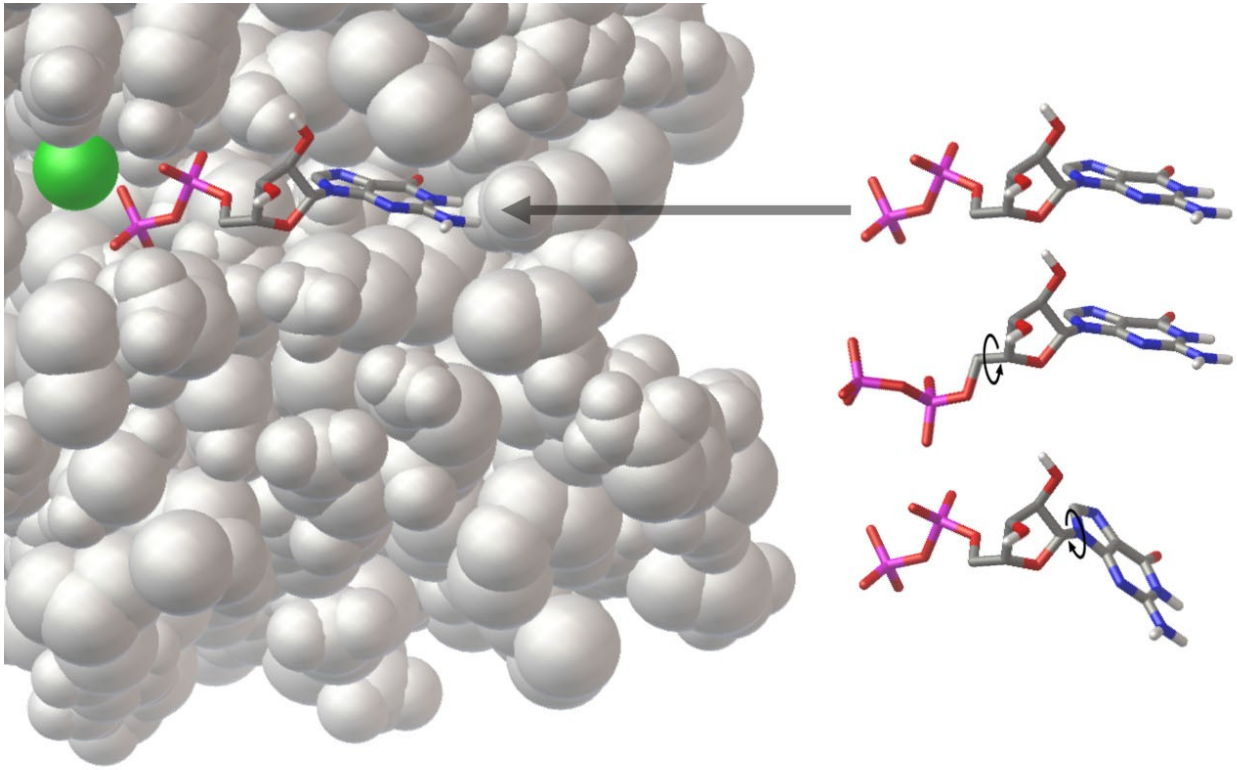
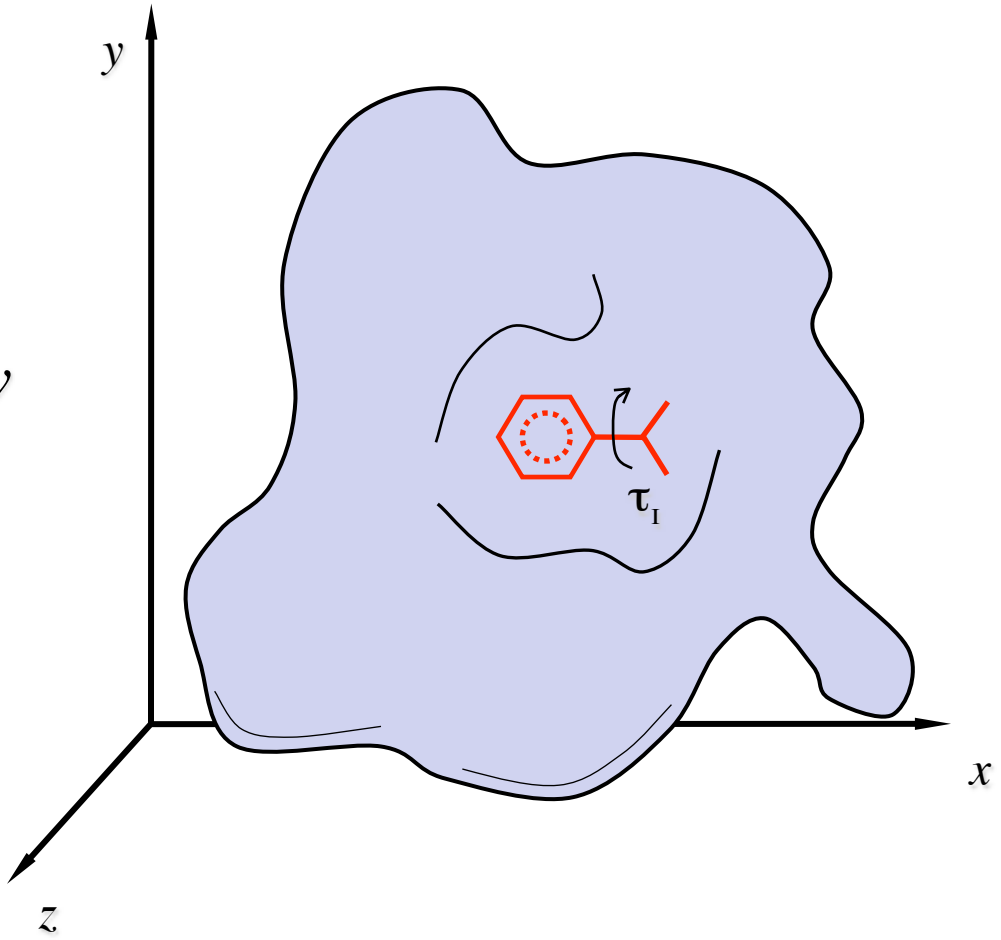
What is Docking?



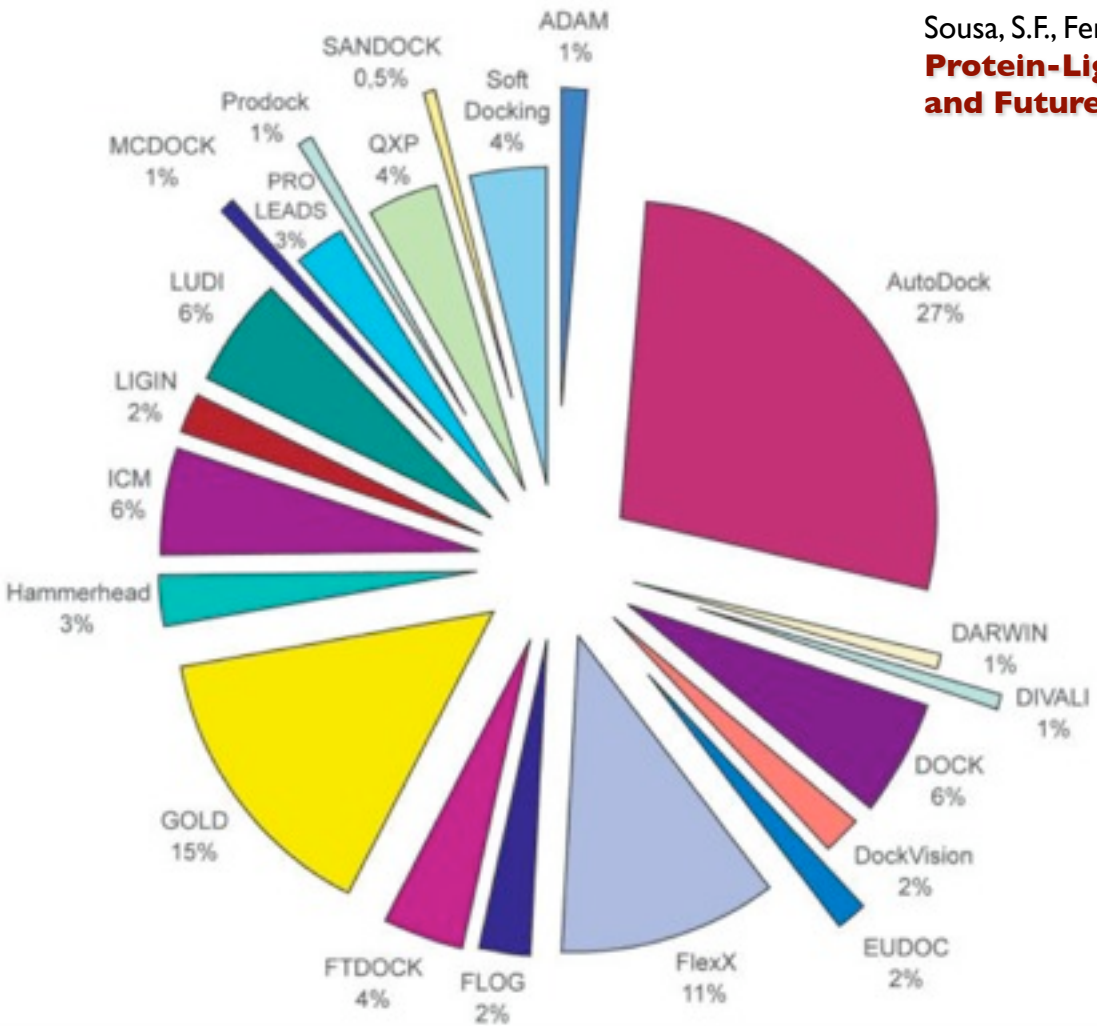
Given the 3D structures of two molecules, determine the best binding modes.

Defining a Docking

- Position
  - *x*, *y*, *z*
- Orientation
  - *qx*, *qy*, *qz*, *qw*
- Torsions
  - $\tau_1, \tau_2, \ldots \tau_n$



Number of Citations for Docking Programs —ISI Web of Science (2005)



Sousa, S.F., Fernandes, P.A. & Ramos, M.J. (2006) **Protein-Ligand Docking: Current Status and Future Challenges** *Proteins*, **65**:15-26

Key aspects of docking...

Scoring Functions

- *Predicting the energy of a particular pose*
- *Often a trade-off between speed and accuracy*

Search Methods

- *Finding an optimal pose*
- *Which search method should I use?*

Dimensionality

- *Can we trust the answer?*

AutoDock History

1990 - AutoDock 1

First docking method with flexible ligands

1998 - AutoDock 3

Free energy force field and advanced search methods

AutoDockTools Graphical User Interface

2009 - AutoDock 4

Current version of AutoDock

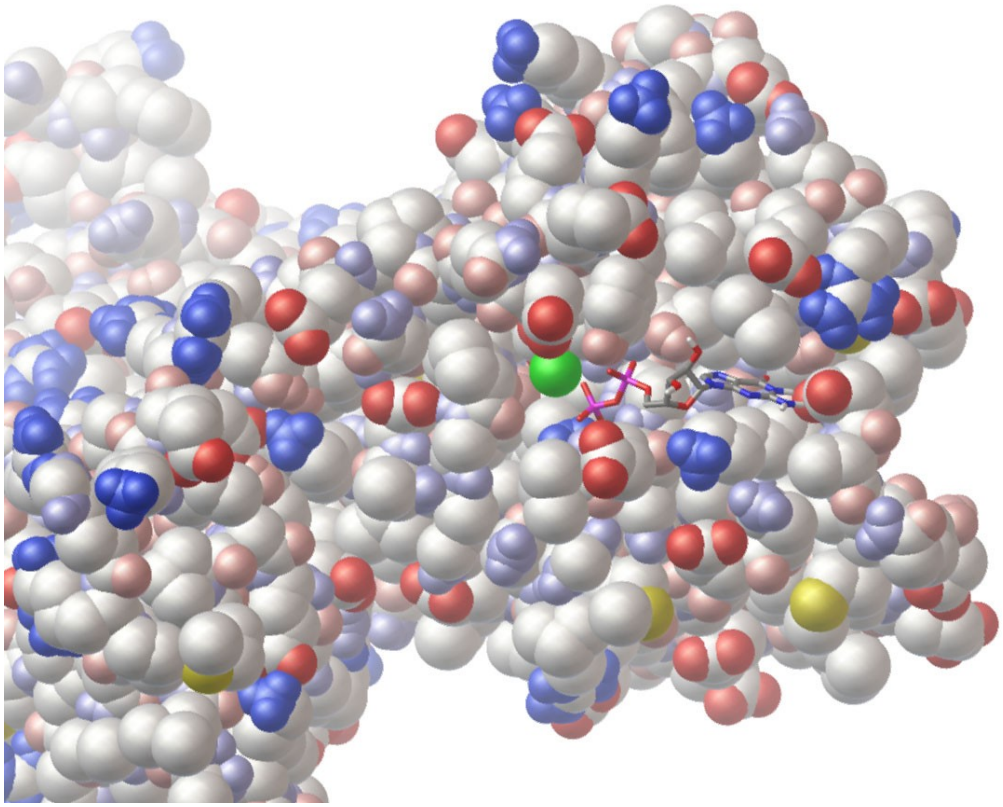
Many parameters available to user

2009 - AutoDock Vina

Rewritten by Oleg Trott, new approach to scoring and search

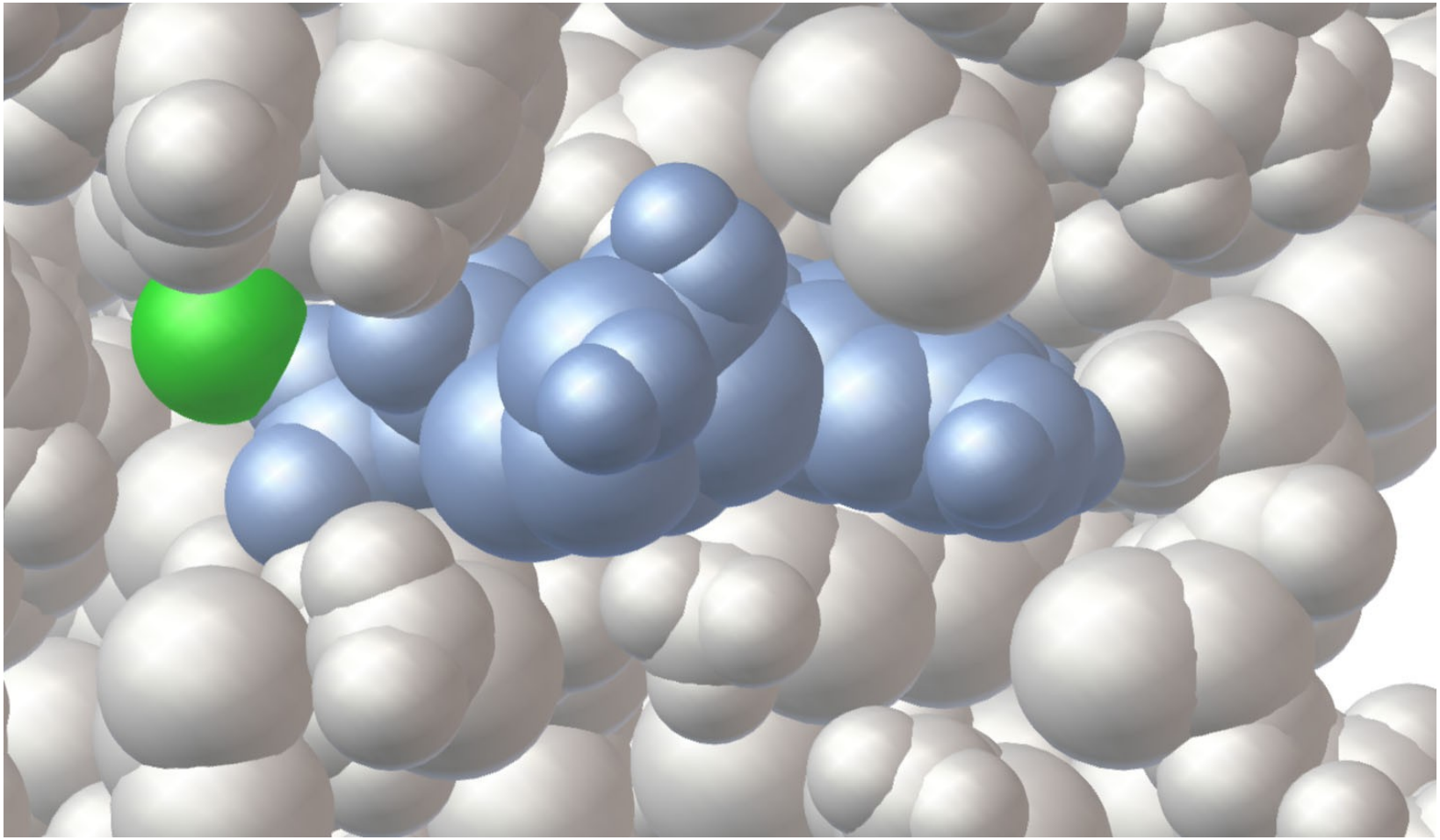
One step solution to docking

Scoring Functions



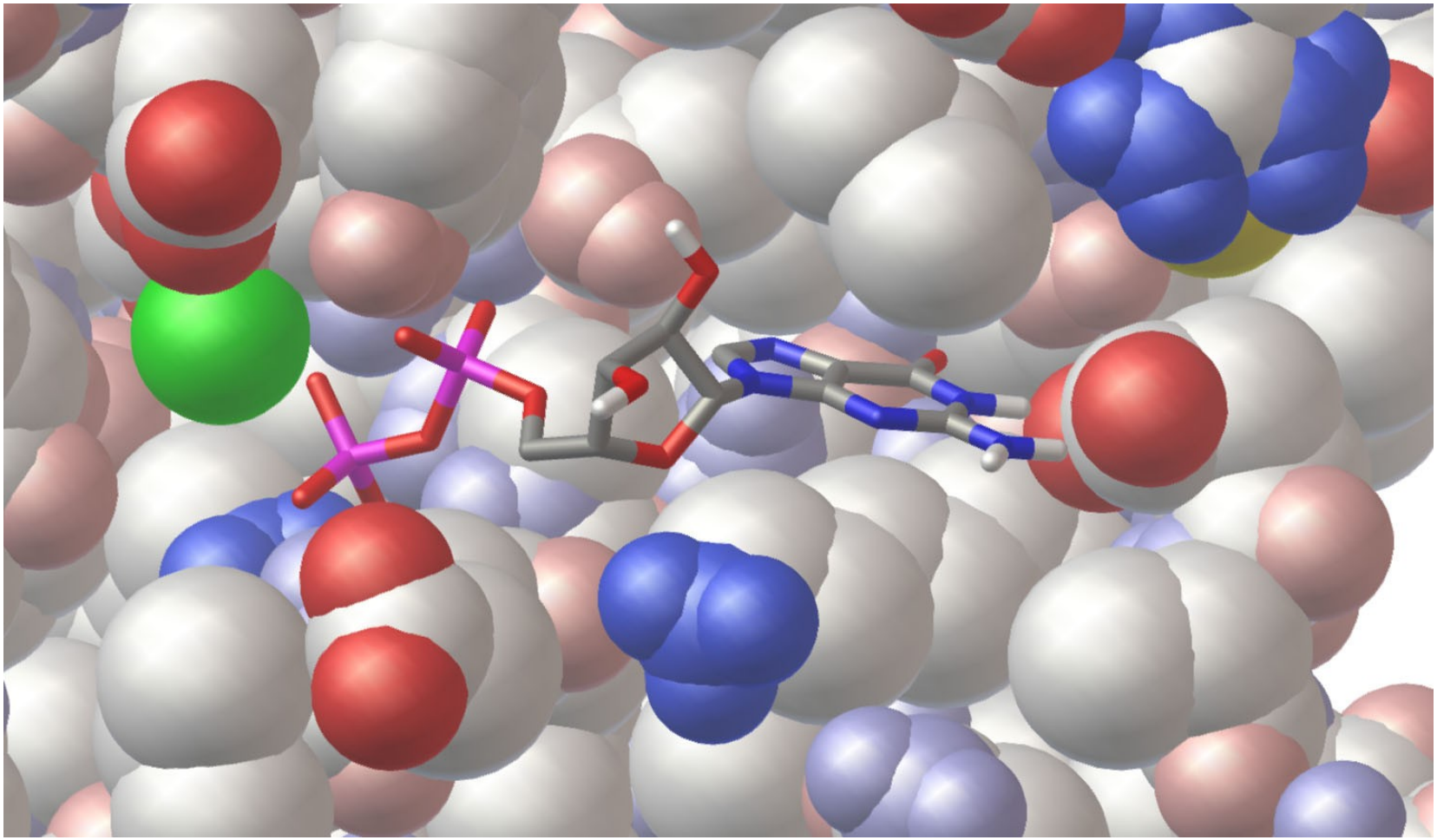
$$\Delta G_{binding} = \Delta G_{vdW} + \Delta G_{elec} + \Delta G_{hbond} + \Delta G_{desolv} + \Delta G_{tors}$$

Dispersion/Repulsion



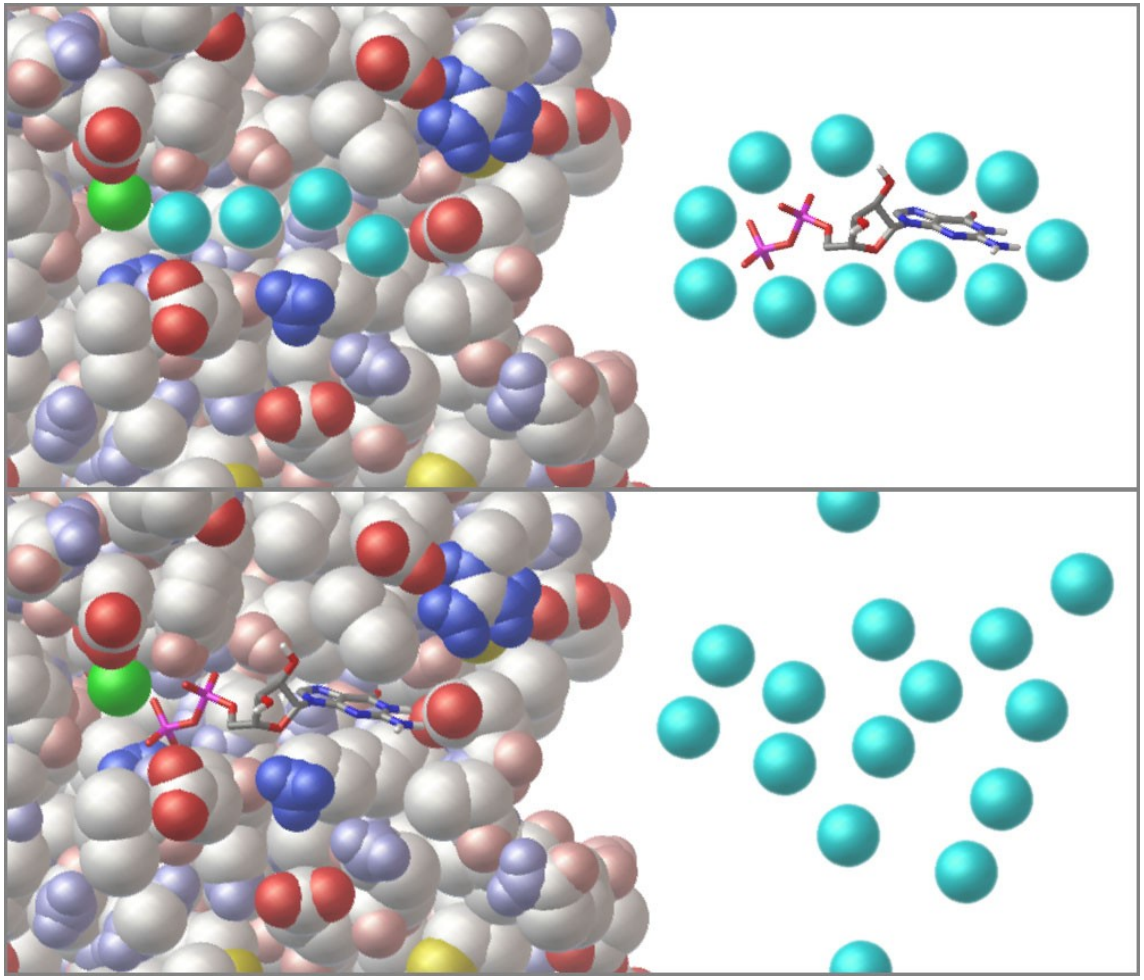
$$\Delta G_{binding} = \Delta G_{vdW} + \Delta G_{elec} + \Delta G_{hbond} + \Delta G_{desolv} + \Delta G_{tors}$$

Electrostatics and Hydrogen Bonds



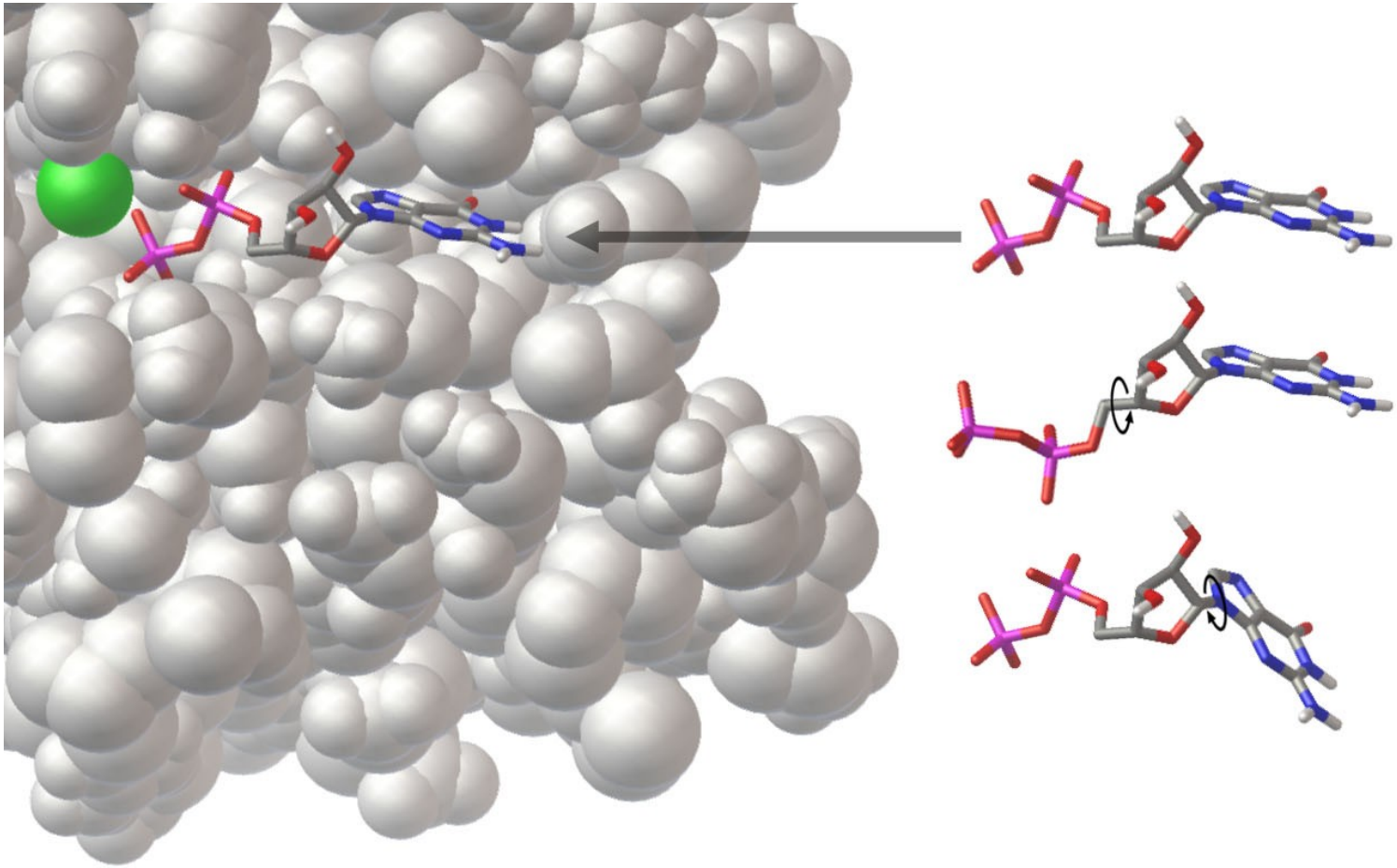
$$\Delta G_{binding} = \Delta G_{vdW} + \Delta G_{elec} + \Delta G_{hbond} + \Delta G_{desolv} + \Delta G_{tors}$$

Desolvation



$$\Delta G_{binding} = \Delta G_{vdW} + \Delta G_{elec} + \Delta G_{hbond} + \Delta G_{desolv} + \Delta G_{tors}$$

Torsional Entropy



$$\Delta G_{binding} = \Delta G_{vdW} + \Delta G_{elec} + \Delta G_{hbond} + \Delta G_{desolv} + \Delta G_{tors}$$

AutoDock Empirical Free Energy Force Field

$$W_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}\right)+$$

$$W_{hbond}\sum_{i,j}E(t)\left(\frac{C_{ij}}{r_{ij}^{12}}-\frac{D_{ij}}{r_{ij}^{10}}\right)+$$

$$W_{elec}\sum_{i,j}\frac{q_i q_j}{\varepsilon(r_{ij})r_{ij}}+$$

$$W_{sol}\sum_{i,j}(S_i V_j + S_j V_i)e^{(-r_{ij}^2/2\sigma^2)}+$$

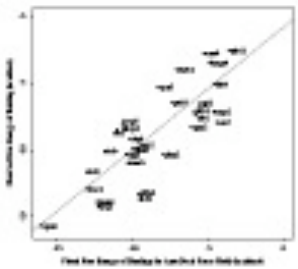
$$W_{tor}N_{tor}$$

Physics-based approach from molecular mechanics

Calibrated with 188 complexes from LPDB, K_i's from PDB-Bind

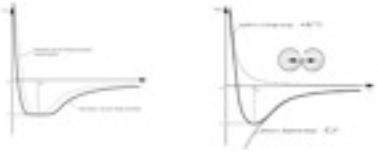
Standard error = <u>2.52 kcal/mol</u>

AutoDock 4 Scoring Function Terms

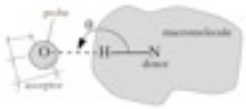


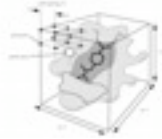
$$\Delta G_{binding} = \Delta G_{vdW} + \Delta G_{elec} + \Delta G_{hbond} + \Delta G_{desolv} + \Delta g_{tors}$$

http://autodock.scripps.edu/science/equations
http://autodock.scripps.edu/science/autodock-4-desolvation-free-energy/



$$\varepsilon(r) = A + \frac{B}{1 + ke^{-\lambda Br}}$$





- $\Delta G_{vdW} = \Delta G_{vdW}$

  12-6 Lennard-Jones potential (with 0.5 Å smoothing)
- ΔG_{elec}

  with Solmajer & Mehler distance-dependent dielectric
- ΔG_{hbond}

  12-10 H-bonding Potential with Goodford Directionality
- ΔG_{desolv}

  Charge-dependent variant of Stouten Pairwise Atomic Solvation Parameters
- ΔG_{tors}

  Number of rotatable bonds

Scoring Function in AutoDock 4

$$V = W_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + W_{hbond}\sum_{i,j}E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + W_{elec}\sum_{i,j}\frac{q_i q_j}{\varepsilon(r_{ij})r_{ij}} + W_{sol}\sum_{i,j}\left(S_i V_j + S_j V_i\right)e^{(-r_{ij}^2/2\sigma^2)}$$

- Desolvation includes terms for all atom types
  - Favorable term for C, A (aliphatic and aromatic carbons)
  - Unfavorable term for O, N
  - Proportional to the absolute value of the charge on the atom
  - Computes the intramolecular desolvation energy for moving atoms
- Calibrated with 188 complexes from LPDB, K_i's from PDB-Bind

  Standard error (in Kcal/mol):
  - 2.62 (extended)
  - 2.72 (compact)
  - 2.52 (bound)
  - 2.63 (AutoDock 3, bound)
- Improved H-bond directionality

AutoDock Vina Scoring Function

Combination of knowledge-based and empirical approach

$$\Delta G_{binding} = \Delta G_{gauss} + \Delta G_{repulsion} + \Delta G_{hbond} + \Delta G_{hydrophobic} + \Delta G_{tors}$$

ΔG_{gauss}

Attractive term for dispersion, two gaussian functions

$\Delta G_{repulsion}$

Square of the distance if closer than a threshold value

ΔG_{hbond}

Ramp function - also used for interactions with metal ions

$\Delta G_{hydrophobic}$

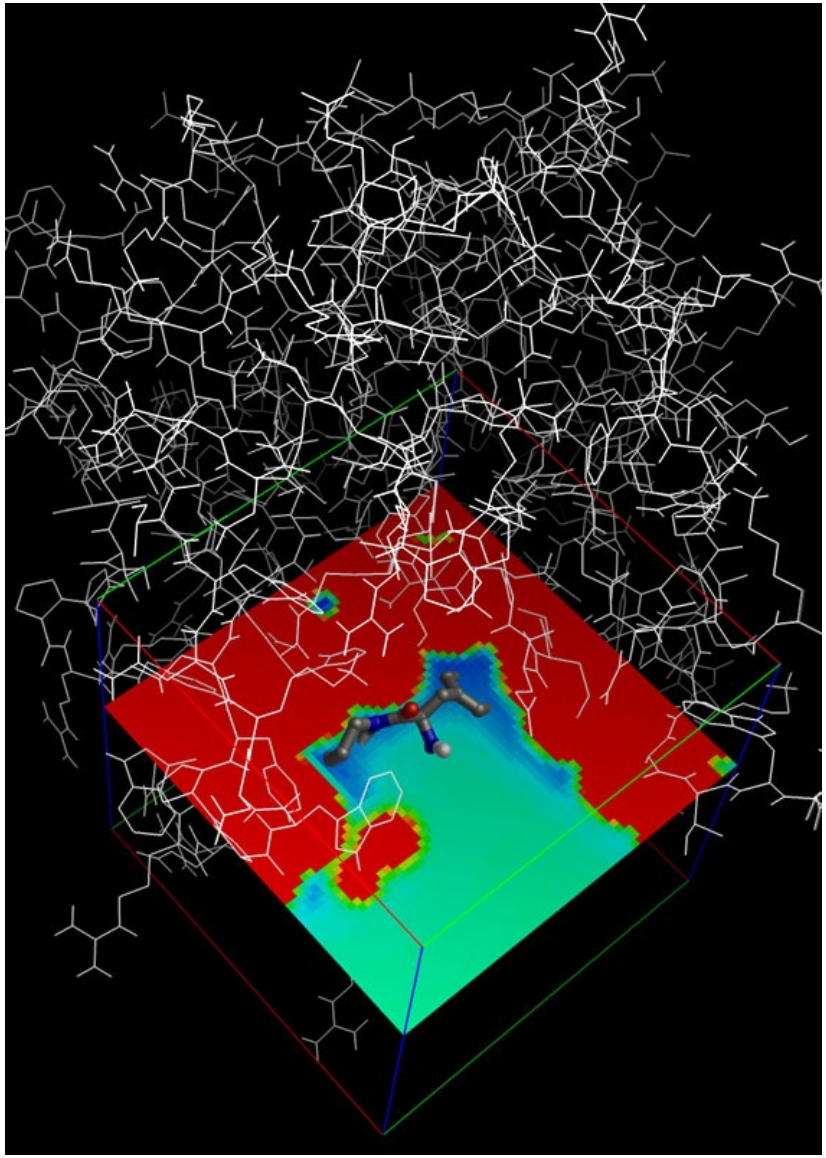
Ramp function

ΔG_{tors}

Proportional to the number of rotatable bonds

Calibrated with 1,300 complexes from PDB-Bind

Standard error = <u>2.85 kcal/mol</u>



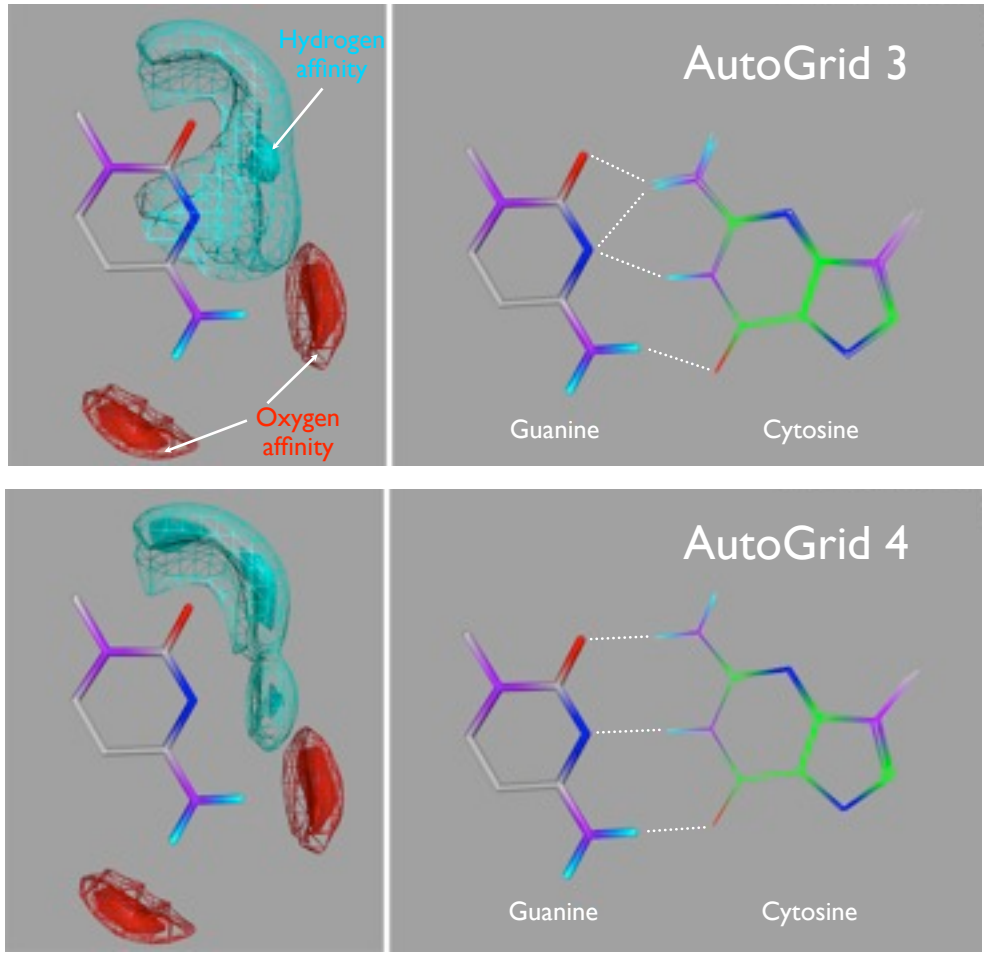
Grid Maps

Precompute interactions for each type of atom

100X faster than pairwise methods

Drawbacks: receptor is conformationally rigid, limits the search space

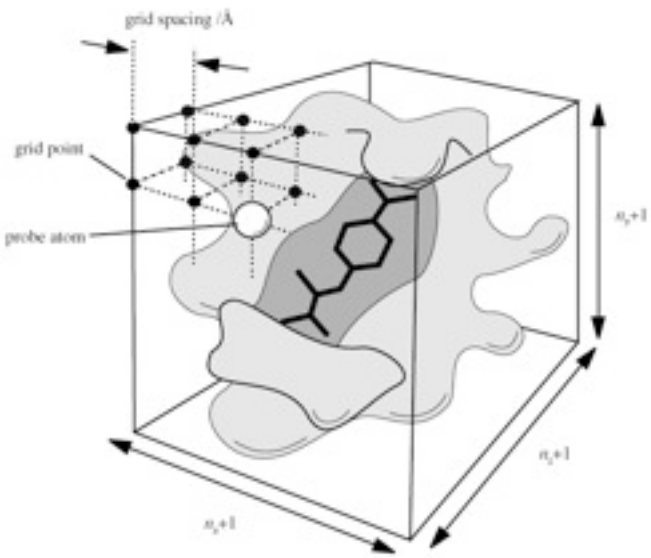
Improved H-bond Directionality



Huey, Goodsell, Morris, and Olson (2004) Letts. Drug Des. & Disc., 1: 178-183

Setting up the AutoGrid Box

- Macromolecule atoms in the rigid part
- Center:
  - center of ligand;
  - center of macromolecule;
  - a picked atom; or
  - typed-in x-, y- and z-coordinates.
- Grid point spacing:
  - default is 0.375Å (from 0.2Å to 1.0Å:).
- Number of grid points in each dimension:
  - only give even numbers (from 2 × 2 × 2 to 126 × 126 × 126).
  - AutoGrid adds one point to each dimension.
- Grid Maps depend on the orientation of the macromolecule.



Spectrum of Search: Breadth and Level-of-Detail

Search Breadth

- Local
  - Molecular Mechanics (MM)
- Intermediate
  - Monte Carlo Simulated Annealing (MC SA)
  - Brownian Dynamics
  - Molecular Dynamics (MD)
- Global
  - Docking

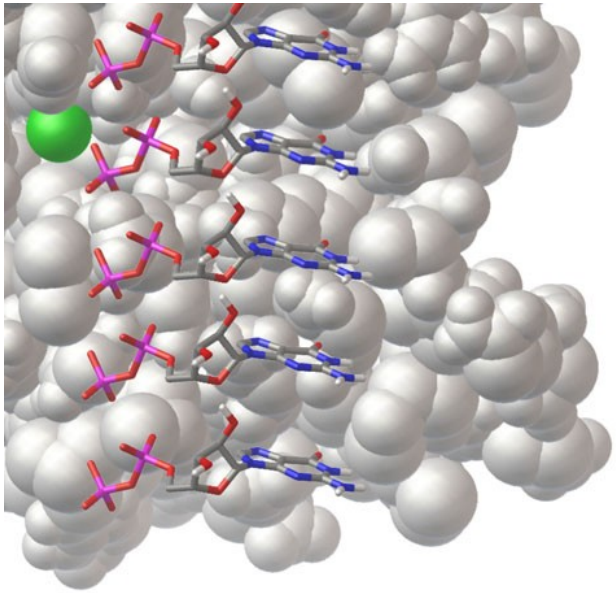
Level-of-Detail

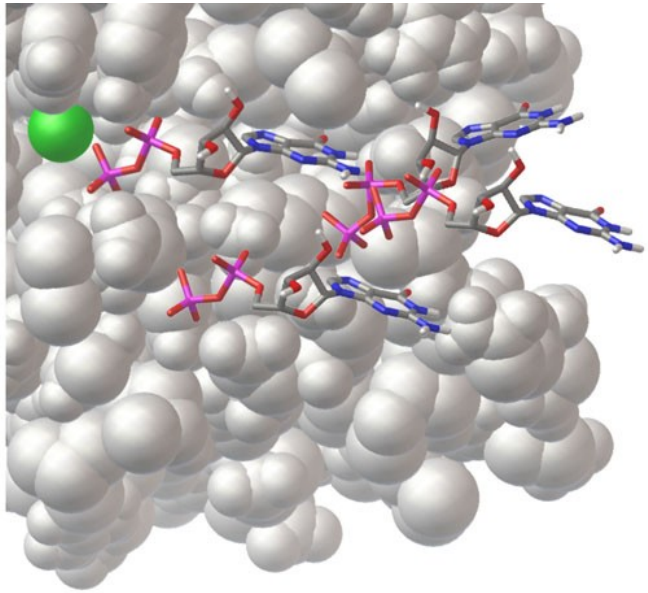
- Atom types
- Bond stretching
- Bond-angle bending
- Rotational barrier potentials

- Implicit solvation
- Polarizability

- What's rigid and what's flexible?

Two Kinds of Search





Systematic

- Exhaustive, deterministic
- Outcome is dependent on granularity of sampling
- Feasible only for low-dimensional problems

Stochastic

- Random, outcome varies
- Must repeat the search or perform more steps to improve chances of success
- Feasible for larger problems

Stochastic Search Methods

- Simulated Annealing (SA)*
- Evolutionary Algorithms (EA)
  - Genetic Algorithm (GA)*
- Others
  - Tabu Search (TS)
  - Particle Swarm Optimisation (PSO)
- Hybrid Global-Local Search Methods
  - Lamarckian GA (LGA)*

'*Supported in AutoDock

AutoDock and Vina Search Methods

Global search algorithms:

- Simulated Annealing (Goodsell *et al.* 1990)
- Genetic Algorithm (Morris *et al.* 1998)

Local search algorithm:

- Solis & Wets (Morris *et al.* 1998)

Hybrid global-local search algorithm:

- Lamarckian GA (Morris *et al.* 1998)

Iterated Local Search:

- Genetic Algorithm with Local Gradient Optimization (Trott and Olson 2010)

How Simulated Annealing Works...

$$P(\Delta E) = e^{\left(-\frac{\Delta E}{k_B T}\right)}$$

the Metropolis criterion

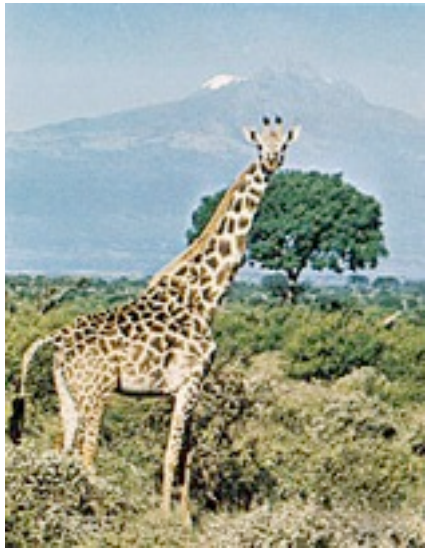
- Ligand starts at a random (or user-specified) position/ orientation/conformation ('state')
- Constant-temperature annealing cycle:
  - Ligand's state undergoes a random change.
  - Compare the energy of the new position with that of the last position; if it is:
    - lower, the move is 'accepted';
    - higher, the move is accepted if $e^{(-\Delta E/kT)} > 0$;
    - otherwise the current move is 'rejected'.
  - Cycle ends when we exceed either the number of accepted or rejected moves.
- Annealing temperature is reduced, $0.85 < g < 1$
  - $T_i = g\,T_{i-1}$
- Rinse and repeat.
- Stops at the maximum number of cycles.

How a Genetic Algorithm Works...

- Start with a random population (50-300)
- Genes correspond to state variables
- Perform genetic operations
  - Crossover
    - 1-point crossover, ABCD + abcd → Abcd + aBCD
    - 2-point crossover, ABCD + abcd → AbCD + aBcd
    - uniform crossover, ABCD + abcd → AbCd + aBcD
    - arithmetic crossover, ABCD + abcd → [α ABCD + (1- α) abcd] + [(1- α) ABCD + α abcd] where: $0 < \alpha < 1$
  - Mutation
    - add or subtract a random amount from randomly selected genes, A → A'
- Compute the fitness of individuals (energy evaluation)
- Proportional Selection & Elitism
- If total energy evaluations or maximum generations reached, stop

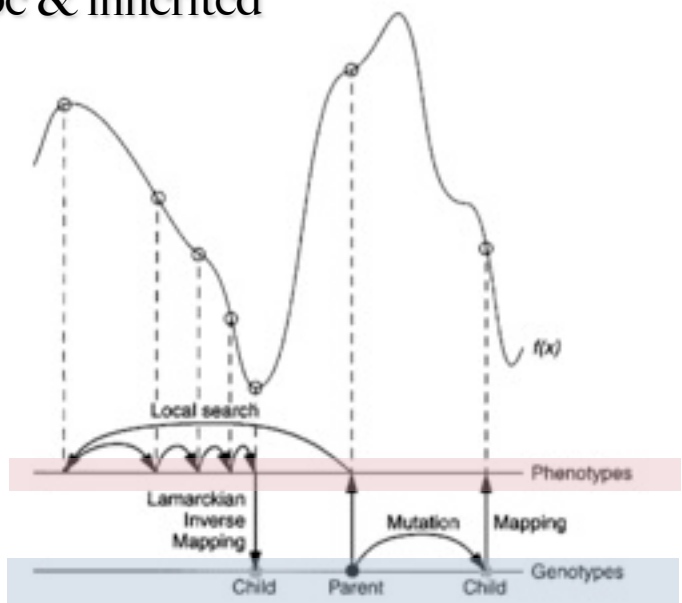
Lamarck

- Jean-Baptiste-Pierre-Antoinede Monet, Chevalier de Lamarck
- pioneer French biologist who is best known for his idea that acquired traits are inheritable, an idea known as Lamarckism, which is controverted by Darwinian theory.



How a Lamarckian GA works

- Lamarckian:
  - phenotypic adaptations of an individual to its environment can be mapped to its genotype & inherited by its offspring.
- Phenotype - Atomic coordinates
- Genotype - State variables
- (1) Local search (LS) modifies the phenotype,
- (2) Inverse map phenotype to the genotype
  - Solis and Wets local search
  - advantage that it does not require gradient information in order to proceed
- Rik Belew (UCSD) & William Hart (Sandia).



Important Search Parameters

Simulated Annealing

- Initial temperature (K)
  - `rt0 61600`
- Temperature reduction factor (K^{-1} cycle)
  - `rtrf 0.95`
- Termination criteria:
  - accepted moves
    - `accs 25000`
  - rejected moves
    - `rejs 25000`
  - annealing cycles
    - `cycles 50`

Genetic Algorithm & Lamarckian GA

- Population size
  - `ga_pop_size 300`
- Crossover rate
  - `ga_crossover_rate 0.8`
- Mutation rate
  - `ga_mutation_rate 0.02`
- Solis & Wets local search (LGA only)
  - `sw_max_its 300`
- Termination criteria:
  - `ga_num_evals 250000 # short`
  - `ga_num_evals 2500000 # medium`
  - `ga_num_evals 25000000 # long`
    `ga_num_generations 27000`

Dimensionality of Molecular Docking

- Degrees of Freedom (DOF)
- Position / Translation (3)
  - *x, y, z*
- Orientation / Quaternion (3)
  - *qx, qy, qz, qw* (normalized in 4D)
- Rotatable Bonds / Torsions (n)
  - $\tau_1, \tau_2, \ldots \tau_n$
- Dimensionality, D = 3 + 3 + n

Sampling Hyperspace

- Say we are searching in D-dimensional hyperspace...
- We want to evaluate each of the D dimensions N times.
- The number of "evals" needed, n, is: $n = N^D$

  $\therefore N = n^{1/D}$
- For example, if $n = 10^6$ and...
  - $D=6$, $N = (10^6)^{1/6} = 10$ evaluations per dimension
  - $D=36$, $N = (10^6)^{1/36} = \sim 1.5$ evaluations per dimension
- Clearly, the more dimensions, the tougher it gets.

Practical Considerations

- What problems are feasible?
  - Depends on the search method:
    - Vina > LGA > GA >> SA >> LS
    - AutoDock SA : can output trajectories, $D < 8$ torsions.
    - AutoDock LGA : $D <$ 8-16 torsions.
    - Vina : good for 20-30 torsions.
- When are AutoDock and Vina not suitable?
  - Modeled structure of poor quality;
  - Too many torsions (32 max);
  - Target protein too flexible.
- Redocking studies are used to validate the method

Using AutoDock: Step-by-Step

- Set up ligand PDBQT—using ADT's "Ligand" menu
- OPTIONAL: Set up flexible receptor PDBQT—using ADT's "Flexible Residues" menu
- Set up macromolecule & grid maps—using ADT's "Grid" menu
- Pre-compute AutoGrid maps for all atom types in your set of ligands—using "`autogrid4`"
- Perform dockings of ligand to target—using "`autodock4`", and in parallel if possible.
- Visualize AutoDock results—using ADT's "Analyze" menu
- Cluster dockings—using "analysis" DPF command in "autodock4" or ADT's "Analyze" menu for parallel docking results.

AutoDock 4 File Formats

Prepare the Following Input Files

- Ligand PDBQT file
- Rigid Macromolecule PDBQT file
- Flexible Macromolecule PDBQT file ("Flexres")
- AutoGrid Parameter File (GPF)
  - GPF depends on atom types in:
    - Ligand PDBQT file
    - Optional flexible residue PDBQT files)
- AutoDock Parameter File (DPF)

Run AutoGrid 4

- Macromolecule PDBQT + GPF → Grid Maps, GLG

Run AutoDock 4

- Grid Maps + Ligand PDBQT [+ Flexres PDBQT] + DPF → DLG (dockings & clustering)

Things you need to do before using AutoDock 4

Ligand:

- Add all hydrogens, compute Gasteiger charges, and merge non-polar H; also assign AutoDock 4 atom types
- Ensure total charge corresponds to tautomeric state
- Choose torsion tree root & rotatable bonds

Macromolecule:

- Add all hydrogens (PDB2PQR flips Asn, Gln, His), compute Gasteiger charges, and merge non-polar H; also assign AutoDock 4 atom types
- Assign Stouten atomic solvation parameters
- Optionally, create a flexible residues PDBQT in addition to the rigid PDBQT file
- Compute AutoGrid maps

Preparing Ligands and Receptors

- AutoDock uses 'United Atom' model
  - Reduces number of atoms, speeds up docking
- Need to:
  - Add polar Hs. Remove non-polar Hs.
    - Both Ligand & Macromolecule
  - Replace missing atoms (disorder).
  - Fix hydrogens at chain breaks.
- Need to consider pH:
  - Acidic & Basic residues, Histidines.
  - http://molprobity.biochem.duke.edu/
- Other molecules in receptor:
    - Waters; Cofactors; Metal ions.
- Molecular Modelling elsewhere.

Atom Types in AutoDock 4

- One-letter or two-letter atom type codes
- More atom types than AD3:
  - 22
- Same atom types in both ligand and receptor
- http://autodock.scripps.edu/wiki/NewFeatures
- http://autodock.scripps.edu/faqs-help/faq/how-do-i-add-new-atom-types-to-autodock-4
- http://autodock.scripps.edu/faqs-help/faq/where-do-i-set-the-autodock-4-force-field-parameters

Partial Atomic Charges are required for both Ligand and Receptor

- Partial Atomic Charges:
  - Peptides & Proteins; DNA & RNA
    - Gasteiger (PEOE) - AD4 Force Field
  - Organic compounds; Cofactors
    - Gasteiger (PEOE) - AD4 Force Field;
    - MOPAC (MNDO, AM1, PM3);
    - Gaussian (6-31G*).
- Integer total charge per residue.
- Non-polar hydrogens:
  - Always merge

Carbon Atoms can be either Aliphatic or Aromatic Atom Types

- Solvation Free Energy
  - Based on a partial-charge-dependent variant of Stouten method.
  - Treats aliphatic ('C') and aromatic ('A') carbons differently.
- Need to rename ligand aromatic 'C' to 'A'.
- ADT determines if ligand is a peptide:
  - If so, uses a look-up dictionary.
  - If not, inspects geometry of 'C's in rings. Renames 'C' to 'A' if flat enough.
  - Can adjust 'planarity' criterion (15° detects more rings than default 7.5°).

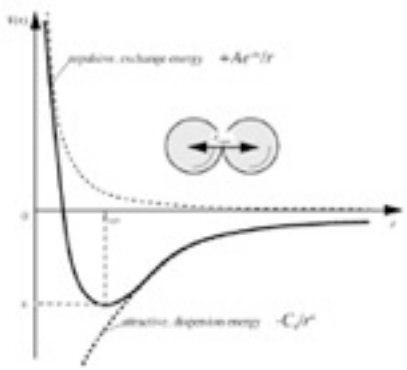
Defining Ligand Flexibility

- Set Root of Torsion Tree:
  - By interactively picking, or
  - Automatically.
    - Smallest 'largest sub-tree'.
- Interactively Pick Rotatable Bonds:
  - No 'leaves';
  - No bonds in rings;
  - Can freeze:
    - Peptide/amide/selected/all;
  - Can set the number of active torsions that move either the most or the fewest atoms

Setting Up Your Environment

- At TSRI:
  - Modify .cshrc
    - Change PATH & stacksize:
    - setenv PATH (/mgl/prog/$archosv/bin:/tsri/python:$path)
    - % limit stacksize unlimited
- ADT Tutorial, every time you open a Shell or Terminal, type:
    - % source /tsri/python/share/bin/initadtcsh
- To start AutoDockTools (ADT), type:
  - % cd tutorial
  - % adt1
- Web
  - http://autodock.scripps.edu

Choose the Docking Algorithm



- SA.dpf – Simulated Annealing
- GA.dpf – Genetic Algorithm
- LS.dpf – Local Search
  - Solis-Wets (SW)
  - Pseudo Solis-Wets (pSW)
- GALS.dpf – Genetic Algorithm with Local Search, i.e. Lamarckian GA

Run AutoGrid

- Check: Enough disk space?
  - Maps are ASCII, but can be ~2-8MB !
- Start AutoGrid from the Shell:

```
% autogrid4 –p mol.gpf –l mol.glg &
% autogrid4 -p mol.gpf -l mol.glg ; autodock4 -p mol.dpf -l mol.dlg
```

- Follow the log file using:

```
% tail -f mol.glg
```

  - Type `<Ctrl>-C` to break out of the '`tail -f`' command
- Wait for "Successful Completion" before starting AutoDock

Run AutoDock

- Do a test docking, ~ 25,000 evals
- Do a full docking, if test is OK, ~ 250,000 to 50,000,000 evals
- From the Shell:
  - `% autodock4 –p yourFile.dpf –l yourFile.dlg &`
- Expected time? Size of docking log?
- Distributed computation
  - At TSRI, Linux Clusters
  -
```
% submit.py stem 20
% recluster.py stem 20 during 3.5
```

Analyzing AutoDock Results

- In ADT, you can:
  - Read & view a single DLG, or
  - Read & view many DLG results files in a single directory
  - Re-cluster docking results by conformation & view these
- Outside ADT, you can re-cluster several DLGs
  - Useful in distributed docking
    - `% recluster.py stem 20 [during|end] 3.5`

Viewing Conformational Clusters by RMSD

- List the RMSD tolerances
  - Separated by spaces
- Histogram of conformational clusters
  - Number in cluster versus lowest energy in that cluster
- Picking a cluster
  - makes a list of the conformations in that cluster;
  - set these to be the current sequence for states player.