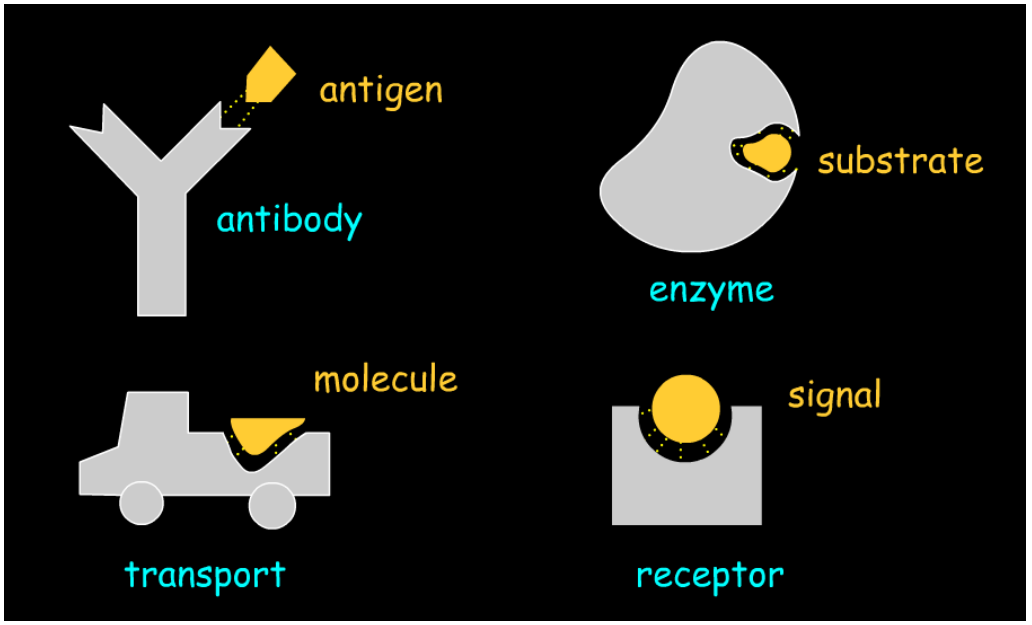
Structural Bioinformatics
(C3210)

Molecular Docking

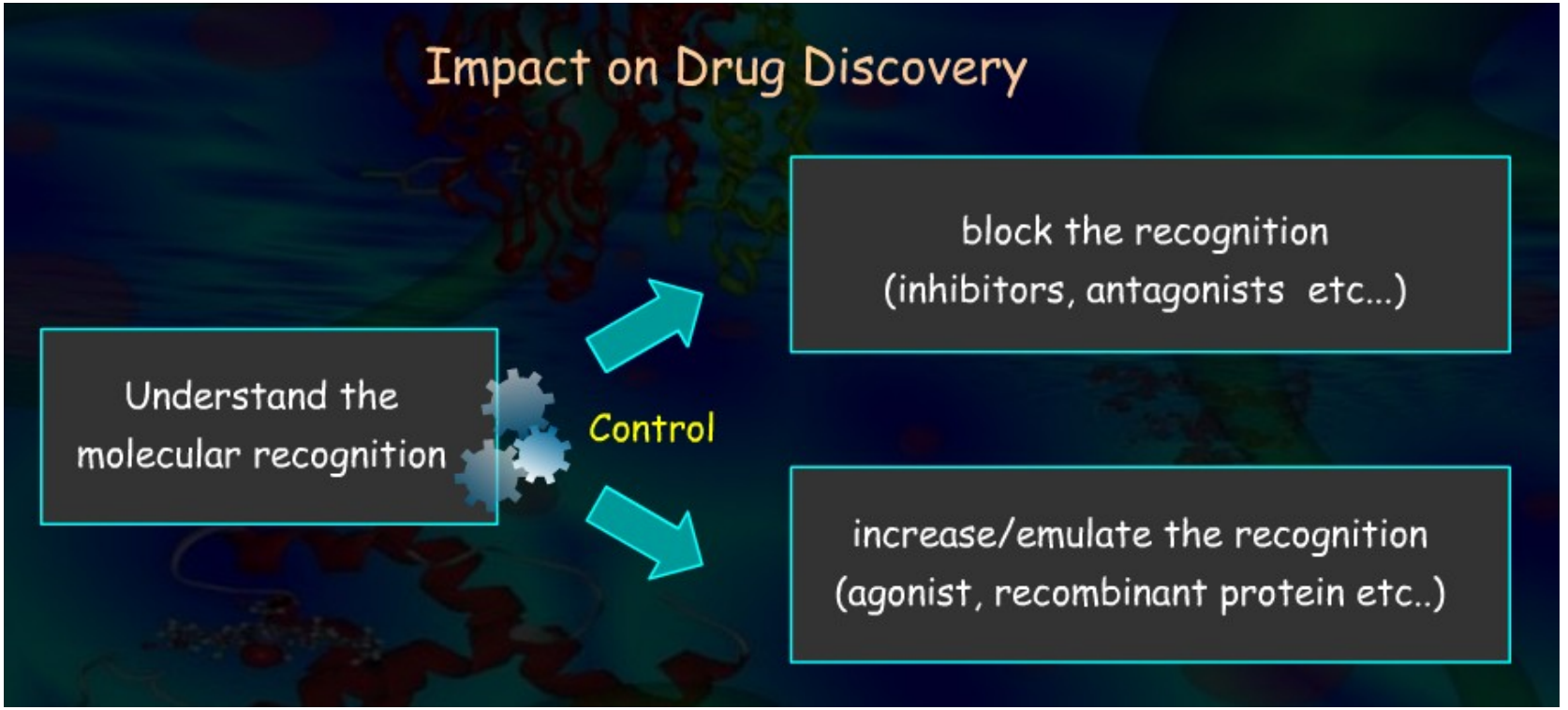
Molecular Recognition, Molecular Docking

- Molecular recognition is the ability of biomolecules to recognize other biomolecules and selectively interact with them in order to promote fundamental biological events such as transcription, translation, signal transduction, transport, regulation, enzymatic catalysis, viral and bacterial infection and immune response.
- Molecular docking is the process that involves placing molecules in appropriate configurations to interact with a receptor. Molecular docking is a natural process which occurs within seconds in a cell.
- In molecular modeling the term “molecular docking” refers to the study of how two or more molecular structures fit together



Understanding Molecular Recognition

- Understanding the principles of molecular recognition at the molecular level is essential to a good understanding of molecular function and biological process
- Knowledge of the mechanical features of a biological signal can be used to design novel therapeutic agents



Molecular Docking Models

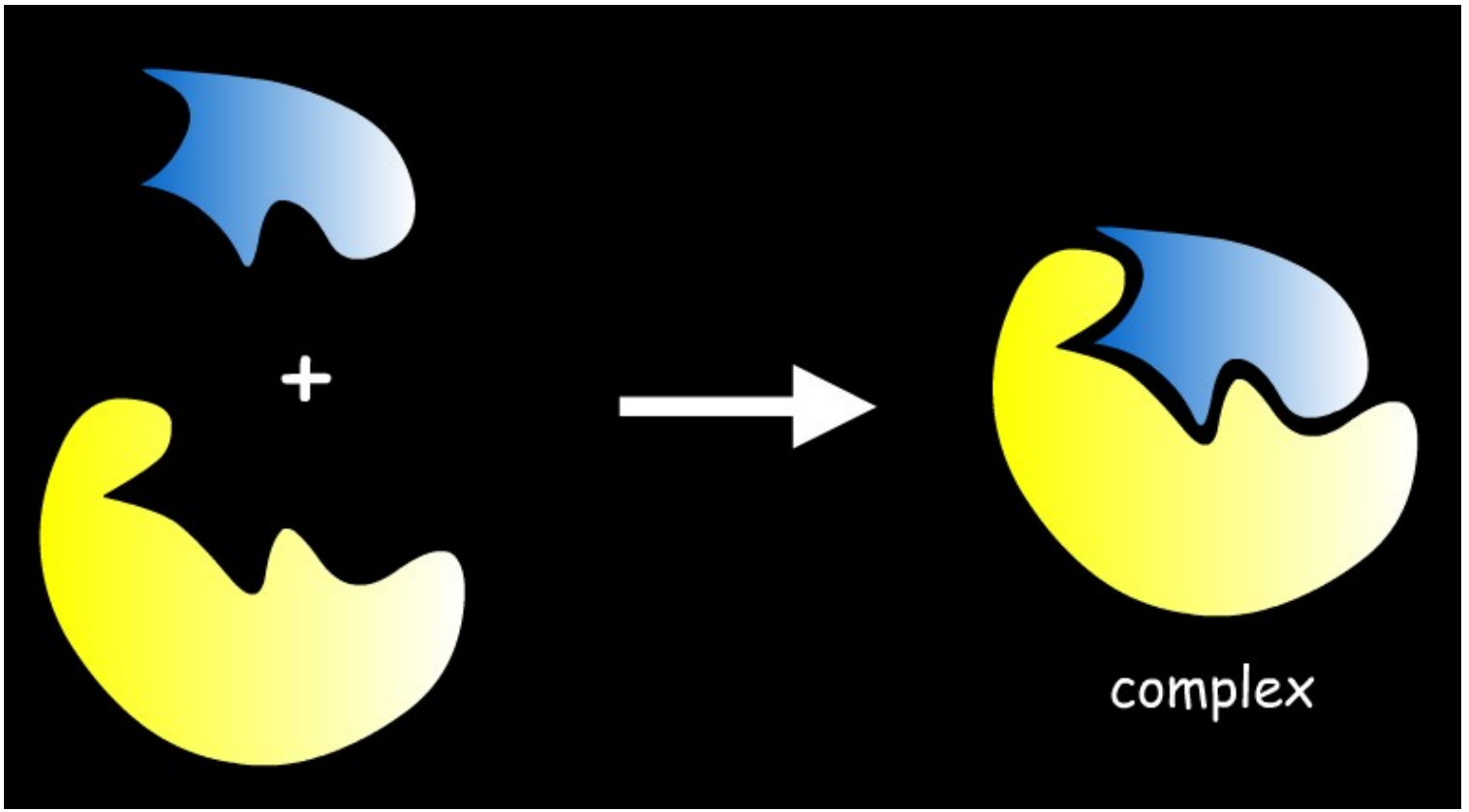
- Over the years biochemists have developed numerous models to capture the key elements of the molecular recognition process. Although very simplified, these models have proven highly useful to the scientific community.

year	model	author(s)
1890	lock-and-key	Emil Fischer
1958	induced-fit	Daniel Koshland
2003	conformation ensemble	Buyong Ma et al.

" All models are wrong, some are useful " (George Box)

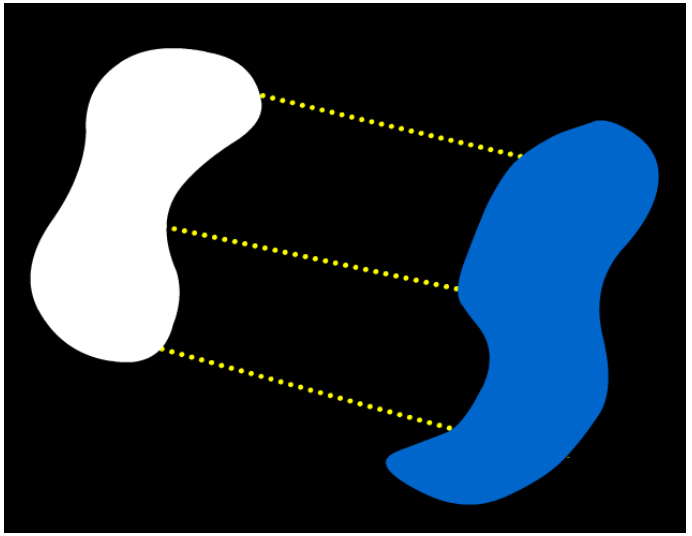
The Lock and Key Theory

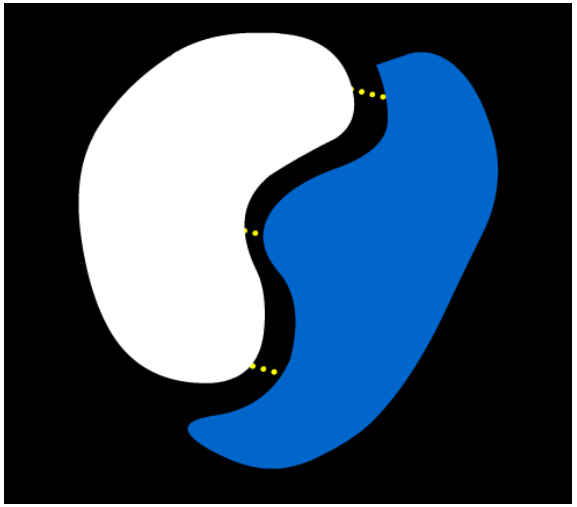
- As far back as 1890 Emil Fischer proposed a model called the "lock-and-key model" that explained how biological systems function. A substrate fits into the active site of a macromolecule, just like a key fits into a lock. Biological 'locks' have unique stereochemical features that are necessary to their function.



The Induced-Fit Theory

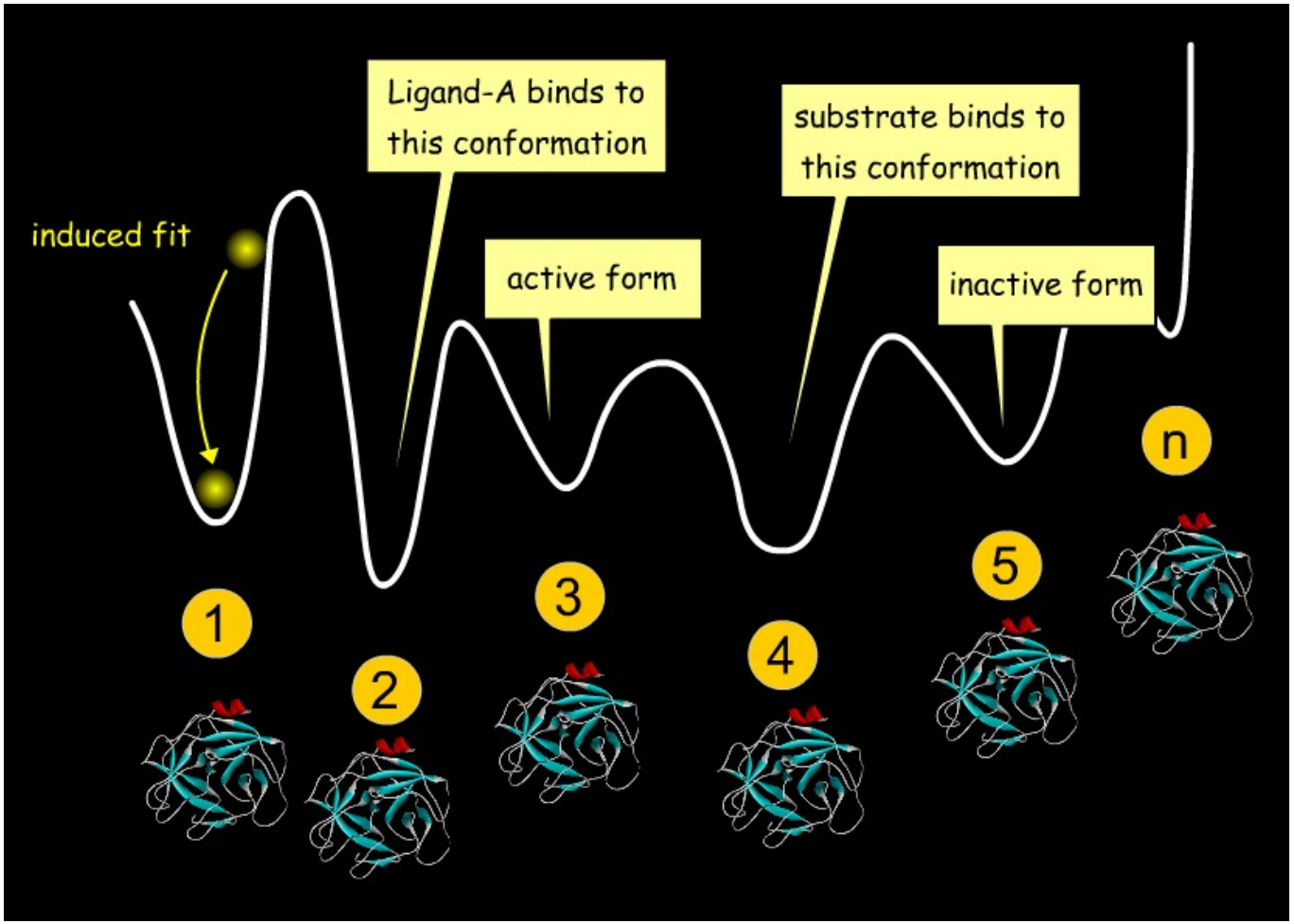
- In 1958 Daniel Koshland introduced the "induced-fit theory". The basic idea is that in the recognition process, both ligand and target mutually adapt to each other through small conformational changes, until an optimal fit is achieved.





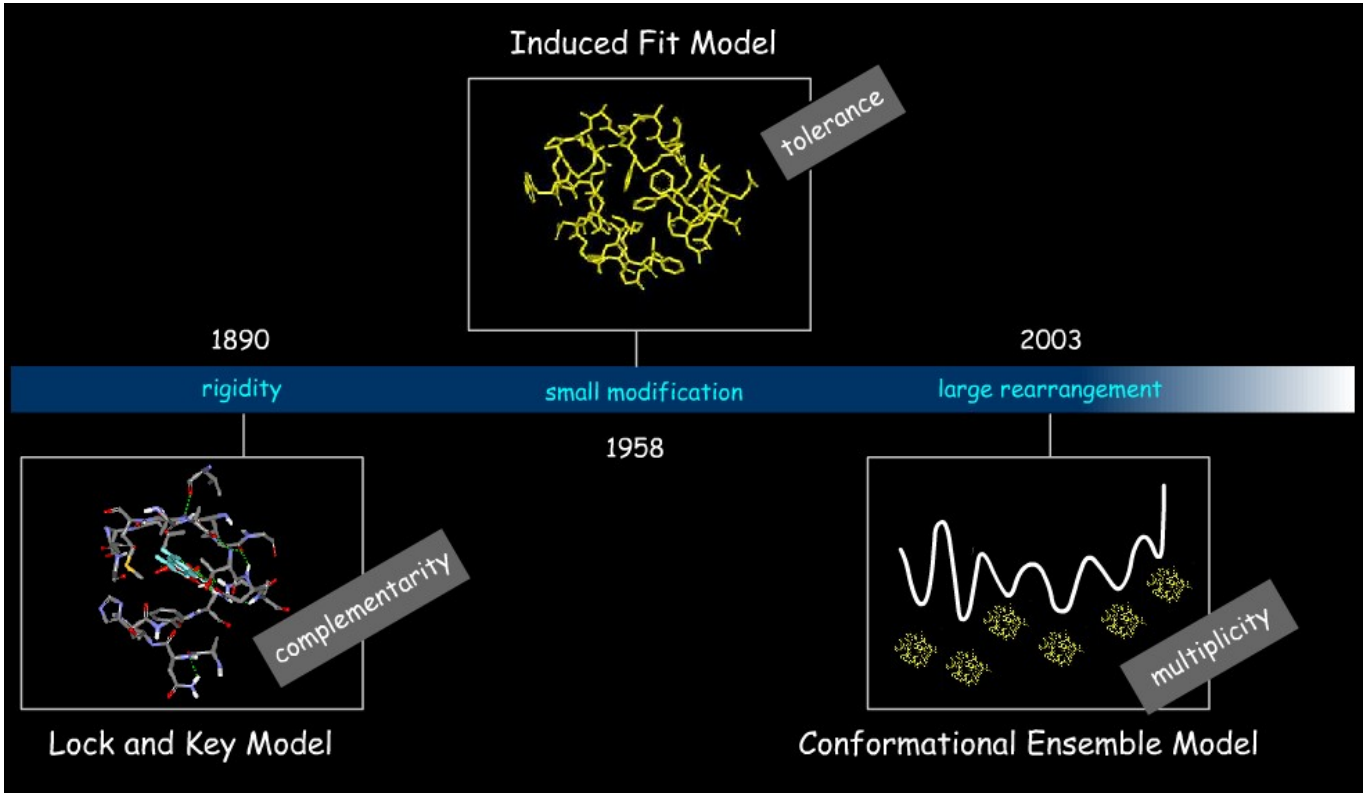
The Conformation Ensemble Model

- In addition to small induced-fit adaptation, it has been observed that proteins can undergo much larger conformational changes. A recent model describes proteins as a pre-existing ensemble of conformational states. The plasticity of the protein allows it to switch from one state to another.



From the Lock and Key to the Ensemble Model

- Lock-and-key, induced-fit and the conformation ensemble model are not contradictory. Each one focuses on a particular aspect of the recognition process. The lock-and-key model introduces the principle of 3D complementarity, the induced-fit model explains how complementarity is achieved, and the ensemble model depicts the conformational complexity of proteins.

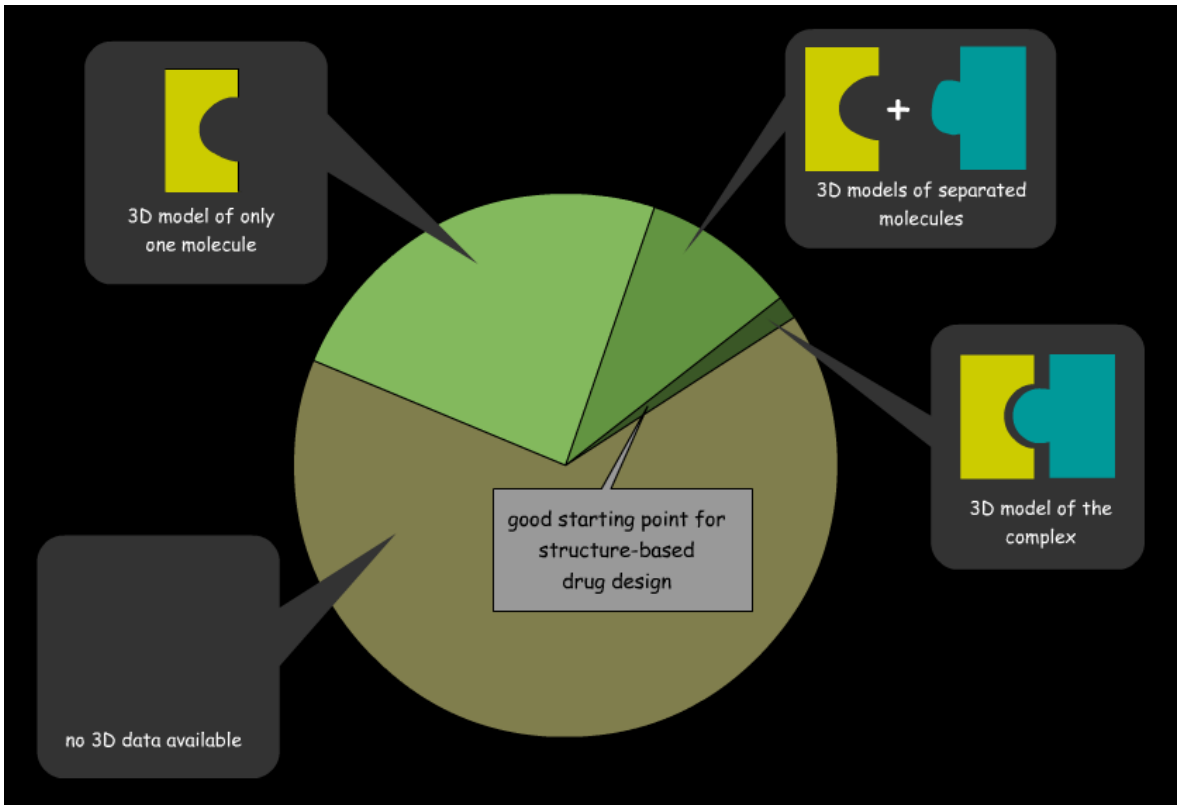


Experimental Methods to Study Molecular Docking

- Experimental techniques for study molecular recognition include X-ray crystallography, NMR, electron microscopy, site directed mutagenesis, co-immuno-precipitation etc...
- They allow us to experimentally solve the detailed 3-dimensional structures of biomolecules in their association form which is a necessary step in identifying crucial residues, study the strength of interaction forces, their energetics, understand how molecular structures fit together, and investigate mechanisms of action

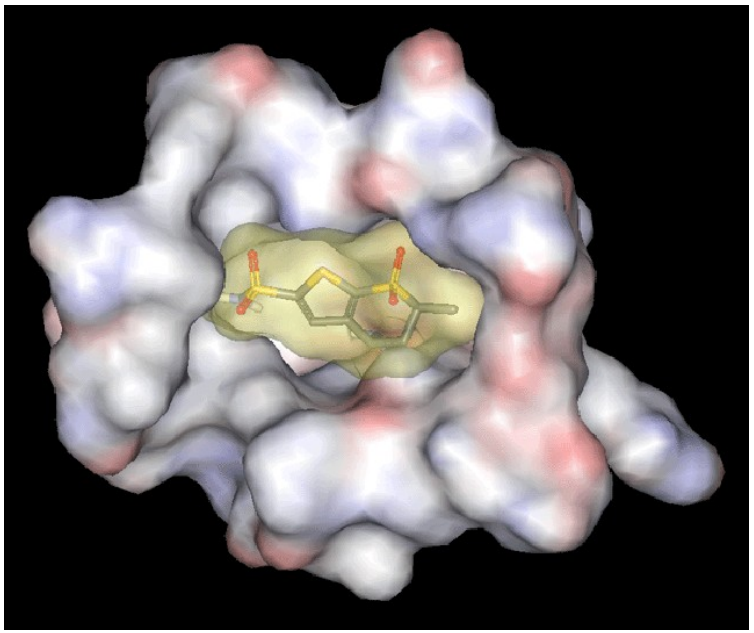
A Bottleneck in Drug Discovery

- Due to the limitations of current experimental methods, 3D structures of complexes are rarely available. But knowledge of the separated molecules in 3D is only weakly informative if we do not know how to assemble them.



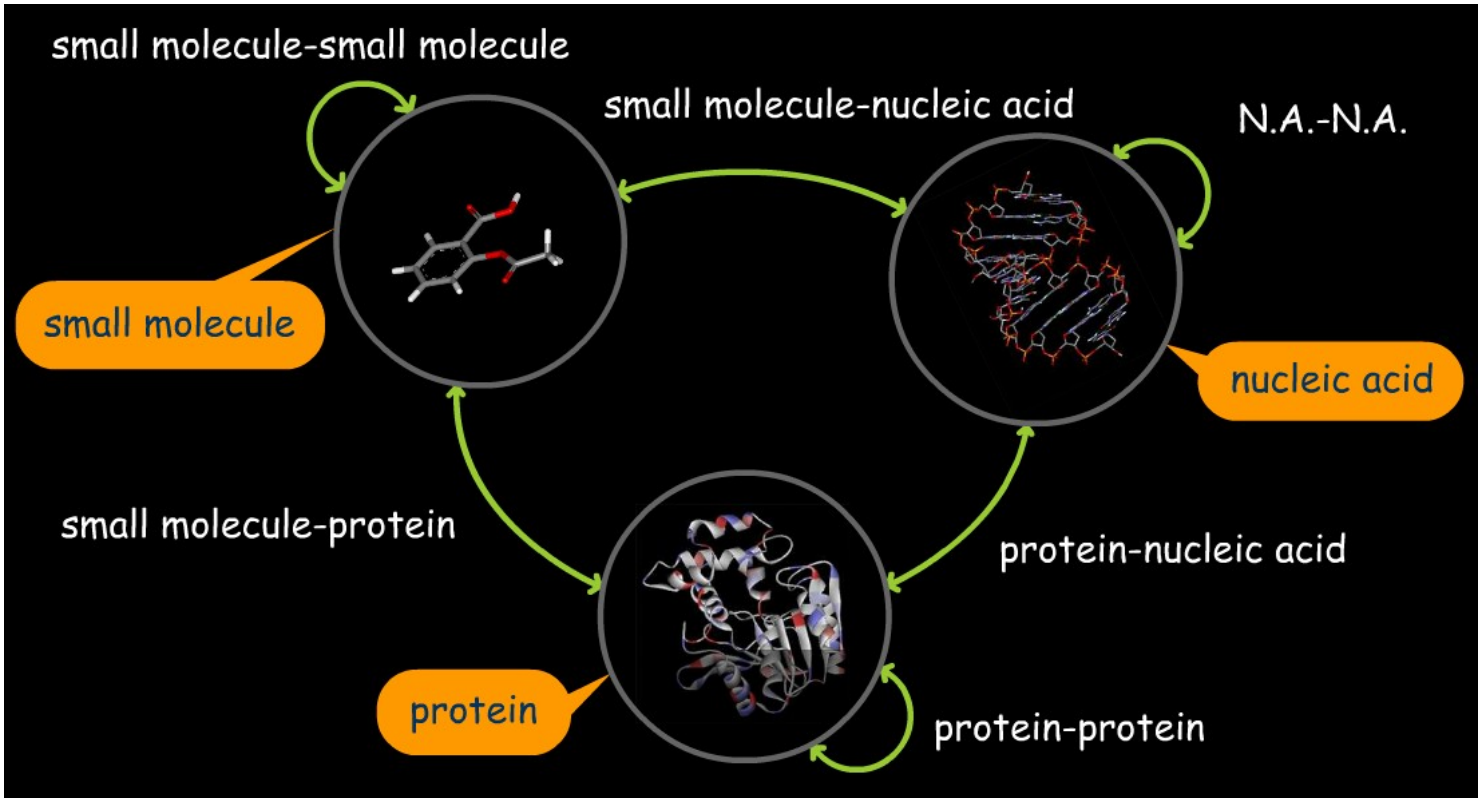
Triggering the Computational Docking Discipline

- The difficulties in obtaining experimentally structural data of macromolecular complexes have triggered the development of computational predictive methods
- Computational docking (also called *in silico* molecular docking or just docking) is a computational science aiming at predicting the optimal binding orientation and conformation of interacting molecules in space, and to estimate the stability of their complex
- Molecular docking predicts whether or not the two molecules interact, the binding affinity and the 3D structure of the complex
- Computational docking is an essential component in modern drug discovery. Over the last few decades, it has been routinely and successfully applied in most pharmaceutical and biotech companies for a large number of applications.



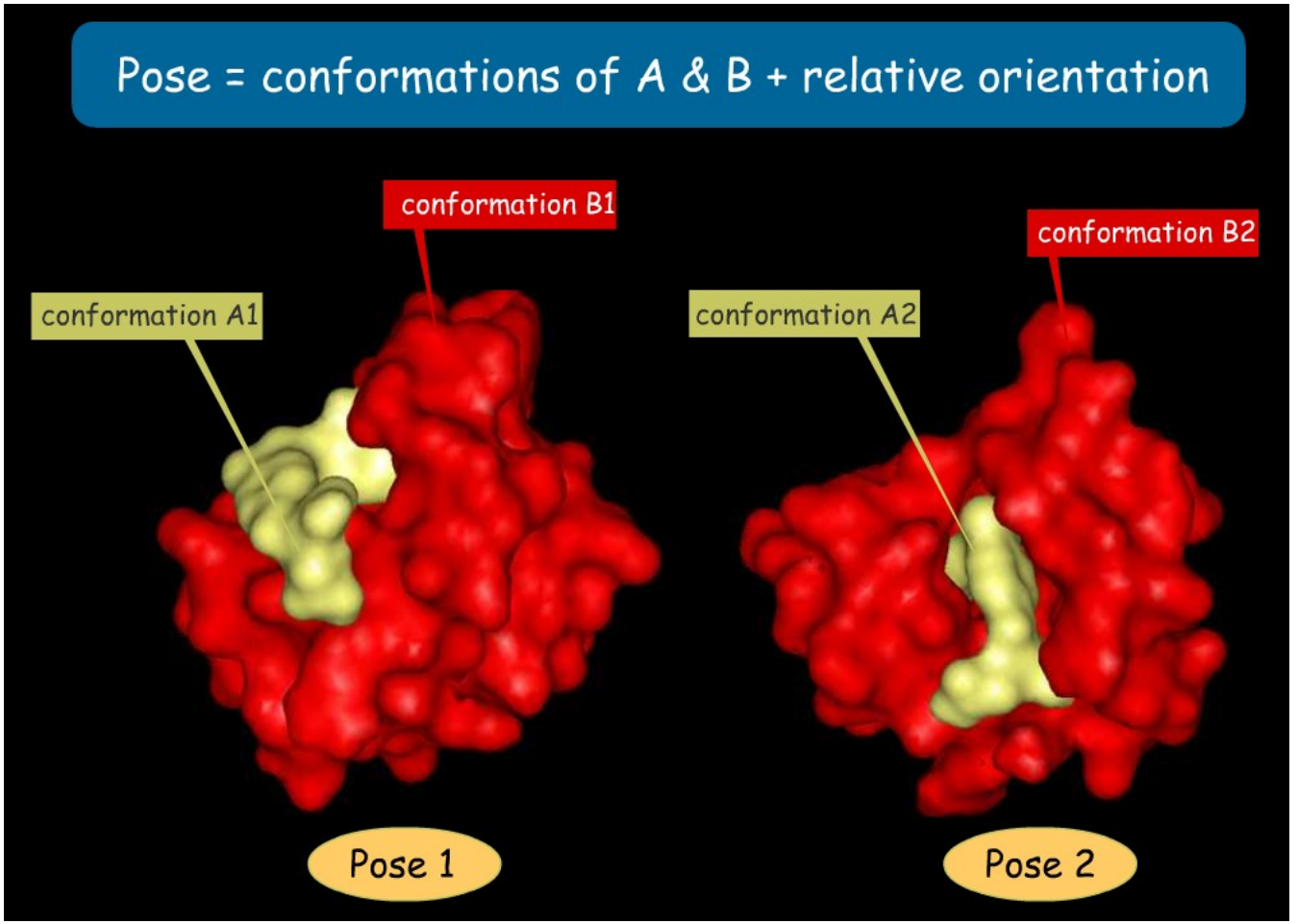
Docking Classification

- Molecular docking classifies biomolecules into three categories: small molecules (also called 'ligands'), proteins, and nucleic acids
- The most important types of docking systems are: protein-ligand, protein-protein and nucleic acid-protein
- The interactions between a small molecule and a protein are by far much better understood than those between a protein and a nucleic acid



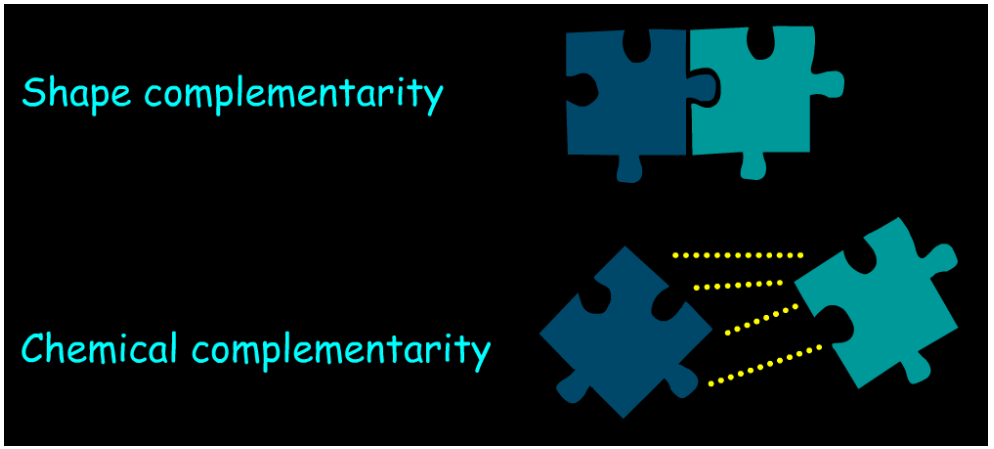
Definition of the "Pose"

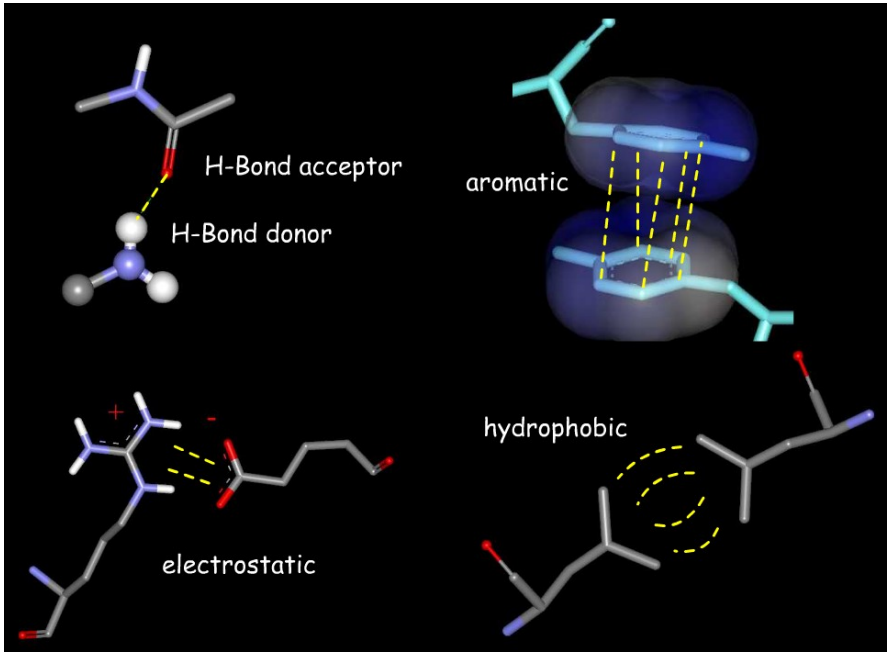
- A "pose" is a term widely adopted for describing the geometry of a particular complex (also called "binding mode")
- It refers to a precise configuration which is characterized not only by the relative orientation of the docked molecules but also their respective conformations



Molecular Complementarity in Computational Docking

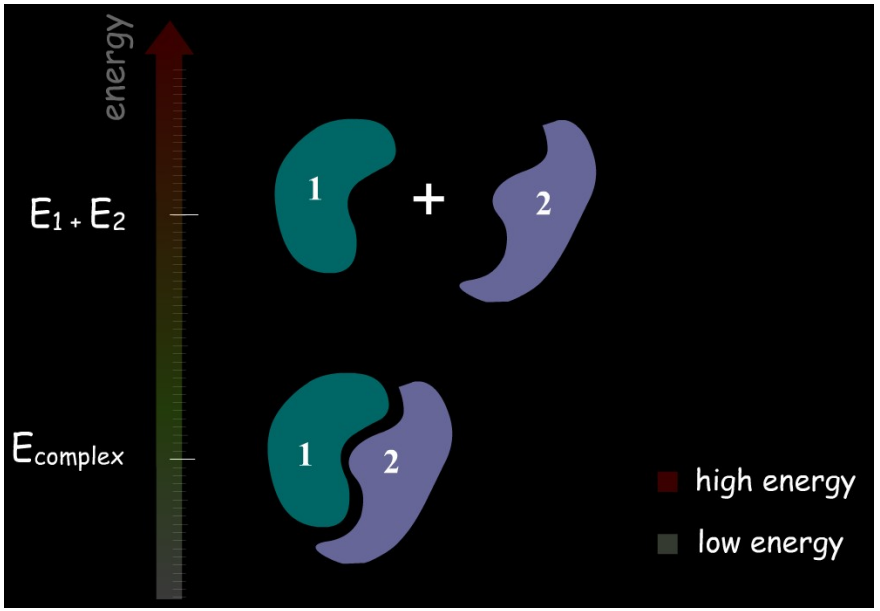
- Computational docking exploit the concept of molecular complementarity. The structures interact like a hand in a glove, where both the shape and the physico-chemical properties of the structures contribute to the fit.
- Shape complementarity is the primary criterion for evaluating the fit in the computational docking of two candidate structures
- In addition to shape compatibility, chemical and physico-chemical complementarity are also important criteria in the docking between candidate structures

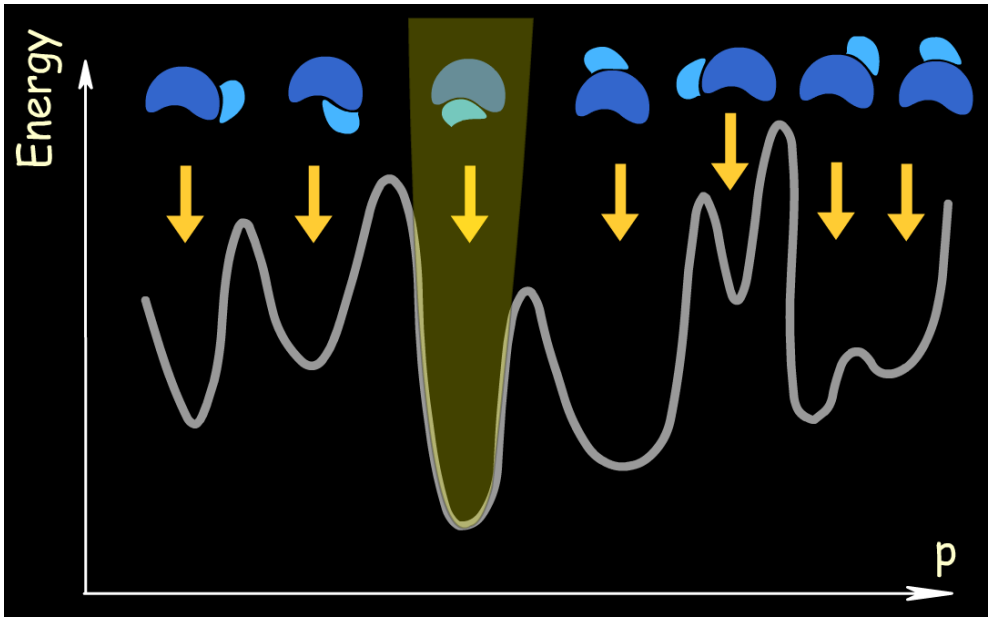




Energy Dictates Molecular Associations

- The process of "self-assembly" is dictated by forces that are energy based: the complex has a lower potential energy than its constituent parts, and this keeps the parts together
- The goal of computational docking is to find the 3D configuration of the complex that minimizes the energy

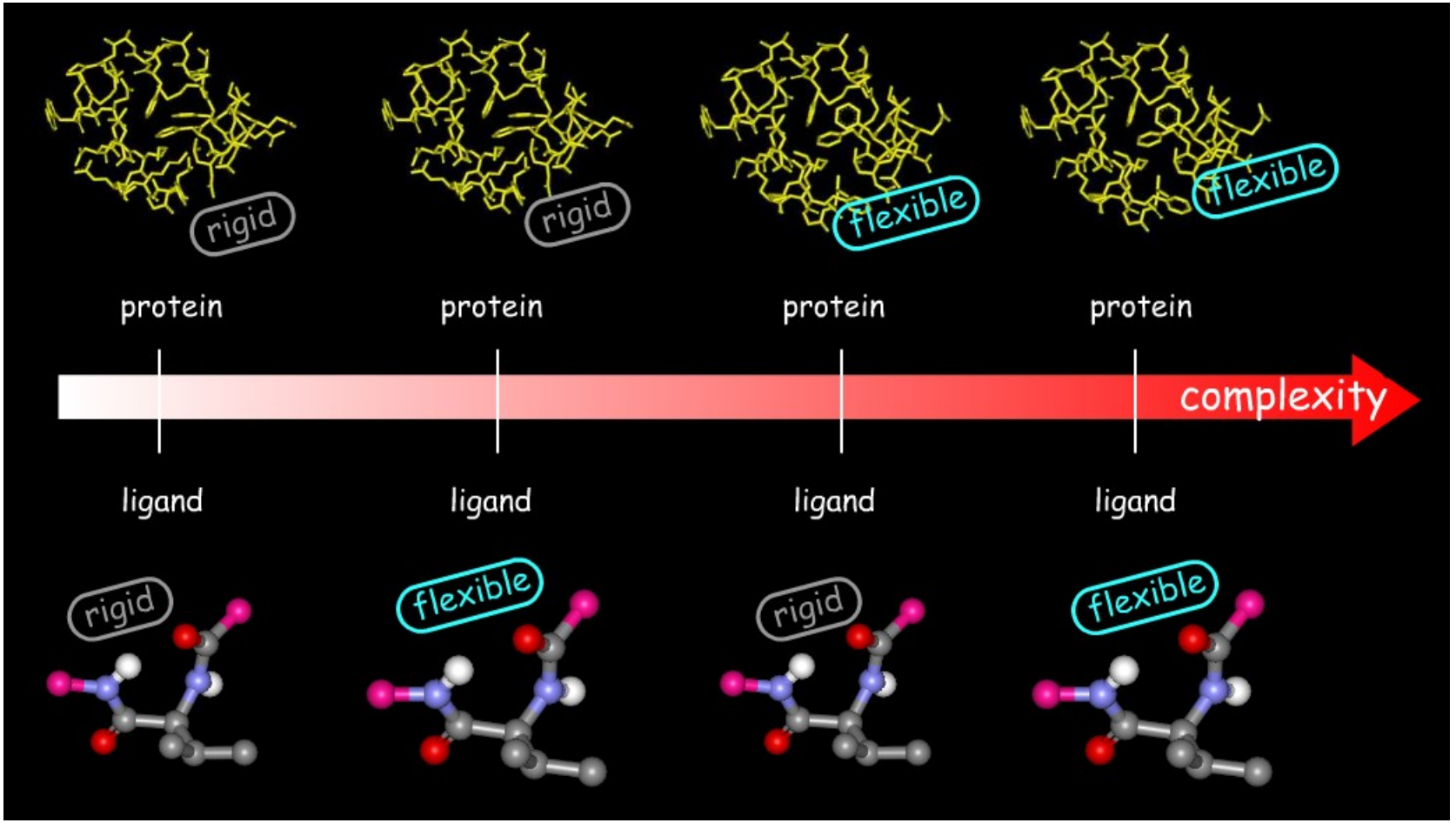




Molecular Flexibility in Protein-Ligand Docking

- The mutual adaptation of a ligand with its receptor is crucial to understanding ligand binding and protein function
- One of the major challenges in molecular docking is how to account for this adaptation in docking calculations
- The docking problem can be classified according to the way flexibility is modeled. In ascending order of complexity:
  1. Rigid body docking ignores the flexibility of the molecules and treats them like rigid objects
  2. Rigid receptor – flexible ligand docking: only the ligand is treated as flexible, receptor is rigid
  3. Flexible receptor – flexible ligand docking: both protein and ligand are treated as flexible.

Molecular Flexibility in Protein-Ligand Docking



Three Components of Docking Software

- Docking software can be categorised based on the following criteria:
  1. Molecular representation - a way to represent structures and properties (atomic, surface, grid representation)
  2. Scoring method - a method to assess the quality of docked complexes (force field, knowledge-based approach, ...)
  3. Searching algorithm - an efficient search algorithm that decides which poses to generate (exhaustive search, Monte Carlo, genetic algorithms, simulated annealing, tabu search)

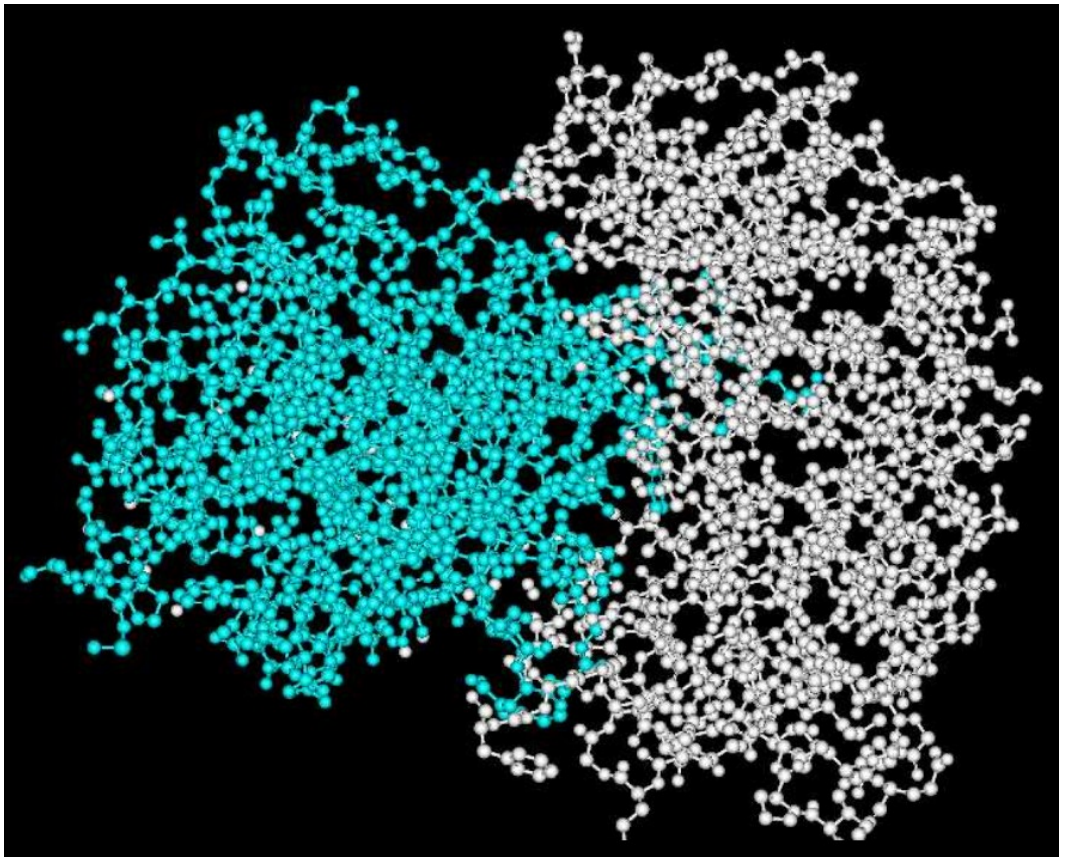
Molecular Representation

Molecular Representation

- There are three representations commonly used in docking programs:
  - the atomic representation (the most common)
  - the surface representation
  - the grid representation
- The choice of representation dictates the way the docking problem will be tackled

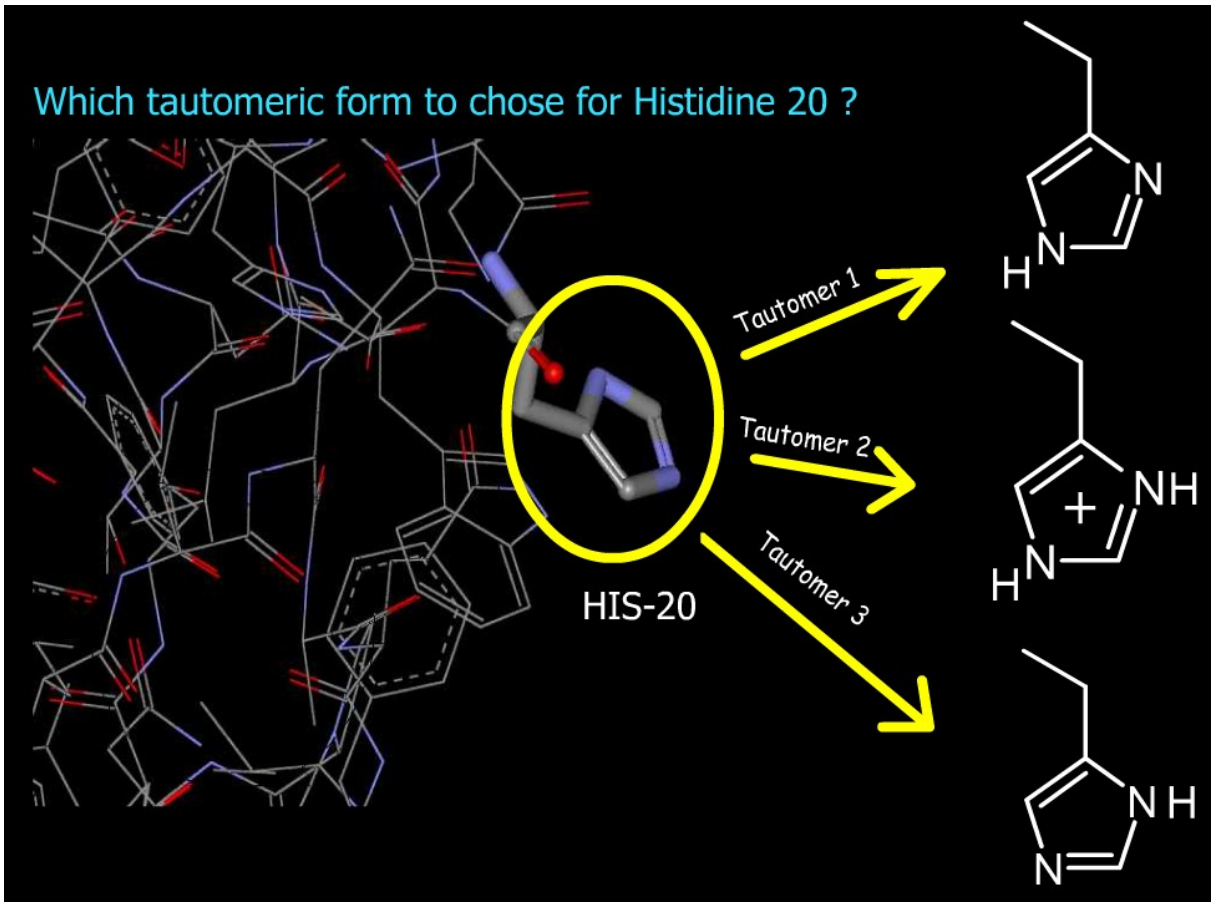
Atomic Representation

- In atomic representation, each atom is characterized by its coordinates and atom-type



Protein Preparation

- The preparation of the protein calls for great care. Important decisions include the choice of the tautomeric forms of histidine residues, the protonation states of amino-acids and conformations of some residues; their incorrect assignments may lead to docking errors.

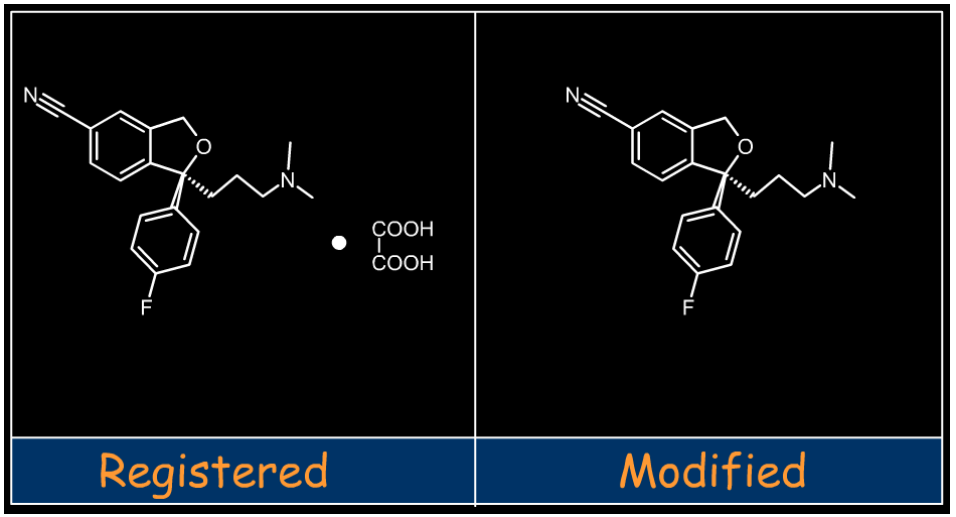


Small Molecule Preparation

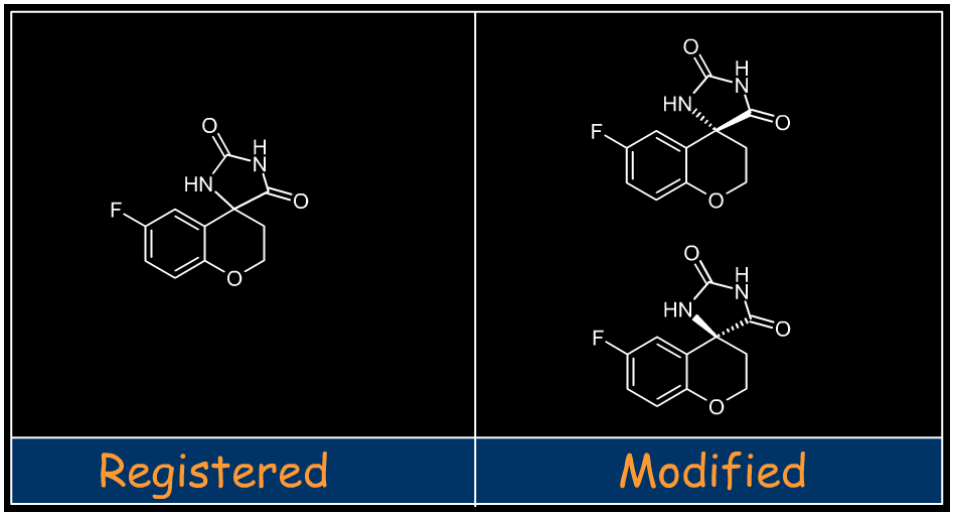
- Before generating and docking the 3D structures of a library of ligands, it is important to "clean up" the 2D structures being used by removing any counter ions, salts, or water molecules that might be part of the registered structure
- All reactive or otherwise undesirable compounds must also be removed
- Possibly generate all optical isomers (enantiomers), cis/trans isomers, tautomers, and protonation states of the structures
- For most docking programs the tautomeric and protonation state of the ligands to be docked is defined by the user; in general the structure considered to be dominant at a neutral pH is generated; here also, incorrect assignments may lead to docking errors

Small Molecule Preparation

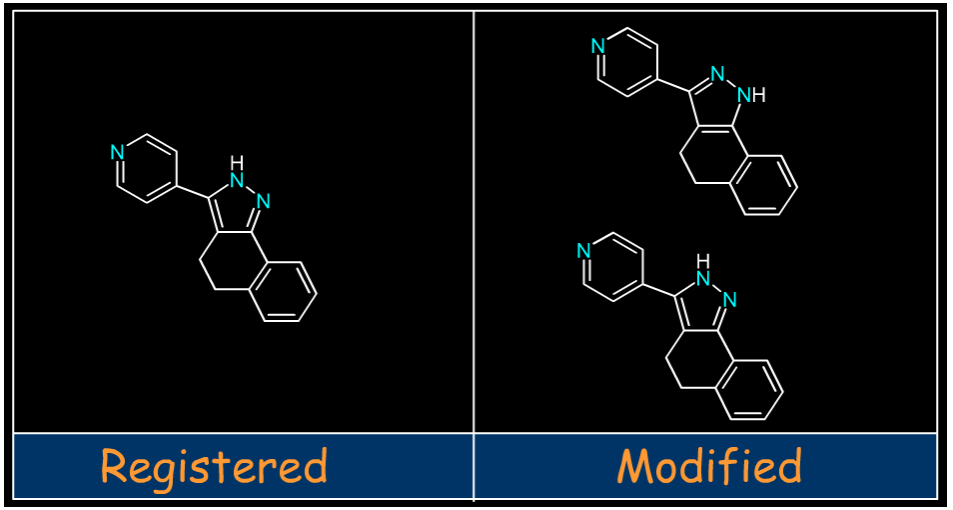
Salt removing



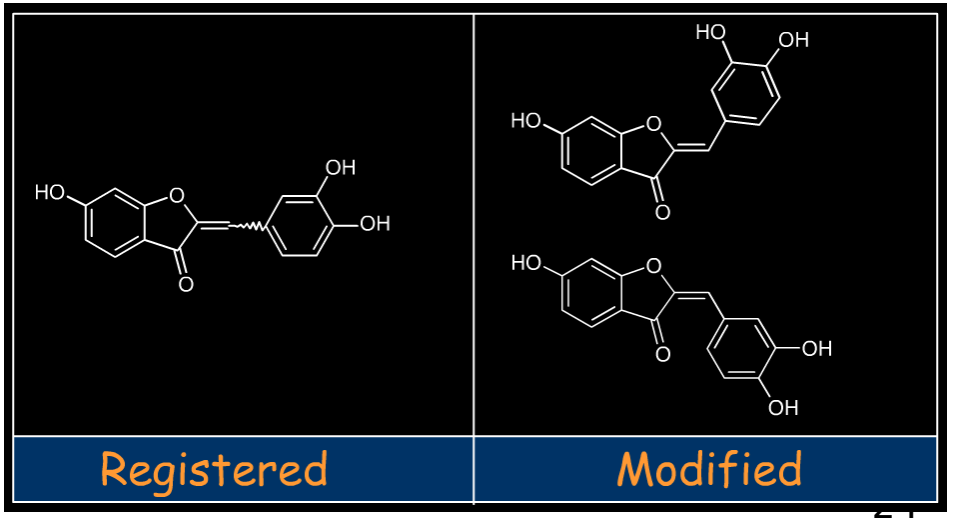
Optical isomers (enantiomers) generation



Tautomers generation



Double bond cis/trans isomers generation



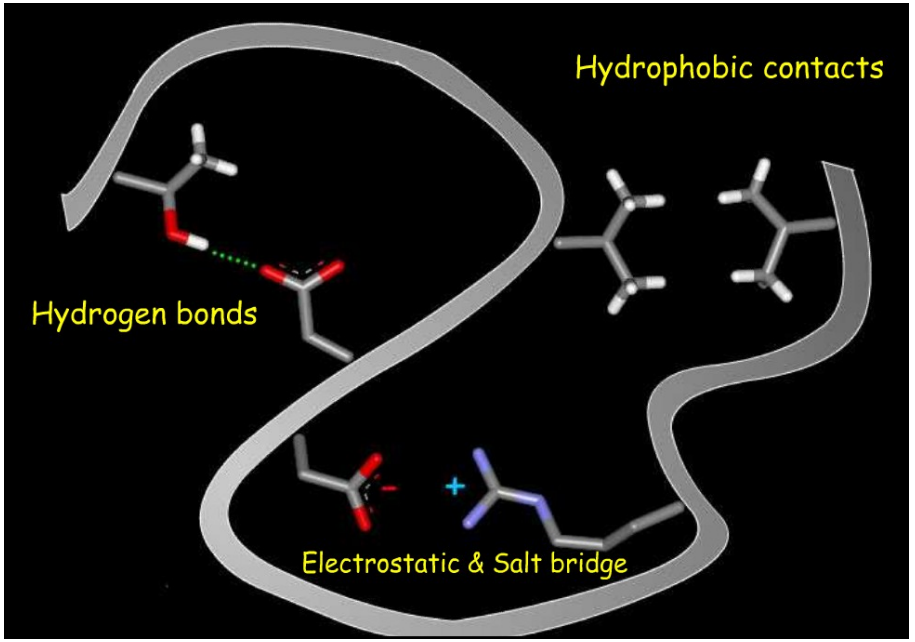
Scoring Methods

Scoring Methods

- Scoring methods aim at assessing the quality of docked complexes and guiding the docking algorithm
- The binding process that leads to the formation of a complex between a ligand and its receptor is controlled by several factors including:
  1. the interaction energies between the two molecules
  2. the desolvation and solvation energies associated with the interacting molecules
  3. the entropic factors that occur upon binding
- The final free energy of binding will depend on the overall balance of these factors

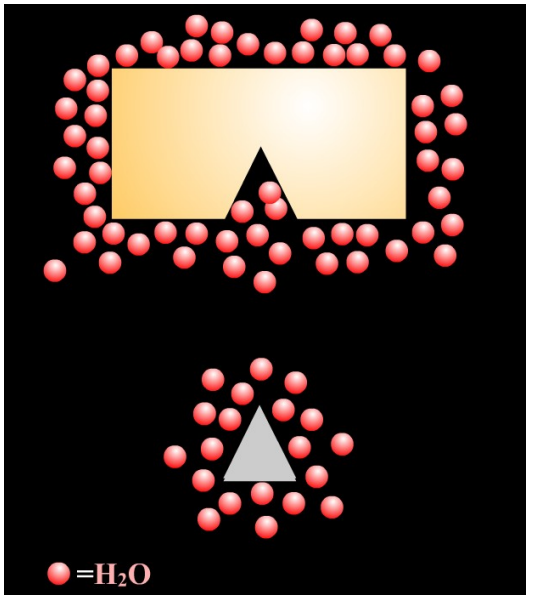
Interaction Energies

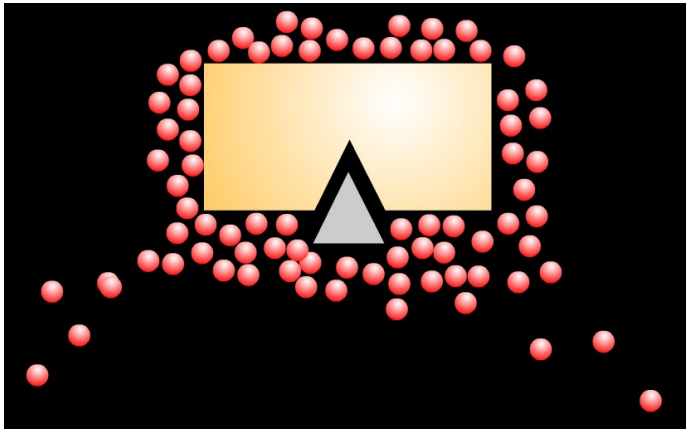
- The interaction forces between two molecules can be divided into:
  - Electrostatic interactions
  - Hydrogen bond interactions
  - Van der Waals interactions
  - Hydrophobic forces



Desolvation Energies

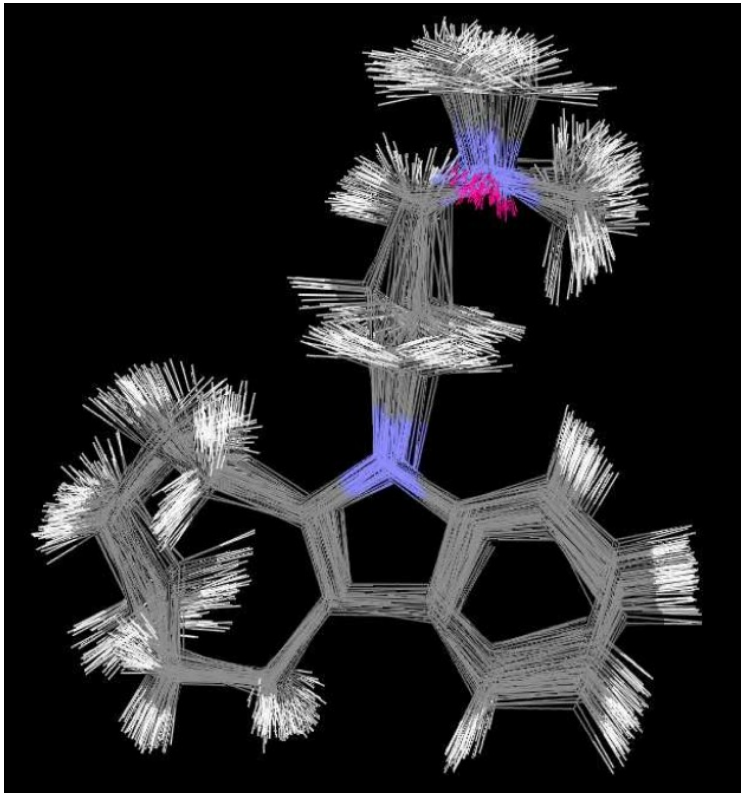
- The binding of a ligand to a protein is a complex process influenced by desolvation and solvation phenomena where the interacting entities become partially desolvated
- This thermodynamically driven chain of events leads to the formation of favorable interactions between the ligand and the protein where hydrophobic contacts are the driving forces: hydrophobic moieties associate together to reduce the interactions with the surrounding water
- Another important energy term is electrostatic interaction between charged atoms and water molecules





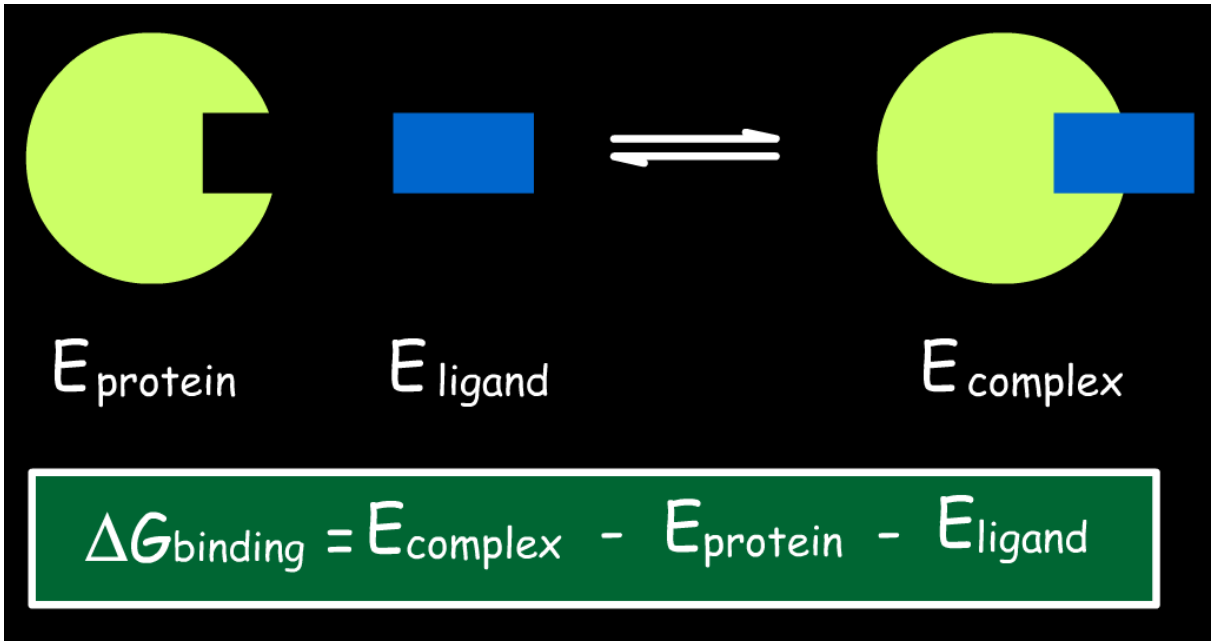
Entropic Effects

- The flexibility of the molecules and the consequences in terms of entropy can have a significant impact on the binding energy of a ligand



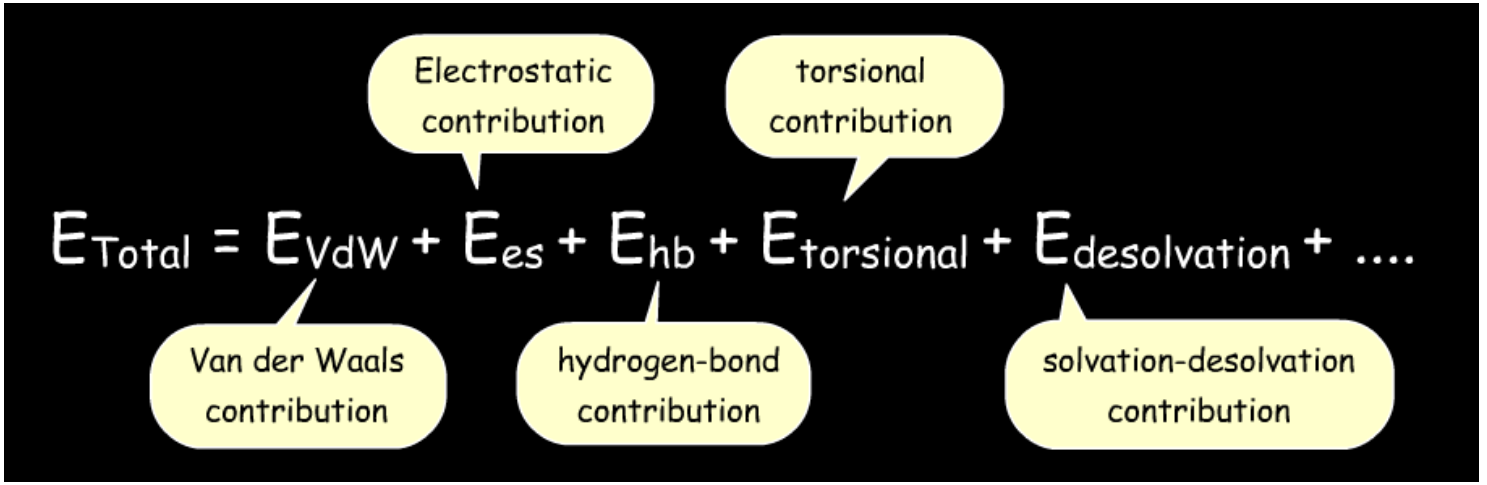
Calculation of the Binding Energies

- The binding energy $\Delta G_{binding}$ is the energy required to separate a complex into separate parts (protein and ligand). It is defined as the difference between the energy of the associated (bound) form ($E_{complex}$) and that of the separated (unbound) molecules ($E_{protein}$ and E_{ligand}).
- A complex has a lower potential energy than its constituent parts. This is what keeps them together.



Force-Field Calculations

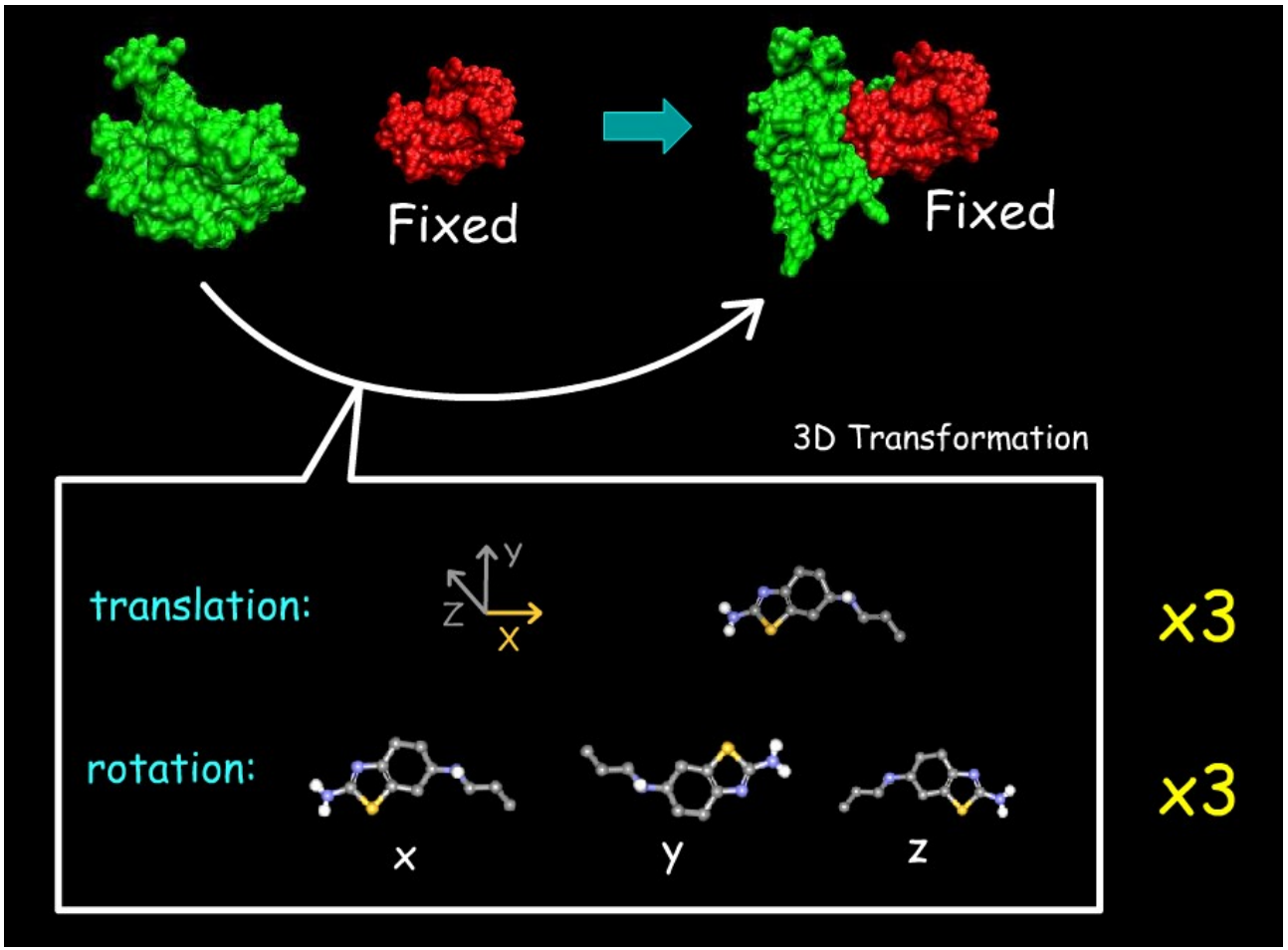
- Molecular mechanics can be used to estimate the internal energy of the system, which makes it useful for calculating ΔG
- The total energy of a system is described as the sum of the independent terms of the force field
- The energies obtained by force field methods can be used directly to approximate free energies of binding



Searching Algorithms

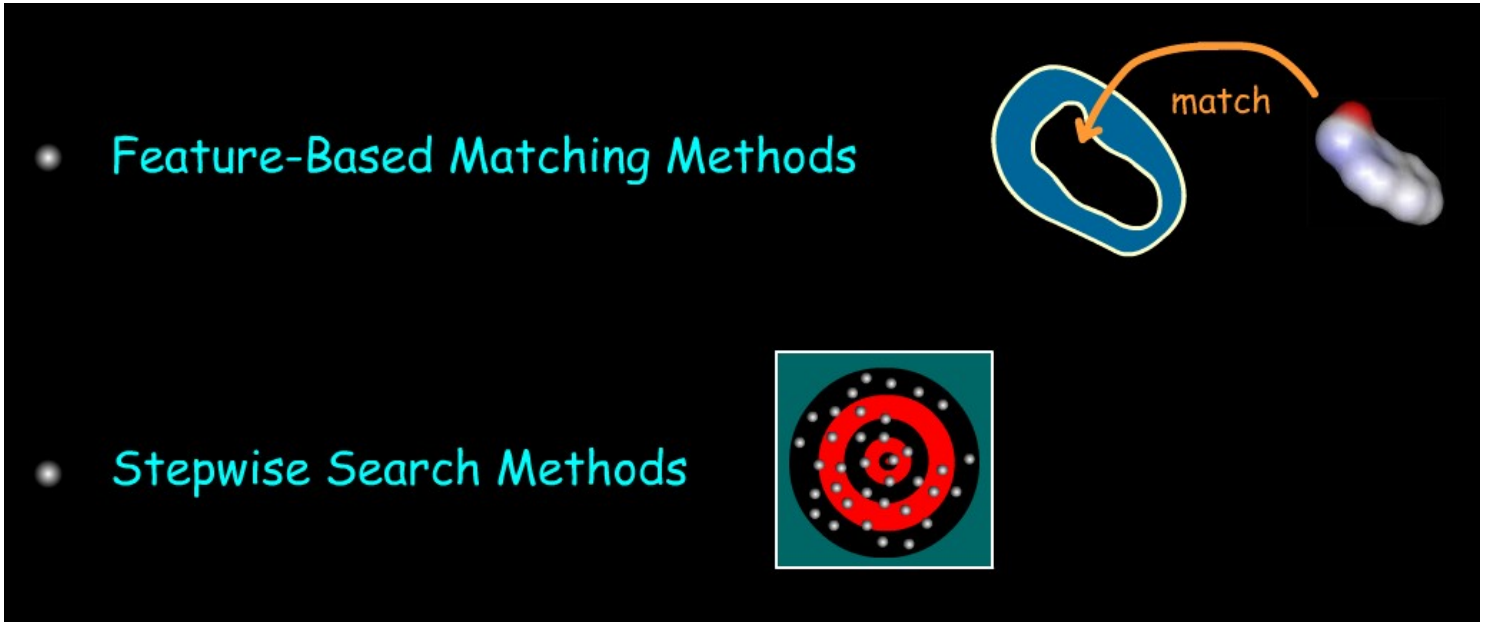
Rigid Docking Methods

- If we assume that the molecules are rigid, we are then looking for a transformation in 3D space of one of the molecules which brings it into optimal fit with the other molecule in terms of a scoring function
- In rigid-body docking, the search space is restricted to three rotational and three translational degrees of freedom



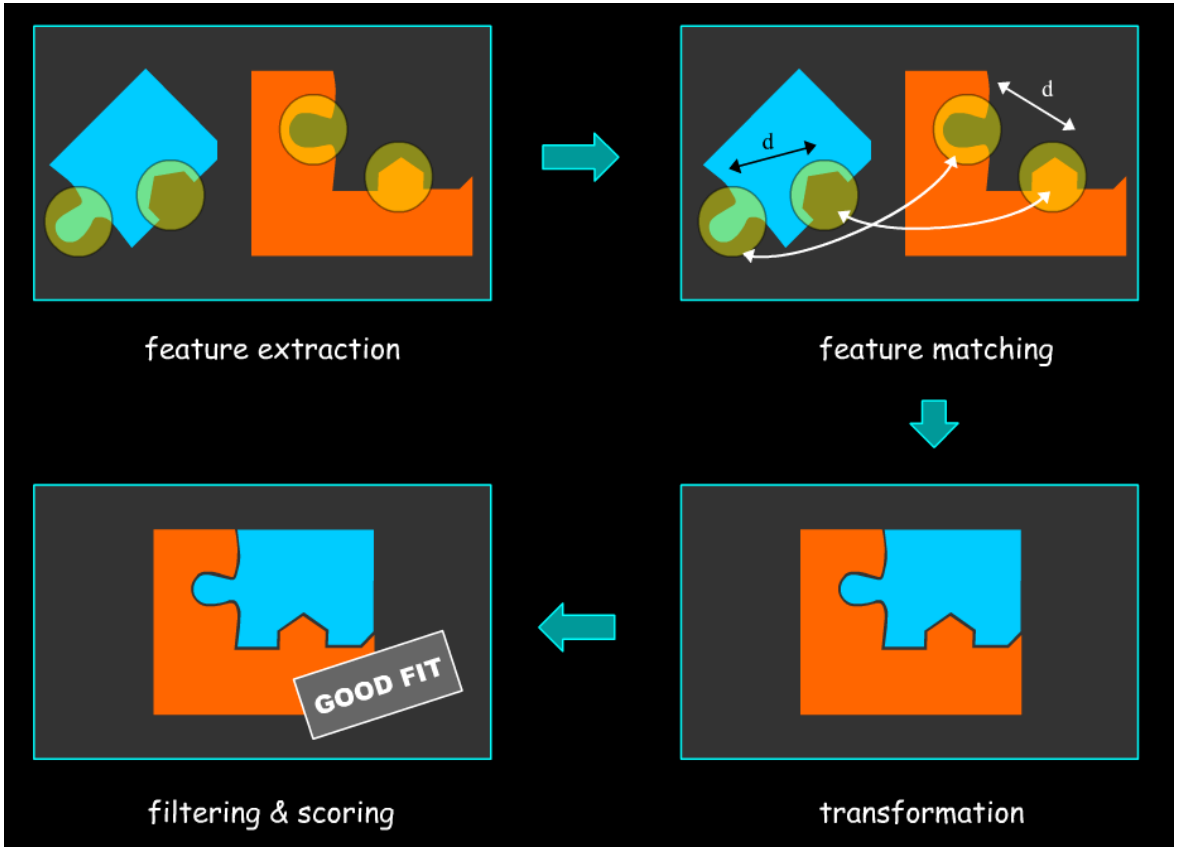
Two Docking Philosophies

- A large number of docking approaches have been developed to predict the formation of molecular complexes. They can be divided into two broad classes that correspond to two different philosophies:
  - Feature-based matching matches local complementarity features among molecules involved in the recognition
  - Stepwise search explores the 'search space' guided by a scoring function



Components of Feature-Based Matching Methods

- The way feature-based matching applies to the docking problem is similar to solving a jigsaw puzzle. You pick a piece and look for a complementary one from among the rest of the pieces (feature extraction). Once a piece is found (feature matching), the elements are assembled (transformation), the solution is then assessed globally for final approval of the compatibility (filtering and scoring).



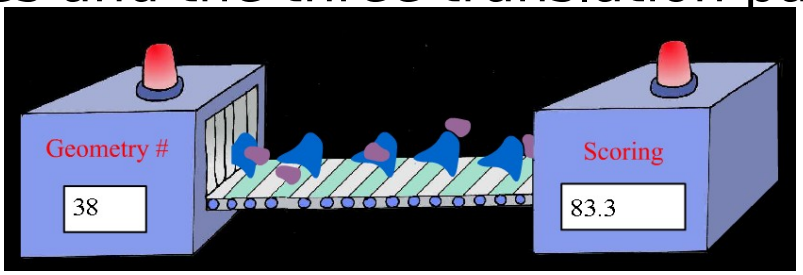
The Stepwise Search Approach

- The stepwise search approach tries to explore the search space (defined as the set of all possible solutions), with the hope of finding an optimal solution
- The approach is driven by a scoring function which guides the search algorithm
- In computational docking the stepwise search involves two components:

    1. A positioning module which generates new complex arrangements

    2. A scoring module that assesses the quality of each individual arrangement

- The positioning module is directly connected to the search module, which dictates the configuration of the next pose to be generated (by appropriate search algorithm)
- In rigid-body docking the variables to optimize are the three rotation angles and the three translation parameters



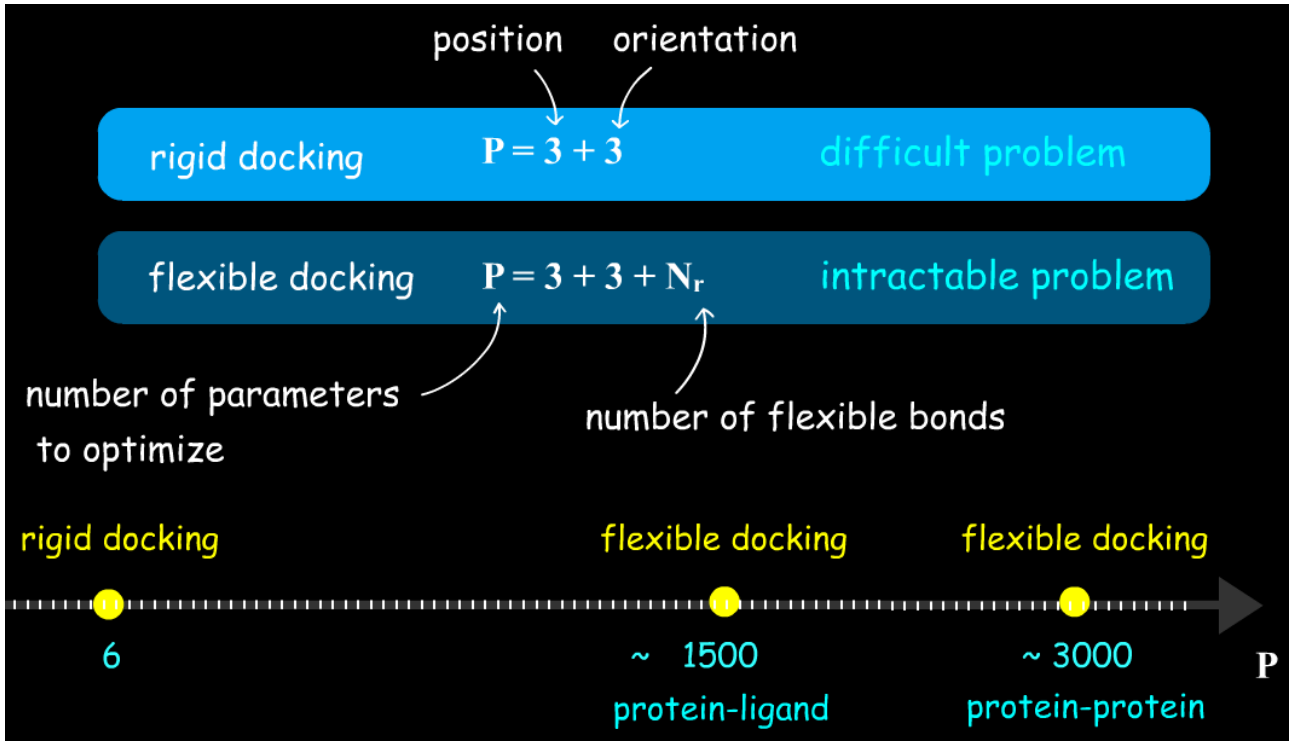
Search Algorithms

- The following approaches are used in computational docking:
  - Exhaustive search (for small systems only)
  - Monte Carlo
  - Genetic Algorithms
  - Simulated Annealing
  - Tabu searches
- Tabu search (TS) is a stochastic searching algorithm that maintains a list of previously visited poses. These poses are forbidden and cannot be revisited (they are "taboo"). TS effectively guides the search process into unvisited areas of the space.

Methods for Incorporating Flexibility

Degrees of Freedom in Flexible Docking

- The rigid-body docking approaches are often not sufficient to predict the structure of a protein complex from the separate unbound structures
- The incorporation of molecular flexibility into docking algorithms requires to add conformational degrees of freedom to translations and rotations
- Approximation algorithms need to be introduced to reduce the dimensionality of the problem and produce acceptable results within a reasonable computing time

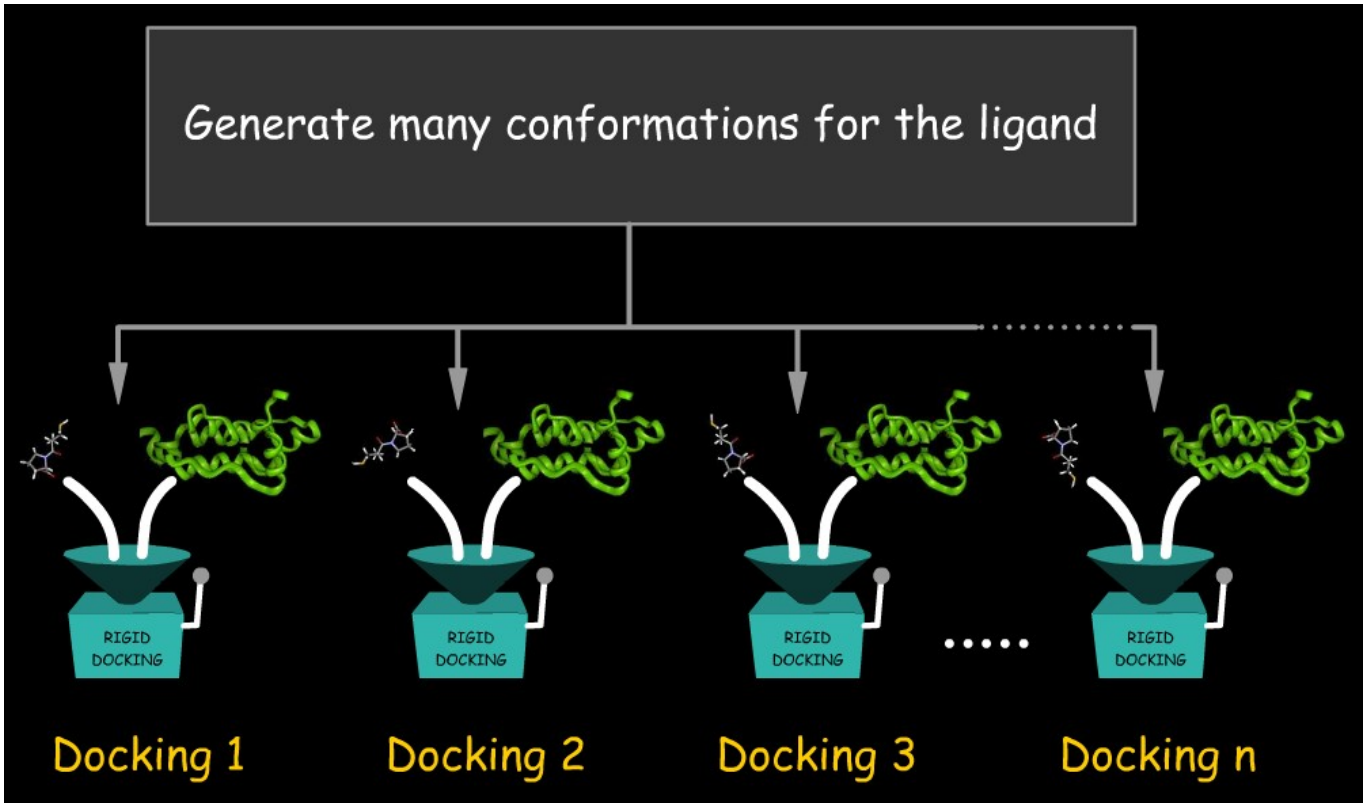


Methods for Handling Ligand Flexibility

- Many methods have been developed for incorporating flexible small molecules into docking software; they include:
  1. Ligand-ensemble docking method
  2. Fragmentation method
  3. Stochastic conformational search method

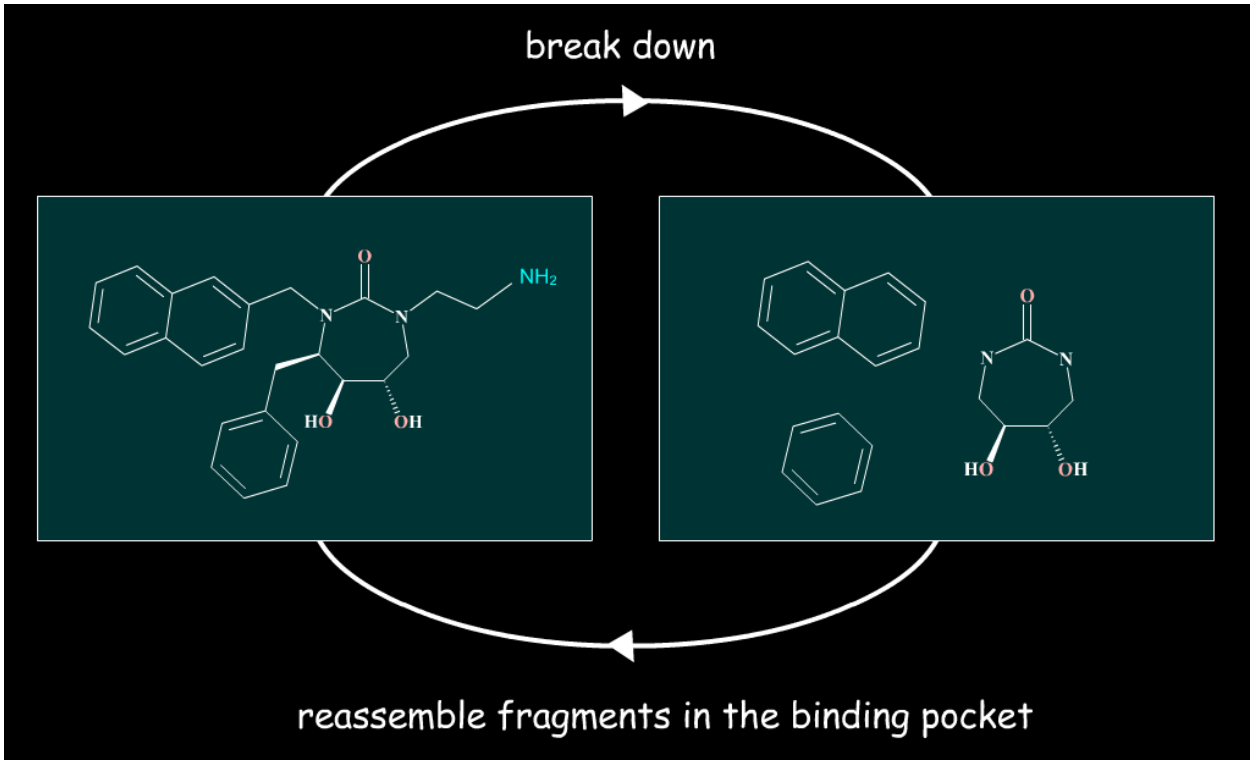
Ligand-Ensemble Docking Method

- The simplest method to account for small molecule flexibility is to consider it as an "ensemble" of rigid and independent ligand conformations
- In the first step low energy conformers of ligand are generated by conformational analysis
- In the second step, rigid docking is applied for each conformer independently in order to find the most favorable small molecule-protein complex



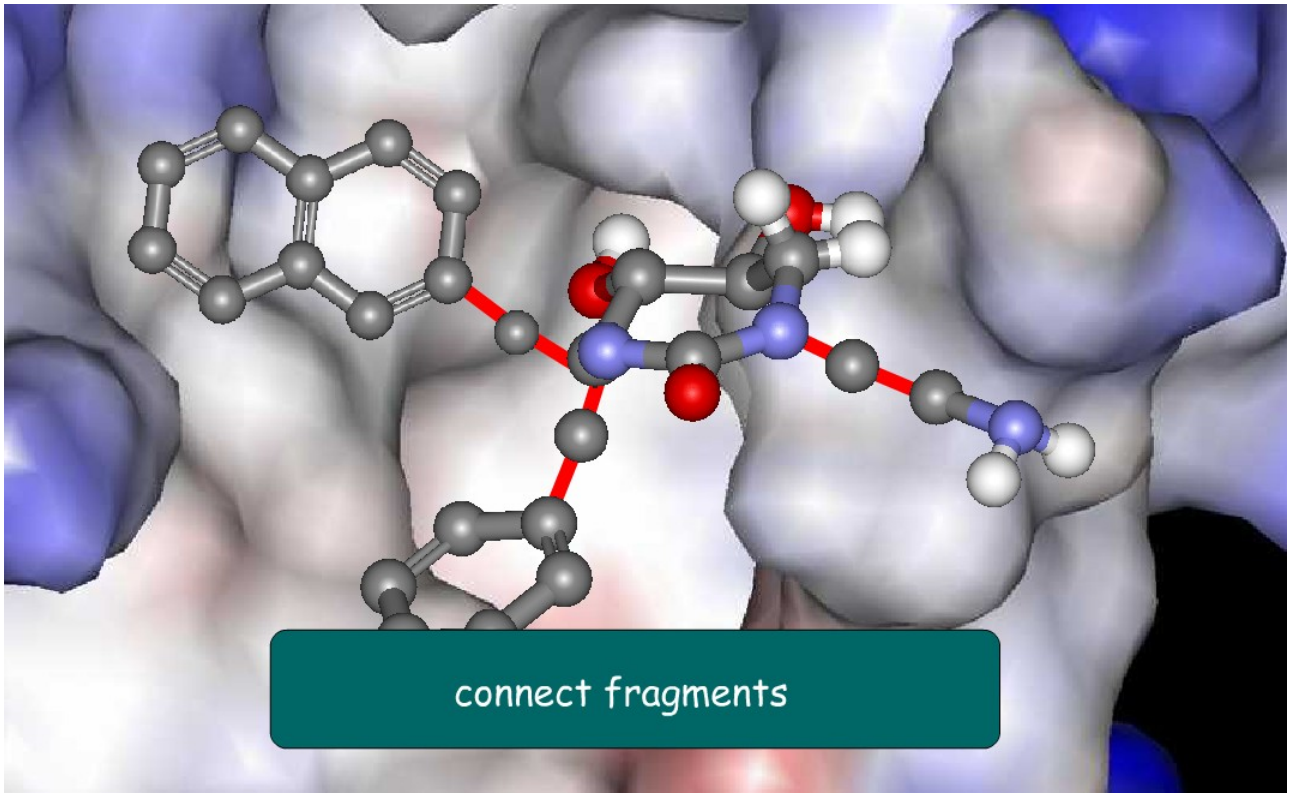
The Fragmentation Docking Method

- Fragmentation methods break down the molecule into small rigid fragments, the fragments are then reassembled in the binding pocket
- Two different approaches are used for reassembling the disconnected moieties: the place-and-join and the incremental approach



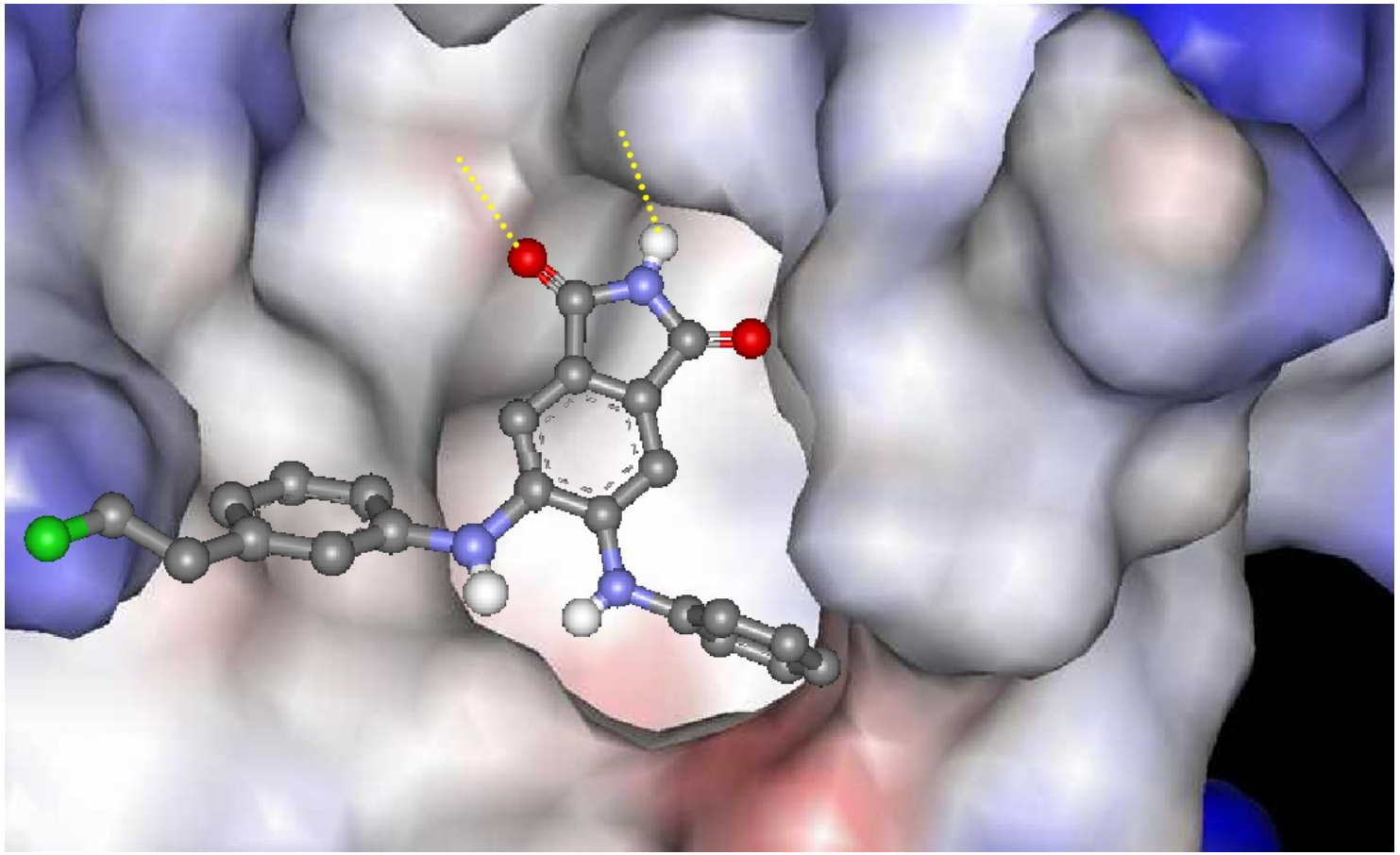
Place-and-Join Algorithm

- The place-and-join method splits the molecule into rigid subparts
- Each subpart is docked independently
- Assembly of the fragments is then done by looking at their relative location and assessing the possibility of re-connecting them with the connectivity of the initial molecule, in geometrically correct conformations



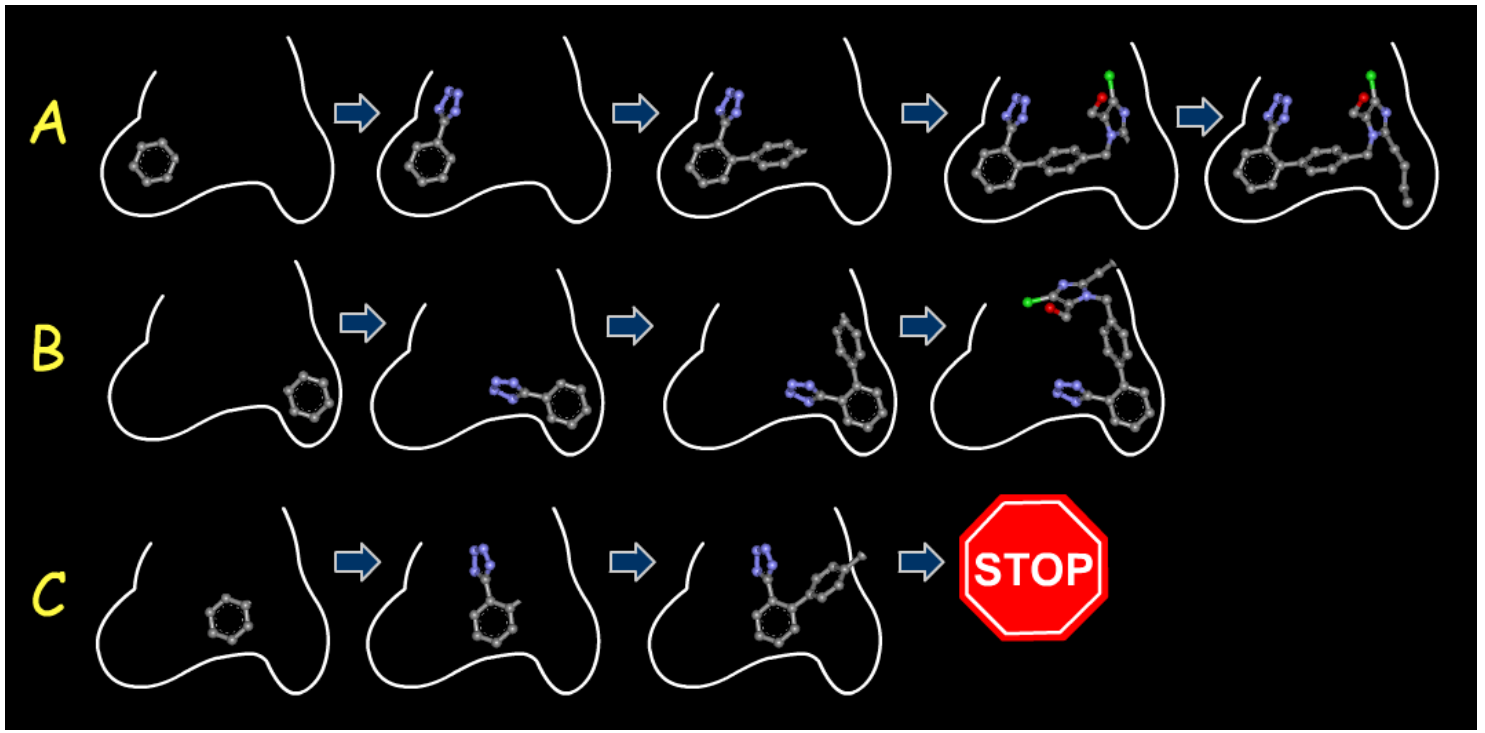
Incremental-Based Methods

- The incremental based approach starts with an initial core docked in the active site, and new fragments are progressively added and minimized; the treatment is terminated when the entire molecule is formed



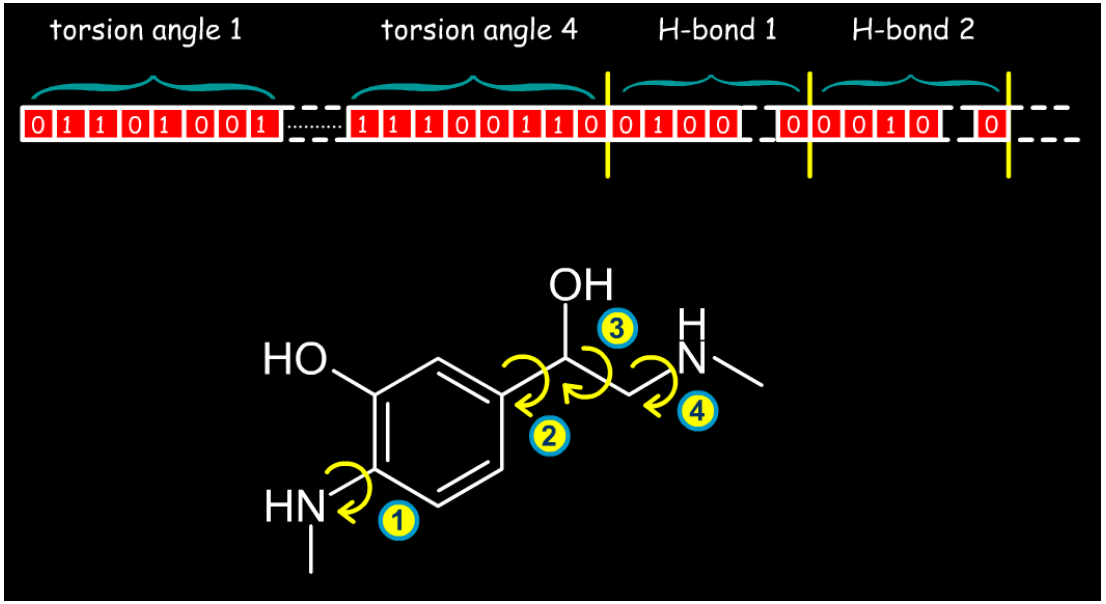
Incremental Algorithm

- The main problem of incremental construction algorithms is their dependence on the choice of the first fragment to be docked, which requires an exploration of several possibilities
- In the example below the initial rigid fragment has been docked three times (A, B, C). For each one, we have a set of potential solutions that can be generated. Each intermediate level is extended to the next one by adding a new next fragment in all possible configurations. Solutions with a bad score are eliminated.



Stochastic Search Methods

- Stochastic search methods modify the conformation of the small molecule in the receptor site and assess it on the fly
- In the case of genetic algorithms, the torsion angles are added on the chromosomes; in Monte Carlo based methods they are set as parameters for optimization



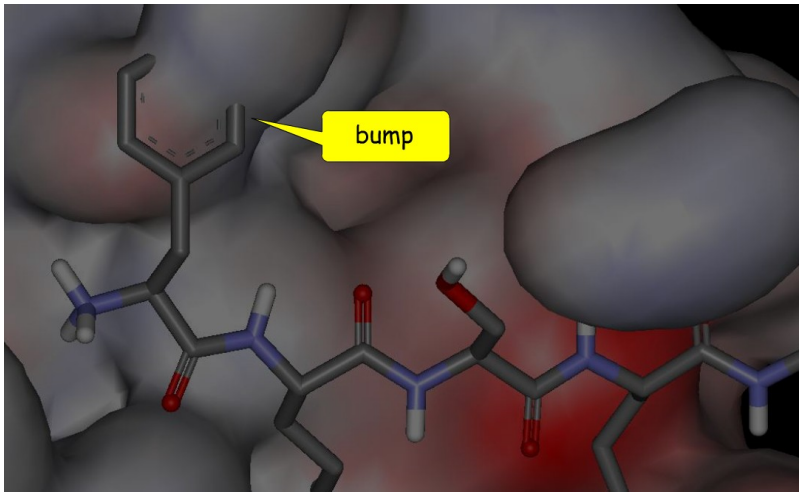
Incorporating Protein Flexibility

Incorporating Protein Flexibility

- Incorporating protein flexibility into molecular docking software is a difficult optimization problem involving a huge number of degrees of freedom that represent the receptor flexibility
- For practical reasons, four types of protein flexibility are recognized:
  - small atom fluctuations (solved by soft docking methods)
  - side-chain flexibility
  - backbone flexibility
  - domain movements

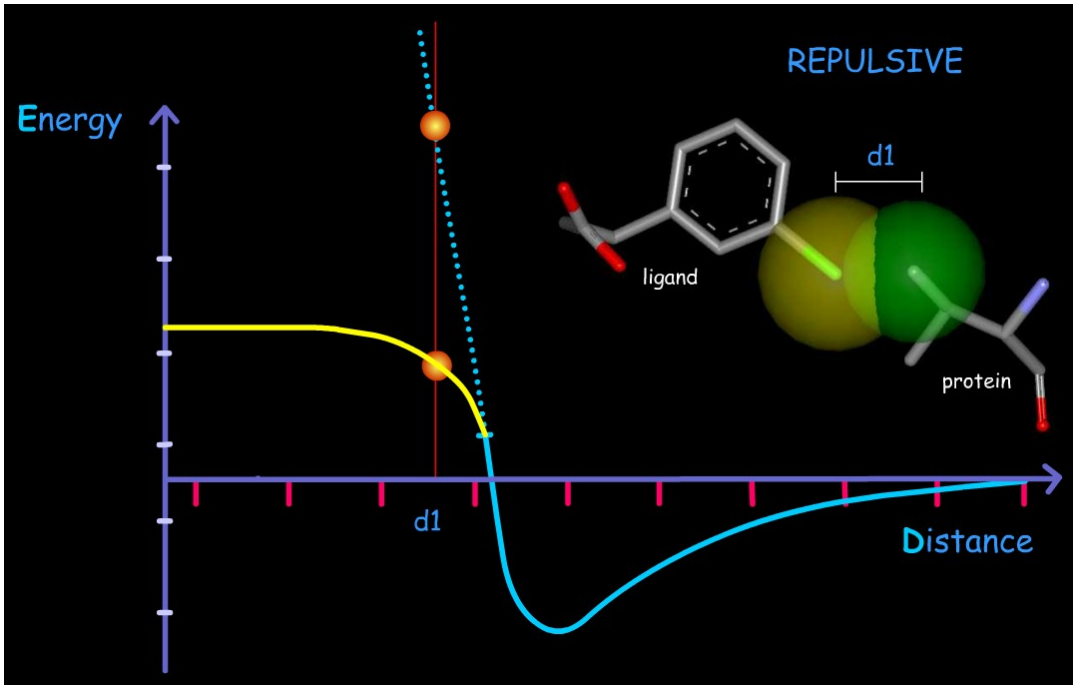
Flexibility Through Soft Docking Methods

- The simple approach to tackle the protein flexibility problem is the 'soft docking' method
- It allows for slight penetrations between the receptor and the ligand molecules; this is a mathematical trick where the receptor and the ligand are held rigid and a 'soft' scoring function is used, allowing some overlap between them.
- In the example below, the fit between the substrate and the protein is acceptable, except however for a bump with the phenyl group. The idea behind a soft scoring function is that instead of resolving bumps by conformational changes, it is possible to reduce their importance by using softer interaction energy functions. A soft scoring function increases the chances of not overlooking good solutions.



Soft Van der Waals Repulsion Functions

- When calculated with force fields from molecular mechanics, steric clashes correspond to high energies
- Modifying the normal Van der Waals repulsion function (in blue) into a softer curve (such as the yellow one) enables them to be less dominant and to simulate the plasticity of the receptor without changing its geometry



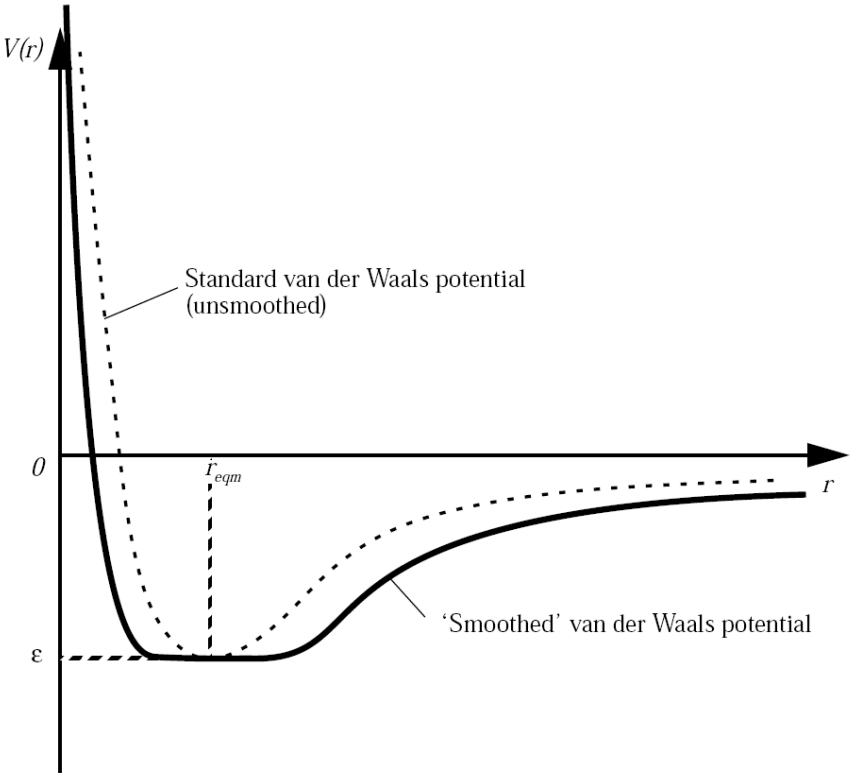
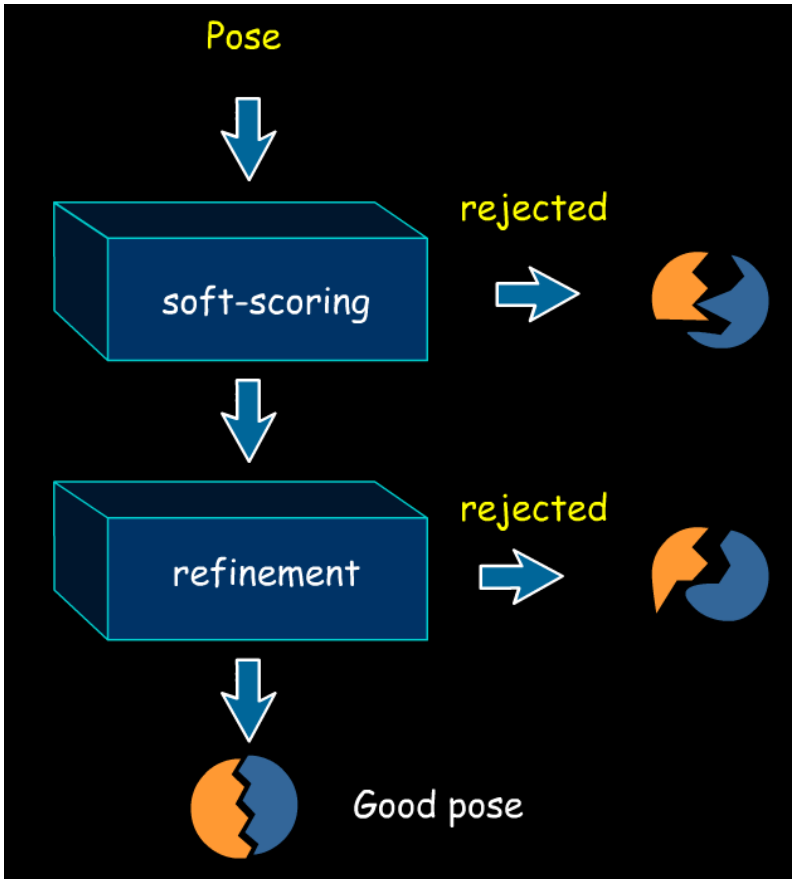


Fig. 2: Another approach implemented in AutoDock software

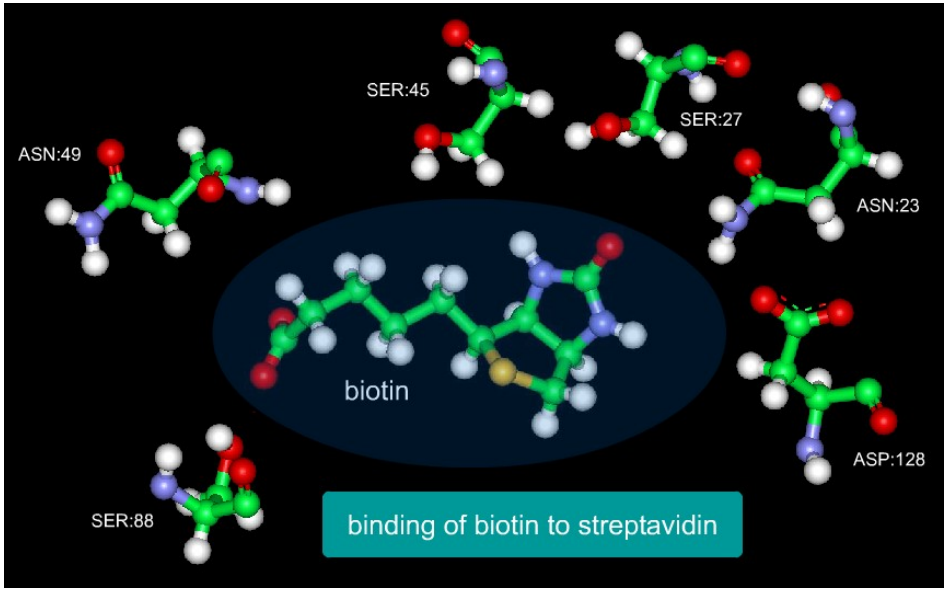
Problems with Soft Scoring

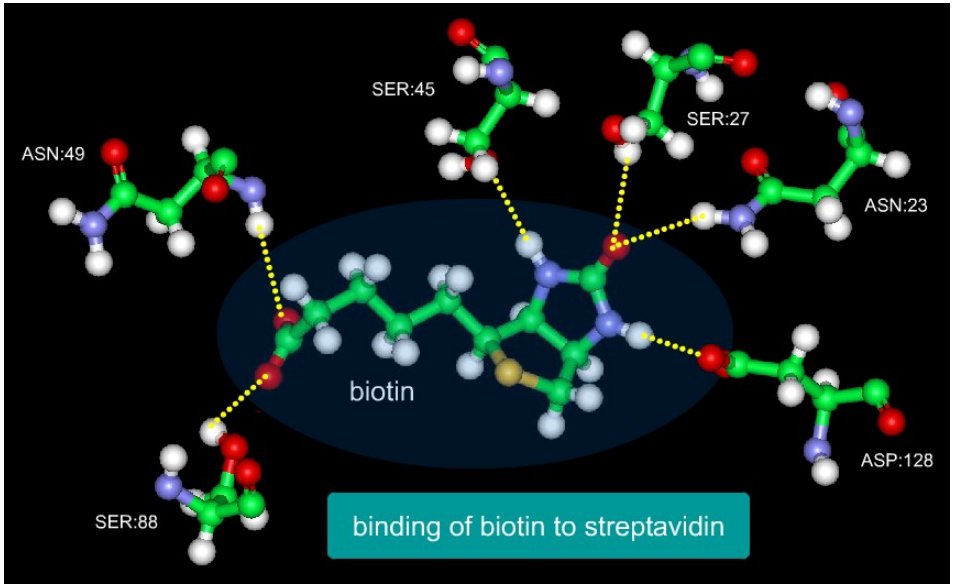
- The major disadvantage of the soft scoring approach is the large number of false-positive hits it produces which therefore makes it harder to discriminate the near native solution from the other candidates
- It has been suggested that the soft scoring function approach should be used as a first filtering algorithm to be complemented by more refined quantitative methods



Protein Side-Chains Flexibility

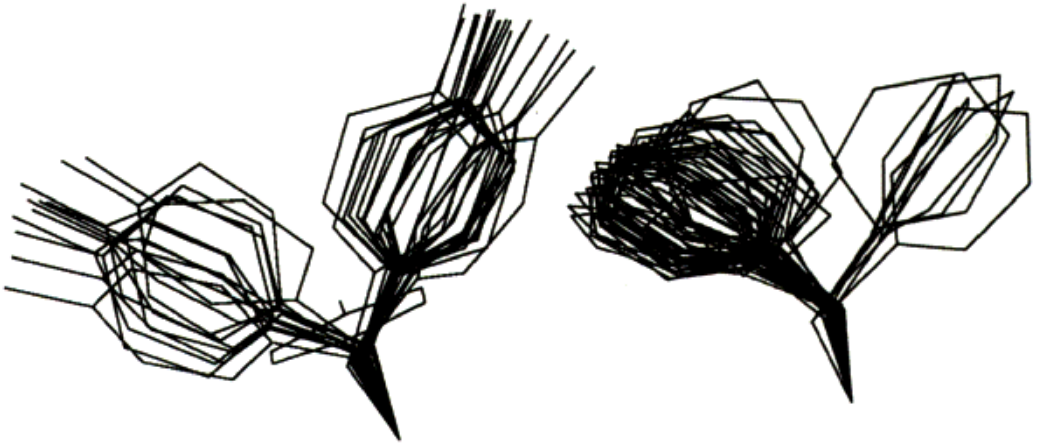
- Side chains are critical for the binding of the ligand to proteins
- Altering the side chain conformations in docking calculations enables the maximization of favourable interactions with the protein
- Studies comparing the X-ray structures of complexes with that of the corresponding unbound proteins (in protein-protein and protein-ligand associations) reveal that about 60% of side chains modify their conformations upon binding





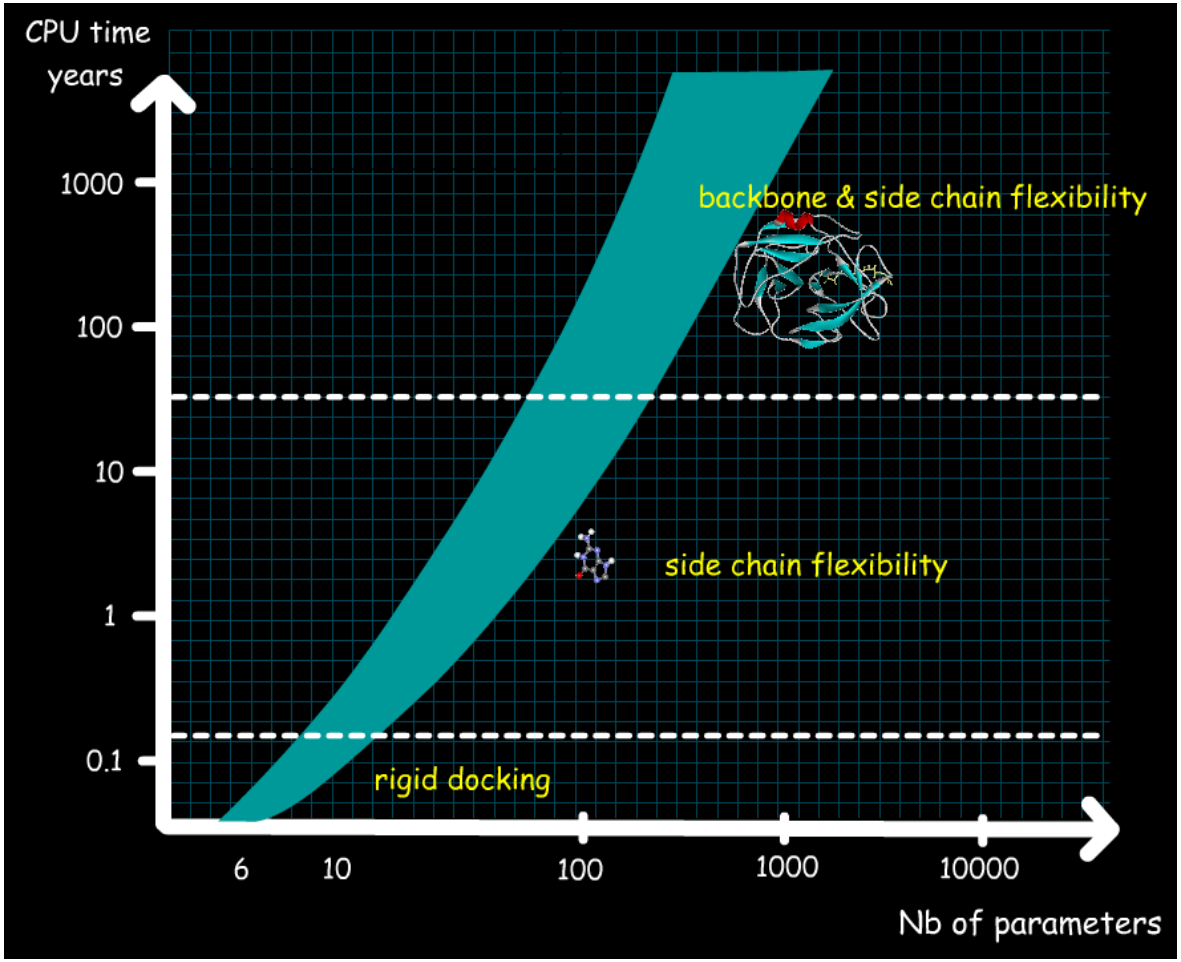
Side Chain Rotamer Libraries

- Exploring the conformational space of protein side chains is a complex optimization problem that leads to a combinatorial explosion of conformers
- Protein side chains can be represented adequately by a small set of discrete rotamers
- Analysis of side-chain Ramachadran plots of pdb structures show that 17 of the 20 amino acids can be represented adequately by 67 side-chain rotamers
- This approach greatly simplified the problem and enabled side-chain flexiblity to be tackled



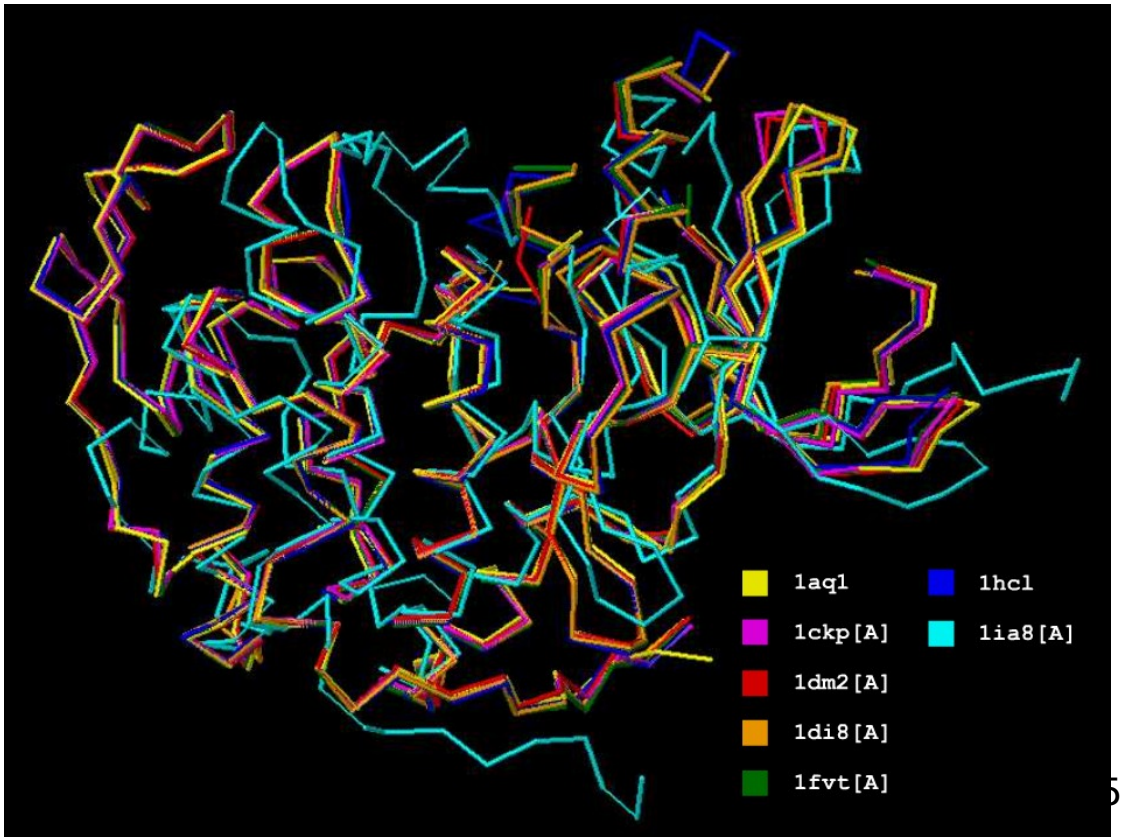
Backbone Flexibility

- One of the greatest challenges in molecular docking is the incorporation of backbone flexibility in docking algorithms
- Due to the complexity level introduced by the huge number degrees of freedom, traditional methods such as the systematic approach or stochastic algorithms cannot be used as a general method



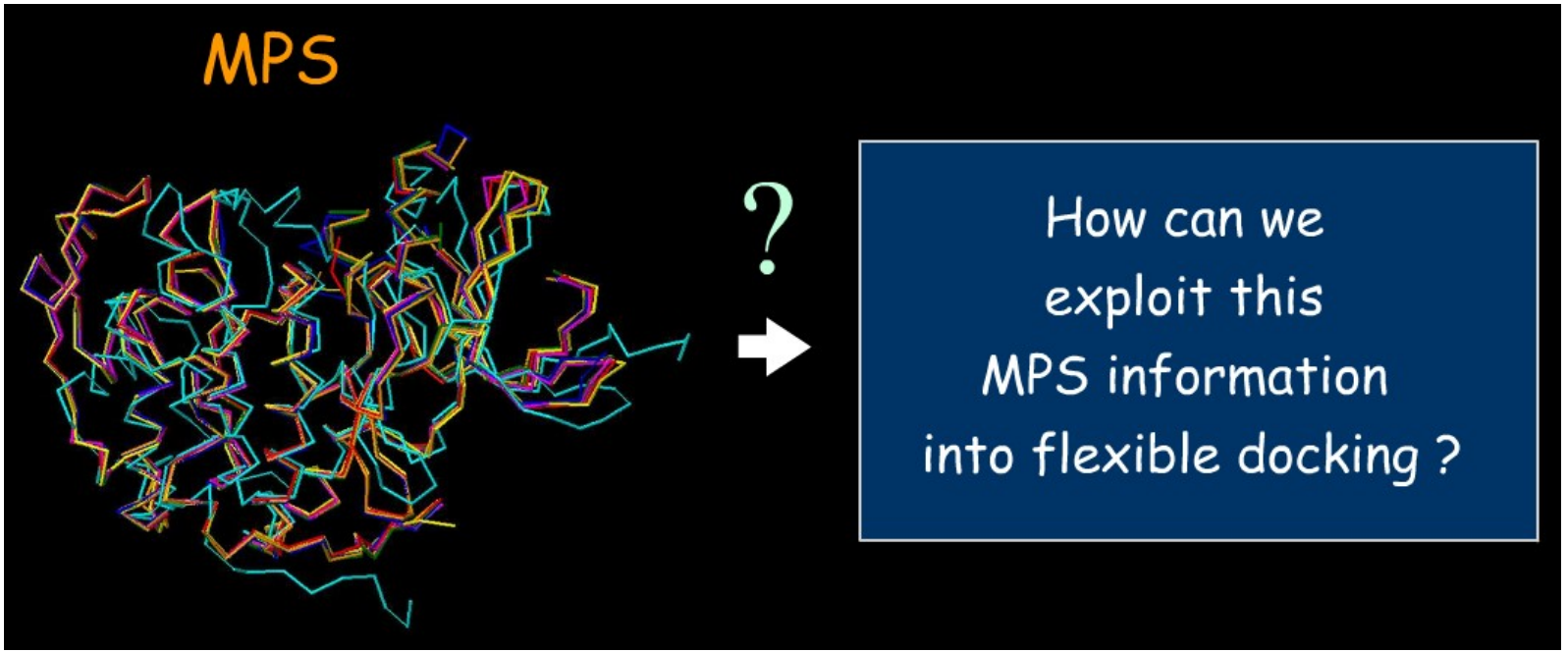
The Multiple Protein Structure (MPS) Approach

- The most efficient method for considering the full flexibility of the protein is the "Multiple Protein Structure" (MPS) approach (also called "multiple receptor structure" or "ensemble approach" or "multiple copy approach")
- The MPS approach is based on the use of multiple structures of the target protein as obtained either from experimental (X-ray, NMR) determinations or generated by theoretical simulations (molecular dynamics)
- The example illustrates the CDK2 protein as determined by seven independent X-ray studies



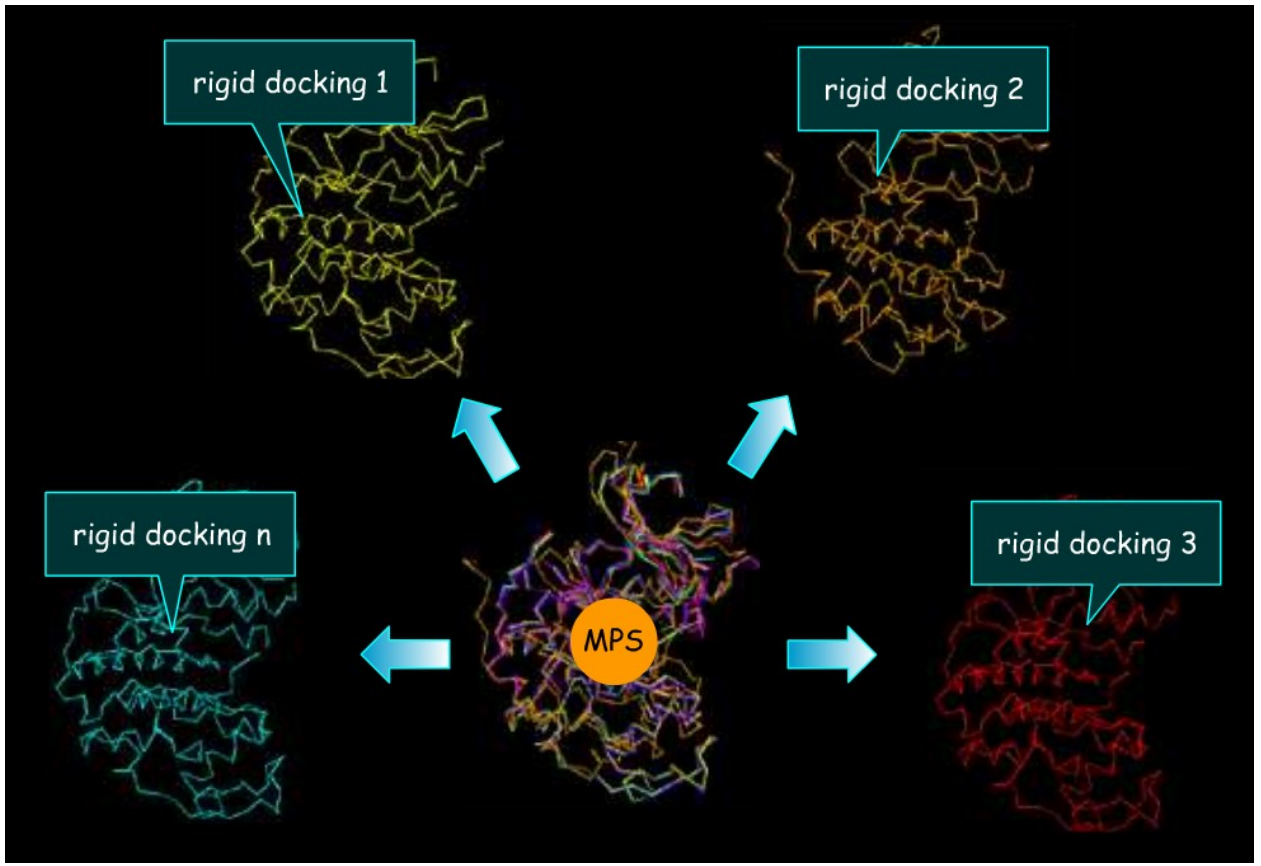
How the MPS are Exploited?

- Diverse approaches have been developed to exploit multiple protein structure (MPS):
  - Successive and Independent Docking Treatments
  - The United Protein Approach
  - The Average Grid Approach



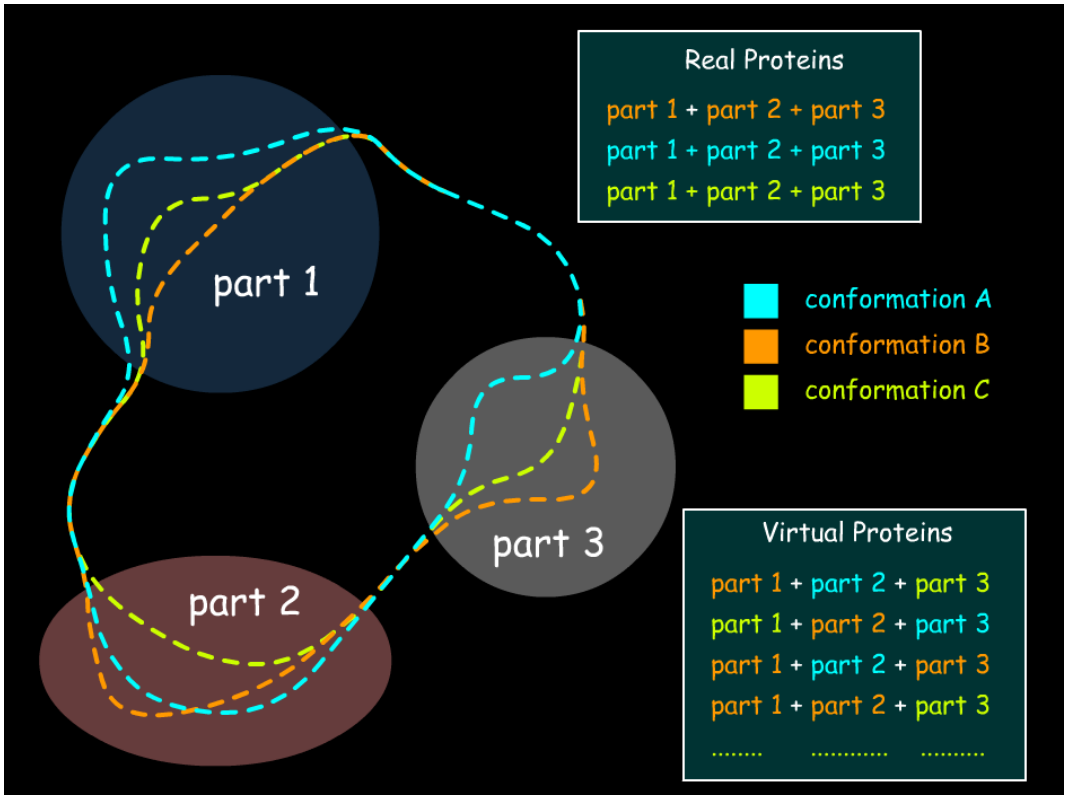
Successive and Independent Docking Treatments

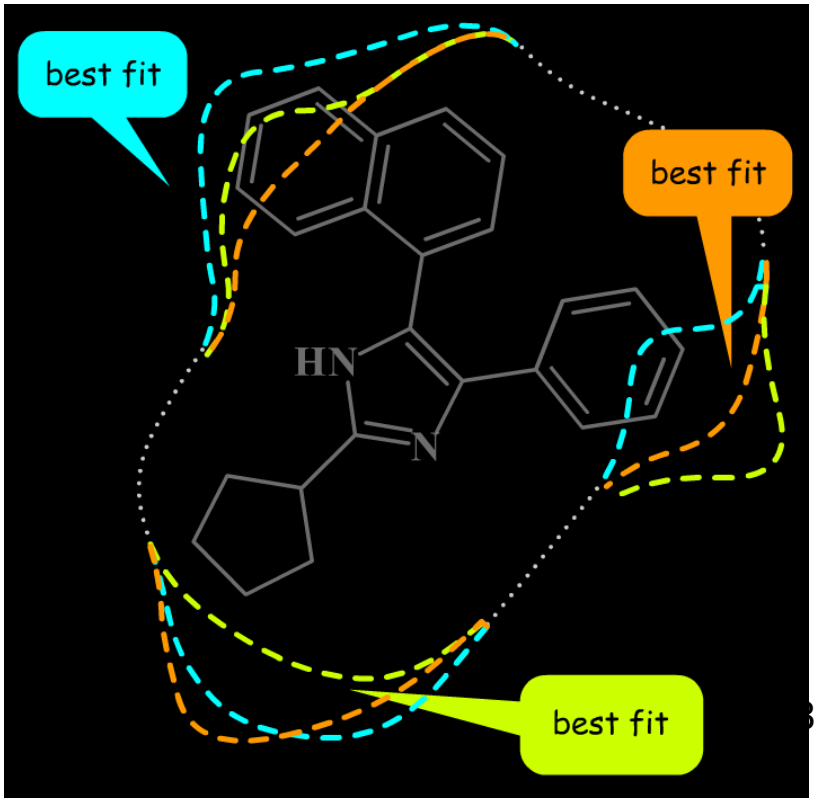
- The most trivial method of exploiting the MPS is to treat each member of the MPS independently, by applying a rigid docking for each protein of the ensemble
- The advantage of this approach is its ease of implementation since the rigid docking software does not need to be modified



The United Protein Approach

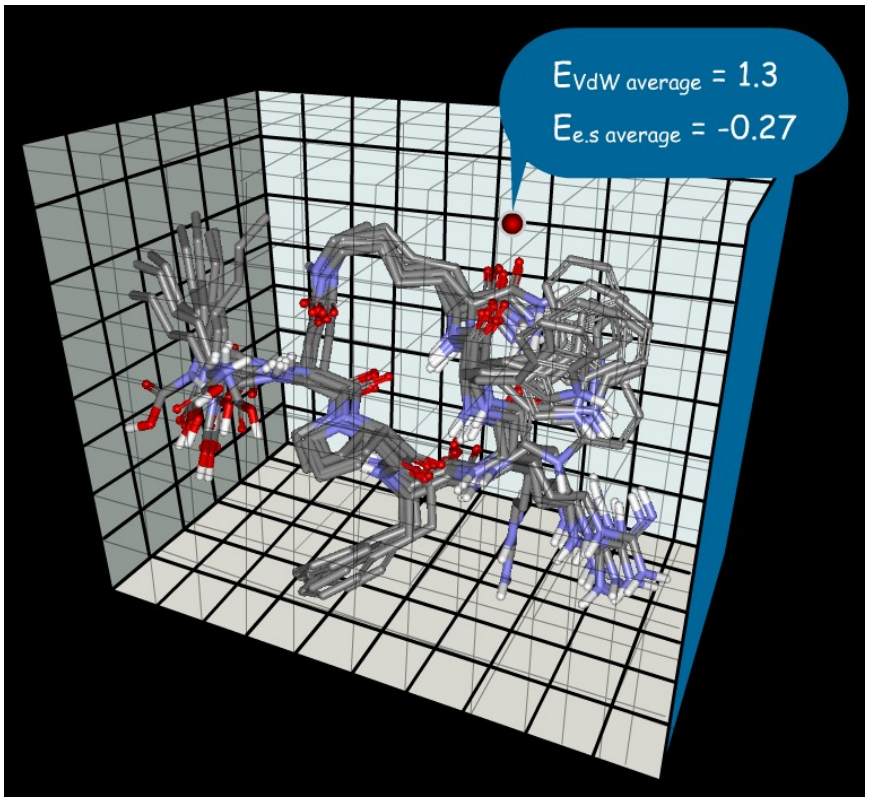
- The united protein approach consists of combining the multiple receptor conformations into a united protein description obtained from the superimposition of all the structures of the ensemble
- Induced-fit is considered locally, each residue (backbone included) in the united protein structure is independent of the others. When a new candidate molecule is docked, the best scoring combination for each residue is selected.





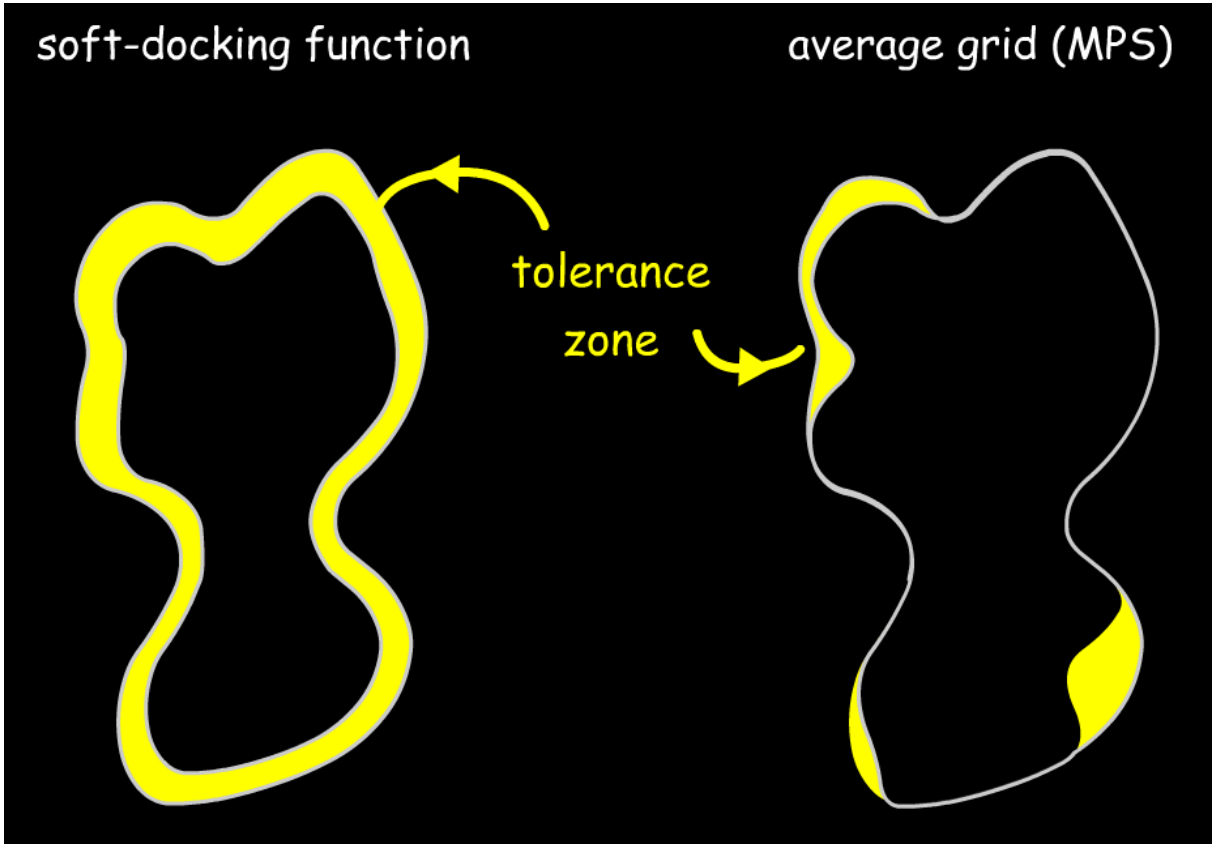
The Average Grid Approach

- A grid is used to characterize the shape and 3D specificity of each protein of the ensemble. The space is systematically explored by calculating the interaction energy (generally Van der Waals and electrostatic energies) between a chemical probe and each protein structure, at each grid point
- The multiple structures are aligned in 3D, the information from aligned multiple protein structures is combined into "energy-weighted" averages of the interaction energies and "geometry-weighted" averages.



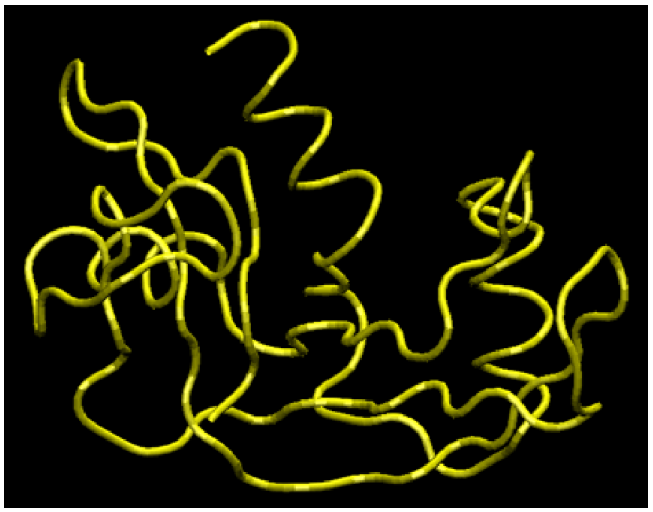
Average Grid Approach vs. Soft Scoring

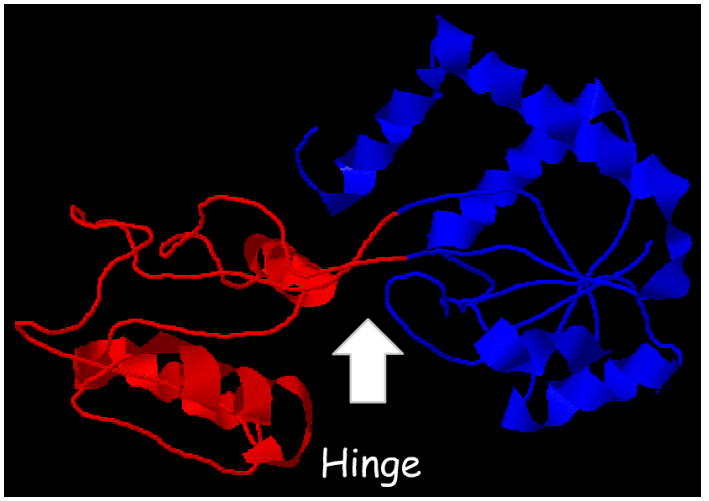
- The average grid approach can be considered as an improvement on soft-scoring function calculations. Both methods alter the scoring function to allow for slight penetrations between the interacting molecules. In the average grid approach, information from the MPS results from many local scoring functions, depending on the flexibility regions of the protein; whereas in the soft-docking approach a mathematical trick is used to modify the scoring function globally.

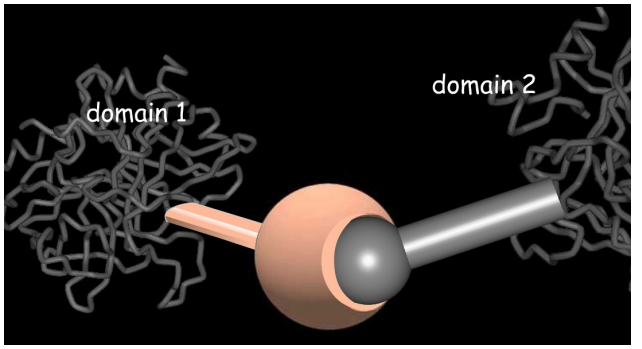


Domain Movements

- Domains motions are often key determinants in the mechanism and the functions of a protein
- Three mechanisms were identified in the movement of domains:
  - intrinsic flexibility can lead some proteins to encounter large motions of domains
  - hinged domain movements are similar to rotations around an articulated joint, in that they involve a small number of residues
  - ball-and-socket motions - involve the rotation of a variable domain with respect to a constant one by a combination of hinge and shear motions
- It is still a challenge to simulate this type of flexibility in computational docking







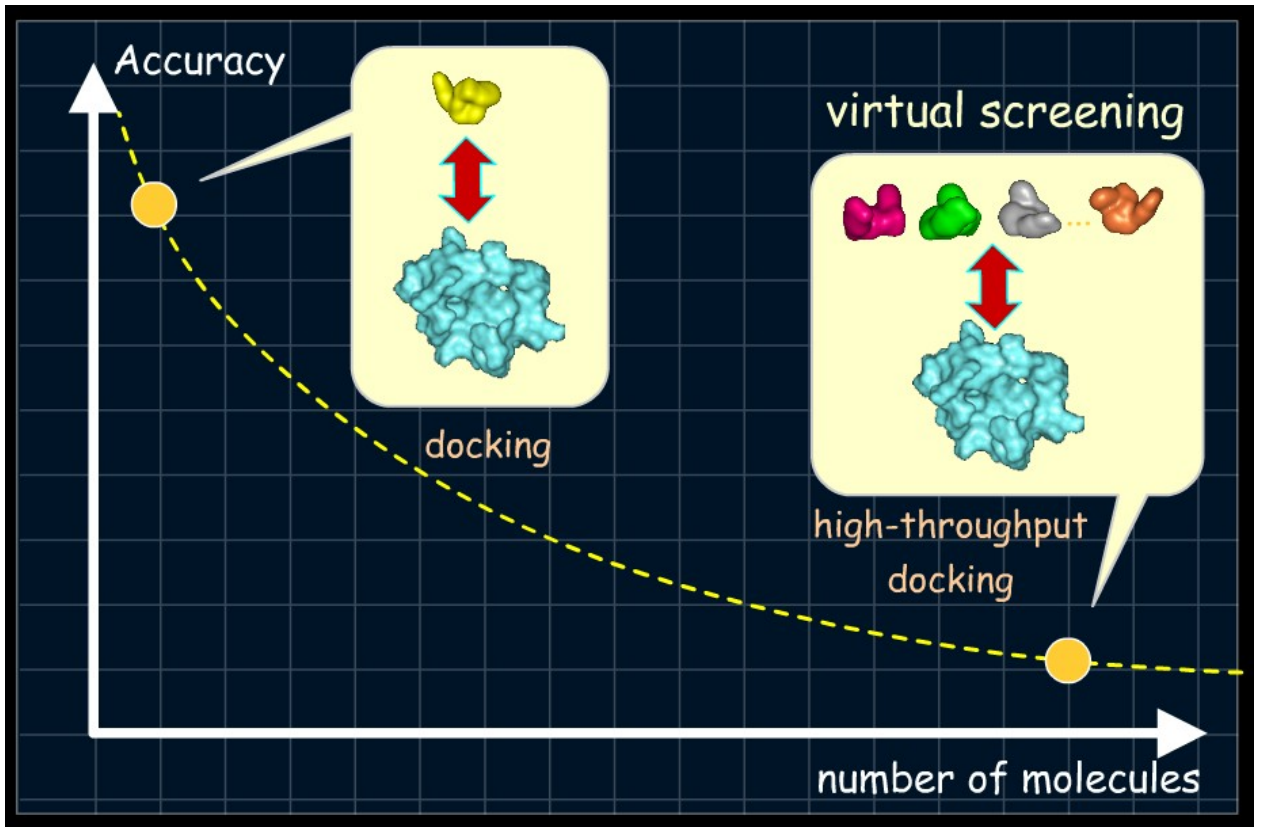
Computational Docking in Drug Discovery

Computational Docking in Drug Discovery

- Docking methods are used by scientists dealing with 3D mechanisms occurring in cellular events: molecular modelers and computational chemists use docking for drug discovery, biochemists to elucidate mechanisms of action, crystallographers for structure refinement, combinatorial chemists for designing new libraries etc...
- Docking has become a key tool in the pharmaceutical and biotech industry
- Typical applications of molecular docking in drug discovery:
  - Virtual Screening
  - Lead Hopping
  - Increasing HTS hit rates

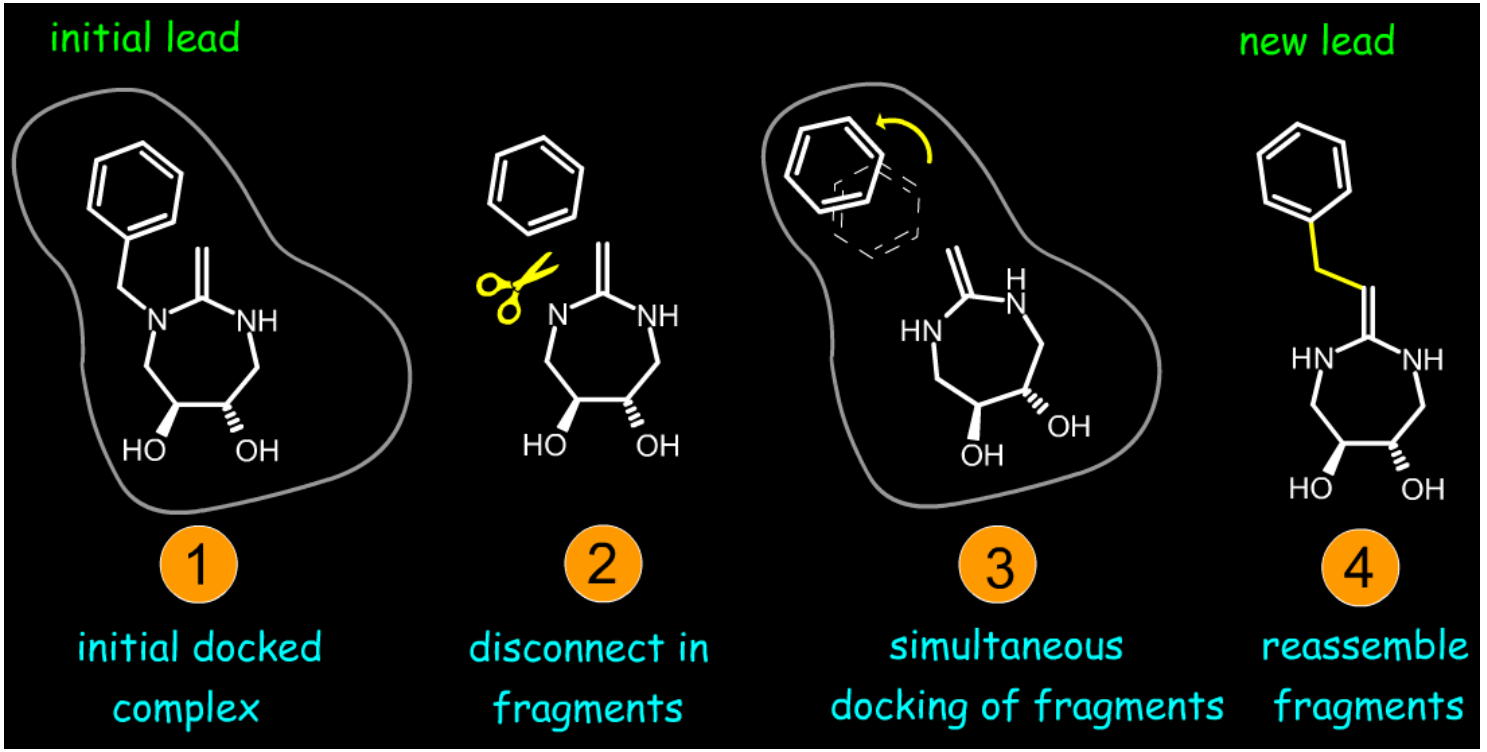
Virtual Screening

- When the goal of docking is to dock all the compounds of a library (the molecules being available or not yet synthesized), the process is called virtual screening or high throughput docking
- Virtual screening identifies active compounds in a large database and ranks them by their affinity to the receptor



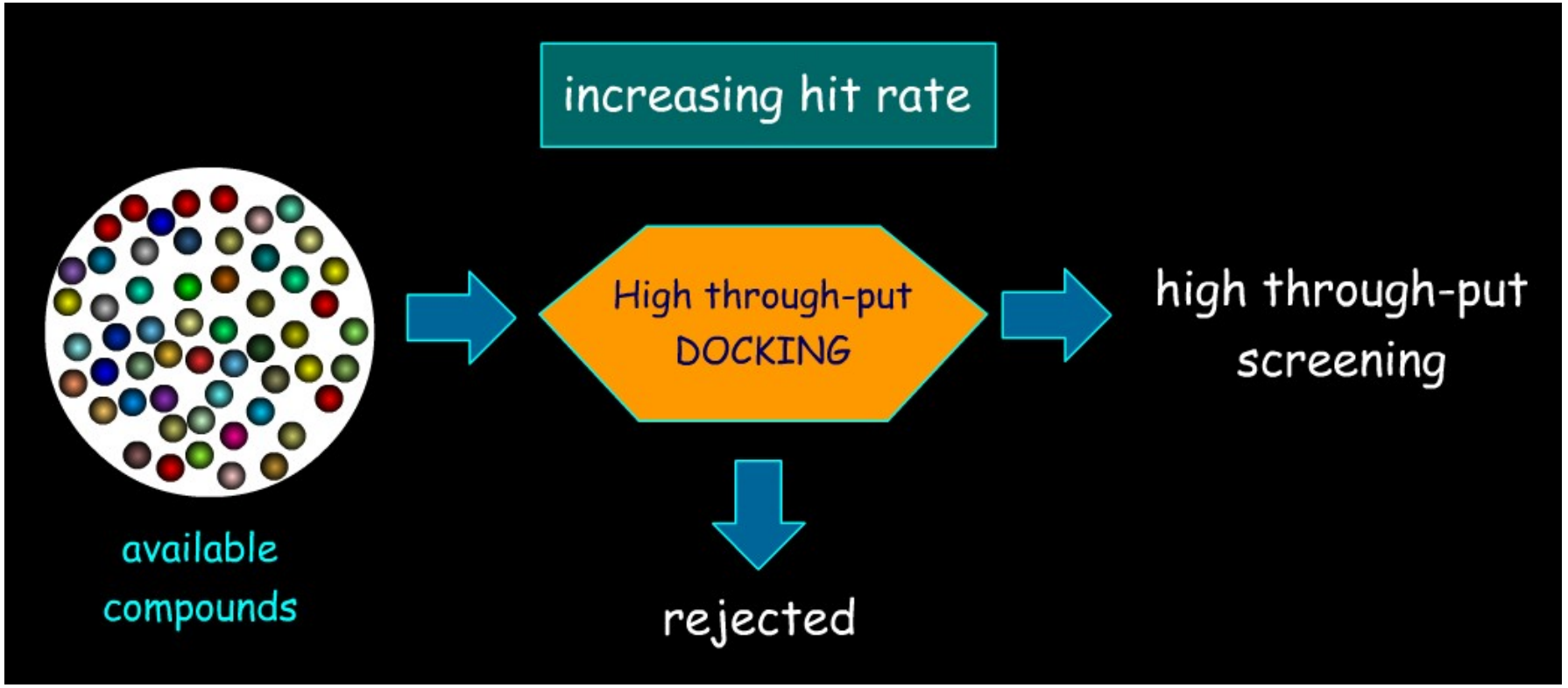
Lead Hopping

- Based on the structure of an active compound, "lead hopping" consists of the identification of novel structures with different topologies that still show the same activity
- Molecular docking can be used for that purpose
- The resulting molecule has a novel structure and may show increased activities, as compared to that of the initial lead



Increasing HTS Hit Rates

- The most widely used type of docking today involves the selection of molecules to be processed for subsequent high through-put screening (HTS)
- The method is not used to recognize active molecules but to eliminate those that are likely to be inactive



Limitations in Computational Docking

- Computational docking has emerged recently as a new discipline. Despite the important achievements that have been obtained, substantial progress remains to be made to exploit the full potential of this approach
- Current challenges of docking methods are:
  - Trade off between efficiency and accuracy
  - More effective scoring functions
  - Better model of flexibility