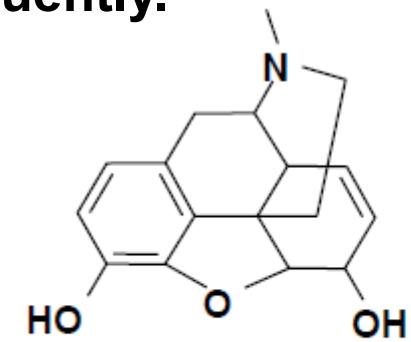
Similarity Search

Uwe Koch

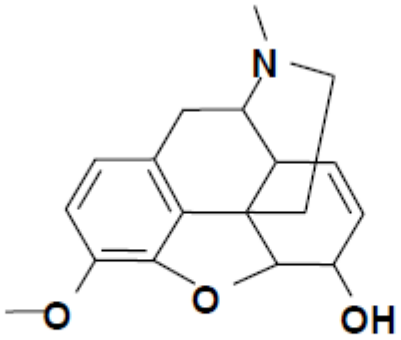
Lead
Discovery
Center
LDC

Similarity Search

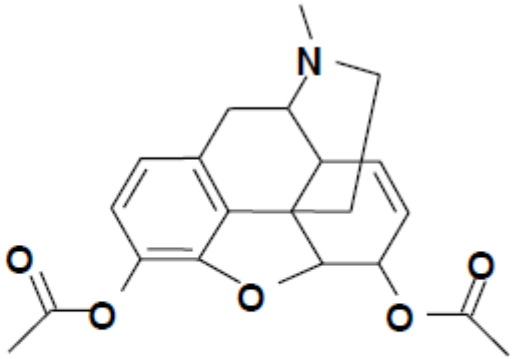
The similar property principle: strurally similar molecules tend to have similar properties. However, structure property discontinuities occur frequently.



Morphine



Codeine



Heroin

Relevance in medicinal chemistry:

- ligand based virtual screening
- identify new series to avoid liabilities in back up
- Identify dissimilar compounds for screening collection

Similarity Search

Lead Discovery Center LDC

Similarity search – probably, together with substructure searches, the cheminformatic method most used by chemists

All similarity measures comprise three basic components:

- the *representation* that characterizes each molecule, the molecular descriptors
- the *weighting scheme* that is used to (de)prioritise different parts of the representation to reflect their relative importance
- the *similarity coefficient* that provides a numeric value for the degree of similarity between two weighted representations

Similarity Search:Descriptors

Representation of a molecule – molecular descriptors: numerical values describing the properties of a molecule

Descriptors representing properties of complete molecules:

- log P, dipole moment, polarizability

Descriptors calculated from 2D graphs:

- topological indices, 2D fingerprints, maximum common substructures

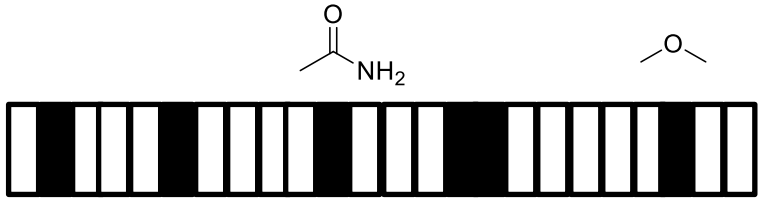
Descriptors requiring 3D representations:

- Pharmacophore descriptors, fingerprints, molecular fields, shape

Todeschini & Consonni, Handbook of molecular descroptors, Wiley-VCH 2009

Similarity Search:Descriptors

2D fingerprints: binary vectors



MACCS: Each bit in the bit string represents one molecular fragment, where 1 indicates the presence of the functional group, 0 ist absence. 166 strcutural fragments.

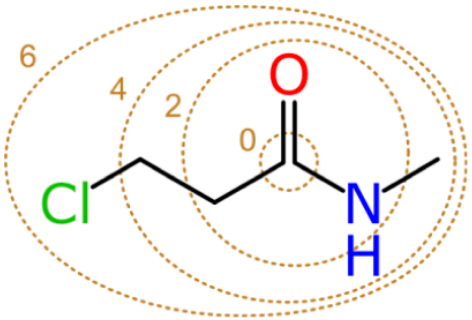
Other types of fingerprint: path (daylight), circular (EFCP, Morgan)

Similarity is quantified by determining the number of common bits

Todeschini & Consonni, Handbook of molecular descroptors, Wiley-VCH 2009

Similarity Search:Descriptors

ECFP2, ECFP4, FCPF2 …: encoding circular substructure



Encodes information on the arrangement of heavy atoms around each central atom.

Layer 0: Carbonyl C (sp2)

Layer 1: Aliphatic carbon (sp3)

Oxygen (sp2)

Nitrogen (sp2)

ECFP fragments encode atomic type, charge and mass

FCFP fragments encode six generalized atom types

2, 4, 6 denotes the diameter (in bonds) of the circular substructure

Todeschini & Consonni, Handbook of molecular descroptors, Wiley-VCH 2009

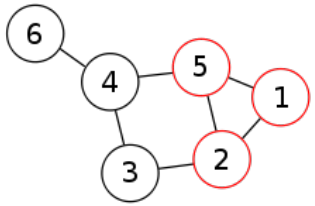
Maximum Common Subgraph similarity

Lead Discovery Center LDC

Pattern (graph) matching.

Graph matching is based on node (atom) and edge (bond) correspondence using the graph theoretic concept of a clique.

Clique is a sub-graph in which every node is connected to every other node



Application: Maximum common substructure

Similarity search

Clustering

Reaction mapping

Molecule Alignment

Similarity Coefficient

The similarity coefficient measures the overlap between descriptors for two compounds.

The Tanimoto coefficient (Tc) is calculated as the ratio between conserved features and the total number of features of each molecule.

The reaches from 1 (total similarity) to 0 no overlap. But neighbourhood behavior depends on both the fingerprint being uses and the similarity coefficient.

a: features compound A

b: features compound B

c: features common to A and B

Tanimoto: $Tc = c/(a+b-c) = c/((a-c)+(b-c)+c)$

Tc is the fraction of features shared by A and B and the total number of features

Similarity Coefficient

a: features compound A

b: features compound B

c: features common to A and B

Tanimoto: $Tc = c/(a+b-c) = c /((a-c)+(b-c)+c)$

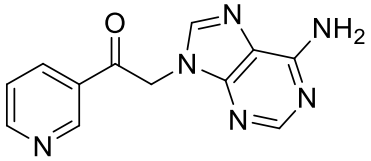
Tc is the fraction of features shared by A and B and the total number of features

Tversky: $Tv = c / (\alpha (a-c) + \beta (b-c) + c)$

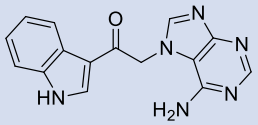
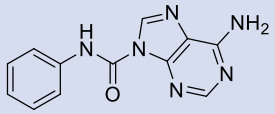
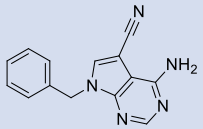
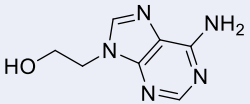
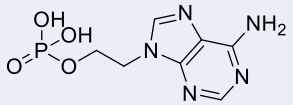
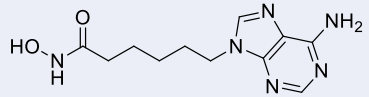
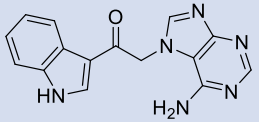
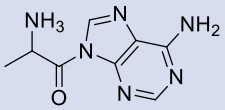
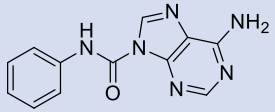
Introduces user defined weighing factors.

If $\alpha = 1$ and $\beta = 0$, $Tv = c / a$: fraction of features it has in common with reference compound. It would be 1 if all features of A are present also in B.

Similarity Search: An example



Reference compound
Search for similars using the same Chembl data set

Descriptor	Highest ranked	2nd	3rd
Fp atom pairs (AP)	 Tanimoto = 0.73 (AP) 0.11 (rad), 0.96 (MACCS)	 Tanimoto = 0.6 (AP) 0.14 (rad), 0.83 (MACCS)	 Tanimoto = 0.55 (AP) 0.12(rad), 0.82 (MACCS)
FP radial (ECFP4)	 Tanimoto = 0.26 (rad) 0.36 (AP), 0.69 (MACCS)	 Tanimoto = 0.24 (rad) 0.29 (AP), 0.58 (MACCS)	 Tanimoto = 0.23 (rad) 0.29 (AP), 0.6 (MACCS)
MACCS	 Tanimoto = 0.96 (MACCS) 0.11 (rad), 0.73 (AP)	 Tanimoto = 0.86 (MACCS) 0.07 (rad), 0.28 (AP)	 Tanimoto = 0.83 (MACCS) 0.14 (rad), 0.6 (AP)

Ranking depends on descriptors used

Similarity Search: An example

Lead Discovery Center LDC

Similarity values are strongly dependent on molecular description.

Many MACCS features are often found in compounds.

ECFP4 encodes often rare molecular environments. Thus Tc(MACCS) is often larger than Tc(ECFP4).

Similarity values also tend to increase with molecular size and complexity due to increasing fingerprint bit density.

Different combinations of fingerprints and and similarity coeffcients produce different similarity value distributions.

Difficult to relate a specific similarity value to a probability of having similar biological activity.

Scaffold hopping

Usually the compounds most similar to a reference are close structural analogs

To identify alternative lead series if problems due to ADME, Tox or IP arise requires identification of structurally different compounds by modifying the core stcucture of the molecule **(Scaffold hopping)**

Descriptors suitable for scaffold hopping:

-Reduced graphs

-Topological pharmacophore keys

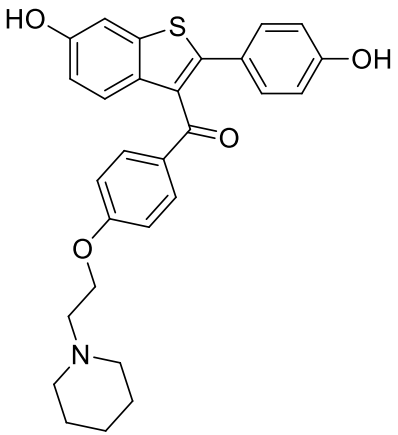
-3D descriptors

Lagdon, Ertl & Brown, Molecular Informatics 2010, 29, 366

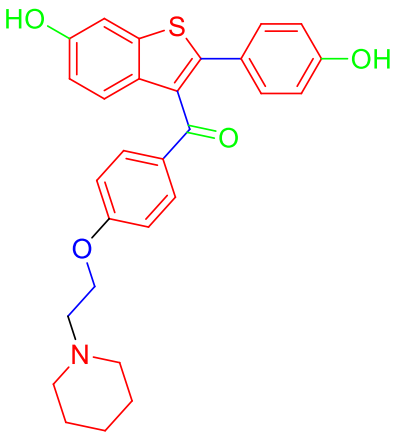
Scaffold hopping

What is a scaffold?

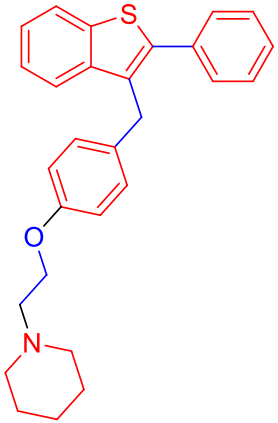
Definition by Murcko & Bemis: The molecule is dissected into ring systems, linkers and side chains. The Murcko framework is the union of ring systems and linkers.



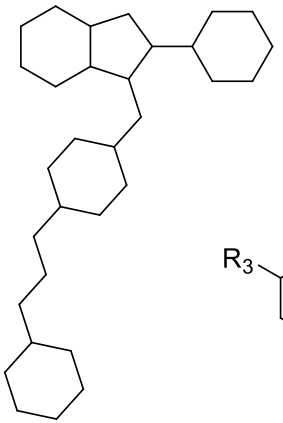
Raloxifen (SERM)
Osteoporosis and
Invasive breast canver



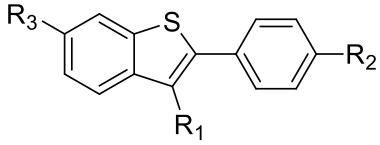
Red:rings
Blue: linkers
Green: side chains



Murcko scaffold



Graph

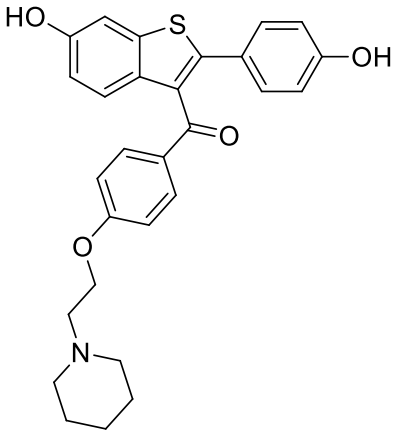


Markush scaffold

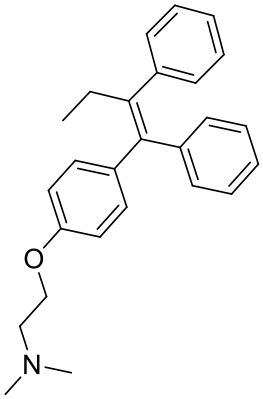
Bemis & Murcko, *J. Med. Chem*. 1996, **39**, 2887.

Scaffold hopping

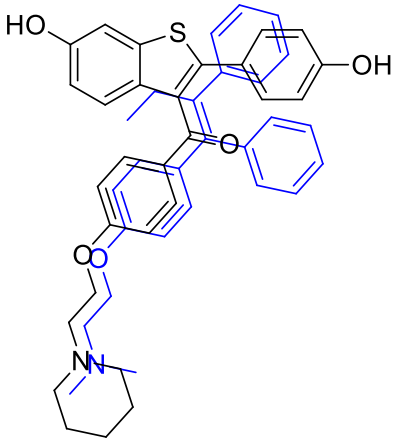
Scaffold hopping example:



Raloxifen (SERM)
Osteoporosis and
Invasive breast cancer



Tamoxifen (SERM)
Invasive breast cancer



2D fingerprints:

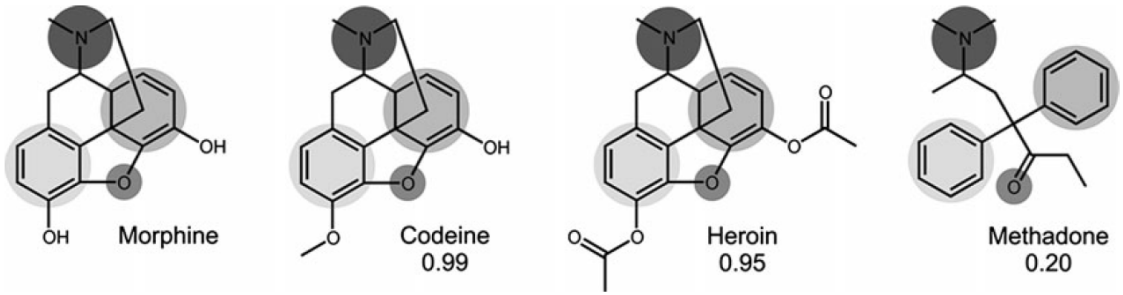
For several cases it has been shown that the top 1% of a screening database select on average 25% of the scaffolds -> iterative focused screening.

However, it is not possible to identify a generally preferred similarity value range.

Bajorath et al., *J. Med. Chem*. 2010, **53**, 5707.

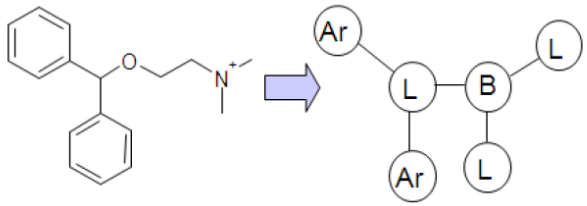
Reduced graph

Reduced graphs provide summary representations of chemical structures by collapsing groups of connected atoms into single nodes while preserving the topology of the original structures.



Shaded spheres map common functional groups revealing similarities notevident when using coventional 2D fingerprints

Similarities based on daylight FPs and the Tanimoto coefficient



Linkers and feature nodes are used to describe molecules.

Different levels of hierarchy according to a more detailed description of node properties

Gillet et al. JCICS 2003, 43, 338

3D similarity search

Molecules with similar 3D shape and properties could share biological activity, even while their 1D and 2D representations are not similar.

Conformational flexibility to be taken into account.

Five classes of 3D representations of molecules

- Pharmacophore
- Atomic distance
- Gaussian function
- Surface
- Field

Pharmacophore searching

Pharmacophore is an abstract description of molecular features that are necessary for molecular recognition of a ligand by a biological macromolecule.

A pharmacophore model explains how structurally diverse ligands can bind to a common receptor site (Wikipedia).

Pharmacophore generation:

•Requires set of actives

•Molecular features: HB Donors, HB Acceptors, hydrophobic groups …

•Active Ligands aligned such that corresponding features are overlaid

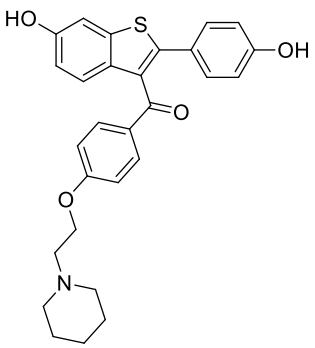
•Conformational space explored

•Scoring of pharmacophore hypothesis taking into account: number of features, goodness of fit, volume of overlay…

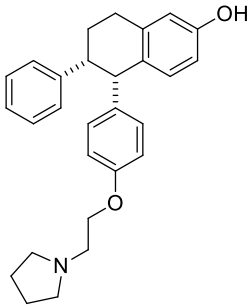
Pharmacophore searching

Lead Discovery Center LDC

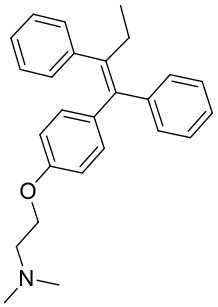
Raloxifene

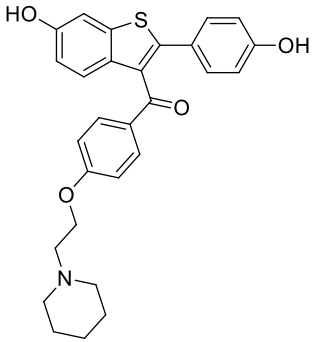


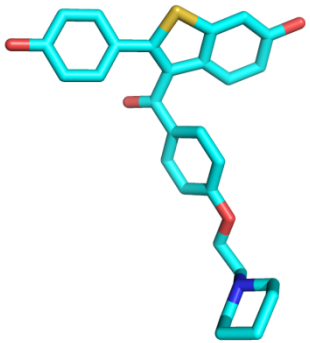
Lasofoxifene

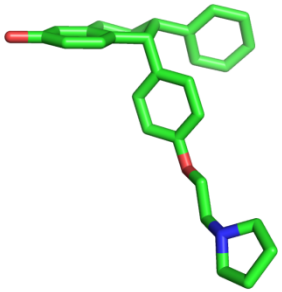


Tamoxifene





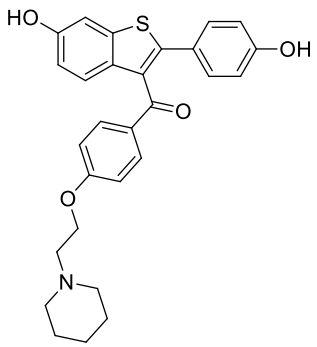




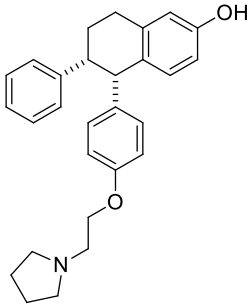
Pharmacophore searching

Lead Discovery Center LDC

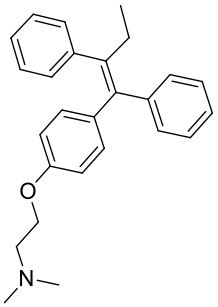
Raloxifene

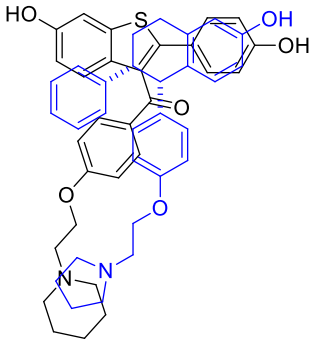


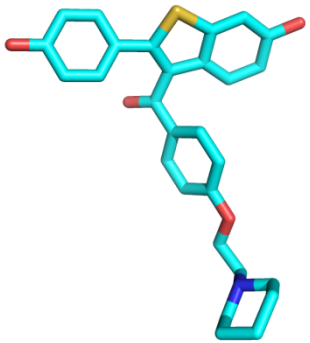
Lasofoxifene

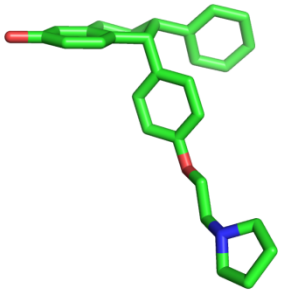


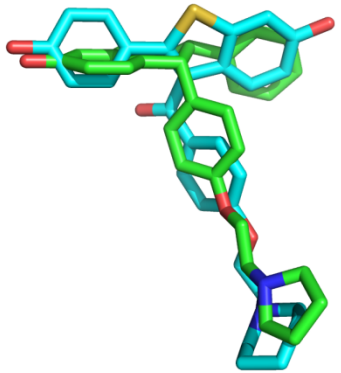
Tamoxifene







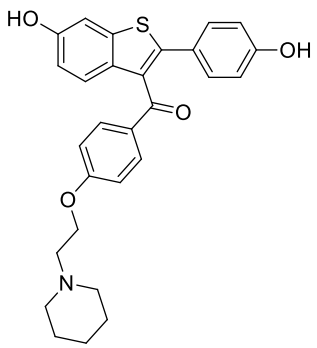




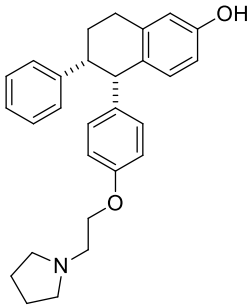
Pharmacophore searching

Lead Discovery Center LDC

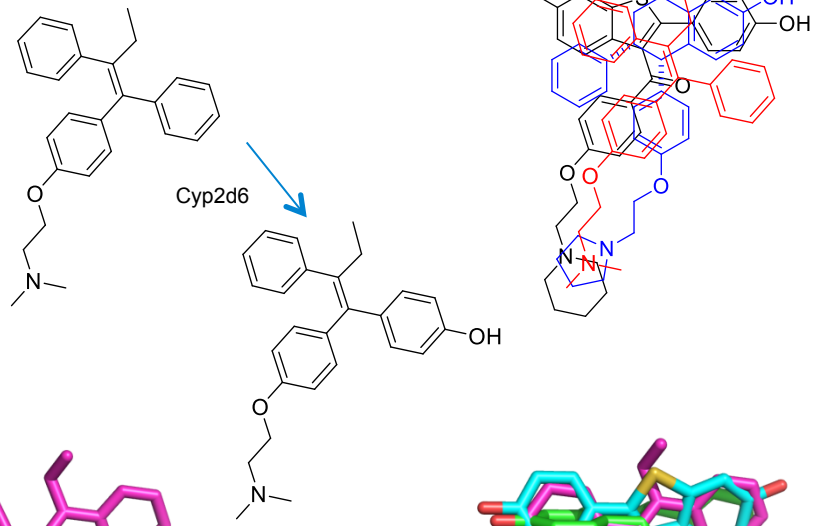
Raloxifene

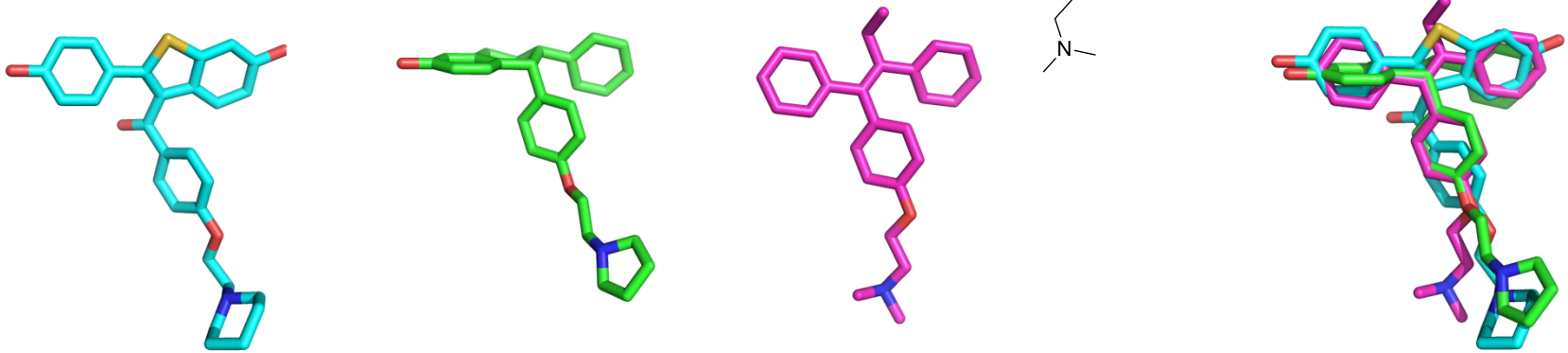


Lasofoxifene



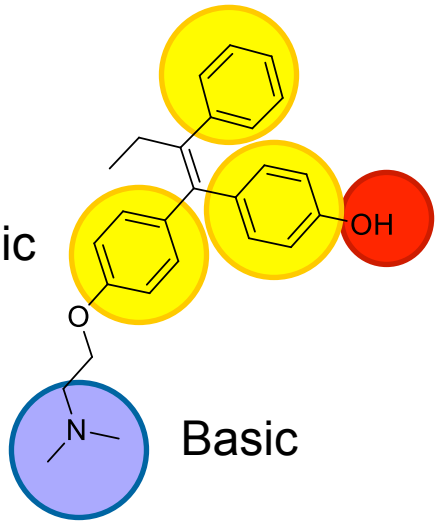
Tamoxifene





Pharmacophore searching

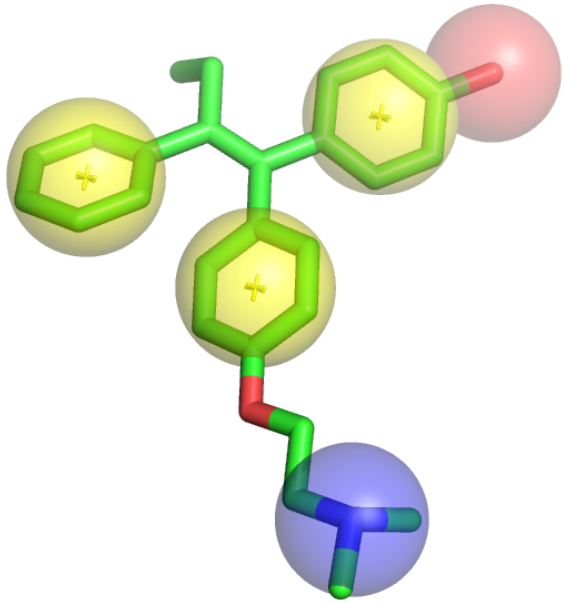
Tamoxifene



Lipophilic

HB-Donor

Basic



Features: Lipohilic, ring, HB donor, basic.

Each feature is represented as a sphere. The radius indicates the toleranceon the deviation from the exact position.

The pharamcophore features are used as queries for searching databases.

3D similarity search

Lead Discovery Center LDC

Atomic distance:

ESHAPE3D uses a distance matrix with distances between all heavy atoms. Eigenvalues are calculated. Difficult to include physicochemical properties.

Gaussian functions:

Similarity as volume overlap between two molecules after superimposition.

ROCS searches for superposition that maximizes Volume overlap. Similarity is quantified by Volume Tamimoto. Several atom types are used and considered for similarity.

Surface similarity:

Surface calculation eg by triangularization. Surface points may be characterized by interaction energies using GRID. Each surface point associated by a vector. Similarity by comparing the vectors.

3D similarity search

Lead Discovery Center LDC

Field based:

Blaze (Cresset): Electrostatic, hydrophobic and shape properties of molecules are represented by field patterns. Field patterns of a reference comcpound are compared to a database of precalculated field patterns of commercial compounds.

Pharmacophore based:

Initiates with identification of pharmacophore including features such as HB-donors and –acceptors, hydrophobicity core and charges. Three of four points are connected to form a triangle of tetrahedron.

Molecular similarity depends on molecular representation, similarity measures and compound class.

It is not always clear how molecular similarity is related to biological activity.

However as a tool to identify groups of compounds to further the knowledge on structure property relationships similarity search is indispensable.

Thank you very much

Questions?