Similarity Search

Canonicalization of Chemical Data

Molecular Descriptor

TABLE 1.2. Different types of molecular descriptors

Descriptor category	Examples
Physical properties	Molecular weight
	logP(o/w)
Atom and bond counts	Number of nitrogen atoms
	Number of aromatic atoms
	Number of rotatable bonds
Pharmacophore features	Number of hydrogen bond acceptors
	Sum of van der Waal surface areas of basic atoms
Charge descriptors	Total positive partial charge
	Dipole moment from partial charges
Connectivity and shape descriptors	Kier and Hall molecular shape indices
Surface area and volume	Solvent-accessible surface area

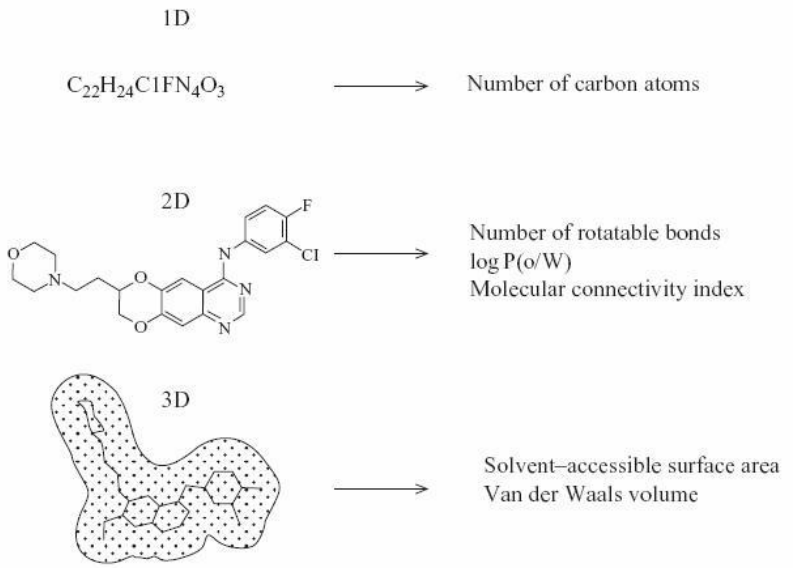


Figure 1.3. Examples of descriptors classified according to dimensionality (adapted from Bajorath 2002)

No generally preferred descriptor spaces - Context Dependent

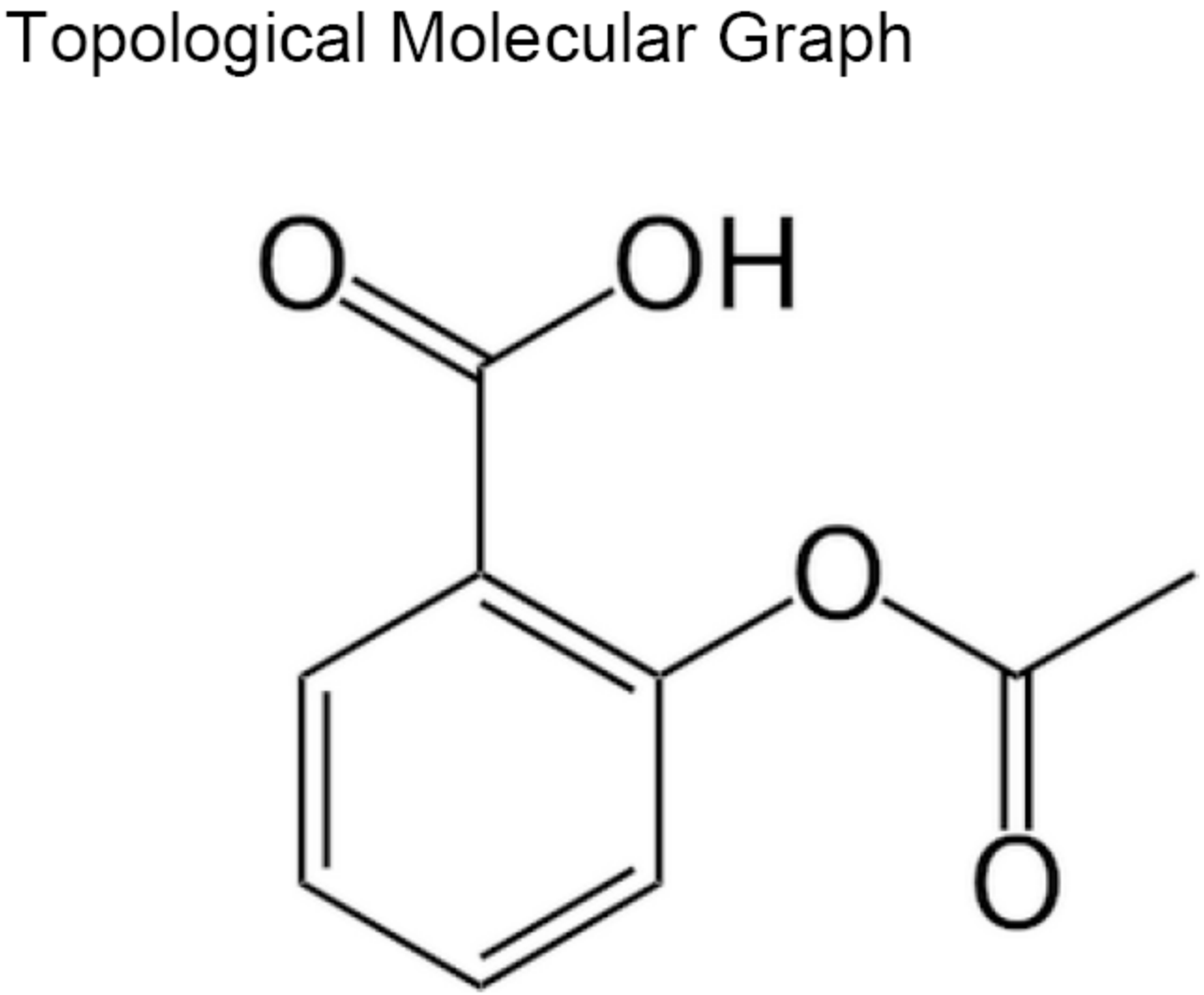
Drug Likeness

TABLE 1.5. Drug-like versus lead like compound characteristics

Drug-like	Lead-like
MW $<$ 500	MW $<$ 350
ClogP $<$ 5	ClogP $<$ 3.0
Hydrogen bond donors $<$ 5	Chemically stable
Hydrogen bond acceptors $<$ 10	
Number of rotatable bonds $\leq$ 10	
PSA $\leq$ 140Å^2	
Peptides not suitable	Non-substrate peptides suitable
Eliminate reactive functional groups, promiscuous inhibitors, and metabolically unstable compounds	

Canonicalisation of Data

Convert the molecular graph to computationally amenable representations

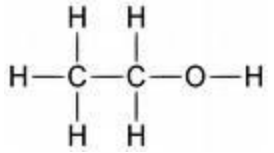


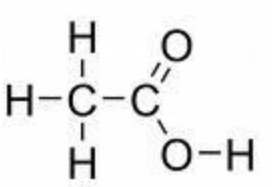
Canonicalization of chemical data……

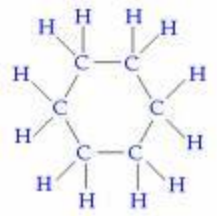
SMILES

Simplified Molecular Input Line Entry System

Ethanol	CCO
Acetic acid	CC(=O)O
Cyclohexane	C1CCCCC1
Pyridine	c1cnccc1
Trans-2-butene	C/C=C/C
L-alanine	N[C@@H](C)C(=O)O
Sodium chloride	[Na+].[Cl-]
Displacement reaction	C=CCBr>>C=CCI







concept of a graph with nodes as atoms and edges as bonds to represent a molecule.

Parentheses are used to indicate branching points and numeric labels designate ring connection points.

David Weininger 1980

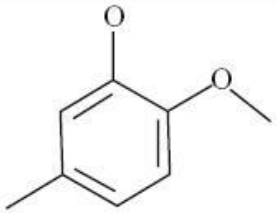
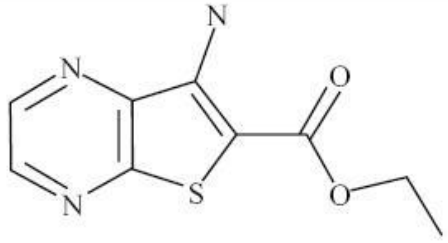
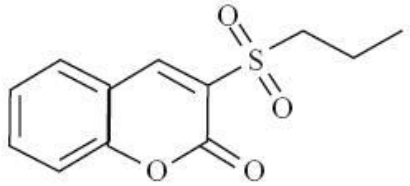
Similarity Searching –Structural queries and graphs

- Contemporary substructure search methods are mostly based on dictionaries of predefined molecular fragments.

- Queries can be transformed into an machine-readable format such as Simplified Molecular Input Line Entry Specification (SMILES) code.

- SMILES encodes 2D representation of molecules as linear strings of alpha-numeric characters.

1D String Representation … SMILES

Structures	Strings
	c1ccccc1
	Oc1cc(C)ccc1OC
	s1c2[nH0]cc[nH0]c2c(N)c1C(=0)OCC
	[S+2]([O–])([O–])(CCC)C1=Cc2ccccc2OC1=O

Scaffolds, Linkers and Sidechains (Functional Groups)

Rings Systems: Cycles within the molecular graphs or rings sharing an edge or vertex in the molecular graph.

Linkers: Edges (bonds) that connect two ring systems

Sidechains: Those atoms that are neither rings or linkers

Frameworks: Ring Systems connected by linkers

Similarity Searching –Structural queries and graphs

- Detection of structural fragments or substructures is a simple but popular form of similarity searching.

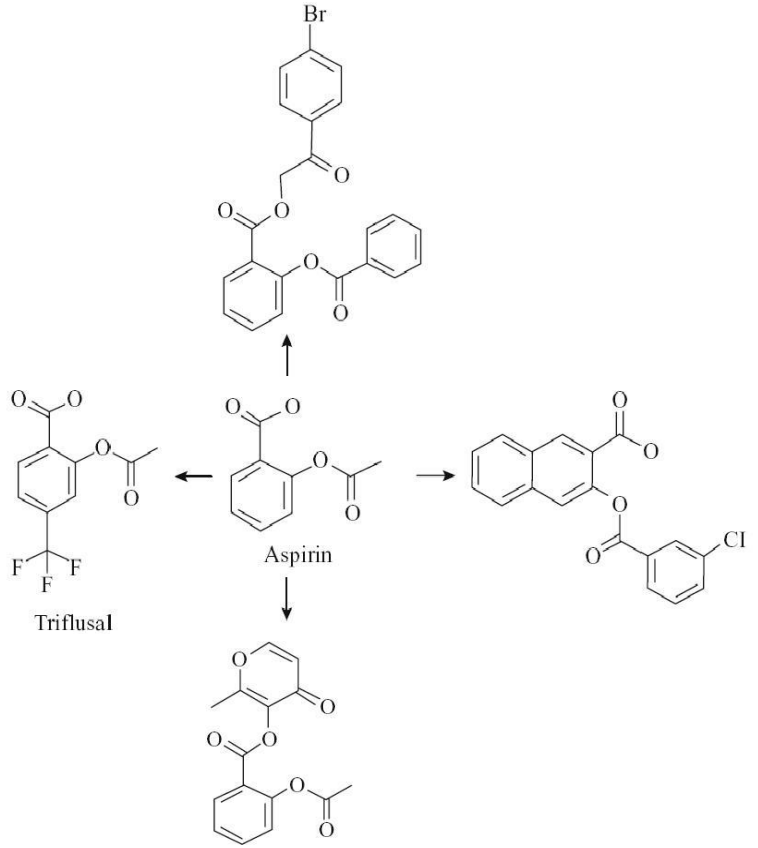


Figure 1.10. Example of compounds containing Aspirin as a substructure that can be used as a query for database searching

Similarity Searching –Structural queries and graphs (Reduced graph)

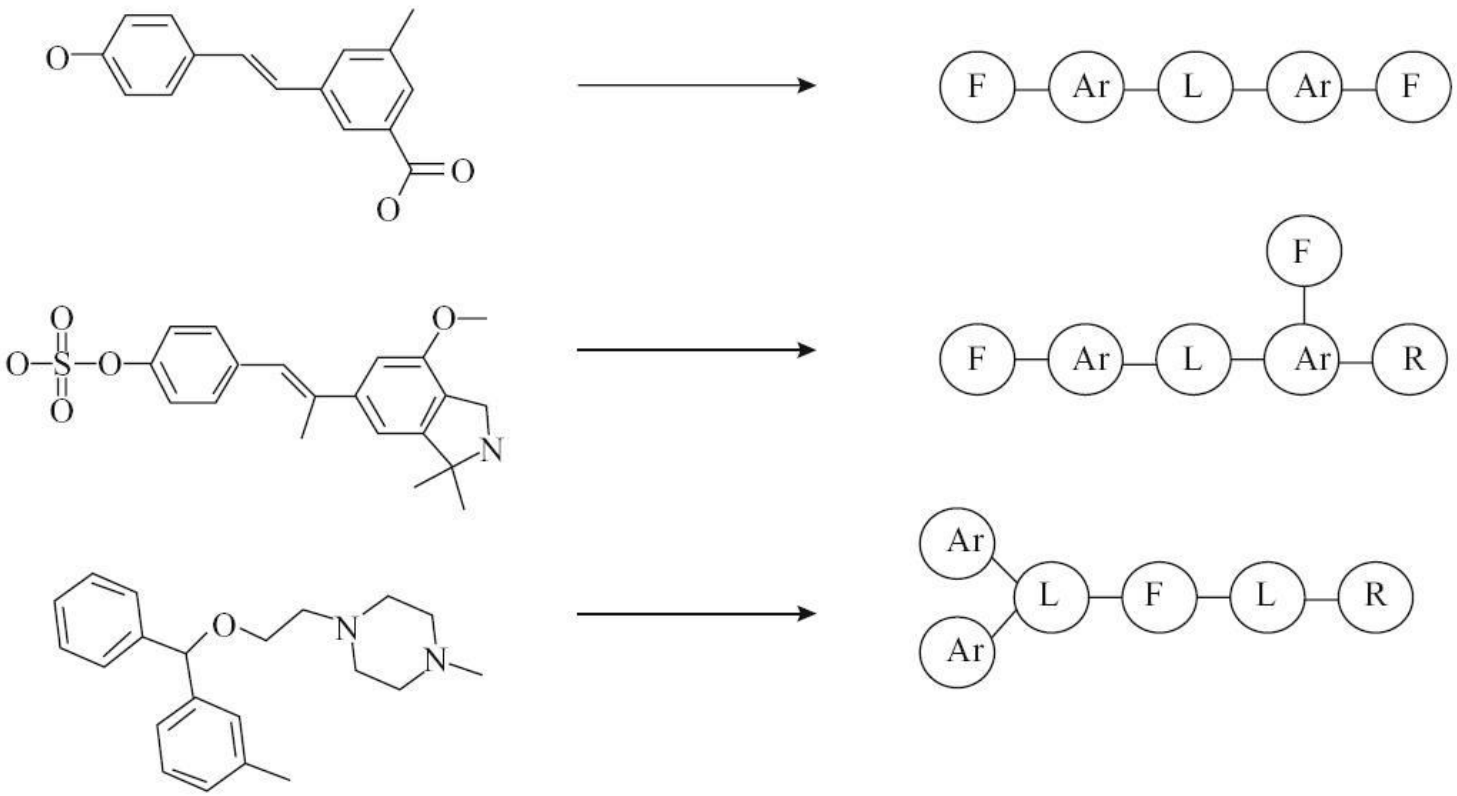


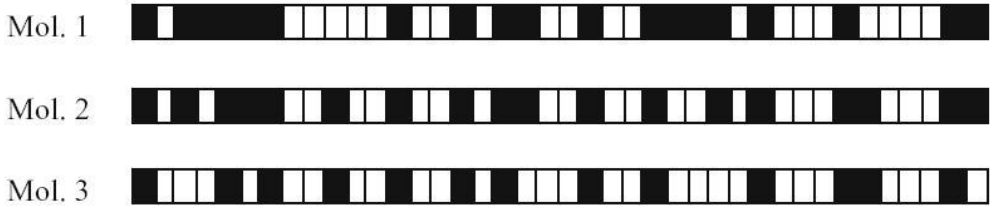
Figure 1.12. Examples of reduced graphs. Nodes corresponding to aromatic rings (Ar), aliphatic rings (R), functional groups (F) and linking groups (L) are shown (adapted from Gillet *et al.* 2003)

Similarity Searching – Pharmacophore

- A molecular framework that carries the essential features responsible for drug's biological activity
- Spatial arrangements of atoms or groups that are responsible for biological activity
- Often used as 3D queries for database searching

Similarity Searching –Fingerprints

- Fingerprints ：
  - widely used similarity search tools.
  - consist of various descriptors that are encoded as bit strings
  - Bit strings of query and database compared using similarity metric such as Tanimoto coefficient



$$Tc(Mol1, Mol2) = \frac{16}{(19+18)-16} = 0.76 \qquad Tc(Mol1, Mol3) = \frac{11}{(19+13)-11} = 0.52$$

Figure 1.14. Model fingerprints and Tc comparisons

Scaffold Hopping

The Concept of scaffold-hopping aims at finding molecules that possess ***different scaffolds*** but exhibit identical or very ***similar pharmocological*** activity

BIOACTIVE CONFORMATION

- Configuration and conformation are different.
- Receptor bound conformation is the bio-active conformation.
- Increase exponentially with the number of rotatable bonds.
- In general receptor bound conformations are almost impossible to predict from the ensemble of possible conformers.