DATA MINING FOR DRUG DISCOVERY, EXPLORING THE UNIVERSES OF CHEMISTRY AND BIOLOGY

BIOTECHNO 2015
Rome

Modest Korff, Thomas Sander

ACTELION

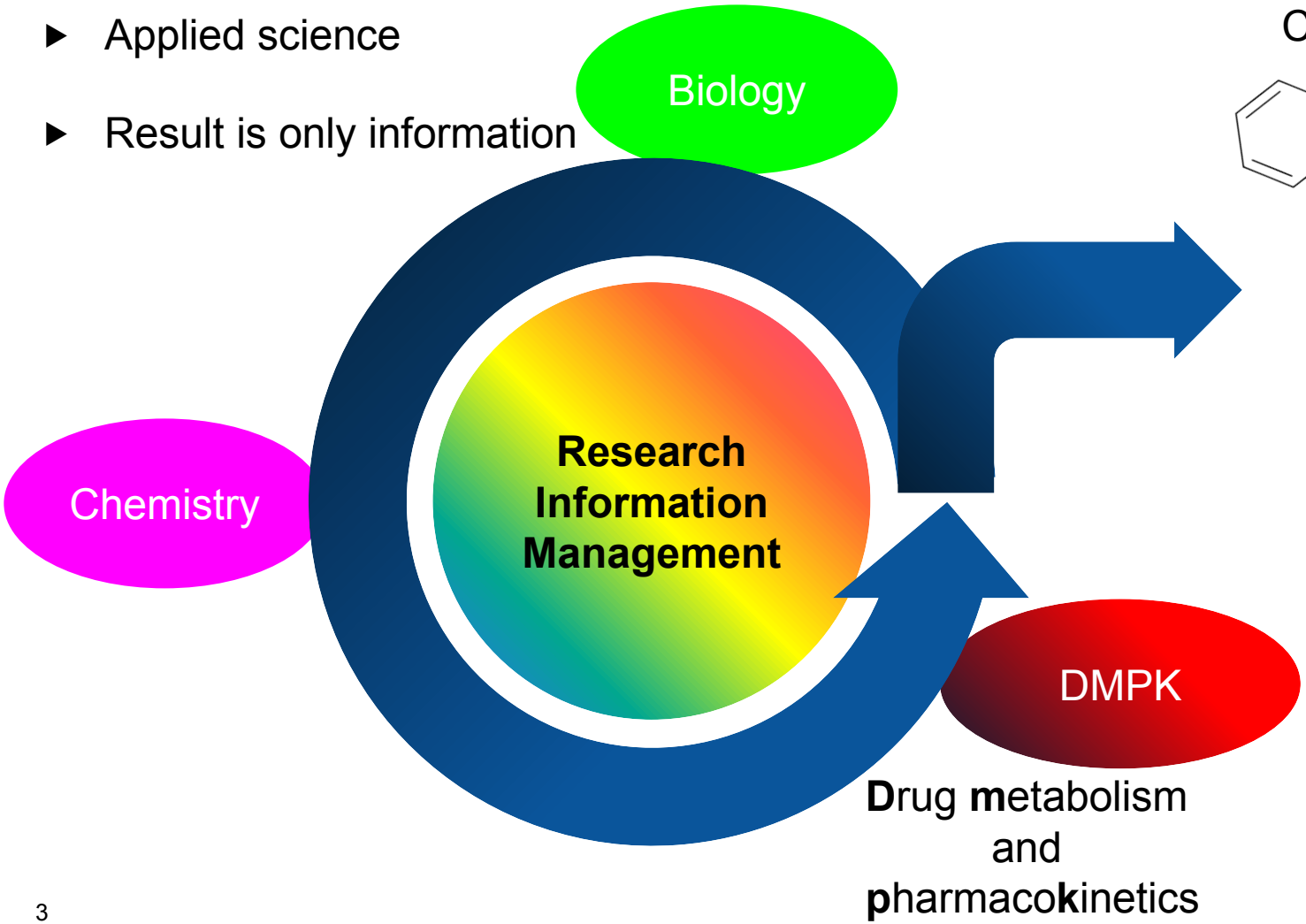
TOPICS

- Drug discovery
- Data mining chemistry
  - Molecular complexity
- Data mining biology
  - Gene2Disease
- What next?

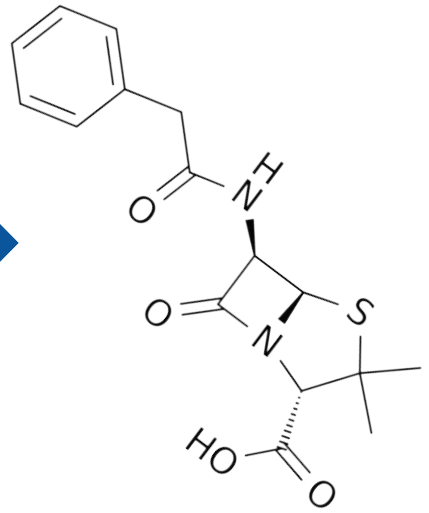
DRUG DISCOVERY

Goal: Deliver clinical candidate structures

- Needle in haystack
- Applied science
- Result is only information

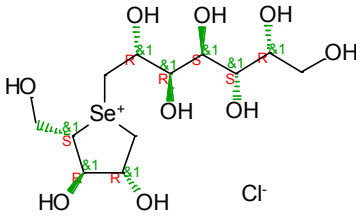


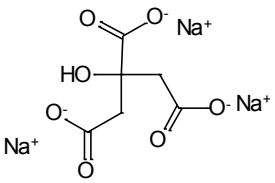
Clinical candidate



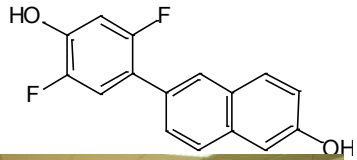
CHEMISTRY

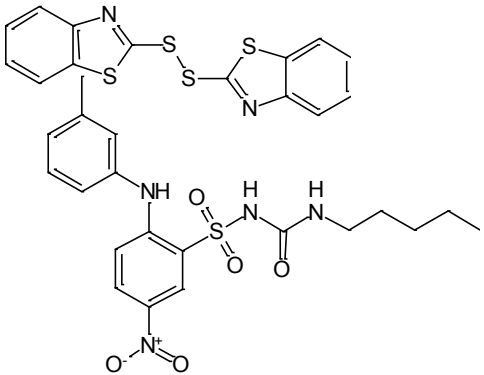
- Chemists feed chemical structures into process





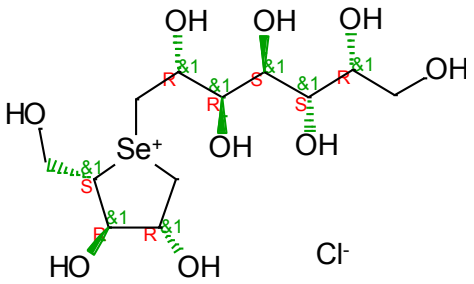
racemate

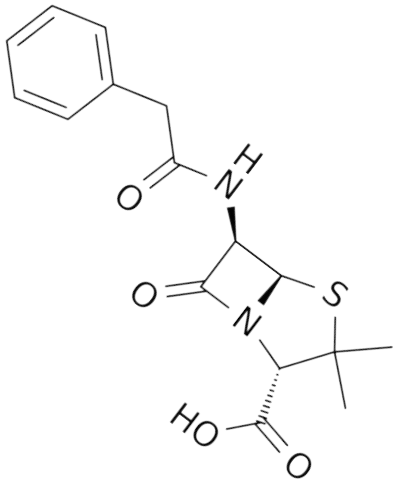


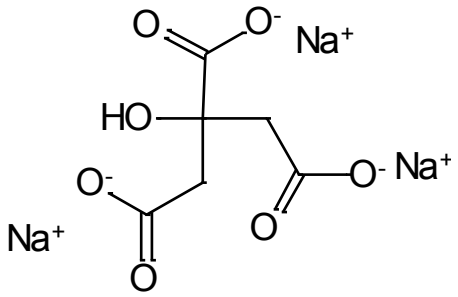


ACTELION

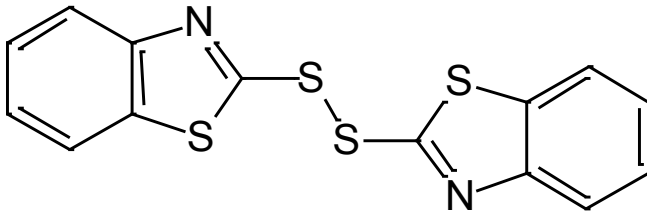
CHEMICAL SPACE

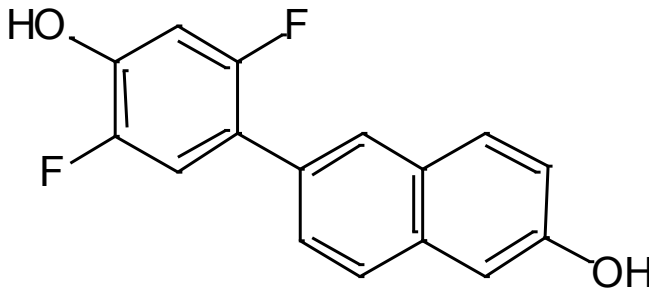


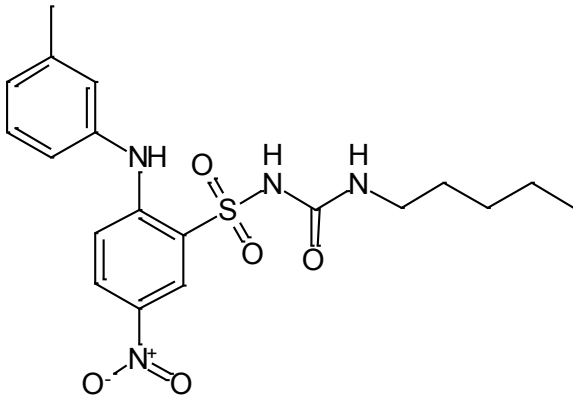


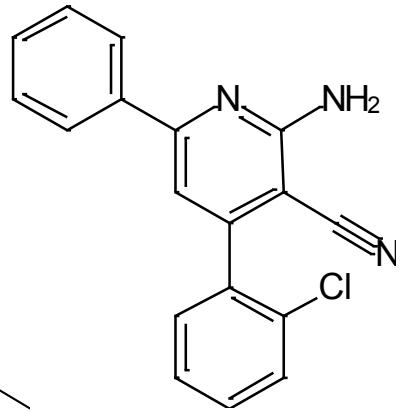


What is the source of chemical structures?









ACTELION

Chemical Space

Universe of chemical structures

Total 10^{60} structures

Described by PubChem 50 Mio.

Commercially available: 8 Mio.

In-house 500,000

Chemist / week: 1-100

WHAT TO MINE IN THE CHEMICAL SPACE?

Chemists

gut feeling for

complexity

Calculate

the information content

of a chemical structure

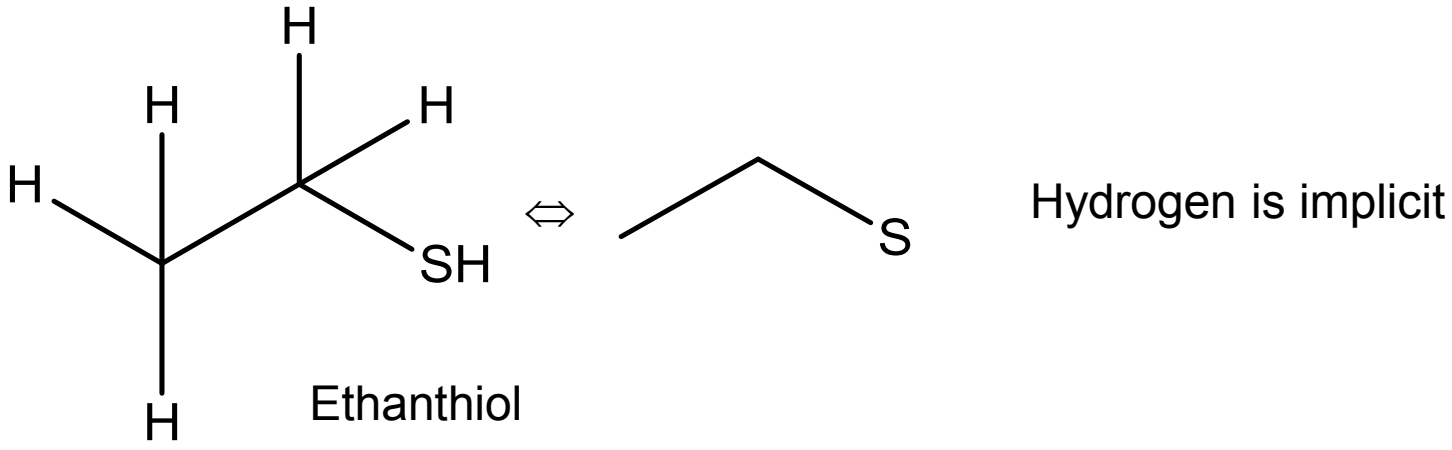
ACTELION

Molecular graphs

Graph	Molecule
Vertex	Atom
Edge	Bond

Atoms: H, C, O, N, S, P, F, Cl, Br

Bonds: single, double, triple, delocalized, up, down

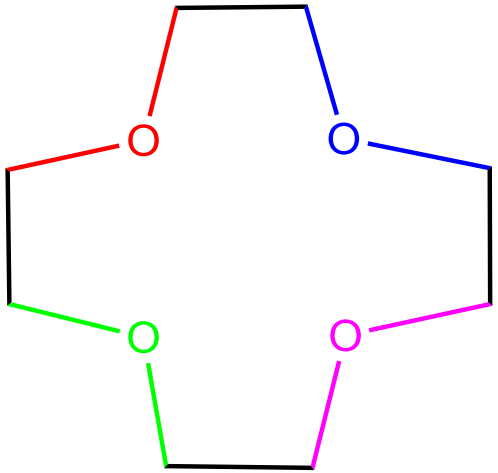


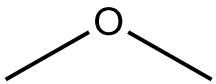
Ethanthiol

Hydrogen is implicit

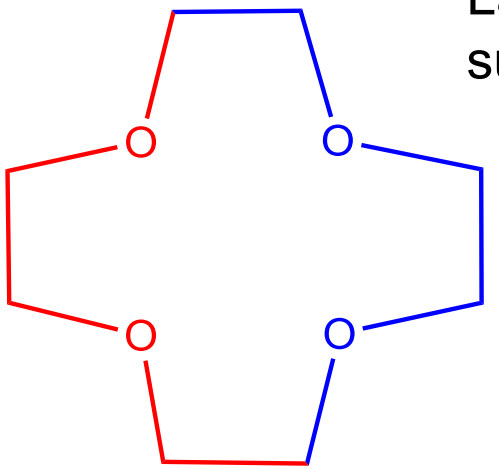
SYMMETRY MEANS REDUNDANCY

Isomorphic, non overlapping subgraphs

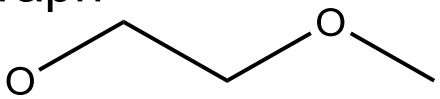




- Subgraph edges: 2
- Frequency: 4
- Ratio bonds covered ≈ 0.7



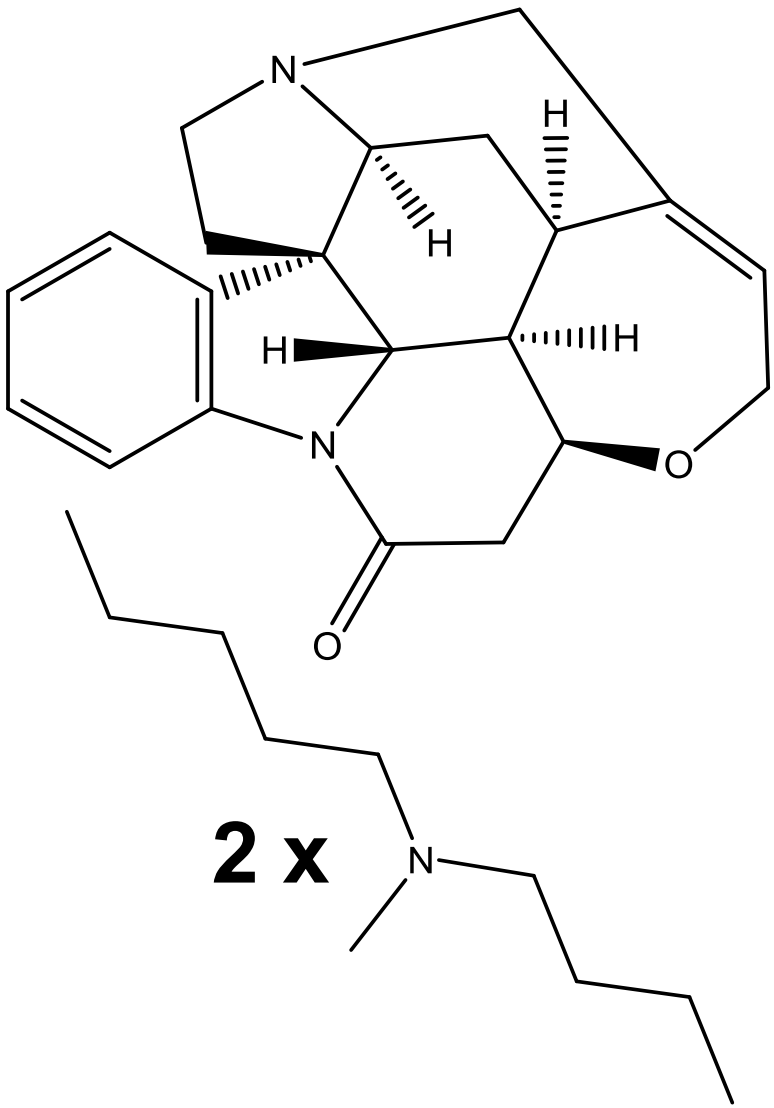
Largest possible isomorphic non overlapping subgraph



- Subgraph edges: 4
- Frequency: 2
- Ratio bonds covered: 1.0

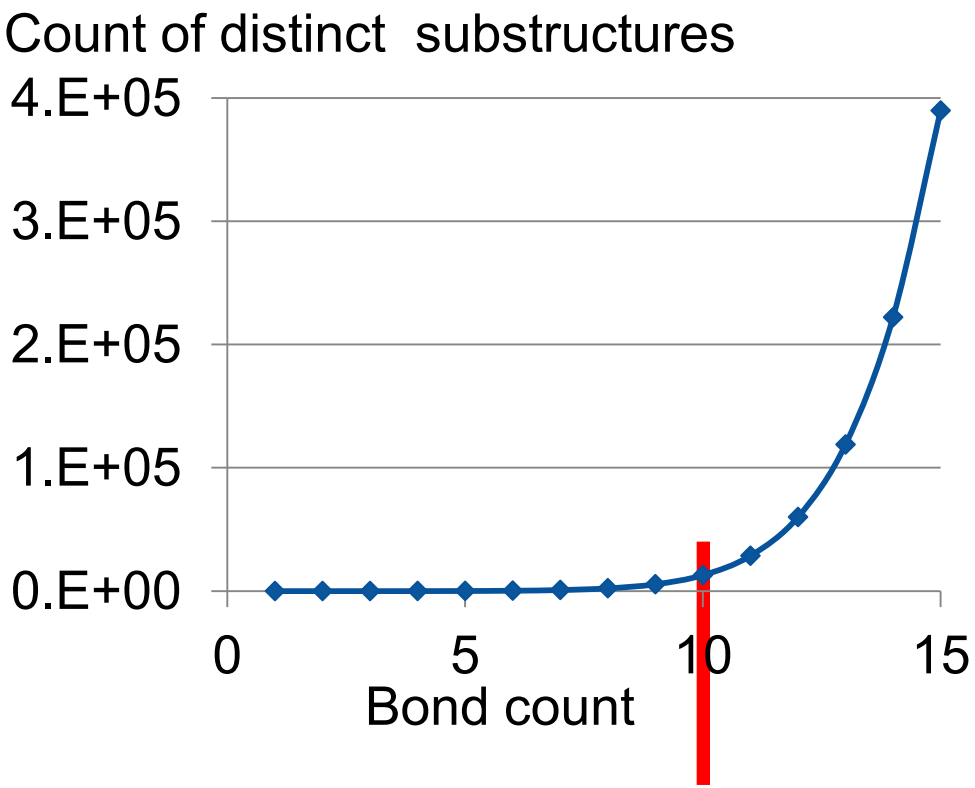
ACTELION

STRYCHNINE I

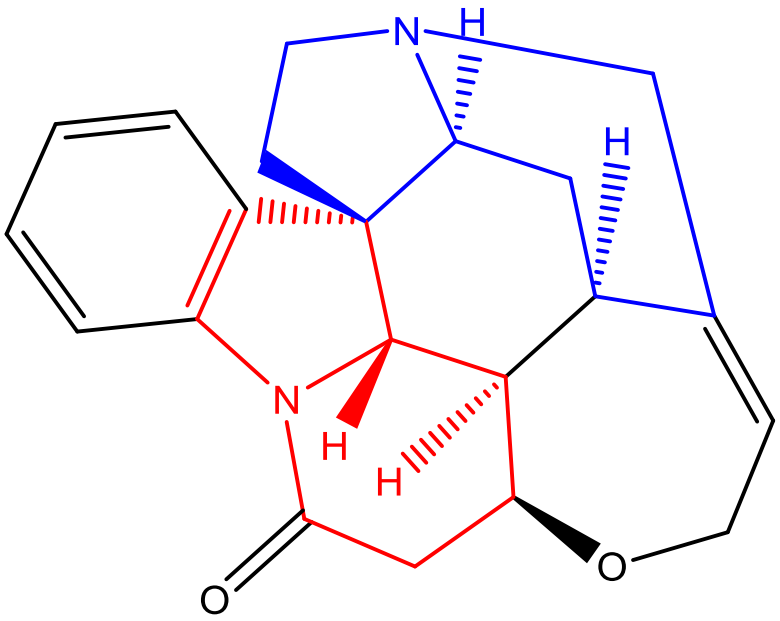


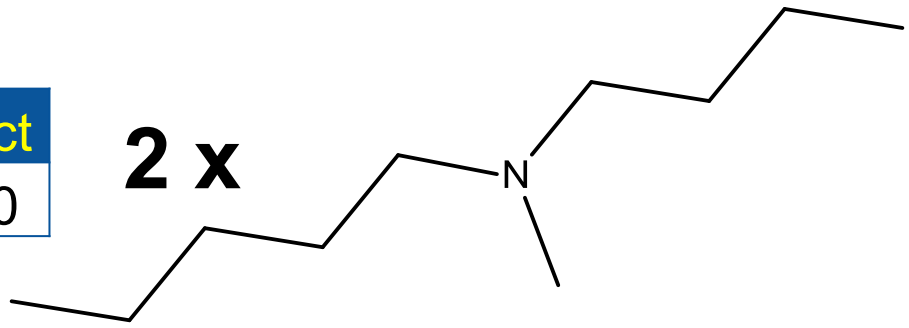
Bond count	Frags	Distinct	Isomo non overlapping	Ratio covered bonds
1	31	6	23	0.742
2	51	11	11	0.71
3	100	26	7	0.677
4	219	56	4	0.516
5	505	137	2	0.323
6	1172	352	2	0.387
7	2709	901	2	0.452
8	6167	2288	2	0.516
9	13666	5557	2	0.581
10	**29323**	**12940**	**2**	**0.645**
11	60560	28659	0	0
12	119880	60064	0	0
13	226229	118926	0	0
14	404703	222216	0	0
15	682196	389935	0	0

STRYCHNINE II



Frags	Distinct
29323	12940





COMPLEXITY CALCULATION

$$c = \frac{\sum_{b=b\min}^{b=a0.5} \frac{u_i}{b^2}\lambda}{a_{0.5} - b_{\min}}$$

$$\lambda = \begin{cases} 1 \text{ for } o_b < 2 \\ (1-r)(1-p_b) \end{cases}$$

λ = redundancy correction factor

a= # *number atoms molecule*

$a_{0.5}$ = # atoms molecule / 2

b=# *bonds in fragment*

b_{min} = minimum fragment size

c=complexity of molecule

o_b = # of non overlapping multiple identical fragments with b bonds

p_b = ratio bonds covered by o_b.

r = b / $a_{0.5}$

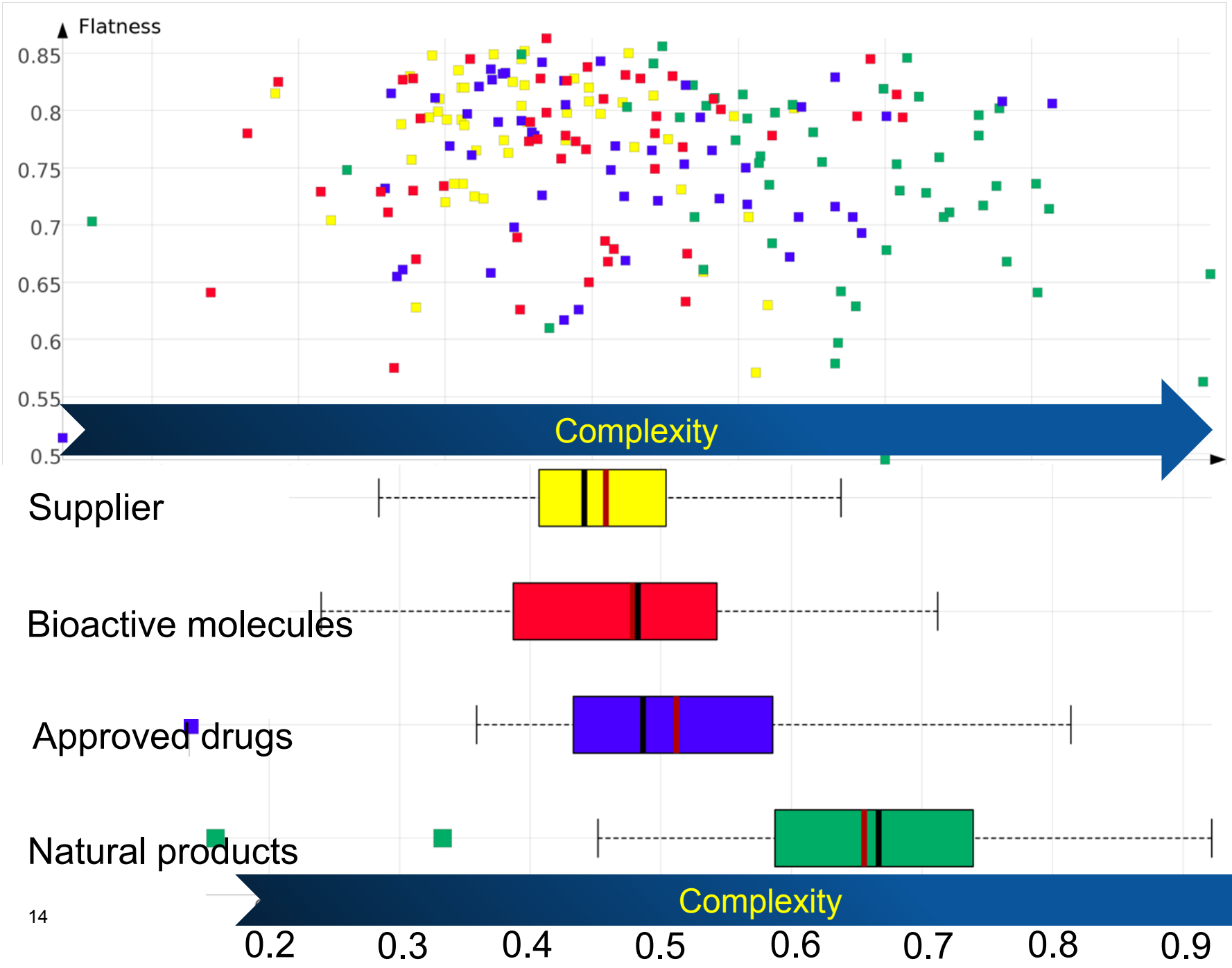
u = # unique fragments

DATA SETS

50 molecules each and 33 bonds / molecule

- Library 1: commercial meta supplier (5 Mio molecules)
  - 200,000 structures 33 bonds→ rnd sampling → 50
- Library 2: highly bio-active compounds (ChEMBL)
  - 66 structures 33 bonds→ most diverse sampling → 50
- Library 3: approved drugs (Drugbank)
  - 170 structures 33 bonds→ most diverse sampling → 50
- Library 4: bioactive natural products (Handbook of pharmaceutical natural products)
  - 70 structures 33 bonds→ Most diverse sampling → 50

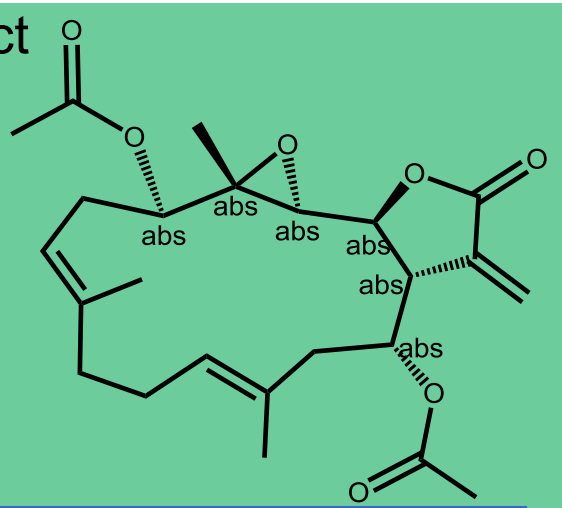
ACTELION

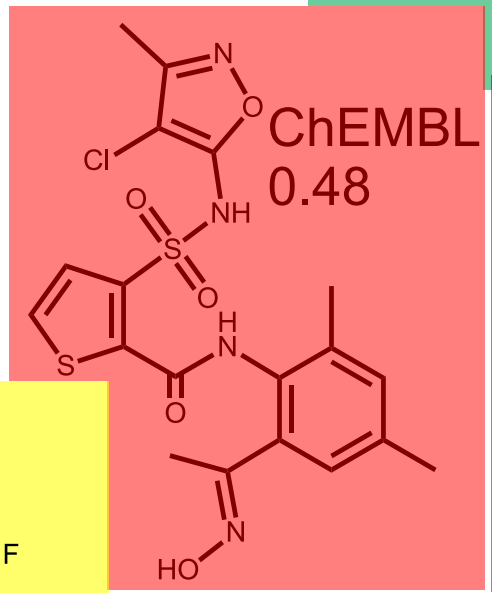

Flatness
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Complexity
Supplier
Bioactive molecules
Approved drugs
Natural products
Complexity
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

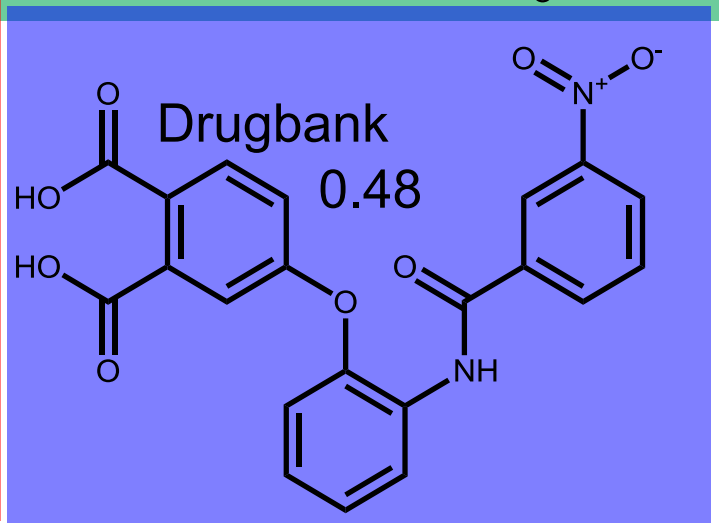
RESULTS: COMPLEXITY MEDIAN MOLECULES

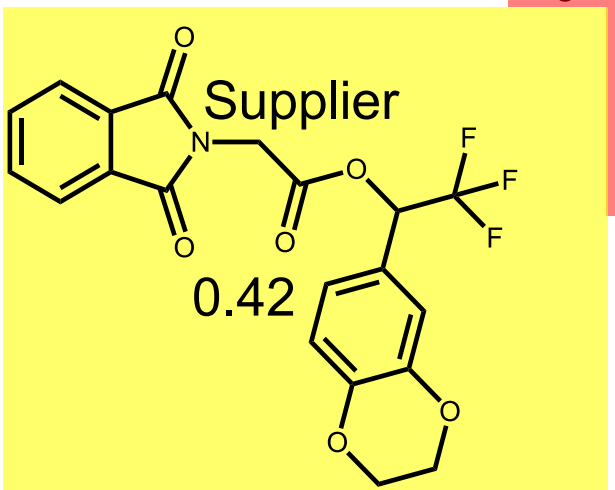
Complexity

Natural product
Cytotoxic
0.67









ACTELION

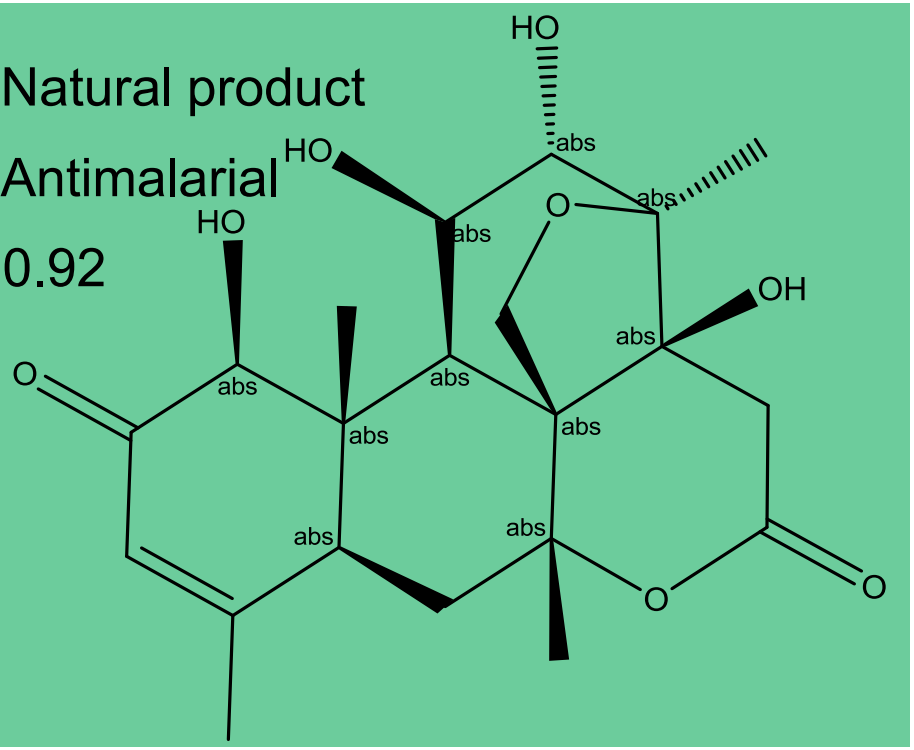
MINIMUM - MAXIMUM COMPLEXITY

Complexity

Natural product

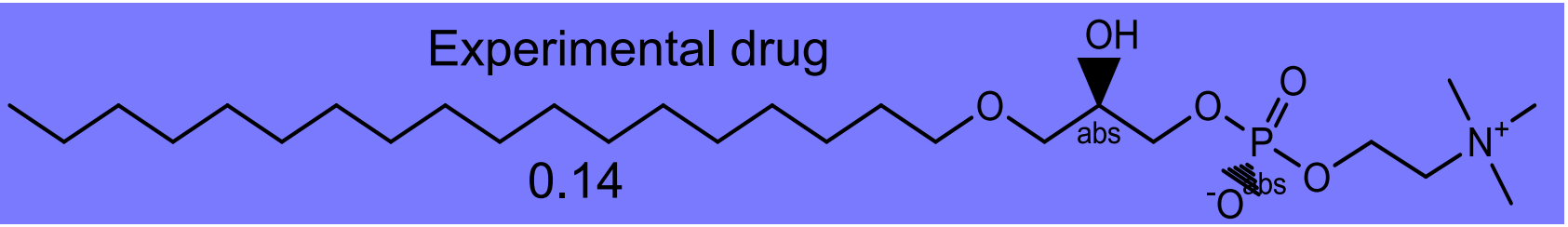
Antimalarial

0.92



Experimental drug

0.14



ACTELION

COUNTERCHECK

Two datasets

identical number of

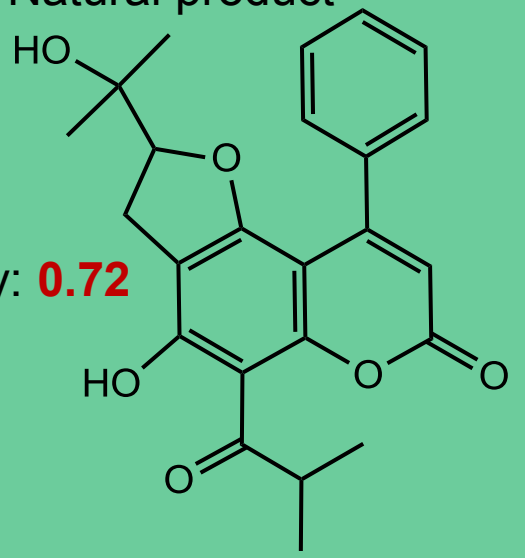
- Atoms
- Bonds
- Hetero atoms

Rings: 4

Hetero atoms: 6

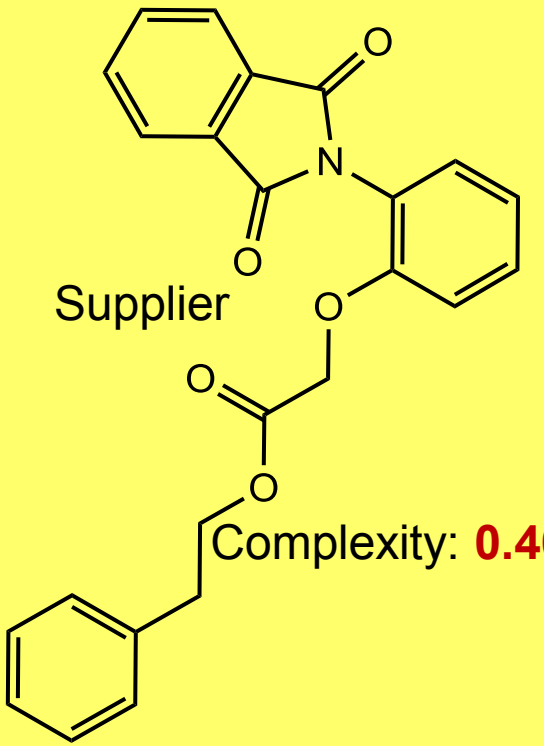
Carbon: 24

Natural product



Complexity: **0.72**

Supplier



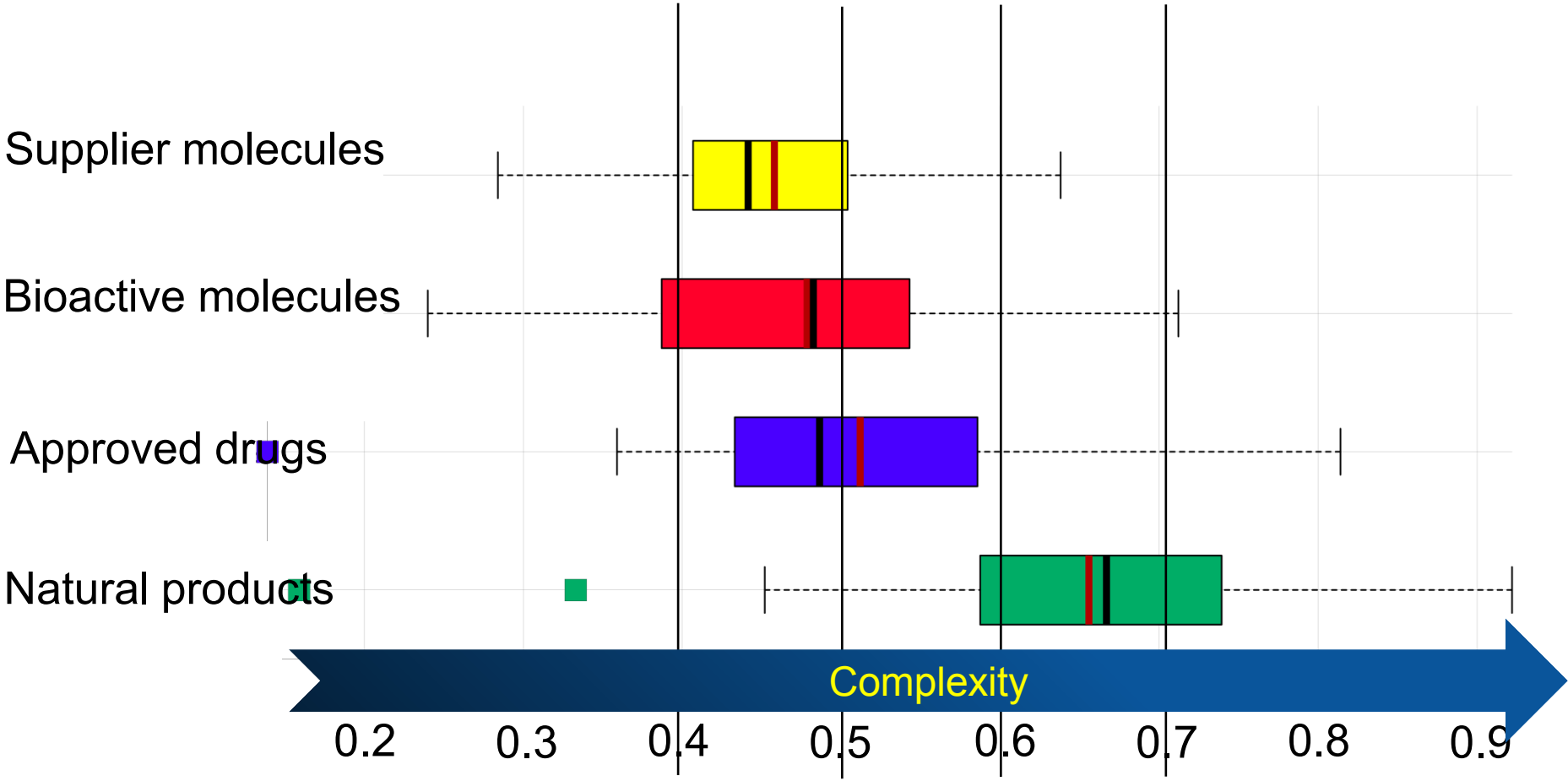
Complexity: **0.46**

Complexity

ACTELION

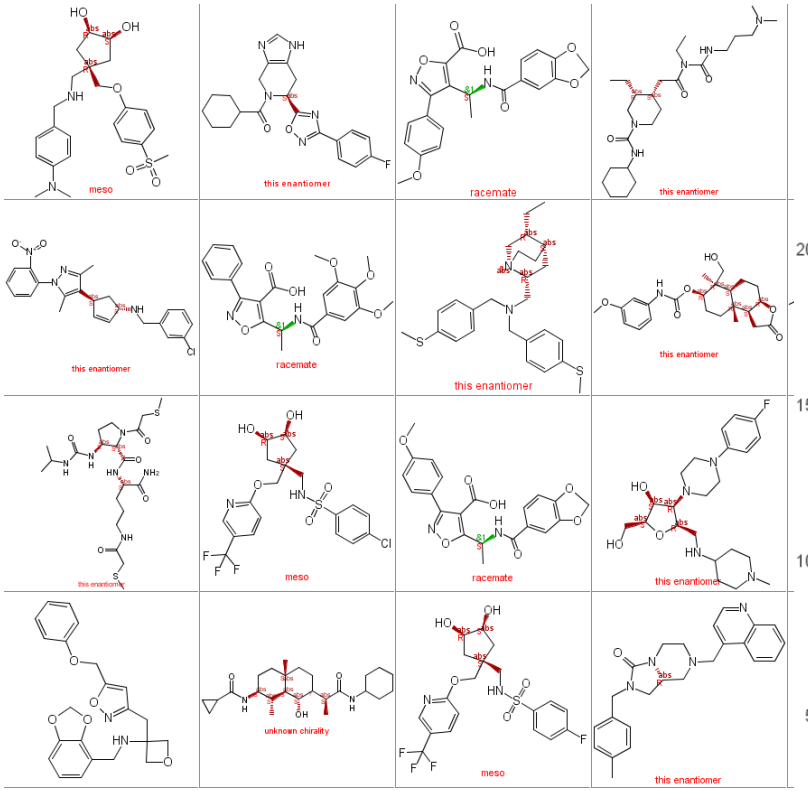
SUMMARY

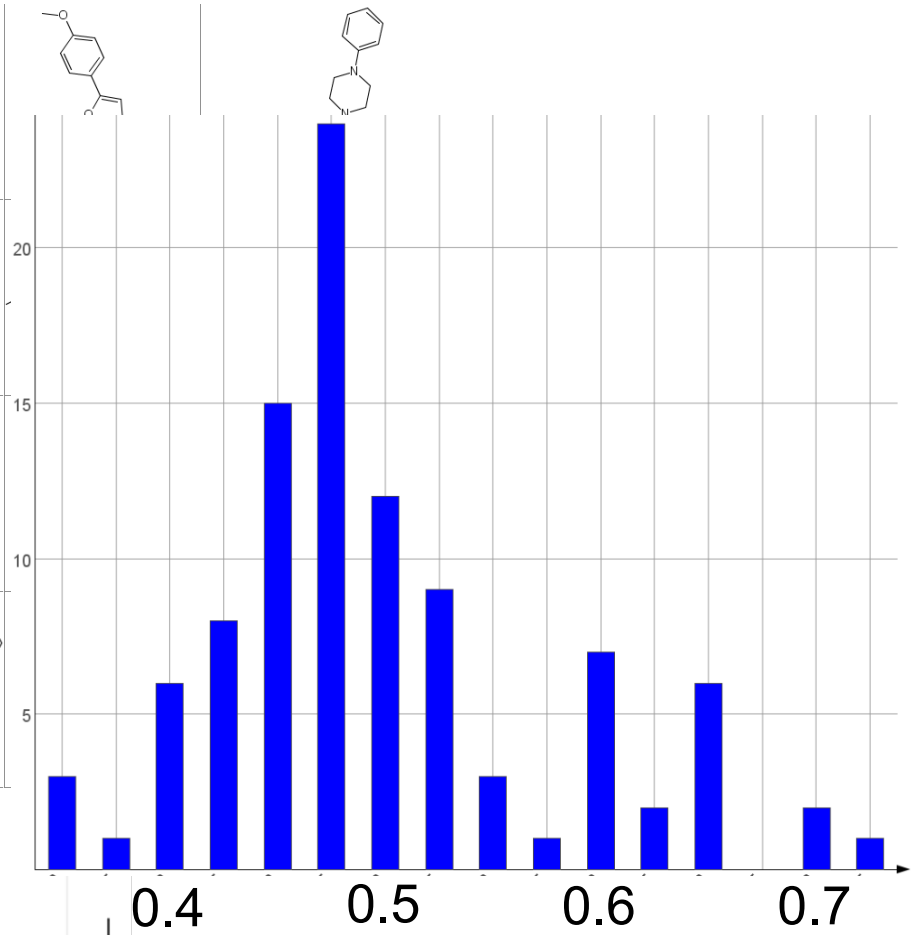
Clear differentiation between

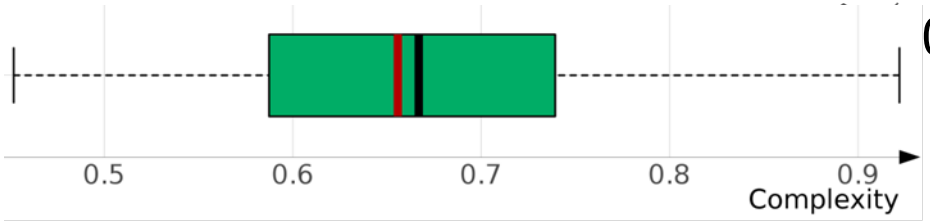


APPLICATION

- Supplier library analysis







ACTELION

CONCLUSIONS

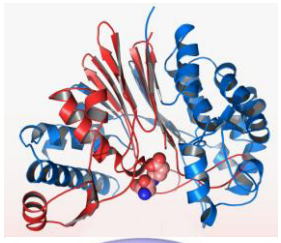
- Low complexity score indicates compounds without innovative character
- High complexity of natural products as a result of evolution
- Complexity score is a figure of merit for compound acquisition and synthesis

Tracing steps of molecular evolution

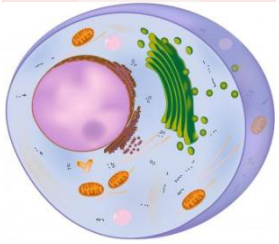
BIOLOGY, A MULTIVERSE

Genes
21,000

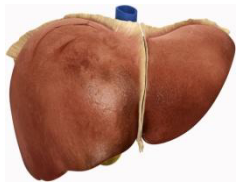
Proteins
50,000



Cell types
200



Organs
100

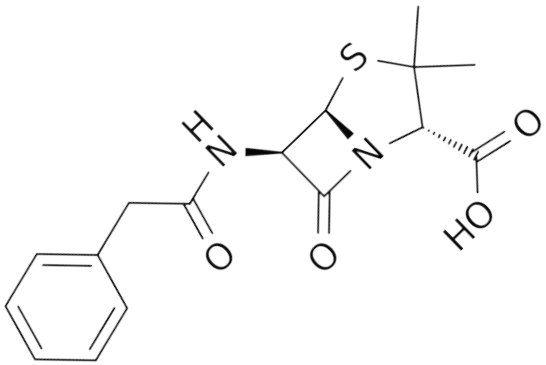


Organism
1 (Human)

Interaction

with

small molecules



Protein: Bacillus anthracis - CapD

DISTORTED FUNCTION

Genes

Proteins

Cells

Organs

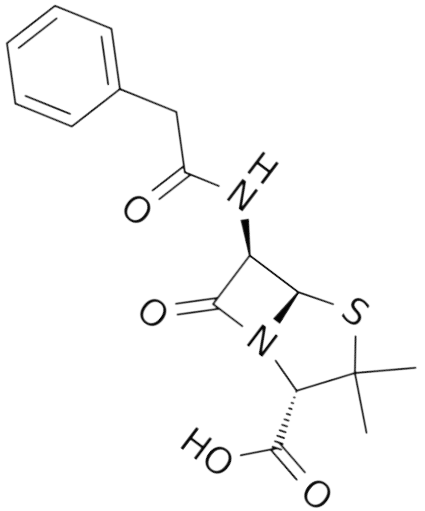
Organisms

Disease

Interaction

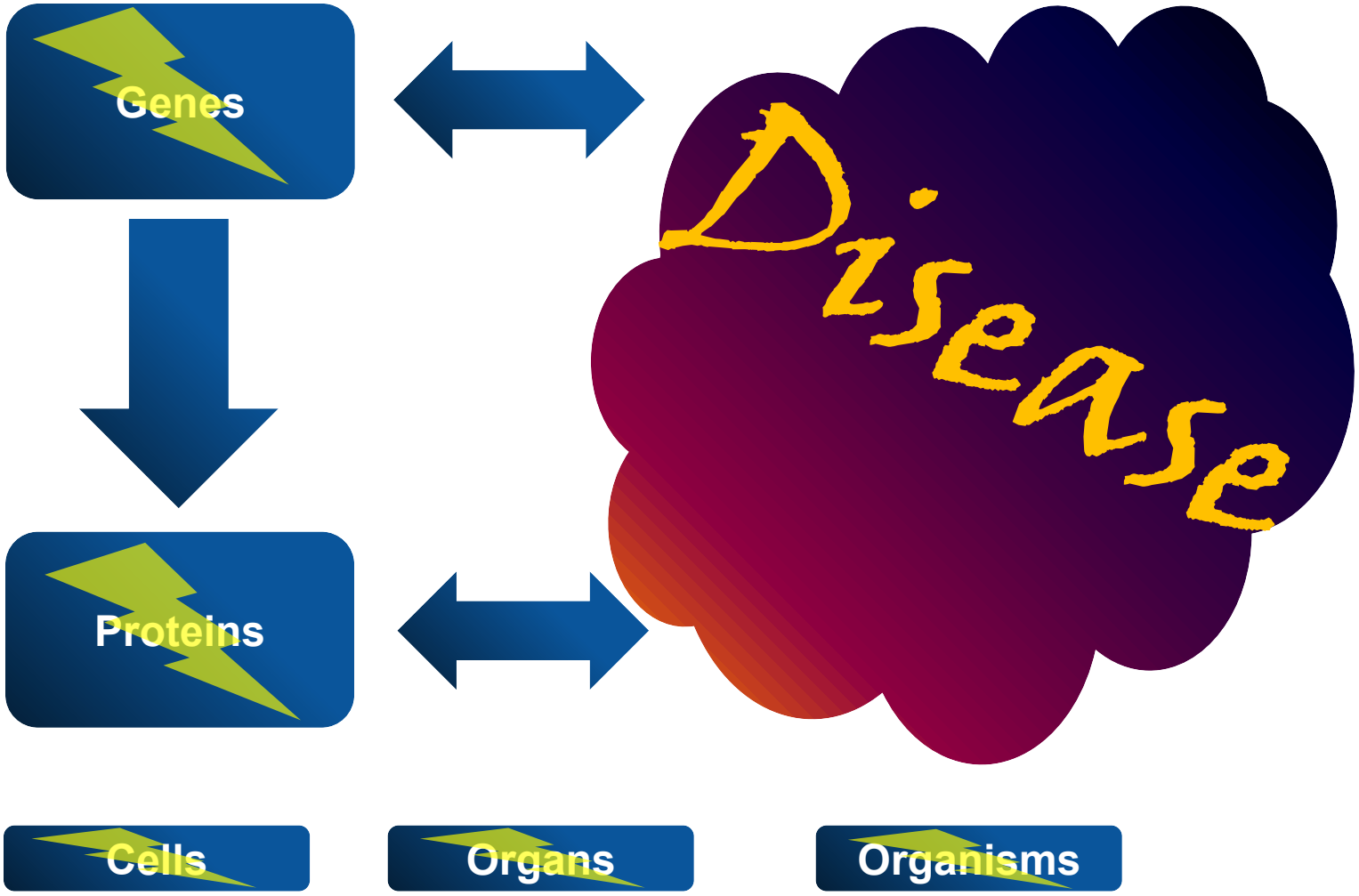
with

small molecules



ACTELION

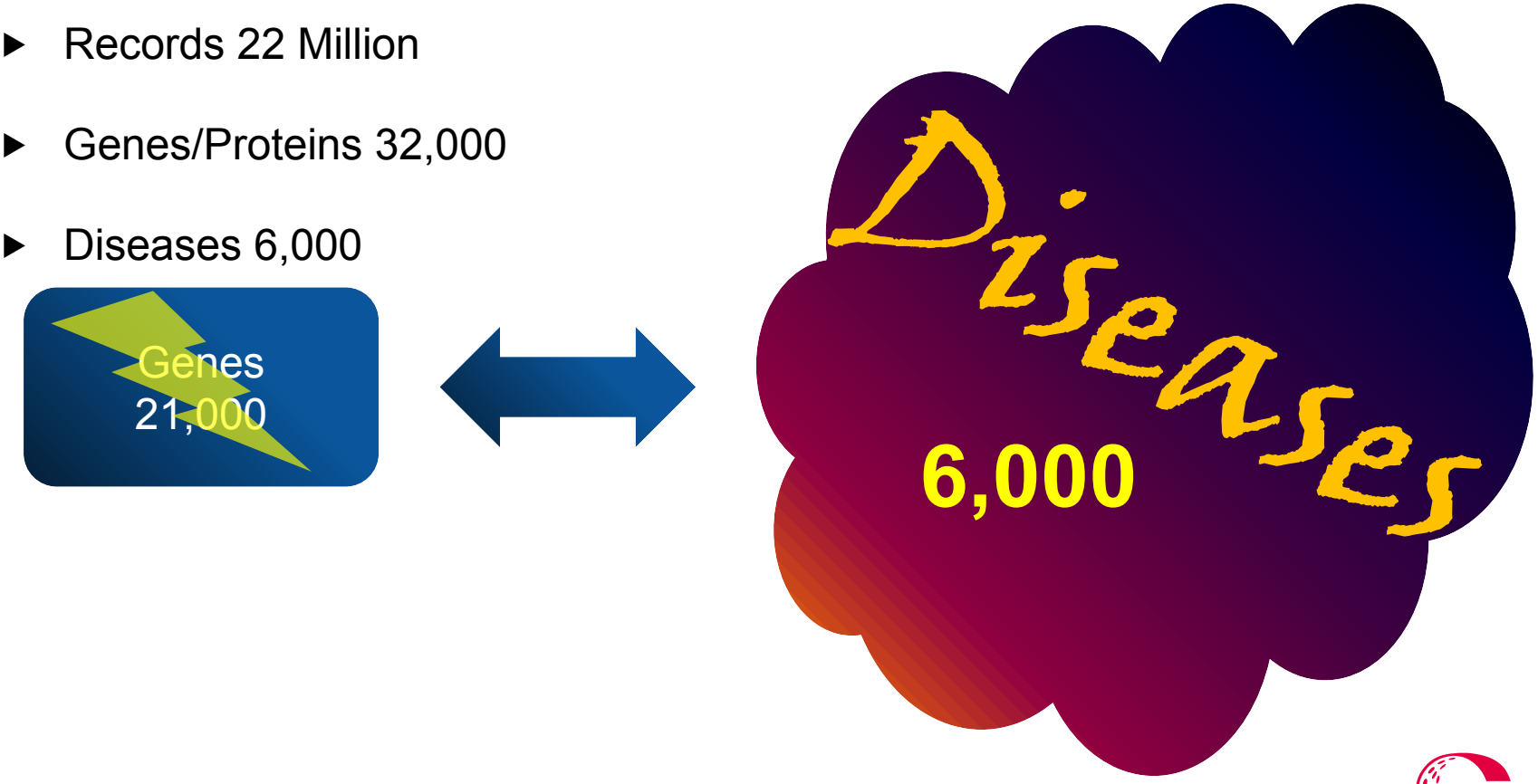
RELATIONSHIPS



ACTELION

SEARCHING GENE2DISEASE RELATIONS

- PubMed database
- Collection of publications with life-science relevant information
- Records 22 Million
- Genes/Proteins 32,000
- Diseases 6,000

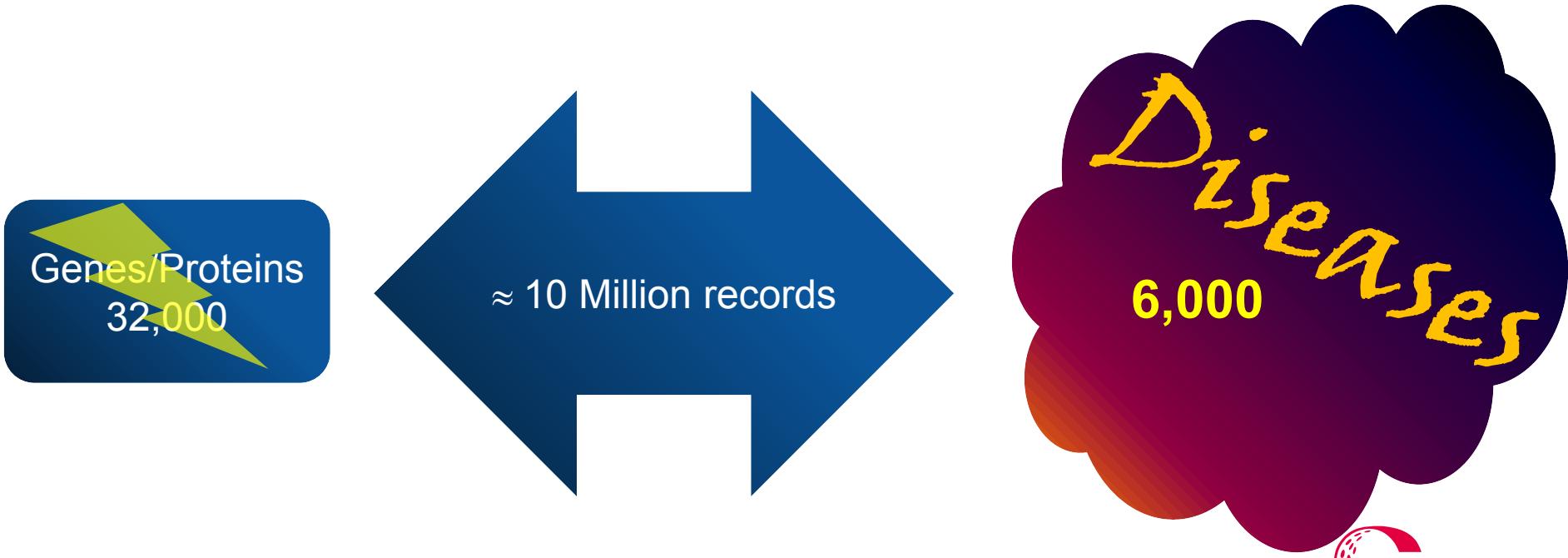


SEARCH SPACE FOR GENE2DISEASE

PubMed: 278,000 publications for

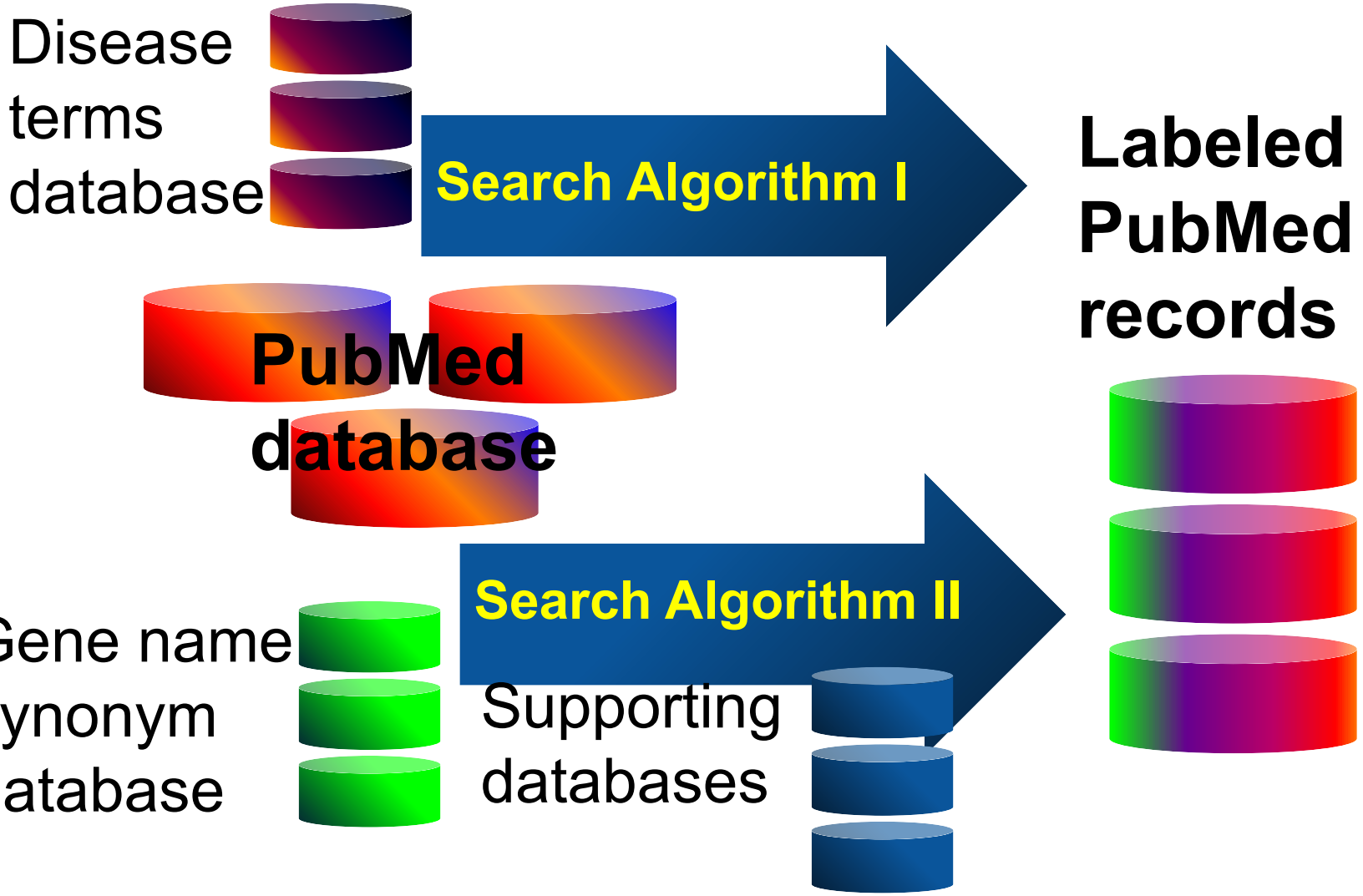
‚Gene-Disease associations‘

Majority of relevant publication does not mention ‚Gene-Disease associations‘



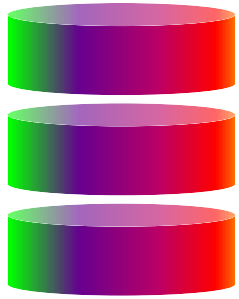
WHAT IS NEEDED FOR MINING?

Tools & know how



ACTELION

LABELED PUBMED RECORDS



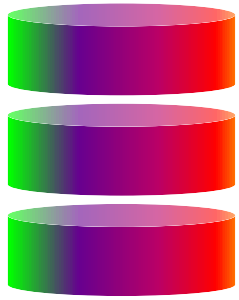
ACLY (ATP citrate lyase)

Reversal of obesity-induced hypertriglyceridemia by (R)-α-lipoic acid in ZDF (fa/fa) rats.

Controlling elevated blood triacylglycerol translates into substantial health benefits. ....analyzed the expression of genes and proteins involved in fatty acid and triacylglycerol metabolism in liver, epididymal fat, and skeletal muscle. Feeding LA to ZDF rats (a) corrected severe **hypertriglyceridemia**, (b) lowered abdominal fat mass, (c) raised circulating fibroblast growth factor-21 and Fgf21 liver gene expression, (d) repressed lipogenic gene expression of **ATP-citrate synthase (Acly),** acetyl-coA carboxylase 1 (Acaca), fatty acid synthase (Fasn), sn-glycerol-3-phosphate acyltransferase 1 (Gpam), adiponutrin (Pnpla3) in the liver and adipose tissue, (e) ...

ACTELION

MAKE A LIST FOR EACH GENE



ACLY

Rank	Disease	Frequency of occurence
1	Body Weight	71
2	Diabetes Mellitus	28
3	Obesity	22
4	Weight Gain	20
5	Diabetes Mellitus...	16
6	Fatty Liver	16
10	Hyperlipidemias	
		1
88	Carcinoma, Transi...	1
89	Leukemia, Myeloid	1

ACTELION

HOW TO MEASURE SUCCESS?

Metric

to assess the quality of the

gene-disease mapping

So far: known disease genes as test cases

How do you know?

Is change in gene expression related to disease?

Drug - Target protein → Gene - Disease

You reall know if you cure the disease

ACTELION

GENE DISEASE ASSOCIATION METRIC WITH DRUGS

Clinical trial phases

- I: healthy patients
- II: small patient group, carefully chosen, successful treatment of illness
- III: larger patient group, successful treatment of illness

- Drug, successful clinical phase II or III
- Approval for defined disease(s) → Disease
- Mode of action is required for approval → Target protein → Gene

ACTELION

TEST SET

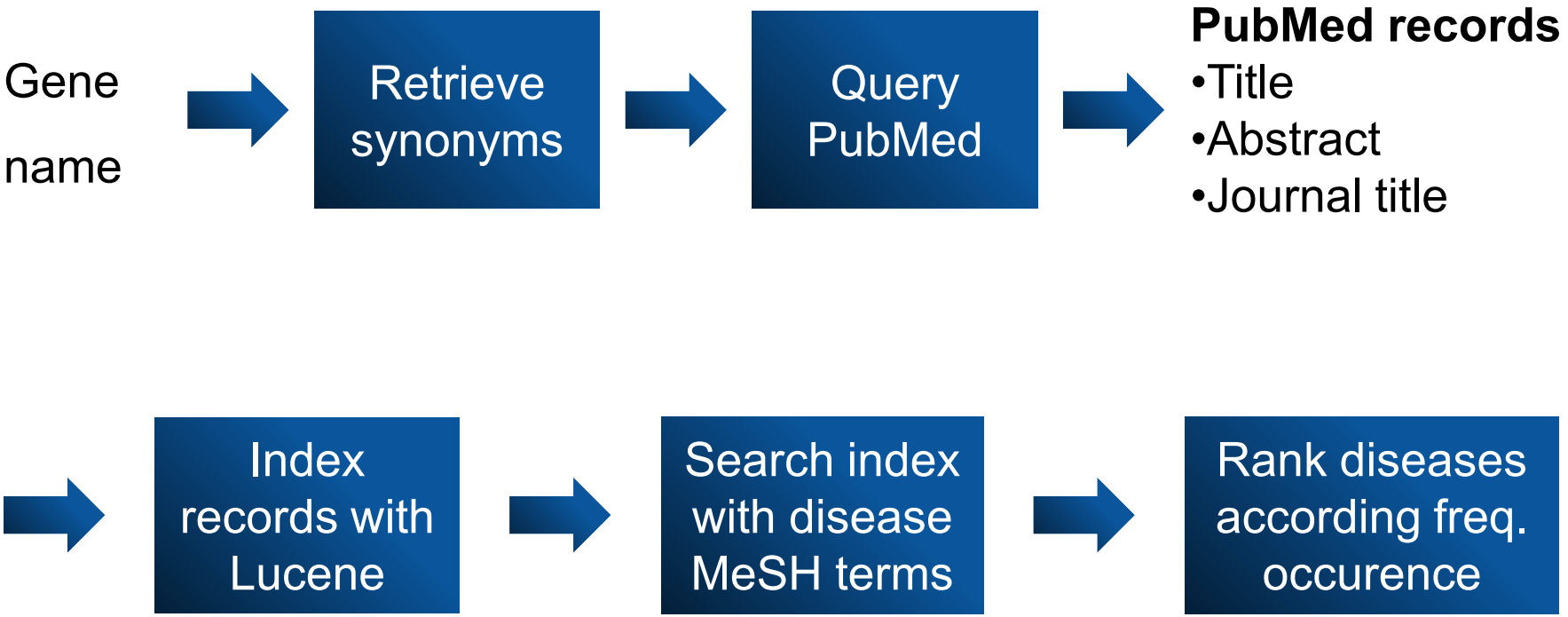
- Compiled from Centerwatch databases 2012 and 2013.
- Random selection to cover different treatment areas
- Clinical Phase II: 18 drugs
- Clinical Phase III: 21 drugs
- **Gene disease associations are not a one to one relation**
  - Some drugs target more than one gene
    - Xeljanz → JAK1, JAK2 and JAK3
  - Some genes are related to more than one disease
    - ERBB2 -> Non-Small-Cell Lung Carcinoma **AND** Heart Failure

ACTELION

EXAMPLES FOR DRUGS FROM CLINICAL PHASE

Drug (Company)	Treatment Area	MeSH descriptor	Associated genes
ETC-1002 (Esperion Therapeutics)	Elevated levels of low-density lipoprotein	Hypercholesterolemia	ACLY, PRKAA1, PRKAA2
ISIS-APOCIIIRx (Isis Pharmaceuticals)	High triglycerides	Hyperlipoproteinemias	APOC3
MK-8931 (Merck)	Alzheimer's Disease	Alzheimer Disease	BACE1

METHOD FOR PUBMED



Input: Genename

Output:Ranked list of diseases

ACTELION

FIGURE OF MERIT FOR TEST DATA

- Retrive ranked list of diseases for test gene
- Calculate relative rank *r* for test gene

$$r = 1 - p / n$$

p position in ranked disease list

n total number of diseases

Rank 1.0: top off the list

Rank 0.0: not in the list

ACTELION

RESULT QUERIES

	Gene name synonyms	PubMed records	Disease MeSH terms
First quartile	8.5	1739	392
Median	**12**	**3972**	**653**
Third quartile	14.5	10298	1243

ACTELION

RESULTS SUMMARY RANKING

- Genes 47
- Drugs 39

Median ranks

	DisGeNET	MalaCards	JensenLab	HuGENavigator	NextBio	Ingenuity	G2DPubMedMiner
Median	**0.64**	**0.61**	**0.86**	**0.50**	**0.42**	**0.85**	**0.94**

ACTELION

CONCLUSIONS

- Straightforward
- Robust

PubMed + Lucene + genename synonyms

➔meaningful gene-disease associations

ACTELION

AND NOW?

Chemical space

- Chemical structures
- Structured data
- Exact description

Biological space

- Medical/biological data
- Unstructured text
- Some fuzziness in the data

MERGE THE UNIVERSES

Gene2Drug

Genes

Proteins

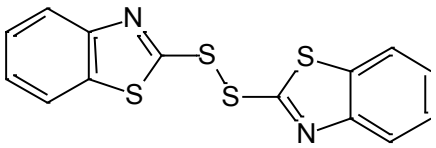
Cells

Organs

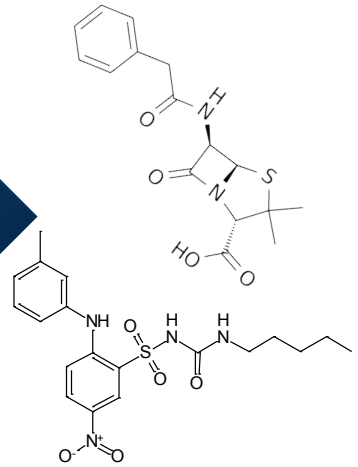
Organisms

Disease

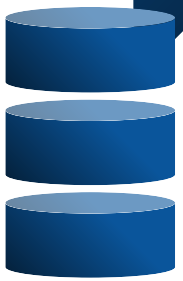
PubMed database



Mine for relations



Supporting databases



Described by PubChem 50 Mio.

Total 10^{60} structures

Commercially available: 8 Mio.

In-house 500,000

Chemist / week: 1-100

ACTELION

THANK YOU!