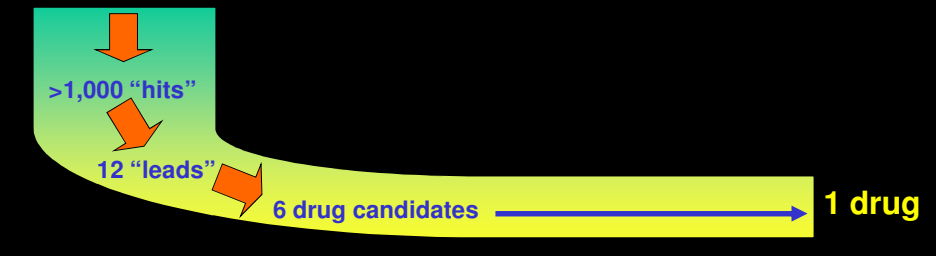
Overview

- Drug Discovery & Development Process
- Computational (Informatics) Approaches
- Chem-informatics
  - QSAR
  - Virtual Combinational Chemistry
  - Molecular Diversity Analysis
  - Virtual Screening

Drug Discovery Process

Time and Money

50,000 - 5,000,000 compounds are often screened to find a single drug



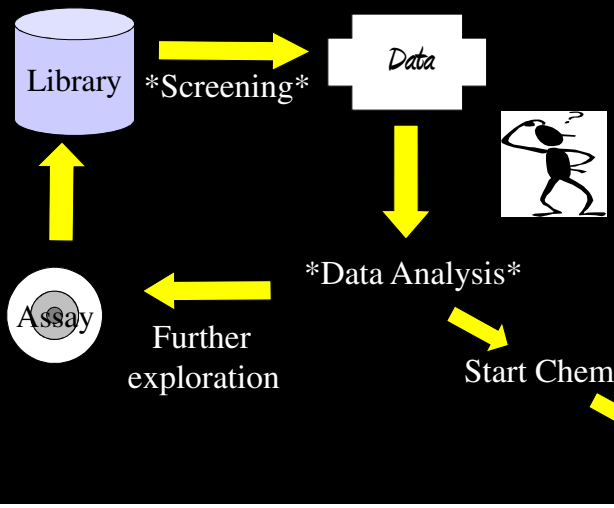
Discovery & Preclinical trials

Clinical trials: Phase I, Phase II, Phase III

8 to 14 years

$750 to >$900 million

Traditional Drug Discovery Process



The Drug Discovery & Development Process

pharmaceutical companies are in the business of identifying compounds that may be useful new drugs

- **tens or hundreds of thousands of compounds are made and tested every year ("screening")**
  - **tests are usually simple binding assays (*does the molecule bind to a target protein?*)**
- **testing is done in two stages**
  - **Lead Generation (find a compound that binds)**
  - **Lead Optimization (find a compound that binds better)**
- **chemical similarity is important at both these stages**

Lead Generation

- when testing a large number of compounds to identify a new "lead", it is obviously desirable to have them as different from each other as possible
  - pharmaceutical companies purchase large numbers of compounds from 3rd party suppliers (often Eastern European) to test
  - they also synthesize combinatorial "libraries" of compounds
- chemical "diversity" is an important feature of such compound collections and libraries
  - the idea is to cover as much of "chemical space" as possible

Standard Drug Lead Approach

- Identify target (e.g., enzyme, receptor, ion channel, transporter)
- Determine DNA and protein sequence
- Elucidate structure and function of protein
- Prove therapeutic concept in animals ("knock-outs")
- Develop assay for high-throughput molecular screen
- Mass screening and/or directed synthesis program
- Select one or more lead structures

Drug Targets and Mechanisms of Drug Action

- Enzymes – inhibitors (reversible, irreversible)
- Receptors – agonists and antagonists
- Ion Channels – blockers
- Transporters – uptake inhibitors
- DNA – intercalating agents, minor groove binders, antisense drugs

Lead Optimization

- when a "lead" has been identified, the next stage is to find compounds that are similar to it, which might bind even better
  - this can involve similarity searching to find compounds previously made, or available commercially for purchase
- in later stages, medicinal chemists make specific changes to the lead compound which they hope will improve its binding affinity

Drug Development

- Patents applications are filed as soon as a good compound (or class of compounds) is identified
  - need to get in before the competition
  - patent life (20 years) starts counting down from here
- Much development work remains to be done
  - animal tests
  - clinical trials (several phases)
  - regulatory requirements
  - many drugs may "fail" during this process
- Patents may have only 10 years left to run by the time a new drug is marketed

Failure of Compounds in Development

- Poor biopharmaceutical properties, 39%
- Lack of efficacy, 29%
- Toxicity, 21%
- Market reasons, 6%

Drug-Like Properties of Candidate Compound Should Be Considered Early

To avoid serious ADME liabilities as early as possible in the drug discovery process

- **Empirical rules**
  - **Lipinski rules of 5 (MW, cLogP, #HD, #HA)**
- **Drug-likeness**
  - **Ajay & Murcko (J Med Chem, 1998, 41, 3314-3324)**
  - **Sadowski & Kubinyi (J Med Chem, 1998 , 41, 3325-3329)**

ADME: Absorption, Distribution, Metabolism, Excretion

DRUG-LIKE BEHAVIOR

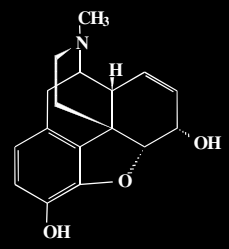
The Lipinski "Rule of Five" [1]

- **Molecular Weight ≤ 500 (opt = ~350)**
- **# Hydrogen Bond Acceptors ≤10 (opt = ~5)**
- **# Hydrogen Bond Donors ≤ 5 (opt = ~2)**
- **-2 < cLog P < 5 (opt = ~3.0)**
- **# Rotatable Bonds ≤ 5**

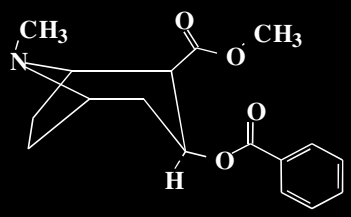
1: C. Lipinski et al, Adv. Drug. Del. Rev, 23, 3-25 (1997)

Bioavailability and LogP

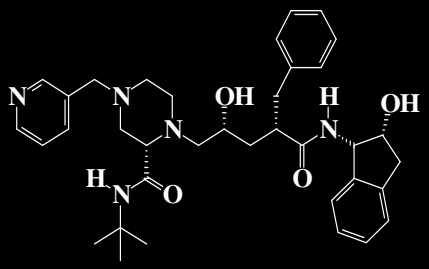
- Log of 1-Octanol/water partition coefficient ($\log K_{OW}$)
- Estimates bioavailability of molecules
  - hydrophobic/hydrophilic balance
  - hydrophilic enough to be soluble in water (blood)
  - hydrophobic enough to transport across cell membranes
- Typical values range from -3 (very hydrophilic) to +7 (very hydrophobic)
- Most drug-like molecules have LogP values in 2-4 range



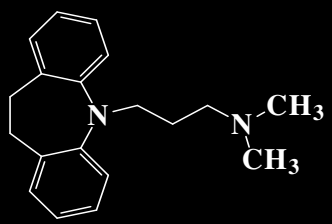
Morphine (clog P = 0.24)



Cocaine (clog P = 2.72)



Indinavir (clog P = 2.78)



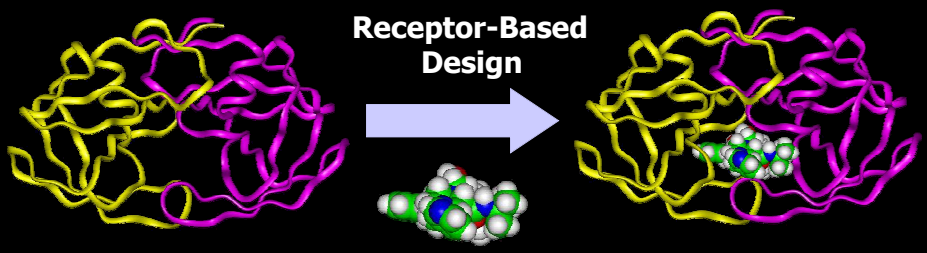
Imipramine (clog P = 4.49)

Properties Related to Good Pharmacokinetics (PK)

- Aqueous solubility
- Membrane passive permeability
- Cytochrome P450 activities (metabolism)
- Plasma protein binding
- Efflux pumping and active transport

Basic Modeling Strategies

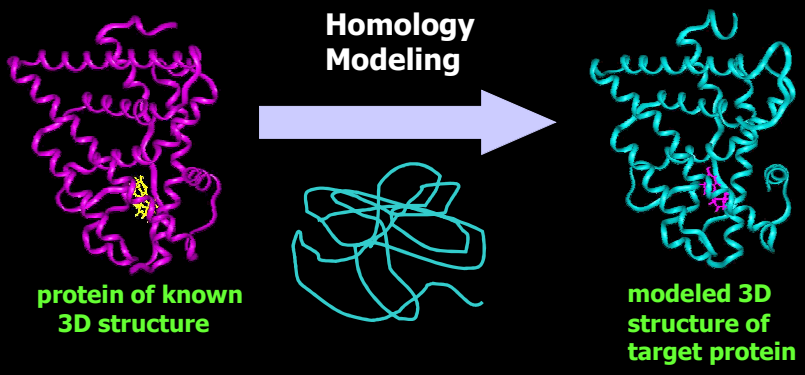
If you **DO** know the 3D structure of the target receptor



Build or Find the key that fits the lock

Basic Modeling Strategies

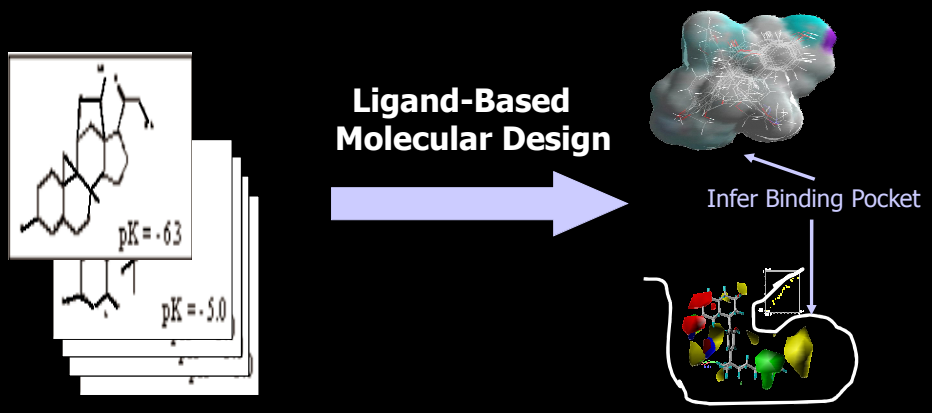
If you DON'T know the 3D structure of the target receptor



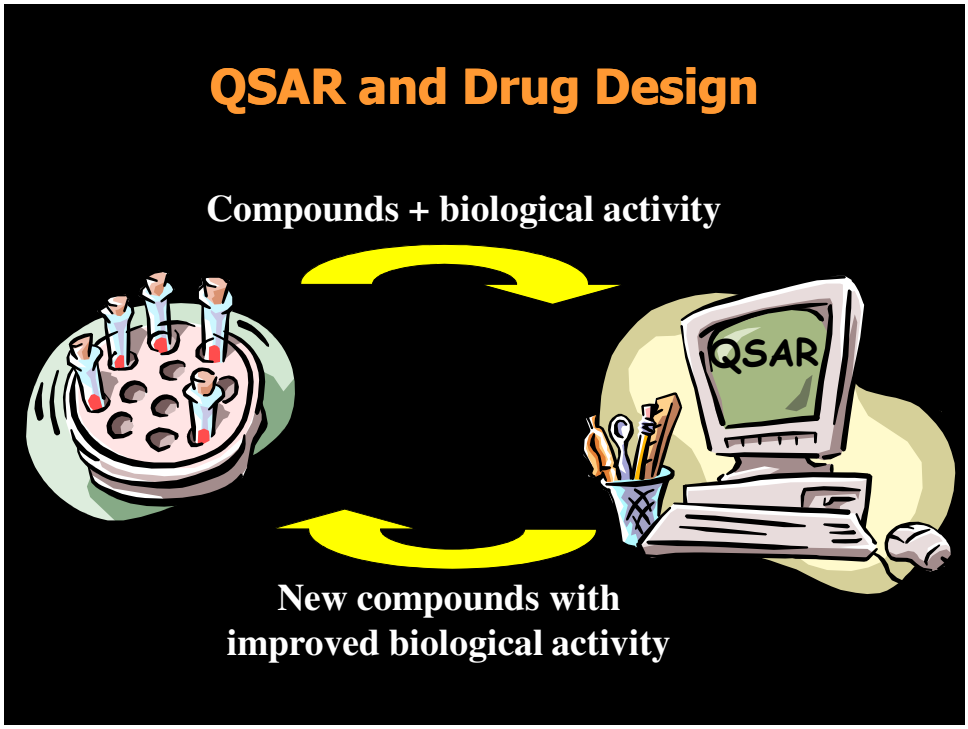
Build the lock, then find the key

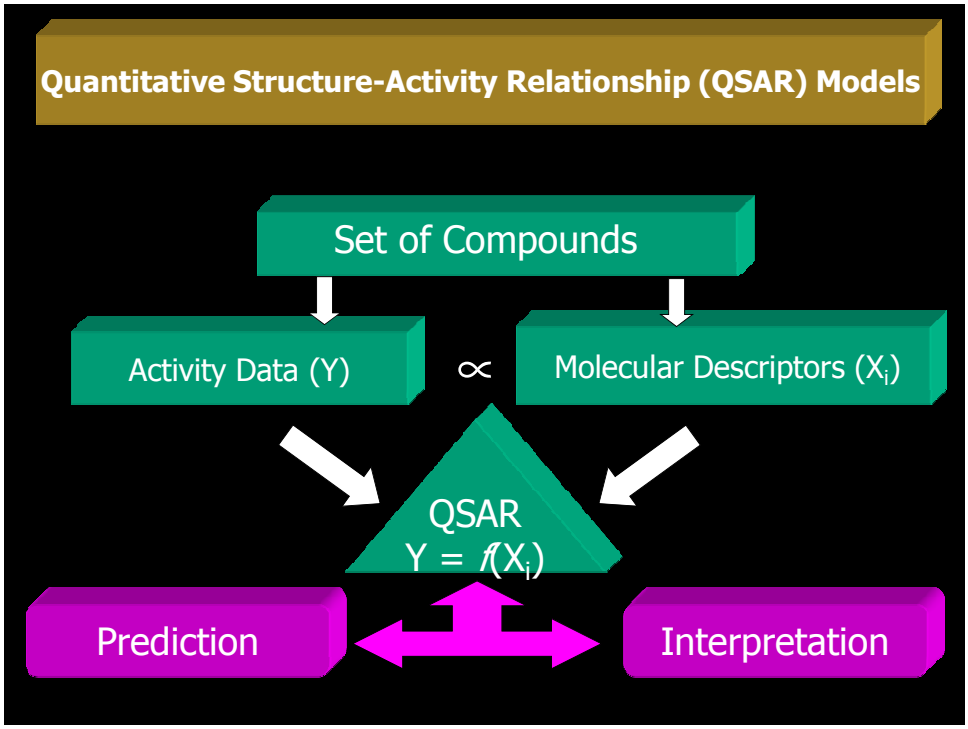
Basic Modeling Strategies

If you DON'T know the 3D structure of the target receptor



Infer the lock by inspecting the keys

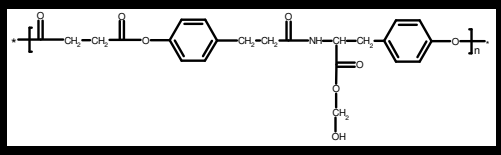

QSAR and Drug Design
Compounds + biological activity
QSAR
New compounds with
improved biological activity


Quantitative Structure-Activity Relationship (QSAR) Models
Set of Compounds
Activity Data (Y)
∝
Molecular Descriptors (Xi)
QSAR
Y = f(Xi)
Prediction
Interpretation

Types of Molecular Descriptors

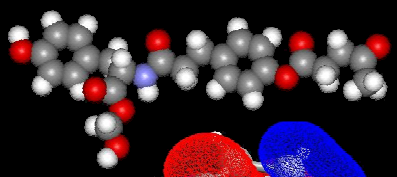
Constitutional, Topological

2-D structural formula

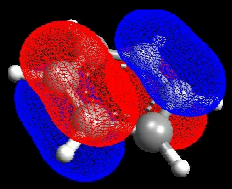


Geometrical

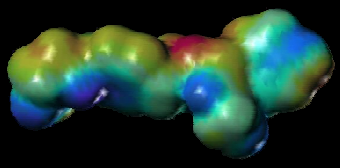
3-D shape and structure



Quantum Chemical

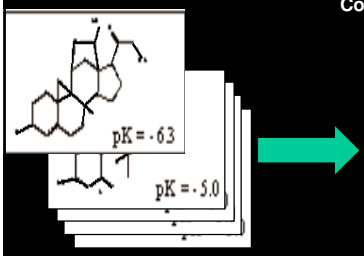


Electrostatic



Thermodynamic

Quantitative Structure-Activity Relationship (QSAR) Models



Extract and Tabulate Descriptors

Compound	Activity (pK_i) "Y"	Descriptors (X_j)		
		Mol. Vol. (Å^3)	LogP	Dipole Mom (μ)
1	2.34	420	2.8	0.97
2	1.89	332	4.6	2.23
3	0.23	198	-0.3	3.36
4	3.67	467	3.7	0.45
5	2.55	359	-1.5	1.77
etc.	etc.	etc.	etc.	etc.

Building QSAR Models

Δ(obs. property or activity) $\propto$ Δ(molecular descriptors)

$Y = f(X_i)$

Simple (Univariate) Linear Regression Hammett, 1939

$pK_i = a_o + a_1 (\text{Mol Vol}_i)$

Multiple Linear Regression (MLR) Hansch, 1969

dependent variable *independent variable* *"best-fit" constants*

$pK_i = a_o + a_1 (\text{Mol Vol}_i) + a_2 (\text{logP}) + a_3 (\mu_i) + \ldots$

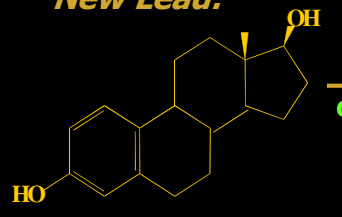
Partial Least-Squares (PLS) Regression Wold, et al. 1984

$pK_i = a_o + a_1 (\text{PC1}) + a_2 (\text{PC2}) + a_3 (\text{PC3}) + \ldots$

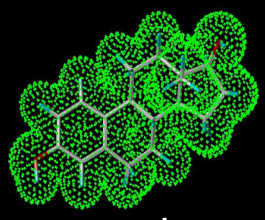
Predicting Activities of Untested Compounds

Using QSAR models as a predictive tool

New Lead:



extract descriptors



V logP μ

Validated QSAR model: $pK_i = 0.52 (V_i) + 0.27 (\text{logP}_i) - 0.38 (\mu_i)$

Predicted activity of new lead

Role of Screening Chemical Libraries in Drug Discovery

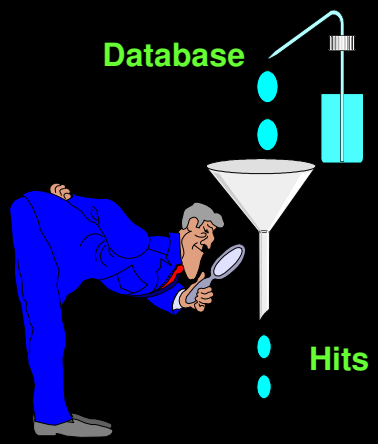
- Lead discovery libraries: high chemical diversity
  - exploiting real and virtual compound databases
- Lead optimization library: high similarity
  - exploiting high throughput combinatorial chemistry
- Drug like character more important than synthetic accessibility
  - importance of medicinal chemistry and the application of known filters

Finding New Lead Compounds

Mining Structural Databases

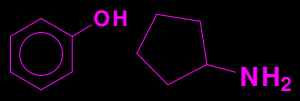
- Maybridge Database ~ 118,000 chemicals
- NCI Database ~ 60,000 chemicals
- ACD Database ~ 230,000 chemicals
- WDI Database ~ 100,000 chemicals

… but how do you find new "leads"?

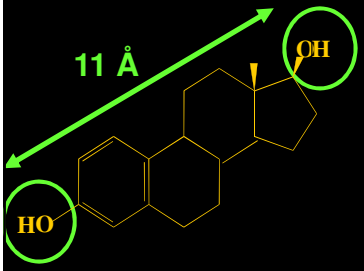


Screening Structural Databases

1. (Sub)structure Searching



2. Pharmacophore Matching



3. Optimal Properties

(e.g., Vol, SA, μ, ... hundreds more)

4. Lipinski's "Rule of 5"

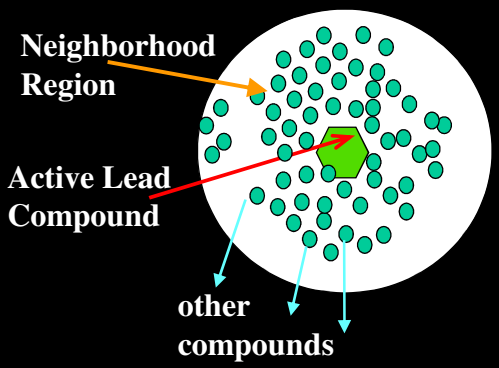
(C. A. Lipinski, et al., Advanced Drug Delivery Reviews (1997), 23, 3-25)

Drug-like molecules share these features:

1) Maximum of 5 H-bond donors (OH, NH)
2) Maximum of 10 H-bond acceptors (O:, N:)
3) Molecular Weight < 500
4) LogP < 5

5. Apply QSAR Models

Structural Similarity



The property of a Compound is shared by *most* other compounds within its Neighborhood Region

i.e. neighbors of an active compound have a higher probability of behaving in a 'similar' way

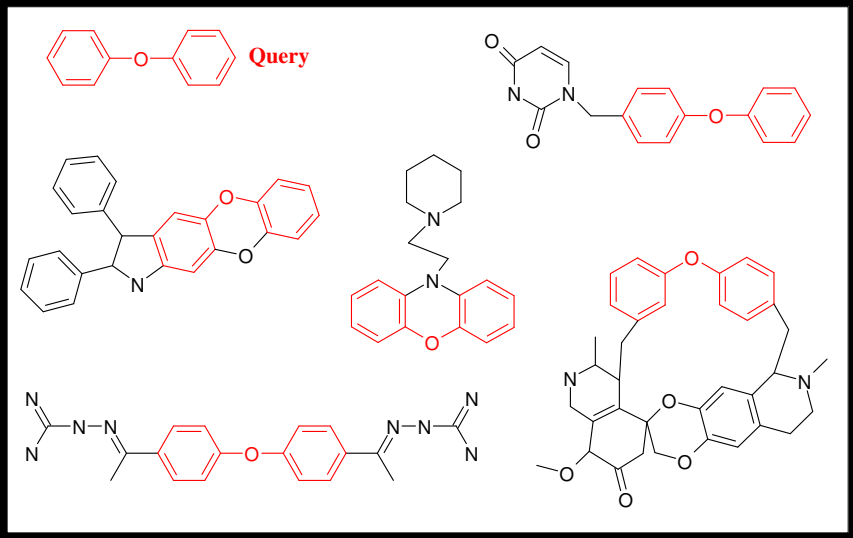
Virtual Screening

- Need to prioritize the many molecules that *could* be tested
- Increasingly sophisticated level of filtering to maximize the numbers of potential leads
  - "Drugability" considerations
  - Similarity searching (both 2D and 3D) using initial weak leads
  - 3D substructure searching once possible pharmacophoric patterns have been identified
  - Docking once the 3D structure of the biological target is available

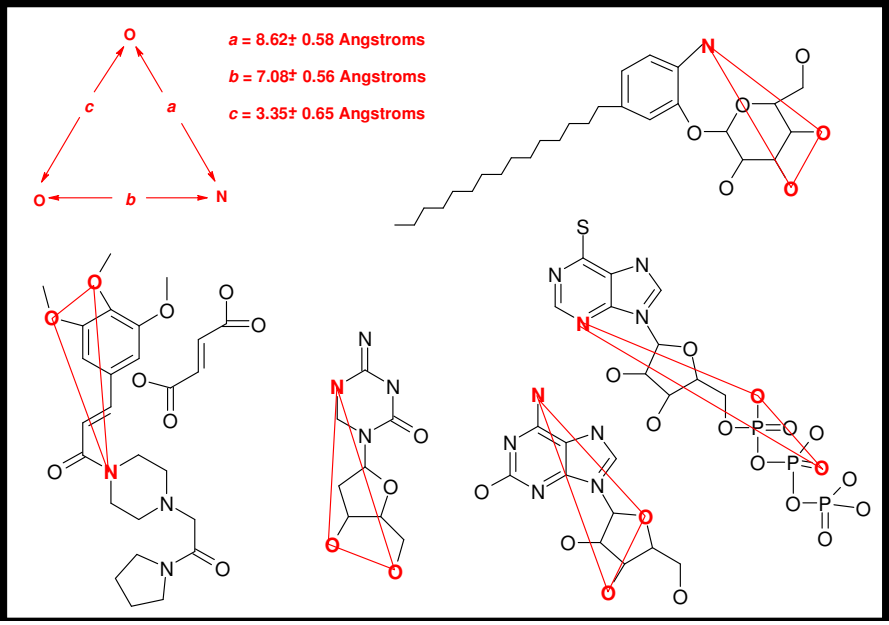
Structure Searches

- 2D Substructure searches
- 3D Substructure searches
  - single conformation
  - multiple conformation (flexible)

2D Substructure Searching

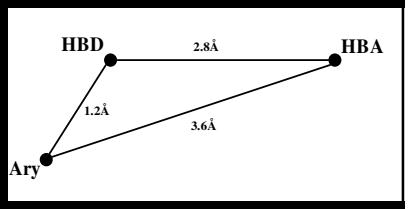


3D Substructure Searching



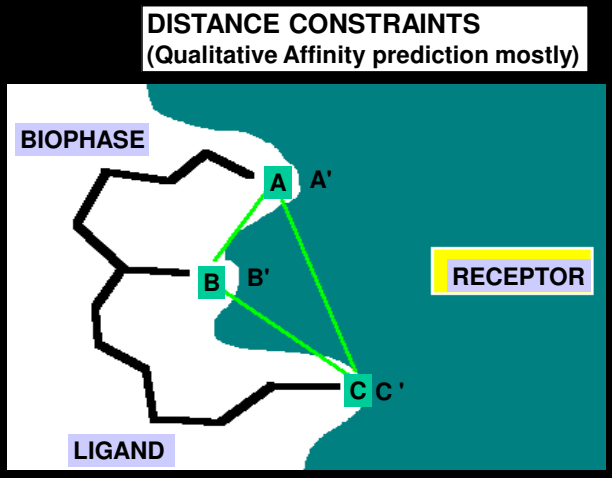
3D Fragments

- each fragment consists of 3 pharmacophoric points
  - the distances between each pair of these points are binned and used to set fingerprint bits



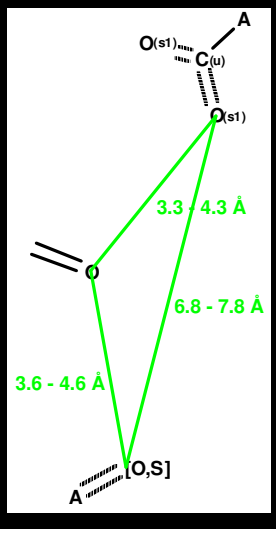
- 4-point pharmacophore fragments are also used
- Different people have used slightly different definitions of pharmacophoric points

Suppose a drug needs to satisfy the distances below to bind and activate a receptor – *how can we search a large database for 'similar' molecules ?*



3D Substructure Searches

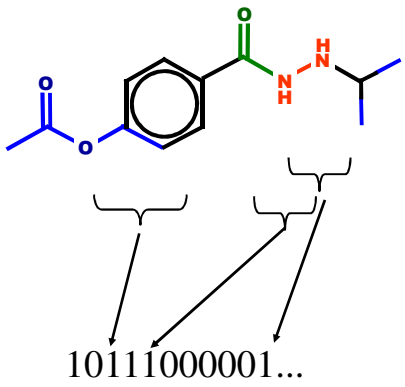
- Spatial Relationships
- Define ranges for distances and angles
- Stored conformation
  - usually lowest energy



Numerical Similarity Measures

- Calculate some numerical measure of similarity between molecules
- Query structure is a "target" molecule
- Database structures can be ranked in decreasing order of similarity to target
  - find all molecules within threshold similarity to target
  - find N most similar molecules to target

Substructure Keys

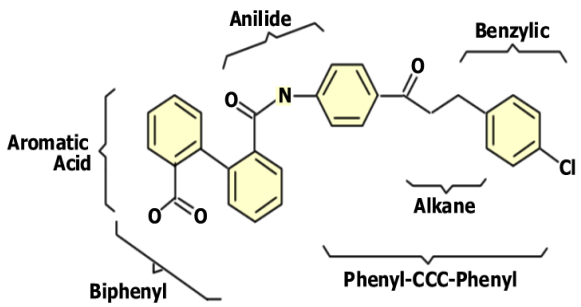


Dictionary of Keys

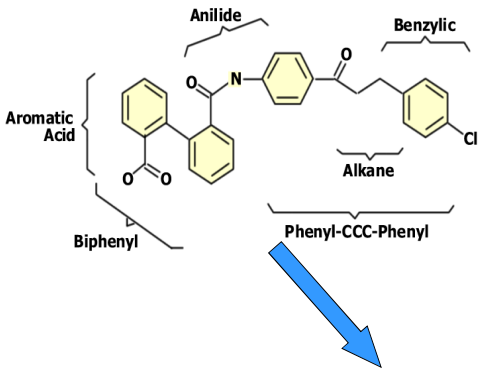
- N-N
- O-C(-N)-C
- CH_3-Ar-CH_3
- C-N-N
- N-Ar-Ar-O
- N-C-O
- N-Ar-O
- OH > 1
- CH_3 > 1
- N > 1
- NH
- ...

Substructural Keys

- **Compounds are multi-domain:**
  - **multiple occurrences of a key/substructure**
  - **y**



"Bit Strings" of Substructure Keys



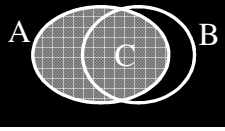
"How" a key hits?

Similarity from Fingerprints

- similarity measures are most commonly calculated from structure fingerprints
  - count the bits that are "on" in both molecules ("C")
  - count the bits that are "on" in each molecule separately

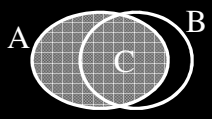
struct A:	00010100010101000101010011110100	13 bits on (A)
struct B:	00000000100101001001000011100000	8 bits on (B)
A & B:	00000000000101000001000011100000	6 bits on (C)

  - similarity coefficient can be calculated from A, B and C



Tanimoto Coefficient

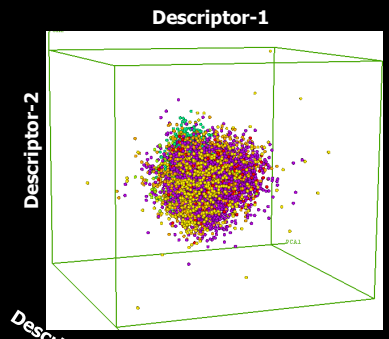
- similarity = $\frac{C}{A + B - C}$



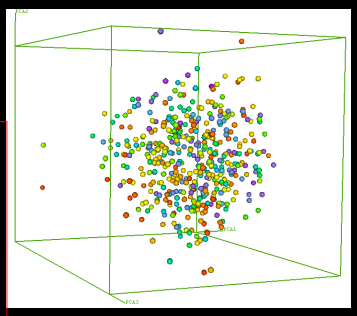
$= 6 / (13 + 8 - 6) = 0.4$

- the number of bits set in both molecules divided by the number of bits set in either molecule
- The Tanimoto Coefficient is the most commonly used similarity coefficient in chemical informatics
  - also called the Jaccard coefficient

Chemical Space

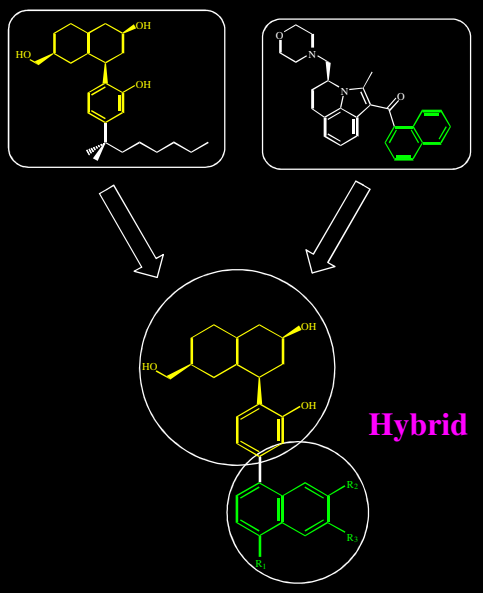


- Supplier A: purple
- Supplier B: yellow
- Supplier C: green
- Supplier D: red

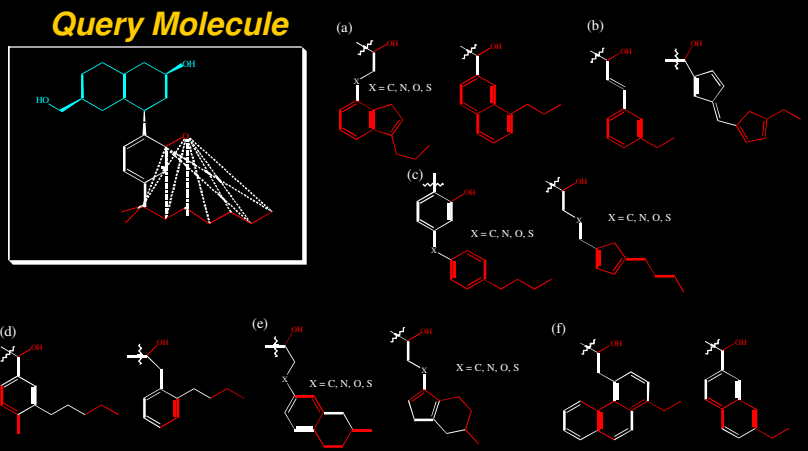


FOCUSING OF LIBRARY
1. Drug-Like
2. Structural Diversity

"Mix-and-Match" Approach to Design

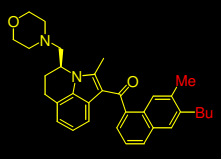


Searching Structural Databases for Lead Compounds

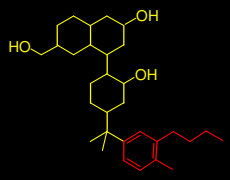


Design Strategies: Novel Cannabimimetics

(A) Mix-and-Match

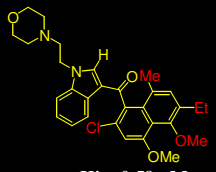


pKi = 0.50 nM

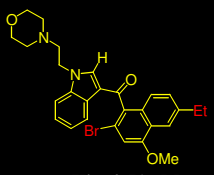


pKi = 1.16 nM

(B) Combinatorial

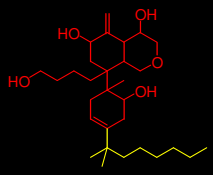


pKi = 0.59 nM



pKi = 0.51 nM

(C) Database Searching



pKi = 2.01 nM

cf. WIN-55212, pKi = -0.04 nM
CP-55244, pKi = 0.96 nM