Protein-Ligand Docking

Alejandro Giorgetti

Outline

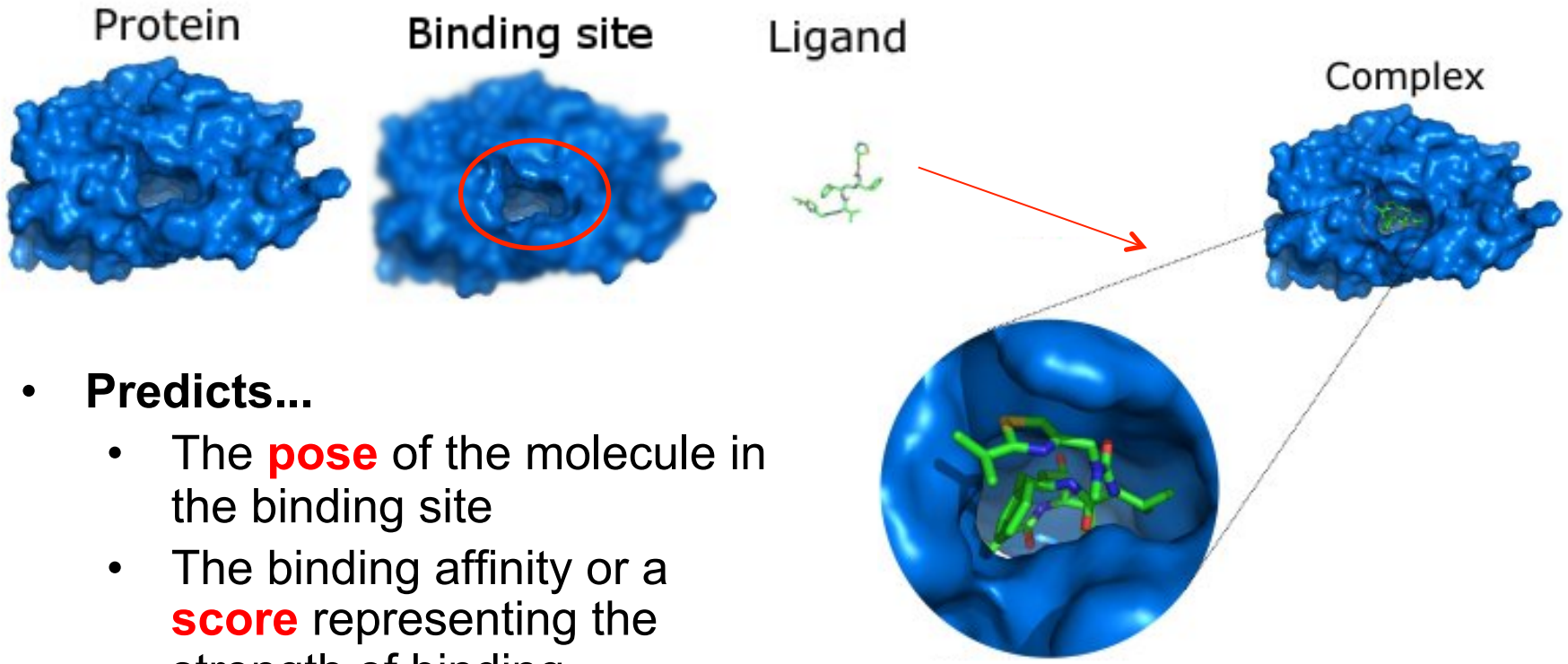
- Introduction to protein-ligand docking
- Searching for poses
- Scoring functions
- Assessing performance
- Practical aspects

Outline

- Introduction to protein-ligand docking
- Searching for poses
- Scoring functions
- Assessing performance
- Practical aspects

Protein-ligand docking

- A Structure-Based Drug Design (SBDD) method
  - "structure" means "using protein structure"
- Computational method that mimics the binding of a ligand to a protein
- **Given...**

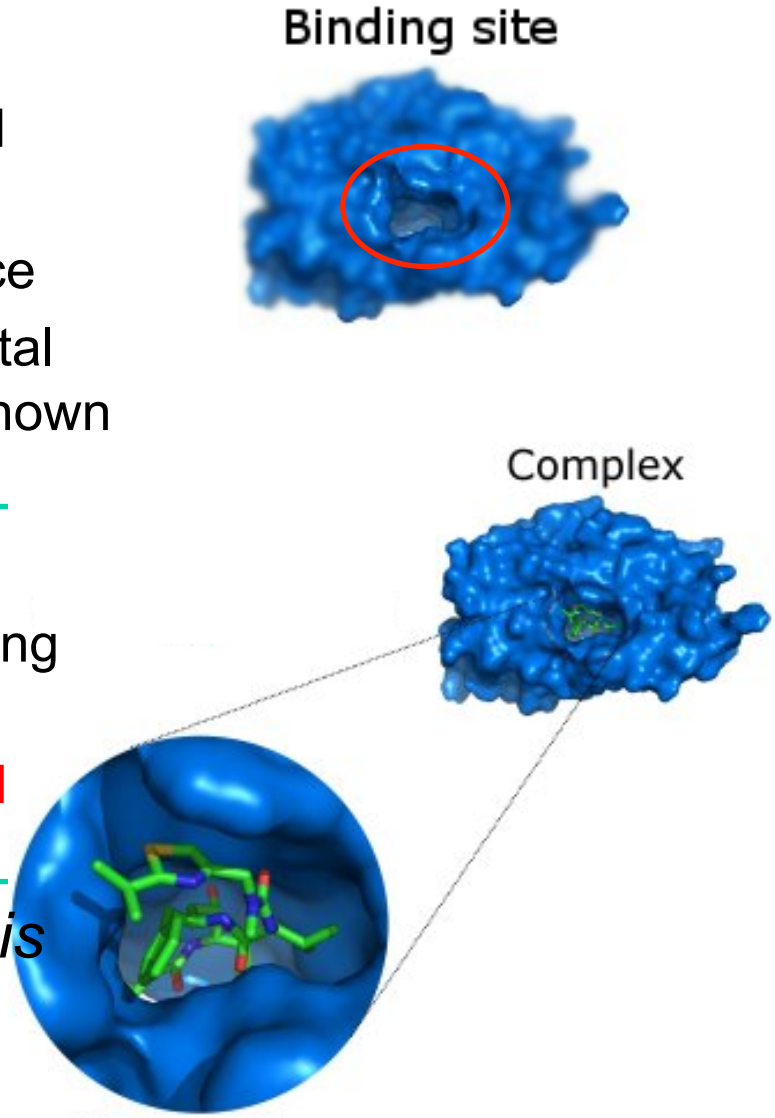


- **Predicts...**
  - The **pose** of the molecule in the binding site
  - The binding affinity or a **score** representing the strength of binding

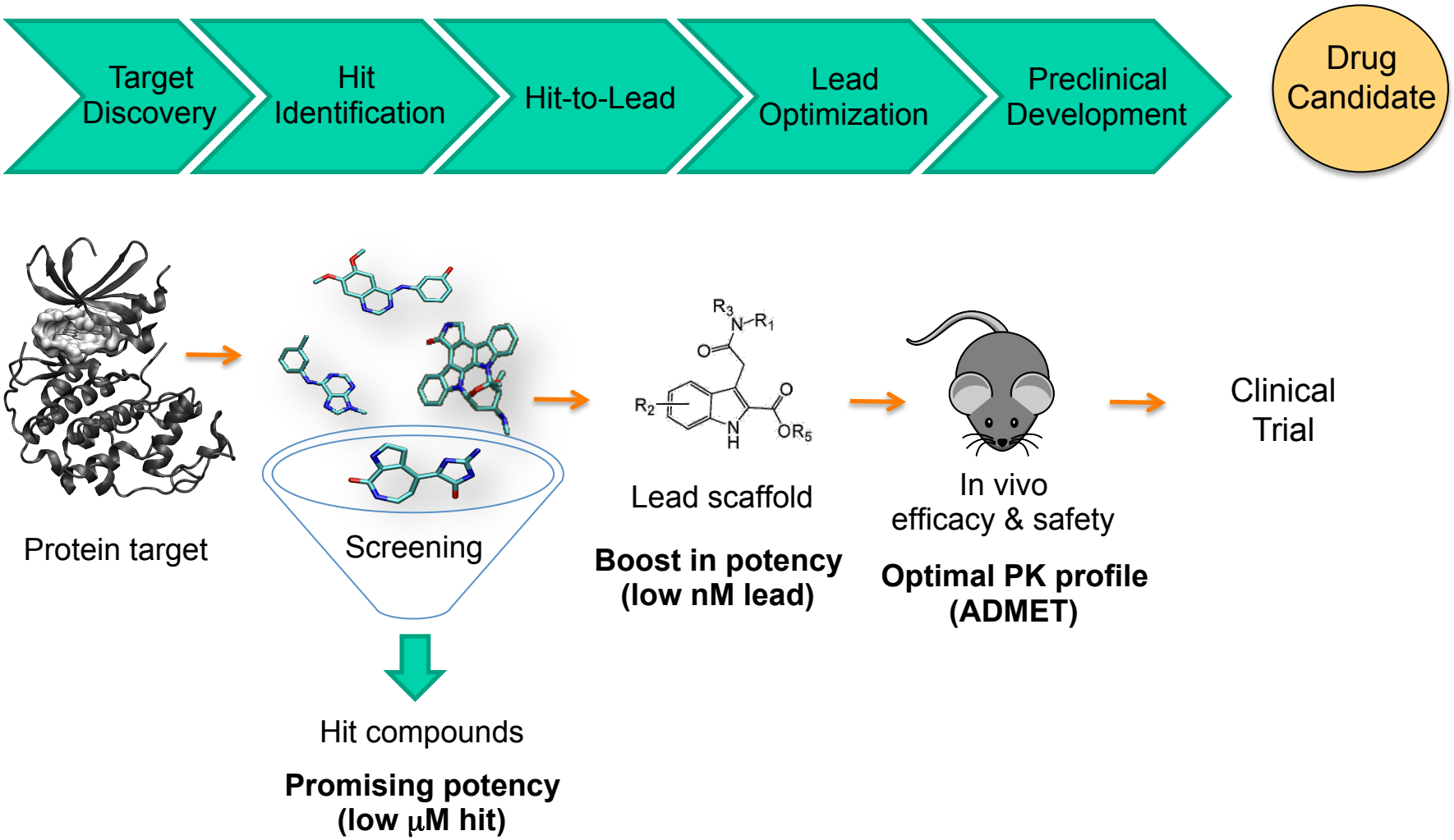
Images from Charaka Goonatilake's web page, Glen Group, Unilever Centre, Cambridge

Pose vs. binding site

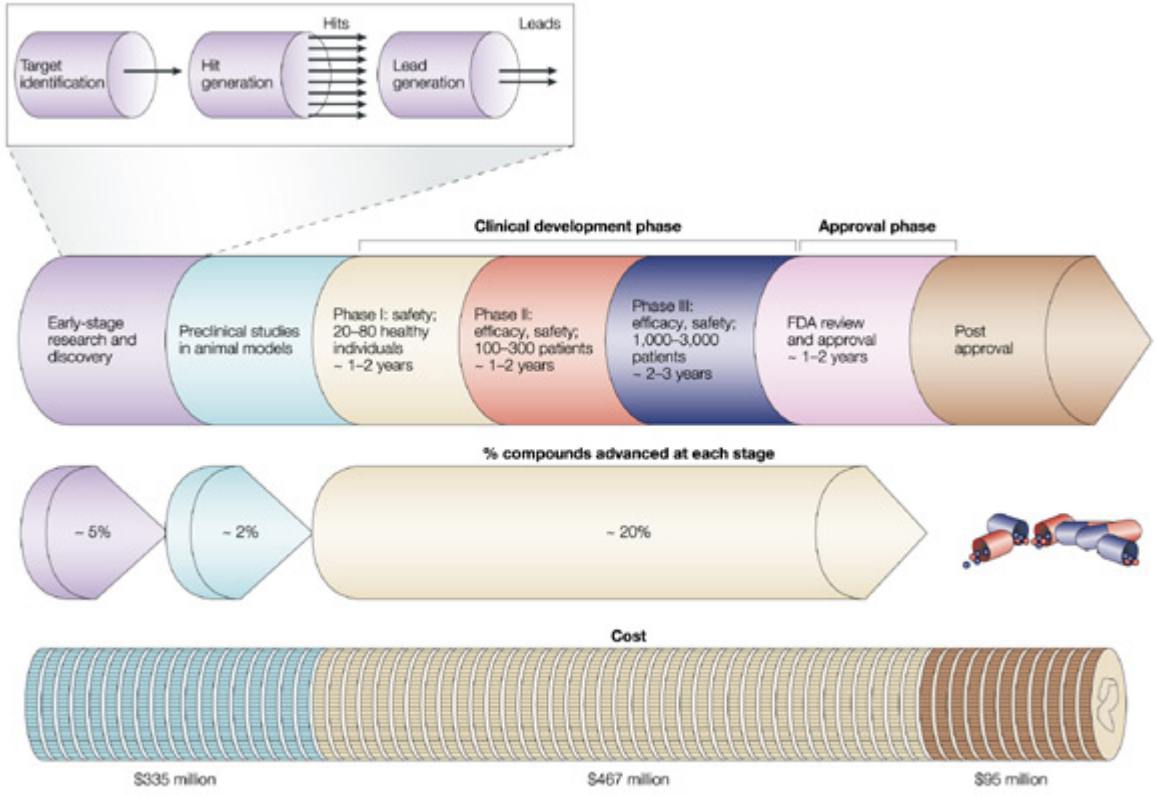
- **Binding site** (or “active site”)
  - the part of the protein where the ligand binds
  - generally a cavity on the protein surface
  - can be identified by looking at the crystal structure of the protein bound with a known inhibitor
- **Pose** (or “binding mode”)
  - The *geometry* of the ligand in the binding site
  - Geometry = **location, orientation and conformation**
- *Sometimes Protein-ligand docking is for identifying the binding site*



Drug Discovery process



Drug Discovery Pipeline



C. O'Driscoll. http://www.nature.com/horizon/chemicalspace/background/pdf/odyssey.pdf

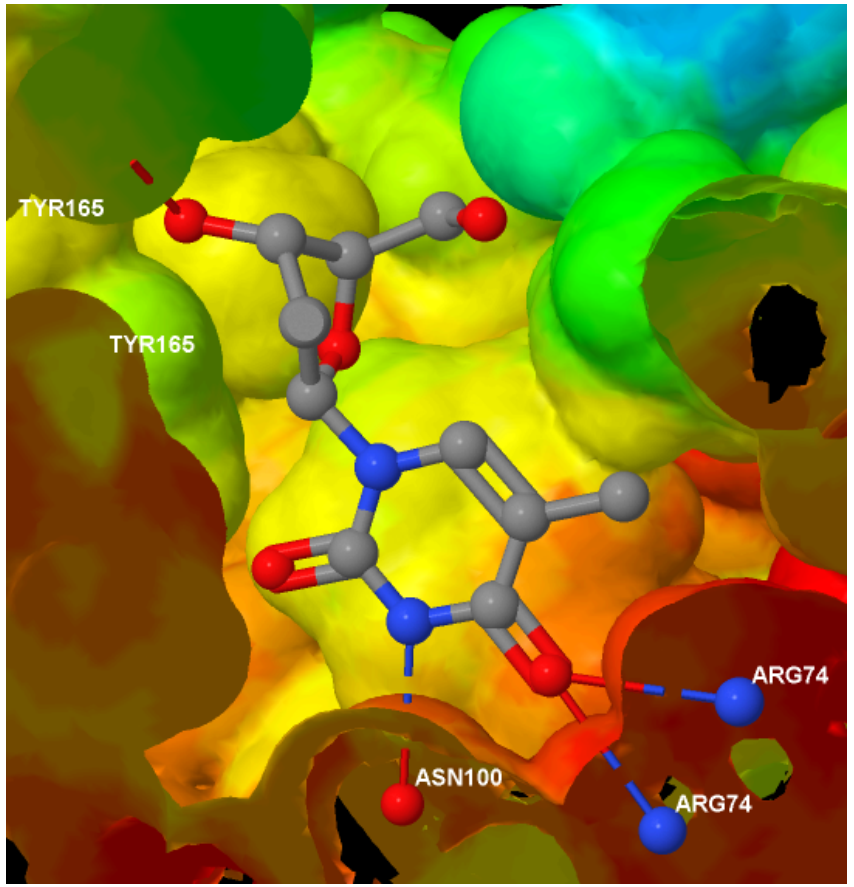
Computer-aided drug design (CADD)

	Known ligand(s)	No known ligand
Known protein structure	**Structure-based drug design (SBDD)** Protein-ligand docking	***De novo* design**
Unknown protein structure	**Ligand-based drug design (LBDD)** *1 or more ligands* • Similarity searching *Several ligands* • Pharmacophore searching *Many ligands (20+)* • Quantitative Structure-Activity Relationships (QSAR)	Need experimental data of some sort

Uses of docking

- The main uses of protein-ligand docking are for
  - Virtual screening, to identify potential lead compounds from a large dataset (next slides)
  - Pose prediction

- **Pose prediction**
- If we know exactly where and how a known ligand binds...
  - We can see which parts are important for binding
  - We can suggest changes to improve affinity
  - Avoid changes that will ‘clash’ with the protein



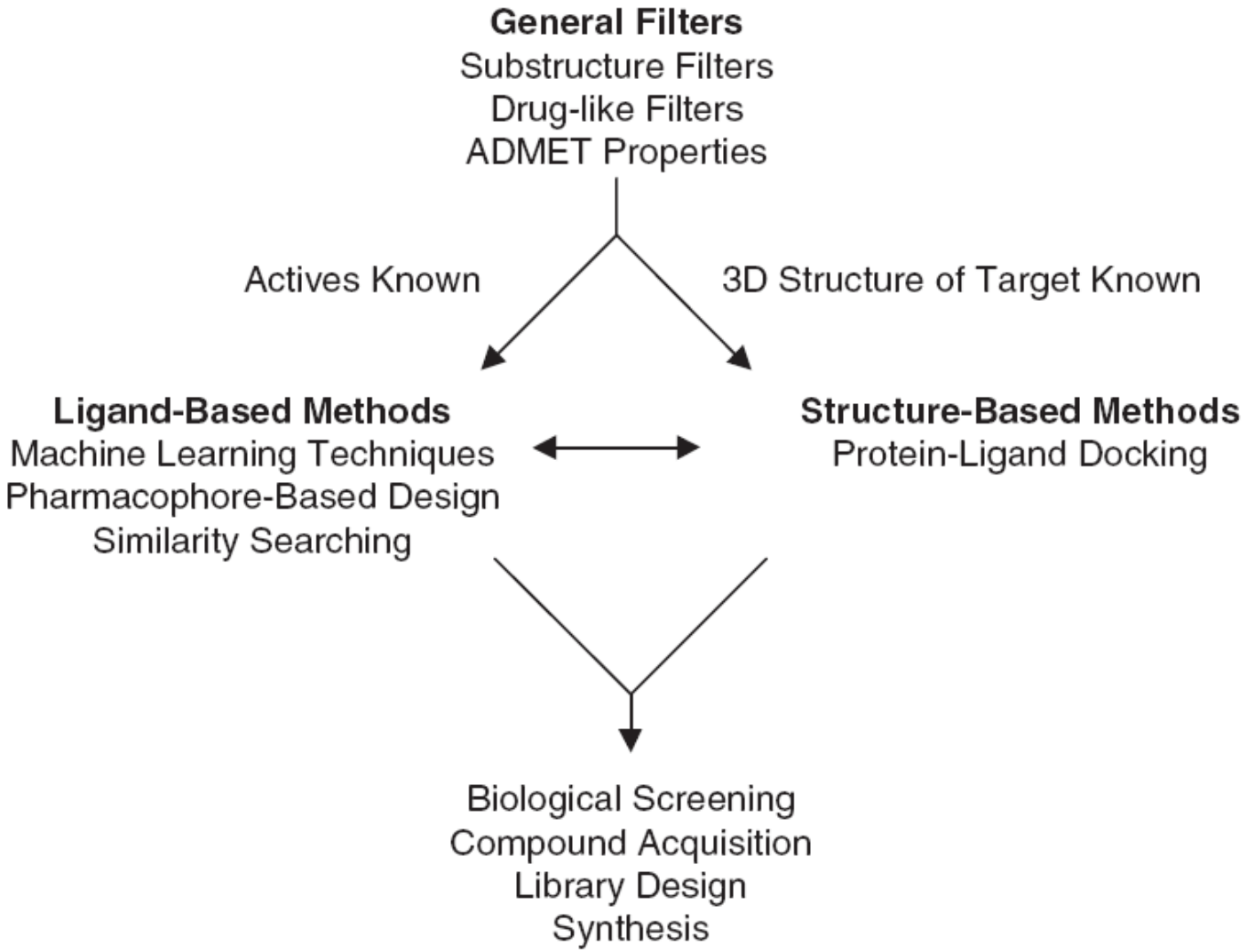
Virtual screening

- Virtual screening is the computational or *in silico* analogue of biological screening

- The aim is to **score**, **rank** or **filter** a set of structures using one or more computational procedures
  - Docking is just one way to do this (see next slide)

- It can be used
  - to help decide which compounds to screen (experimentally)
  - which libraries to synthesise
  - which compounds to purchase from an external company
  - to analyse the results of an experiment.

Virtual screening



AR Leach, VJ Gillet, An Introduction to Cheminformatics

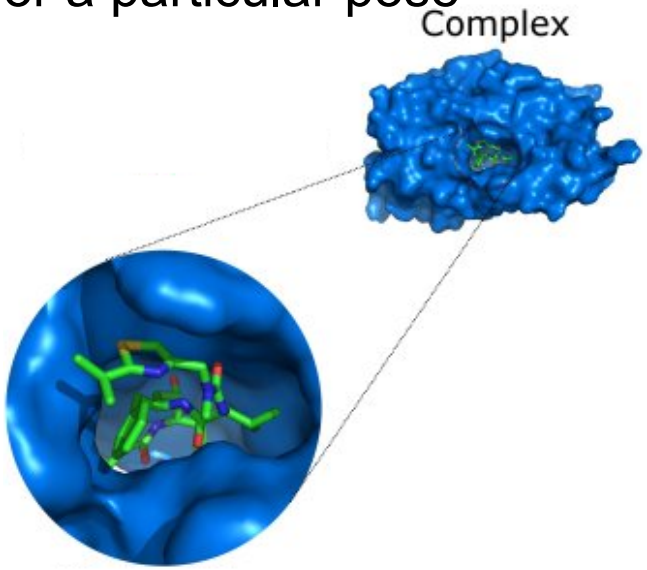
Components of docking software

- Typically, protein-ligand docking software consist of two main components which work together:

- **1. Search algorithm**
  - Generates a large number of poses of a molecule in the binding site

- **2. Scoring function**
  - Calculates a score or binding affinity for a particular pose

- **To give:**
  - The **pose** of the molecule in the binding site
  - The binding affinity or a **score** representing the strength of binding



Final points

- Large number of docking programs available
  - AutoDock, DOCK, e-Hits, FlexX, FRED, Glide, GOLD, LigandFit, QXP, Surflex-Dock…among others
  - Different scoring functions, different search algorithms, different approaches
  - See Section 12.5 in DC Young, Computational Drug Design (Wiley 2009) for good overview of different packages

- Note: protein-ligand docking is not to be confused with the field of protein-protein docking (“protein docking”)

Outline

- Introduction to protein-ligand docking
- Searching for poses
- Scoring functions
- Assessing performance
- Practical aspects

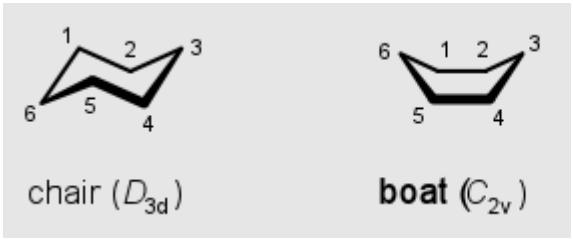
The search space

- The difficulty with protein–ligand docking is in part due to the fact that it involves many degrees of freedom
  - The translation and rotation of one molecule relative to another involves six degrees of freedom
  - There are in addition the conformational degrees of freedom of both the ligand and the protein
  - The solvent may also play a significant role in determining the protein–ligand geometry (often ignored though)

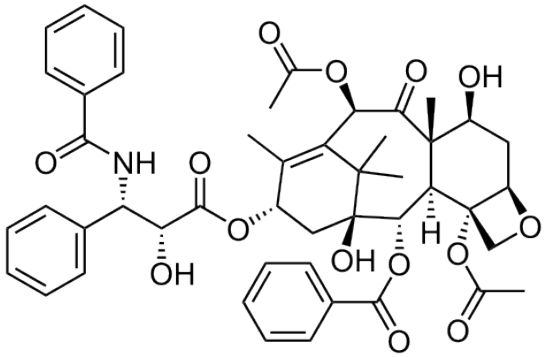
- The search algorithm generates poses, orientations of particular conformations of the molecule in the binding site
  - Tries to cover the search space, if not exhaustively, then as extensively as possible
  - There is a tradeoff between time and search space coverage

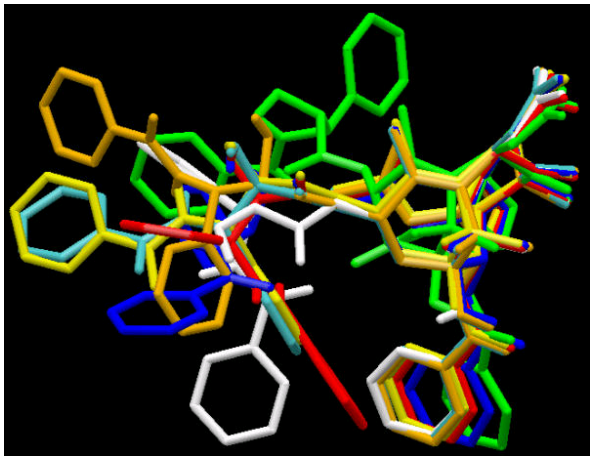
Ligand conformations

- Conformations are different three-dimensional structures of molecules that result from rotation about single bonds
  - That is, they have the same bond lengths and angles but different torsion angles
- For a molecule with N rotatable bonds, if each torsion angle is rotated in increments of θ degrees, number of conformations is $(360º/\theta)^N$
  - If the torsion angles are incremented in steps of 30º, this is 12^N
  - Having too many rotatable bonds results in "combinatorial explosion"
- Also ring conformations
  - Bioactive conformation may not be lowest energy ring conformer

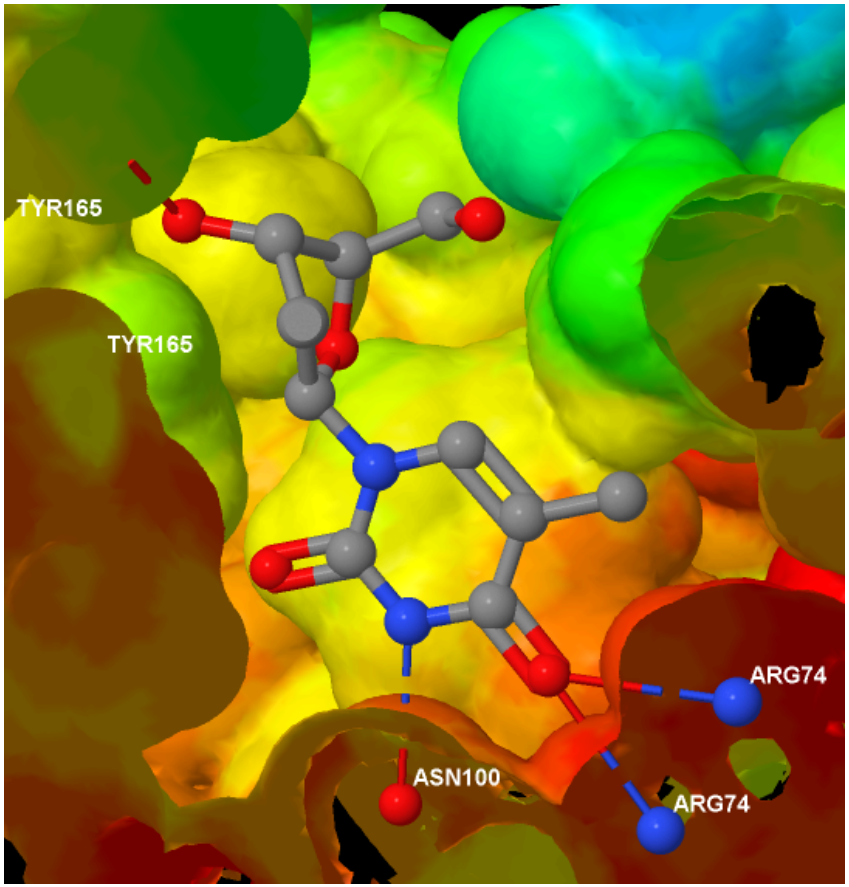


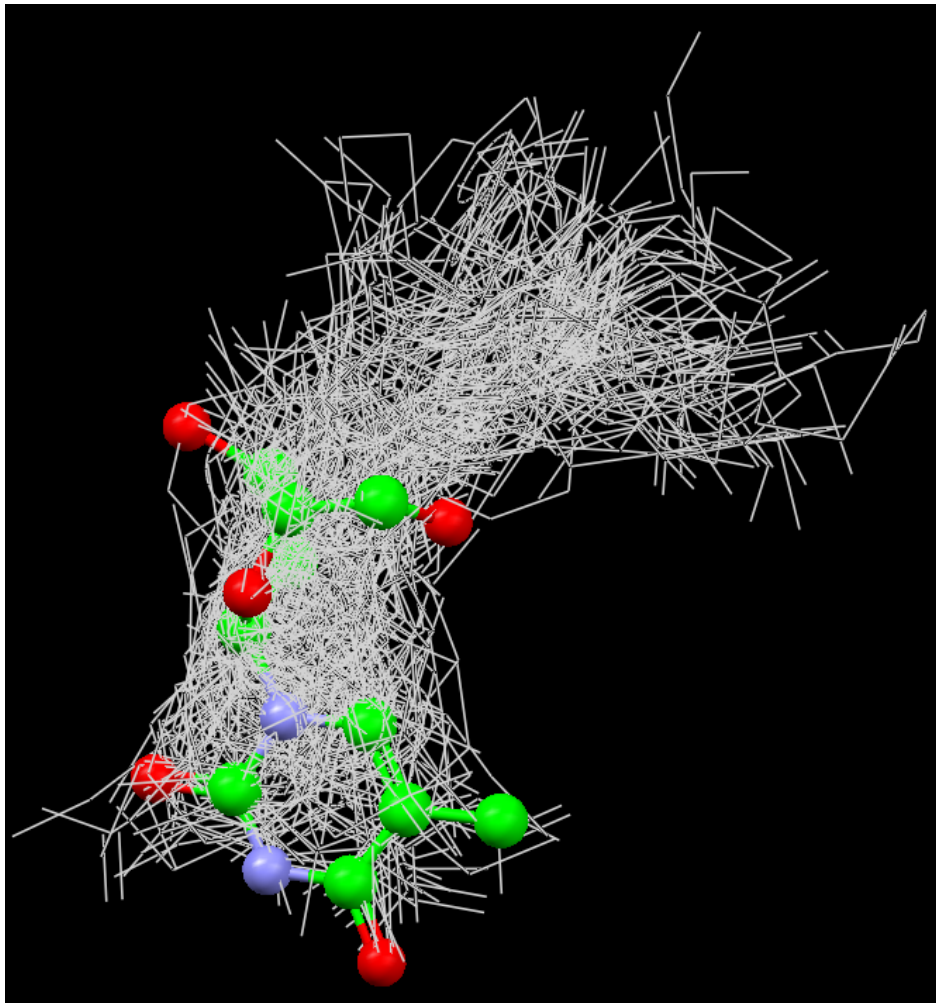
Taxol





Images (from left): IUPAC Gold Book "chair, boat, twist"; "Wikipedia "Taxol"; Lakdawala et al, *BMC Chem Biol*, **2001**, *1*, 2.


TYR165
TYR165
ARG74
ASN100
ARG74



Search Algorithms

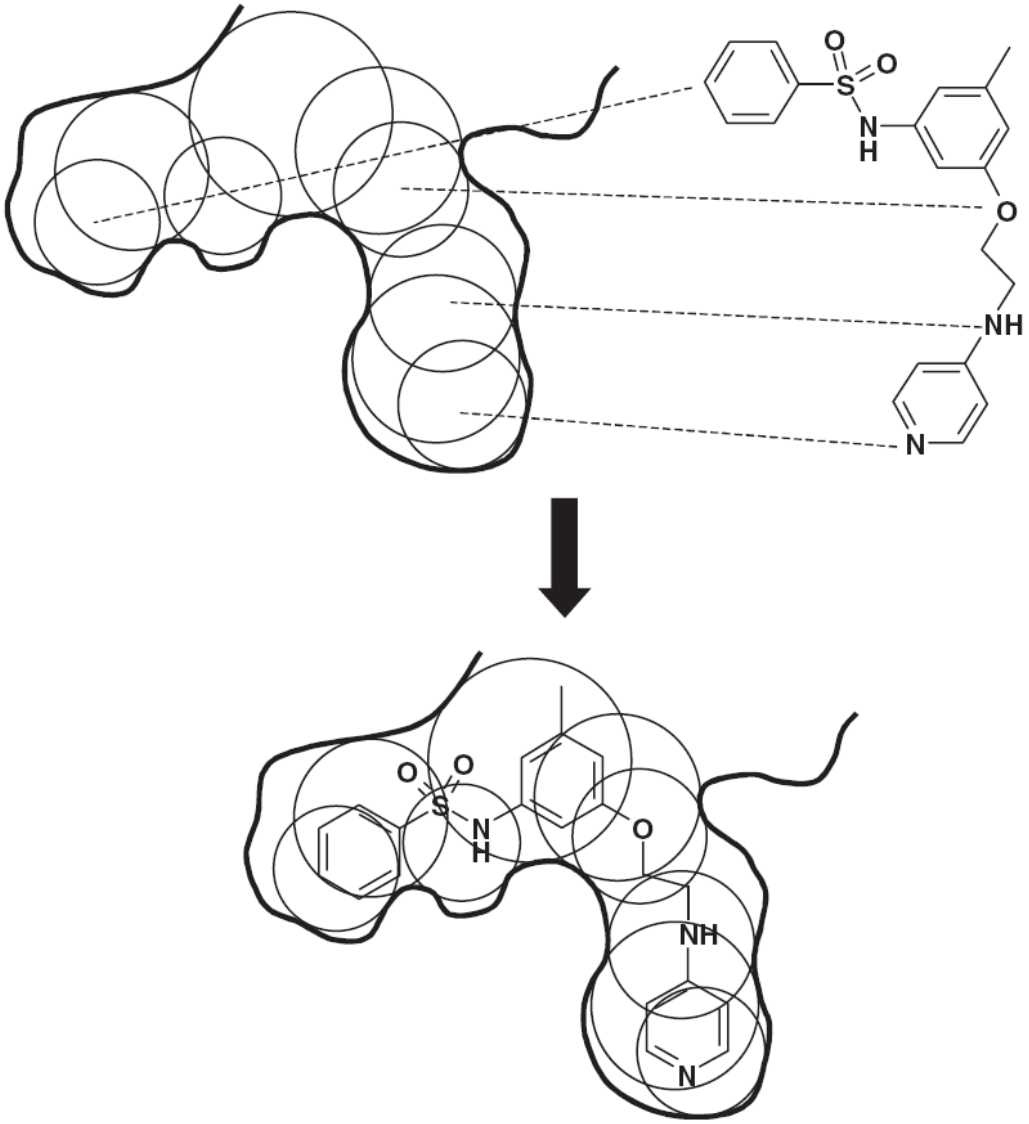
- We can classify the various search algorithms according to the degrees of freedom that they consider
- Rigid docking or flexible docking
  - With respect to the ligand structure
- **Rigid docking**
- The ligand is treated as a rigid structure during the docking
  - Only the translational and rotational degrees of freedom are considered
- To deal with the problem of ligand conformations, a large number of conformations of each ligand are generated in advance and each is docked separately
- Examples: FRED (Fast Rigid Exhaustive Docking) from OpenEye, and one of the earliest docking programs, DOCK

The DOCK algorithm – Rigid docking

- The DOCK algorithm developed by Kuntz and co-workers is generally considered one of the major advances in protein–ligand docking [Kuntz et al., *JMB,* **1982**, *161*, 269]

- The earliest version of the DOCK algorithm only considered rigid body docking and was designed to identify molecules with a high degree of shape complementarity to the protein binding site.

- The first stage of the DOCK method involves the construction of a “negative image” of the binding site consisting of a series of overlapping spheres of varying radii, derived from the molecular surface of the protein



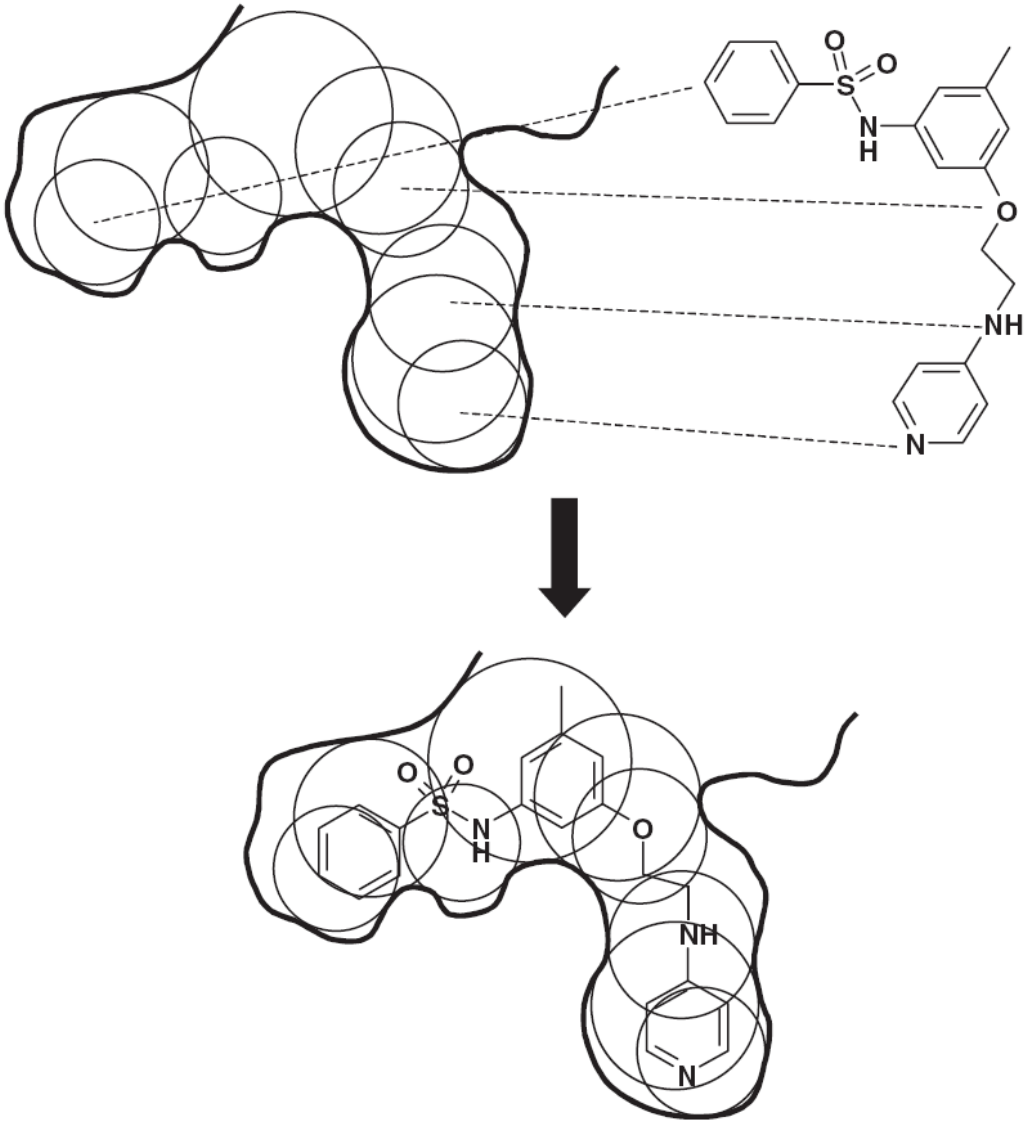
AR Leach, VJ Gillet, An Introduction to Cheminformatics

The DOCK algorithm – Rigid docking

- Ligand atoms are then matched to the sphere centres so that the distances between the atoms equal the distances between the corresponding sphere centres, within some tolerance.

- The ligand conformation is then oriented into the binding site. After checking to ensure that there are no unacceptable steric interactions, it is then scored.

- New orientations are produced by generating new sets of matching ligand atoms and sphere centres. The procedure continues until all possible matches have been considered.



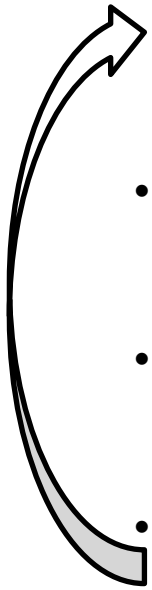
AR Leach, VJ Gillet, An Introduction to Cheminformatics

Flexible docking

- **Flexible docking** is the most common form of docking today
  - Conformations of each molecule are generated on-the-fly by the search algorithm during the docking process
  - The algorithm can avoid considering conformations that do not fit
- Exhaustive (systematic) searching computationally too expensive as the search space is very large
- One common approach is to use stochastic search methods
  - These don't guarantee optimum solution, but good solution within reasonable length of time
  - Stochastic means that they incorporate a degree of randomness
  - Such algorithms include genetic algorithms (GOLD), simulated annealing (AutoDock)
- An alternative is to use incremental construction methods
  - These construct conformations of the ligand within the binding site in a series of stages
  - First one or more "base fragments" are identified which are docked into the binding site
  - The orientations of the base fragment then act as anchors for a systematic conformational analysis of the remainder of the ligand
  - Example: FlexX

Flexible docking using a Genetic Algorithm (GA)

- A genetic algorithm is a stochastic algorithm that can be used to solve global optimisation problems
  - It is based on ideas from Darwinian evolution
- **An initial population of chromosomes is generated randomly**
  - Each chromosome represents a pose, so a large number of poses are randomly generated in the binding site
- **Pairs of high-scoring chromosomes ("parents") are combined to generate "children"**
  - For example, the location of one high-scoring pose may be combined with the torsion angles of another high-scoring pose to generate a new 'child' pose
- **Children are randomly mutated**
  - For example, a torsion angle or the orientation of the child pose might be altered randomly
- **Selection of the fittest to produce the next generation**
  - The highest scoring of the new poses are combined with the highest scoring of the original poses to make the next generation
- **Repeat for N generations or until no significant improvement is observed**
  - We have identified a high scoring pose



Handling protein conformations

- Most docking software treats the protein as rigid
  - Rigid Receptor Approximation
- This approximation may be invalid for a particular protein-ligand complex as...
  - the protein may deform slightly to accommodate different ligands (ligand-induced fit)
  - protein side chains in the active site may adopt different conformations
- Some docking programs allow protein side-chain flexibility
  - For example, selected side chains are allowed to undergo torsional rotation around acyclic bonds
  - Increases the search space
- Larger protein movements can only be handled by separate dockings to different protein conformations

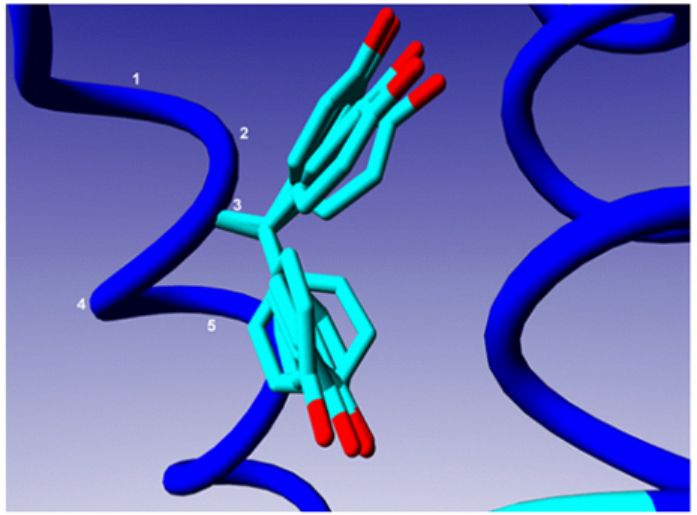


Image of Tyrosine rotamers from CMBI, Univeristy of Nijmegen at http://www.cmbi.kun.nl/mcsis/richardn/explanation.html

Outline

- Introduction to protein-ligand docking
- Searching for poses
- Scoring functions
- Assessing performance
- Practical aspects

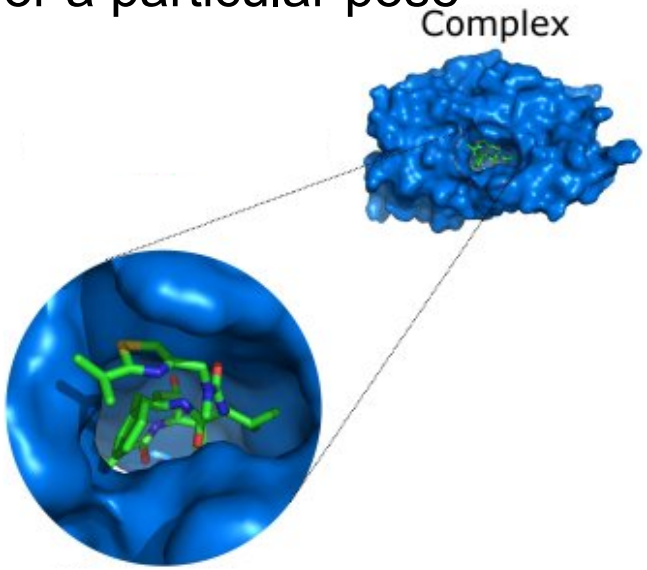
Components of docking software

- Typically, protein-ligand docking software consist of two main components which work together:

- **1. Search algorithm**
  - Generates a large number of poses of a molecule in the binding site

- **2. Scoring function**
  - Calculates a score or binding affinity for a particular pose

- **To give:**
  - The **pose** of the molecule in the binding site
  - The binding affinity or a **score** representing the strength of binding



The perfect scoring function will…

- Accurately calculate the binding affinity
  - Will allow actives to be identified in a virtual screen
  - Be able to rank actives in terms of affinity
- Score the poses of an active higher than poses of an inactive
  - Will rank actives higher than inactives in a virtual screen
- Score the correct pose of the active higher than an incorrect pose of the active
  - Will allow the correct pose of the active to be identified

- "actives" = molecules with biological activity

Classes of scoring function

- Broadly speaking, scoring functions can be divided into the following classes:
  - Forcefield-based
    - Based on terms from molecular mechanics forcefields
    - GoldScore, DOCK, AutoDock
  - Empirical
    - Parameterised against experimental binding affinities
    - ChemScore, PLP, Glide SP/XP
  - Knowledge-based potentials
    - Based on statistical analysis of observed pairwise distributions
    - PMF, DrugScore, ASP

Empirical scoring functions

Journal of Computer-Aided Molecular Design, 8 (1994) 243–256
ESCOM

J-CAMD 247

The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure

Hans-Joachim Böhm

BASF AG, Central Research, D-67056 Ludwigshafen, Germany

Received 20 September 1993
Accepted 15 December 1993

Key words: Proteins; Protein–ligand interaction; De novo design; Scoring function

SUMMARY

A new simple empirical function has been developed that estimates the free energy of binding for a given protein–ligand complex of known 3D structure. The function takes into account hydrogen bonds, ionic interactions, the lipophilic protein–ligand contact surface and the number of rotatable bonds in the ligand. The dataset for the calibration of the function consists of 45 protein–ligand complexes. The new energy function reproduces the binding constants (ranging from $2.5 \cdot 10^{-2}$ to $4 \cdot 10^{-14}$ M, corresponding to binding energies between –9 and –76 kJ/mol) of the dataset with a standard deviation of 7.9 kJ/mol, corresponding to 1.4 orders of magnitude in binding affinity. The individual contributions to protein–ligand binding obtained from the scoring function are: ideal neutral hydrogen bond: –4.7 kJ/mol; ideal ionic interaction: –8.3 kJ/mol; lipophilic contact: –0.17 kJ/mol Å^2; one rotatable bond in the ligand: +1.4 kJ/mol. The function also contains a constant contribution (+5.4 kJ/mol) which may be rationalized as loss of translational and rotational entropy. The function can be evaluated very fast and is therefore also suitable for application in a 3D database search or de novo ligand design program such as LUDI.

Böhm's empirical scoring function

- In general, scoring functions assume that the free energy of binding can be written as a linear sum of terms to reflect the various contributions to binding

$$\Delta G_{bind} = \Delta G_0 + \Delta G_{hb} \sum_{h-bonds} f(\Delta R, \Delta\alpha) + \Delta G_{ionic} \sum_{\substack{\text{ionic} \\ \text{interactions}}} f(\Delta R, \Delta\alpha) + \Delta G_{lipo} |A_{lipo}| + \Delta G_{rot} NROT$$

- Bohm's scoring function included contributions from hydrogen bonding, ionic interactions, lipophilic interactions and the loss of internal conformational freedom of the ligand.

- The ΔG values on the right of the equation are all constants (see next slide)
- $\Delta \mathbf{G_o}$ is a contribution to the binding energy that does not directly depend on any specific interactions with the protein
- The **hydrogen bonding** and **ionic terms** are both dependent on the geometry of the interaction, with large deviations from ideal geometries (ideal distance R, ideal angle α) being penalised.
- The **lipophilic term** is proportional to the contact surface area (Alipo) between protein and ligand involving non-polar atoms.
- The **conformational entropy term** is the penalty associated with freezing internal rotations of the ligand. It is largely entropic in nature. Here the value is directly proportional to the number of rotatable bonds in the ligand (NROT).

Böhm's empirical scoring function

$$\Delta G_{bind} = \Delta G_0 + \Delta G_{hb} \sum_{h-bonds} f(\Delta R, \Delta\alpha) + \Delta G_{ionic} \sum_{\substack{\text{ionic} \\ \text{interactions}}} f(\Delta R, \Delta\alpha) + \Delta G_{lipo} |A_{lipo}| + \Delta G_{rot} NROT$$

- This scoring function is an empirical scoring function
  - Empirical = incorporates some experimental data
- The coefficients (ΔG) in the equation were determined using multiple linear regression on experimental binding data for 45 protein–ligand complexes
- Although the terms in the equation may differ, this general approach has been applied to the development of many different empirical scoring functions

Knowledge-based potentials

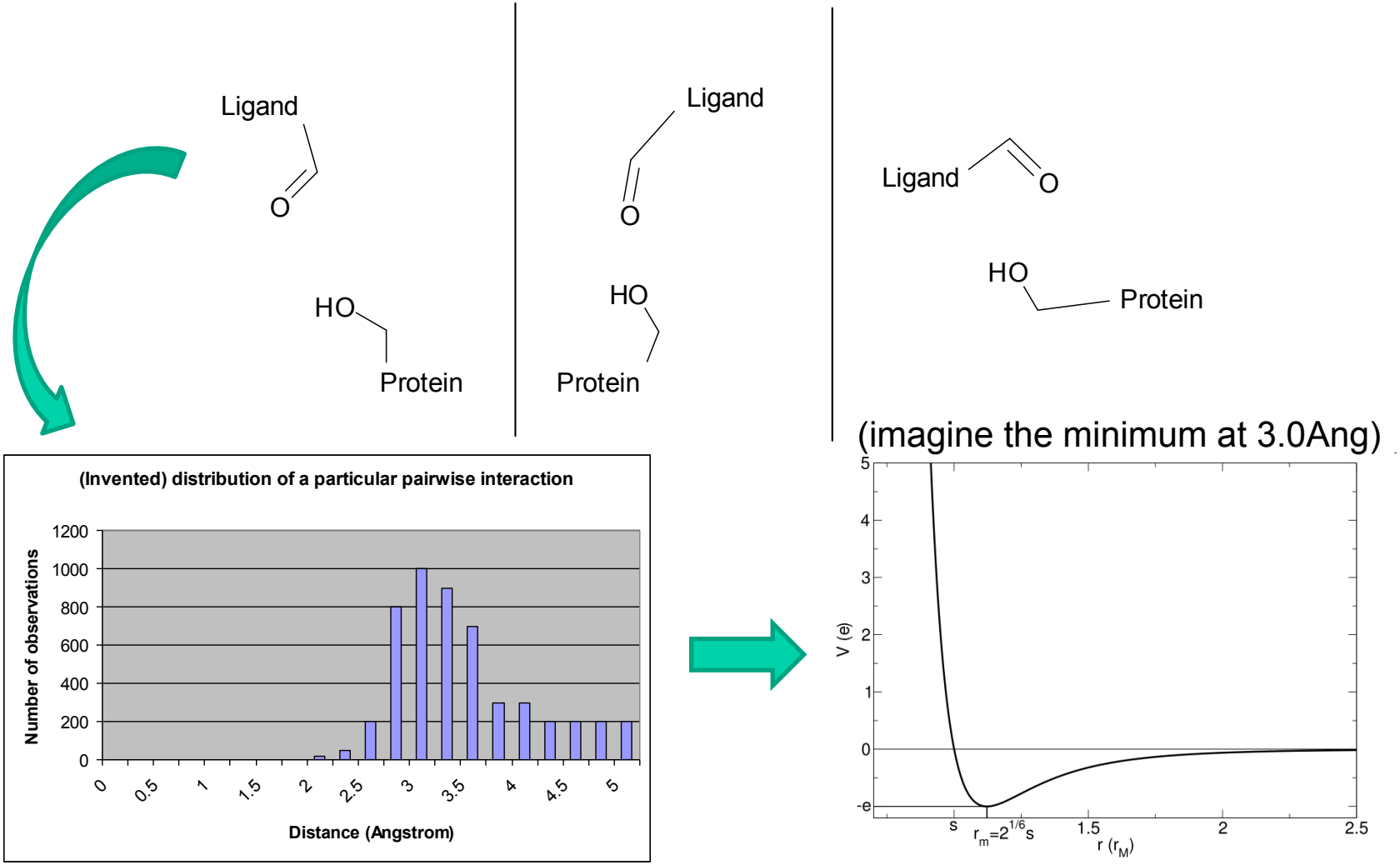
- Statistical potentials

- Based on a comparison between the observed number of contacts between certain atom types (e.g. sp^2-hybridised oxygens in the ligand and aromatic carbons in the protein) and the number of contacts one would expect if there were no interaction between the atoms (the reference state)

- Derived from an analysis of pairs of non-bonded interactions between proteins and ligands in PDB
  - Observed distributions of geometries of ligands in crystal structures are used to deduce the potential that gave rise to the distribution
  - Hence “knowledge-based” potential

Knowledge-based potentials

For example, creating the distributions of **ligand carbonyl oxygens** to **protein hydroxyl groups**:



Knowledge-based potentials

- Some pairwise interactions may occur seldom in the PDB
  - Resulting distribution may be inaccurate
- Doesn't take into account directionality of interactions, e.g. hydrogen bonds
  - Just based on pairwise distances
- Resulting score contains contributions from a large number of pairwise interactions
  - Difficult to identify problems and to improve
- Sensitive to definition of reference state
  - DrugScore has a different reference state than ASP (Astex Statistical Potential)

Alternative scoring strategies

- Consensus docking
  - Carry out two (or more) separate docking experiments and rank based on the mean of the ranks from the two dockings (rank-by-rank)
  - Rationale: a true active will be scored well by both scoring functions
- Rescoring
  - Use one scoring function during the docking, but evaluate the final poses using another scoring function
  - Rationale: One scoring function is better at pose prediction, the other is better at ranking actives versus inactives
- Consensus scoring
  - Combine the results from several rescorings

Outline

- Introduction to protein-ligand docking
- Searching for poses
- Scoring functions
- Assessing performance
- Practical aspects

Pose prediction accuracy

- Given a set of actives with known crystal poses, can they be docked accurately?
- Accuracy measured by RMSD (root mean squared deviation) compared to known crystal structures
  - RMSD = square root of the average of (the difference between a particular coordinate in the crystal and that coordinate in the pose)^2
  - Within 2.0Å RMSD considered cut-off for accuracy
  - More sophisticated measures have been proposed, but are not widely adopted
- In general, the best docking software predicts the correct pose about 70% of the time
- Note: it's always easier to find the correct pose when docking back into the active's own crystal structure
  - More difficult to cross-dock

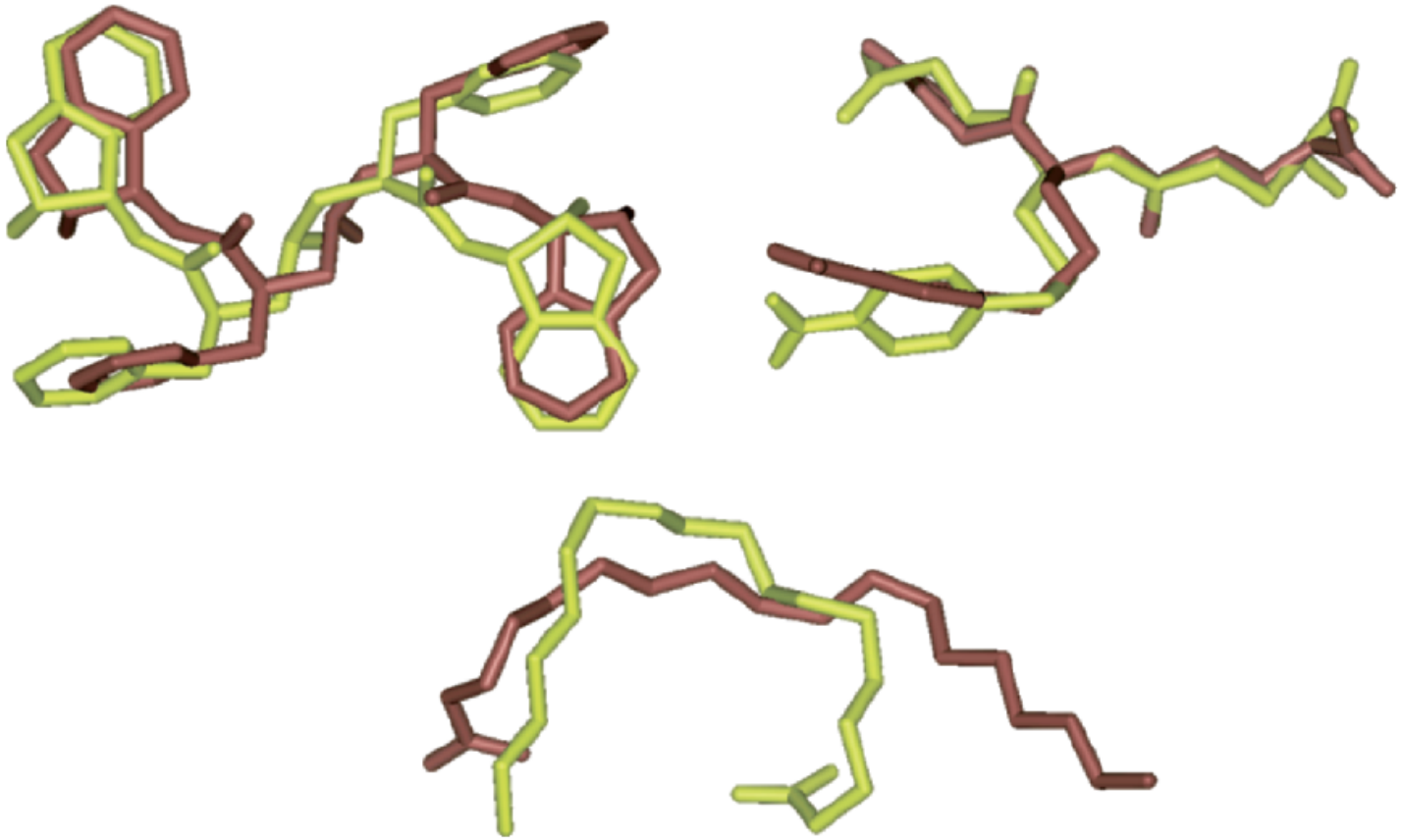


Figure 8-4. Illustration of the range of results produced by a typical docking program. Here we show the results obtained by running the GOLD program [Jones et al. 1995b] on three ligands from the PDB. In each case the x-ray conformation is shown in dark grey and the top-ranked docking result in light grey. The PDB codes for the three ligands are (clockwise, from top left) 4PHV (a peptide-like ligand in HIV Protease), 1GLQ (a nitrophenyl-substituted peptide in glutathione S transferease) and 1CIN (oleate in fatty acid binding protein). These dockings were classified as "good", "close" and "wrong" [Jones et al. 1997].

Assess performance of a virtual screen

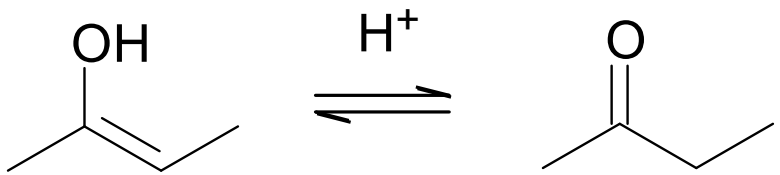
- Need a dataset of N_{act} known actives, and inactives
- Dock all molecules, and rank each by score
- Ideally, all actives would be at the top of the list
  - In practice, we are interested in any improvement over what is expected by chance
- Define enrichment, E, as the number of actives found (N_{found}) in the top X% of scores (typically 1% or 5%), compared to how many expected by chance
  - $E = N_{found} / (N_{act} * X/100)$
  - E > 1 implies "positive enrichment", better than random
  - E < 1 implies "negative enrichment", worse than random
- Why use a cut-off instead of looking at the mean rank of the actives?
  - Typically, the researchers might test only have the resources to experimentally test the top 1% or 5% of compounds
- More sophisticated approaches have been developed (e.g. BEDROC) but enrichment is still widely used

Outline

- Introduction to protein-ligand docking
- Searching for poses
- Scoring functions
- Assessing performance
- Practical aspects

Ligand Preparation

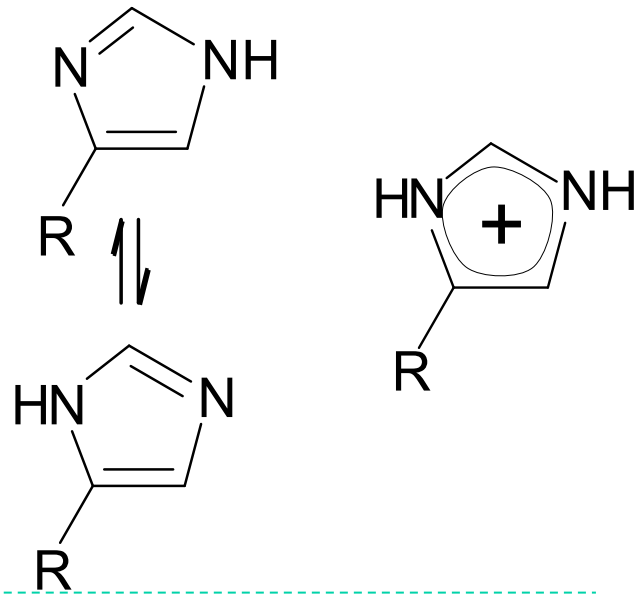
- A reasonable 3D structure is required as starting point
  - Even during flexible docking, bond lengths and angles are held fixed
- The protonation state and tautomeric form of a particular ligand could influence its hydrogen bonding ability
  - Either protonate as expected for physiological pH and use a single tautomer
  - Or generate and dock all possible protonation states and tautomers, and retain the one with the highest score



Enol Ketone

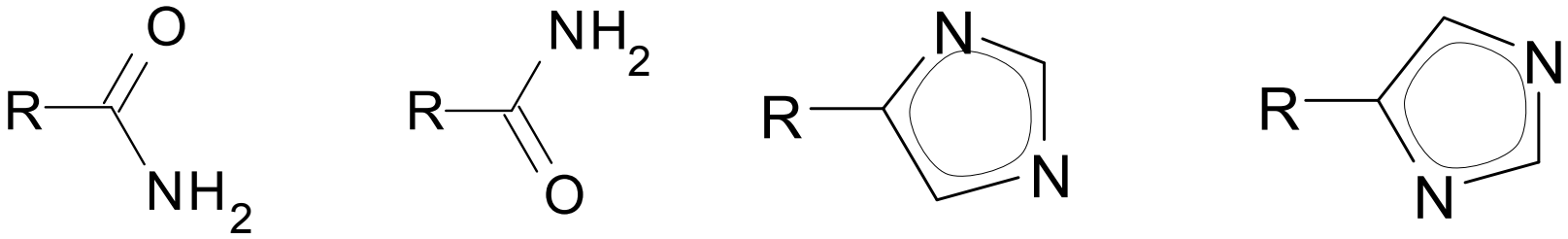
Preparing the protein structure

- The **Protein Data Bank** (PDB) is a repository of protein crystal structures, often in complexes with inhibitors
- PDB structures often contain water molecules
  - In general, all water molecules are removed except where it is known that they play an important role in coordinating to the ligand
- PDB structures are missing all hydrogen atoms
  - Many docking programs require the protein to have explicit hydrogens. In general these can be added unambiguously, except in the case of acidic/basic side chains
- An incorrect assignment of protonation states in the active site will give poor results
- Glutamate, Aspartate have COO- or COOH
  - OH is hydrogen bond donor, O- is not
- Histidine is a base and its neutral form has two tautomers



Preparing the protein structure

- For particular protein side chains, the PDB structure can be incorrect
- Crystallography gives electron density, not molecular structure
  - In poorly resolved crystal structures of proteins, isoelectronic groups can give make it difficult to deduce the correct structure



- Affects asparagine, glutamine, histidine
- Important? Affects hydrogen bonding pattern
- May need to flip amide or imidazole
  - How to decide? Look at hydrogen bonding pattern in crystal structures containing ligands

Final thoughts

- Protein-ligand docking is an essential tool for computational drug design
  - Widely used in pharmaceutical companies
  - Many success stories (see Kolb et al. *Curr. Opin. Biotech.*, **2009**, *20*, 429)
- But it's not a golden bullet
  - The perfect scoring function has yet to be found
  - The performance varies from target to target, and scoring function to scoring function
- Care needs to be taken when preparing both the protein and the ligands
- The more information you have (and use!), the better your chances
  - Targeted library, docking constraints, filtering poses, seeding with known actives, comparing with known crystal poses

Questions?

Protein-Ligand Docking

Dr. Noel O'Boyle
University College Cork
n.oboyle@ucc.ie

Jan 2010
EBI Hands-on Training
Small Molecule Bioactivity Resources at the EBI