Quantitative Structure-Property Relationships Part 2

Tudor I. Oprea
Division of Biocomputing
University of New Mexico School of Medicine
toprea@salud.unm.edu
and
Center for Biological Sequence Analysis
Technical University of Denmark, Dept. Systems Biology
tuop@cbs.dtu.dk

Funding: 1R21GM095952-01, 5U54MH084690-03, 2P30CA118100-06

IPAM Seminar Series
17 March 2011, UCLA, Los Angeles

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

DTU

The University of New Mexico ♦ Health Sciences Center
SCHOOL OF MEDICINE

Outline

Part 1

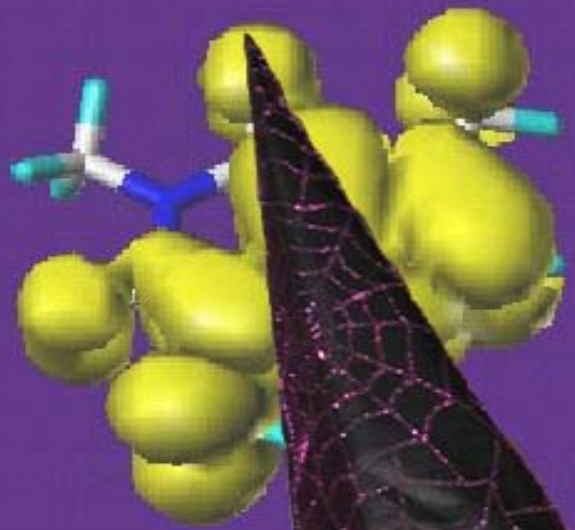
- Overview of Drug Discovery
  - Focus on Informatics
- Model-Free Drug-Like Filters
- A Brief QSAR Adventure
  - An attempt to be Rational about QSAR

Part 2

- Continuing the QSAR Adventure
  - An Attempt to get more Data from the Industry
- Caveat Emptor
  - Beware of What's Lurking in the File(s)
- Exhaustive Enumeration of Scaffold Topologies
  - How Large is Chemical Space?
- Black Swans and Blue Pills

Continuing the QSAR Adventure

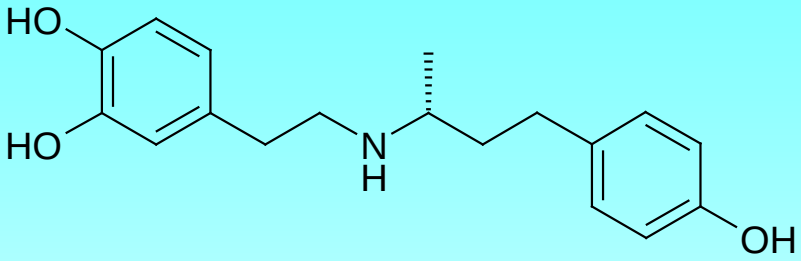
An Attempt to get more Data from the Industry



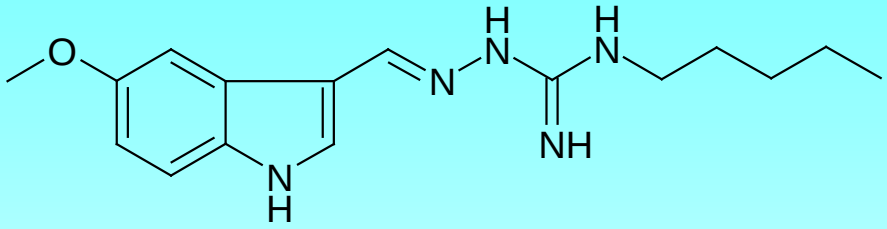
O=C1N(C) C)C(=C12)N=CN2C

CN1C(=O)N(C)C(=O)c(c12)n(C)cn2

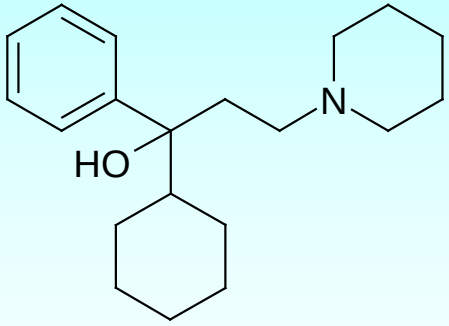
What do these drugs have in common?



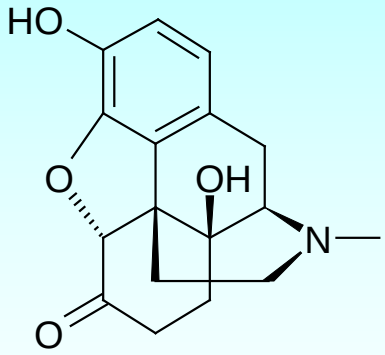
Dobutamine, MW = 301.39



Tegaserod, MW = 301.39

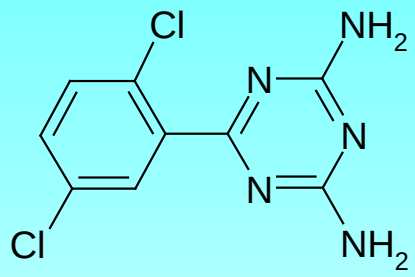


Trihexylphenidyl, MW = 301.48

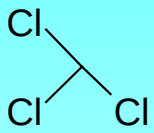


Oxymorphone, MW = 301.34

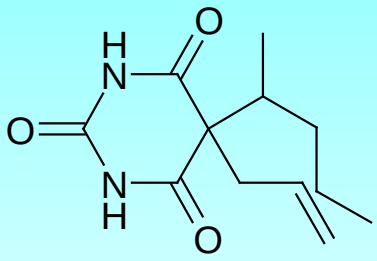
What do these drugs have in common?



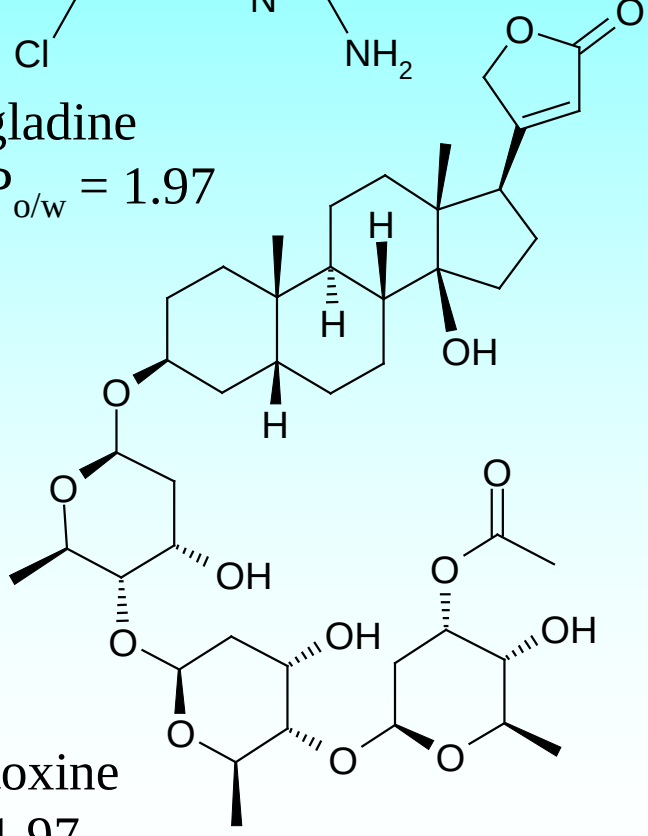
Irsogladine
$LogP_{o/w}$ = 1.97



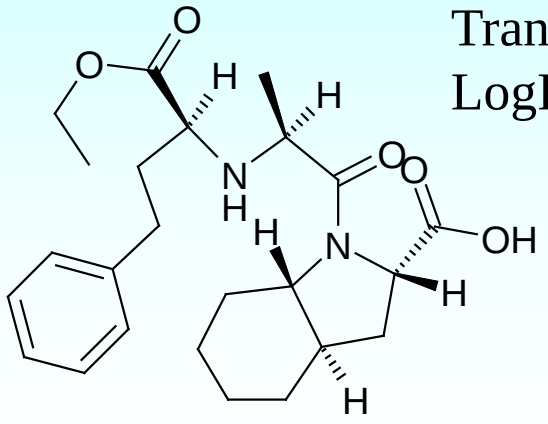
Chloroform
$LogP_{o/w}$ = 1.97



Secobarbital
$LogP_{o/w}$ = 1.97



Acetyldigitoxine
$LogP_{o/w}$ = 1.97



Trandolapril
$LogP_{o/w}$ = 1.97

Safe Exchange of Chemical Information

- **The NIH Roadmap acquired and is screening 0.4 million chemicals for biological activity** (update: Feb 2011). This effort requires storing and handling of physical samples; however, there is no provision to measure these for, e.g., water & DMSO solubility. How do we know the sample was screened? [it could simply be insoluble!]
- **Private sector data can directly benefit public science in this case**
- Computer models will improve when 100,000 endpoints become available; such predictions can become part of the PubChem system
- **One can also envision plates with undisclosed structures, made available by private companies, for screening thru the NIH Roadmap.**
- *Descriptors to mask chemical structure (DMCS) are a necessity*
- We can make a directed effort to derive DMCS in source code (public)
- **Source code has to be public in order to circulate in the industry**
- Tables of descriptors & measured properties can become available
- **Potentially the biggest contribution from cheminformatics to society & public benefit**

Challenge posed by T. Oprea and C. Lipinski, ACS 2005

The (now defunct) ChemMask project

http://pimento.health.unm.edu/index.html

- Data-sharing schemes required. The private sector needs to find mechanisms for releasing data to the public.
- Computational models sharing biological and physico-chemical properties for many compounds across the private sector can lead to significant improvements of computational models.
- The proprietary value of chemical structures is a major impediment to sharing information and testing of computational models. Methods that provide for the safe exchange of chemical information could perhaps address these problems.

The issue was if we could derive models for safe chemical information exchange, in such a way that it does not allow for the reverse engineering of the original chemical structures

ACS CINF/COMP Symposium, San Diego 2005

C.A. Lipinski & T.I. Oprea organizers

E. Wilson, C&E News, April 25, 2005

The (now defunct) ChemMask challenge

http://pimento.health.unm.edu/about.html

- **Can we mask the chemical structure of a molecule, while sharing chemical properties that are relevant to this molecule in a given context?**

  1. We challenge anyone who is interested to reverse engineer the masked chemical structures from a set of molecular descriptors available to this website. *What we intended to offer:* Chemical descriptors according to your method of choice. We would have provided the descriptors, from which one can guess the chemical structures.

  2. We challenge anyone who is interested to offer descriptors that pass the first challenge, but have to be relevant to a given set of properties. *What we intended to offer:* We assembled 13 datasets with the Y column (target property) and the X block (chemical descriptors according to methods that pass the first challenge). We require the source code [exception: the binary code is freely available to the public sector, and commercially available otherwise].

  The source code requirement comes from a legal perspective (the private sector will need assurance that the code used to generate the masking descriptors does that, only that and nothing but that).

Descriptor Confusion Test Sets

- To establish how reliable is the fingerprint technology in mapping structures, we investigated structures from WOMBAT [Sunset Molecular] and the iResearch Library [ChemNavigator] with several descriptor systems.
- For each database, we extracted all the **unique, non-stereoisomeric** (i.e., no R/S or E/Z isomers) SMILES, and found 98,575 structures in WOMBAT, and 13,334,014 structures in iResearch Library, respectively.
- WOMBAT is a prototype database for medicinal chemistry , whereas ChemNavigator collects virtual (but feasible) and existing commercial compounds

Bologa C et al., J. Comput-Aided Mol Design, **2005**, 19:625-635

Descriptor Collision Rates

Fingerprint type/size	WOMBAT duplicates(*)		iResearch Library™ duplicates	
	Number	%	Number	%
MDL – 320	4,526	4.7	1,943,712	14.5
DY – 512	8,443	8.8	2,029,981	15.2
DY – 1024	5,809	6.1	974,395	7.3
DY – 2048	4,732	4.9	702,333	5.3

(*) For WOMBAT we also computed descriptor collisions using 2D descriptors, and found 433 (0.5%) duplicates

Bologa C et al., J. Comput-Aided Mol Design, **2005**, 19:625-635

Daylight
Chemical Information Systems, Inc.

Daylight Fingerprints Collision

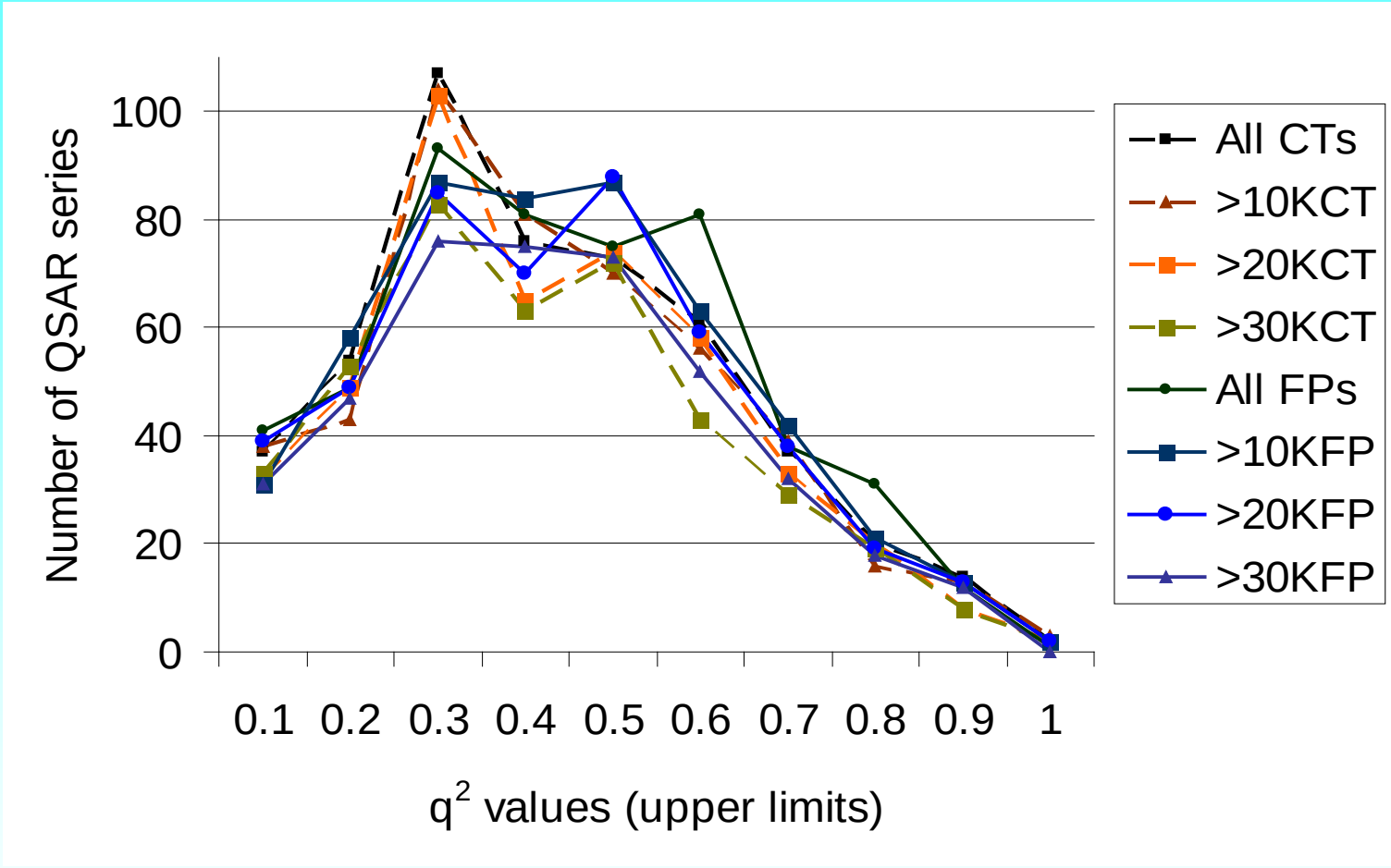
- WOMBAT (2005.1) has 98,575 unique non-isomeric SMILES and 70,211 unique graphs (atom & bond types are suppressed).

DY FPsize	uniqueFP	uniqueFP (graph)
32,768	94,148	178
16,384	94,123	178
8,192	94,087	178
4,096	93,801	178
2,048	93,596	178
1,024	93,079	178
512	91,891	178
256	88,231	126
128	69,183	108
64	12,939	74
32	776	43

- The upper limit of 94,141 unique fingerprints for the DY-32,768 bits leads to ~4434 duplicates (i.e., ~5% collision rate)

Bologa C et al., J. Comput-Aided Mol Design, **2005**, 19:625-635

QSAR Validation of Confused Descriptors



The validation dataset had 1277 series - 948 unique SAR series, or targets (some had multiple Y columns), totaling 50,925 activity–structure pairs

Each series had >=25 cpds, q^2 CV7, PLS (WOMBAT-PLS)

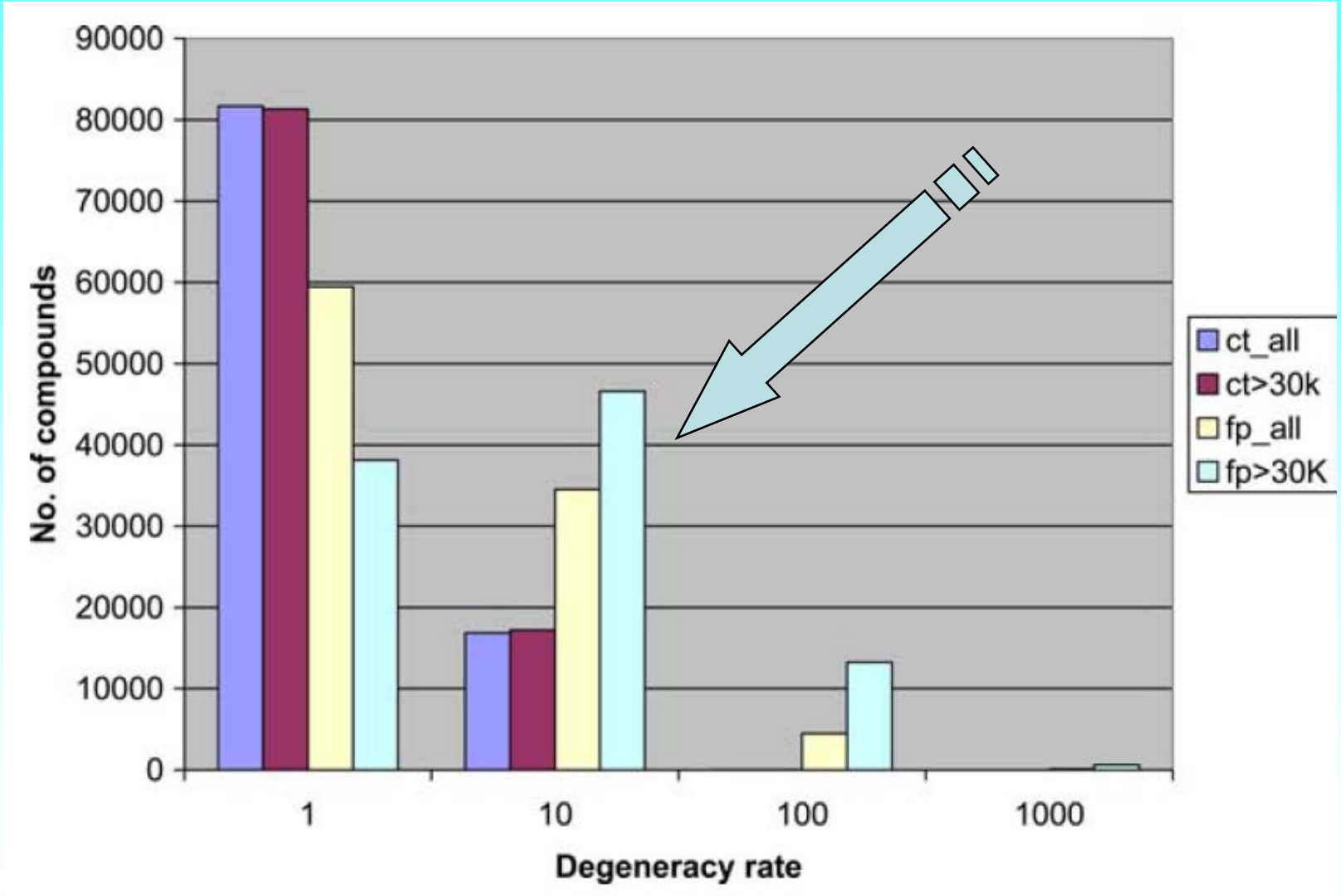
The University of New Mexico
SCHOOL OF MEDICINE

QSAR Validation of Confused Descriptors

Binary set	Patterns	Dup%	$q^2 \geq 0.3$	Dup%
All FPs (403)	71,634	27.33	283	22.16
>10KFP (232)	68,824	30.18	278	21.76
>20KFP (170)	62,667	36.43	260	20.36
>30KFP (114)	**53,563**	**45.66**	**236**	**18.48**
Counts set				
All CTs (403)	88,768	9.94	403	31.55
>10KCT (250)	88,759	9.95	400	31.32
>20KCT (219)	88,647	10.06	390	30.54
>30KCT (194)	88,550	10.07	394	30.85

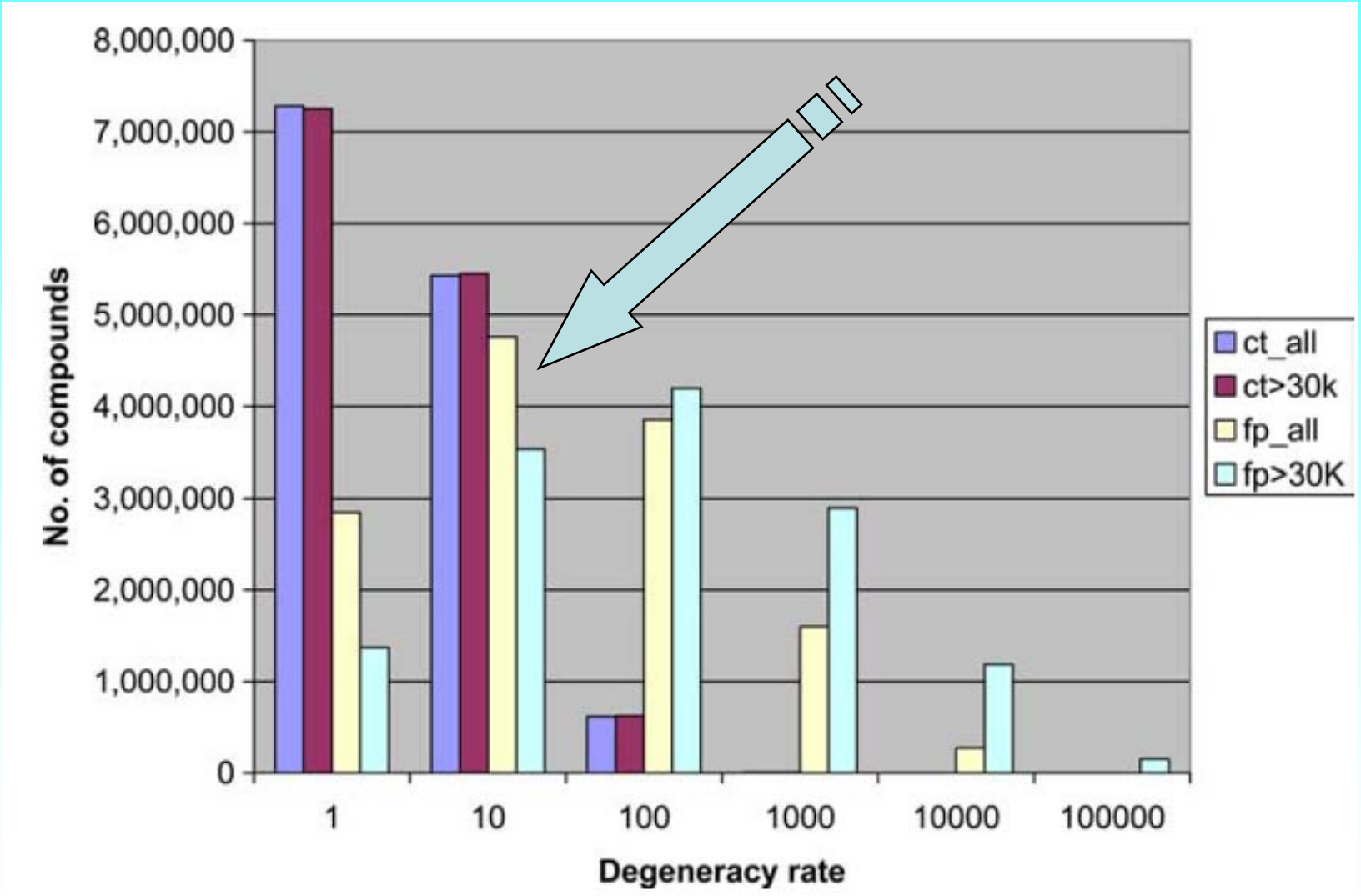
Bologa C et al., J. Comput-Aided Mol Design, **2005**, 19:625-635

Degeneracy Rate in WOMBAT



Bologa C et al., J. Comput-Aided Mol Design, **2005**, 19:625-635

Degeneracy Rate in ChemNavigator



Bologa C et al., J. Comput-Aided Mol Design, **2005**, 19:625-635

Solution: Surrogate Chemicals

		prediction performance for the PHYSPROP set		
Indices/Method	training set size	r^2	RMSE	MAE
real datasets				
E-state/NN[1]	1949	0.87	0.69	0.47
E-state/NN[2]	1671	0.88	0.65	0.46
3D/MLRA[1]	1949	0.45(0.84)[3]	1.75(0.72)	0.74(0.56)
surrogate datasets				
E-state/NN[1]	1949	0.85	0.72	0.50
E-state/NN[2]	1671	0.67	1.1	0.64
3D/MLRA[1]	1949	0.69(0.82)[3]	1.04(0.73)	0.71(0.57)

[1]Real and surrogate sets selected by mapping PHYSPROP molecules to the Pub_NZ (NCI, ZINC) database.

[2]Real and surrogate sets selected by mapping PHYSPROP molecules to the iResearch Library.

[3]Statistical parameters after filtering heavy outliers with absolute error > 2 log units (647 and 559 out of 12903 PHYSPROP molecules for real and surrogate models, respectively) given in brackets. E-state/NN models were developed using 75 E-state indices and ASNN. 3D/MLRA models were developed using 3D indices and multiple linear regression analysis. r^2 is square of Pearson correlation coefficient between predicted and calculated values.

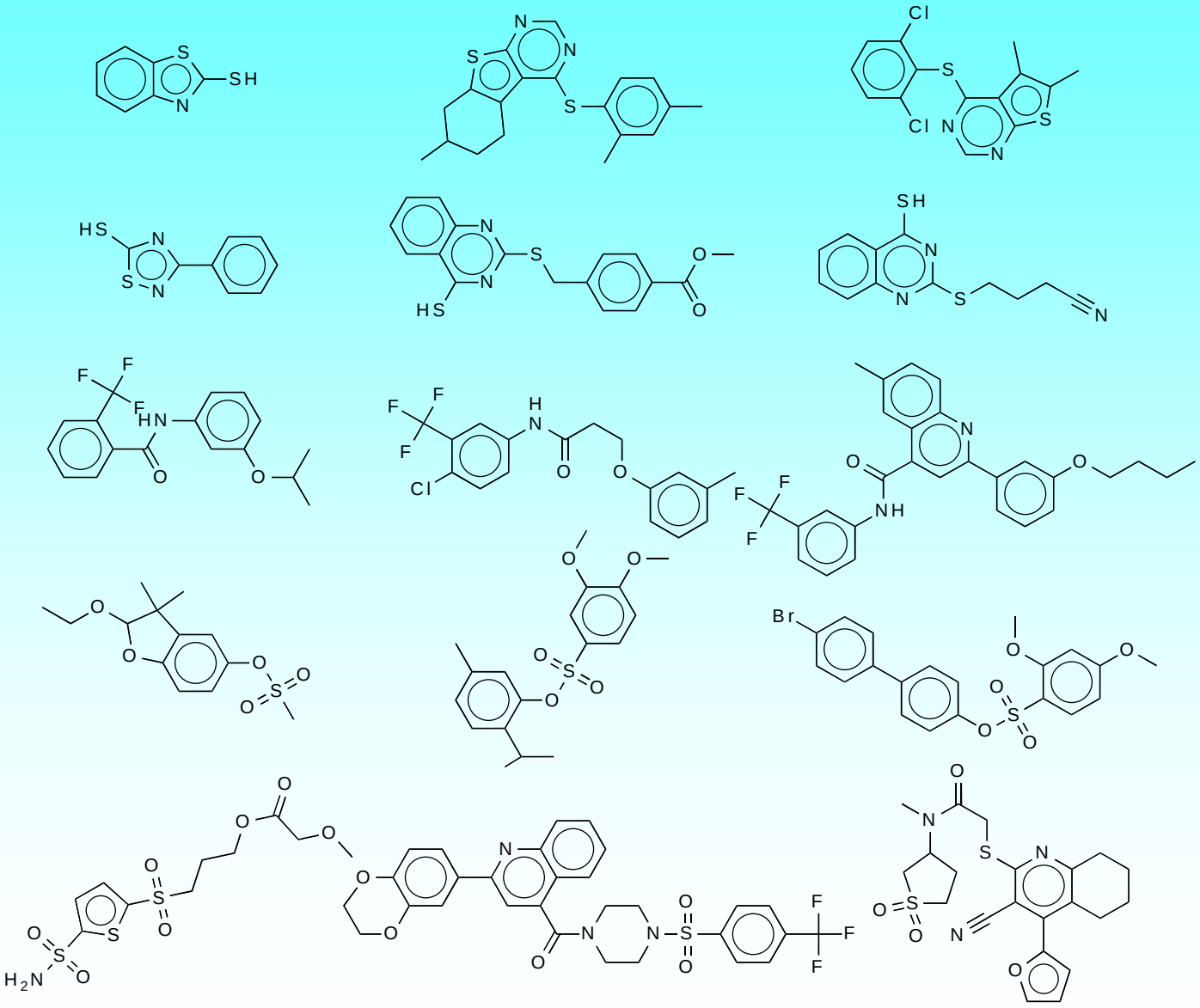
Tetko I et al., J. Comput-Aided Mol Design, **2005**, 19:749-764

Surrogate Chemicals: Examples

PHYSPROP molecule

100th similar molecule

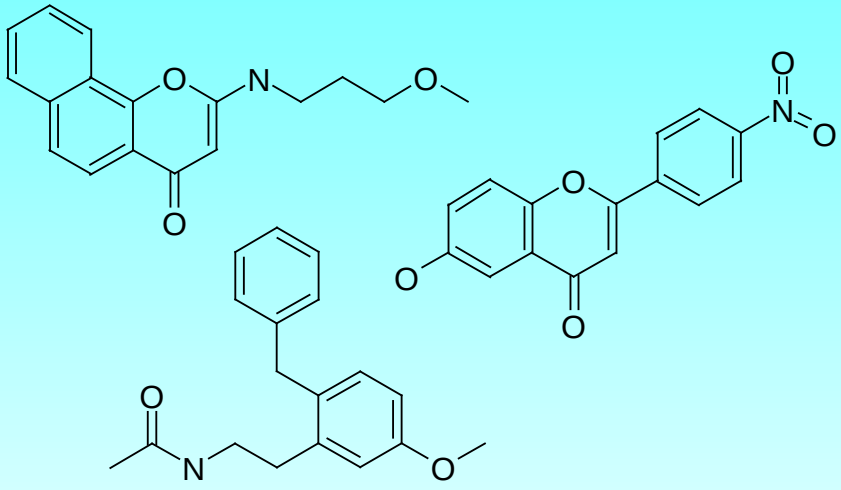
1000th similar molecule



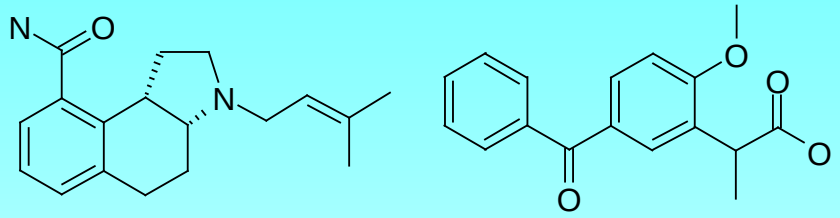
Tetko I et al., J. Comput-Aided Mol Design, **2005**, 19:749-764

Interesting neighbors

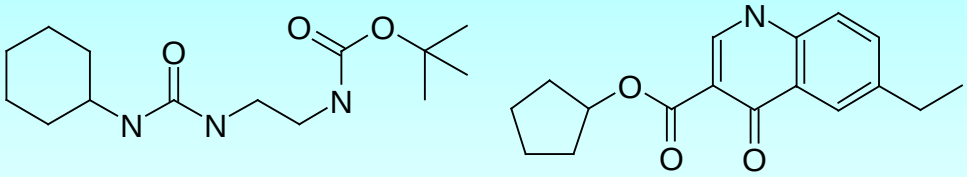
MW between 283.2 and 285.4, CLogP between 2.89 and 2.99



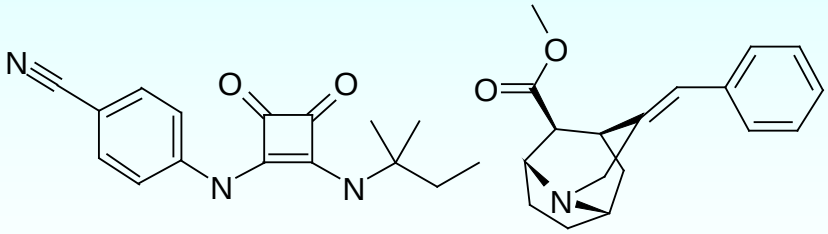
MW ~283.25, CLogP = 2.95



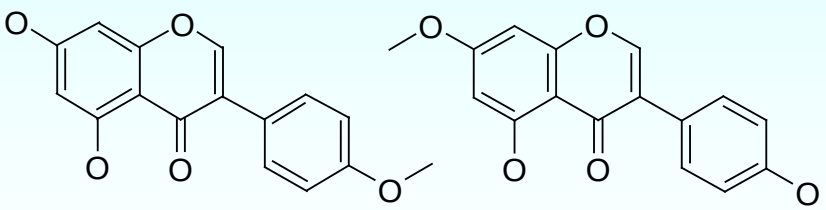
MW ~284.35, CLogP = 2.98



MW ~285.35, CLogP = 2.9



MW ~283.35, CLogP = 2.89



MW =284.26, CLogP ~2.99

Safe Exchange: Why?

Who needs to exchange chemical information?

- **There is a large amount of experimental data in the private sector**.
- **This is rarely, if ever, made available for public-sector science**
- Solubility (DMSO; water); melting points; $LogD_{74}$; cytochrome P450 inhibition/substrates; metabolic stability; toxicity – in vitro; in animals; human data [clinical PK for failed and successful trials], etc.
- *Everyone wants good software to predict all of the above*
- Everyone blames computational chemistry for failure to deliver
- Everyone thinks computer predictions are unreliable
- **In fact, everyone wants good models but is not willing to share data**
- The major issue is lost IP position and the fear that competitors will use this data to forge ahead with similar/competing products
- What's needed is a reliable way to exchange chemical information, without the possibility of reverse engineering chemical structures

D. Bradley, ***Nature Rev Drug Discov* 2005, 4:180-181**

Issues with “Safe Exchange”

– At what level would the community feel that DMCS are safe and “uncrackable”?

– Some degree of chemical information required (otherwise models are useless)

– Assuming “safe exchange” criteria are accepted, how many companies would agree to release data? *Answer: none so far*

– FDA-deposited data are not public domain, but they contain a wealth of very useful input data which can be used to build good models.

– *LeadScope and LHASA Ltd are selling such data (they have successfully established partnerships with the FDA and the industry)*

– No funding & efforts are dedicated to this issue

– This could have become a “win-win” situation; it’s a people issue.

Academic vs. Industrial Research

- For most academics, research begins with an idea, which is then implemented – usually by example on a well-known case, and published 1x, 2x, …Nx by the same group, sometimes by others.
- Goal is peer-recognition via publications. Often, the project stops with publication.
- For industrial scientists, research *sometimes* begins by reading academic papers; and sometimes, it begins by addressing unmet needs in-house. It's often developed to function in production mode, and tested on-the-fly on novel projects.
- Goal is to solve company needs. This may lead to peer recognition. Often, the real project begins at the stage where academics would have submitted a paper.

Academic vs. Industrial Research (2)

- Academic research in cheminformatics leads to software and technology that is often unique; until recently, academics naively believed that what works on 100 molecules works on 10,000; now they believe that if it works on 100,000, it will work on 10,000,000.
- Academics keep developing software with funny/flashy names, and consider discovery an intellectual exercise.
- Software companies function as “translators” by taking academic software, re-writing and adapting it, to provide it industrial-grade quality (for production mode). Thus, they play an important role in disseminating academic research.
- Industrial cheminformatics research is rarely (RASmol) made available to the public. Patents, papers and other publications are the outcome. This software has to reach chemists, and discovery is their main goal.

A natural history of leadership

Stage	*Time period*	*Society*	*Group size*	*Leadership structure*	*Leader*	*Leader-follower relations*
1	> 2.5 million years ago	Pre-human	Any size	Situational or dominance hierarchy	Any individual or alpha	Democratic or despotic
2	2.5 million- to 13,000 years ago	Hominid bands, clans, tribes	Dozens to hundreds	Informal, situational, prestige-based	Big man, head man	Egalitarian and consensual
3	13,000- to 250 years ago	Chiefdoms, kingdoms, warlord societies	Thousands	Formal, centralized, hereditary	Chiefs, kings, warlords	Hierarchical and unilateral
4	250 years ago to the present	Nations, states, businesses	Thousands to millions	Structural, centralized, democratic	Heads of state, managers and executives	Hierarchical but participatory

Leadership abilities evolved before humans, but current society faces novel challenges. Potential pay-offs are estimated by followers before leadership is accepted. Leaders must be socially astute, competent and benevolent. Dominance is not a long-term situation in science, as followers are the pool for future leaders.

Vugt, Hogan, Kaiser – American Psychologist 2008

Safe Exchange: Why indeed?!

Chemical data vs. knowledge

- **G. Maggiora argues that "safe exchange" is a moot point since knowledge, not data, should be transferred**.
- **The lack of good models for solubility, to take the simplest of properties key to FDA drug approval, is limited by lack of quality data**
- How many academics are even aware of the **FDA definition for solubility**? Hint: it's probably none of the definitions used by chemists
- *FDA: A drug is soluble if the maximum strength dose is soluble in a glass of water between pH 1 and 6.8*
- Models can be developed to address, e.g., solubility for most common approved FDA strengths (rarely over say from 0.1 to 1000 mg), over a pH range, in order to improve solubility prediction tools
- **Everyone wants models, but *very few are sharing data***
- The EU now anticipates QSAR tools to assist with environmental regulations
- In some companies, models are now "excellent" & have replaced experiments
- In an ideal world, chemical structures would accompany data
- In the real world, reliable and good models can be "black box"
- I do not need to know how my car works, I just need to learn to drive

To conclude…

- Sixteen groups have conclusively shown that DMCS and “safe exchange” are possible
- Funding, partnerships – these need to be developed and could grow over time
- **It is possible that “paranoia” will never disappear**
- Consortia on precompetitive areas are being developed (e.g., LHASA Ltd on Tox; MolDiscovery on P450s; the NIH plans to fund an initiative on docking/scoring)
- Open competitions, e.g., CASP; the recent docking one from OpenEye; the upcoming solubility contest – these can lead to visible improvements and better quality science
- Pre-doctoral, Ph.D. and post-doctoral industrial training of young scientists is of great benefit to both sectors.
- Academic & Industrial Researchers often alternate between leaders & followers. Expect them to stay INVOLVED!

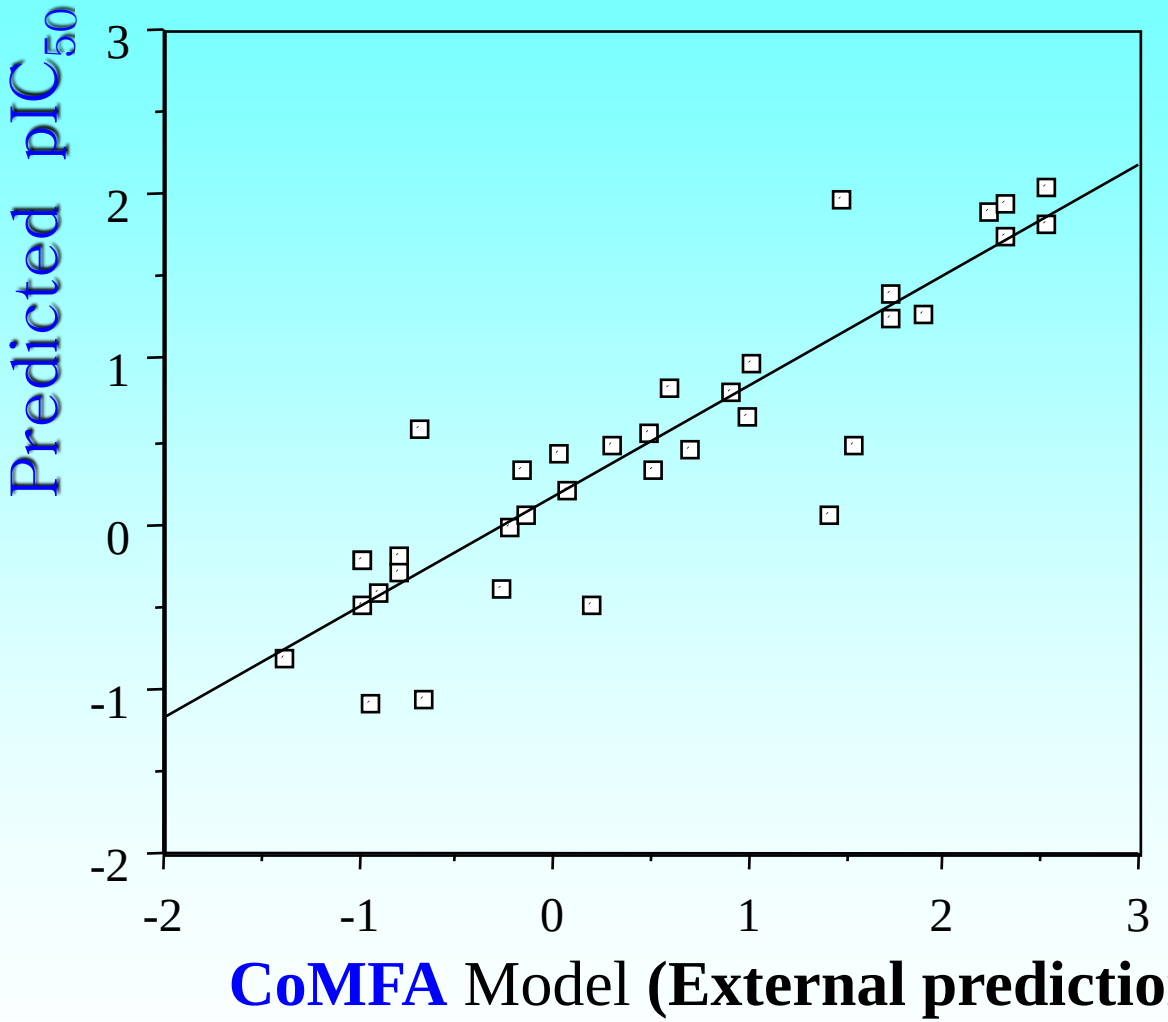
Take Home Message for IPAM community

- **At what level does the community feel that descriptor systems are safe and uncrackable?**
- Problem: some degree of chemical information is required (otherwise models are useless)
- Assuming “safe exchange” criteria are accepted, how many companies would agree to release data?
- How about FDA-deposited data (not public domain, but very useful for all competitors)? *LeadScope and LHASA Limited are already selling such data (being mined by summer students)*
- Right now, little or no funding & efforts dedicated to this issue
- If enough momentum is gathered, this can become a win-win situation for everyone involved; however, *it remains a people issue and, if not adopted, will simply reside in the “poor me” domain* (i.e., people complain about lack of data, etc.)
- Can advanced mathematics (data encryption?) help?

Caveat Emptor

Beware of What's Lurking in the File(s)

Reliability of Biological Data (1)



HIV-1 Protease Inhibitors

N = 36

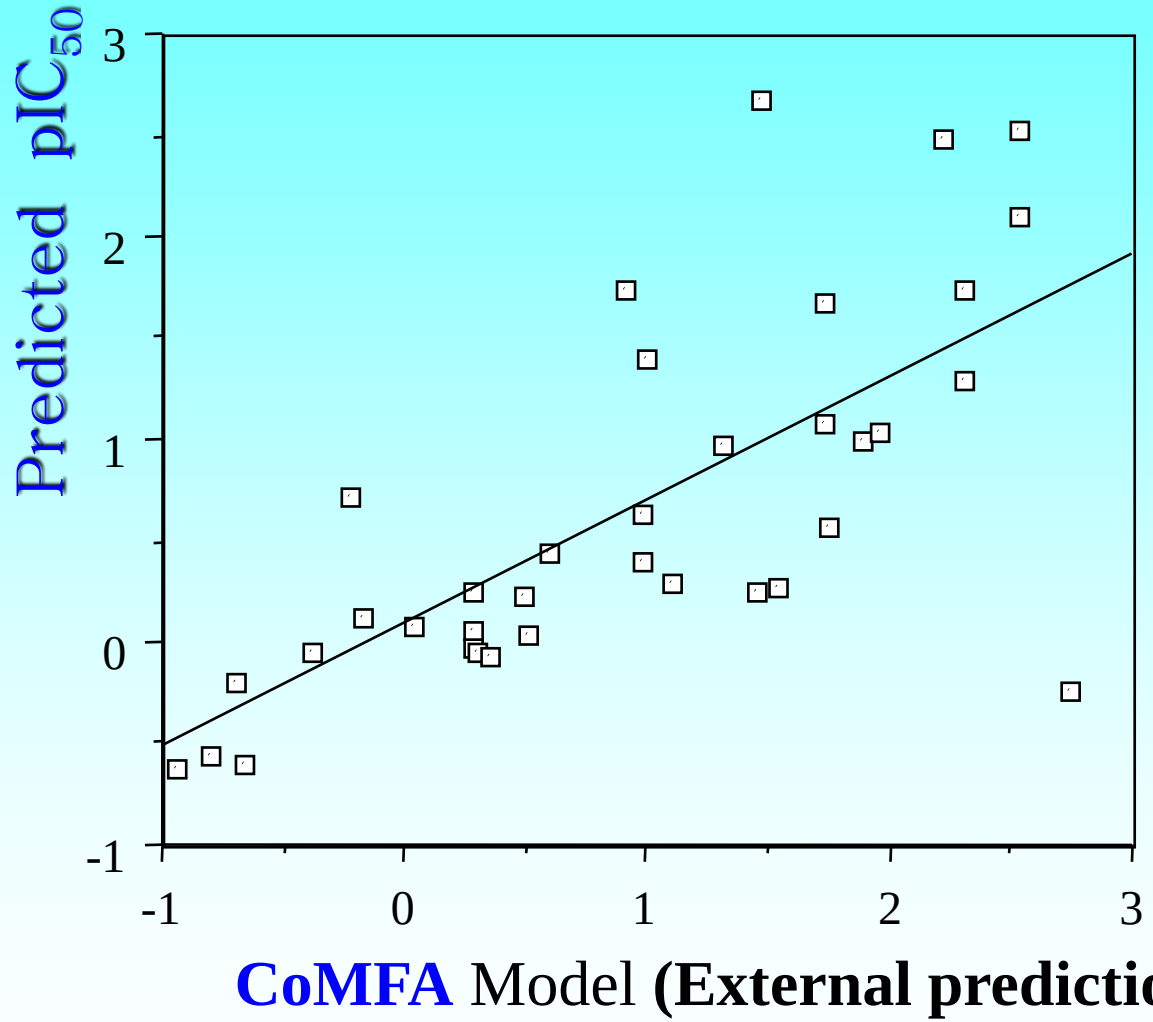
$Q^2 = 0.814$

$K_M = 0.064$ mM

Actual pIC_{50}

Oprea et al., in Ghose & Viswhanandan, Dekker 2001, 233-266

Reliability of Biological Data (2)



HIV-1 Protease Inhibitors
N = 36
Q^2 = 0.487
K_M = 2.0 mM

Oprea et al., in Ghose & Viswhanandan, Dekker 2001, 233-266

QSAR of Oral Absorption

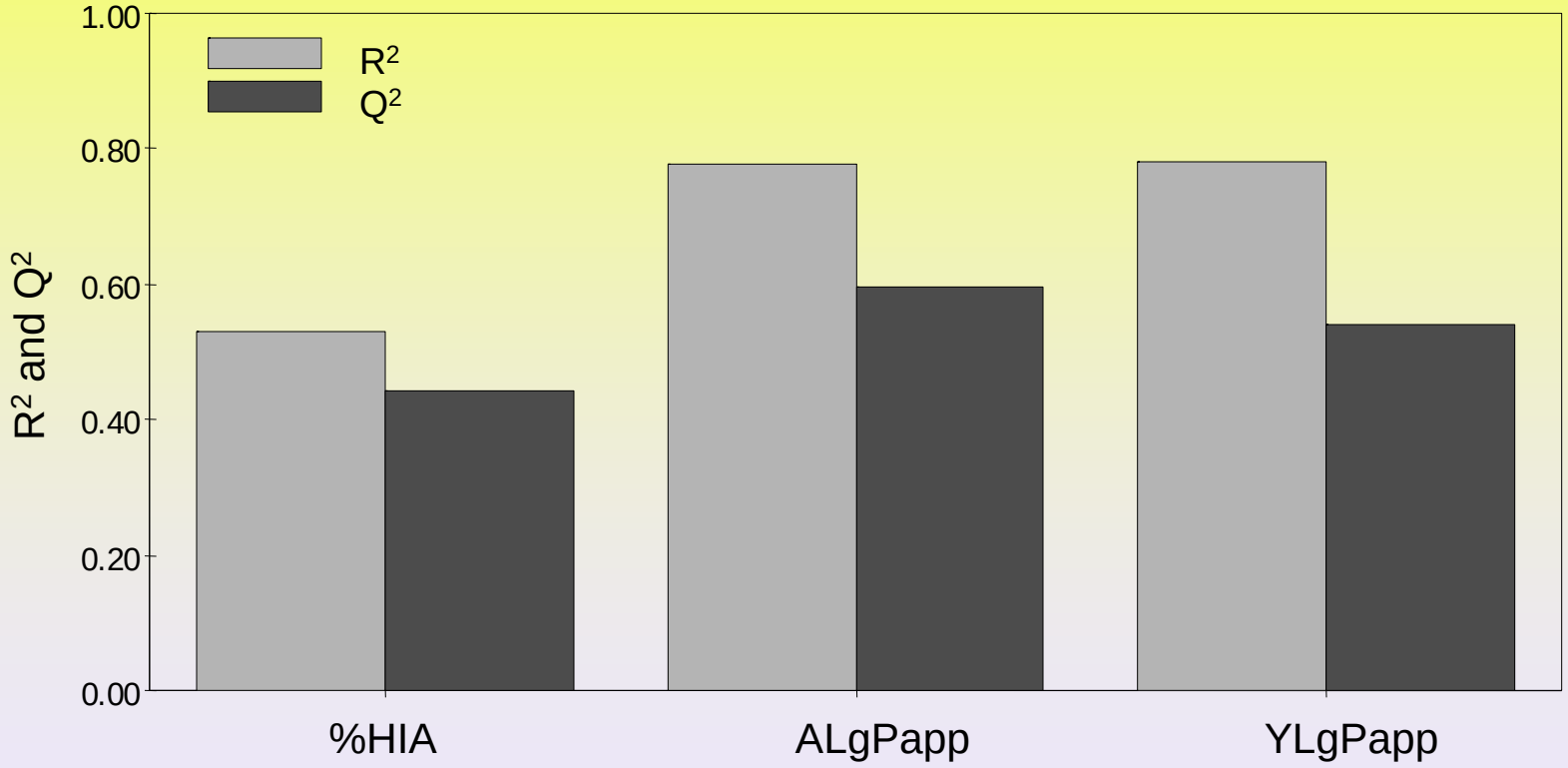
- oral absorption is a major bottleneck in forwarding lead compounds towards clinical development
- Current QSAR paradigm can be summarized as:

  Passive oral absorption = f (LogD7.4, molecular size, H-bond capacity)
- All literature data includes H-bond factors, as well as a measure for hydrophobicity (e.g., LogP o/w).
- HIA = human intestinal absorption; available for 85 drugs
- Caco2 = cell-based assay to evaluate drug permeability in an intestinal cell monolayer; 2 papers available (1998) with 16 and 29 drugs respectively (“LgPapp”)
- Training set: 16 drugs; the rest were predicted.

AstraZeneca

T.I. Oprea, J. Gottfries. *J. Mol. Graph. Model.*, 17, 261-274, 1999

PLS Models



AstraZeneca

T.I. Oprea, J. Gottfries. *J. Mol. Graph. Model.*, 17, 261-274, 1999

Loadings by Name:

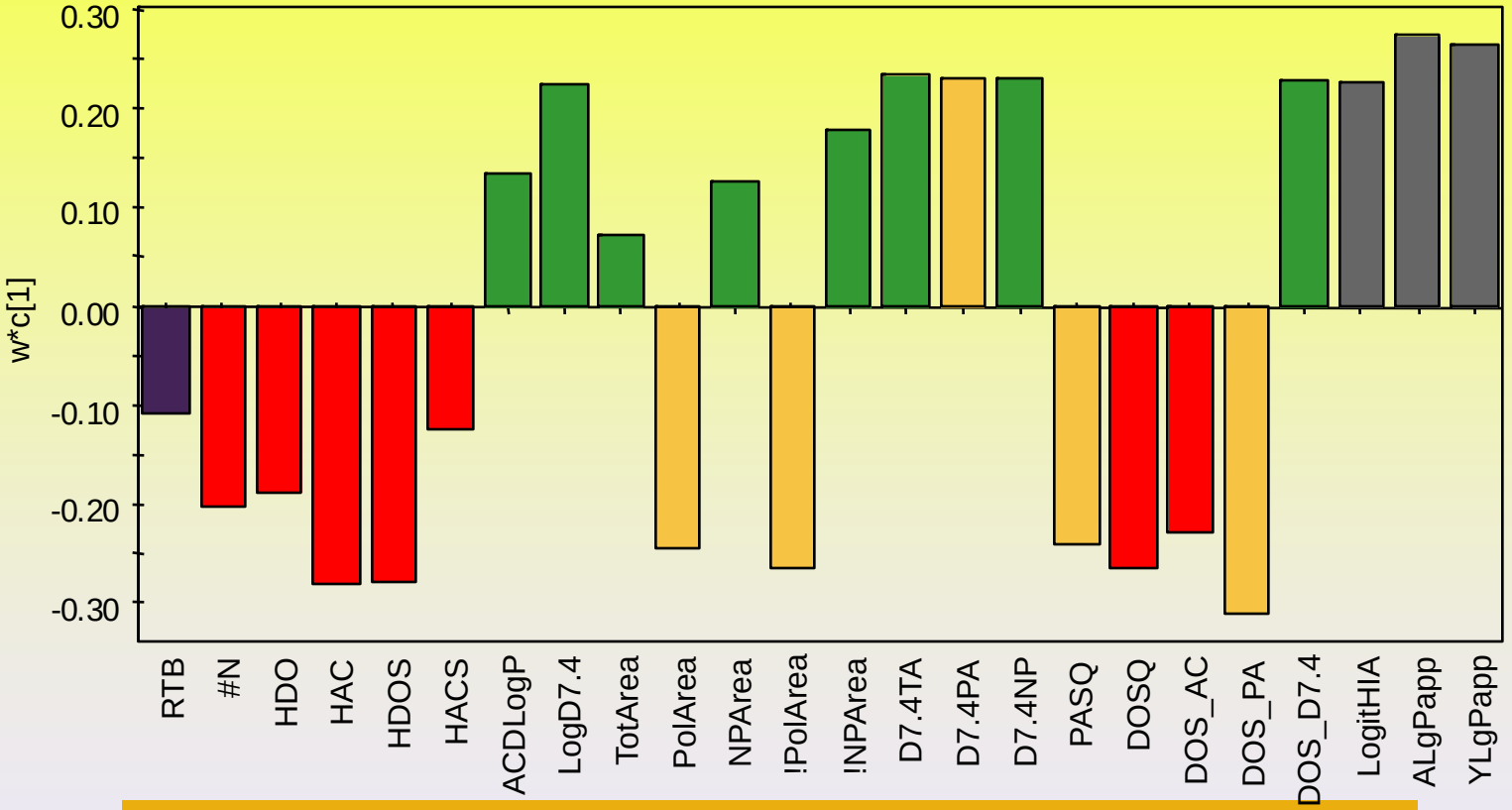
- H-bond donors (number, sum of acidity strength); **HDON[sum] squared**
- H-bond acceptors (number, sum of basicity strength)
- ACDLogP
- RTB & nr Nitrogens
- **LogD7.4**

- Total Area
- **Polar Area** (heteroatoms) and its **squared** value
- Nonpolar area (C, H)
- *cross-terms:*
- **LogD7.4** with all areas
- **HDON[sum]** with HACC[sum], LogD7.4 and with Polar Area

AstraZeneca

T.I. Oprea, J. Gottfries. *J. Mol. Graph. Model.*, 17, 261-274, 1999

%HIA and LgPapp Models



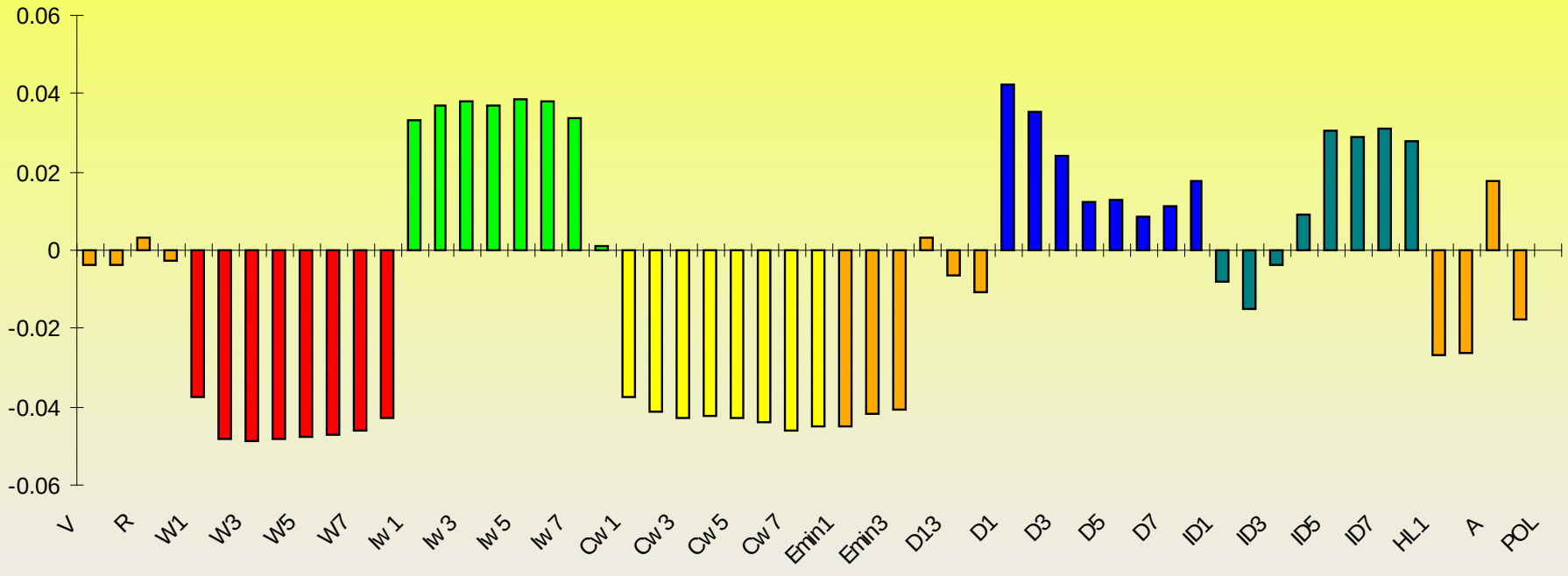
Loadings: w*c [1]

Training Set: 16 compounds; 1 component model

AstraZeneca

T.I. Oprea, J. Gottfries. *J. Mol. Graph. Model.*, 17, 261-274, 1999

%FA Model: VOLSURF Loadings



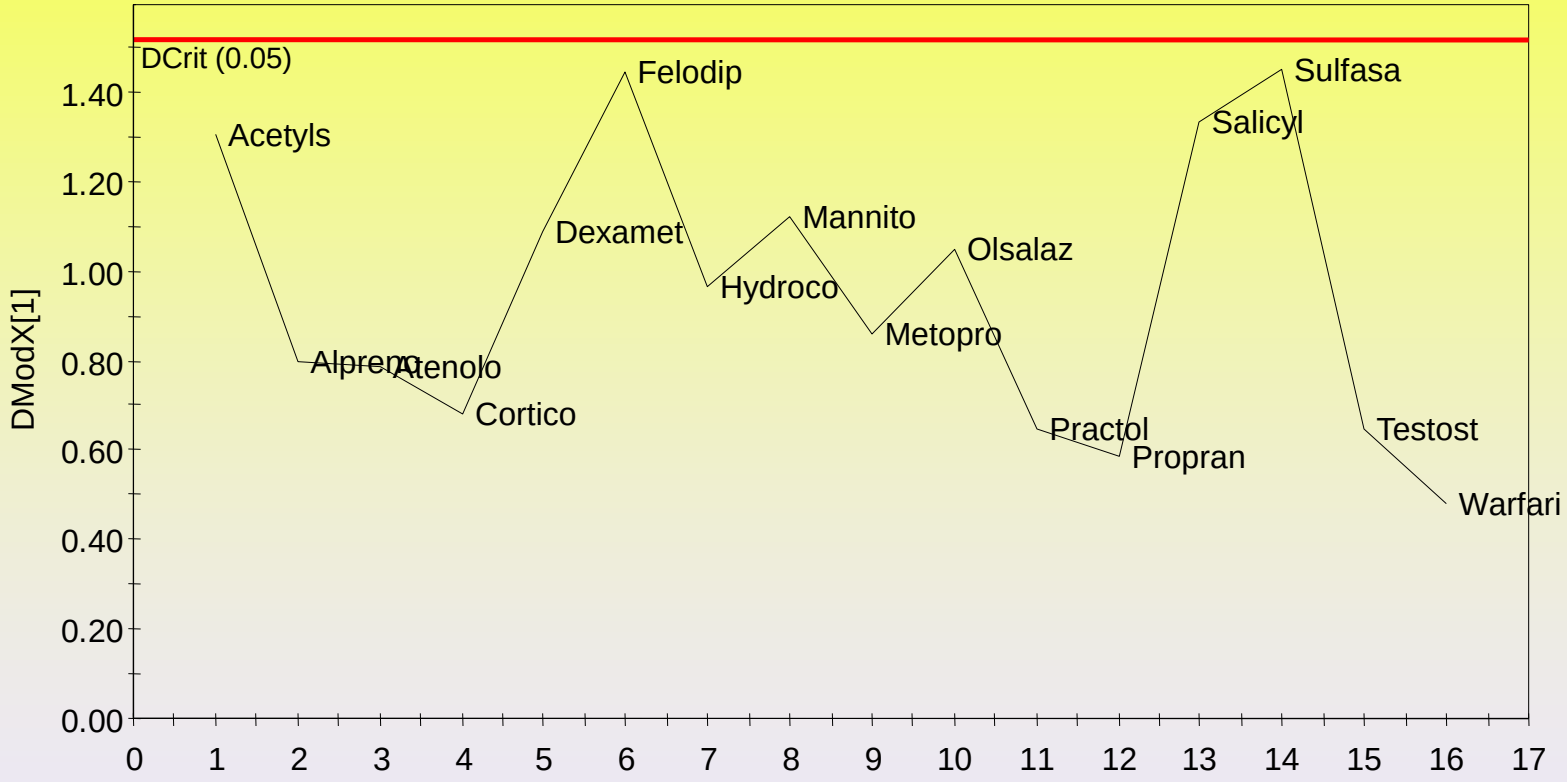
- Hydrophilic Vol. (8 levels)
- Integy moments (8 levels)
- HydroVol./Total Surf. (8 levels)
- Hydrophobic Vol.(8 levels)
- Integy moments Hydroph (8 levels)
- Others

Guba & Cruciani, in Gundertofte & Jørgensen, Kluwer 2000, 89-94

AstraZeneca

T.I. Oprea, J. Gottfries. *J. Mol. Graph. Model.*, 17, 261-274, 1999

Dmod(X): Any Outliers?

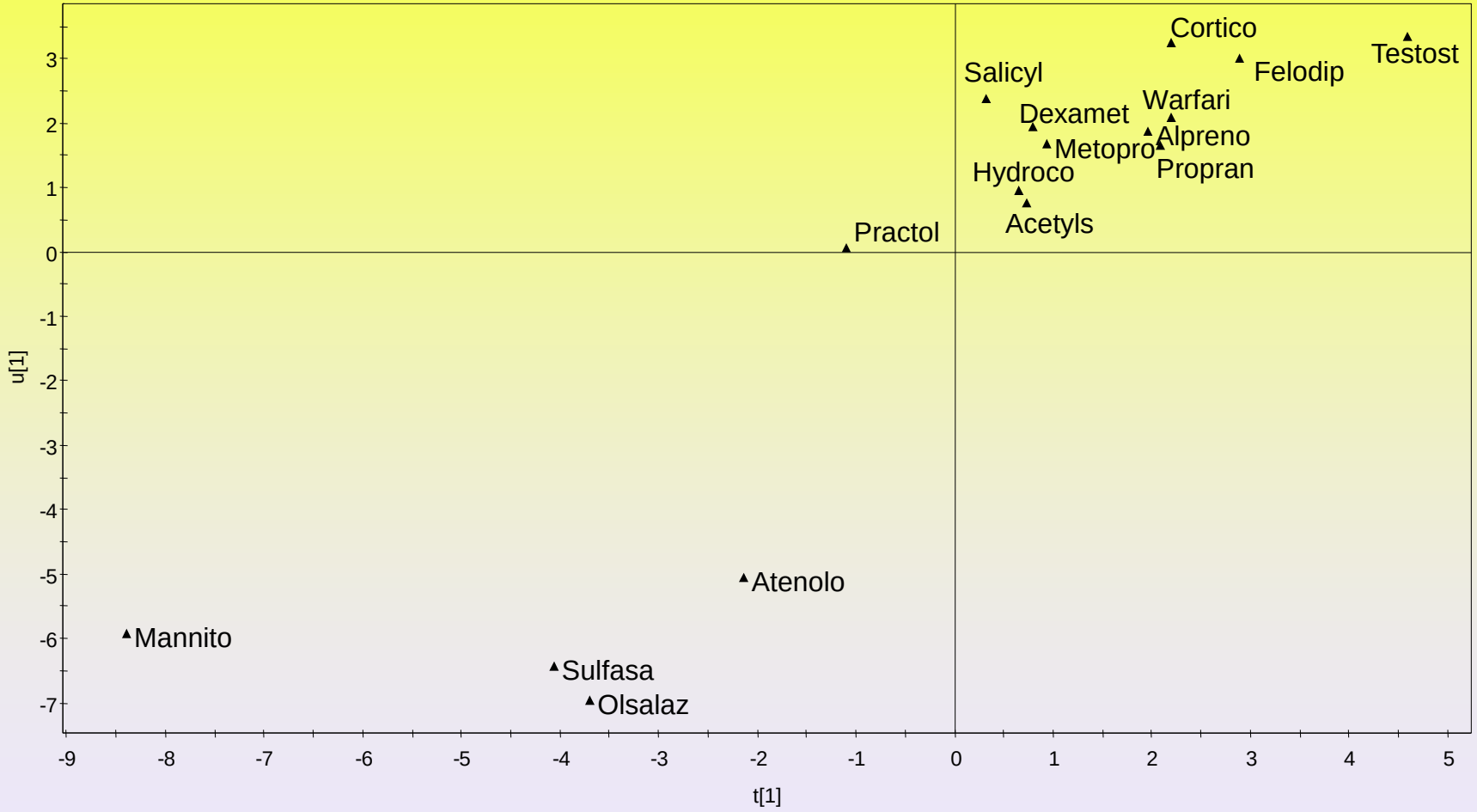



```
Dcrit [1] =  1.5192   , Normalized distances, Non weighted residuals
```

AstraZeneca

T.I. Oprea, J. Gottfries. *J. Mol. Graph. Model.*, 17, 261-274, 1999

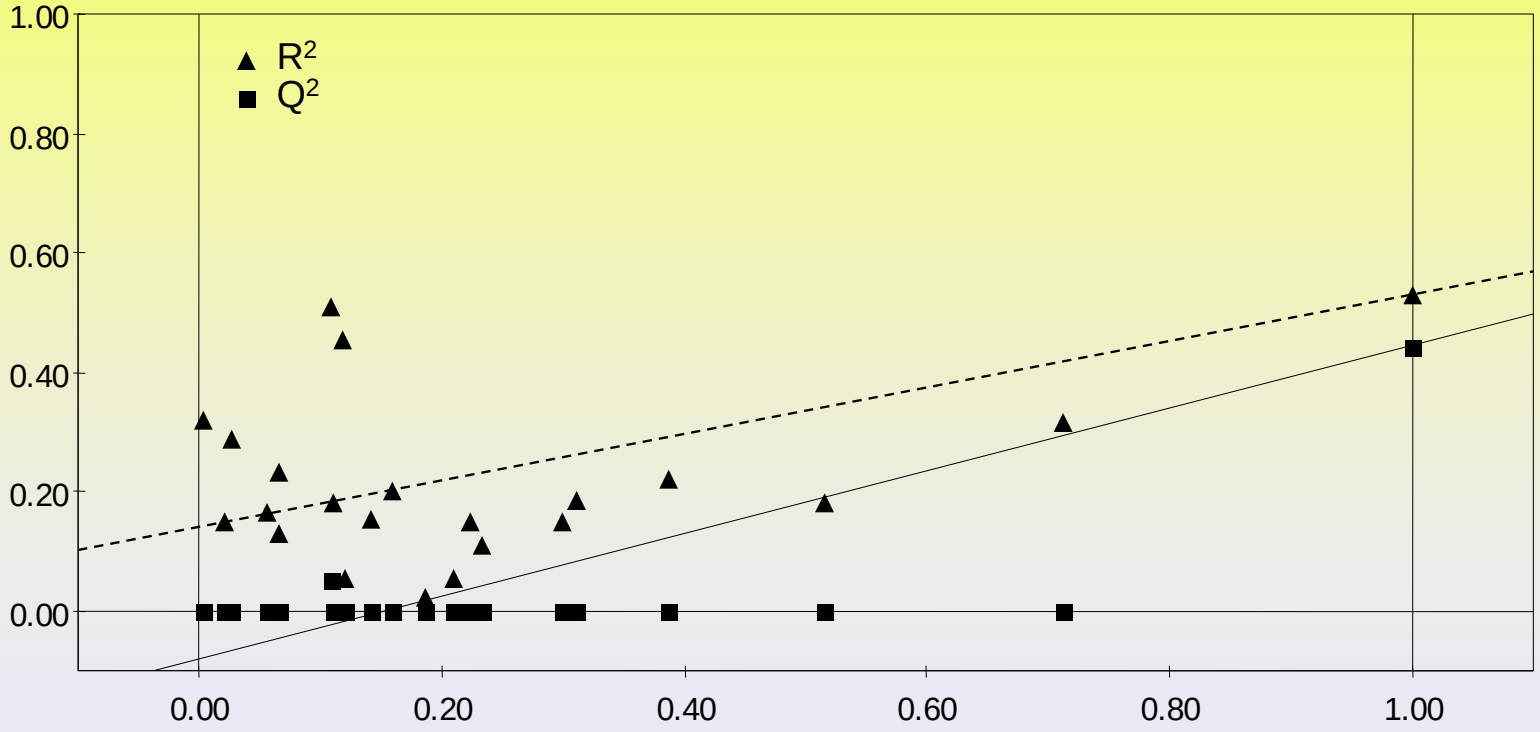
Scores: t[1]/u[1]



AstraZeneca

T.I. Oprea, J. Gottfries. *J. Mol. Graph. Model.*, 17, 261-274, 1999

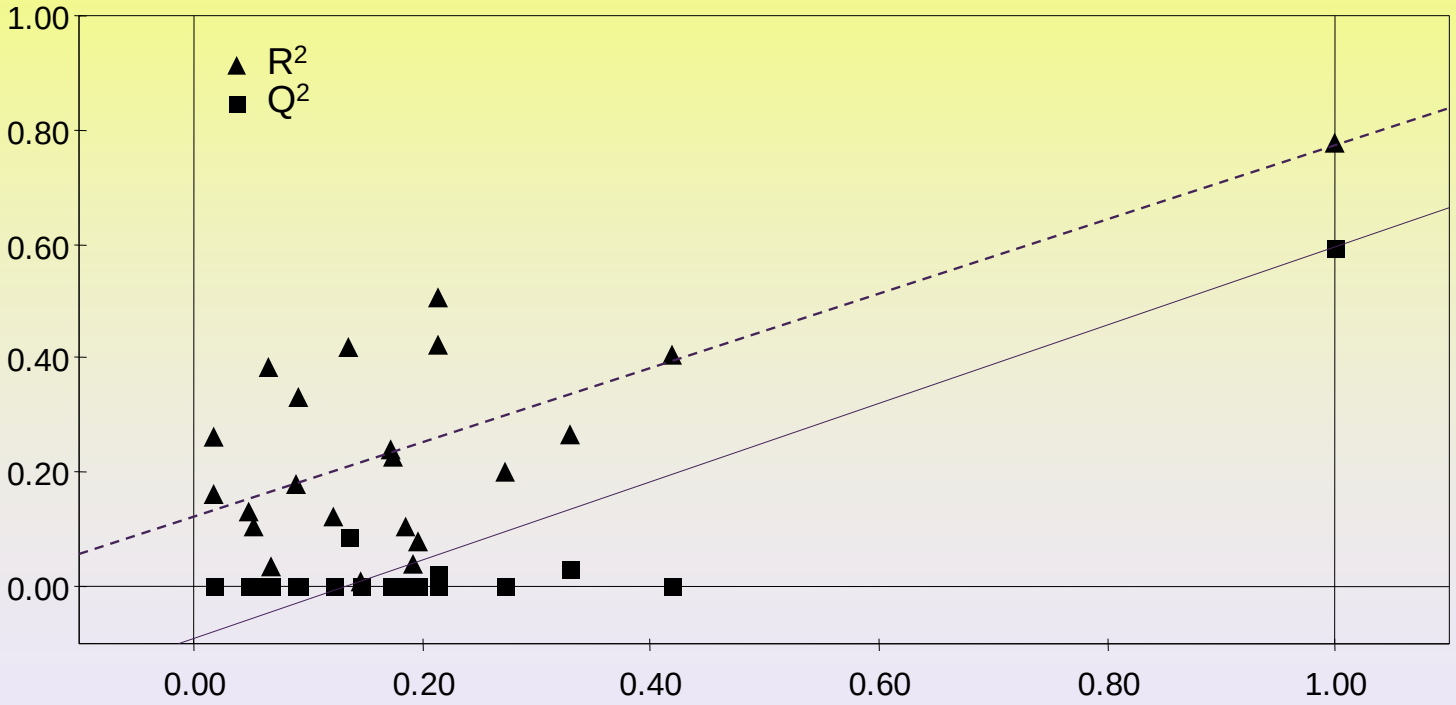
Scrambling HIA Values



AstraZeneca

T.I. Oprea, J. Gottfries. *J. Mol. Graph. Model.*, 17, 261-274, 1999

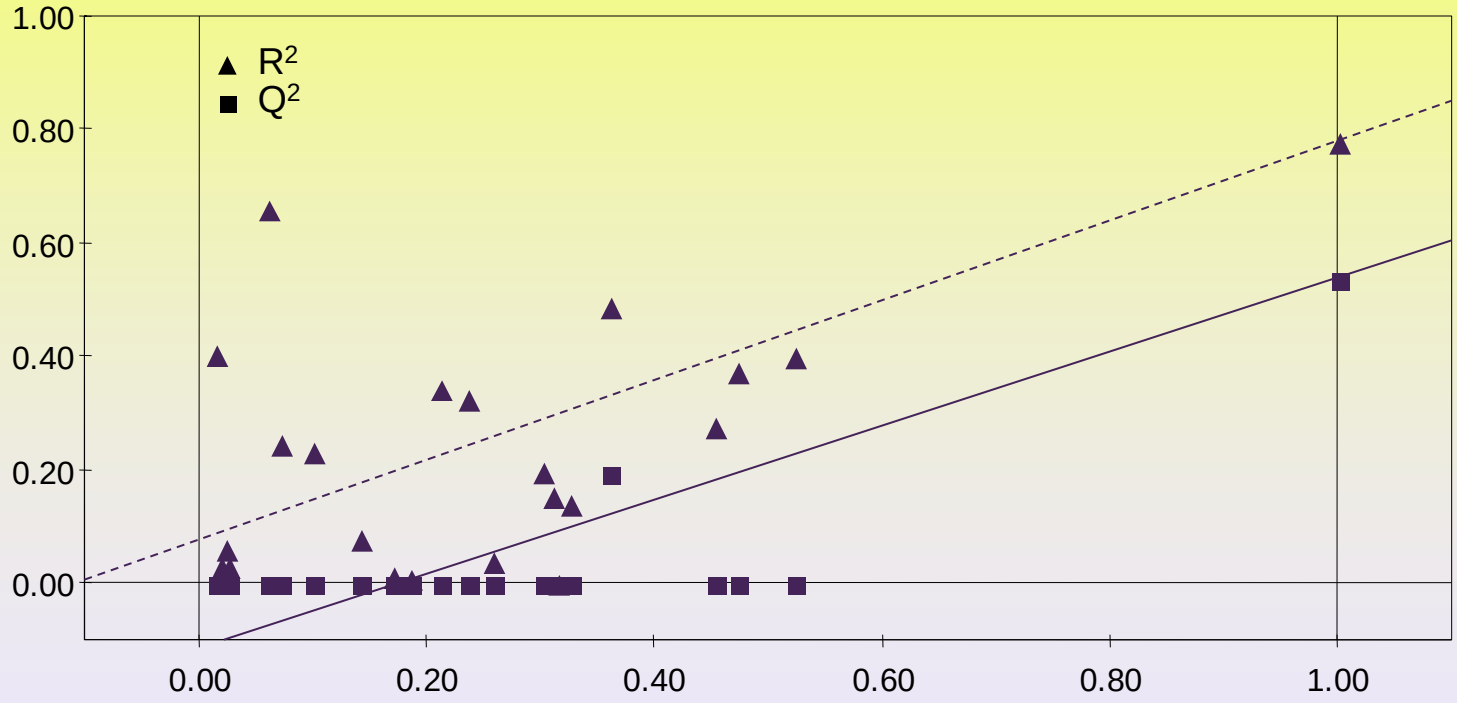
Scrambling ALgPapp Values



AstraZeneca

T.I. Oprea, J. Gottfries. *J. Mol. Graph. Model.*, 17, 261-274, 1999

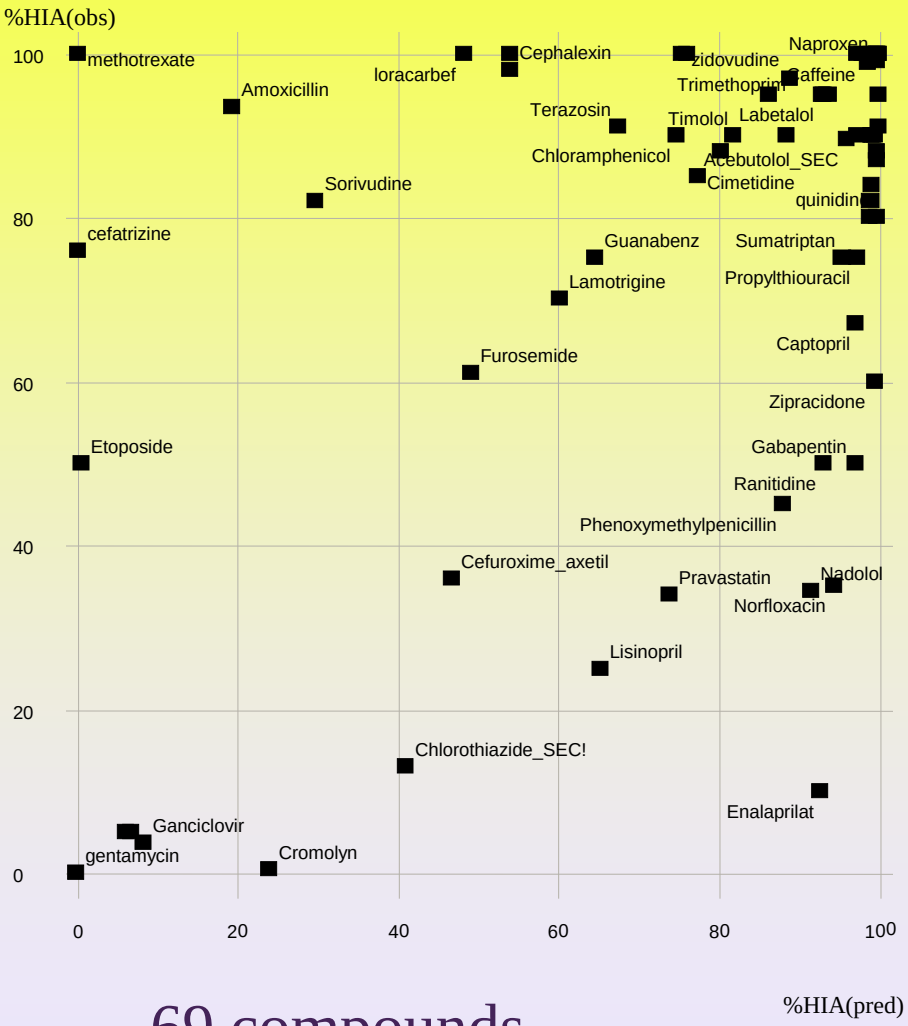
Scrambling YLgPapp Values



AstraZeneca

T.I. Oprea, J. Gottfries. *J. Mol. Graph. Model.*, 17, 261-274, 1999

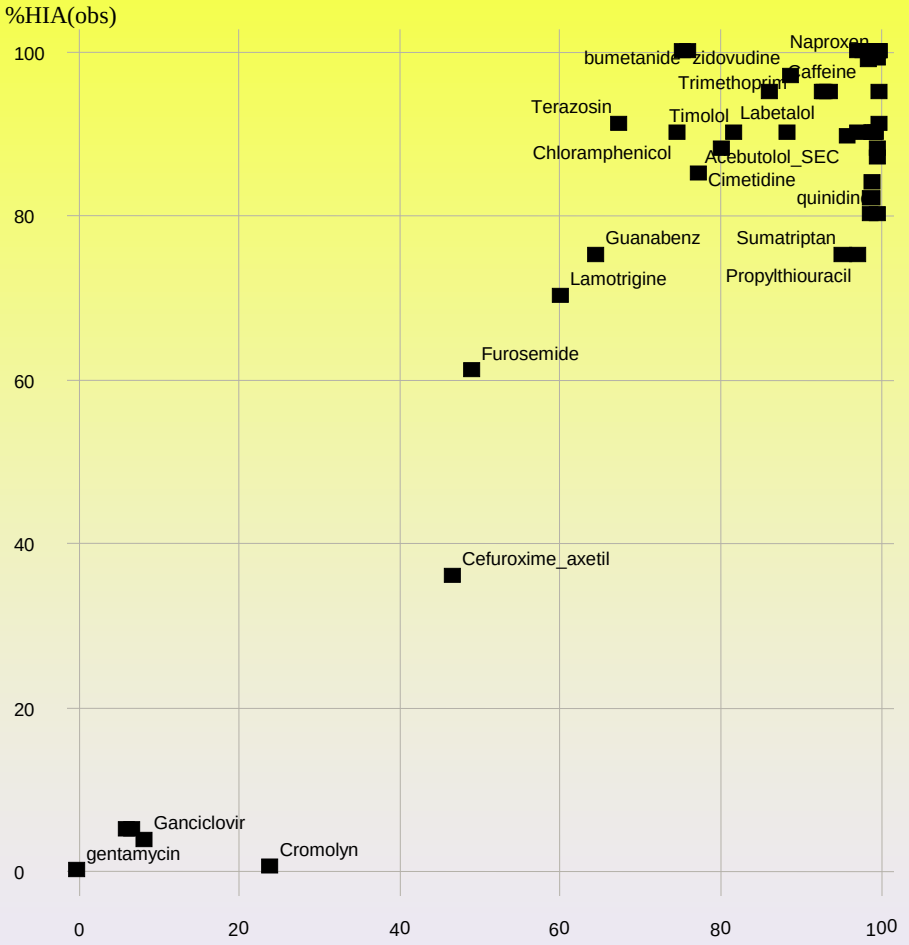
External Prediction for HIA



69 compounds

T.I. Oprea, J. Gottfries. *J. Mol. Graph. Model.*, 17, 261-274, 1999

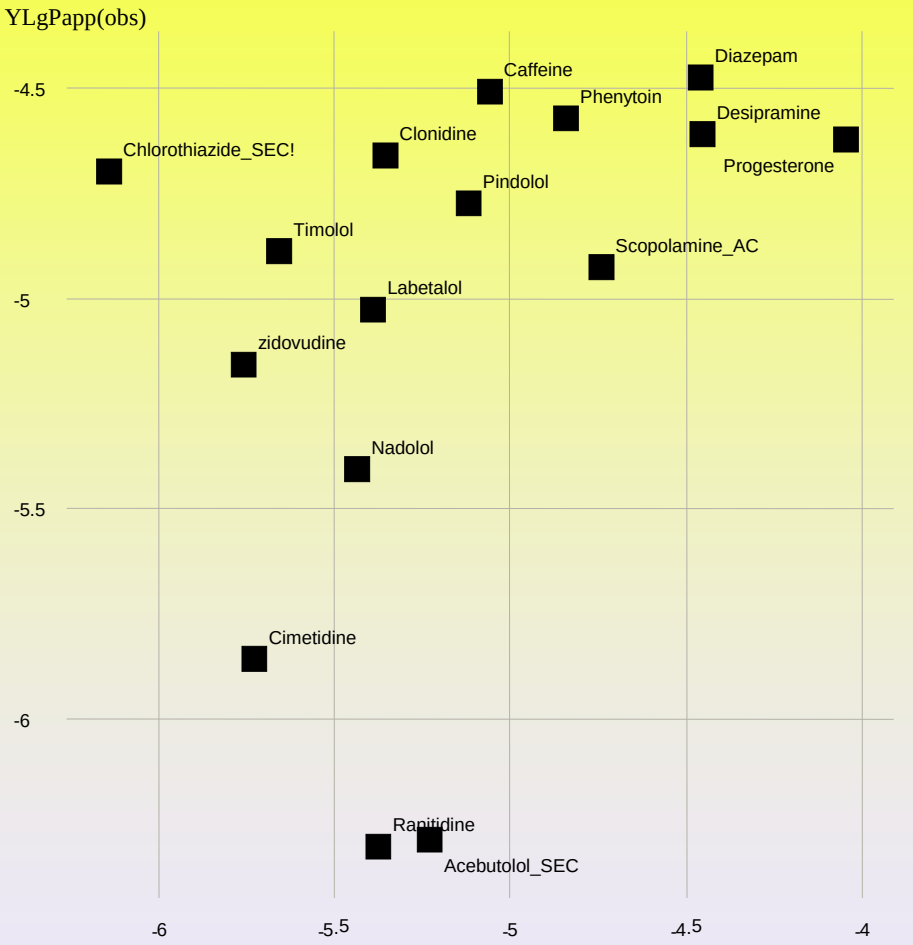
External Prediction for HIA (2)



50 compounds (72.46%)
+/- 25% error

AstraZeneca

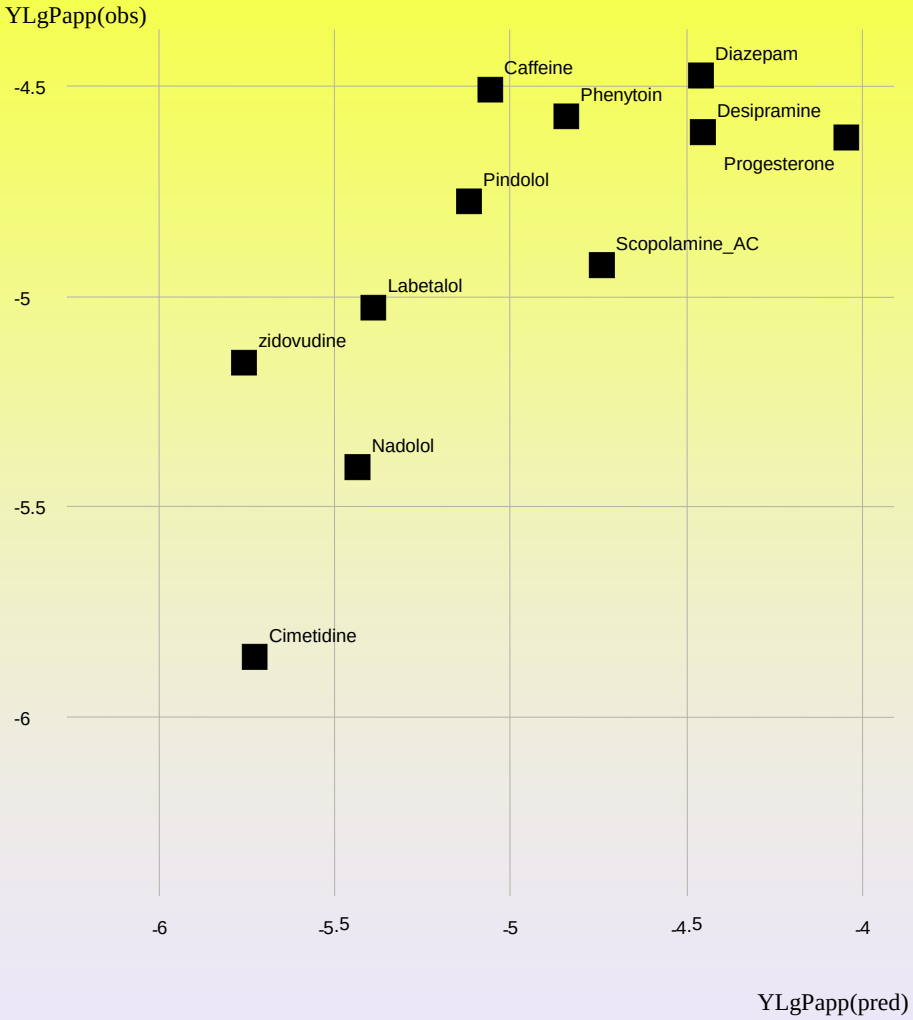
External Prediction for YLgPapp



16 compounds

AstraZeneca

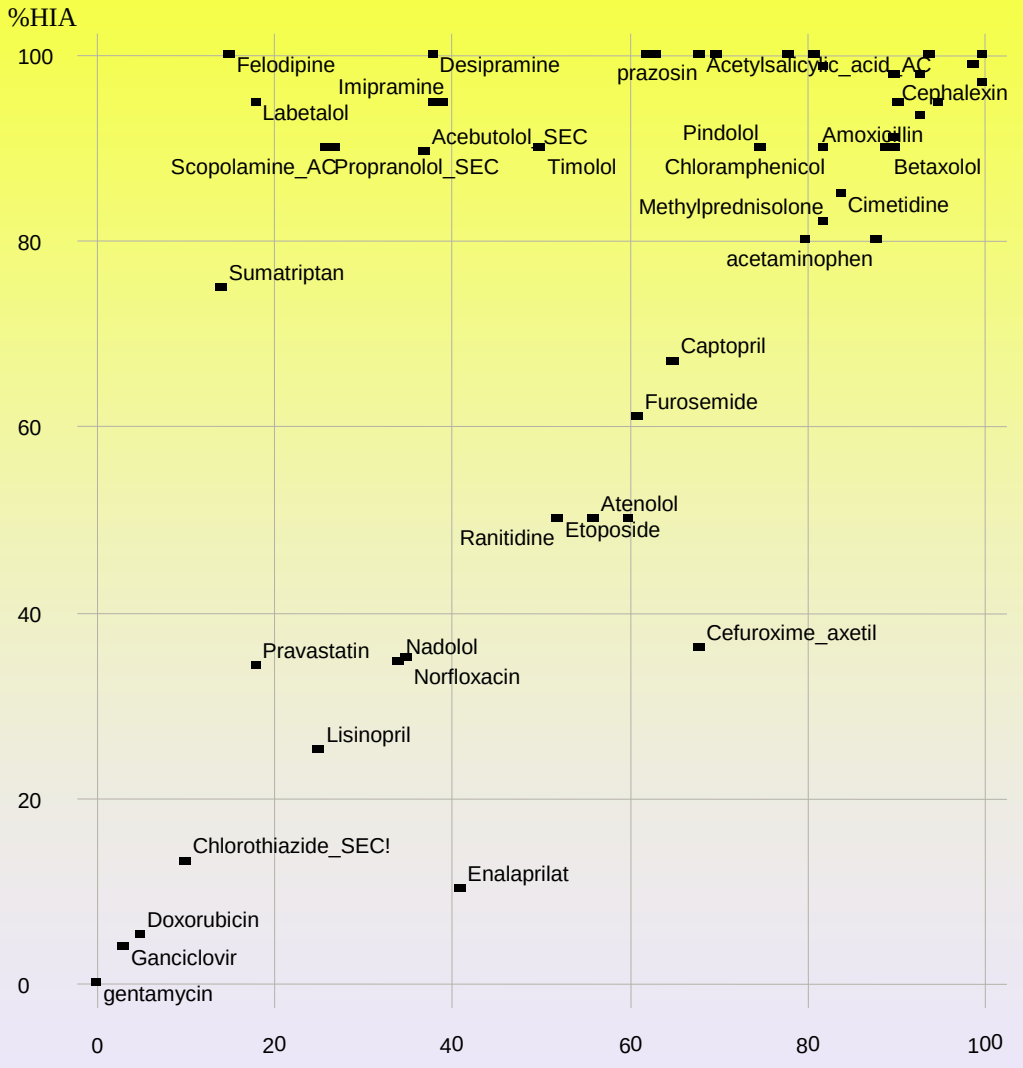
External Prediction for YLgPapp (2)



11 compounds (68.75%)
+/- 0.6 log units error

AstraZeneca

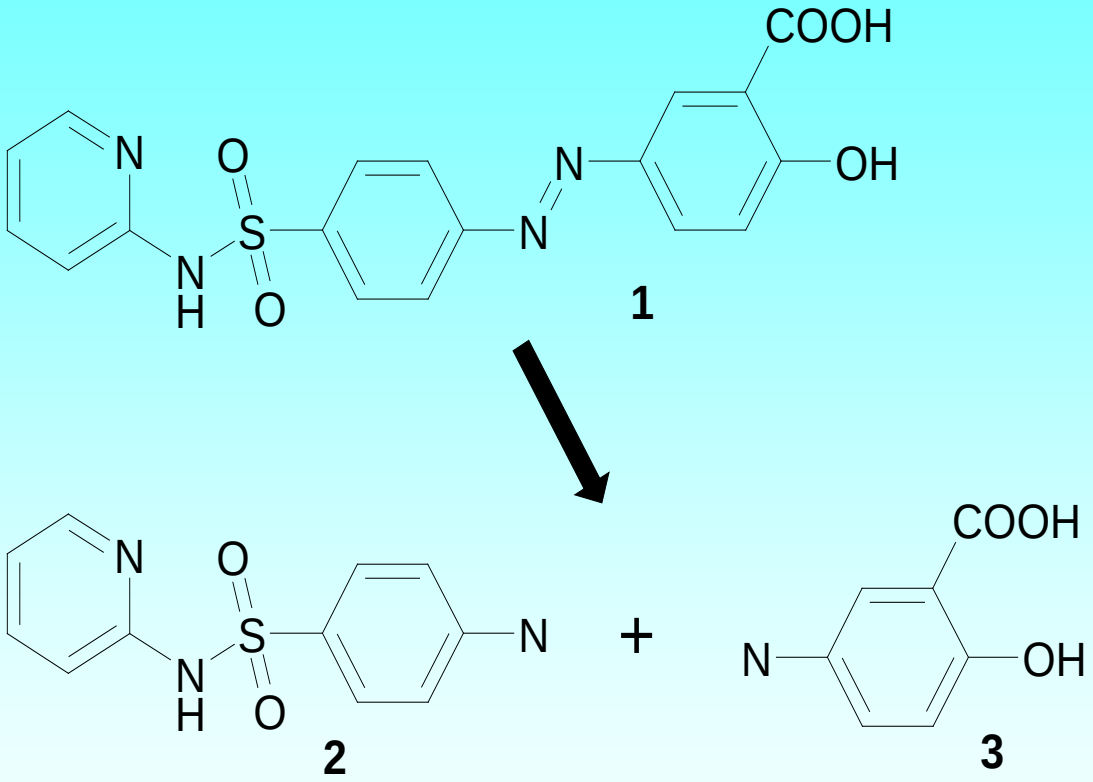
%HIA vs. %Oral absorption



N = 58; R^2 = 0.42; SE= 23%

AstraZeneca

Reliability of Biological Data (3)

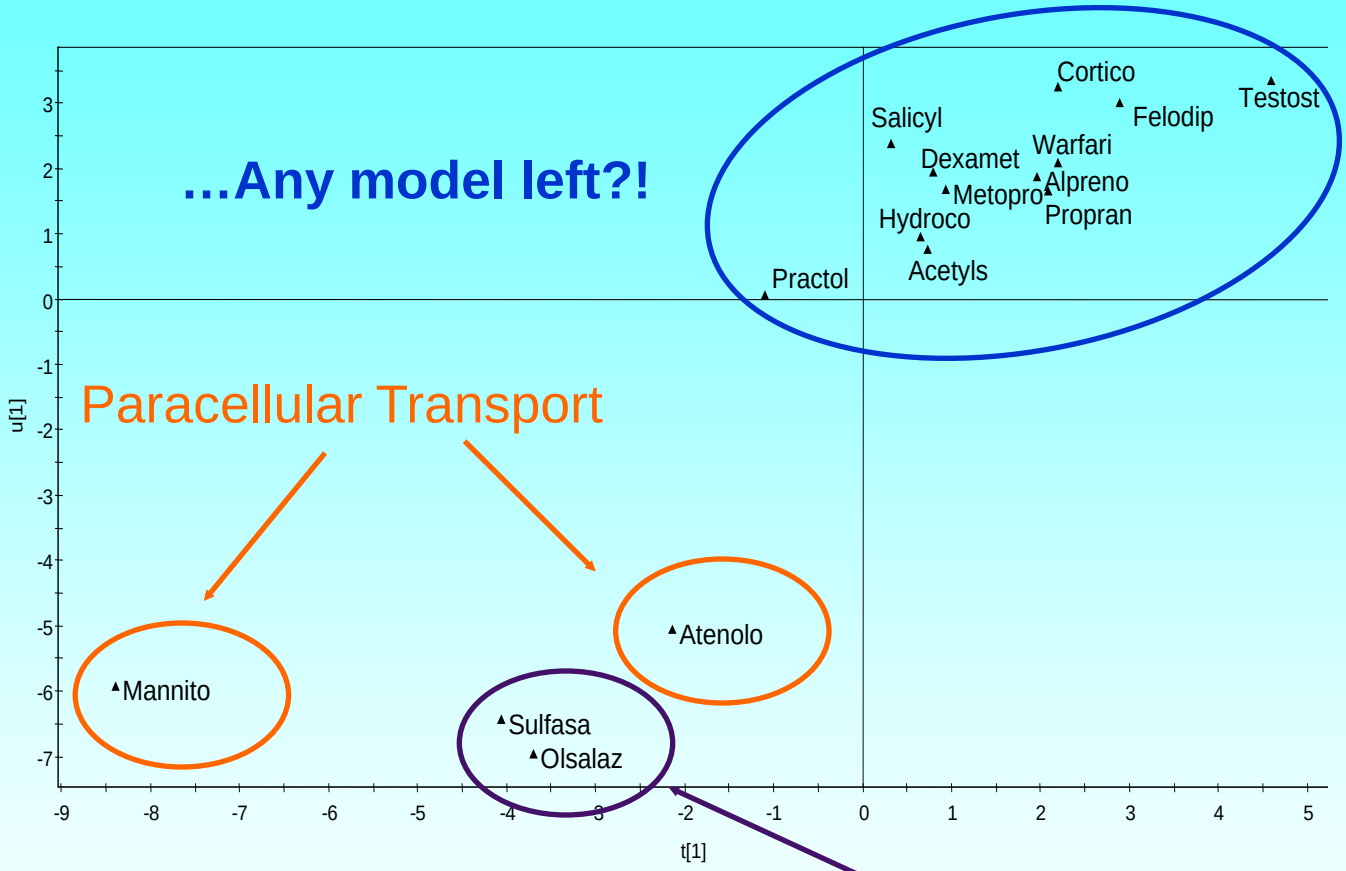


Human Intestinal Absorption:
1: Sulfasalazine (%HIA is 12, not 65)
2: Sulfapyridine (%HIA is 93)
3. 5-aminosalicylic
Bacterial azo bond reduction occurs in the intestine

%HIA Model **(Training set)**

Oprea & Gottfries, J. Mol. Graphics Mod. 1999, 17, 261-274

Reliability of Biological Data (4)



Oral Drug Absorption
N = 16
Q^2 = 0.44 (%HIA)
Q^2 = 0.71 (Caco)

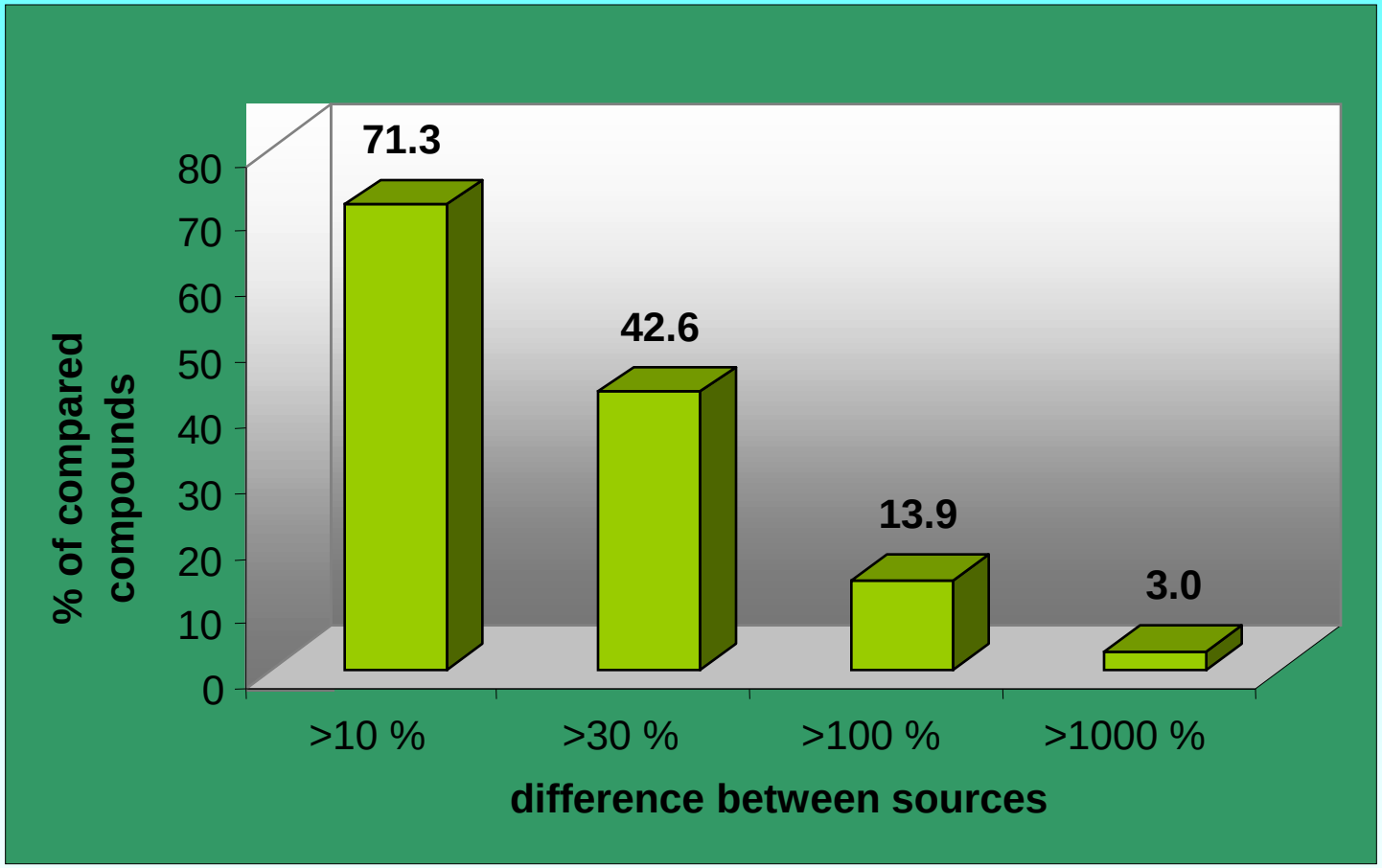
This sigmoidal relationship is frequently described in the literature

Azo bond reduction in the large intestine

%HIA Model **(Training set)**

Oprea & Gottfries, J. Mol. Graphics Mod. 1999, 17, 261-274

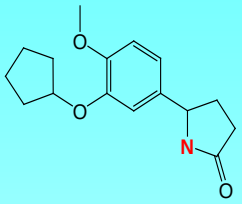
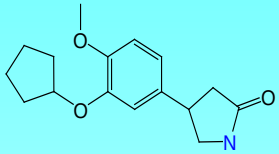
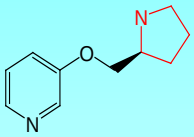
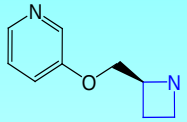
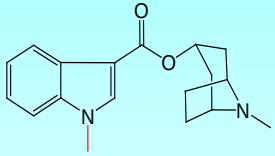
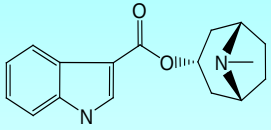
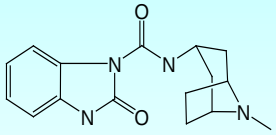
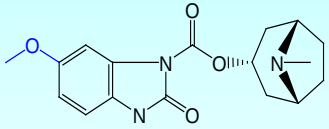
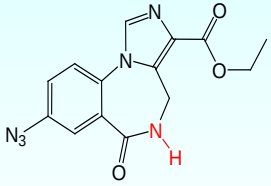
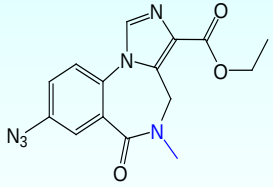
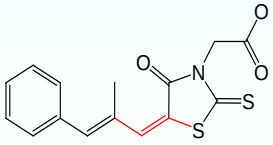
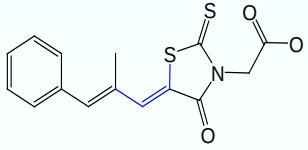
Reliability of Biological Data (5)



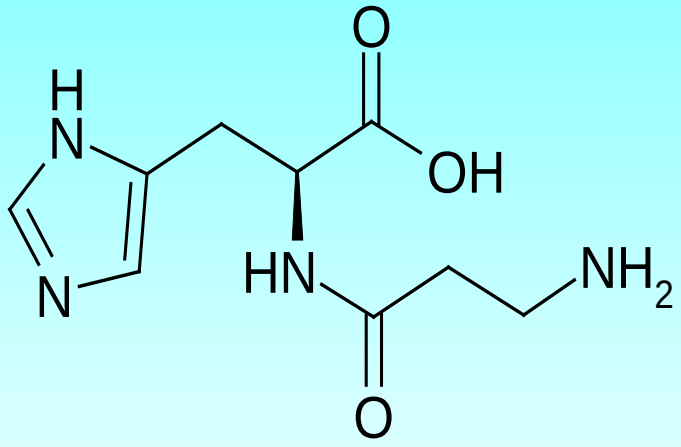
Sources:
Goodman and Gilman 1996 vs. Avery's 1997
Data:
Clearance (202 drugs)

There is poor agreement in terms of clearance data - over 42% of the compounds differ more than 30%

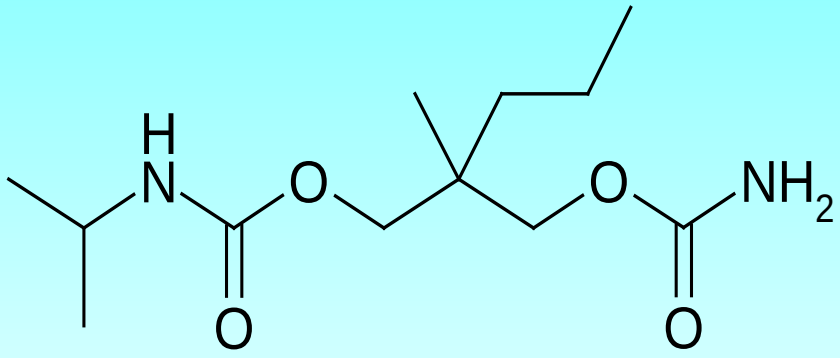
Reliability of Chemical Data (1)

Reference	Published Structure	Corrected Structure	Comment
JMC 37-476 chart 1			*rolipram*: incorrect N atom position
JMC 43-2217 chart 1			*A-85380*: incorrect ring size
-‖- & JMC 36-2645			*tropisetron*: methyl group in plus
-‖-			*DAU-6285*: missing methoxy; N instead O
JMC 37-758 chart 1			*Ro-15-4513*: methyl group missing
JMC 37-787 figure 1			*epalrestat*: E/Z config: E instead Z

Reliability of Chemical Data (2)



"Carisoprodol"
Merck Index 13th ed #1854



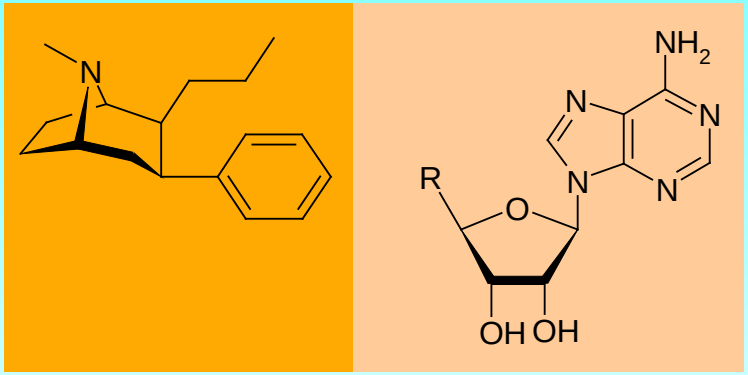
Carisoprodol
correct structure

Disclaimer:
The above error have been corrected in Merck Index 14th edition.
In general, the Merck Index is a reliable source of information.

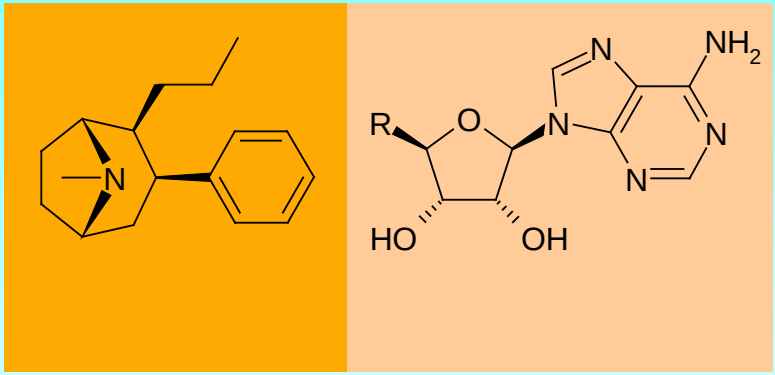
Reliability of Chemical Data (3)

- Chirality: What chemists can interpret, computers are not always able (the "above/below the plane" must be strictly enforced)

Not machine-readable



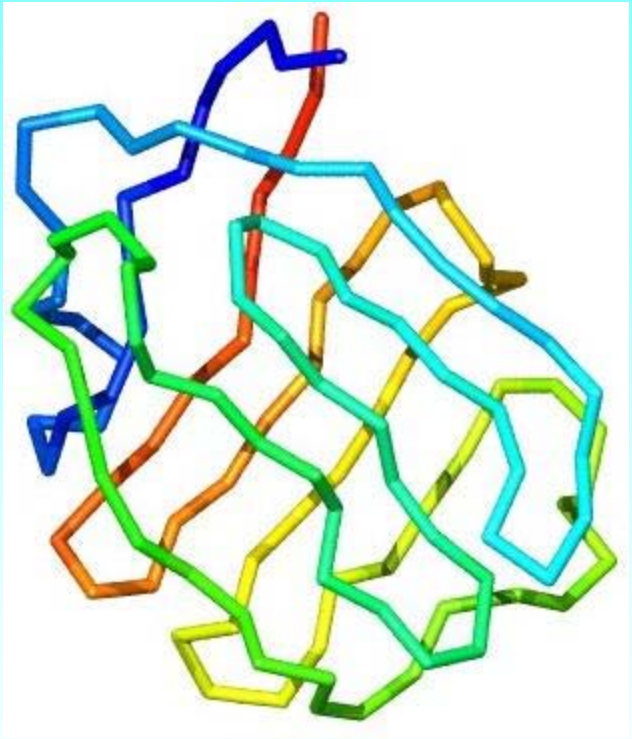
Machine-readable



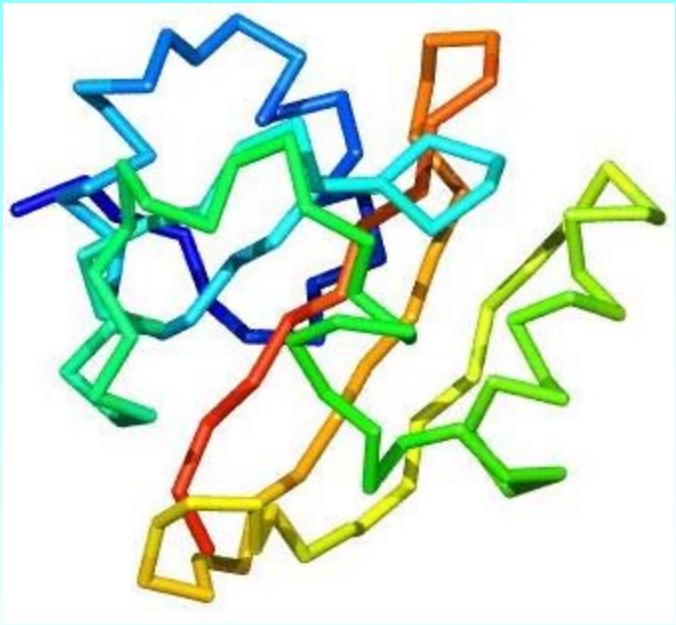
- Missing/altered atoms/substituents – overall error rate above 9%
  - Incorrectly drawn or written structures (3.4%); incorrect molecular formula or molecular weight (3.4%);
  - Unspecified binding position for substituents or ambiguous numbering scheme for the heterocyclic backbone (0.91%);
  - Structures with the incorrect backbone (0.71%);
  - Incorrect generic names or chemical names (0.24%);
  - Incorrect biological activity (0.34%);
  - Incorrect references (0.2%).

Reliability of Structural Data

1PHY 1989 2.4Å



2PHY 1995 1.4Å

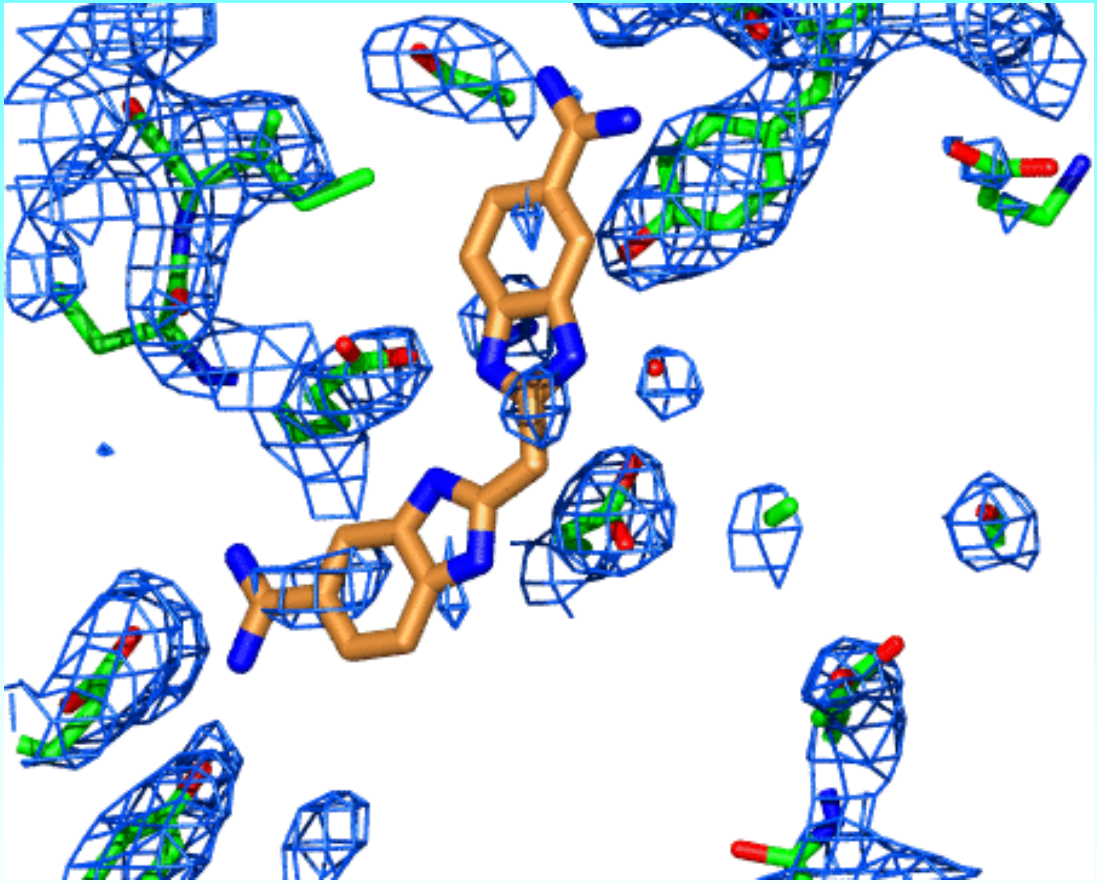


- Photoactive yellow protein from E. Halophila
  - First structure 1PHY.. Wrong
  - Subsequently corrected at higher resolution

A Davis, S Teague G Kleywegt Angew. Chem. Int. Ed. 2003

Reliability of Structural Data (2)

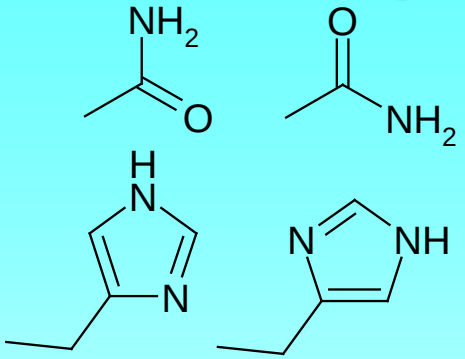
"Where there is no chicken wire, there are no electrons..atoms"



- 1FQH now withdrawn from PDB

A Davis, S Teague G Kleywegt Angew. Chem. Int. Ed. 2003

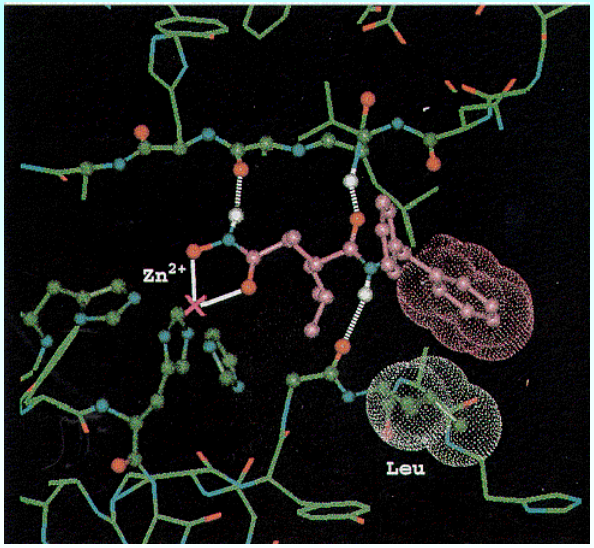
Reliability of Structural Data (3)

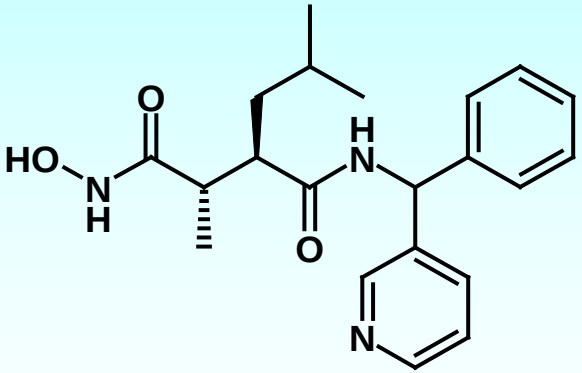


e.g. glutamine, asparagine

e.g. histidine

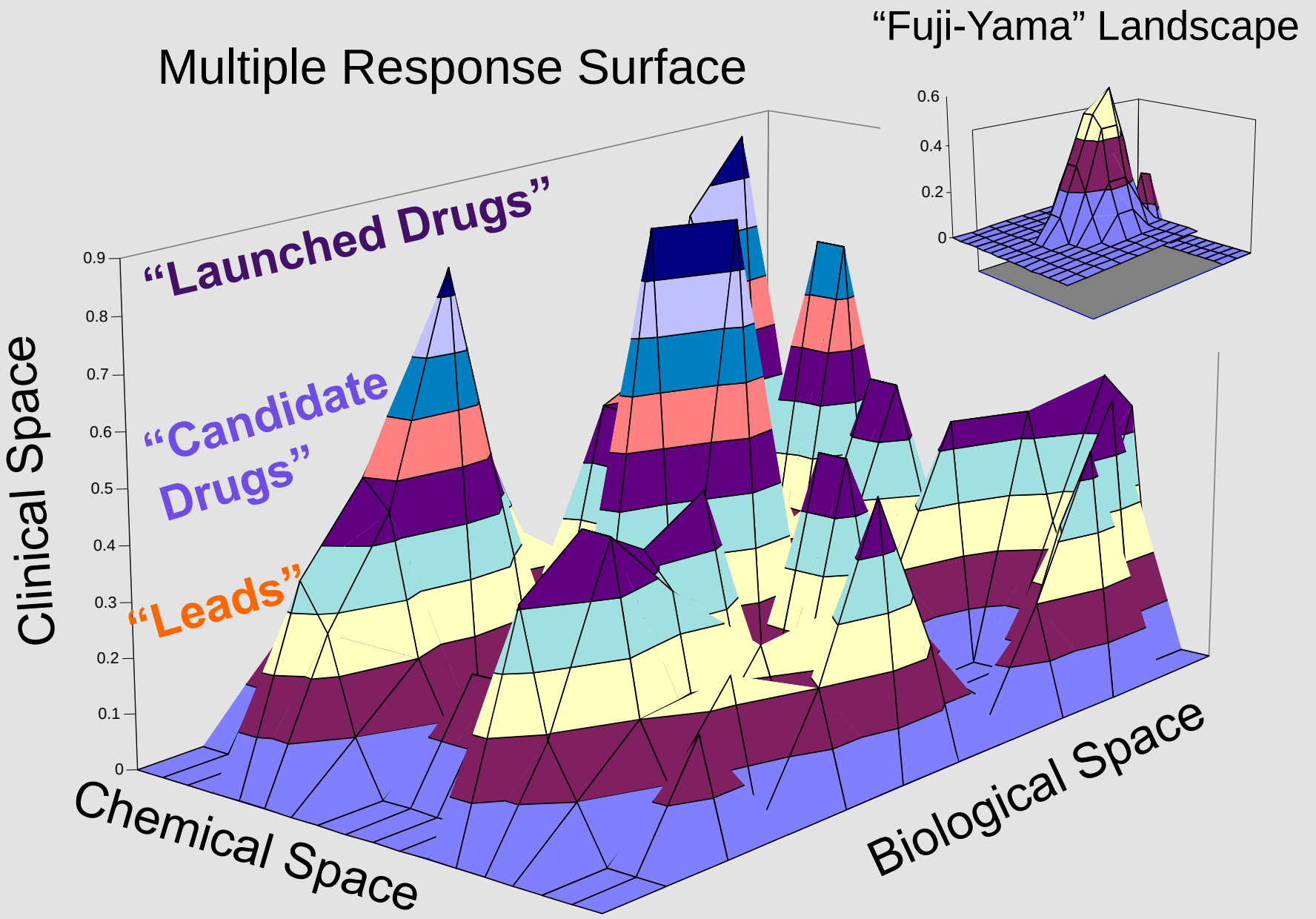
- NH_2 & O can't be distinguished from density as isoelectronic
- PDBREPORT suggest 15% in Protein databank likely incorrect





- N/C cannot normally be distinguished from density

Why Drug Discovery is Difficult



Take Home Message 5 (or 6…)
The Importance of Accurate Information

- With one source, we have information
- With two sources, we can have confirmation or confusion
- With several sources, knowledge emerges
- **Accurate Information** is important – hence we tend to "trust" certain newspapers, TV stations or scientific journals (the peer-review system regulates that).
- If something is really important to you (*) then consult multiple sources and verify that your assumptions are correct

- (*) e.g., who is on my PhD committee; what's my girlfriend's birthday; what are the side-effects for the medicine my parents are taking; is this formula/algorithm correct? *etc.*

Exhaustive Enumeration of Scaffold Topologies

How Large is Chemical Space?

Lord of the Rings

- The chemical space of small molecules (CSSM) has, to date, been systematically mapped for MW <= 160
- *Fink, T., Brugesser, H., Reymond, J.L.*
  Angew. Chem. Int. Ed. ***2005****, 44:1504-1508*
- The Quest for the Rings led to the generation of a virtual library of ~600k aromatic heterocycles which has been characterized
- *Ertl, P.; Jelfs, S.; Muhlbacher, J.; Schuffenhauer, A.; Selzer, P.*
  J Med Chem ***2006****, 49:4568-4573*

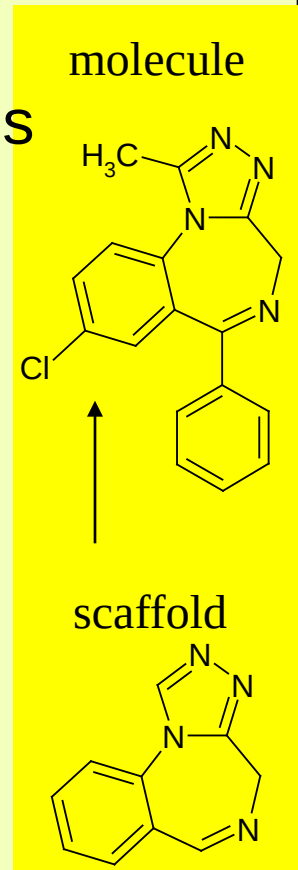
Lord of the Rings [2]

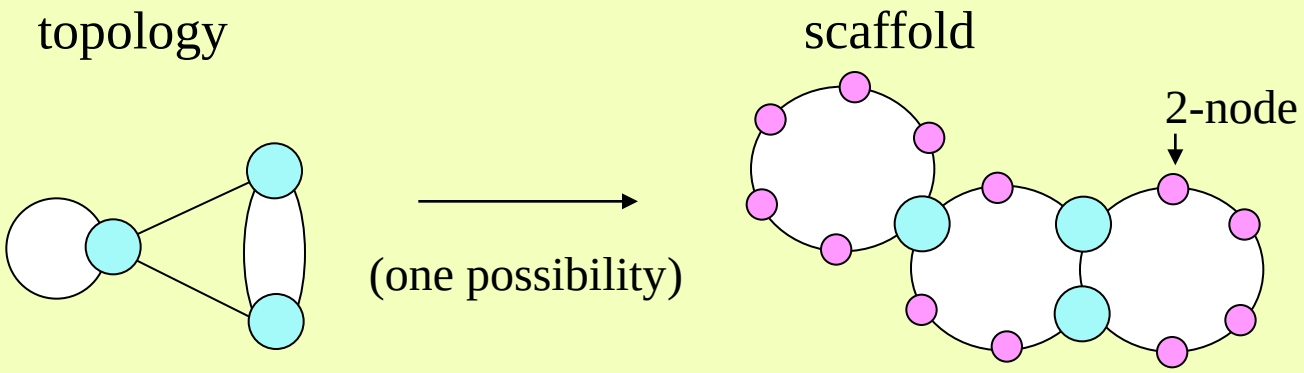
- **Systematic Enumeration of Molecular Topologies**
- *Sara Pollock, Vagelis Coutsias, Mike Wester, T. Oprea, in preparation*
- Two independent methods have been designed and used to perform the exhaustive enumeration of all possible topologies for up to 8 rings:
  - Start with the two 2-ring graphs consisting of all 3-nodes.
  - From there, generate all 3-ring graphs consisting of all 3-nodes.
  - Continue in this manner until arriving at the desired number of rings.
  - At that point, systematically fuse 3-nodes together to form 4-nodes.

S.N. Pollock, E.A. Coutsias, M.J. Wester, T.I. Oprea. *J. Chem. Info. Model.* 48:1304 - 1310, 2008

Lord of the Rings [3]

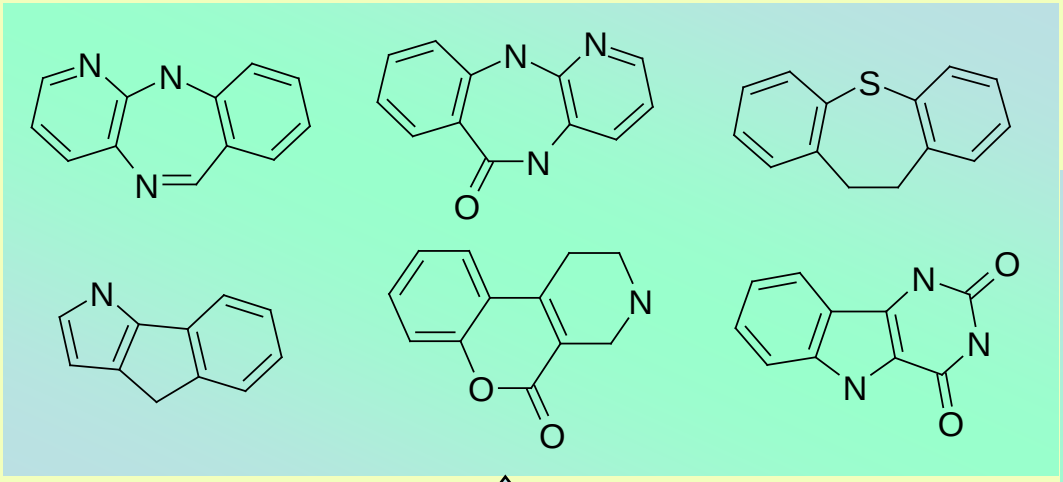
- A topology represents the basic geometry of a molecule by defining the number of rings and how they connect.
- A chemical structure can be reduced to its scaffold by removing its branches.
- A scaffold can be reduced to its topology by removing geometrically irrelevant nodes (2-nodes).
- To obtain a chemical scaffold from a topology, 2-nodes may be added to any edge of the topology.
- This is a straightforward combinatorial problem.

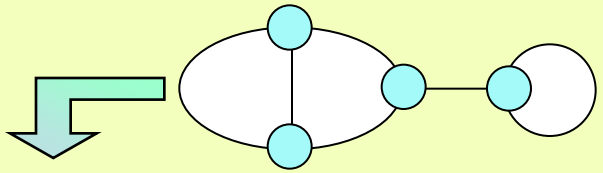


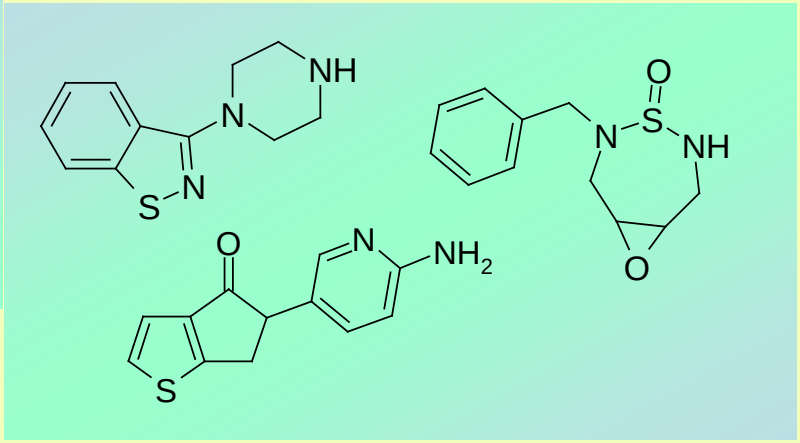


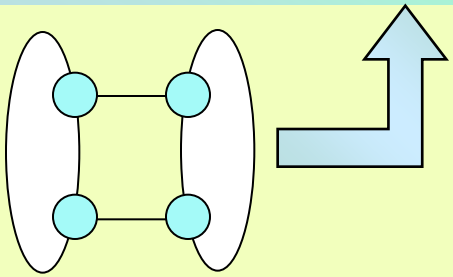
S.N. Pollock, E.A. Coutsias, M.J. Wester, T.I. Oprea. *J. Chem. Info. Model.* 48:1304 - 1310, 2008

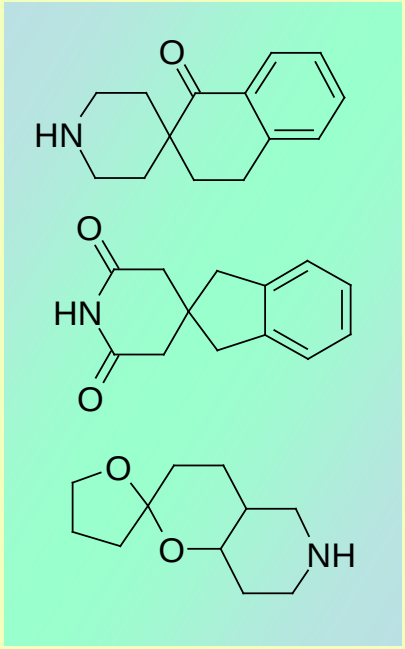
Topologies vs. Scaffolds: Examples

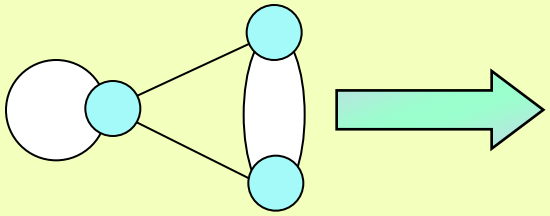


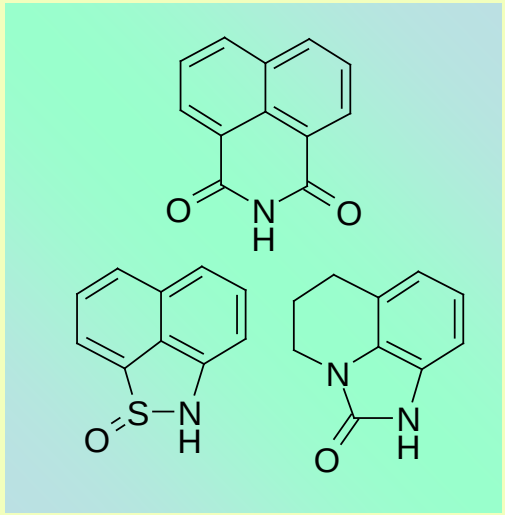


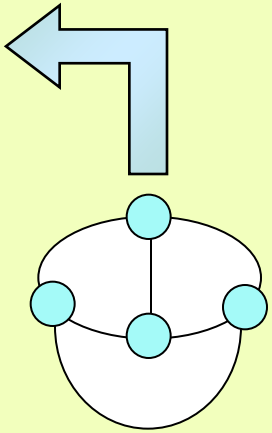










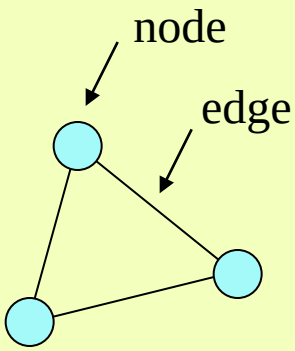


S.N. Pollock, E.A. Coutsias, M.J. Wester, T.I. Oprea. *J. Chem. Info. Model.* 48:1304 - 1310, 2008

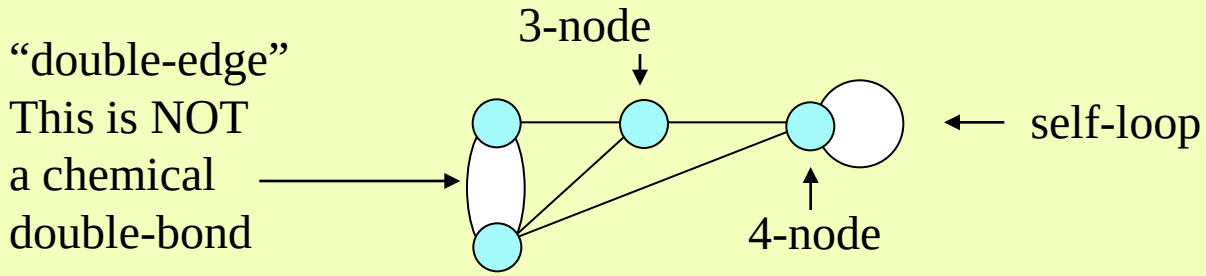
Lord of the Rings [4]

GRAPH

- A topology is a graph in which every node has either 3 or 4 associated edges.

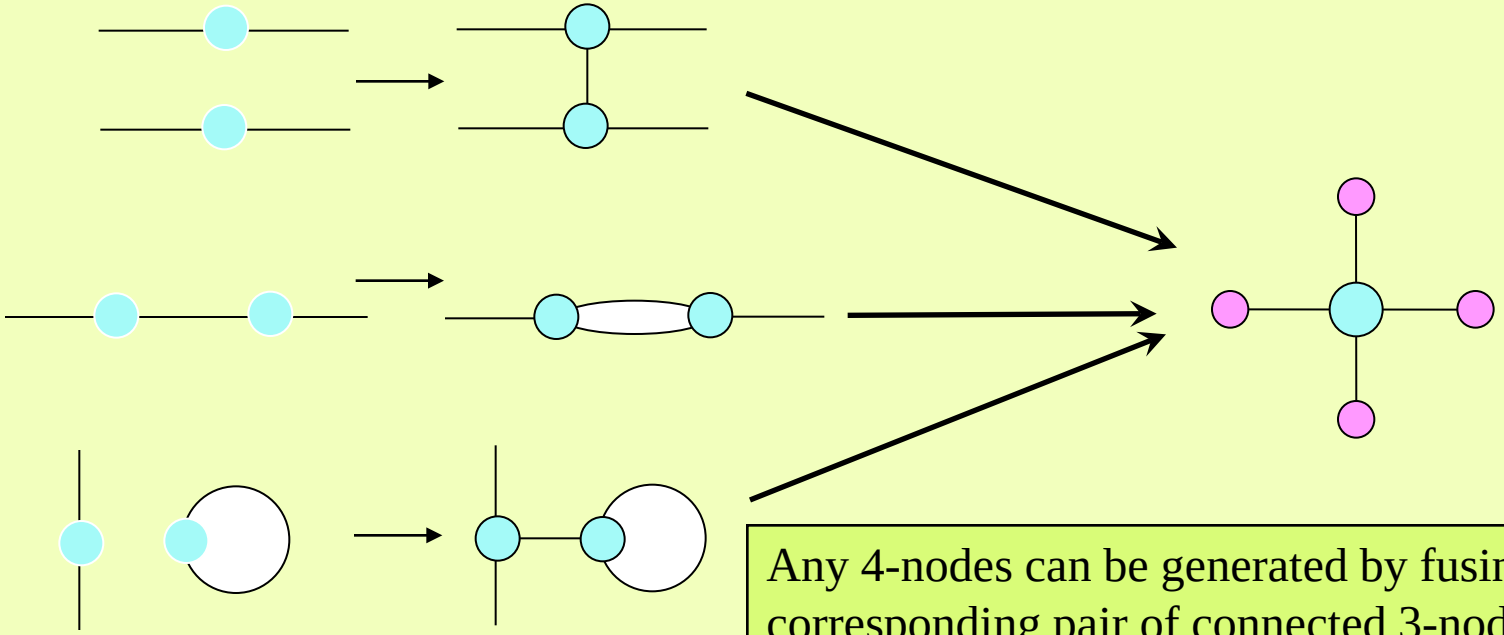


TOPOLOGY



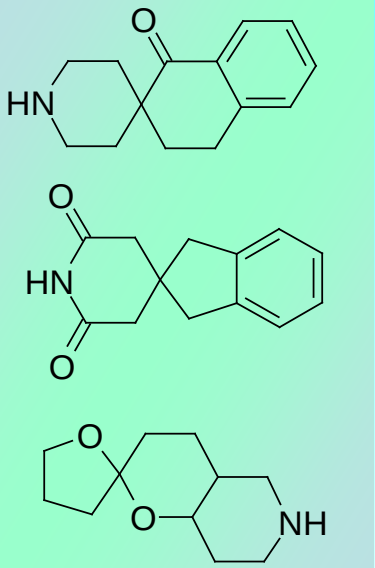
We generated all possible configurations of 3-nodes (3-nodes increment in pairs).

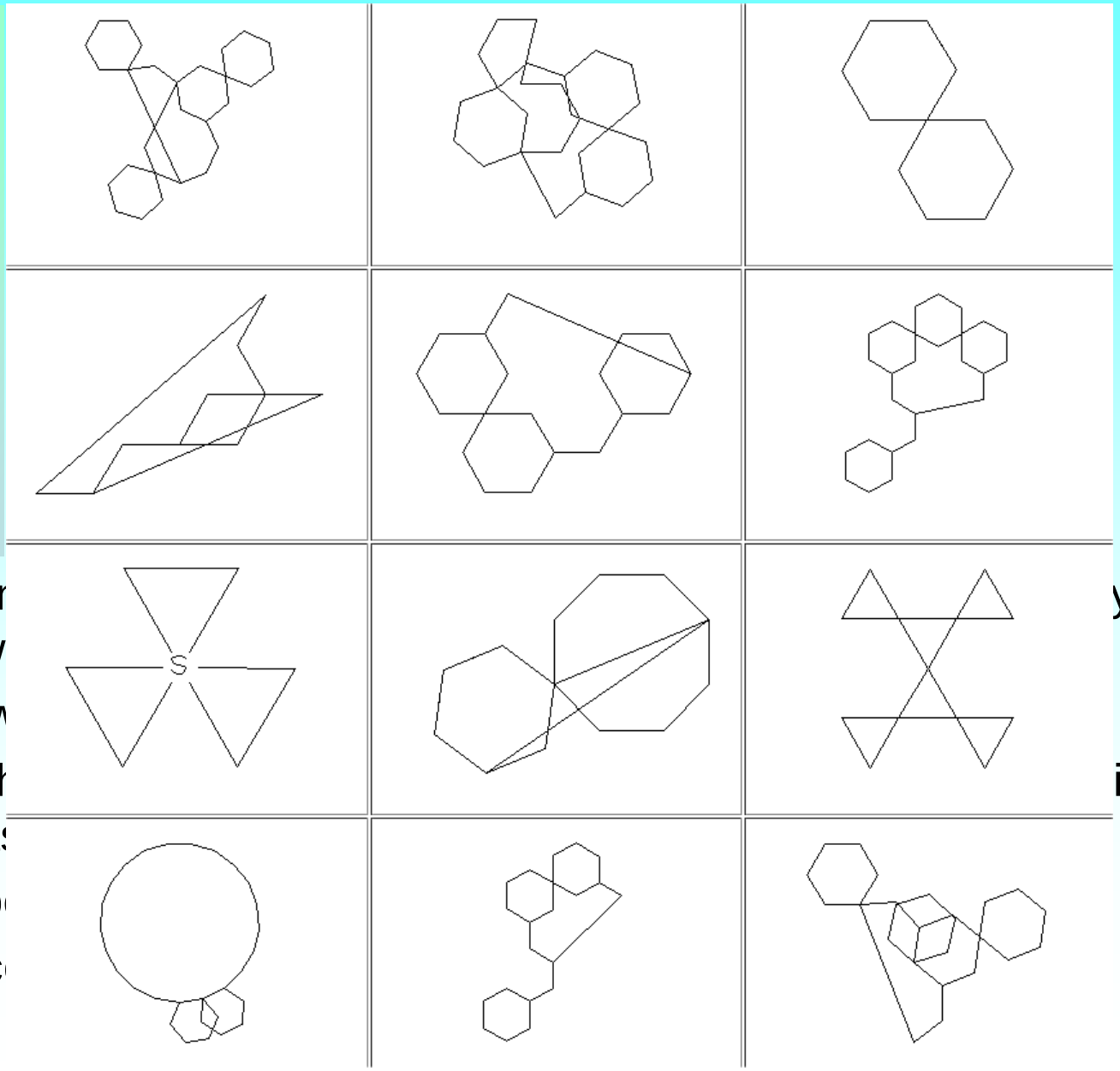
There are three ways to increase the number of rings in a graph by adding 3-nodes:



Any 4-nodes can be generated by fusing together its corresponding pair of connected 3-nodes

Small Problem: What is a Spiro Atom?





- Strict defi y exactly tw
- http://www
- But what h ing system, as
- Do the ab
- …Yes, ac

Databases Scanned for Unique Scaffold Topologies, up to 8 Rings

Database Source	Initial Count	Unique Topologies
ChemNavigator	14,041,970	3,880
DrugBank	2,742	**155**
Natural Products	132,434	**3,199**
PubChem	11,595,690	22,612
WOMBAT	149,451	**1333**
All Databases	25,029,900	**23,737**
GDB 2005	26,434,571	**76**

- **PubChem** covers the highest number of unique topology samples [its function as general repository reflects this]
- **DrugBank** has a *really small* number of unique topologies
- The gap between existing and missing topologies increases rapidly beyond 5 rings
- We had http://topology.health.unm.edu (not updated)

MJ Wester, SN Pollock, EA Coutsias, TK Allu, S Muresan, TI Oprea. *J. Chem. Info. Model.* 48: 1311 - 1324, 2008

Take Home Message 6

- Chemical space is only limited by our ability to map it
- At the low-end molecular weight region, it is clearly finite
- It diverges quickly into massive diversity; up to 8 rings we have performed an exhaustive map
- Beyond that, uncertainty…

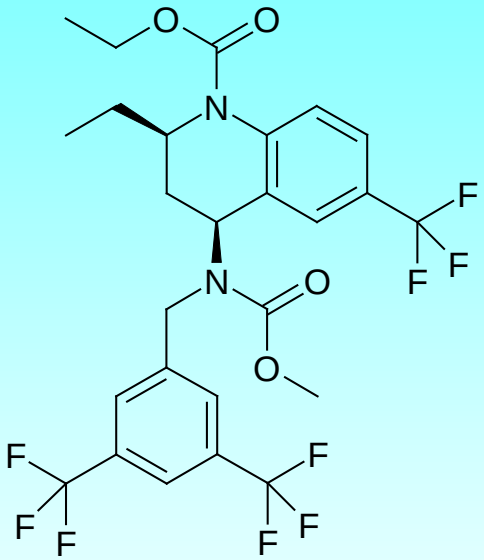
Black Swans and Blue Pills

Tudor I. Oprea
UNM School of Medicine
Andrew L. Hopkins
University of Dundee College of Life Sciences
Manuscript used to be in preparation for Nature Rev. Drug Discov.

The University of New Mexico
SCHOOL OF MEDICINE

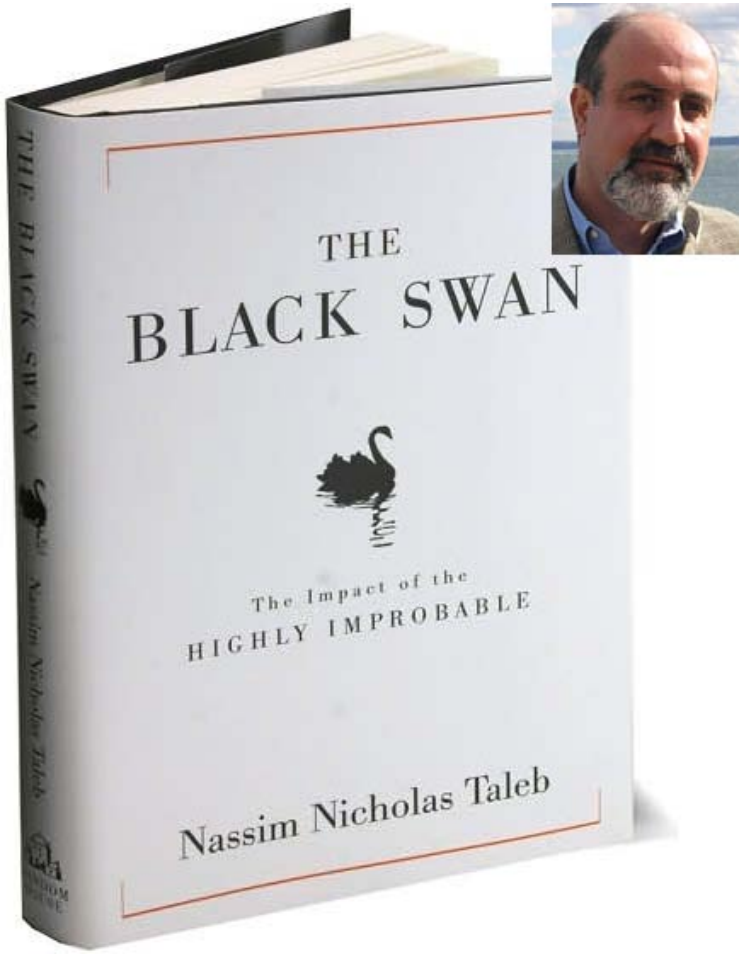
The Impact of the Unpredictable

- On November 30 2006, Jeff Kindler, Pfizer's CEO, was quoted as saying about **Torcetrapib** that "…this will be one of the most important compounds of our generation."
- On December 2, 2006, Pfizer cut off **Torcetrapib**'s ILLUMINATE trial because of "an imbalance of mortality and cardiovascular events" associated with its use (82 vs 51 deaths in a 15,000 patients clinical trial, comparing Lipitor/Torcetrapib vs Lipitor alone)
- This was the most late-stage important compound in Pfizer's portfolio. The event not only threw the financial projections of the world's largest pharmaceutical plans into flux (e.g., closing Ann Arbor), but it even raised questions about utility of raising HDL, the main hypothesis governing anti-arthrosclerosis therapy for the past decade.
- This event was a **Black Swan**: an unpredictable event, which had a massive impact.



Cholesterylester Transfer Protein inhibitor... Is this still a valid drug target?...

Source of Inspiration

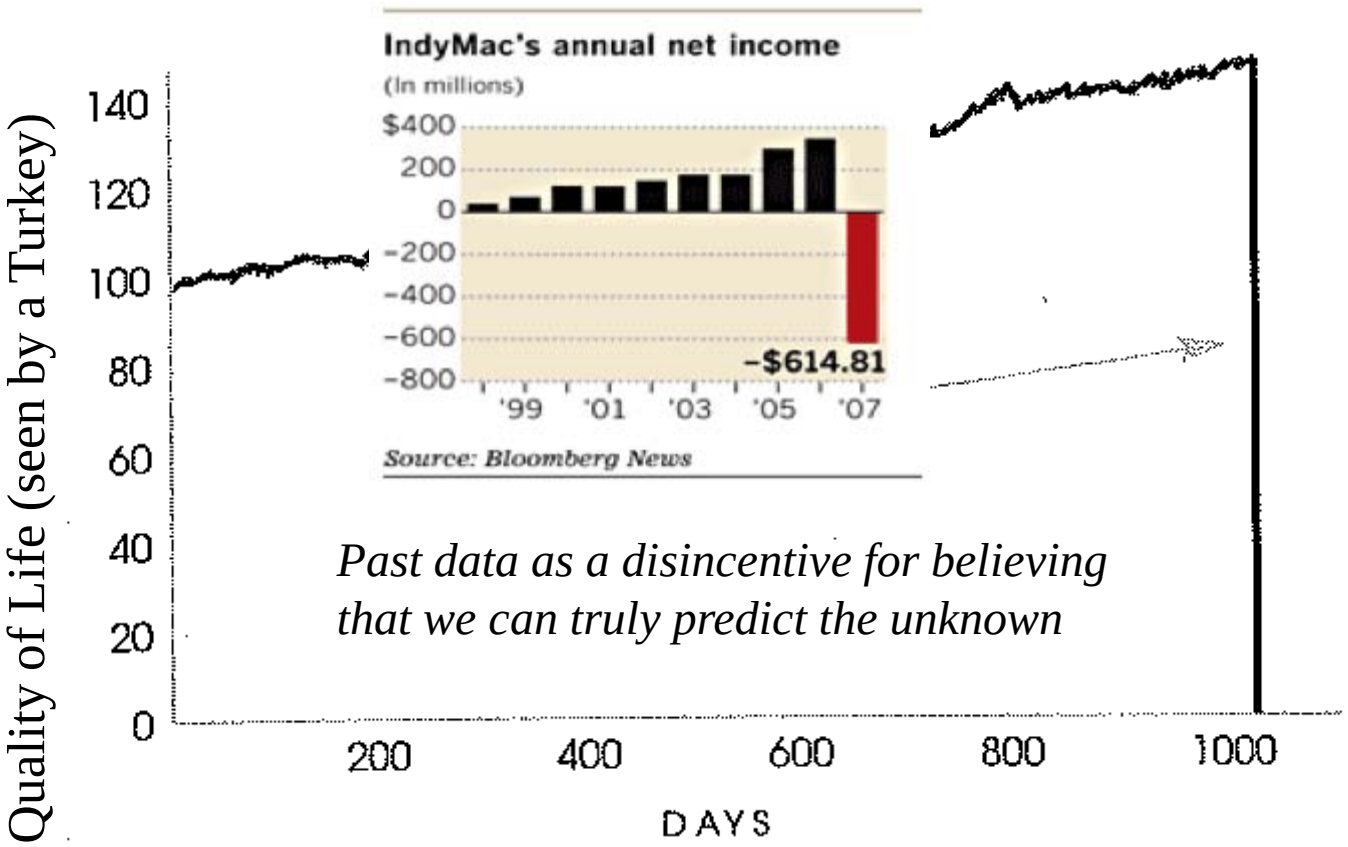


- NNT is a financial trader from Amioun, Lebanon, who learned that life, war and science are unpredictable through years of practice
- The Black Swan is a metaphor for the first sighting of the black swan, which (a) invalidated the assumption that all swans are white – based on millions of previous observations; (b) changed our perception of those birds and (c) was retrospectively "assimilated" as a highly predictable event.

See also:

http://www.fooledbyrandomness.com/

One Thousand and One Days Of History



- The Turkey Surprise (NNT), also referred to as [David] "Hume's Problem" was contemplated by Sextus Empiricus, Al-Ghazali, P.D. Huet, Francis Bacon, Bertrand Russell...
- Sir Francis extended it to *silent evidence* (we can't learn from the hidden; we don't learn that we never learn)

Fig. 1: *A turkey before and after Thanksgiving. The history of a process over a thousand days tells you nothing about what is to happen next. This naïve projection of the future from the past can be applied to anything*

Adapted from "*The Black Swan: the impact of the highly improbable*", by Nassim Nicholas Taleb

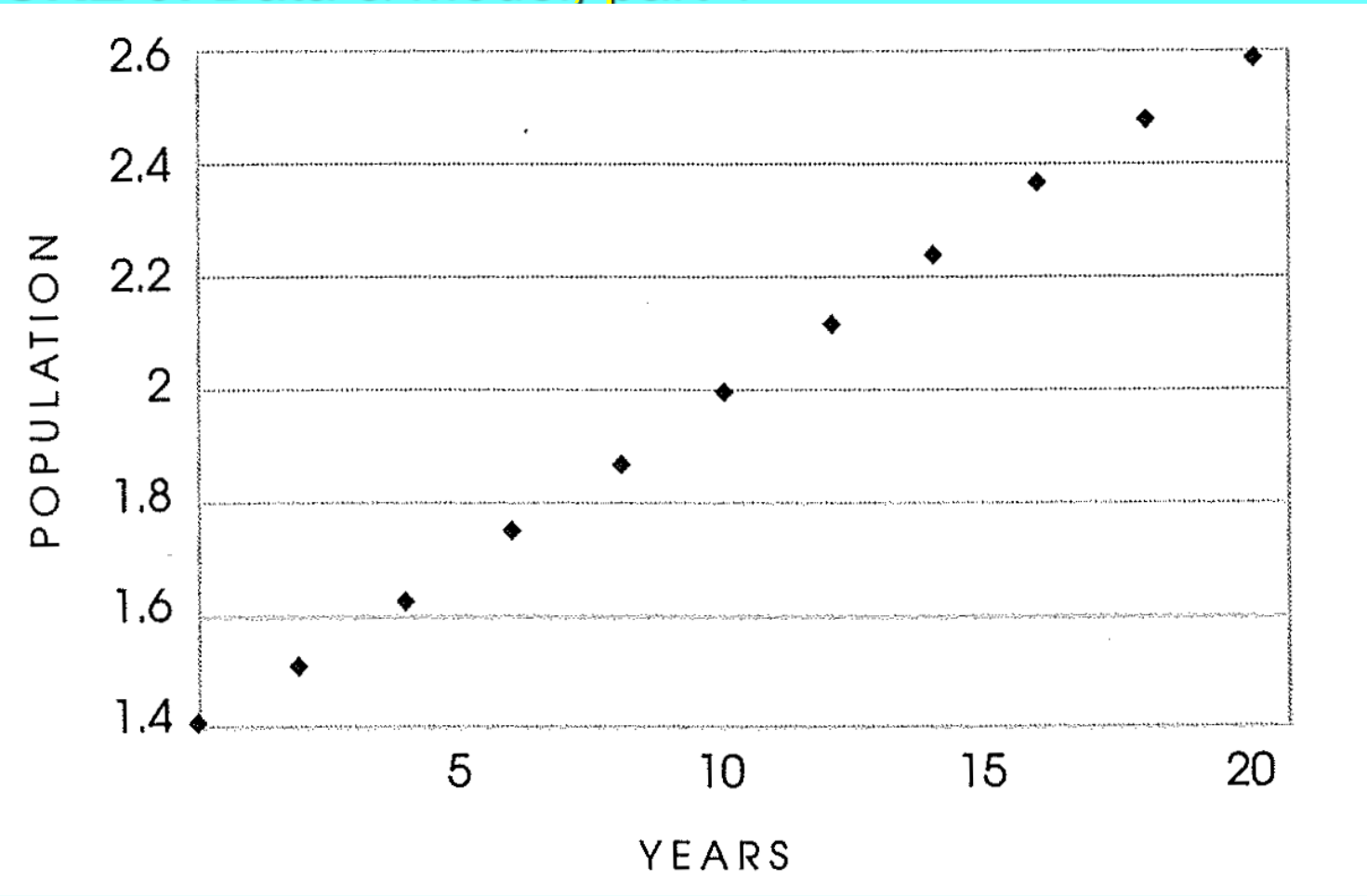
Random House, Inc., New York: 2007, pg. 41

The Tale of the Unknown Unknowns

Ridiculed by the media, Sec. Donald Rumsfeld (once CEO of a pharma company, Searle) was, in fact, making a very serious point about the impossible-to-predict situations.

- "There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know."
- *Known Knowns*
  - Sets & models that we have, which were validated externally.
- *Known Unknowns*
  - Sets & models that we think we can prove, but lack external validation.
- *Unknown Unknowns*
  - Sets & models that we do not have, which lack any validation. These include that part of model space that we don't even know exists

FIGURE 3: Data & Model, part 1

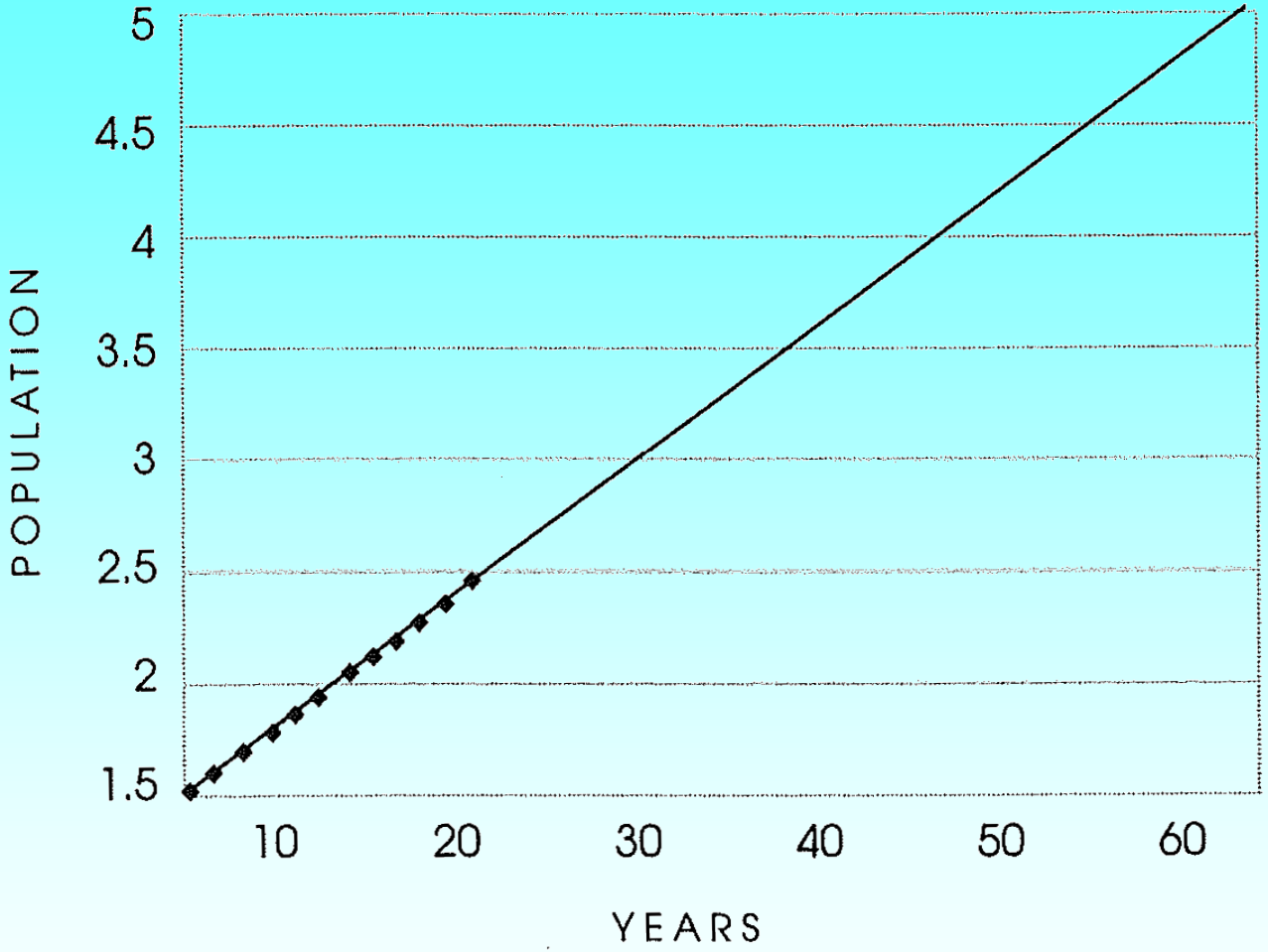


A series of a seemingly growing bacterial population (or of sales records, or of any variable observed through time – such as the total feeding of the turkey in Chapter 4

Taken from "*The Black Swan: the impact of the highly improbable*", by Nassim Nicholas Taleb

Random House, Inc., New York: 2007, pg. 186

FIGURE 4: Data & Model, part 2

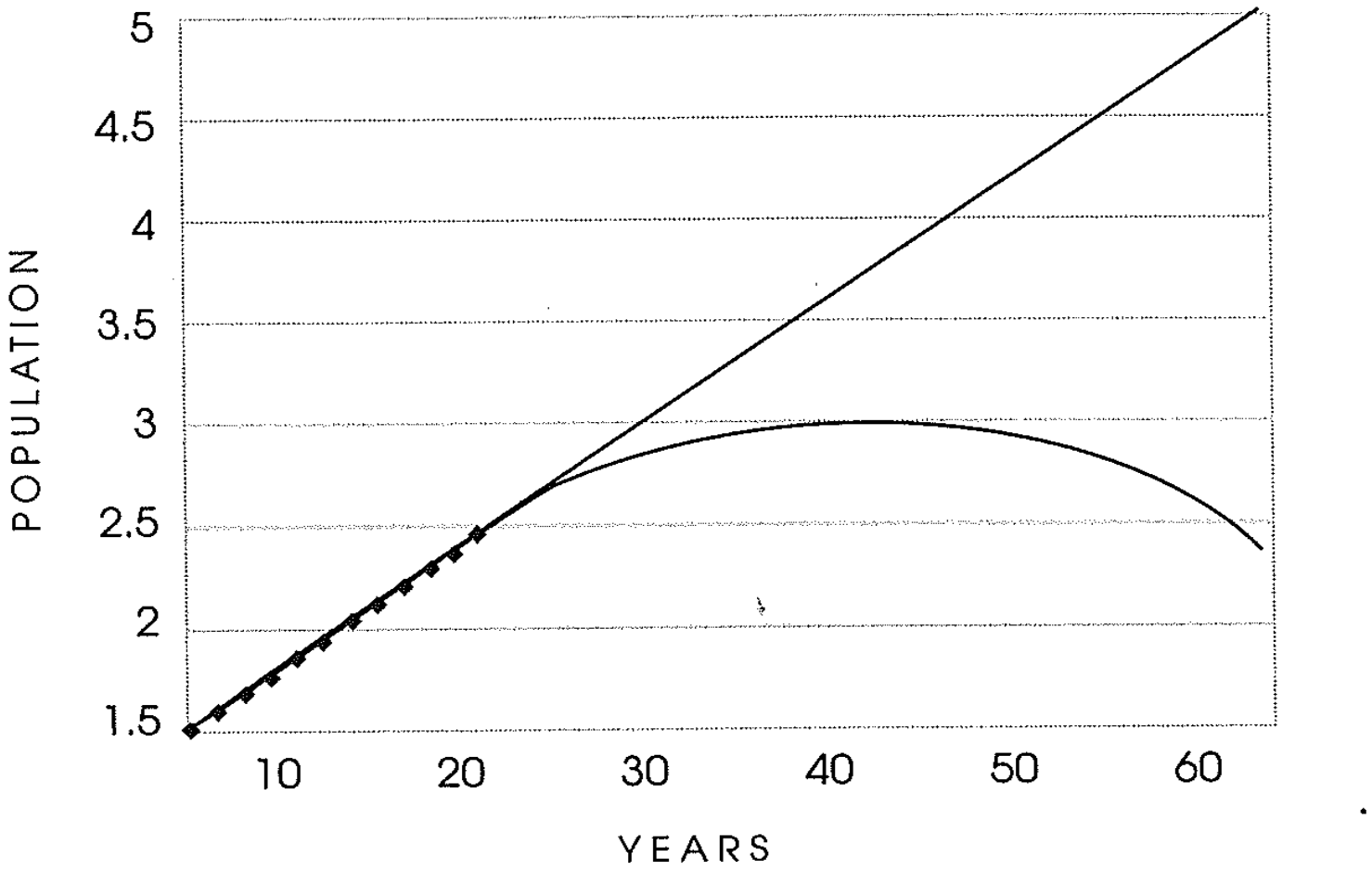


Easy to fit the trend – there is one and only one linear model that fits the data. You can project a continuation into the future

Taken from "*The Black Swan: the impact of the highly improbable*", by Nassim Nicholas Taleb

Random House, Inc., New York: 2007, pg. 186

FIGURE 5: Data & Model(s), part 3

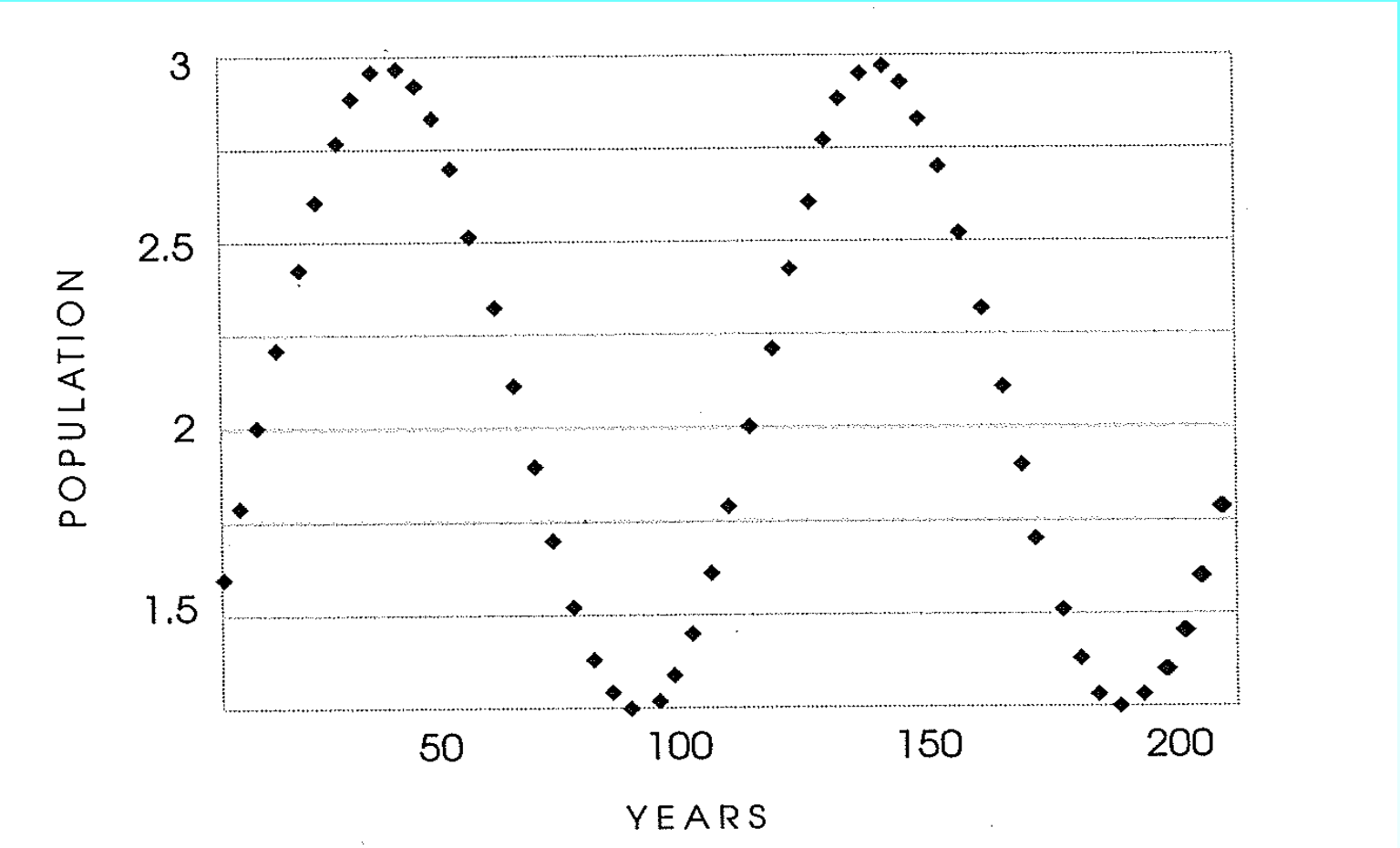


We look at a broader scale. Hey, other models also fit it rather well

Taken from "*The Black Swan: the impact of the highly improbable*", by Nassim Nicholas Taleb

Random House, Inc., New York: 2007, pg. 187

FIGURE 6: Data & Models, part 4

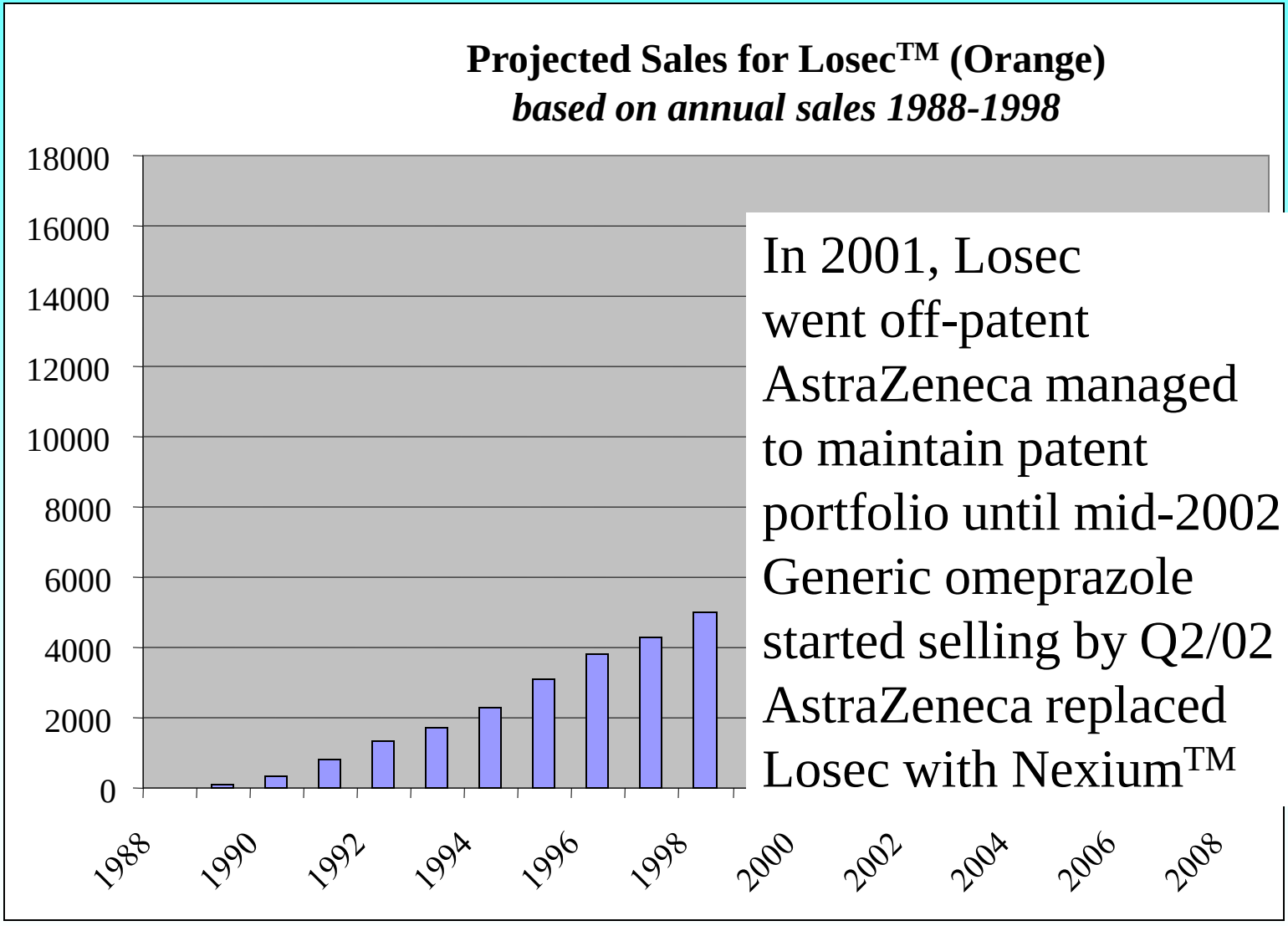


And the real "generating process" is extremely simple but it had nothing to do with a linear model! Some parts of it appear to be linear and we are fooled by extrapolating in a direct line.*

Taken from "*The Black Swan: the impact of the highly improbable*", by Nassim Nicholas Taleb

Random House, Inc., New York: 2007, pg. 187

A Past-time Prediction (is it True?)



Never Tell Me the Odds!

Han Solo, aboard the Millenium Falcon

- We focus on preselected segments of the seen and generalize from it to the unseen: Error of confirmation (absence of evidence is not evidence of absence)
- We fool ourselves with stories that cater to our "idealized" thirst for patterns: Narrative fallacy (people have selective memory & tend to generalize from single observations)
- We behave as if the Black Swan does not exist (people go on living as if death is the unlikeliest of events. *Mors certa, ora incerta*)
- What we see is not all there is… The odds are not always evaluated properly: Silent evidence distortion (e.g., people play the Lottery focusing on winners, & forget the odds)
- We "tunnel" on specific Black Swans (e.g., floods, fires) and rarely prepare for the unknown unknowns

Adapted from "*The Black Swan: the impact of the highly improbable*", by Nassim Nicholas Taleb

Random House, Inc., New York: 2007, pg. 50

We think we live in Mediocristan… but live in Extremistan

- Non-Scalable
- One datapoint does not influence the average
- Tyrrany of the collective
- Winner-get-a-slice
- Ancestral environment
- Applies to humans in physical context (weight)
- Easy to predict from what you see to what you don't
- 80/20 principle not obvious
- "Bell curve" (Gaussian)
- Impervious to Black Swans

- Scalable
- One datapoint can have high impact and great influence
- Tyrrany of the accidental
- Winner-takes-the-pizza
- Modern environment
- Applies to humans in social context (money)
- Past information rarely assists in making predictions
- 80/20 principle rules
- Power law / Mandelbrotian
- Vulnerable to Black Swans

Adapted from "*The Black Swan: the impact of the highly improbable*", by Nassim Nicholas Taleb

Random House, Inc., New York: 2007, pg. 36

We think we live in Mediocristan… but live in Extremistan

- Solar system during our lifetime (except for comets)
- Peoples' weight & height
- Peoples' 1:1 conversations
- The world before Guttenberg
- Your untold stories
- Your hamburger & coffee
- The world before Bell
- Your website
- Your "usual" scientific journals
- Your emails
- Your local music band
- Your hotdog

- Solar system in the long run (comets, supernovas, aliens, etc.)
- Peoples' money & social networks
- TV & radio show-hosts
- The world of printed books
- Tom Clancy, JK Rowling
- McDonald's & Starbucks
- The world after the Internet
- Google
- Science, Nature – high-impact
- The blogosphere
- Pink Floyd
- The one with everything

Adapted from "*Black Swans & White Tablets*", by T.I.O. & A.L.H.

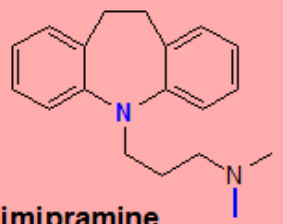
(the point is: once you get it, you got it)

The Fitness Landscape

Similar molecules act in a similar manner

…or do they?! We're beginning to realize that similar molecules may have very different activities, leading to what Gerry Maggiora calls activity cliffs.

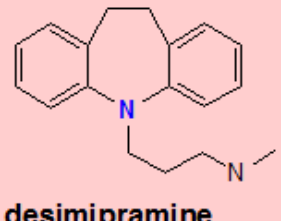
Bioactivity Cliffs



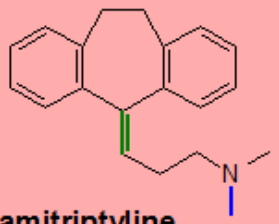
imipramine
MW = 280.4
pK_a = 9.5
$LogD_{74}$ = 2.2
$LogS_w$ = -4.19

8.85	(5HTT)
7.43	(NAT)
5.07	(DAT)
6.82	($5\text{-}HT_{2A}$)
6.80	($5\text{-}HT_{2C}$)
7.49	(α_{1A})
7.61	(H1)
7.38	(M1)
8.47	(M2)
5.9	(hERG)
<4.3	(AOX1)

7.75	(5HTT)
9.08	(NAT)
5.50	(DAT)
6.78	($5\text{-}HT_{2A}$)
6.50	($5\text{-}HT_{2C}$)
7.00	(α_{1A})
7.12	(H1)
6.76	(M1)
6.27	(M2)
5.86	(hERG)
<4.3	(AOX1)



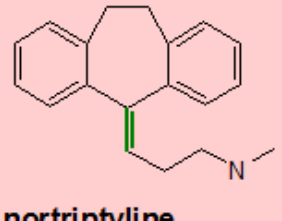
desimipramine
MW = 266.4
pK_a = 10.2
$LogD_{74}$ = 1.4
$LogS_w$ = -3.9



amitriptyline
MW = 277.4
pK_a = 9.4
$LogD_{74}$ = 2.95
$LogS_w$ = -4.46

8.37	(5HTT)
7.46	(NAT)
5.49	(DAT)
7.96	($5\text{-}HT_{2A}$)
8.09	($5\text{-}HT_{2C}$)
7.90	(α_{1A})
9.30	(H1)
7.83	(M1)
8.10	(M2)
5.0	(hERG)
6.59	(AOX1)

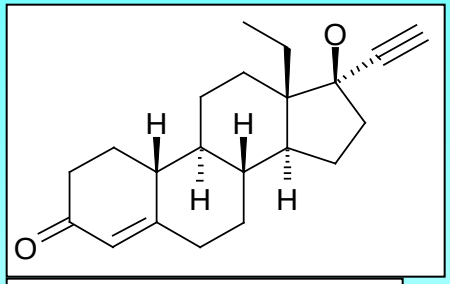
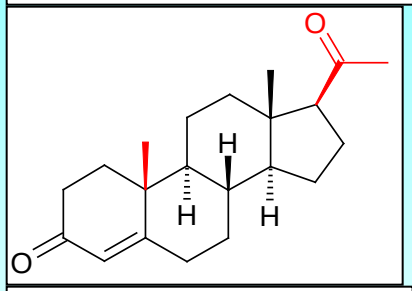
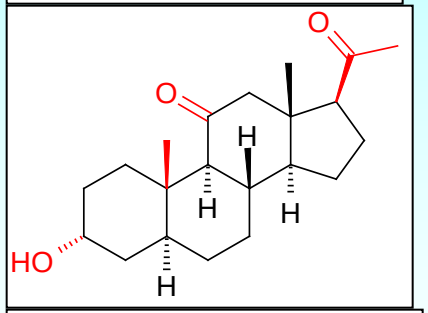
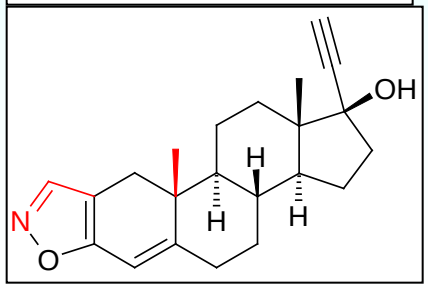
7.75	(5HTT)
8.36	(NAT)
5.94	(DAT)
7.36	($5\text{-}HT_{2A}$)
8.05	($5\text{-}HT_{2C}$)
7.26	(α_{1A})
8.03	(H1)
7.40	(M1)
6.96	(M2)
N/A	(hERG)
6.07	(AOX1)



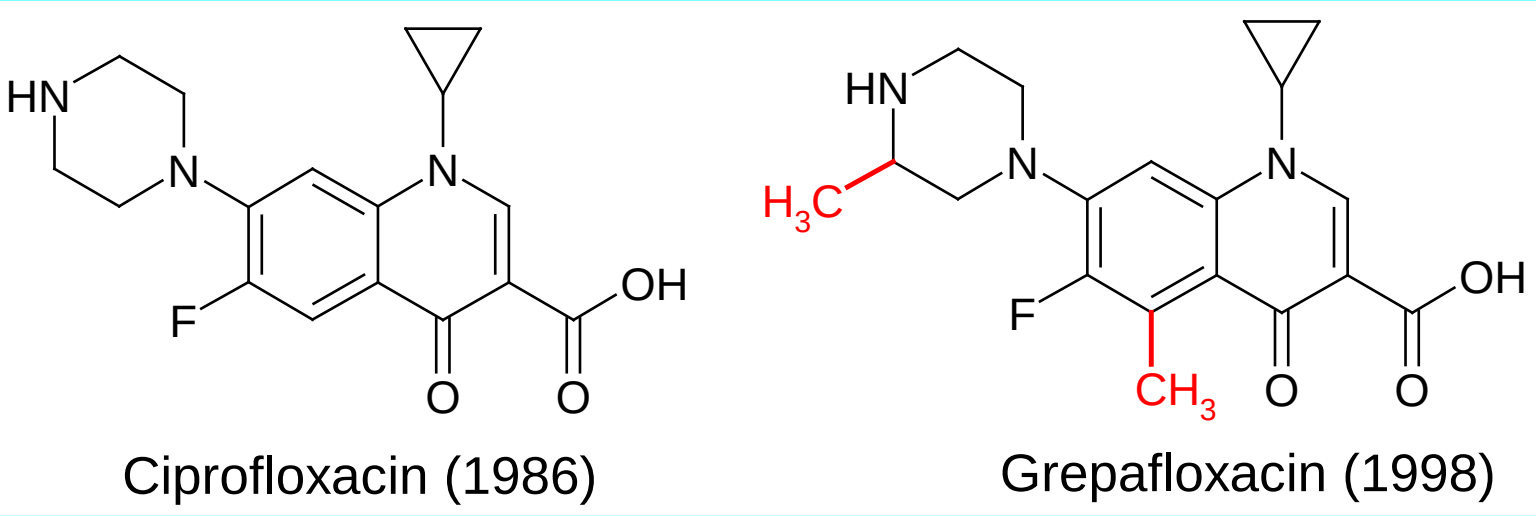
nortriptyline
MW = 263.4
pK_a = 9.73
$LogD_{74}$ = 1.69
$LogS_w$ = -4.15

Target and antitarget affinity profile for four antidepressants.
Differences in affinity to the intended targets, the monoamine transporters 5HTT and NAT, and the serotonin 5-HT2 receptors, are likely to result in distinct clinical efficacy. Differences in antitarget activity on α_{1A}, H1, M1-M2, hERG and AOX1 impact the safety profile of these drugs

Target Affinity Cliffs

Chemical Structure	MolName	Target 1	Target 2	Target 3	Target 4	MW	AlogS	AlogP	ClogP
	Norgestrel	Progesterone receptor	Estrogen receptor			312.46	-4.74	3.25	3.5
	Progesterone	Progesterone receptor	Estrogen receptor	Membrane progestin receptor alpha	Mineralo-corticoid receptor	314.47	-4.77	3.58	3.96
	Alphaxalone	Chloride channel protein, skeletal muscle, ClC-1	GABA-A receptor alpha-1 subunit	GABA-A receptor alpha-2 subunit	GABA-A receptor alpha-5 subunit	332.49	-4.15	3.28	3.73
	Danazol	Progesterone receptor	Estrogen sulfatase	Androgen receptor		337.47	-4.27	3.63	3.93

Antitarget Affinity Cliffs

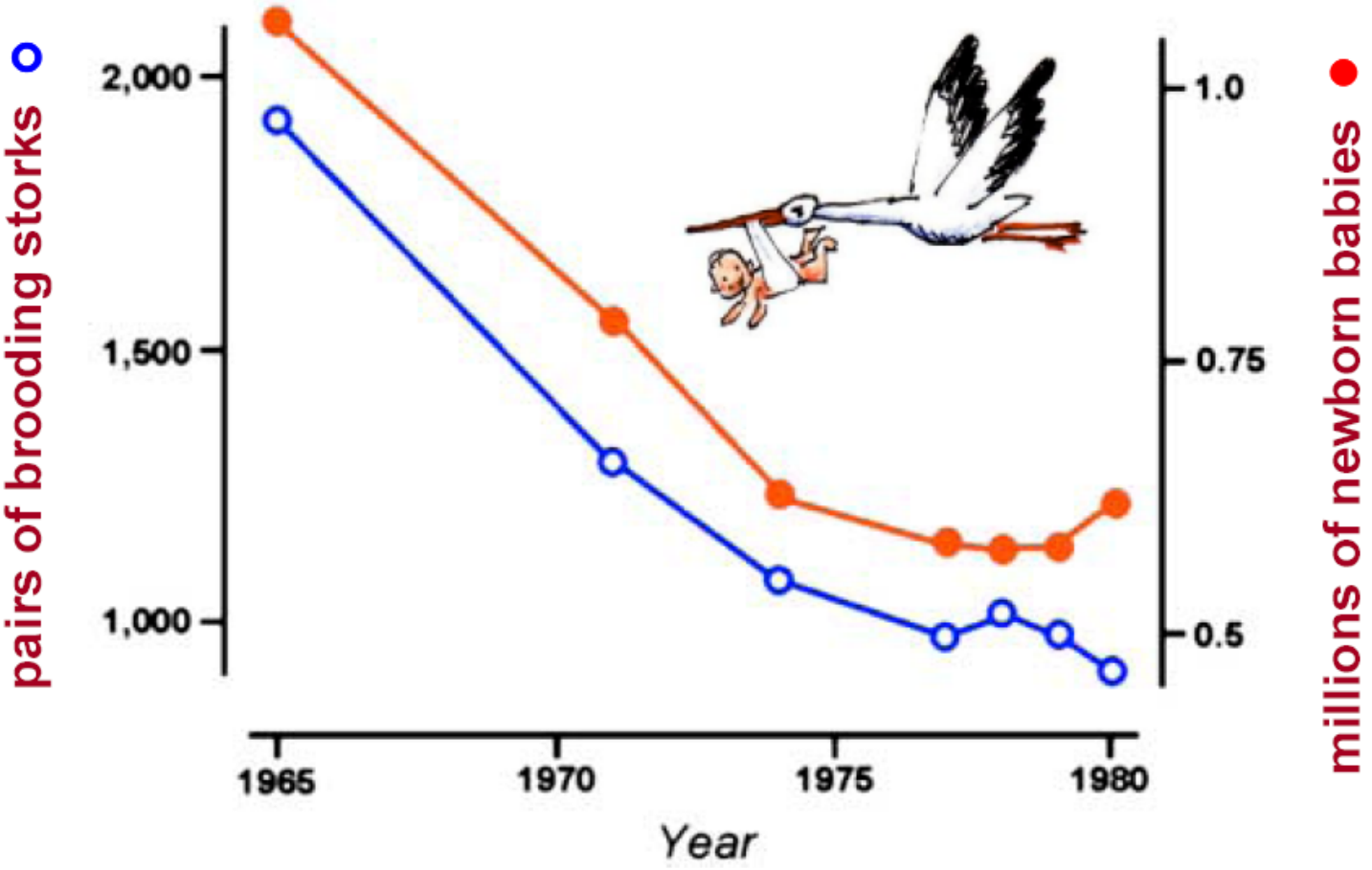


- Grepafloxacin, launched as Vaxar in Germany and Denmark (1998) by Otsuka (Japan) was withdrawn in 1999, following reports of severe cardiovascular events (hERG binder, causes QT prolongation which may lead to fatal ventricular arrhythmias)
- Ciprofloxacin, a slow-seller, turned into a positive Black Swan for Bayer during the Anthrax scare

Failed Predictions, Rationalized Post-Hoc

- *Post hoc ergo propter hoc* is a logical fallacy that permeates in science: the appearance of correlation is often thought to relate to causality (just because the R-square is high does not mean that variables X & Y share a causality relationship)
- **We blame outliers** (chemical; biological; statistical), **we blame the descriptor system** (or the length of the simulation time), sometimes **we blame the experiments**, we forgot to take water or hydrophobics into account, even worse *we delete the data when it does not fit the model!*
- **We rarely query the outliers.** That wealth of data, that unexpected, which we're unfortunately trained to ignore has given the world many "serendipitously" discovered medicines such as Penicillin and Viagra.

The Storks and the Babies



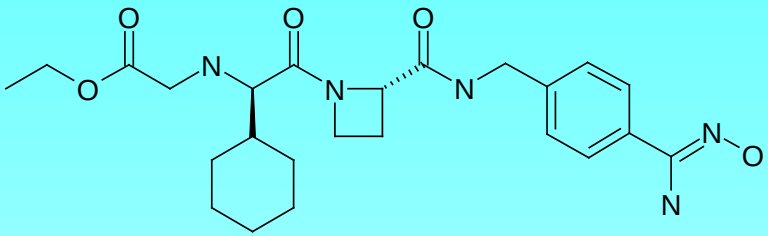
Sir – There is concern in West Germany over the falling birth rate. The accompanying graph might suggest a solution that every child knows makes sense.

H. Sies, Nature 332, 495 (1988)

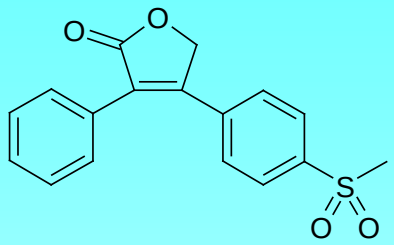
The Impact of the Highly Improbable

- Recall the financial outlook that Pfizer's CEO Jeffrey Kindler presented at an Analyst's Meet 2 days before clinical data forced them to stop the **Torcetrapib** ILLUMINATE Trial (2006).
- "Pfizer, Pfizer, Pfizer. Depending on your point of view, it's ironic, inspiring, or merely interesting that the company that staggered out of 2006 with its every vulnerability and vanity exposed in the media glare nonetheless finishes in *Pharm Exec's* winner's circle for the eighth year running. " *Pfizer ranked #1, with $45.08 Billion USD sales in 2006*
- Imagine you worked at Merck on September 26 2004, benefiting from the $2.5 billion/year sales of **Vioxx**. By September 30, Merck announced a voluntary worldwide withdrawal of Rofecoxib (Vioxx) following results from the 3-year APPROVe trial, which showed an increased risk of cardiovascular events such as heart attack and stroke beginning after 18 months of treatment . Can you predict chronic (ab)use in clinical trials?! *Merck ranked #7, with $22.64 Billion USD sales in 2006*
- Bayer voluntarily withdrew **Baycol** (cerivastatin) on August 7 2001, following reports of side-effects of potentially fatal myopathy and rhabdomyolysis, particularly in patients co-treated with Gemfibrozil. *Bayer ranked #15, with $9.87 Billion USD sales in 2006*

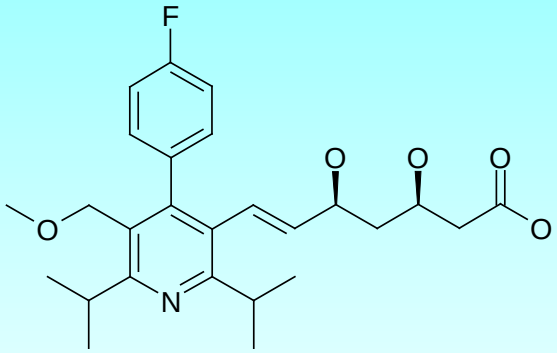
What do these Drugs have in Common?



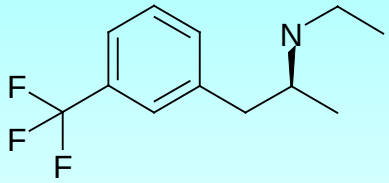
Ximelagatran (Exanta) - AstraZeneca



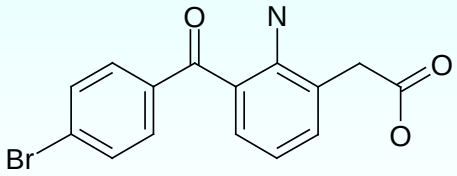
Rofecoxib (Vioxx) – Merck



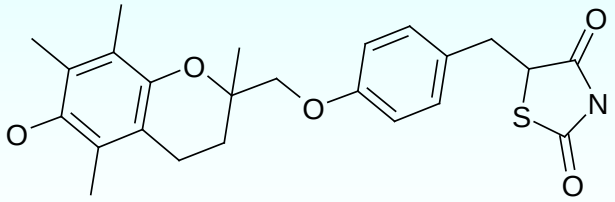
Cerivastatin (Baycol) - Bayer



Dexfenfluramine (Redux) - Servier



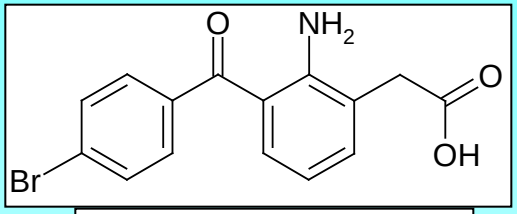
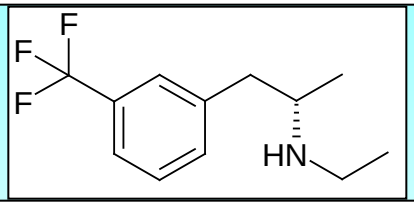
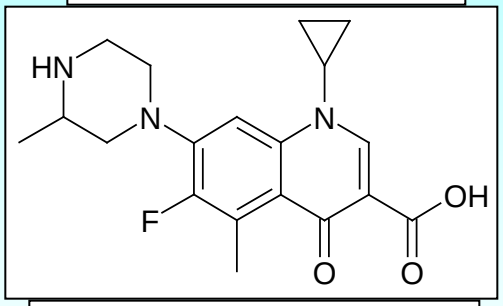
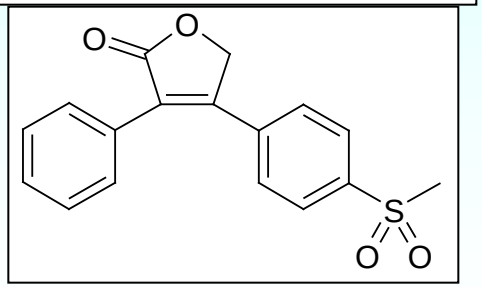
Bromfenac (Duract) - Wyeth



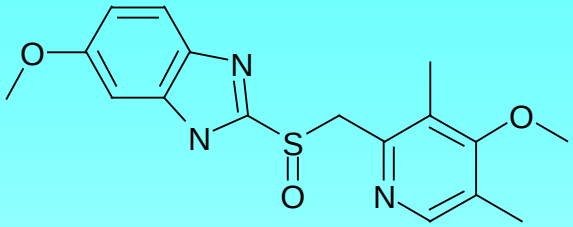
Troglitazone (Rezulin) - Sankyo

All of them were withdrawn globally in the past decade

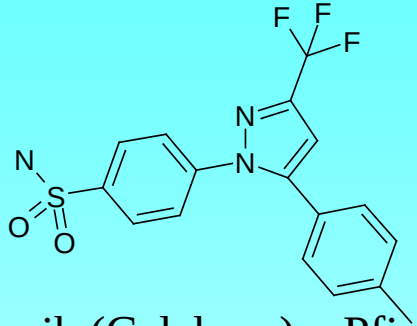
Market Activity Cliffs

DRUG Name (Country first launched) Generic Name	Com-pany	Chemical Structure (active compound)	Period Launched	Reason for withdrawal	Therapetic category	Total Sales 1996-2006
DURACT (USA) Bromfenac	Wyeth		1997-1998	following several cases of severe liver failure in patients taking bromfenac for >10 days	anti-inflammatory; analgesic	169.76 million USD
ADIFAX (France); REDUX (USA) Dexfenfluramine	Servier		1970-1997	following a US FDA recommendation due to abnormalities in heart valves	anorexic, for the treatment of obesity	346.27 million USD
VAXAR (Germany and Denmark) Grepafloxacin	Otsuka		1998 - 1999	following reports of severe cardiovascular events	antibacterial, quinolone	47.236 million USD
VIOXX (several countries) Rofecoxib	Merck & Co		1999-2004	due to increased risk of cardiovascular events such as heart attack and stroke beginning after 18 months of treatment	anti-inflammatory; analgesic; COX-2 inhibitor	13,319.98 million USD

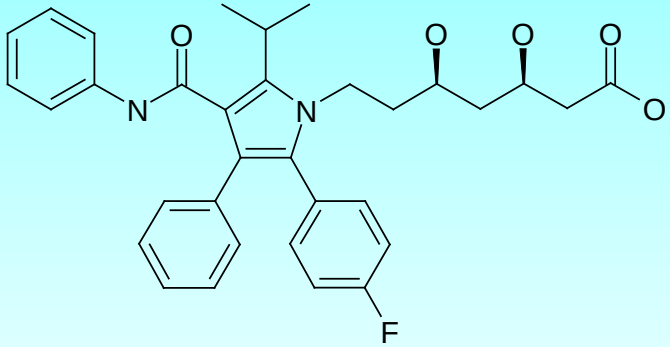
What do these Drugs have in Common?



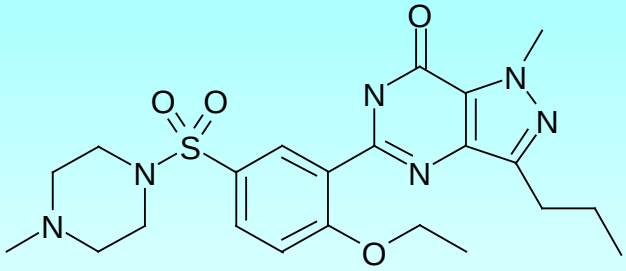
Omeprazole (Losec) - AstraZeneca



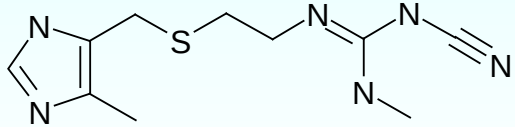
Celecoxib (Celebrex) – Pfizer



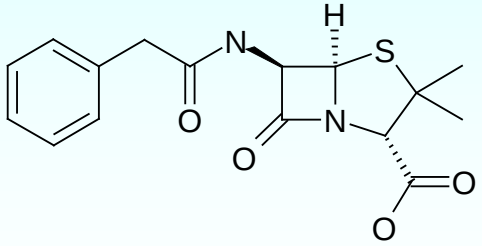
Atorvastatin (Lipitor) - Pfizer



Sildenafil (Viagra) – Pfizer



Cimetidine (Tagamet) – GSK



Penicillin G (generic)

All of them were positive Black Swans

Predictions And Small Molecule Discovery

- One cannot travel forward in time (speeds = 1 sec/sec)
- While informatics (machine learning) specialists assume that the chemical space related to drug discovery belongs to *Mediocristan*, in fact it belongs to *Extremistan* **because** of the highly hierarchical structure of that space (winner-takes-it-all)
- Hint: **The definition of a drug** or **GRAS** (both unknowable quantities) **are done by a regulatory body** based on available evidence submitted by a company

  Drug and Flavor research as businesses are highly impacted by the **social aspect:** The industry (people) decide to petition the regulatory agencies for an NDA/GRAS, and the agency (people) VOTES on that small molecule's safety based on filed data

- **“Drug” and “GRAS” are not a natural property of chemicals**

- GRAS: “generally regarded as safe”

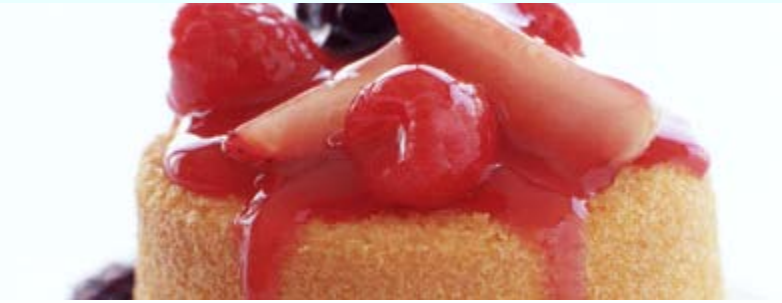
Givaudan – UNM collaboration

Givaudan

The University of New Mexico
Division of BIOCOMPUTING

- UNM maintains FEMA/GRAS & flavour databases of compounds:
  - » Cooling taste
  - » Bitter taste (inhibitors)
  - » Umami taste (enhancers)
  - » Sweetness taste (enhancers)

- UNM develops collaborative and decision making tools

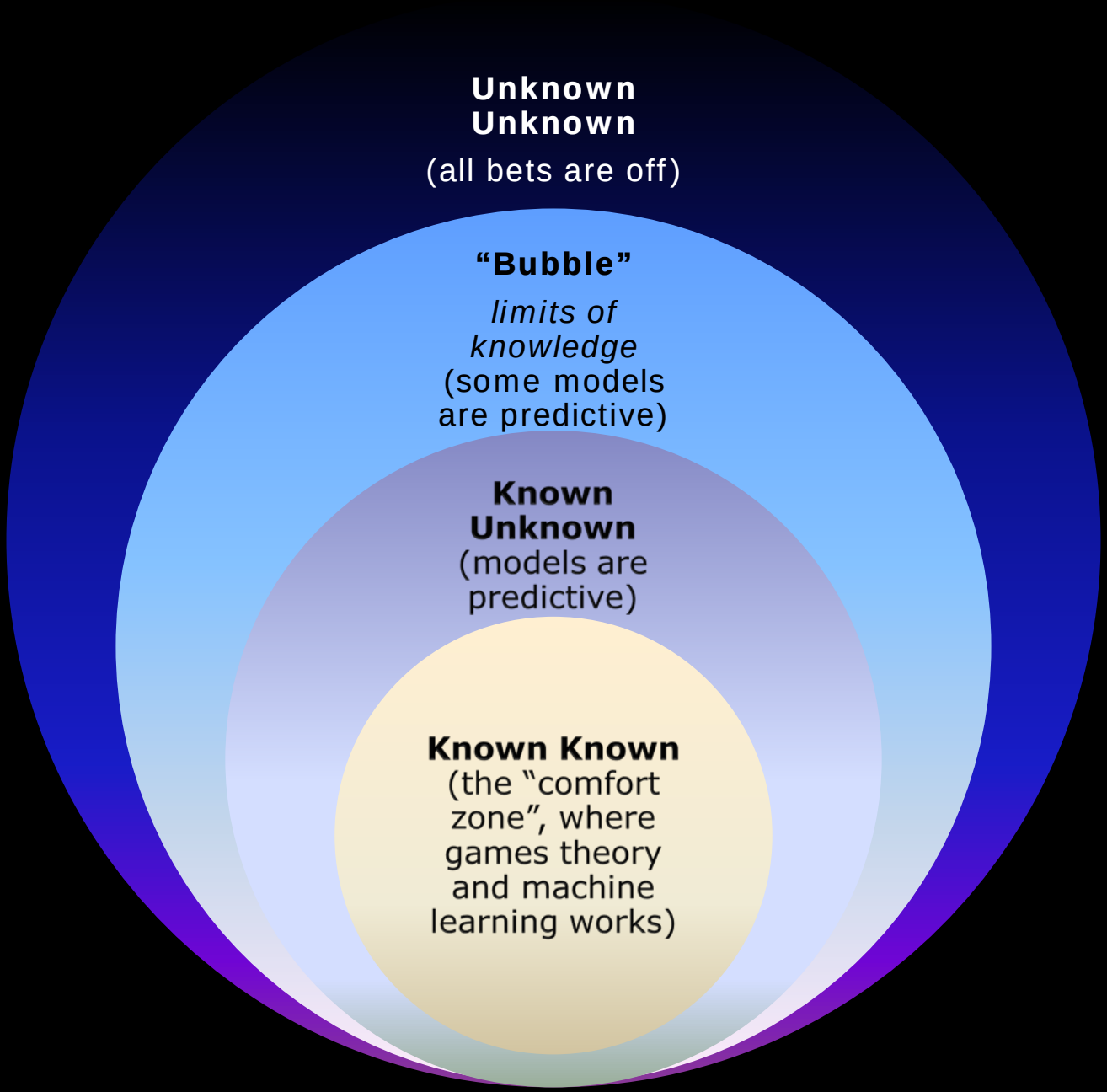


> 50% of the hits (since 2003) were discovered with help from UNM

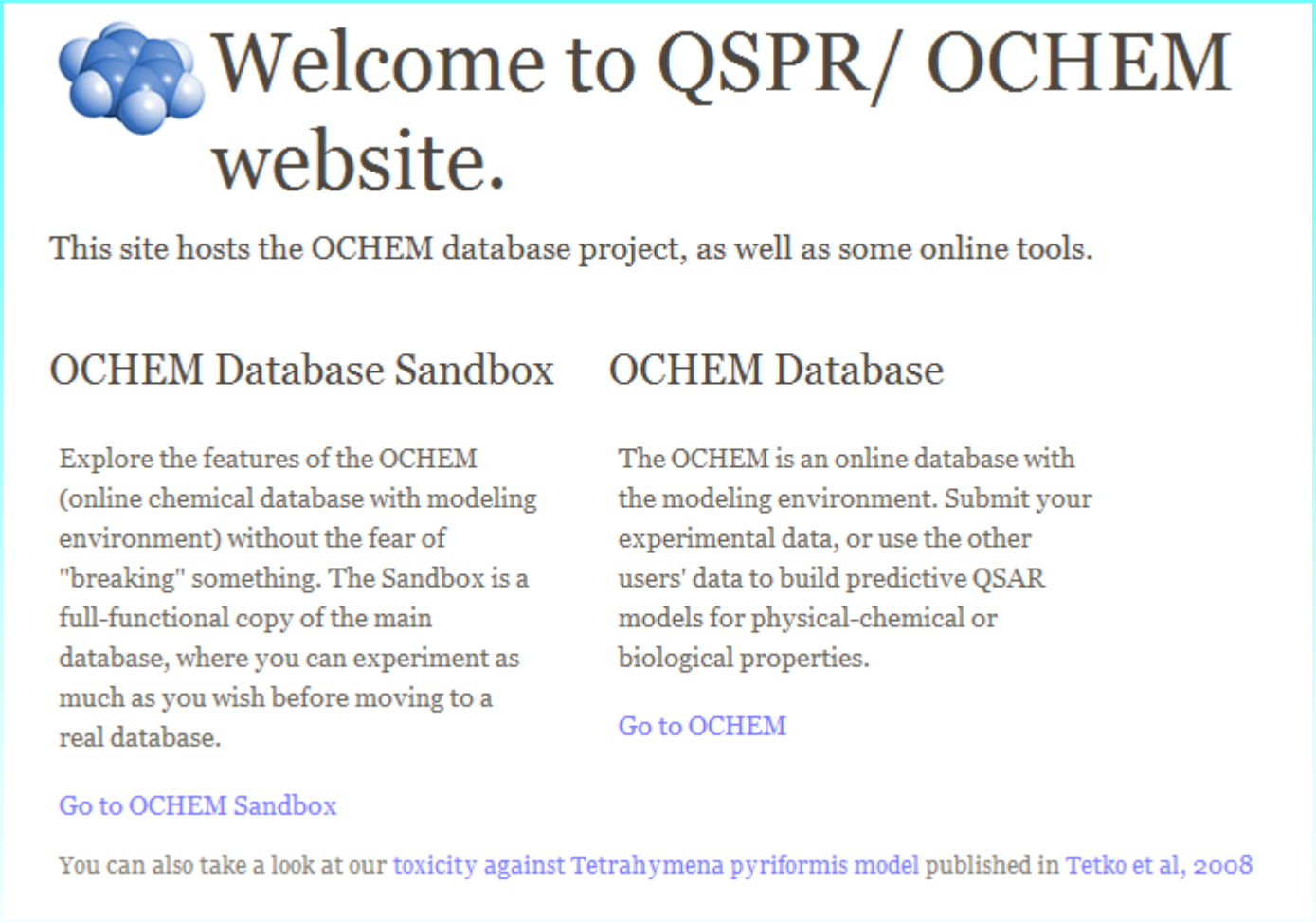
On Machine Learning

- Predictions Are Based on the Past
- What follows is a highly personal opinion:
  - The known chemical and biological space can be mapped.
  - This implies that, within limits and given appropriate descriptors, we can generate (wrong) models that are useful in small increments
  - These models do not allow us to understand what Gerry Maggiora calls “affinity cliffs” – the “turkey surprise” in medicinal chemistry space
  - There needs to be a balanced choice between good coverage of chemical space (“diversity”) and biological space (“targets”) in order to build reliable models
  - As reliable as these models may be, *within the constraints of the already-mapped (i.e., the past) chemical space,*
  - … no machine learning method can predict the previously-unmapped chemical and biological space [the unknown unknown]
- *We keep pretending that Machine Learning / QSAR works*
- **We are biased because most scientific papers we read are success stories!** We are left to our own devices when it comes to failures – we have to learn the hard way (experience!)

Limits of Knowledge (Why QSAR Fails)



Final Home Tools

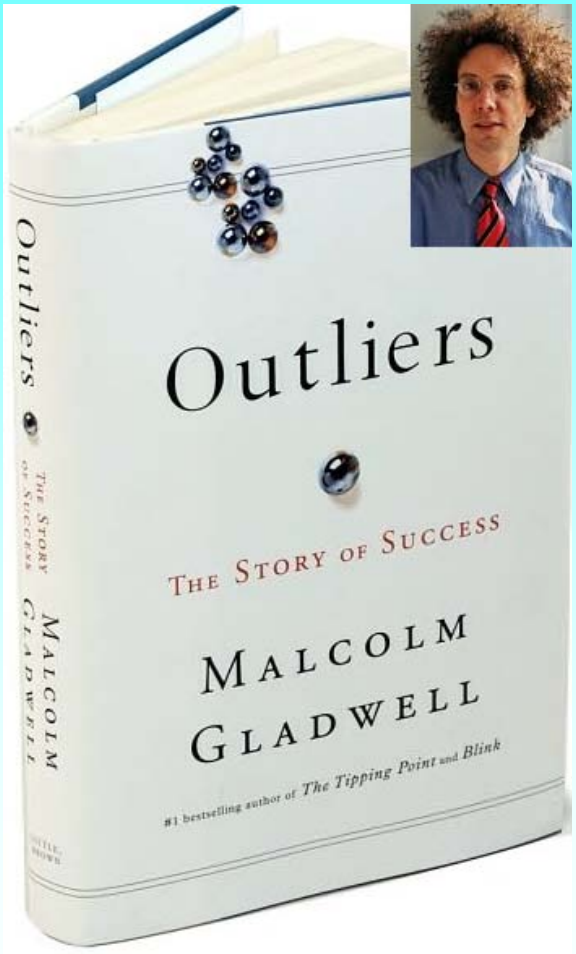


This online chemistry & modeling resource, OCHEM, is offered free to the on-line community. It builds on the VCCLAB.ORG experience (good source for LogP and LogS models, as well as QSAR engines!!!)

Reality Check

- Pharma & Flavor Industries are not bailed out by governments
  - like the airline industry, where governments jump to bail out bankrupting companies despite lack of foresight from management [e.g., nobody anticipated that oil prices would go up, that terrorists would scare travelers, etc.]…
  - or the banking industry [governments continue to rush to save failing banks e.g., when South American countries defaulted on their loans in 1998; when the subprime loans system collapsed in 2008].
  - nor is it even similar to the software industry [e.g., if a computer is rendered useless because of Dell or Microsoft, your chances to get anything than a new computer are slim.
  - … not so if you happened to take Vioxx & died of heart attack - your family will seek damages from the drug manufacturer.
- It requires lobbying by professional associations (PhRMA)
- Carefully perform damage control should crises arise (i.e., have crises management plan/team, act swiftly)
- We need to move away from the schizophrenic attitude of curing the world (R&D) and making quick profits (marketing)

Source of Inspiration (2)



- English-born Jamaican Malcolm Gladwell studies outliers in the context of success as it appears in finances, sports, business, law, agriculture
- He posits that there is no such thing as an isolated success story
- Outliers emerge in a certain societal and cultural context, and tend to emerge as “ready” and “experts” when new needs arise
- Most interesting read: How critical communication is when the first officer talks to the captain (and why some airplanes crashed)

- …could this happen in the pharma?

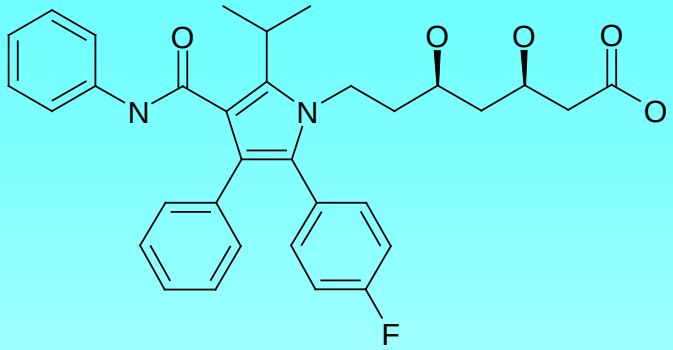
http://www.gladwell.com/outliers/index.html

The Ten Thousand Hours Rule

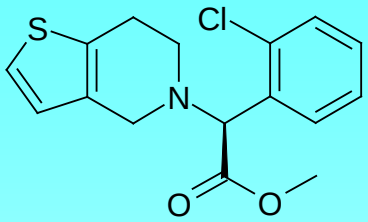
- If it's worth it, invest 10,000 hours of your time
  - Bill Joyce (UNIX, Sun Computers) and Bill Gates (Microsoft) had 10,000 hours of computer-programming *before* starting their companies…
  - The Beatles played 7 days a week, 7-8 hours a night (for a total of 270 nights, 1960-1962) in Hamburg, with an estimated 1200 public appearances before their first 1964 big hit.
  - From hockey players to soccer and chess, to scientists and successful businessmen, there seems to be a 10,000 hours rule
  - … even Mozart's compositions matured at age 21 (started playing at 5).
- Suggestion: If you want to understand what makes a good drug successful, start by staring at them. Each day, every day, for … 10,000 hours

See also: http://www.gladwell.com/outliers/outliers_excerpt1.html

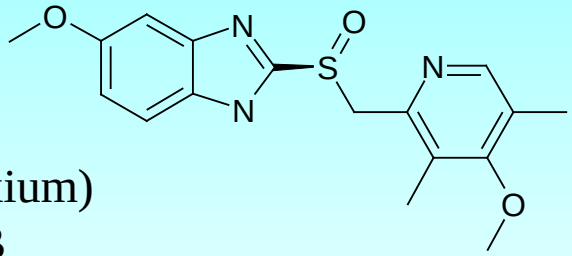
What do these Drugs have in Common?



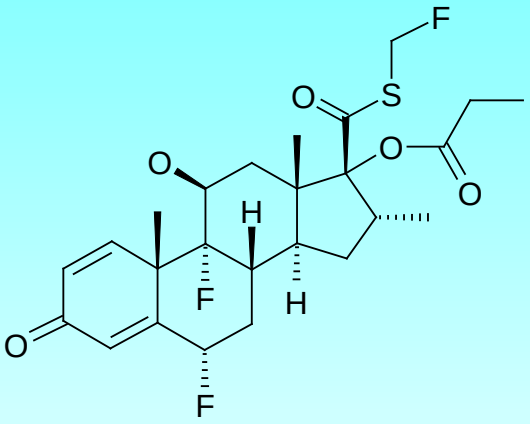
Atorvastatin (Lipitor, Caduet)
Pfizer, $12.9B



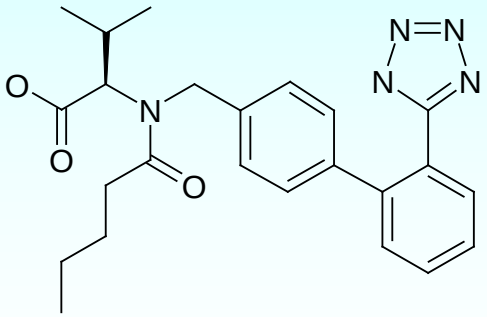
Clopidogrel (Plavix)
Sanofi-Aventis, BMS, $8.3B



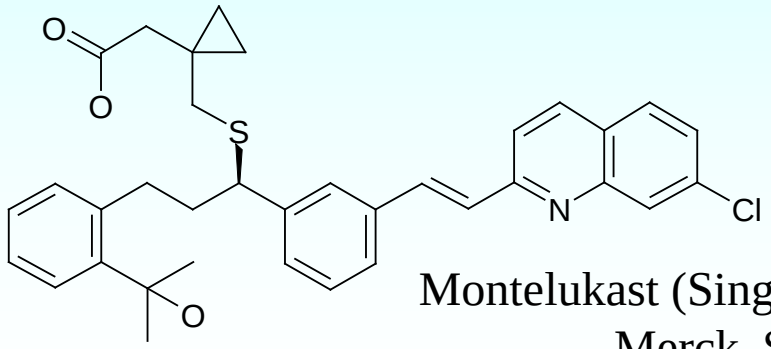
Esomeprazole (Nexium)
AstraZeneca, $5.2B



Fluticasone propionate (Flonase, Advair, Sertide) – GSK, $5.15B



Valsartan (Diovan)
Novartis, $5B



Montelukast (Singulair)
Merck, $4.2B

All of them were sold for over 4 billion USD in FY 2007

Non-Sense in Drug Discovery

- We cannot expect any learning method (machine or human) to perform well when predicting previously unmapped chemical and biological spaces, or the “unknown unknown”
- We like to think that tools to accurately predict the unknown unknown can be designed, perhaps using computers.
  - Some like to point out that computers consistently outclass grand masters in chess competitions. They fail to recall that chess falls into the “known known” category – though large, the number of moves on the chess board is finite, thus knowable.
- In the current climate of competitive pressure and “airplane assembly” analogies, pharma management wants everyone to believe drug discovery is akin to engineering new planes.
  - Emphasis is placed on execution, not creative thinking, because drug discovery is financially more lucrative by imitation (“me-too” drugs) .
- Financial forecasts, critical to pharma decision makers as well, are subject to “Black Swans” as well, and subject to severe limitations (have they ever worked properly?!)

Living with the Black Swan

- **Studies in innovation show that output does not correlate with R&D investment.** Positive Black Swans are breakthroughs, they are *unplanned* and can only happen when out-of-the-box thinking is encouraged
- Environment has to be conducive to creative thinking, and focus on **learning**, not on execution. Pharmaceutical industry analysts & managers should demanding that the search for a novel drug is as fast as a Google search.
  - This is not avionics, car manufacturing & banking.
  - Be flexible and open-minded about what constitutes (or not) good research. Recall wrong financial forecasts (Cimetidine, Omeprazole)
- Follow the market and let it dictate opportunities: While catering to “first-world” diseases remains a priority, the pharma industry needs to keep an eye on Asia and South America: These emerging markets have different therapeutic needs & sometimes very different disease profiles...

Living with the Black Swan… Too

- Option one (not embraced by Dr J. Typical Scientist): *Seek Enlightenment*… become omniscient. Transcendentally, there is no time, hence no need to predict. Quantum physics agrees!
- Option two: Expect the unexpected, but remain nimble, flexible, and open-minded. One cannot *wait* for Black Swans to occur, but one may have to act *swiftly* when such events happen (think earthquake, tsunami, terrorist attack… [global], & computer viruses, identity theft, fire, flood [local]…
  - Be ready to navigate crises swiftly, e.g., reports from adverse events, complete loss of laboratory data, surviving bad intel from competitors or industrial espionage, misleading information from literature/patents.
- Taleb alternates conservative and aggressive behavior.
- Fail early, fail safe(r)… **don't trust "secrets" & "tips"** revealed on the pages of mega-marketed magazines…
  - When stuck in airports because of frequent delays, just think…

    "I'll get there when I get there, not a moment sooner", smile and accept miracles. They do happen.

Considering a Career in Molecular Sciences

- **Science is not a democracy:** just because everyone believes something does not make it true
- **Choose topics that open new possibilities**, not those that (may) lead to dead-ends; don't get stuck with a single technology
- **Always go to the source** (original publication), as information gets to be sometimes selectively presented

- **Make sure you understand the basics** – try to explain what you do to a 5-yr old. If you manage, you grasped the concepts
- **Stay away from "fashionable" science:** just because it might get you funded, it does not mean it's science
- Make sure to **give credit where credit is due…**
- …but **do not be afraid to claim what's yours** (protect your ideas)

Final Take Home Message

- **Nothing is what it seems:** verify what you see, doubt what you find, and *always* get independent confirmation of your observations. For this reason, once you are sure about your findings you can be ready to defend your results (e.g., GPR30 still very contested by others)
- **Don't be afraid to say *I DO NOT KNOW***, omniscient beings are not of this world (think Buddha, Jesus, and other enlightened beings)
- Although you do not know, **be ready to learn**
- Focus on **problem-solving skills**, they are more important than static learning & memory
- **Always find ways to reward creativity and out-of-the box thinking**
- **People are 100x more important than equipment** – as you progress in your career, you will find that people are the most important asset
- If someone steals your ideas (this happends all the time) remember that this is a subtle form of flattery (so is envy). Focus on generating new ideas, and do not turn the stolen-idea-situation into an obsession (this will block your creativity)
- **Learn where your fear comes from:** deal with it inside, and do not take it on other people. Fear leads to anger, anger leads to violence
- Express yourself freely and creatively. The Universe is a friendly place.