CABANA

Capacity building for bioinformatics in Latin America

Introduction to Chemoinformatics

Databases, chemical space, descriptors, data visualization

José L. Medina-Franco
medinajl@unam.mx

CINVESTAV
Unidad Irapuato

DIFACQUIM
UNAM

DIFACQUIM
COMPUTER-AIDED DRUG DESIGN AT UNAM

Chemoinformatics as part of CADD

Computer Aided-Drug Design (CADD)

Chemoinformatics
Design, organization, management, visualization and analysis of chemical information

Molecular modeling
"Real" representation of molecules

Computational Chemistry

Theoretical chemistry

Table 1. Different definitions of chemoinformatics as a field.

Frank Brown[5]	The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization.
Greg Paris[45]	Chemoinformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information.
Johann Gasteiger[2]	Chemoinformatics is the application of informatics methods to solve chemical problems.
Jean-Loup Faulon and Andreas Bender[8]	Chemoinformatics is the field of handling chemical information
This work	Chemoinformatics is a field based on the representation of molecules as objects (graphs or vectors) in a chemical space.

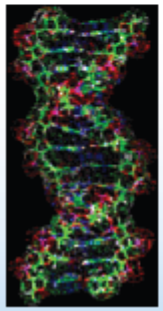
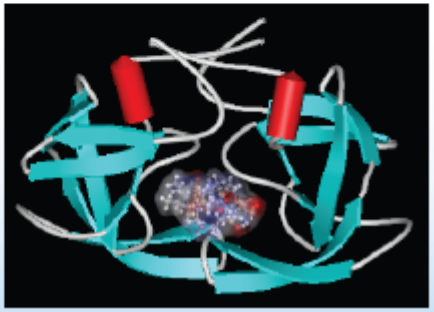
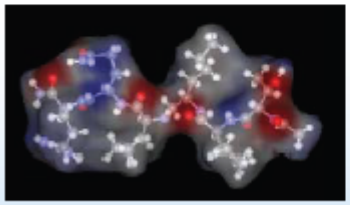
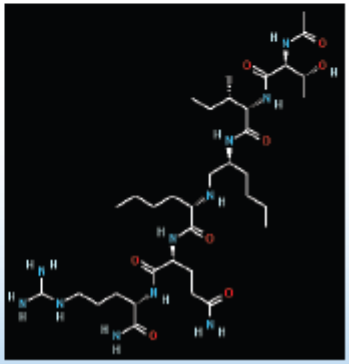
Table 2. Interrelations between three branches of theoretical chemistry.

	Quantum chemistry	Force field based molecular modeling	Chemoinformatics
Molecular model	Electrons and Nuclei	Atoms and bonds	Graphs and descriptor vectors
Inference mechanism	Deductive$\gg$inductive	Deductive$\cong$inductive	Deductive$\ll$inductive
Typically applied to	Individual species or ensemble of a few species	Individual species, complex system representing an ensemble of many species	Ensemble of species (both for knowledge extraction and predictions), individual species (for predictions only)
Basic concept	Wave/particle dualism	Classical mechanics	Chemical space
Basic mathematical approaches	Schrödinger equation and approximate methods (HF, DFT, …)	Force field method and its implementation in molecular mechanics, molecular dynamics, Monte-Carlo and free energy perturbation techniques	Statistical learning, graph theory

Varnek A, Baskin, II. Chemoinformatics as a Theoretical Chemistry Discipline. *Mol Inf* (2011) *30:*20

Bioinformatics Chemoinformatics

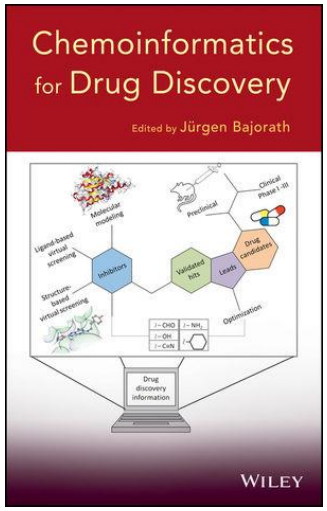
gene ⟷ protein ⟷ drug ⟷ lead

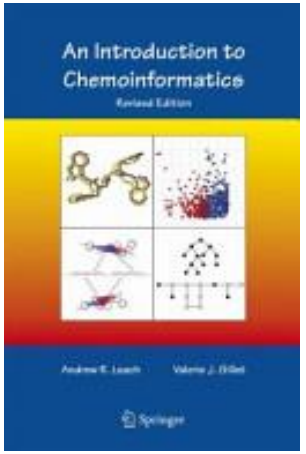





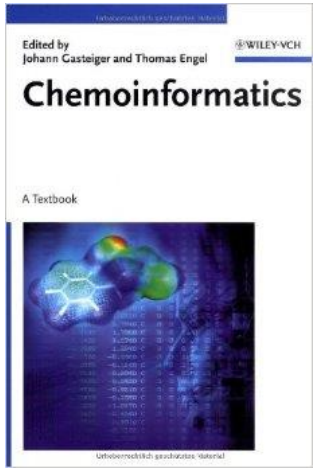
Gastaiger J. History and Challenges of Chemoinformatics

Areas of application of chemoinformatics

- *drug design*
- *analytical chemistry*
- *chemical engineering*
- *inorganic chemistry*
- *medicinal chemistry*
- *organic chemistry*
- *physical chemistry*
- *theoretical chemistry*







Gastaiger J. History and Challenges of Chemoinformatics

COMPOUND DATABASES

General type of compound databases

- Public
- Commercial (MDDR, CRC Press, Prod. Nat.)
- *In-house*
- *De novo,* virtual chemical libraries (on demand)

Journal of Medicinal Chemistry

Perspective

Cite This: *J. Med. Chem.* 2019, 62, 1116–1124

pubs.acs.org/jmc

Virtual Chemical Libraries

Miniperspective

W. Patrick Walters*

Relay Therapeutics, 215 First Street, Cambridge, Massachusetts 02142, United States

Public databases

Few of many examples

- **ZINC**

  Purchasable compounds
  http://zinc.docking.org/

- **National Cancer Institute**

  Send free samples for screening
  http://dtp.nci.nih.gov

For virtual screening

DRUGBANK http://www.drugbank.ca/

PHAROS https://pharos.nih.gov/

Information of drugs, Drug repurposing

Compound databases with biological activity

- ChEMBLdb
  - www.ebi.ac.uk/chembl
- Binding Database
  - www.bindingdb.org
- PubChem
  - https://pubchem.ncbi.nlm.nih.gov/

Bender A. *Nat. Chem. Biol.* (2010) *6:*309
Scior T, Bernard P, Medina-Franco JL, Maggiora GM. *Mini-Rev. Med. Chem.* (2007) *7:*851

Other public databases

- ChemSpider
  - http://www.chemspider.com/
- ChemBank
  - http://chembank.broad.harvard.edu
- World of Molecular Bioactivity (WOMBAT)
  - http://www.sunsetmolecular.com/index.php
- ChemMine
  - http://bioweb.ucr.edu/ChemMineV2/
- French National Chemical Library
  - http://chimiotheque-nationale.enscm.fr
- Chemical Universe Database (GBD)
  - http://www.gdb.unibe.ch/gdb/home.html

Bender A. *Nat. Chem. Biol.* (2010) *6:*309
Scior T, Bernard P, Medina-Franco JL, Maggiora GM. *Mini-Rev. Med. Chem.* (2007) *7:*851

Databases of Natural Products

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: *J. Chem. Inf. Model.* 2018, 58, 1518–1532

pubs.acs.org/jcim

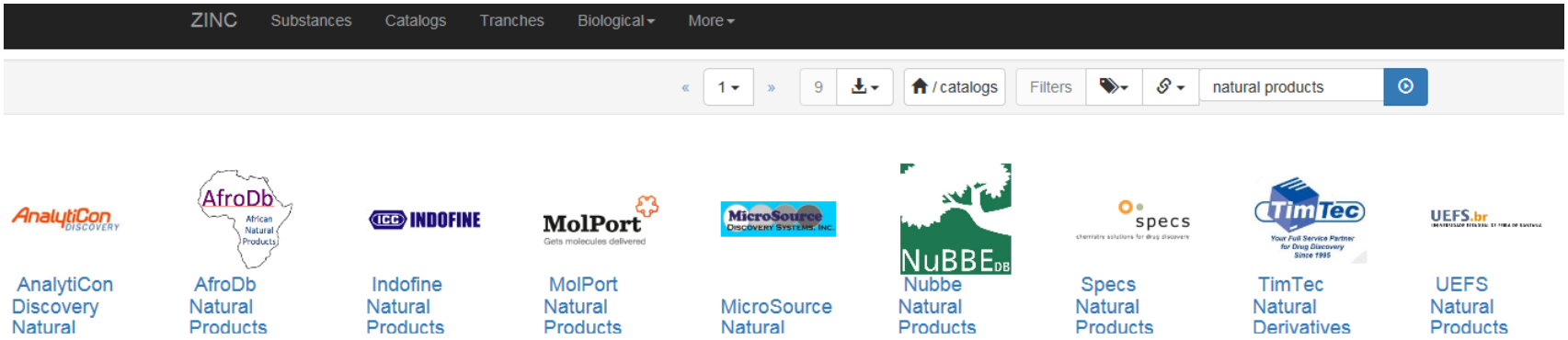
Characterization of the Chemical Space of Known and Readily Obtainable Natural Products

Ya Chen, Marina Garcia de Lomana, Nils-Ole Friedrich, and Johannes Kirchmair*

Center for Bioinformatics, Department of Computer Science, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany

> 250,000 natural products in public databases

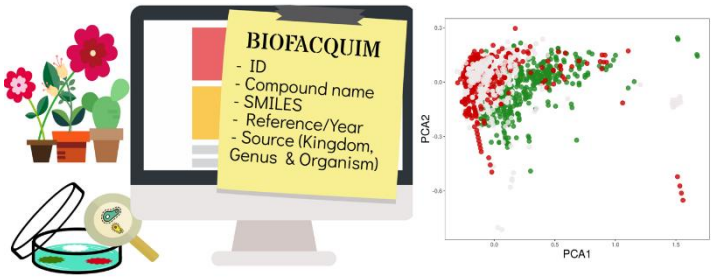
In ZINC



http://zinc15.docking.org/catalogs

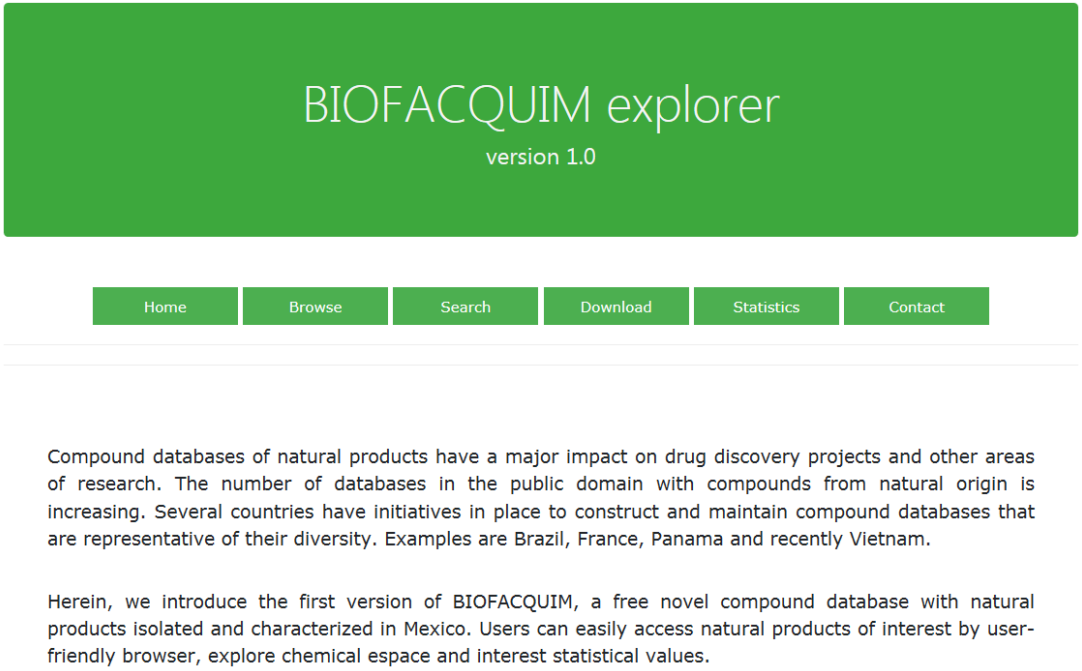
BIOFACQUIM

A database of natural products from Mexico



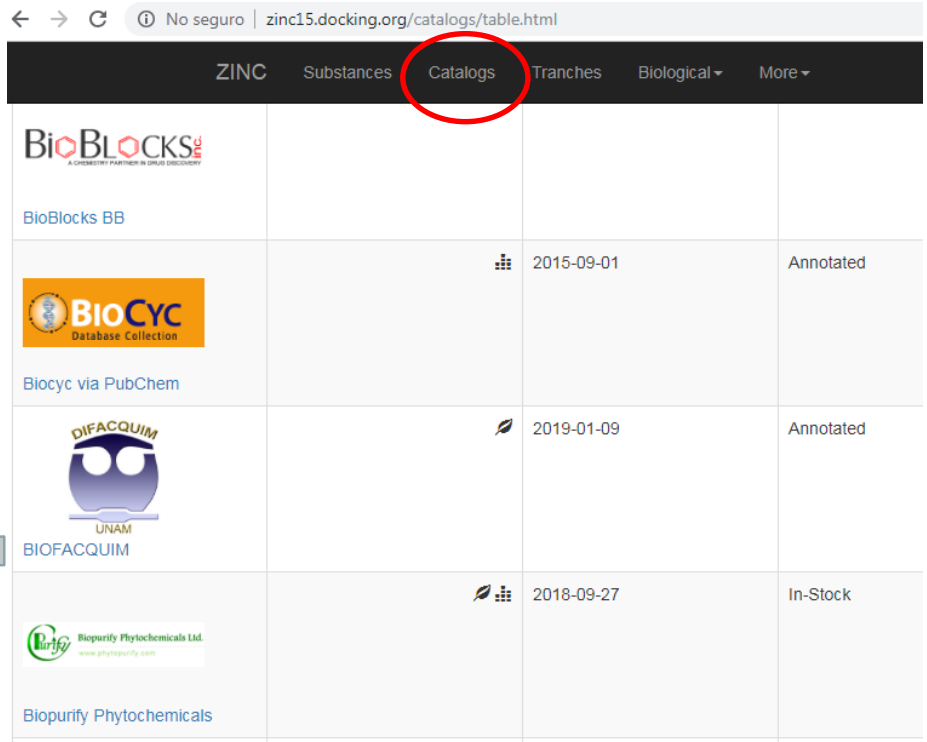
Angélica Pilón

https://biofacquim.herokuapp.com



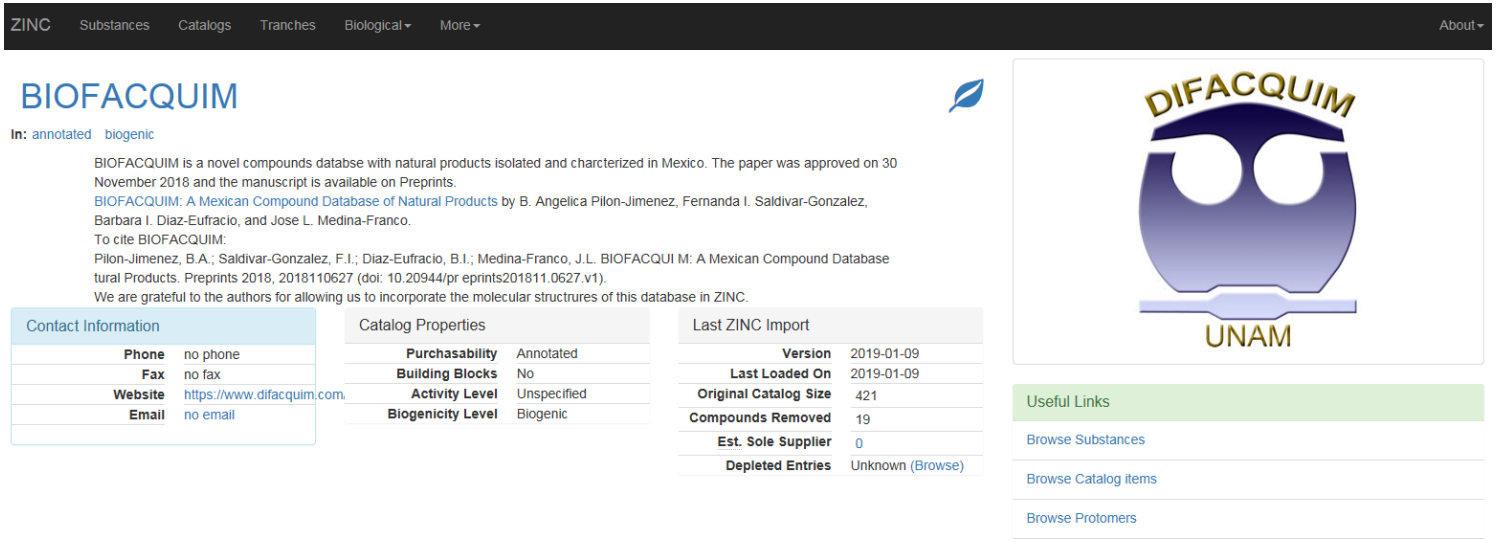
Developed by MScs Bárbara Díaz

Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, et al *Biomolecules* 2019 *9*:31

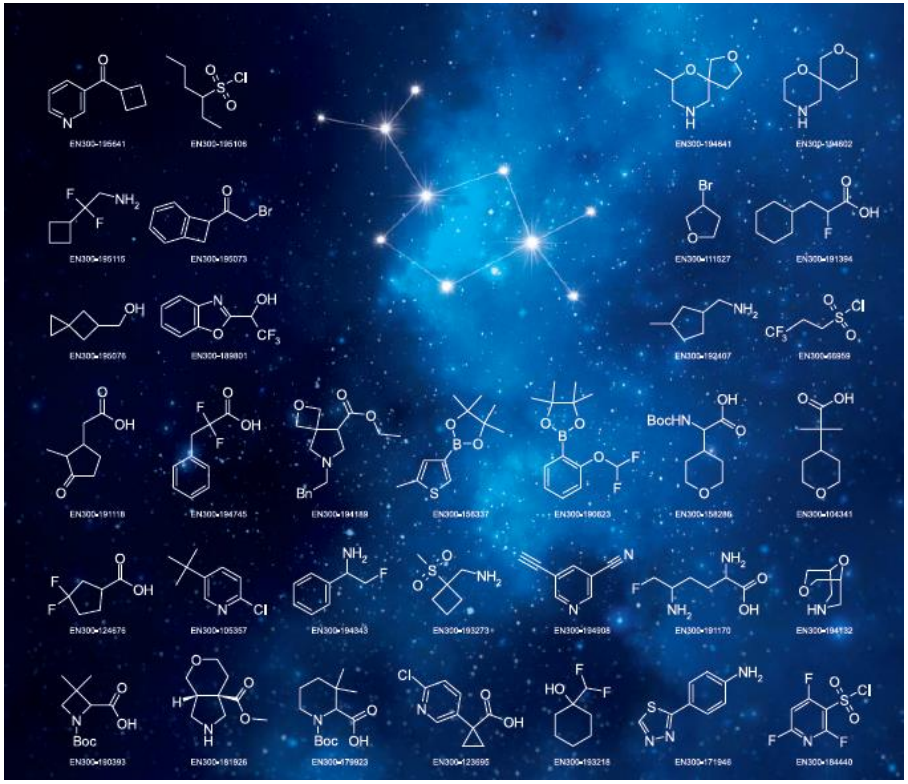


Available in ZINC

Thanks to John Irwin (UCSF)



http://zinc15.docking.org/catalogs/biofacquimnp/



CHEMICAL SPACE

DATA VISUALIZATION

Chemical space

Major concept in Chemoinformatics

'An M-dimensional Cartesian space in which compounds are located by a set of M physicochemical and/or chemoinformatic descriptors'

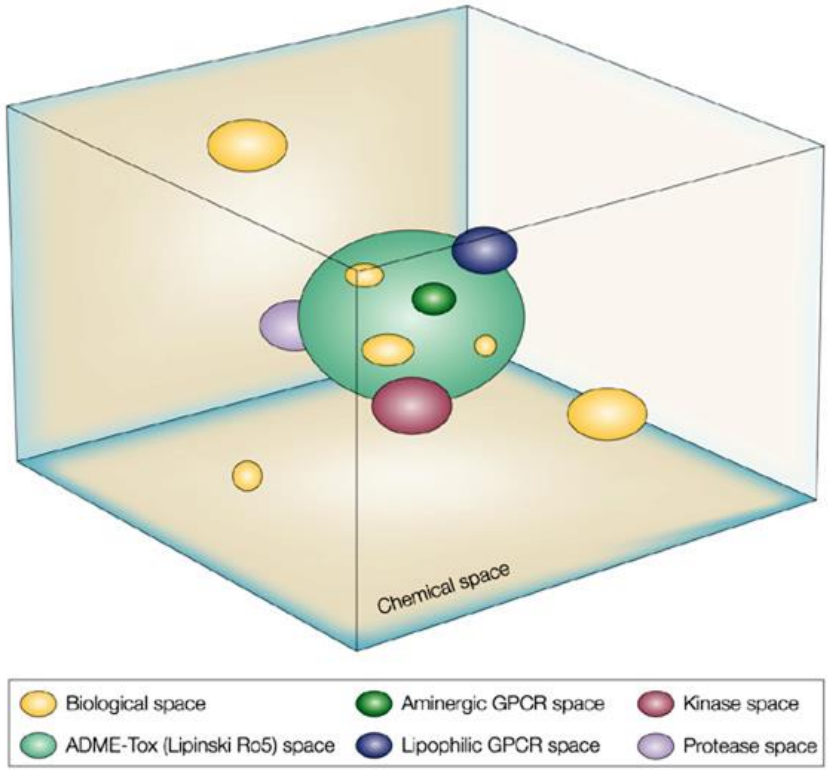
Virshup AM, Contreras-García J, Wipf P, et al. (2013) J Am Chem Soc *135:*7296-303

Other definitions collected in*:

- *"The total descriptor space that encompasses all the small carbon-based molecules that could in principle be created"* (Dobson)
- *"The set of all possible molecular structures"* (Horizon Symposia, Charting chemical space: finding news tools to explore biology")

*Medina-Franco JL. et al. *Curr. Comput.-Aided Drug Des. (*2008) *4:*322

How big is the chemical space



10^{60} (up to 30 C, N, O, S atoms)

Bohacek et al., *Med. Res. Rev.* 1996, *16*, 3

10^{18} - 10^{200}

Petit-Zeman, Horizon Symposia, 2004

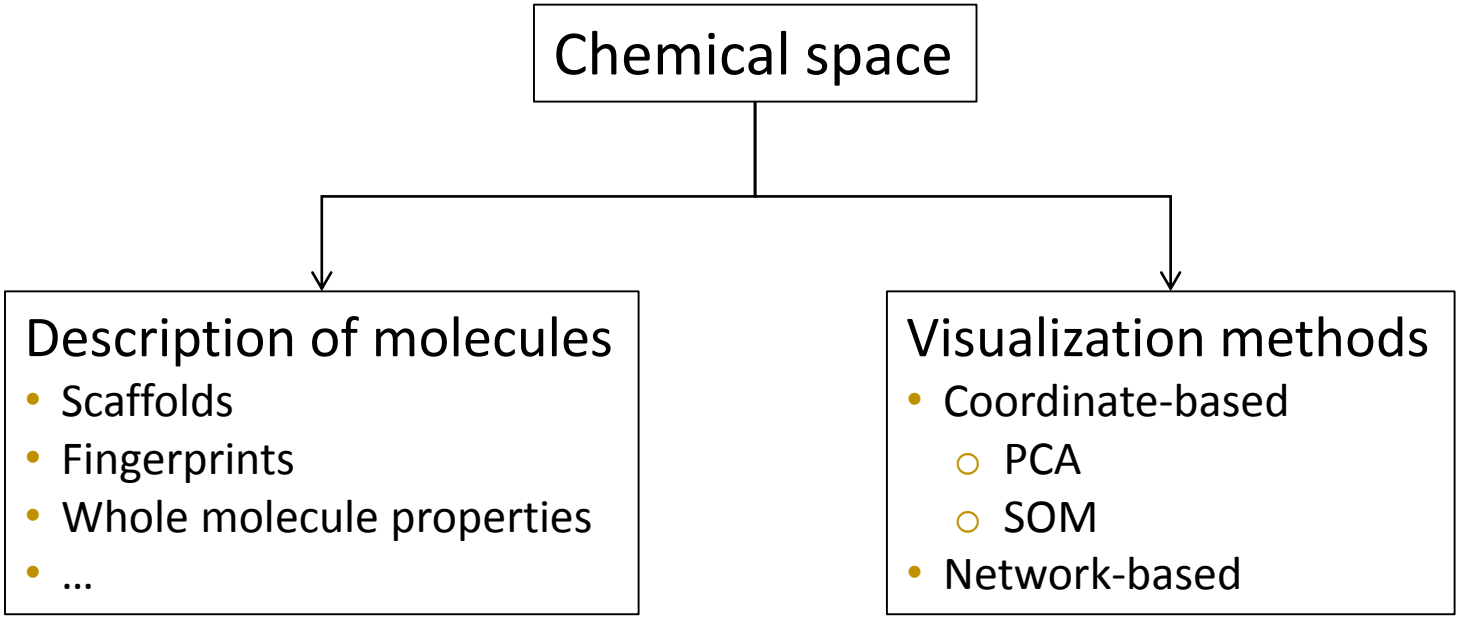
10^{14} - 10^{30}

Geysen et al., *Nat. Rev. Drug Discov.* 2003, *2*, 222

Linpinski & Hopkins, *Nature* (2004) *432*:855

Chemical Space

One of many definitions: 'an M-dimensional Cartesian space in which compounds are located by a set of M physicochemical and/or chemoinformatic descriptors'.



Virshup AM et al. *J Am Chem Soc* 2013 135:7296

Major applications of chemical space

- Compound library design
- Compound selection
- Compound classification
  - Pharmacological effect
  - Molecular target
  - Others

Medina-Franco JL. Martínez-Mayorga, K, Meurice, N. *Expert Opin. Drug Discov.* (2014) *9:*151

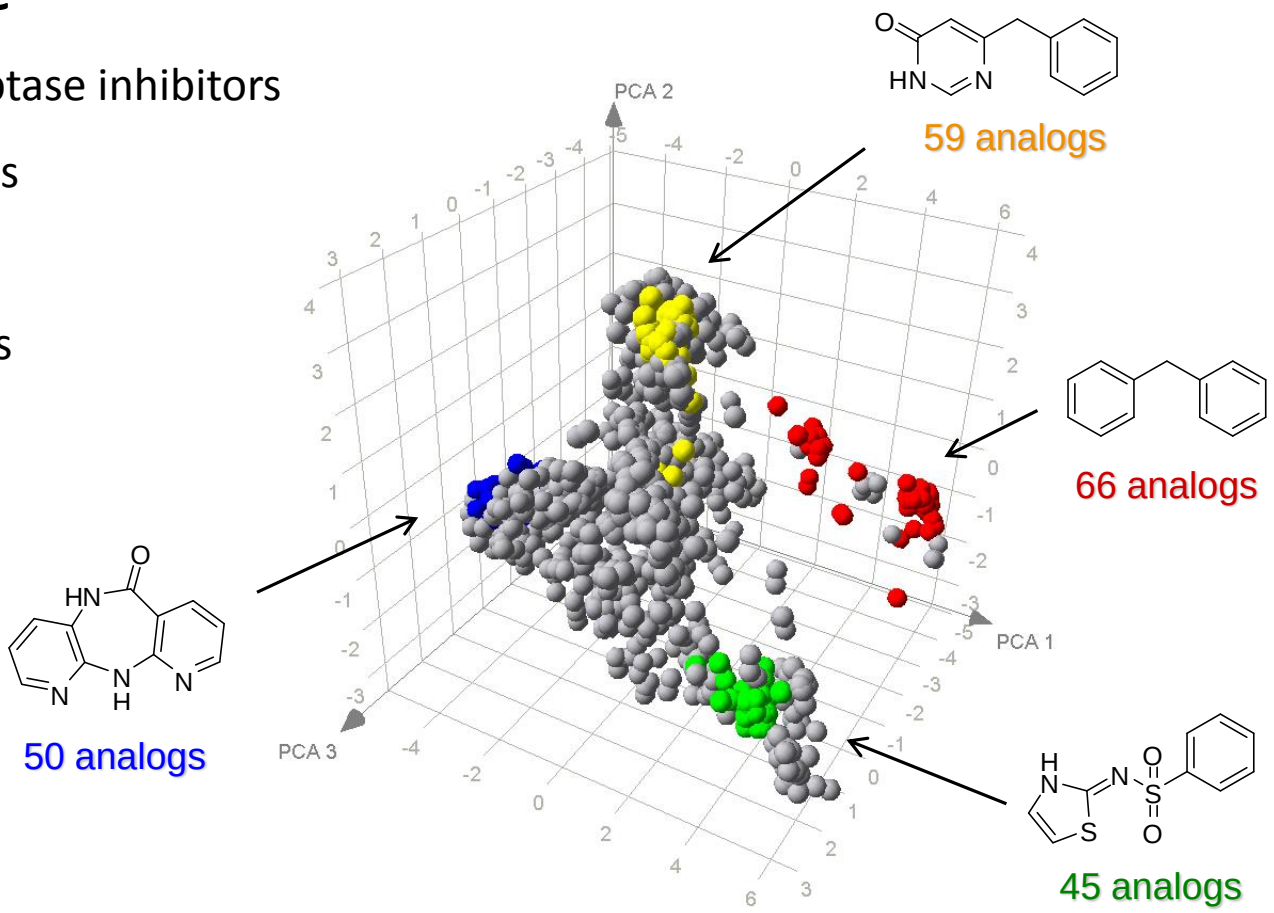
Chemical space of large databases

Binding Database

- Reverse transcriptase inhibitors
- 1,337 compounds

Chemotype analysis

- 353 cyclic systems

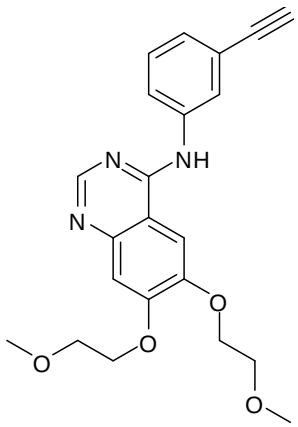


Representation of biological activity

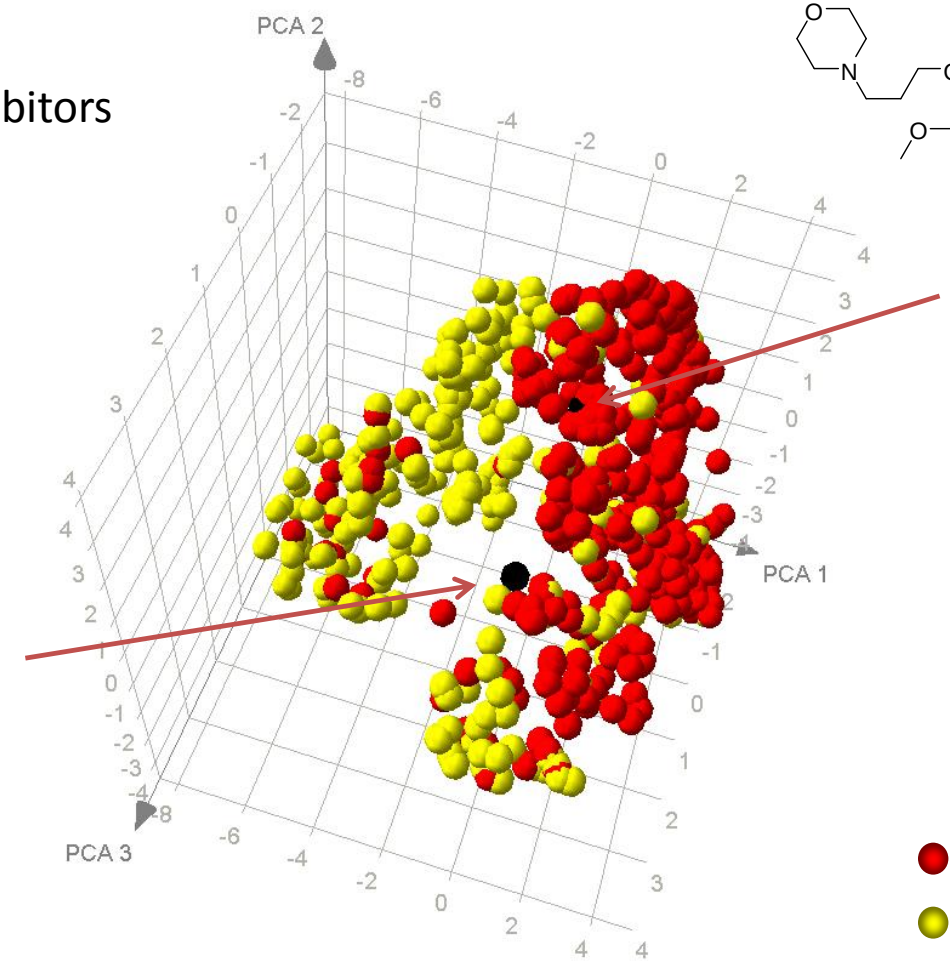
Binding Database

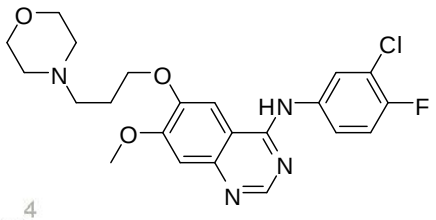
EGFR tyrosine kinase inhibitors

790 compounds



Erlotinib
(Tarceva)



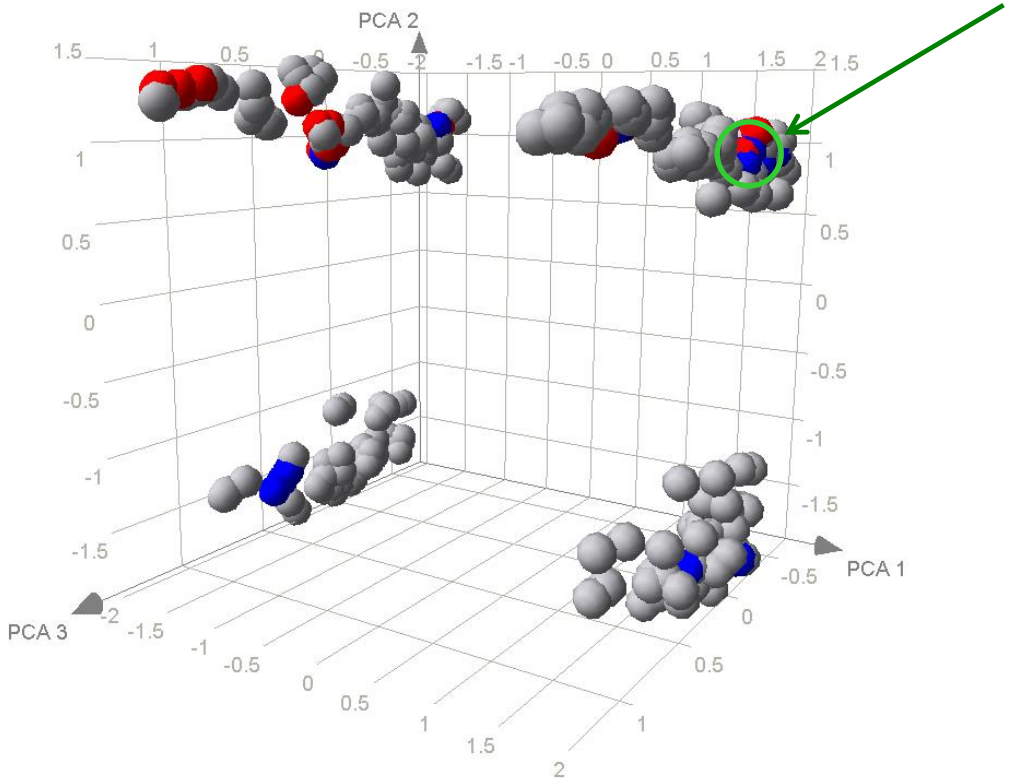


Gefitinib
(Iressa)

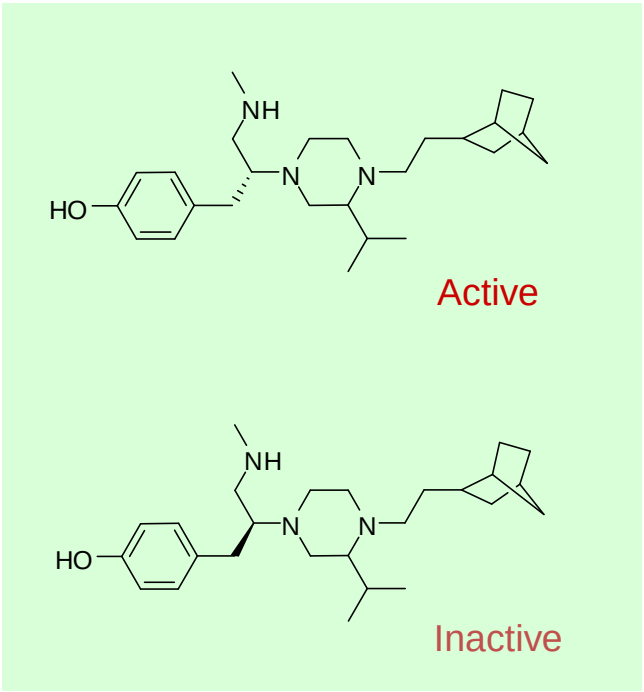
$IC_{50} \leq 1\ \mu M$

$IC_{50} > 1\ \mu M$

Activity Cliffs



High structural similarity but very different activity

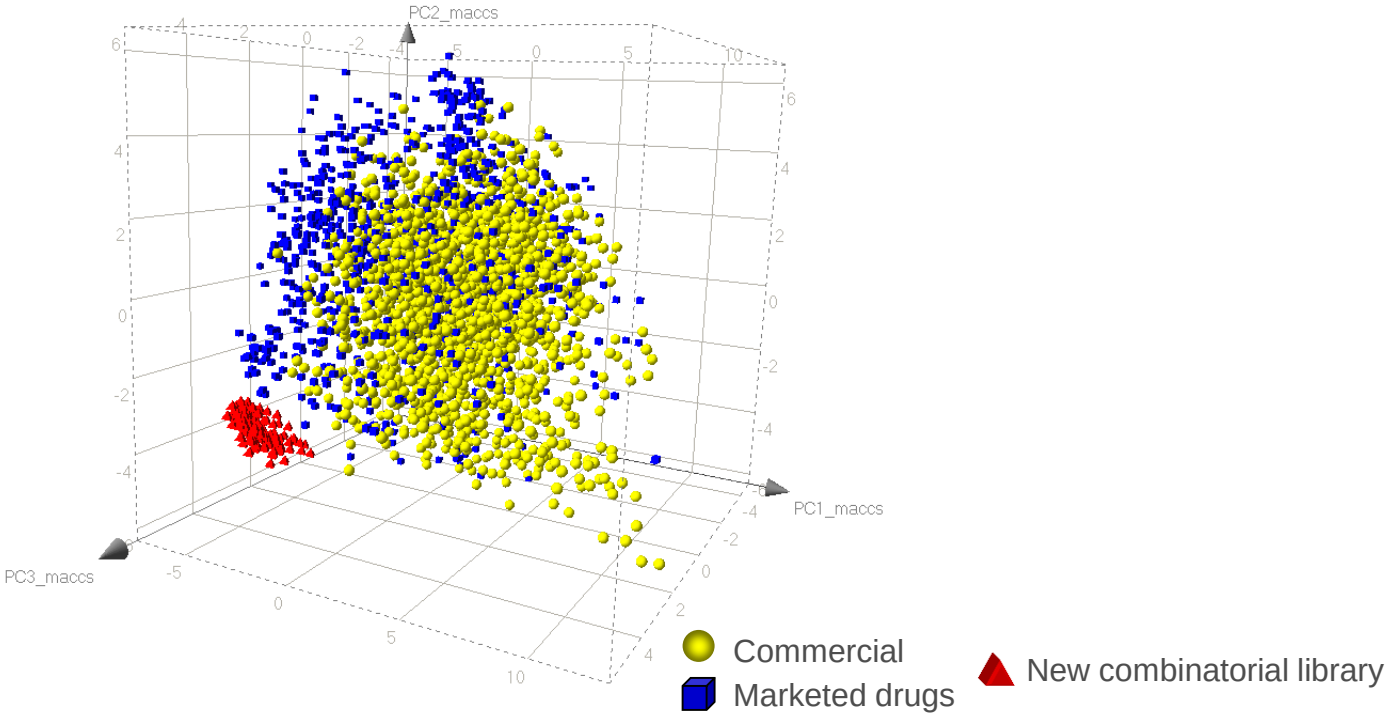


Maggiora GM. *J. Chem. Inf. Model.* (2006) *46:*1535

Classification of Molecular Structures

Chemical space and library design

Structural fragments



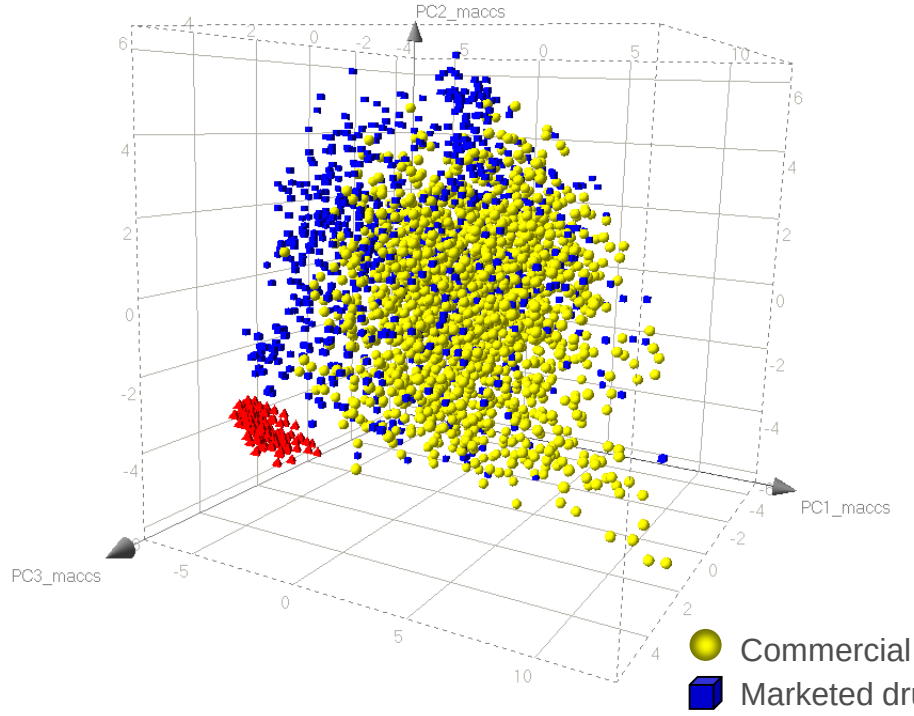
- Chemical structures of new library are novel

Medina-Franco, J.L. et al. *Curr. Comput.-Aided Drug Des.* (2008) *4:*322

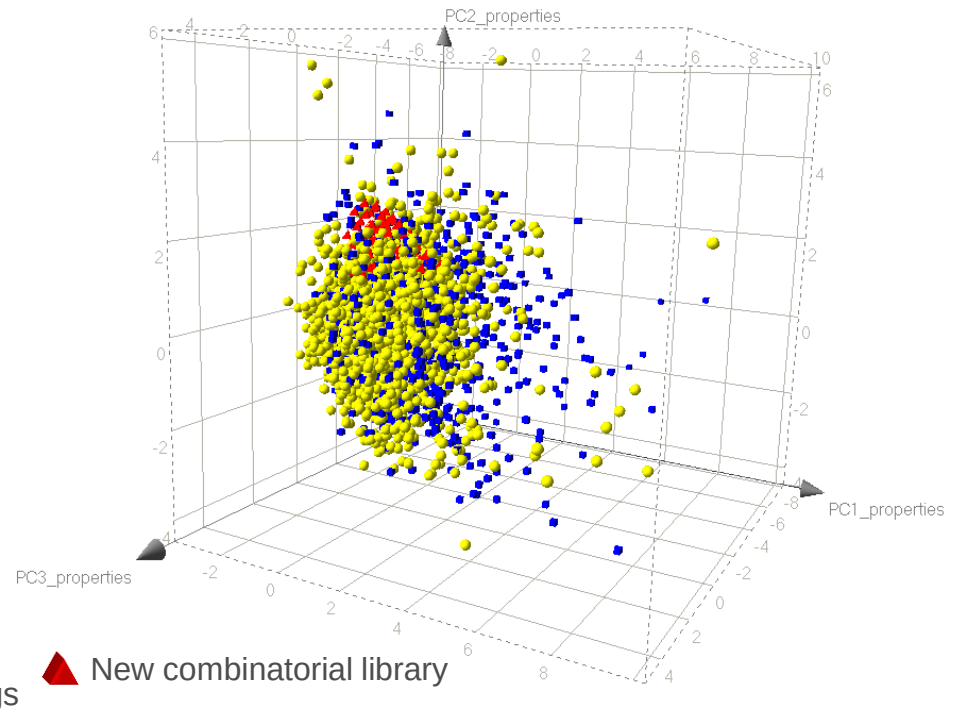
Classification of Molecular Structures

Chemical space and library design

Structural fragments



Physicochemical properties

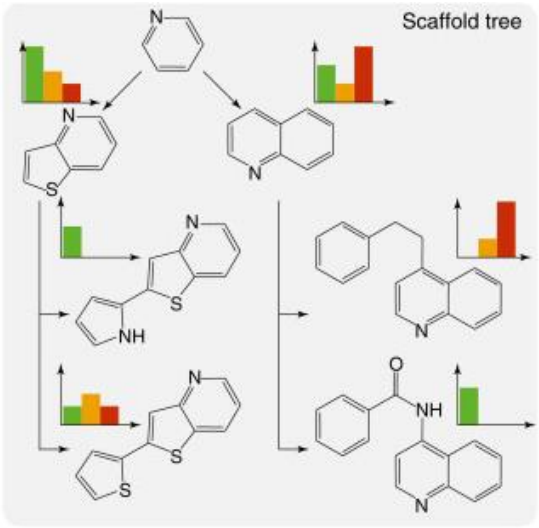


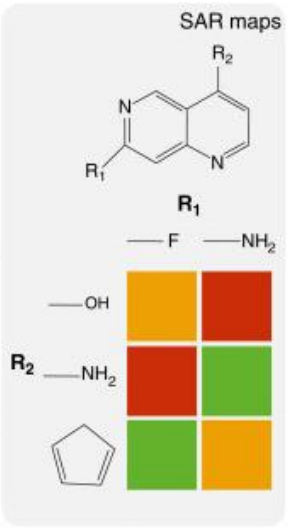
- Chemical structures of new library are novel
- New library is drug-like

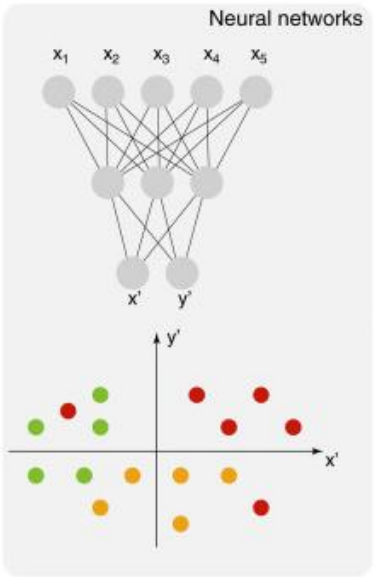
Medina-Franco, J.L. et al. *Curr. Comput.-Aided Drug Des.* (2008) *4*:322

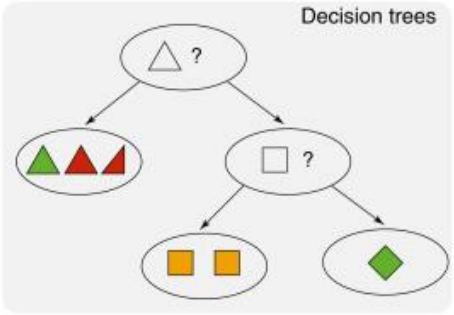
General approaches for SAR analysis

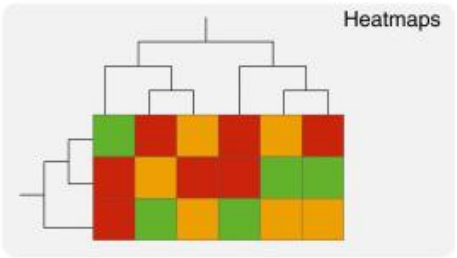
- Dimension reduction.
- Clustering and partitioning.
- Organization and annotation of substructures.
- Structural vs. activity similarity.

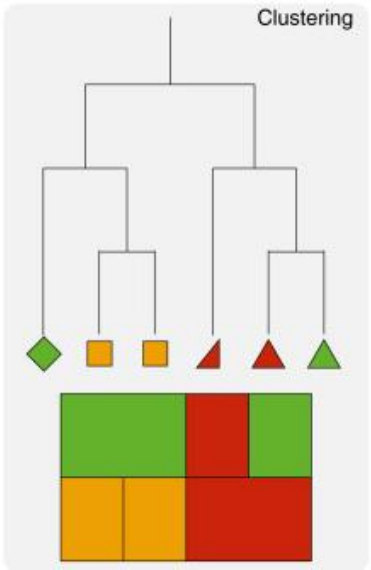


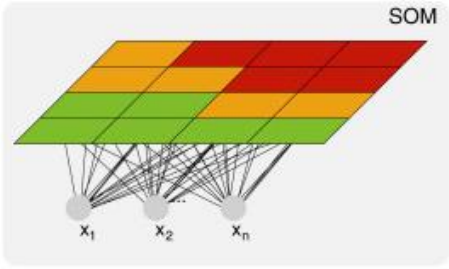


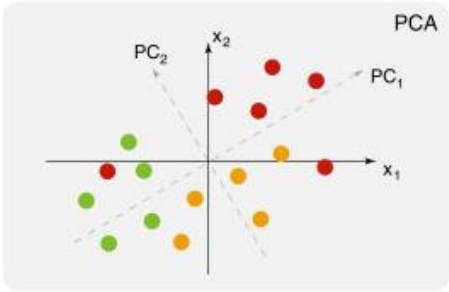












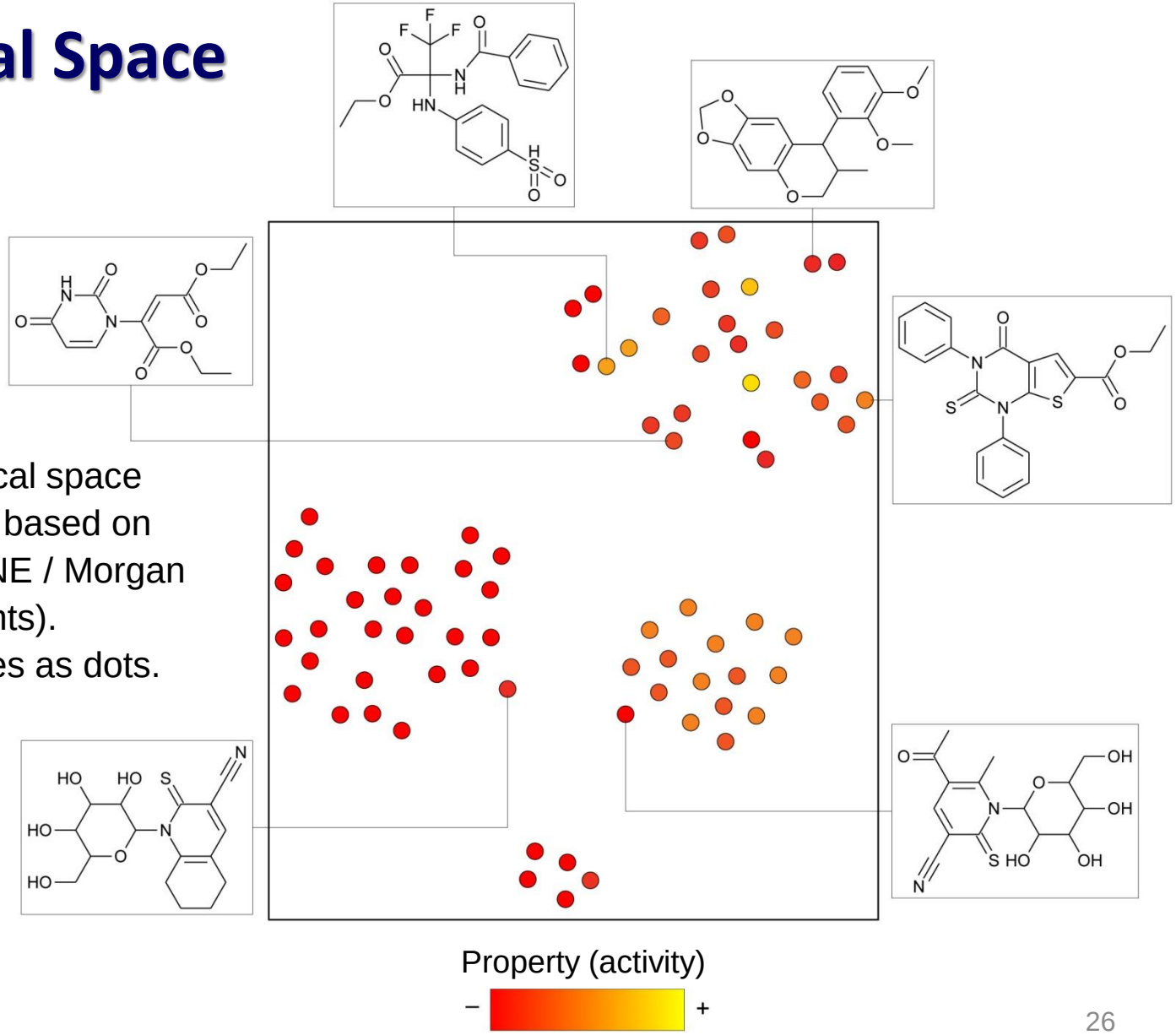
Wawer et al. *Drug Discovery Today* 2010 15:630

Implementations of Methods for SAR

Method	Ref.
SAR matrices (SARMs)	ACS Omega 4:7061
SAR maps and Enhanced SAR maps	J Med Chem 50:5926
SAR Analyzer	J Cheminform 5:31
Consensus Diversity Plots	J Cheminform. 8:63
Chem Maps	JCC 3:157
Shinyheatmap	PLoS ONE. 12:e0176334

Software	Ref.
DataWarrior	Exp Op on Drug Disc 14:335
Tableau	Tableau Soft. WA, USA
SARANEA	JCIM 50:16
Scaffold Hunter	J Cheminf 9 28
Spotfire	TIBCO Soft. CA, USA
Molecular Property eXplorer	JCIM 45: 523
LeadScope	JCICS 40:6
StarDrop	Optibrium. CA,UK
Miner3D	Addinsoft. PA, FR
SAR Report	MOE Soft. MO, CAN
Web-based applications	
VisualiSAR	JMGM 17:85
MOESaic	MOE Soft. MO, CAN

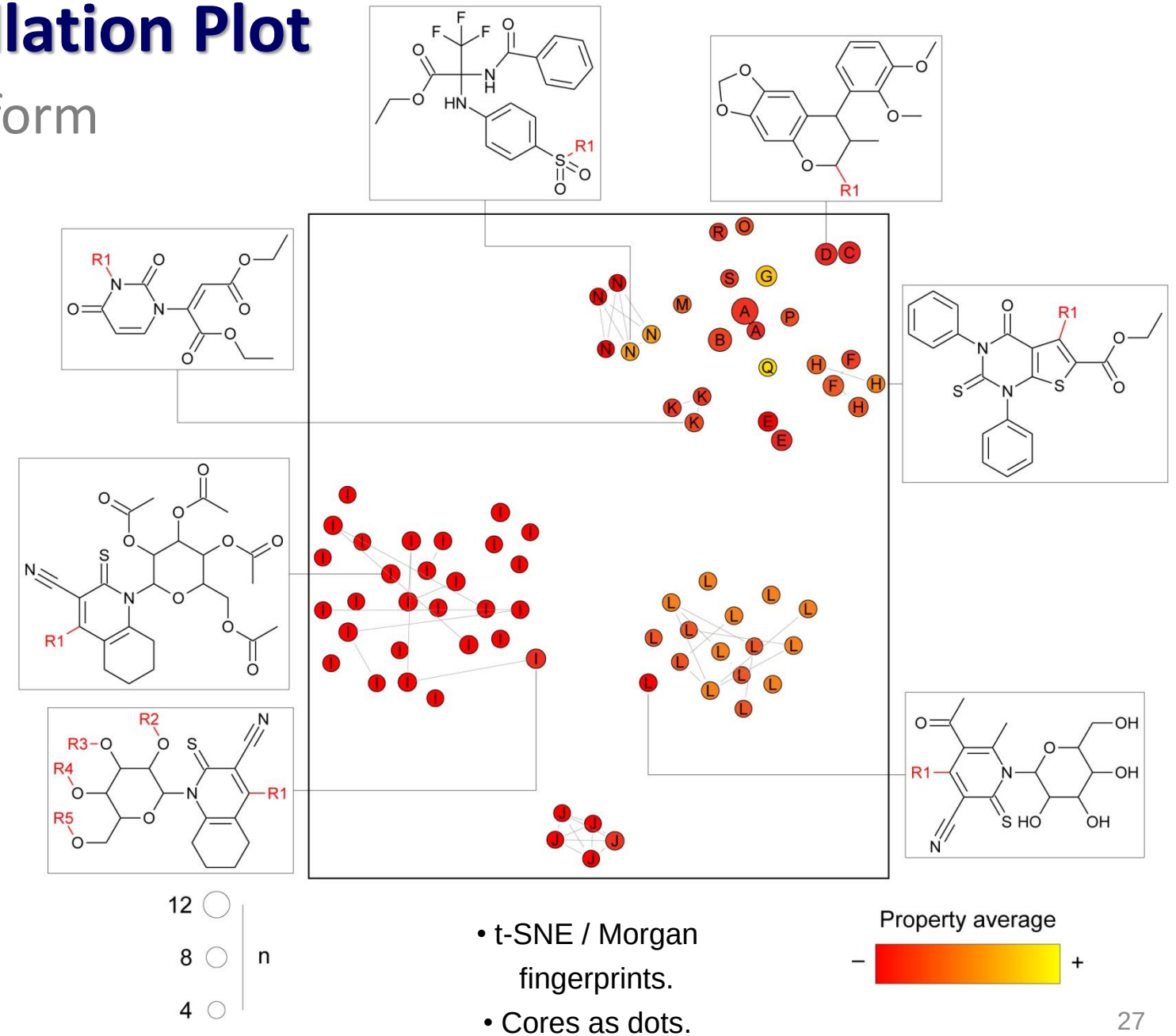
Visualization of SAR in Chemical Space



- Typical chemical space representation based on coordinates (t-SNE / Morgan fingerprints).
- Single molecules as dots.

Constellation Plot

General form



- t-SNE / Morgan fingerprints.
- Cores as dots.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 16 July 2019
doi: 10.3389/fchem.2019.00510

Check for updates

Finding Constellations in Chemical Space Through Core Analysis

DESCRIPTORS

Frequent approaches to describe molecules

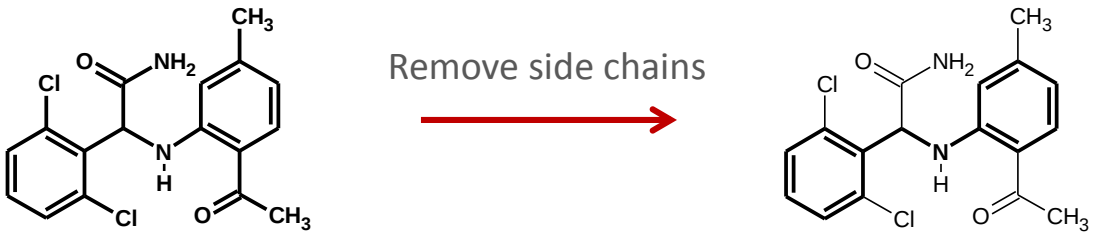
Whole molecule properties e.g., physicochemical properties

Pharmaceutical relevance: MW (~size), SlogP, HBA, HBD, TPSA (~polarity), rotatable bonds (~flexibility)

- ☺ Intuitive, basis of <u>empirical</u> rules: Lipinski, Verver's …
- ☹ Not specific, no direct information on the atom connectivity

Molecular scaffolds and sub-structural features

- ☺ Intuitive, SAR information, guide medicinal chemistry optimization
- ☹ Do not provide compete information, side chains (*decorations*) missing



Frequent approaches to describe molecules

Molecular fingerprints

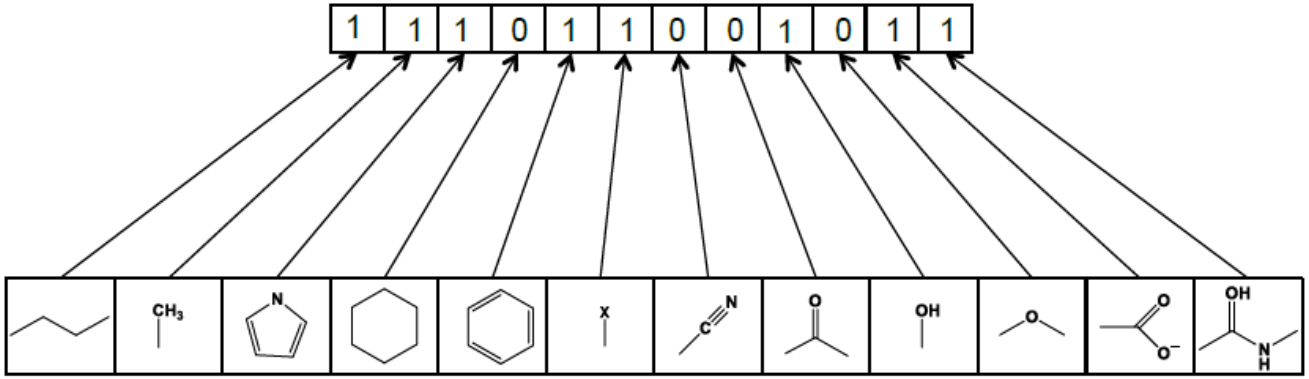
☺ Capture atom connectivity of entire molecules

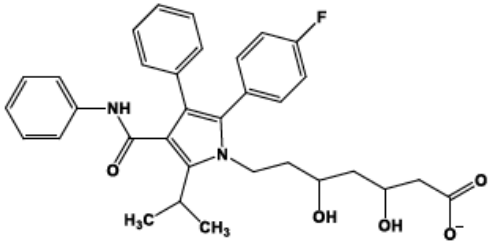
☹ Not necessarily intuitive; the 'best' fingerprint is problem - dependent

MACCS keys

Dictionary based

- Diccionary of structural fragments
- Size of the vector is fixed

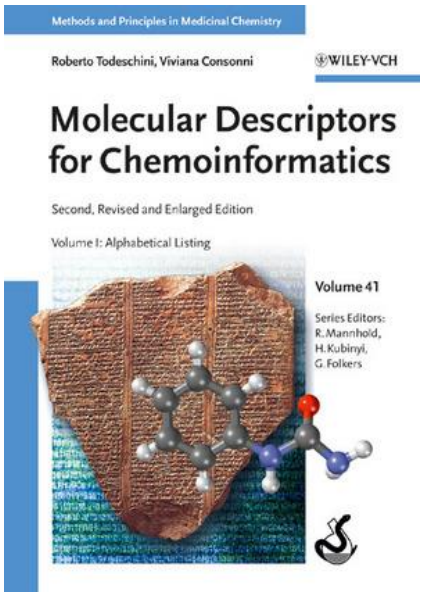




USE THE COMBINATION OF ALL 3 METHODS!

Representation of chemical structures

- Molecular fingerprints (binary)
  - 2D
    - Dicctionary-based: *MACCS keys*
    - Connectivity: *ECFP (radial)*
    - Pharmacophoric: *TGD*
  - 3D
    - Based on volume and shape
    - Pharmacophoric
- Properties (continuous values)
  - Physicochemical
  - Electronic …



Properties of interest: Empiric rules

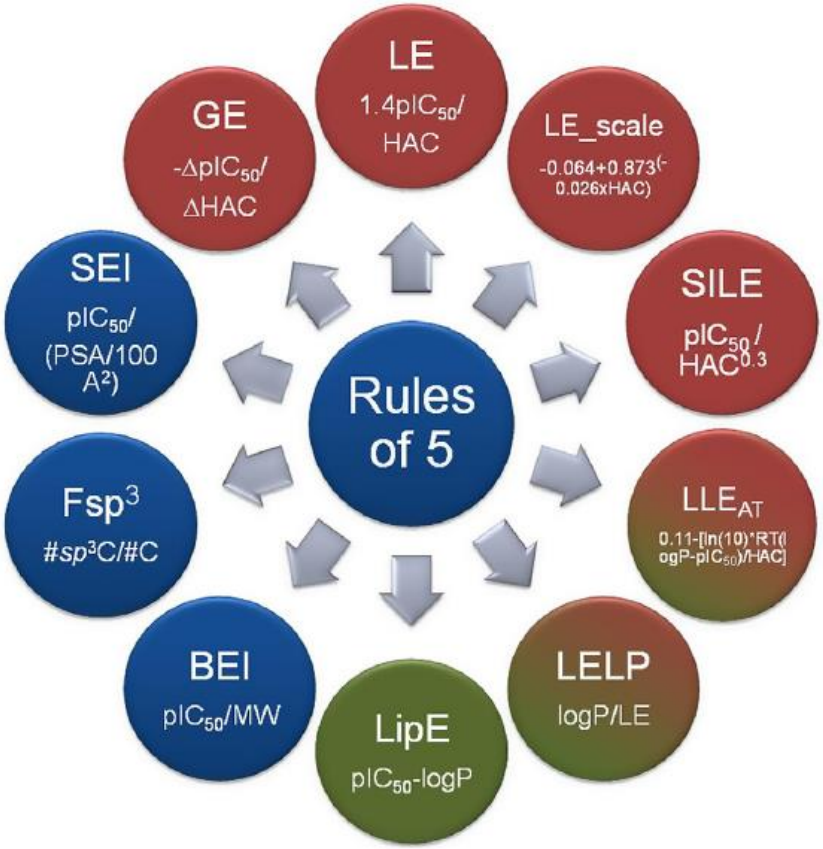


Figure 1. The 'Big Bang' of property based optimization spurred by the work of Lipinski and co-workers. Composite parameters dependent on heavy atom count are in red, log*P* are in green and those dependent on other parameters are in blue.

- *Molecular weight (MW)*
- *Lipophilicity*
- *ionization state,*
- *pKa,*
- *molecular volume,*
- *total polar surface area*
- *Aromatic rings,*
- *oxygen atoms, nitrogen atoms,*
- *sp3 carbon atoms, chiral atoms,*
- *non-hydrogen atoms,*
- *aromatic versus non-hydrogen atoms,*
- *aromatic atoms minus sp3 carbon atoms,*
- *hydrogen bond donors, hydrogen bond acceptors*
- *rotatable bonds*

Shultz MD (2013) *Bioorg Med Chem Lett 23*:5980-5991

Applications of physicochemical profile of compound databases

- Characterization and comparison of the chemical space of compound collections.
  - Ej. Aproved drugs, combinatorial libraries, synthetic compounds, natural products.
- Explore *Structure-Activity Relationships* (StARs).
- Assessment of compound novelty in chemical space.
- Filtering of compounds with right ADME/Tox properties.

Time-Related Differences in the Physical Property Profiles of Oral Drugs

Paul D. Leeson*,† and Andrew M. Davis‡

Department of Medicinal Chemistry and Department of Physical and Metabolic Science, AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, U.K.

Table 1. Mean (Median) Physical Properties of Oral Drugs Launched Pre-1983 and 1983−2002[a]

	oral drugs pre-1983[b] $n = 864$	oral drugs 1983−2002[c] $n = 329$	p pre-1983 vs 1983−2002[d]	difference in mean (median) values
Mol Wt	331 (310)	377 (357)	5.82×10^{-7}	14% (15%)
cLogP	2.27 (2.31)	2.50 (2.36)	0.17	10% (2%)
%PSA	21.1 (18.5)[e]	21.0 (19.4)	0.90	0% (5%)
OH + NH	1.81 (1)	1.77 (1)	0.35	−2% (0%)
O + N	5.14 (4)	6.33 (6)	5.65×10^{-8}	23% (50%)
HBA	2.95 (2)	3.74 (3)	1.34×10^{-7}	27% (50%)
RotB	4.97 (4)	6.42 (6)	2.20×10^{-8}	29% (50%)
Rings	2.56 (3)	2.88 (3)	1.18×10^{-4}	13% (0%)

[a] Mol Wt = molecular weight; cLogP = calculated 1-octanol/water partition coefficient (Daylight method); %PSA = calculated [(polar surface area)/(total surface area)] × 100; OH + NH = sum of OH + NH groups (H-bond donors); O + N = sum of O + N atoms; HBA = sum of H-bond acceptors; RotB = number of freely rotating bonds; Rings = number of rings in structure. All physical property values are taken from ref 8. [b] U.S. drugs (FDA approved to 2002) from Supporting Information in ref 8, excluding 83-02 NCEs. [c] Identified from oral drugs in Supporting Information in ref 8, using the 1983−2002 NCE list in ref 15. [d] Two-tailed, from two-sample t-test assuming unequal variances. [e] $n = 860$ (four compounds have missing values in ref 8).

REVIEWS Drug Discovery Today • Volume 16, Numbers 3/4 • February 2011

ELSEVIER

The impact of aromatic ring count on compound developability: further insights by examining carbo- and hetero-aromatic and -aliphatic ring types

Timothy J. Ritchie[1], Simon J.F. Macdonald[2], Robert J. Young[3] and Stephen D. Pickett[4]

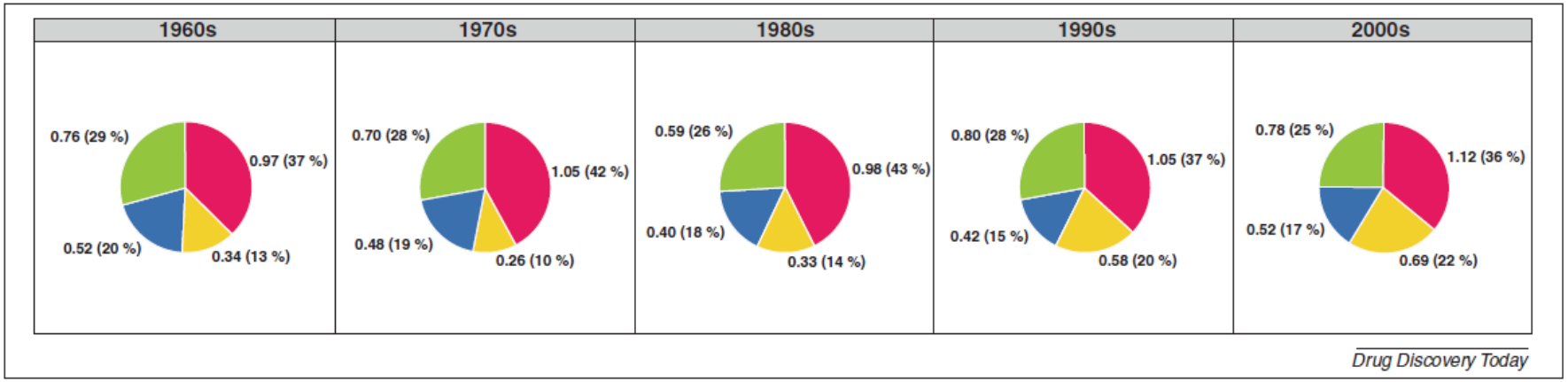


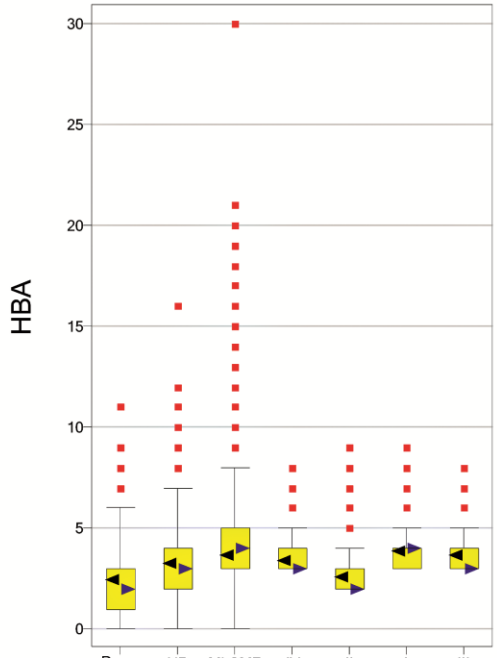
FIGURE 7

Increase in heteroaromatic ring count over time in marketed oral drugs. The heteroaromatic ring count (yellow segment) has increased from 0.34 rings in the 1960s to 0.58 and 0.69 rings in the 1990s and 2000s, respectively. Carboaromatic (red), carboaliphatic (blue) and heteroaliphatic (green) ring counts have remained relatively constant.

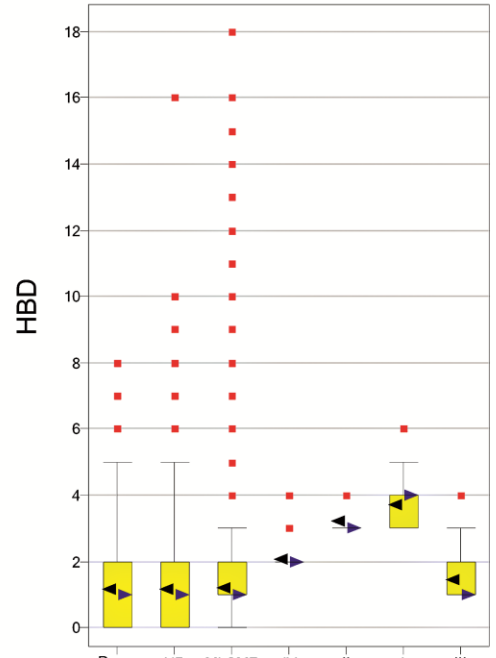
DIFACQUIM UNAM

Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository

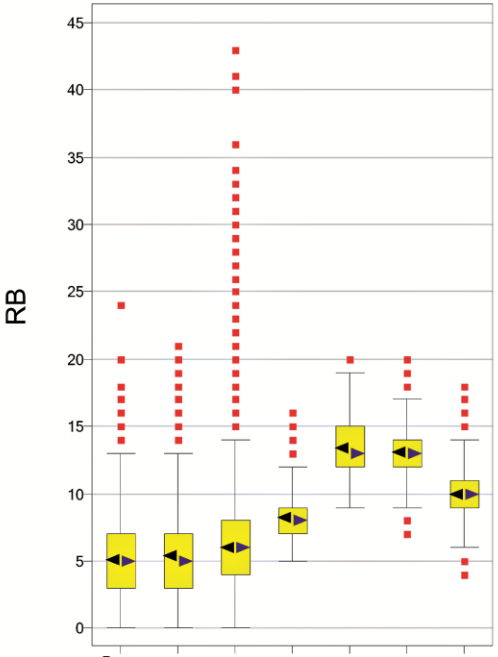
Narender Singh,[†] Rajarshi Guha,[‡] Marc A. Giulianotti,[†] Clemencia Pinilla,[§] Richard A. Houghten,[†,§] and Jose L. Medina-Franco*[,†,⊥]



	Drugs	NP	MLSMR	IV	II	I	III
Max	11	16	30	8	9	9	8
Median	2	3	4	3	2	4	3
Min	0	0	0	3	2	3	3
Mean	2.5	3.3	3.7	3.4	2.6	3.9	3.6
StdDev	1.7	1.6	1.5	0.71	0.93	0.94	0.81



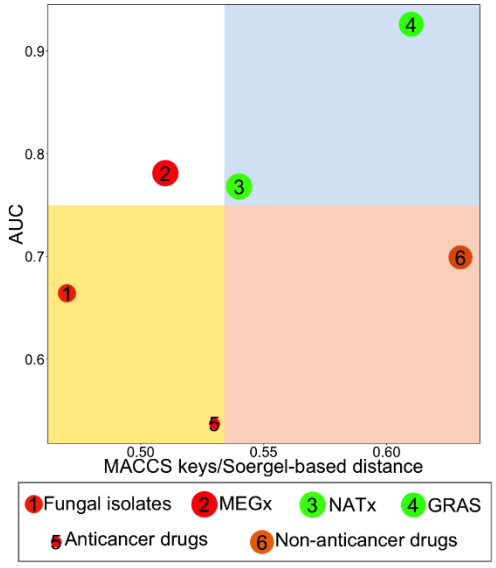
	Drugs	NP	MLSMR	IV	II	I	III
Max	8	16	18	4	4	6	4
Median	1	1	1	2	3	4	1
Min	0	0	0	2	3	3	1
Mean	1.2	1.2	1.2	2.1	3.2	3.7	1.5
StdDev	1.2	1.2	0.99	0.24	0.42	0.73	0.6



	Drugs	NP	MLSMR	IV	II	I	III
Max	24	21	43	16	20	20	18
Median	5	5	6	8	13	13	10
Min	0	0	0	5	9	7	4
Mean	5.1	5.4	6.1	8.3	13	13	9.9
StdDev	3.3	3	2.7	1.6	1.7	1.8	1.9

Diversity Analysis of Natural Products

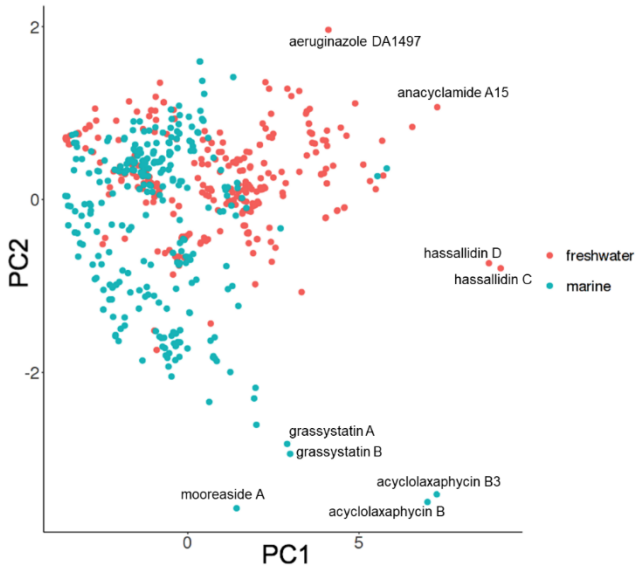
Fungi metabolites



Consensus Diversity Plot

Fut Med Chem 2016:8 1399
Front Pharm 2017:8 180

Cyanobacteria metabolites

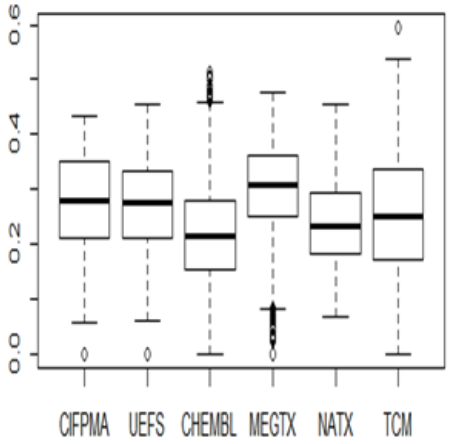


Chemical space

ACS Omega 2019:4 6229

Colaboration: M. Figueroa (UNAM)
N. Oberlies (UNC-Greensboro)

Natural products from Panama



Complexity (Fsp^3) of 324 compounds (CFPMA)

Mol Diversity 2017:21 779

M. Gupta & D. Olmedo
(University of Panama)

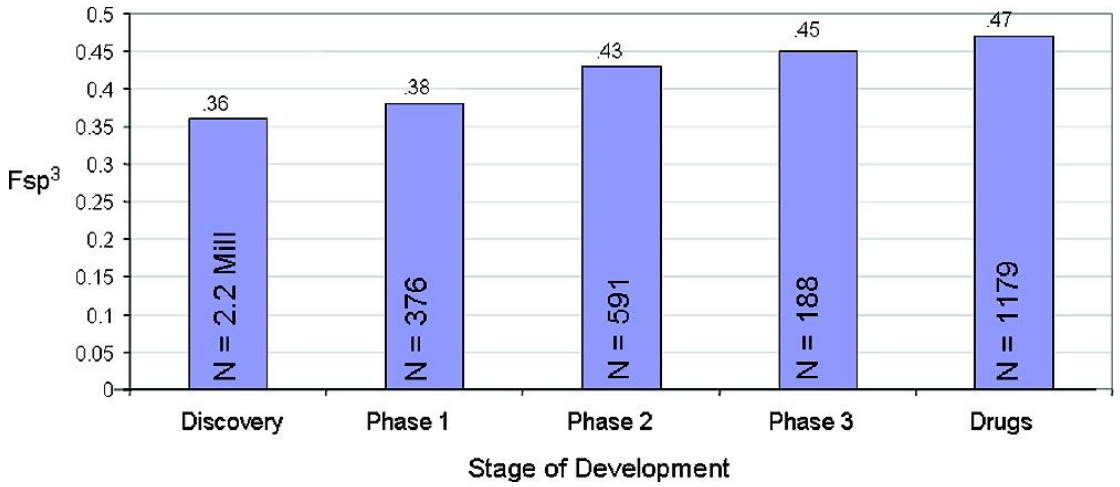
Molecular Complexity

Two simple metrics commonly used (of many more):

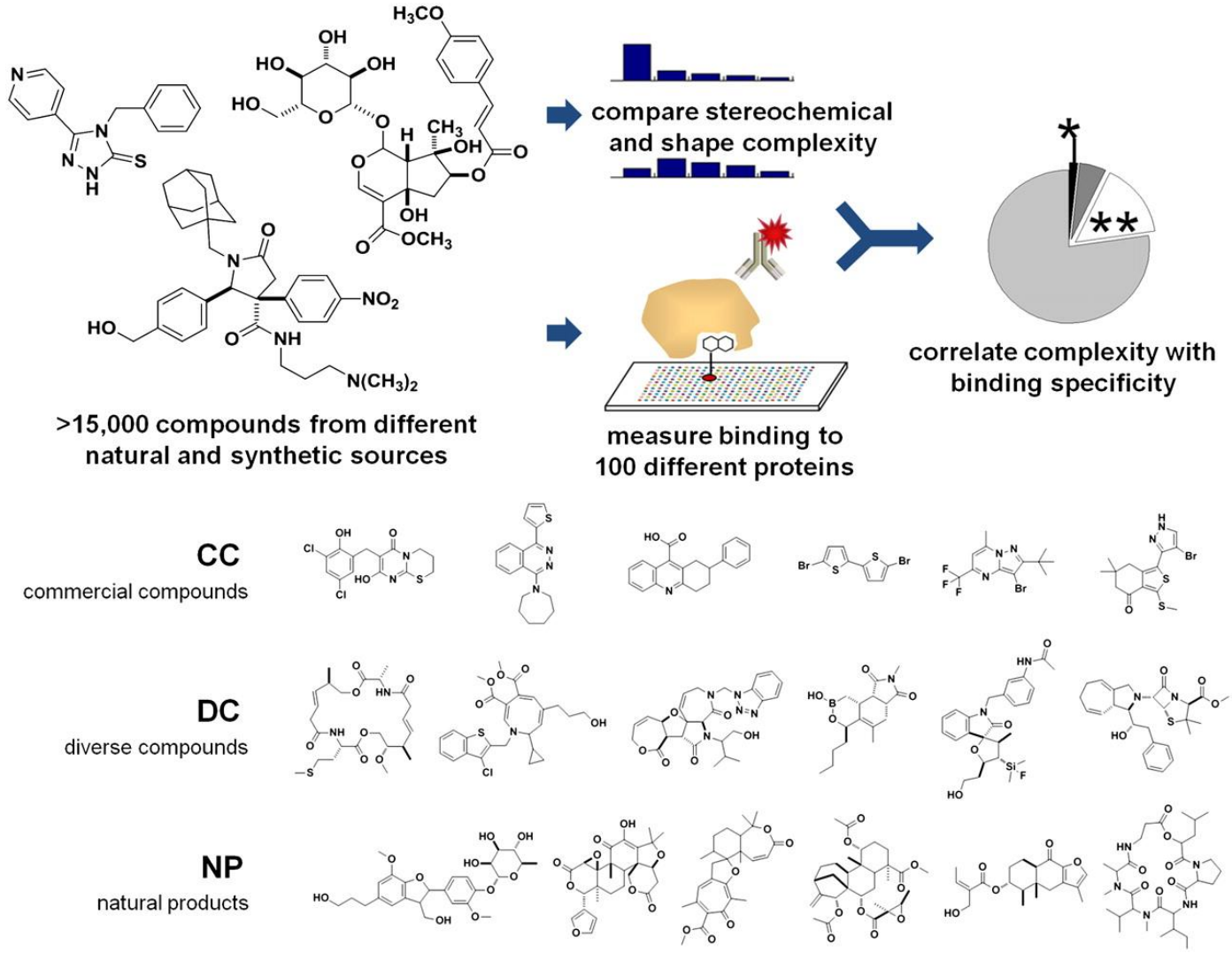
- Fraction of sp^3 carbon atoms (F-sp^3)
- Fraction of chiral carbon atoms

Increased stereochemical content has been associated with Successful progression through clinical trials

Lovering F et al. J Med Chem (2009) 52:6752

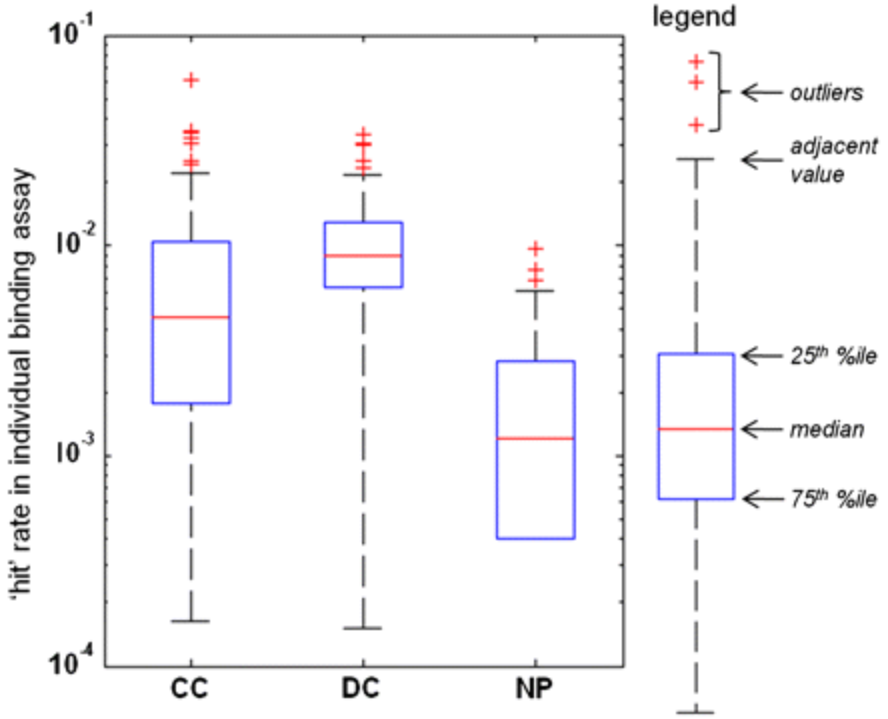


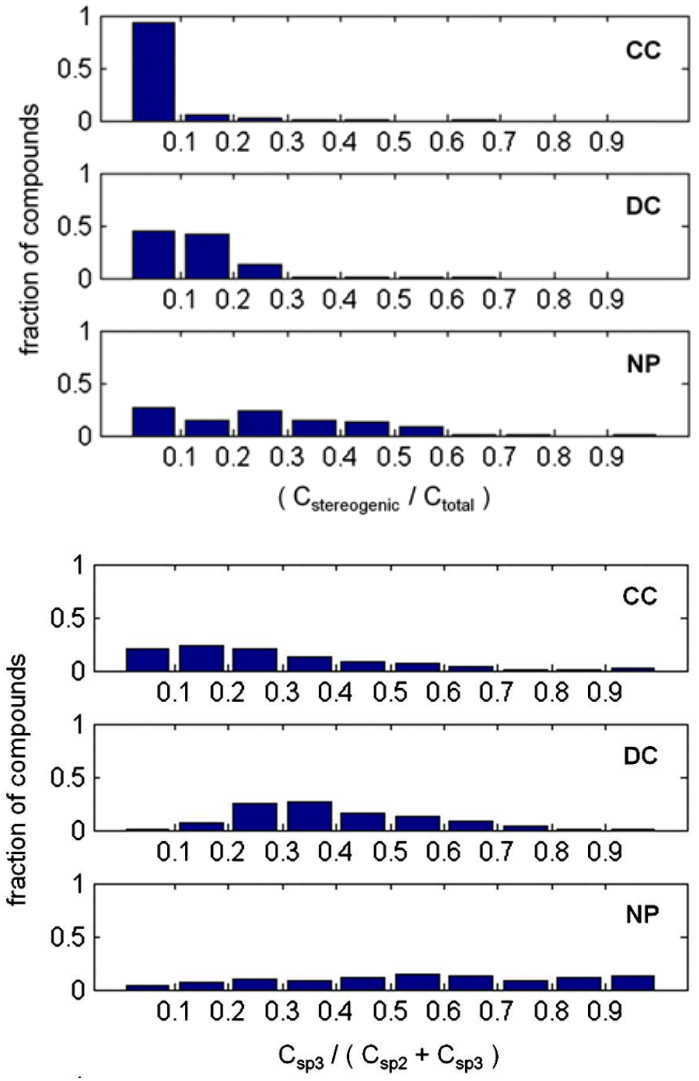
Is Molecular Complexity Related to Selectivity?



Clemons PA et al. Proc Natl Acad Sci USA (2010) 107:18787

Is Molecular Complexity Related to Selectivity?





YES!

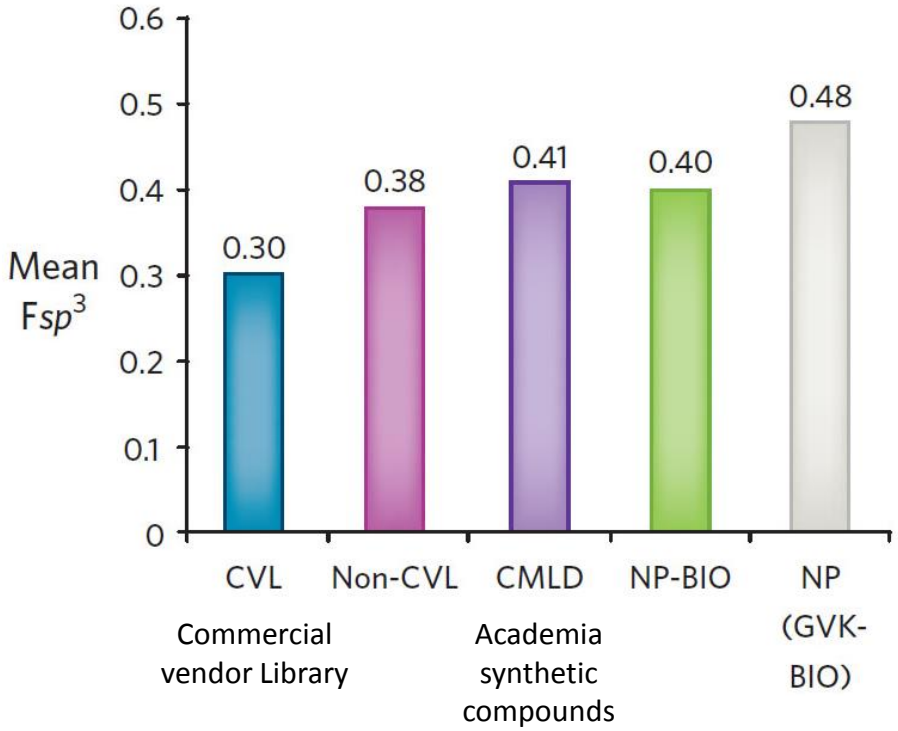
Increased stereochemical content has been associated with Improved binding selectivity

Clemons PA et al. Proc Natl Acad Sci USA (2010) 107:18787

Molecular Complexity

F-sp^3 has been used to illustrate that natural products are more complex than the synthetic drug-like compounds found in commercial screening libraries

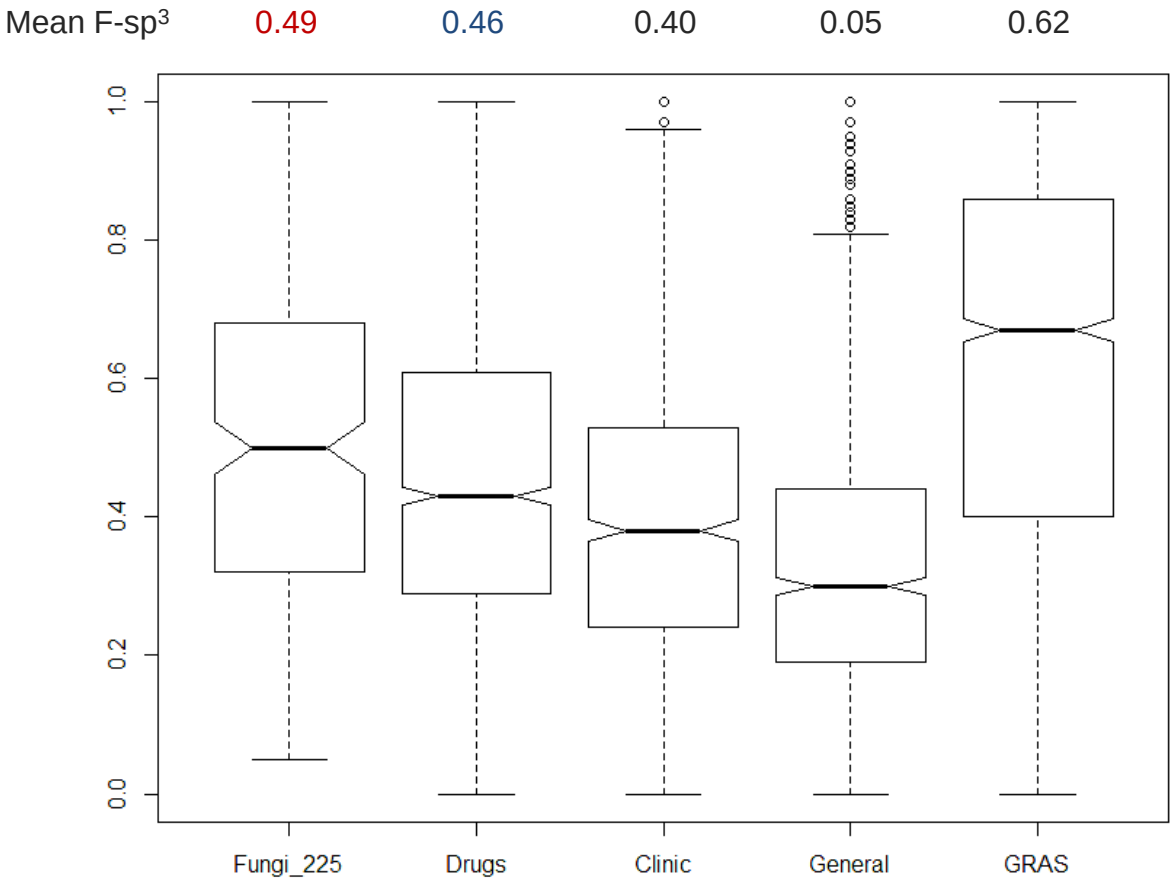
Dandapani S, Marcaurelle LA. Nat Chem Biol (2010) 6:861



What are the F*sp*^{3} values of fungal isolates?

Molecular Complexity of Fungal Isolates

Distribution of F-sp^3



- Approved drugs are ***less flat*** than clinical
- Drugs and clinical are more complex than general screening

Fungal isolates

- More complex than general screening collection and approved drugs.
- Potentially with increased selectivity.

REVIEWS Drug Discovery Today • Volume 17, Numbers 13/14 • July 2012

ELSEVIER

Reviews • INFORMATICS

Expanding the medicinally relevant chemical space with compound libraries

Database

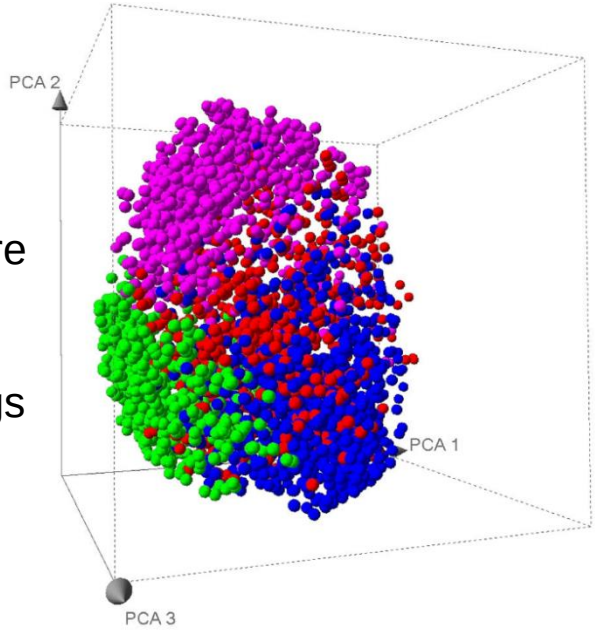
Public: Traditional Chinese Medicine

Representation

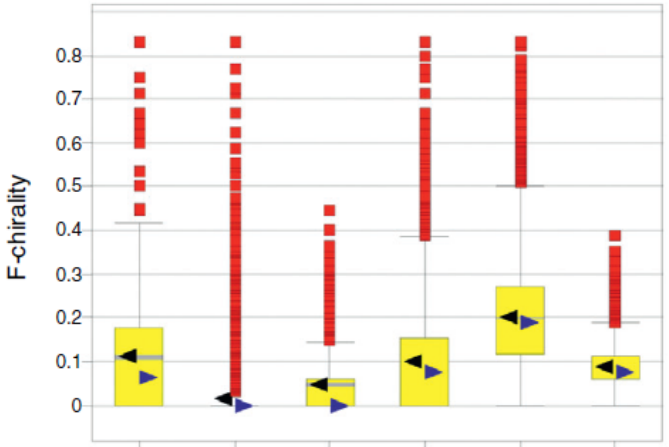
Fingerprints (MACCS keys), physicochemical properties; atom counts

Visualization method: PCA

Metrics: Counts, distribution, Tanimoto



Natural products share and expand chemical space of drugs



	DrugBank	Maybridge	NCI	NP	TCM	TPISR
Min	0.00	0.00	0.00	0.00	0.00	0.00
Q1	0.00	0.00	0.00	0.00	0.12	0.06
Median	0.07	0.00	0.00	0.08	0.19	0.08
Q3	0.18	0.00	0.06	0.15	0.27	0.11
Max	0.83	0.83	0.44	0.83	0.83	0.39
Mean	0.11	0.01	0.05	0.10	0.20	0.09
StdDev	0.13	0.05	0.09	0.10	0.12	0.04

NP (TCM) have more chiral centers than drugs

- Traditional Chinese Medicine (TCM) database
- Drugs
- General screening collection (commercial compounds)
- Combinatorial libraries

DIFACQUIM UNAM