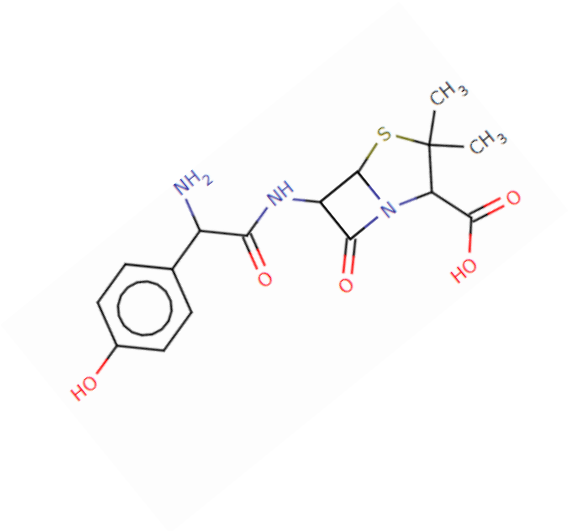
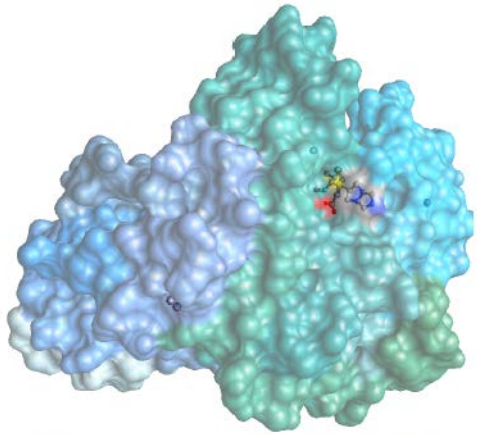
Instructor: Sael Lee

CS549 Spring – Computational Biology

LECTURE 19:
DRUG DISCOVERY & CHEMOINFORMATICS

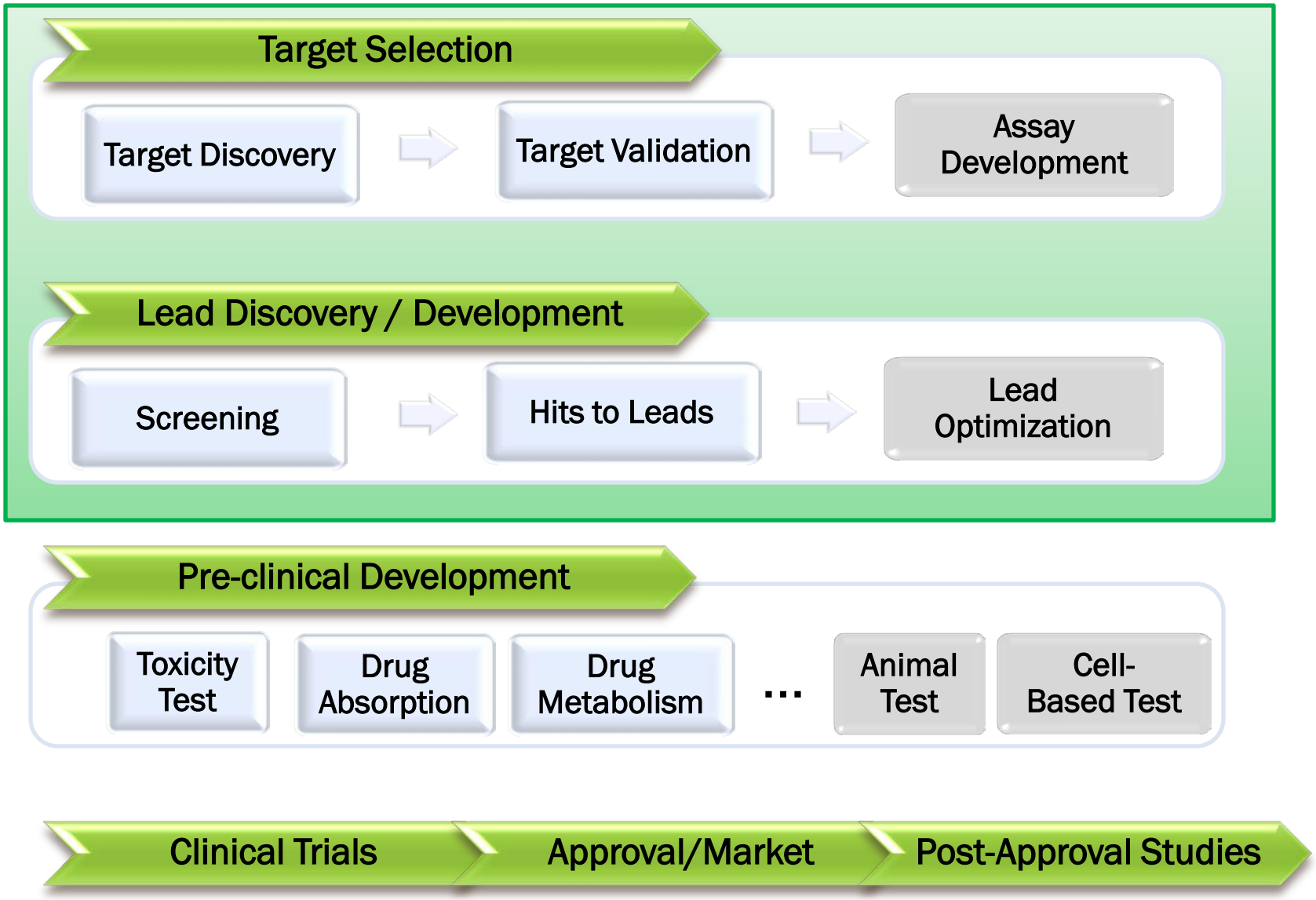
RATIONAL DRUG DISCOVERY





Biapenem in PBP-1A

TYPICAL RATIONAL DRUG DISCOVERY PROCEDURE

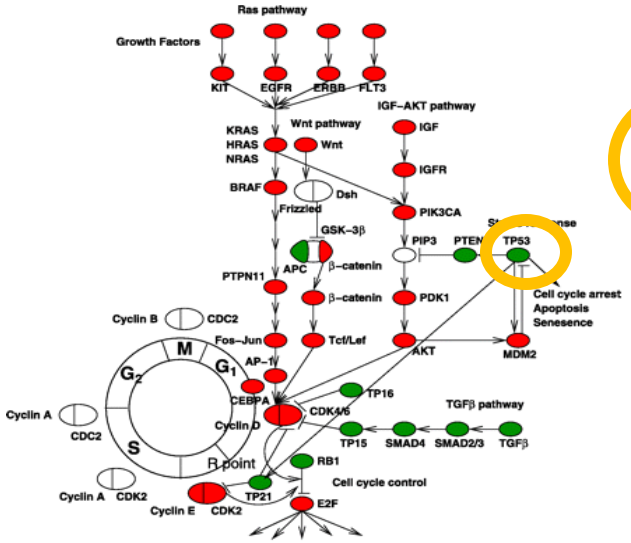


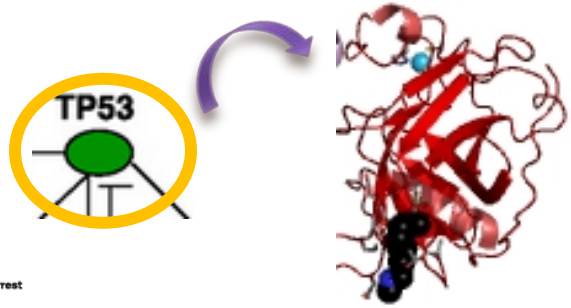
Target Selection

Target Discovery → Target Validation → Assay Development

Computational Functional Genomics

Druggability : Structure Analysis





PDBID: 2VUK
Cellular tumor antigen
p53 core domain

Computational Study

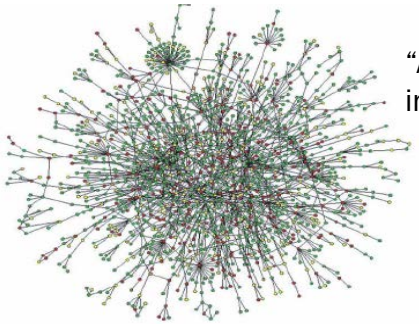
Experimental Study

[Yeang et al. *The FASEB Journal* 2008;22:2605-262]

Computational Functional Genomics

DEF.: Computational methods that make s use of the large scale genomic data to describe gene (and protein) functions and their interactions.

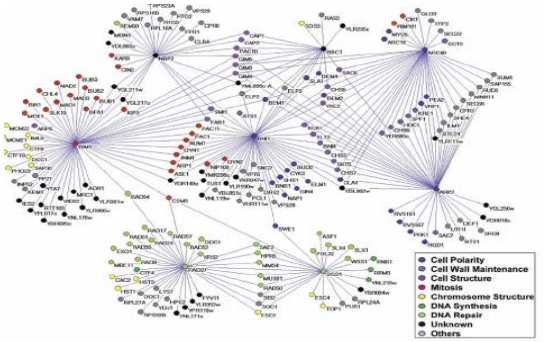
Protein-protein interaction network



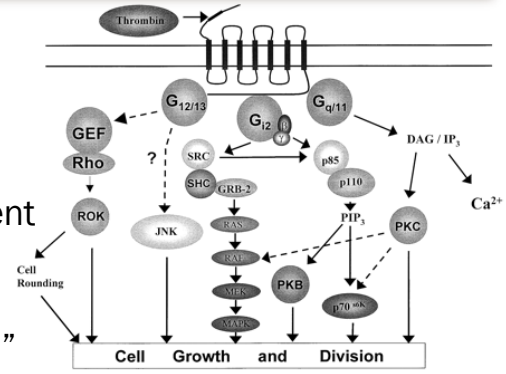
"A yeast protein–protein interaction network"

Gene association networks

"A yeast genetic network "

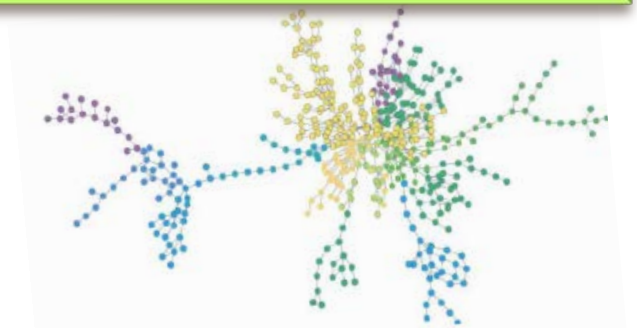


Regulatory pathways



"G-protein-dependent signaling pathways regulated through activation of PAR-1."

Metabolic pathways



"An E. coli metabolic network with 574 reactions and 473 metabolites colored according to their modules"

Druggability : Structure Analysis

DEF.: The suitability of a portion of a protein or protein complex to be targeted by a drug, especially by a small molecule drug.

Protein structure prediction	Prediction of the three-dimensional structure of a protein from its amino acid sequence
Protein-ligand/drug binding site prediction	Identification of potential interaction sites such as cavities or pockets on the structure
Protein surface analysis & searching	Calculation and comparison of physicochemical and geometric properties of the potential interaction sites

Protein structure prediction

- Computational determination of three dimensional structure of macro-molecules given their primary structure (amino acid sequence/DNA sequence/RNA sequence)
- Types of structure prediction
  - Protein structure prediction
    - Ab-initial structure prediction
    - Homology modeling
    - Threading
  - RNA structure prediction
  - DNA structure prediction

} Structural searching is important

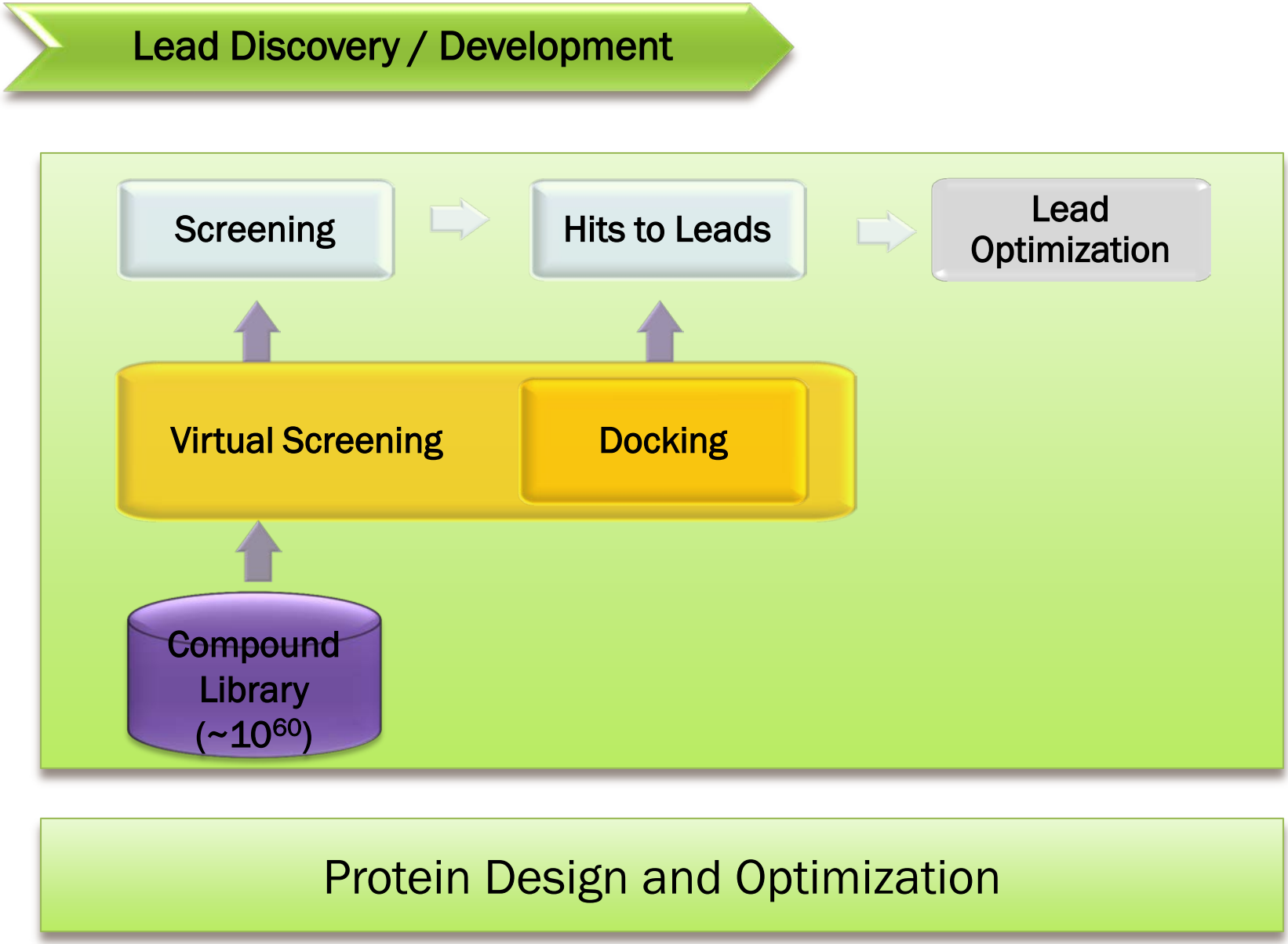
Protein-ligand / drug binding site prediction

Identifying potential ligand/drug binding sites in proteins using geometric properties such as pocket-like shape and evolutionally conservation information.

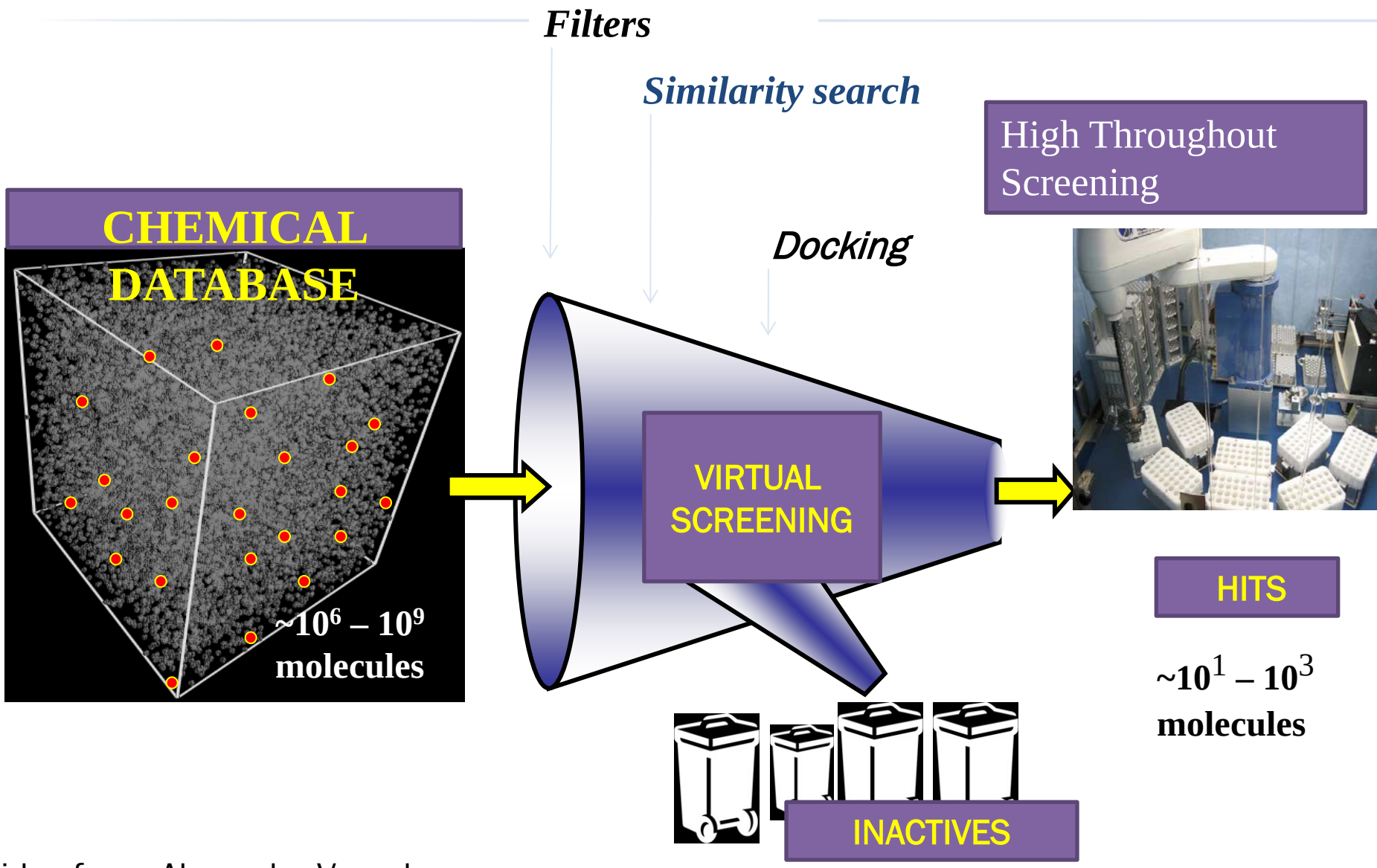
Some methods using geometric properties:

- **SURFNET** searches for a gap in a protein surface by fitting spheres inside the convex hull. [Laskowski RA. J Mol Graph1995;13:323–328]
- **PocketPicker** and **LIGSITE** locate a protein onto a three-dimensional (3D) grid and scan it for protein-void-protein events in many directions [Weisel et al. Chem Cent J 2007;1:7, Hendlich et al. J Mol Graph Model 1997;15:359–363]
- **VisGrid** uses the visibility of surface points to find pockets.
- **PocketDepth** clusters grid cells using information of the depth of the grid cells. [Kalidas & Chandra J Struct Biol 2008;161:31–42]

** Several methods consider additional information, such as sequence **conservation** and **energetics** which are often combined while considering geometrical shape.


Lead Discovery / Development
Screening
Hits to Leads
Lead Optimization
Virtual Screening
Docking
Compound Library (~10^60)
Protein Design and Optimization

Virtual screening "funnel"



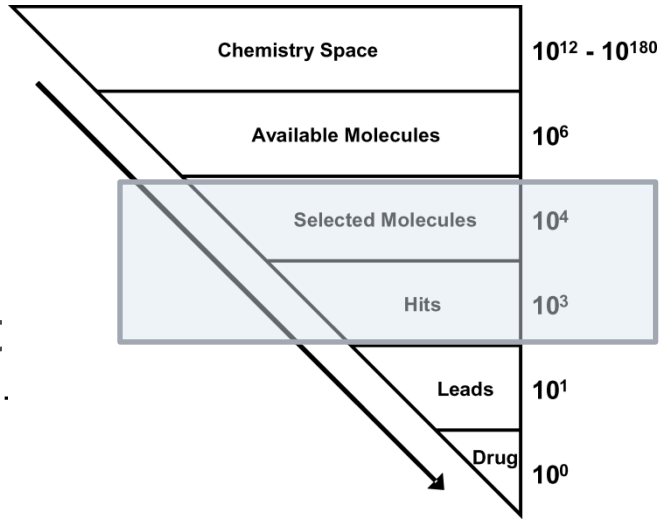
Slides from: Alexandre Varnek

Virtual screening

Computational quick search of large compound libraries in order to identify those structures which are most likely to bind to a drug target, typically a protein receptor or enzyme.

Chemoinformatics

- Similarity between known drugs or ones that have predefined properties



Molecular interaction predict

- Computational determination of whether a interact.

  Types of interaction prediction
  - Protein-small molecule interaction prediction
  - Protein-protein interaction prediction

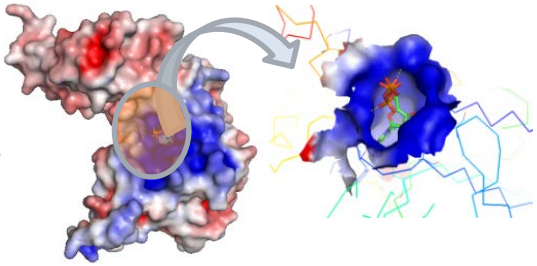
virtual screening are generally good at eliminate the bulk of inactive compounds (negative design). Actual selection of bioactive molecules for a given target requires more improvement(positive design).

Docking

"Computational methods that predict the preferred orientation of one molecule to a second when bound to each other to form a stable complex."[Lengauer & Rarey *Curr. Opin. Struct. Biol.* 1996; **6** (3): 402–6]

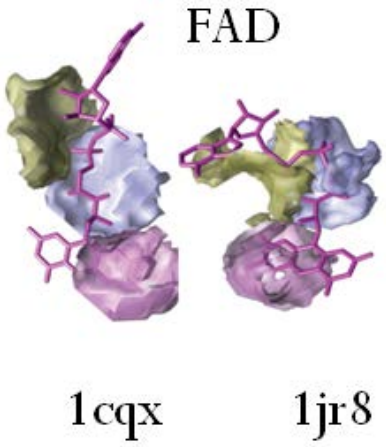
Protein-ligand docking

- Catalyze enzymatic reactions
- Metabolic processes
- Pocket like shapes



1AOI: ATP binding protein

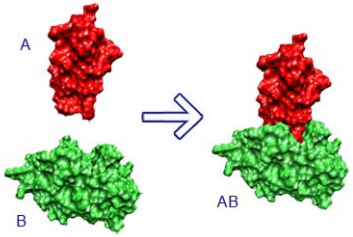
[Chikhi et al *Proteins* 2010]



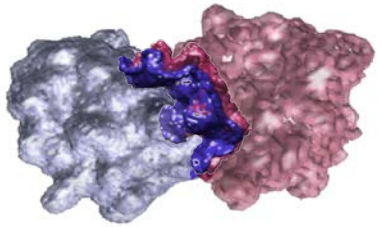
[Sael et al. *IJMS* 2010]

Protein-protein docking

- Permanent complex
- Transient interaction
- Mostly flat region



Z-Dock; LZerD;



1AY7: Ribonuclease Sa/Barstar complex

[Venkatraman et al. *BMC Bioinformatics* 2009]

Many of these problems deals with bio-molecular surface comparison.

CHEMOINFORMATICS & LIGAND-BASED VIRTUAL SCREENING

Resource:

- Brown, N. (2009). Chemoinformatics—an introduction for computer scientists. *ACM Computing Surveys*, *41*(2), 1–38.
- Karsten Borgwardt and Xifeng Yan | Part 8 I: Graph Mining
- Takigawa, I., & Mamitsuka, H. (2013). Graph mining: procedure, application to drug discovery and recent advances. *Drug discovery today*, *18*(1-2), 50–7.

Chemoinformatics: Chemical Similarity Searching

The mixing of information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug lead identification and optimization.

Frank K. Brown [1998] (cited in Russo [2000], page 4)

[Chemoinformatics involves]. . . the computer manipulation of two- or three-dimensional chemical structures and excludes textual information. This distinguishes the term from chemical information, largely a discipline of chemical librarians and does not include the development of computational methods.

Peter Willett, 2002 (cited in Russo [2001], page 4)

. . . the application of informatics to solve chemical problems . . . [and] chemoinformatics makes the point that you're using one scientific discipline to understand another scientific discipline.

Johann Gasteiger, 2002 (cited in Russo [2002], page 5)

"**Chemoinformatics** is an interface science aimed primarily at discovering novel chemical entities that will ultimately result in the development of novel treatments for unmet medical needs, although these same methods are also applied in other fields that ultimately design new molecules."

(N. Brown 2009)

THE SIMILAR-STRUCTURE, SIMILAR-PROPERTY PRINCIPLE

The fundamental assertion of chemoinformatics is the *similar-structure, similar-property principle* (*similar property principle*)

- similar molecules will also tend to exhibit similar properties; this is known as
- ". . . the so-called principle of similitude, which states that systems constructed similarly on different scales will possess similar properties."

[Johnson and Maggiora 1990, page 18]

Problems are solved by determining of structural similarity between two molecules, or a larger set of molecules.

Similarity searching in virtual screening from a problem-centric rather than a method centric perspective is needed, **depending on what is already known about a target and its ligands.**

CHEMICAL SEARCH SPACE

Chemistry space is the term given to the space that contains all of the theoretically possible molecules and is therefore theoretically infinite.

Druglike chemistry space : a set of empirically derived rules is used to define molecules that are more likely to be orally available as drugs.
Reduced druglike chemistry space is estimated to **contain anything from 10^{12} to 10^{180} molecules**

Goal of chemoinformatics is to assist in 1) **filtering** the space of available molecules to something more manageable while also 2) **maximizing** the chances of a) covering the molecules with the most potential to enter the clinic and b) maintaining some degree of structural diversity to avoid prospective redundancies or premature convergence.

Brown, N. (2009).

Representation	Name
Caffeine	Common Name
trimethylxanthine, theine, mateine, guaranine, methyltheobromine	Synonyms
$C_8H_{10}N_4O_2$	Empirical Formula
1,3,7-trimethylpurine-2,6-dione	IUPAC Name
58-08-2	CAS Registry Number
T56 BN DN FNVNVJ B F H	WLN
CN1C=NC2=C1C(=O)N(C(=O)N2C)C	SMILES
CN1C(=O)N(C)c2ncn(C)c2C1=O	SMILES (Aromatic)
1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3	InChI

```
C 0 0 0 1 0 0 0 0 0 0 1 0 2
0 C 0 0 0 0 0 0 0 0 1 1 2 0
0 0 C 0 2 0 0 0 0 1 1 0 0 0
0 0 0 C 0 0 0 0 1 2 0 0 0 0
1 0 2 0 C 0 0 0 1 0 0 0 0 0
0 0 0 0 0 C 0 0 0 0 1 0 0 0
0 0 0 0 0 0 C 0 0 0 0 1 0 0
0 0 0 0 0 0 0 C 1 0 0 0 0 0
0 0 0 1 1 0 0 1 N 0 0 0 0 0
0 0 1 2 0 0 0 0 0 N 0 0 0 0
0 1 1 0 0 1 0 0 0 0 N 0 0 0
1 1 0 0 0 0 1 0 0 0 0 N 0 0
0 2 0 0 0 0 0 0 0 0 0 0 O 0
2 0 0 0 0 0 0 0 0 0 0 0 0 O
```

Adjacency Matrix

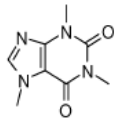
```
Caffeine
  Comment Line

 14 15  0  0  0  0            999 V2000
    3.0312   -2.1688    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.7457   -0.9312    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.3168   -0.9312    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.0473   -1.3437    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.3168   -1.7563    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0312    0.3063    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.4602   -2.1688    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.2773   -2.7958    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.5322   -2.0112    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    1.5322   -0.6763    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    3.0312   -0.5187    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    3.7457   -1.7563    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    4.4602   -0.5187    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.0312   -2.9938    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
 11  3  1  0  0  0  0
 11  2  1  0  0  0  0
  5  1  1  0  0  0  0
```

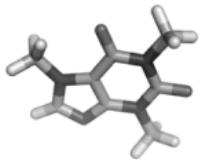
Connection Table (SDF)

Fig. 6. Some of the many ways in which molecules can be represented from simple names, empirical formulae, and line notations, through to computational models of their structure.

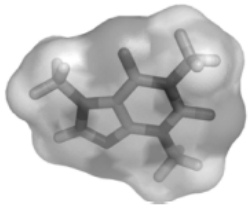
```
  1 12  1  0  0  0  0
 12  2  1  0  0  0  0
  3 10  1  0  0  0  0
  4  9  1  0  0  0  0
  4 10  2  0  0  0  0
  9  5  1  0  0  0  0
  5  3  2  0  0  0  0
 11  6  1  0  0  0  0
  2 13  2  0  0  0  0
 12  7  1  0  0  0  0
  1 14  2  0  0  0  0
  9  8  1  0  0  0  0
M  END
$$$$
```



Topology (2D Structure)



Topography (3D Structure)



Topography (Surface Model)

Fig. 6. Continued.

Caffeine representations

- Define **feature vectors** that record the presence/absence (or number of occurrences) of particular patterns in a given molecular graph

$$\phi(A) = (\phi_s(A))_{s \text{ substructure}}$$

where

$$\phi_s(A) = \begin{cases} 1 & \text{if } s \text{ occurs in } A \\ 0 & \text{otherwise} \end{cases}$$

- Extension of traditional chemical **fingerprints**

FINGERPRINTING METHOD

DIB-09 (2013) Chloé-Agathe Azencott: Data Mining in Bioinformatics

CHEMISTRY AND GRAPH THEORY

The **molecular graph** is a type of graph that is undirected and where the nodes are colored and edges are weighted where the **nodes** are the **atoms** of a molecule and the **edges** are the **bonds**.

- The individual nodes are colored according to the particular atom type: carbon (C), oxygen (O), nitrogen (N), chlorine (Cl),etc.,
- The edges are assigned weights according to the bond order: single, double, triple, and aromatic.

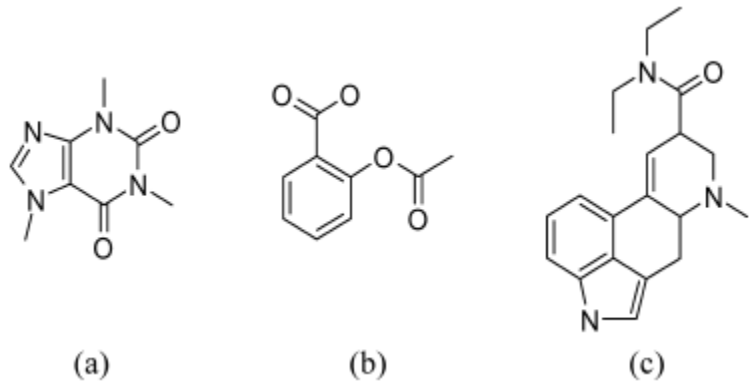
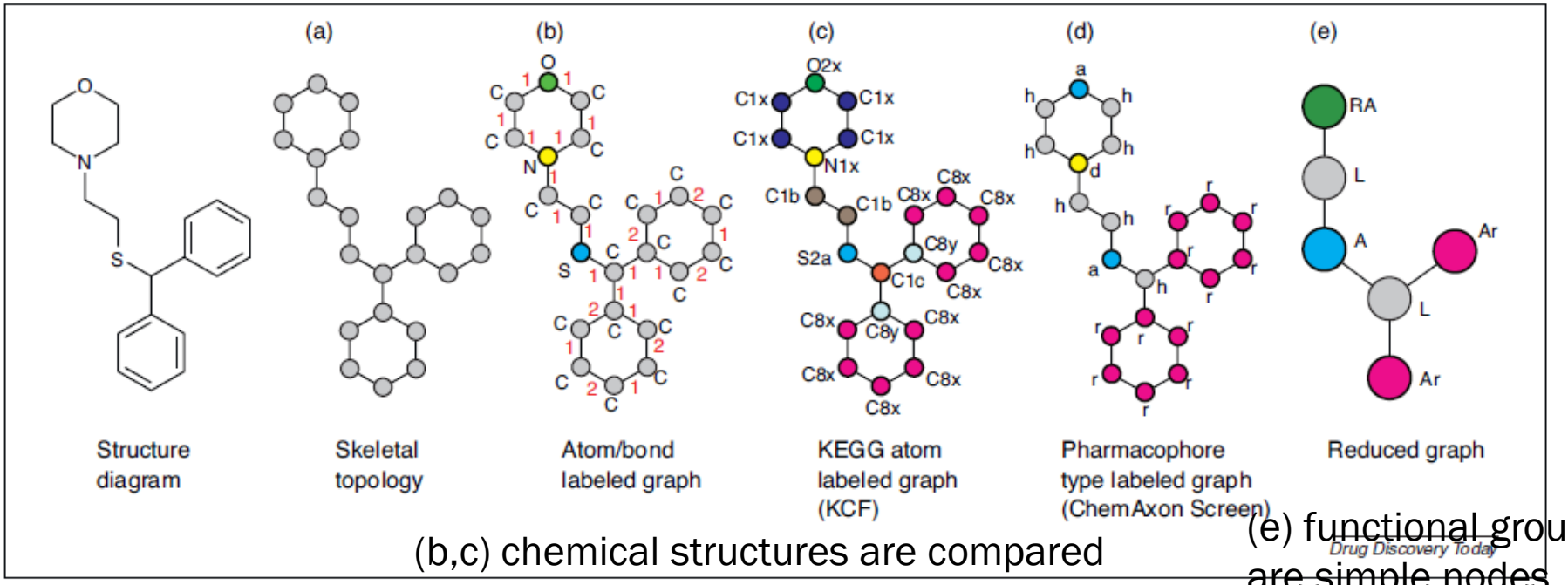


Fig. 5. The hydrogen-depleted molecular graphs of (a) caffeine, (b) aspirin, and (c) D-lysergic acid diethylamide. (N Brown 2009)

FIVE TYPES OF MOLECULAR GRAPHS REPRESENTING

(a) topological features

(d) focus on molecular properties



(b,c) chemical structures are compared

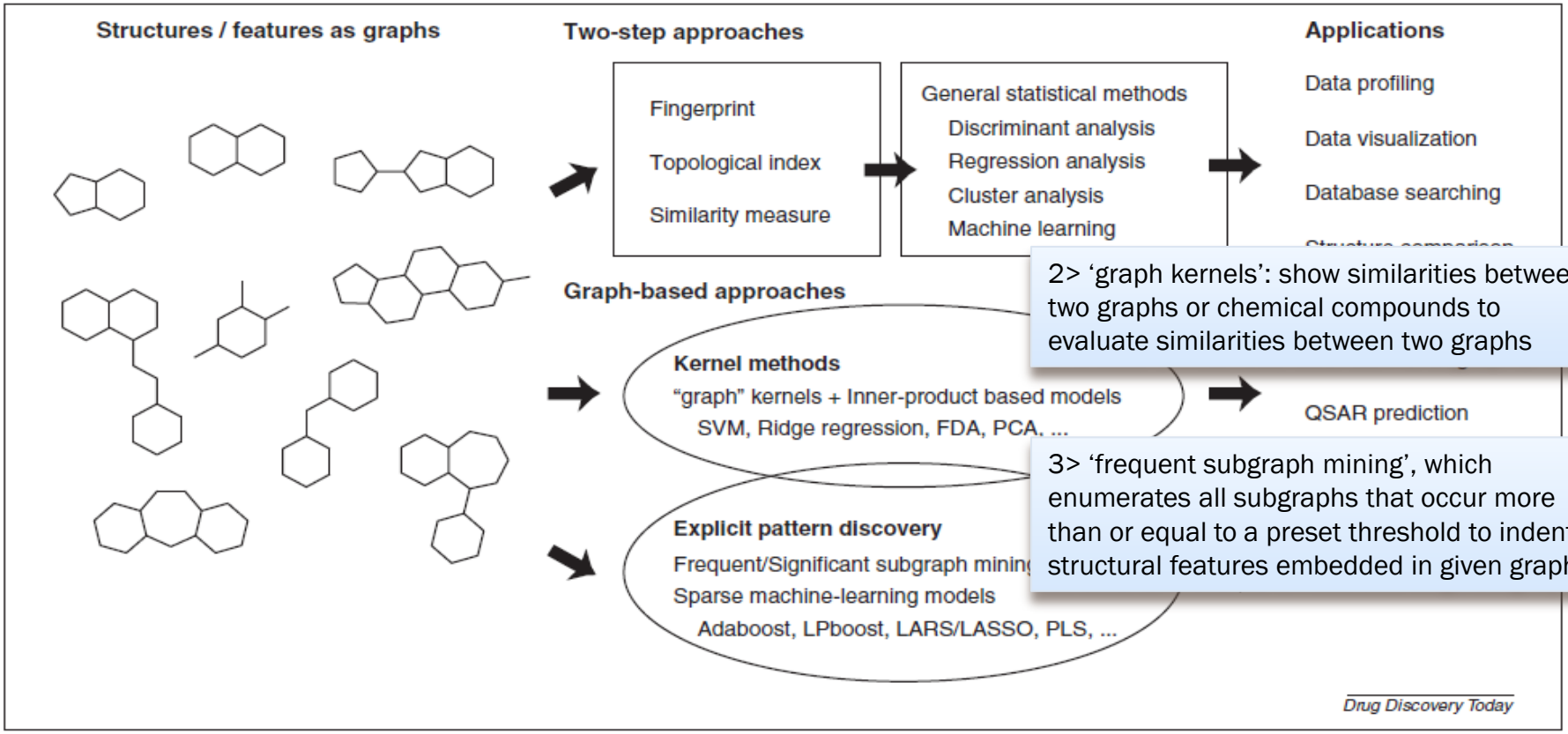
(e) functional groups are simple nodes,

FIGURE 2

Five molecular graph types. *Abbreviations*: KCF: KEGG Chemical Function.

Fig. from Takigawa, I., & Mamitsuka, H. (2013).

VARIOUS GRAPH MINING-BASED APPROACHES



2> 'graph kernels': show similarities between two graphs or chemical compounds to evaluate similarities between two graphs

3> 'frequent subgraph mining', which enumerates all subgraphs that occur more than or equal to a preset threshold to indentify structural features embedded in given graphs.

FIGURE 1

Three types of graph mining approaches. *Abbrevations*: ADME/Tox: absorption, distribution, metabolism, excretion and toxicology; LARS: least square regression; LASSO: least absolute shrinkage and selection operator; PCA: principal component analysis; PLS: partial least squares; QSAR: quantitative structure–activity relationship; SVM: support vector machines.

Fig. from Takigawa, I., & Mamitsuka, H. (2013).

FREQUENT SUBGRAPH MINING

Frequent subgraph mining is used for analyzing structural fragments or partial structures and molecular graphs.

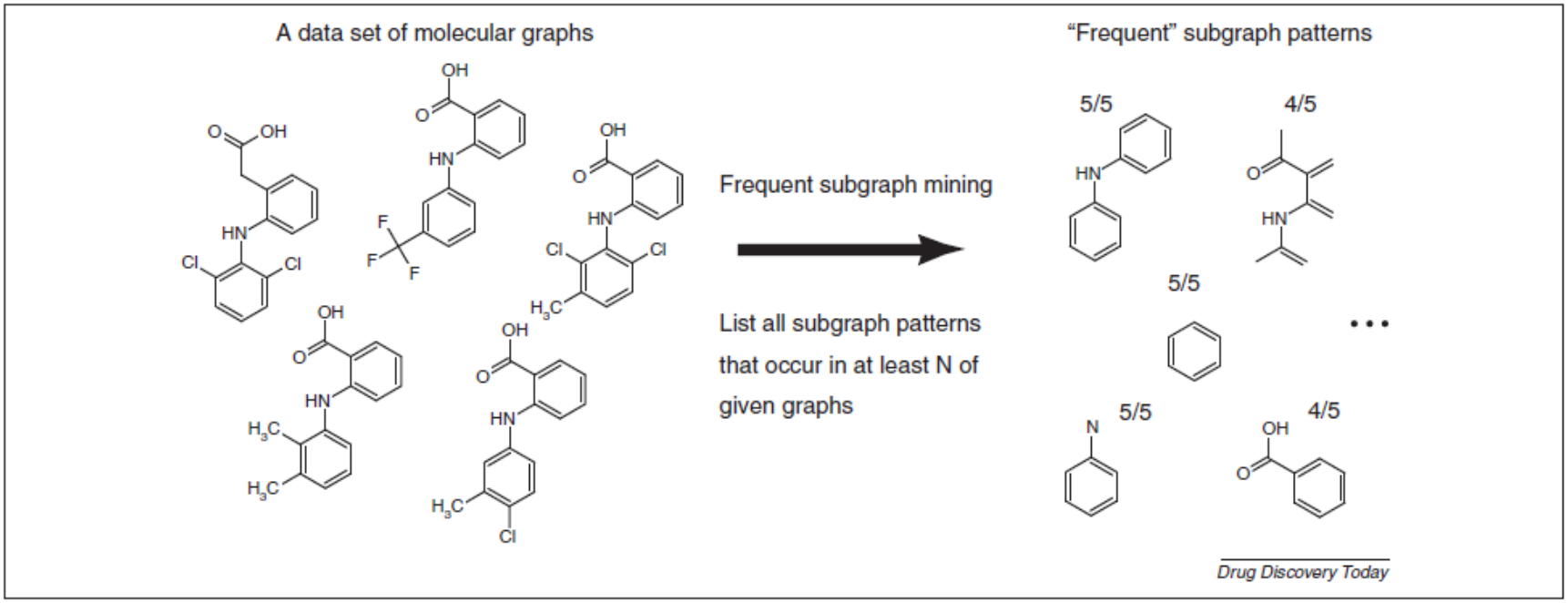


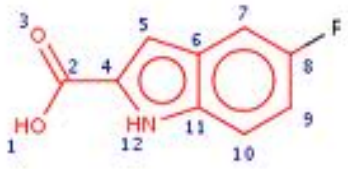
FIGURE 3

Frequent subgraph mining. Note that this example does not consider aromaticity, however, it can be incorporated.

Fig. from Takigawa, I., & Mamitsuka, H. (2013).

SUBGRAPH ISOMORPHISM

Problem: Given two graphs G and H as input, determine whether G contains a subgraph G' that is where two vertices u and v of G' are adjacent in G' if and only if f(u) and f(v) are adjacent in H (isomorphic to H)



1

2

"subgraph isomorphism problem' is theoretically proven to be NP-complete.