Cheminformatics

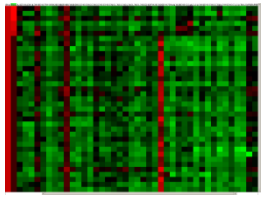
Uwe Koch

Lead
Discovery
Center
LDC

Cheminformatics

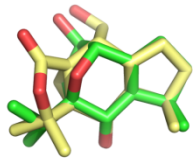
Lead Discovery Center LDC

Target identification



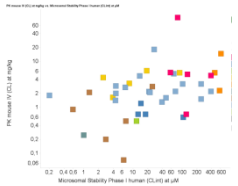
Genomics
Proteomics
Ligand-based
Pathway analysis

Lead finding



Compound acquisition
Combinatorial chemistry
Virtual screening
Data mining
HTS screening support
Filtering

Lead optimization



(Q)SAR
Structure based design
In silico ADME/Tox
Biososters

Process very large datasets - chemical structures, screening results

LDC – Cheminformatics

Lead Discovery Center LDC

Cheminformatic activities at LDC:

- Compound acquisition
- Analysis of screening data: Filtering, Clustering
- Acquisition of Hit analogs: in silico screening, 2D and 3D, Docking
- Support Hit optimization:

    Structure and pharmacophore based modelling

    ADME/T: Identify metabolic hot spots, toxicophores …

- Support ELF: Library optimization, reagent selection, enumeration, calculation of properties of library.

Compound collection

Purchase of commercial compounds

Focus on:

- Diversity (Fingerprint calculation & Clustering)
- Good phys-chem property space

  (eg logP, MW, PSA, HBD & HBA counts, rotatable bonds)
- Avoid problematic substructures, frequent hitters and/or toxicophores
- Favour novel chemistry (eg.number of nearest neighbours or same scffold in sureChem)

Additional criteria for sub-libraries

Project centric – target class libraries

Chemistry centric: novel chemistry, 3D character

Property calculation

ADME related properties

Properties related to biological effect and fate in organism

- water solubility
- pka / protonation state
- log P and log D

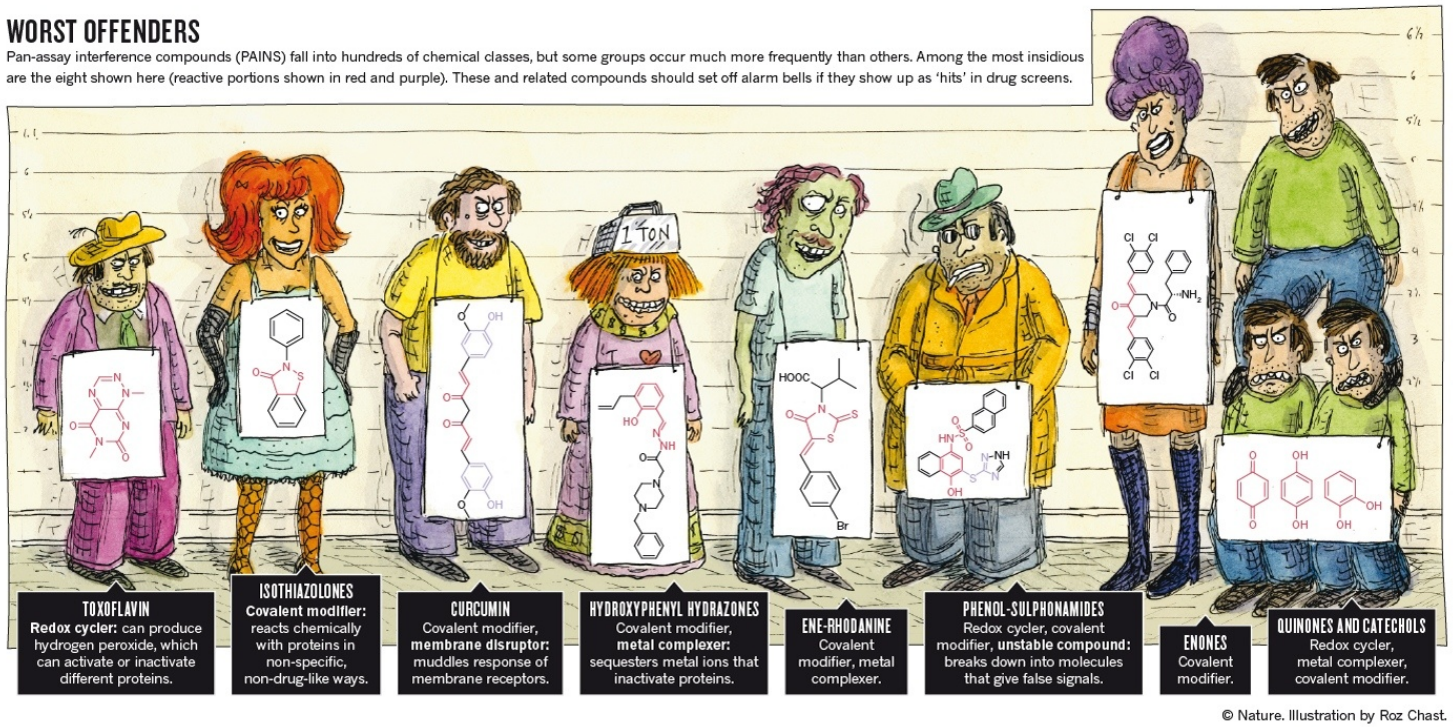
The following properties describe complex biological processes for which it is more difficult to build reliable models.

- toxic and metabolic characteristics
- drug transport characteristics

Compound collection

Avoid problematic substructures:

PAINS – Pan Assay Interference Compounds (eg redox cyclers producing H_2O_2, which inactivates the protein)



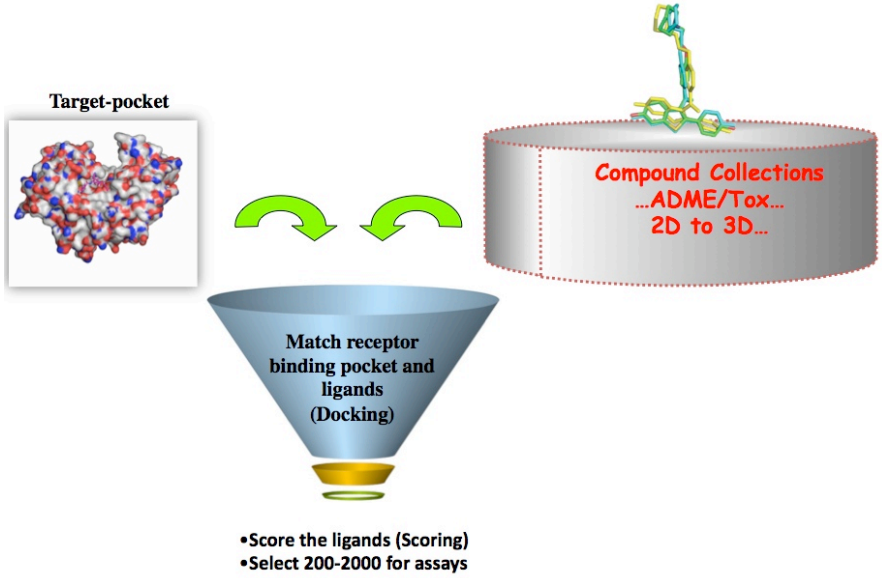
© Nature. Illustration by Roz Chast.

Virtual Screening

Lead Discovery Center LDC

Two major approaches

• Structure based virtual screening requires knowledge of the 3D structure of the biological target (Docking)



• Ligand-based virtual screening requires knowledge of at least some ligands that exhibit the desired bioactivity

Virtual screening

Ligand based approaches:

-**Pharmacophore methods**: identification of the pharmacophoric pattern common to a set of known actives and the use of this pattern in a subsequent 3D substructure search.

- **Machine learning methods**: develops classification rules based on a training set of actives and inactives

- **Similarity methods:** based on the central premise of medicinal chemistry: *Structurally similar molecules exhibit similar biological activities*

A bioactive reference is searched against a database to identify the nearest neighbour molecules

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 *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

Lead Discovery Center LDC

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

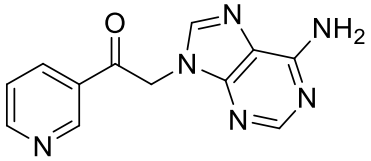
Descriptors calculated from 2D graphs:

- topological indices, 2D fingerprints

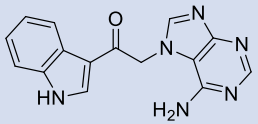
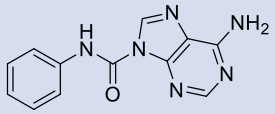
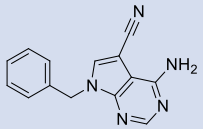
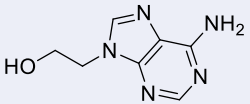
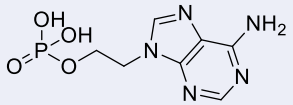
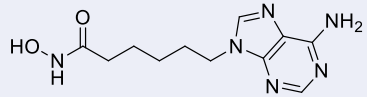
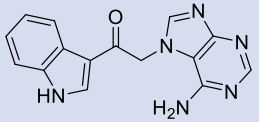
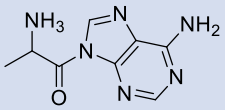
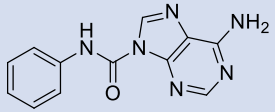
Descriptors requiring 3D representations:

- Pharmacophore descriptors

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	 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

Lead Discovery Center LDC

Search results depends on molecular descriptors

Highly unlikely that any one method performs equally well under all circumstances („No free lunch theorem“ of informatics)

Data fusion: if many virtual screening methods are available combinations of results from multiple methods to prioritise compounds

HTS

Lead Discovery Center LDC

Workflow

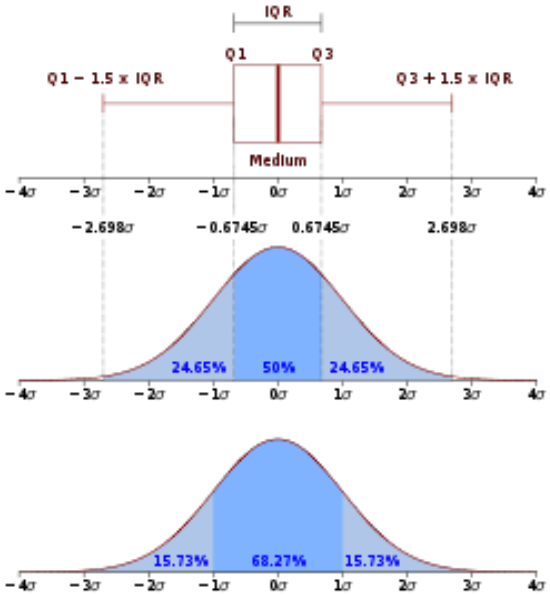
- Run HTS, measure %activity
- Select actives based on activity cut-off
- Filter actives – undesirable substructures, off-target activity, physicochemical and eADME properties
- Cluster actives – based on fingerprints, maximal common substructure
- Identify inactives related to active series (cluster hit rate)
- Hit validation (IC50, orthogonal & secondary assays)
- Hit expansion – ligand based virtual screen for further analogs

Screening data

Large quantity of activity data generated by screening

Select actives

Compounds with activity significantly above average (DMSO)



One measure is the interquartile range (IQR) determined for a reference set, eg DMSO.

Calculation of interquartile range: Q3 – Q1

Actives: %Act < DMSO median – 2 * IQR

Screening data

Filtering

Frequent hitters (eg Pains), substructure based

Toxicophores, substructure based,

eg in a test set it has been shown for mutagenicity*

	compds	mutagens	non-mutagens
polycyclic aromatic	660	614	46
aromatic nitro	632	561	71
aromatic amine	441	380	61
aromatic azo	88	67	21

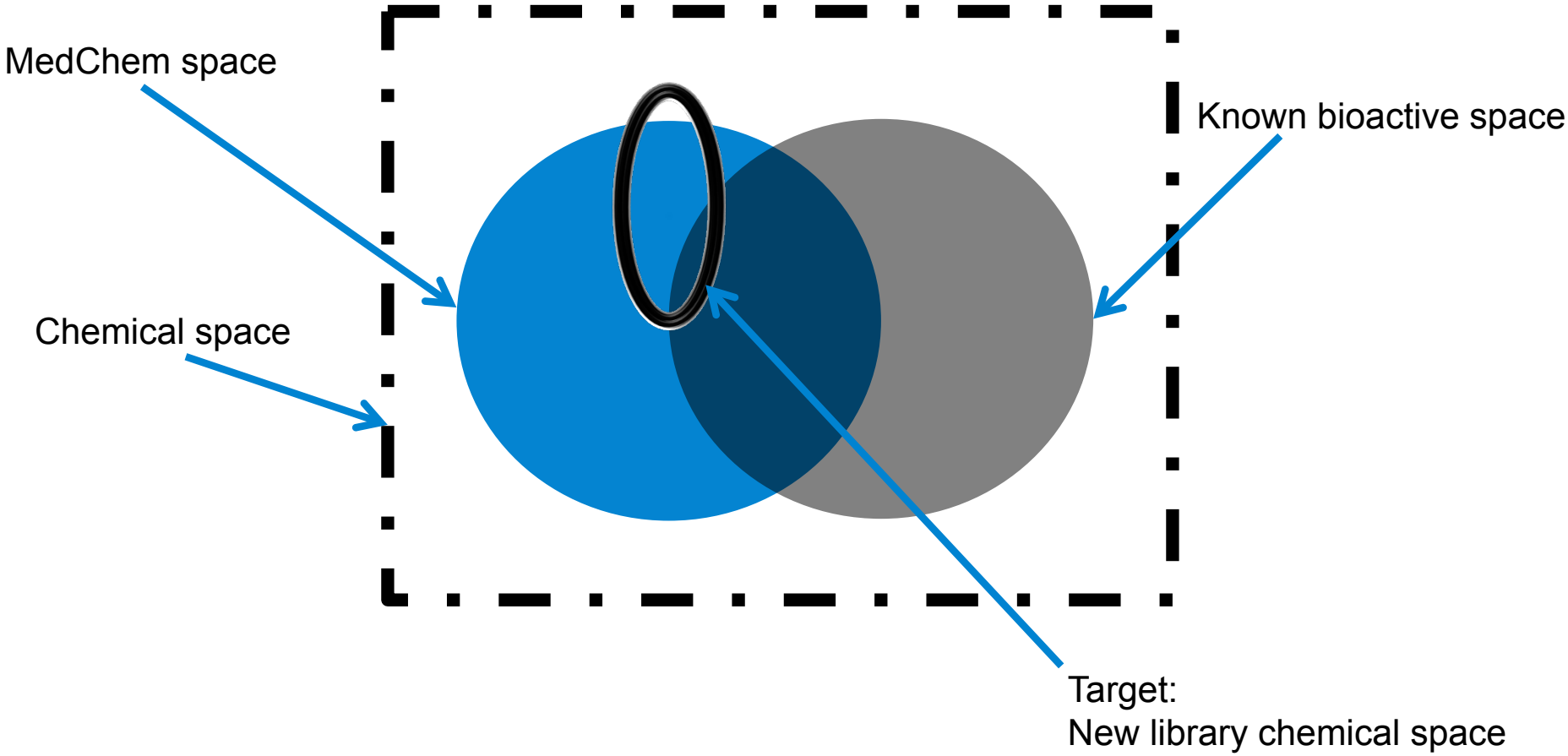
Physicochemical properties (eg. MW > 600, logP >5)

Purity

*Ref

Library design –increasing the compound collection

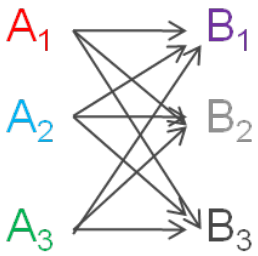
Generate novel MedChem-like molecules



Library design –increasing the compound collection

Introduction to combinatorial libraries

Building blocks



	A1	A2	A3
B1	*A1B1*	*A2B1*	*A3B1*
B2	*A1B2*	*A2B2*	*A3B2*
B3	*A1B3*	*A2B3*	*A3B3*

Library

$A_1 B_1$

$A_1 B_2$

$A_1 B_3$

$A_2 B_1$

$A_2 B_2$

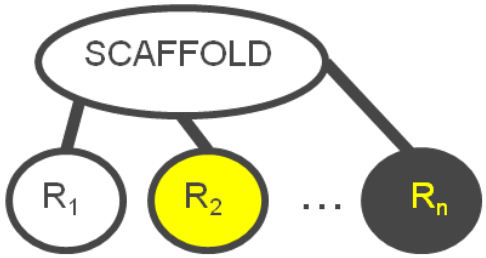
$A_2 B_3$

$A_3 B_1$

$A_3 B_2$

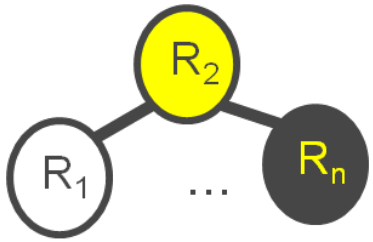
$A_3 B_3$

Type A



Scaffold-based libraries

Type B



Backbone-based libraries

Library design –increasing the compound collection

Library design – cheminformatics

Reagent selection:

- diversity

- fill holes in chemical space of existing screening collections

- Physicochemical properties – MW, lipophilicty, PSA

Two strategies:

Reactand based: select building blocks based on their properties

Product-based: select building blocks based on properties of final library, computationally more demanding

Recent development: smaller libraries with target focus

Cheminformatics: Future directions

Lead Discovery Center LDC

- Extraction of knowledge from increasingly large global databases
- Integration of multiple data sources – biological, pharmacological and chemical (patent) data
- Integration with bioinformatics
- Based on increasingly available data on molecular properties more reliable models for toxicity and eADME prediction
- Open source collaborative software development