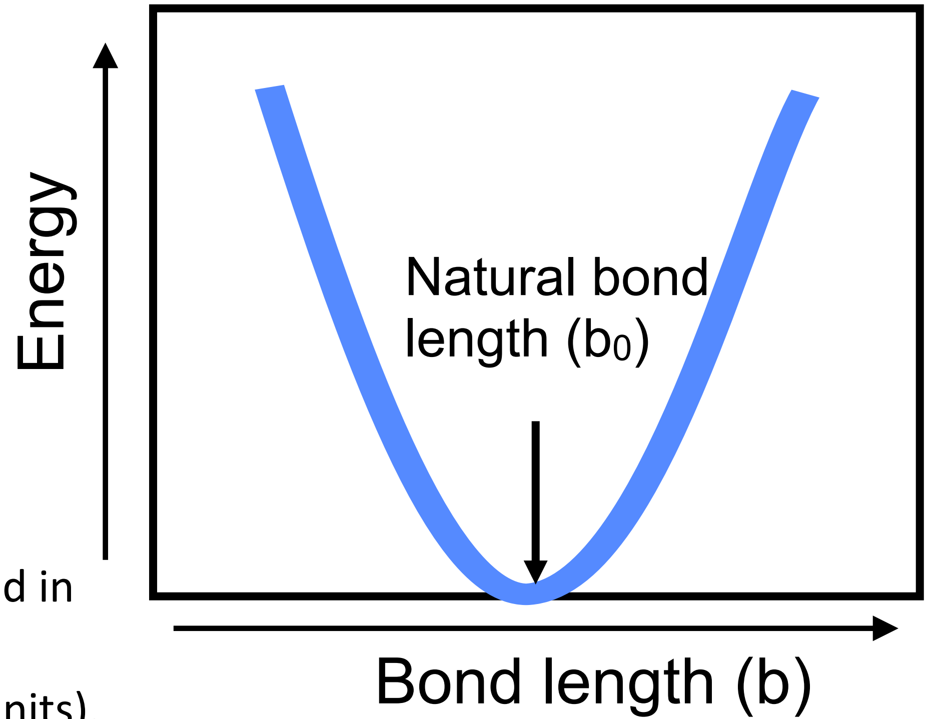
Bond length stretching

- A bonded pair of atoms is effectively connected by a spring with some preferred (natural) length. Stretching or compressing it requires energy.



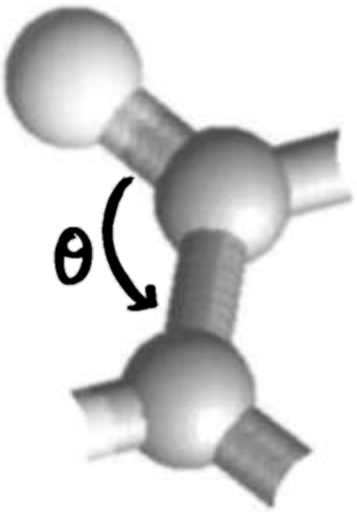
$$U(b)=k_b\left(b-b_0\right)^2$$

Note: A factor of 1/2 is sometimes included in this equation. I'm ignoring such constant factors (they can be folded into *k_b* or the units).

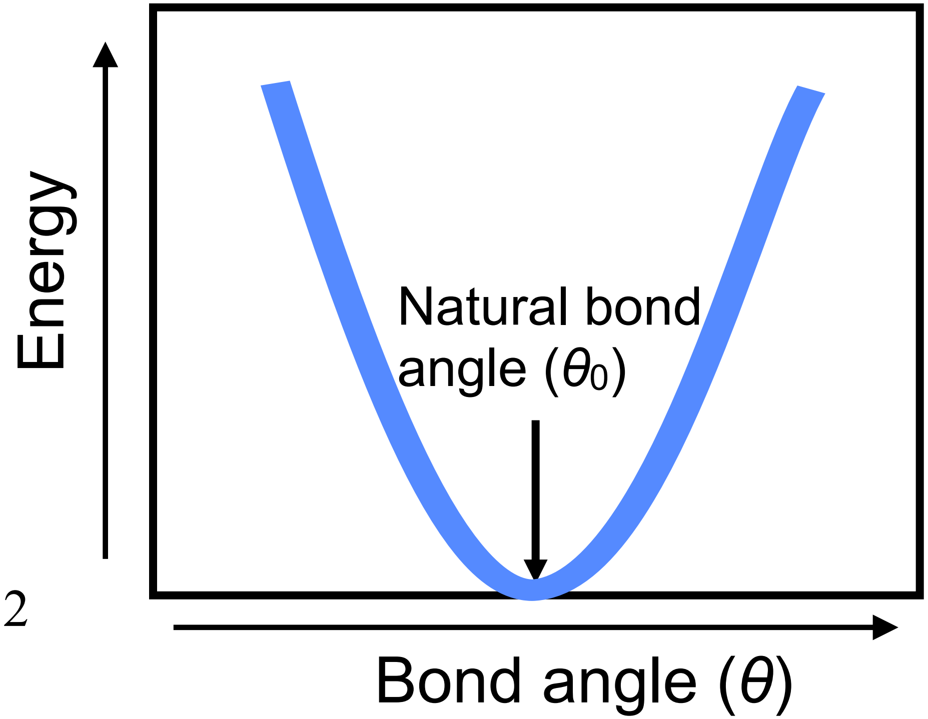


Bond angle bending

- Likewise, each bond angle has some natural value. Increasing or decreasing it requires energy.

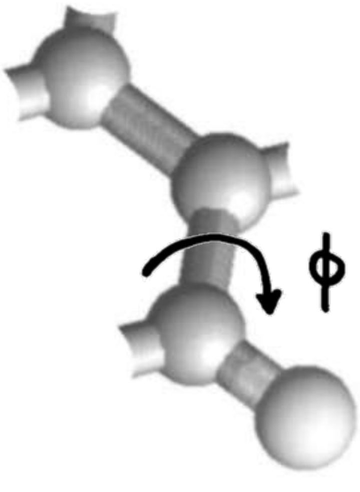


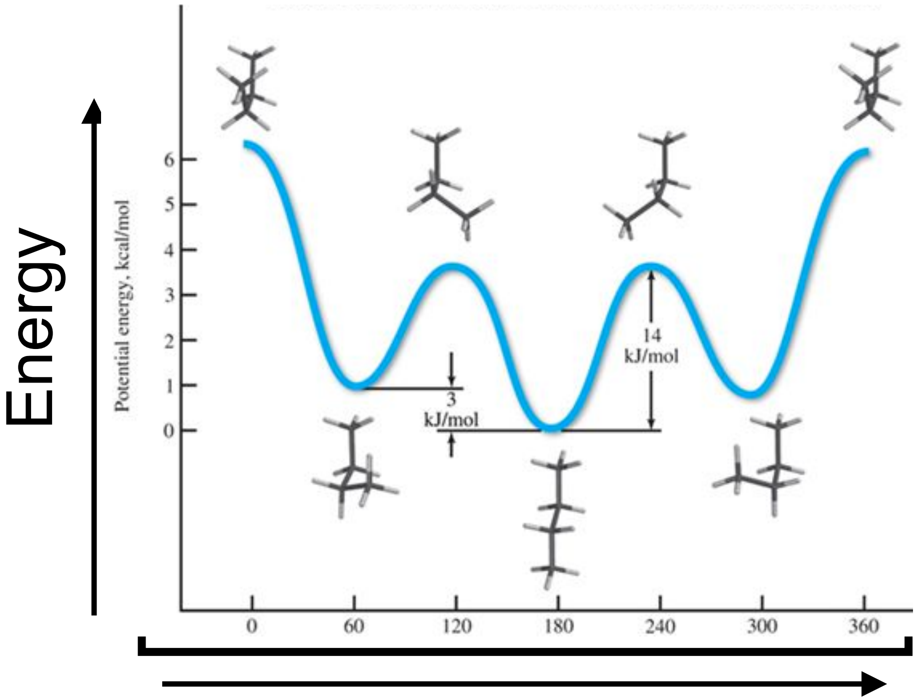
$$U(\theta)=k_\theta\left(\theta-\theta_0\right)^2$$



Torsional angle twisting

- Certain values of each torsional angle are preferred over others.





Energy

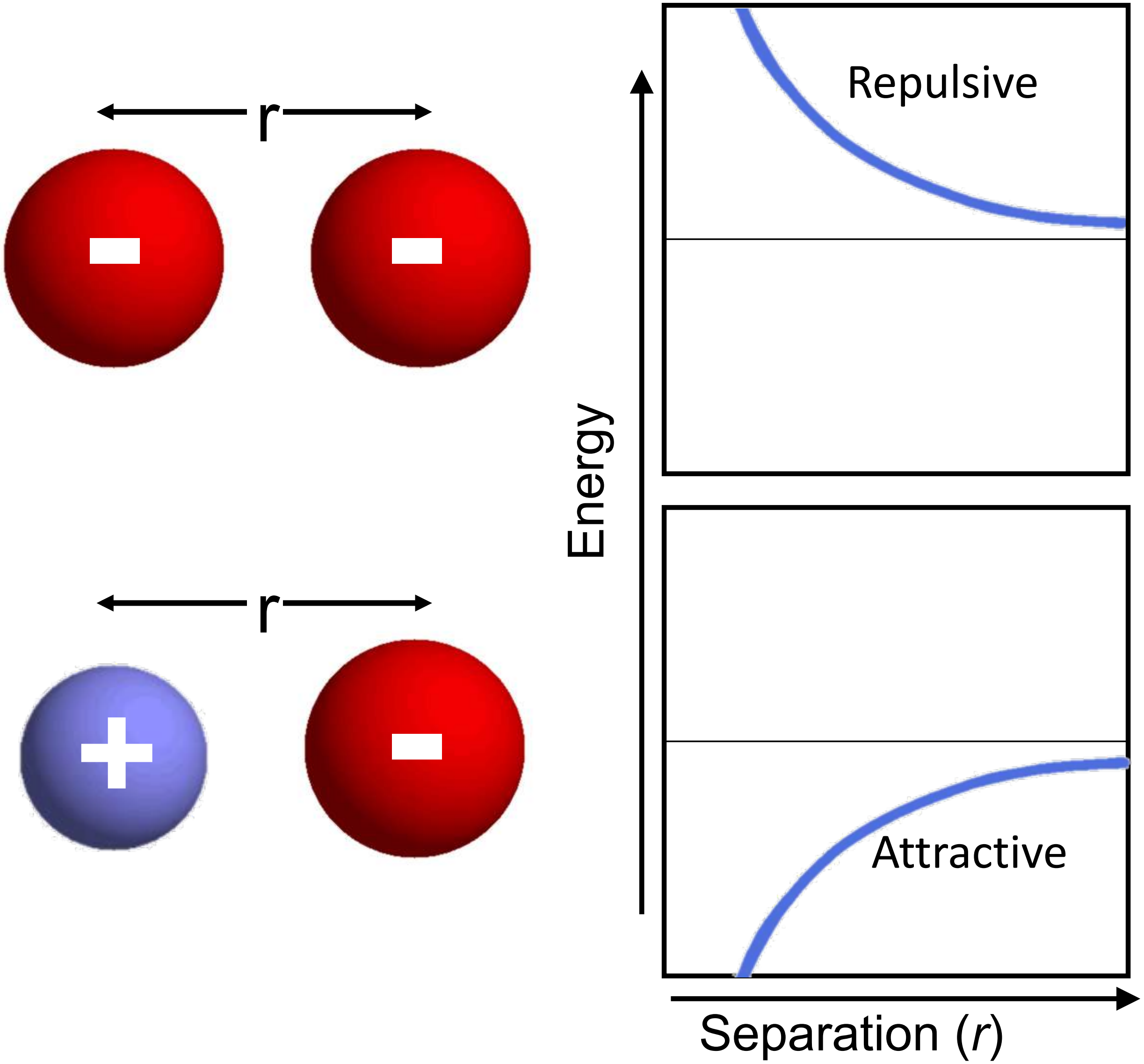
Torsional angle (Φ)

$$U(\phi)=\sum_{n} k_{\phi,n}\left[1+\cos\left(n\phi-\phi_{n}\right)\right]$$

Typically n takes on one or a few values between 1 and 6 (particularly 1, 2, 3, 6)

http://images.slideplayer.com/38/10813003/slides/slide_34.jpg

Electrostatics interaction

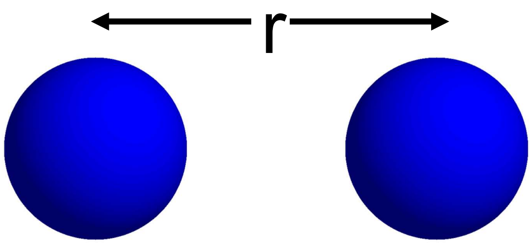


- Like charges repel. Opposite charges attract.
- Acts between all pairs of atoms, including those in different molecules.
- Each atom carries some “partial charge” (may be a fraction of an elementary charge), which depends on which atoms it’s connected to

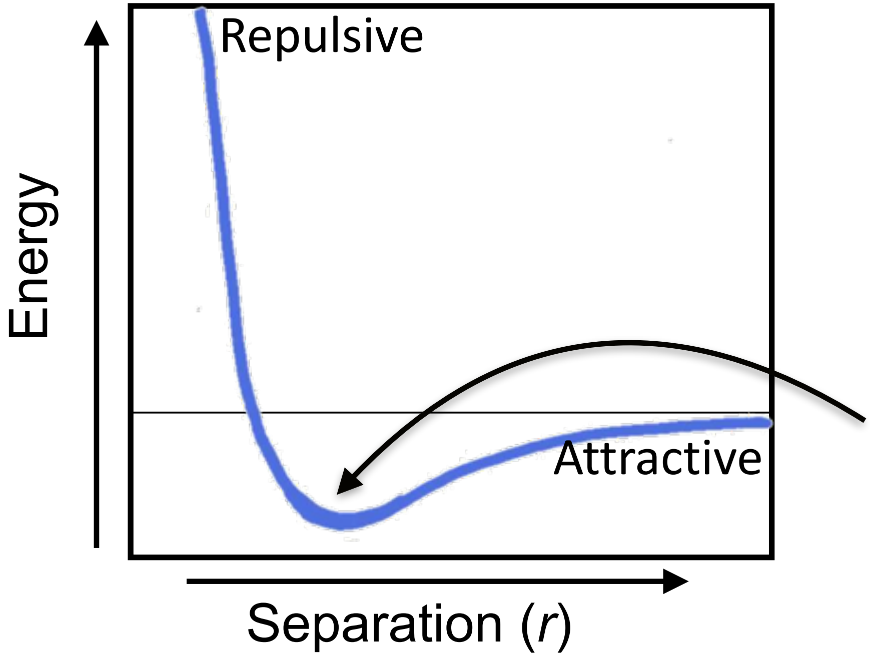
$$U(r) = \frac{q_i q_j}{r}$$

where q_i and q_j are partial charges on atoms *i* and *j*

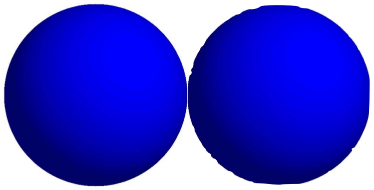
van der Waals interaction



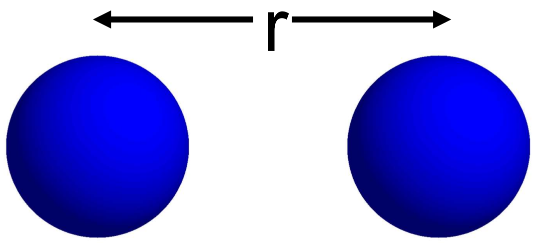
- van der Waals forces act between all pairs of atoms and do not depend on charge.
- When two atoms are too close together, they repel strongly.
- When two atoms are a bit further apart, they attract one another weakly.



Energy is minimal when atoms are “just touching” one another



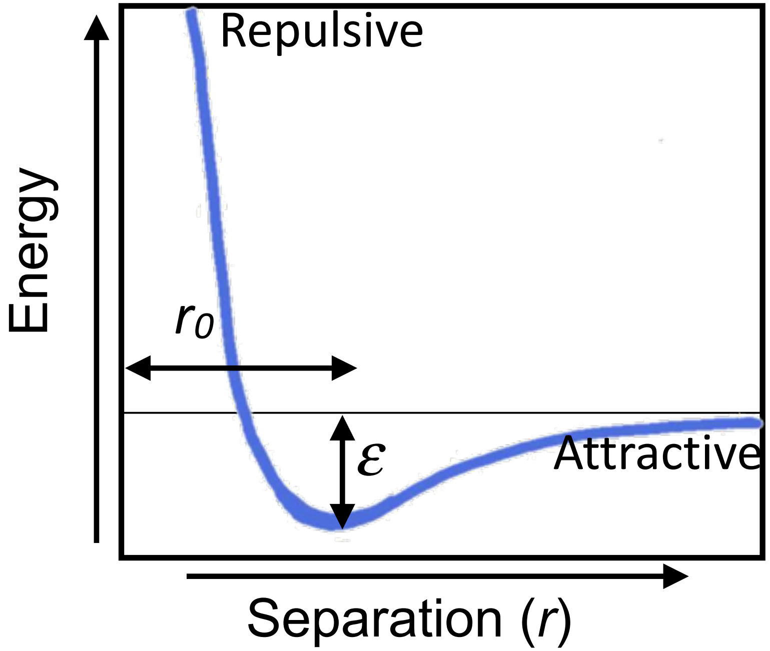
van der Waals interaction



$$U(r) = \frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^{6}}$$

We can also write this as:

$$U(r) = \varepsilon\left[\left(\frac{r_0}{r}\right)^{12} - 2\left(\frac{r_0}{r}\right)^{6}\right]$$



Note: Historically, r^{12} term was chosen for computational convenience; other forms are sometimes used

A typical molecular mechanics force field

$$U = \sum_{\text{bonds}} k_b \left(b - b_0\right)^2$$

Bond lengths (“Stretch”)

$$+ \sum_{\text{angles}} k_\theta \left(\theta - \theta_0\right)^2$$

Bond angles (“Bend”)

Bonded terms

$$+ \sum_{\text{torsions}} \sum_{n} k_{\phi,n} \left[1 + \cos\left(n\phi - \phi_n\right)\right]$$

Torsional/dihedral angles

$$+ \sum_{i} \sum_{j>i} \frac{q_i q_j}{r_{ij}}$$

Electrostatics

$$+ \sum_{i} \sum_{j>i} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}$$

Van der Waals

Non-bonded terms

How are the parameters fit?

- Combination of:
  - Quantum mechanical calculations
  - Experimental data
    - For example: b_0 can be estimated from x-ray crystallography, and K_b from spectroscopy (infrared absorption)

$$U(b) = K_b\left(b - b_0\right)^2$$

- The torsional parameters are usually fit last. They absorb the “slop.” Fidelity to physics is debatable.
- These force fields are approximations!

What does the energy function tell us about biomolecular conformation?

What does the energy function tell us about biomolecular conformation?

The Boltzmann distribution

Relating energy to probability

- Given the potential energy associated with a particular arrangement of atoms (set of atom positions), what is the probability that we'll see that arrangement of atoms?
- Assumptions:
  - System is at constant temperature (so atoms are constantly jiggling around).
  - We watch the system for a really long time (allowing it to fully equilibrate).

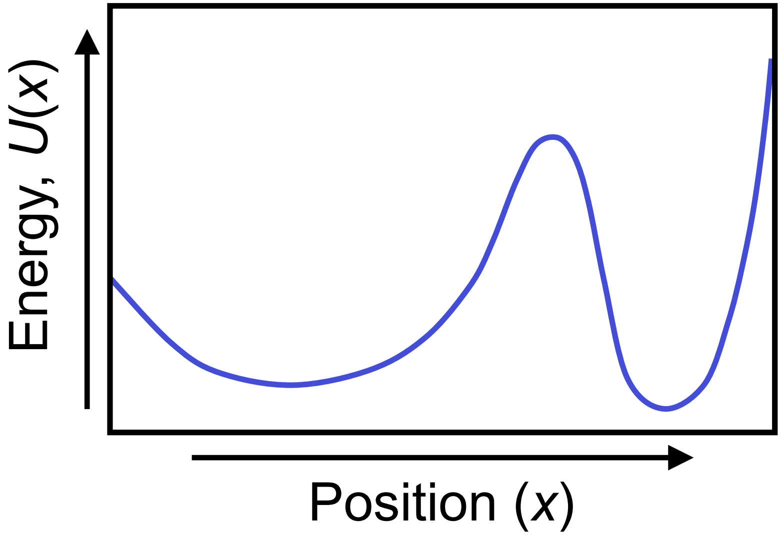
The Boltzmann Distribution

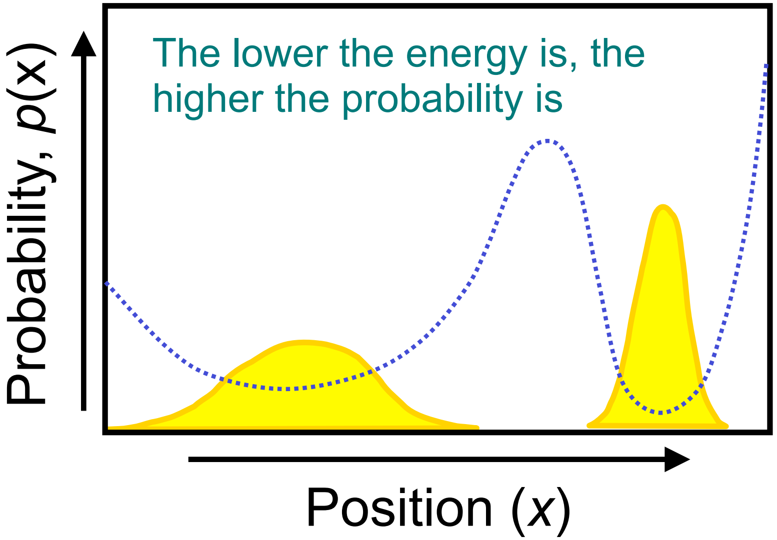
- The Boltzmann distribution relates potential energy to probability

U(x) = potential energy of the system
associated with atom arrangement x

$$p(\boldsymbol{x}) \propto \exp\left(-U(\boldsymbol{x}) \middle/ k_B T \right)$$

where *T* is temperature and k_B is the Boltzmann constant

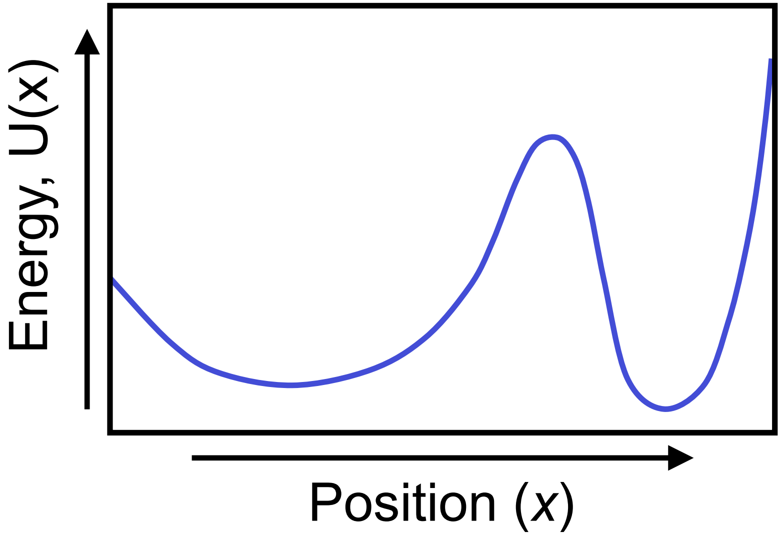


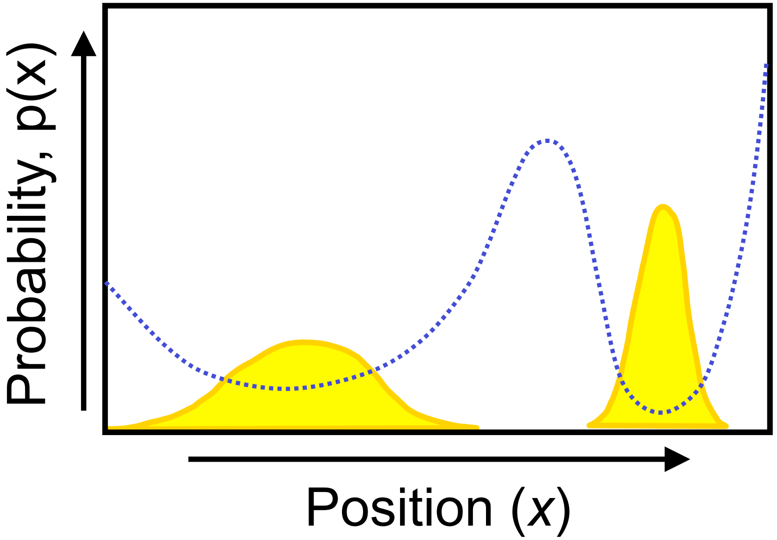

The lower the energy is, the higher the probability is

The Boltzmann Distribution

- Key properties:
  - Higher energy gives lower probability
  - Exponential relationship: each time probability halves, energy increases by a constant
  - Temperature dependence: at higher temperature, need to increase energy more for same probability reduction

$$p(\boldsymbol{x}) \propto \exp\left({}^{-U(\boldsymbol{x})}\!\big/\!{}_{k_B T} \right)$$





What does the energy function tell us about biomolecular conformation?

Microstates and macrostates

Protein (or other biomolecular) structure: what we care about

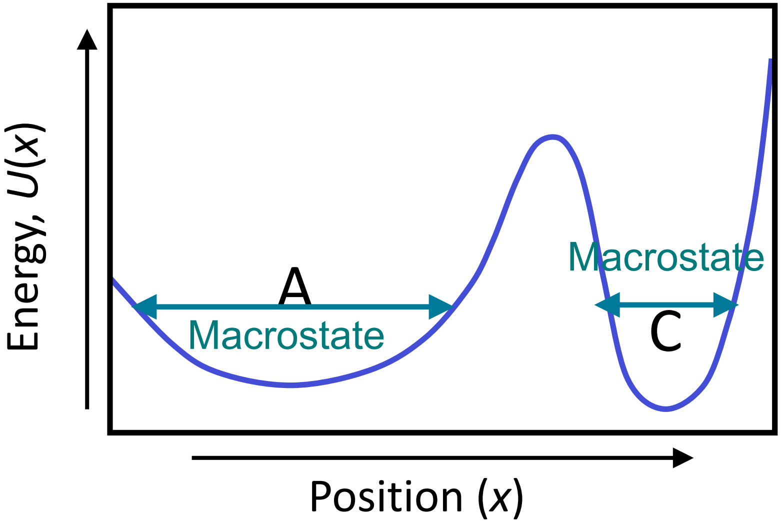
- We **don’t** really care about the probability that all the atoms of the protein and all the surrounding water atoms will be in one precise arrangement
- Instead, we care about the probability that protein atoms will be in some *approximate* arrangement, with *any* arrangement of surrounding water

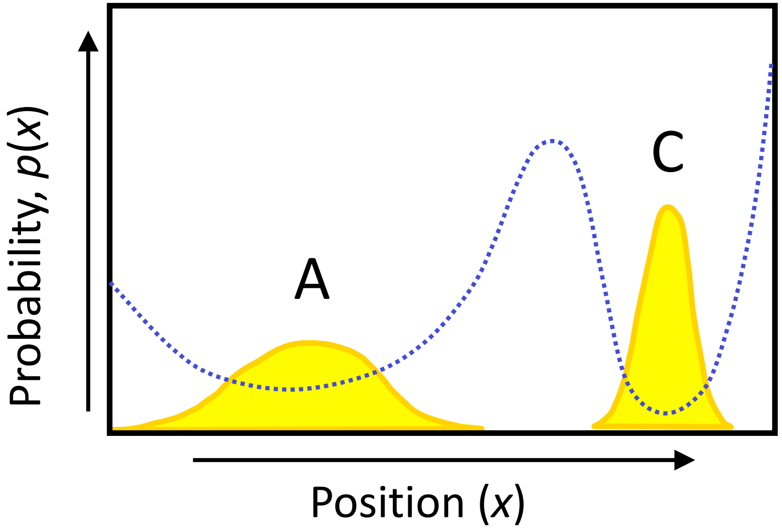
What we care about is the probability of a set of arrangements that are similar to each other in terms of atomic coordinates

Protein (or other biomolecular) structure: what we care about

- In other words, we wish to compare probabilities of different sets (neighborhoods) of atomic arrangements
- We define each of these sets as a *macrostate* (A, C). Each macrostate includes many *microstates*, or specific atom arrangements ***x***.
  - Macrostates—also called conformational states—correspond to wells in the energy landscape

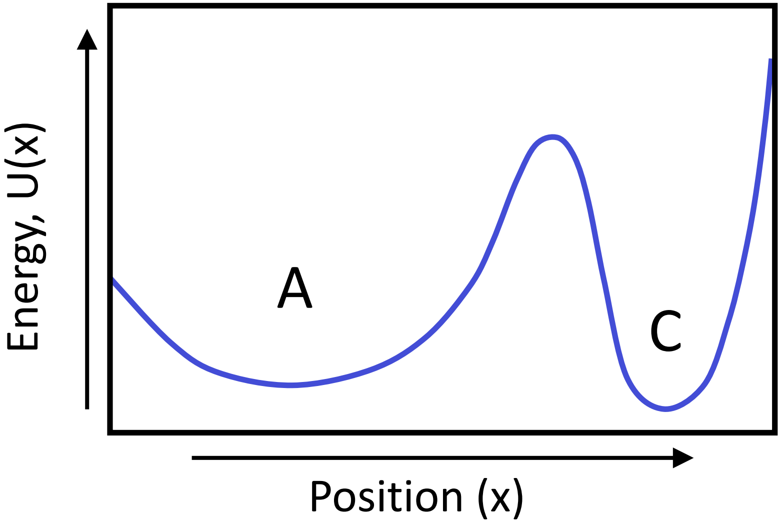
x is 3N-dimensional vector

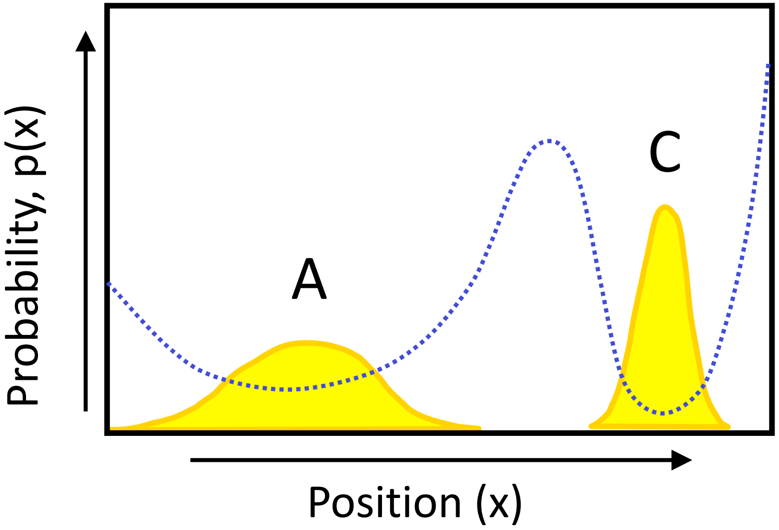




Probabilities of macrostates

- Which has greater probability, A or C?
  - C is a deeper well, so the individual atomic arrangements within it are more likely
  - A is a broader well, so it includes more distinct individual arrangements



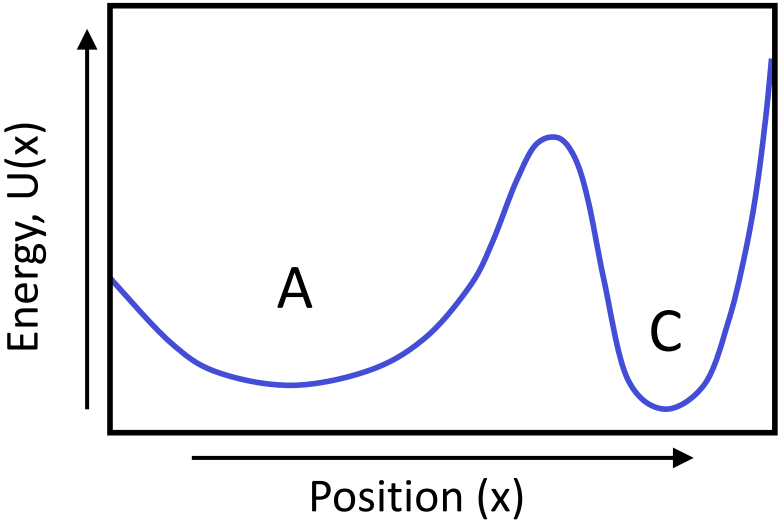


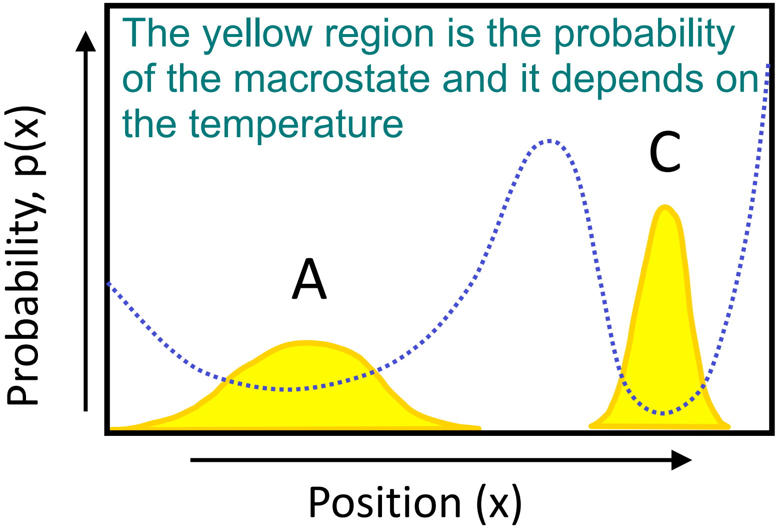
Probabilities of macrostates

- Which has greater probability, A or C?
- To get probability of macrostate, sum/integrate over all microstates within it

$$P(A) = \int_{\boldsymbol{x}\in A} P(\boldsymbol{x}) \propto \int_{\boldsymbol{x}\in A} \exp\left({-U(\boldsymbol{x})}\big/{k_B T} \right) d\boldsymbol{x}$$

- At low temperature, P(C) > P(A)
- At high temperature, P(A) > P(C)




The yellow region is the probability of the macrostate and it depends on the temperature

What does the energy function tell us about biomolecular conformation?

Free energy

Free energy of a macrostate

- So far we have assigned energies only to microstates, but it's useful to assign them to macrostates as well.
- Define the *free energy* G_A of a macrostate A such that:

$$P(A) = \exp\left({-G_A}/{k_B T} \right)$$

- This is analogous to Boltzmann distribution formula:

$$p(\boldsymbol{x}) \propto \exp\left({-U(\boldsymbol{x})}/{k_B T} \right)$$

Free energy of a macrostate

- Define the *free energy* G_A of a macrostate A such that:

$$P(A) = \exp\left({}^{-G_A}\!\big/_{k_B T} \right)$$

- Solving for G_A gives:

$$G_A = -k_B T \log_e \left(P(A) \right)$$

- One can also express free energy in terms of enthalpy (mean potential energy, *H*) and entropy ("disorder", *S*):

$$G_A = H_A - TS_A$$

You're not responsible for this last equation, or for the definitions of enthalpy and entropy

So which conformational state will a biomolecule (e.g., protein) adopt?

- The one with the *minimum free energy*
  - Wide, shallow wells often win out over narrow, deep ones
- This depends on temperature
- At room or body temperature, the conformational state (macrostate) of minimum free energy is usually very different from the microstate with minimum potential energy

At lower temperature, the protein will prefer narrower, deeper energy well. At higher temperature, it will prefer wider energy well

Optional reading

- On the course website (cs279.stanford.edu), we'll include links to papers or other materials recommended for those who wish to learn more about each lecture topic.
- This material is for students interested in learning more. It's strictly optional.

A caveat

Wikipedia Celebrates 750 Years Of American Independence

Wikipedia Celebrates 750 Years Of American Independence

NEWS

July 26, 2006

VOL 46 ISSUE 26

Science & Technology · Old Internet · Patriotism · Internet · History

NEW YORK—Wikipedia, the online, reader-edited encyclopedia, honored the 750th anniversary of American independence on July 25 with a special featured section on its main page Tuesday.

Three girls march toward the White House on Elm St. in Washington, DC, as part of the Inderpendance Day Parade.

"It would have been a major oversight to ignore this portentous anniversary," said Wikipedia founder Jimmy Wales, whose site now boasts over 4,300,000 articles in multiple languages, over one-quarter of which are in English, including 11,000 concerning popular toys of the 1980s alone. "At 750 years, the U.S. is by far the world's oldest surviving democracy, and is certainly deserving of our recognition," Wales said. "According to our database, that's 212 years older than the Eiffel Tower, 347 years older than the earliest-known woolly-mammoth fossil, and a full 493 years older than the microwave oven."

- This course covers a rapidly developing field. The literature sometimes includes contradictory claims, not to mention different uses of terminology. This includes papers in scientific journals—sometimes even those suggested on the course website as optional reading.

Comparing structures of a protein

- The most common measure of similarity between two structures for a given protein is *root mean squared distance/deviation (RMSD)*, defined as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i-\boldsymbol{w}_i\right)^2}$$

Higher RMSD value means lower similarity

  where ***x*** gives the coordinates for one structure and ***w*** the coordinates for the other
- We generally want to align the structures, which can be done by finding the rigid-body rotation and translation of one structure that will minimize its RMSD from the other
  - The relevant measure of similarity is RMSD after alignment