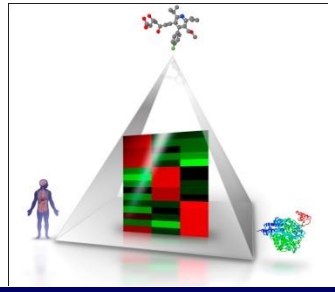
Chemical Space (Relevant to Drug Discovery), Virtual Screening and Library Design

Andreas Bender, PhD
Natural Philosopher for Molecular Informatics, University of Cambridge
Fellow of King's College, Cambridge
CTO, Healx

UNIVERSITY OF CAMBRIDGE



Outline

- Which protein space can we target (with current small molecules)? Which type of chemistry is favourable?

- If a starting point is known… virtual screening (descriptors, similarities)

- If no starting point is known… library design

‘Drug-relevant’ chemical space

Which small molecule chemical might we be interested in for drug discovery?

Difficult to answer… proxy questions:

- Which proteins *are current drug targets*?
- Which proteins *can be liganded*?

- Assumptions: No covalent binders, non-peptidic

Current drug and their targets

- 1,419 unique small-molecule drugs and 250 unique biologic agents were obtained
- 667 human, 189 pathogen targets

Table 1 | **Molecular targets of FDA-approved drugs**

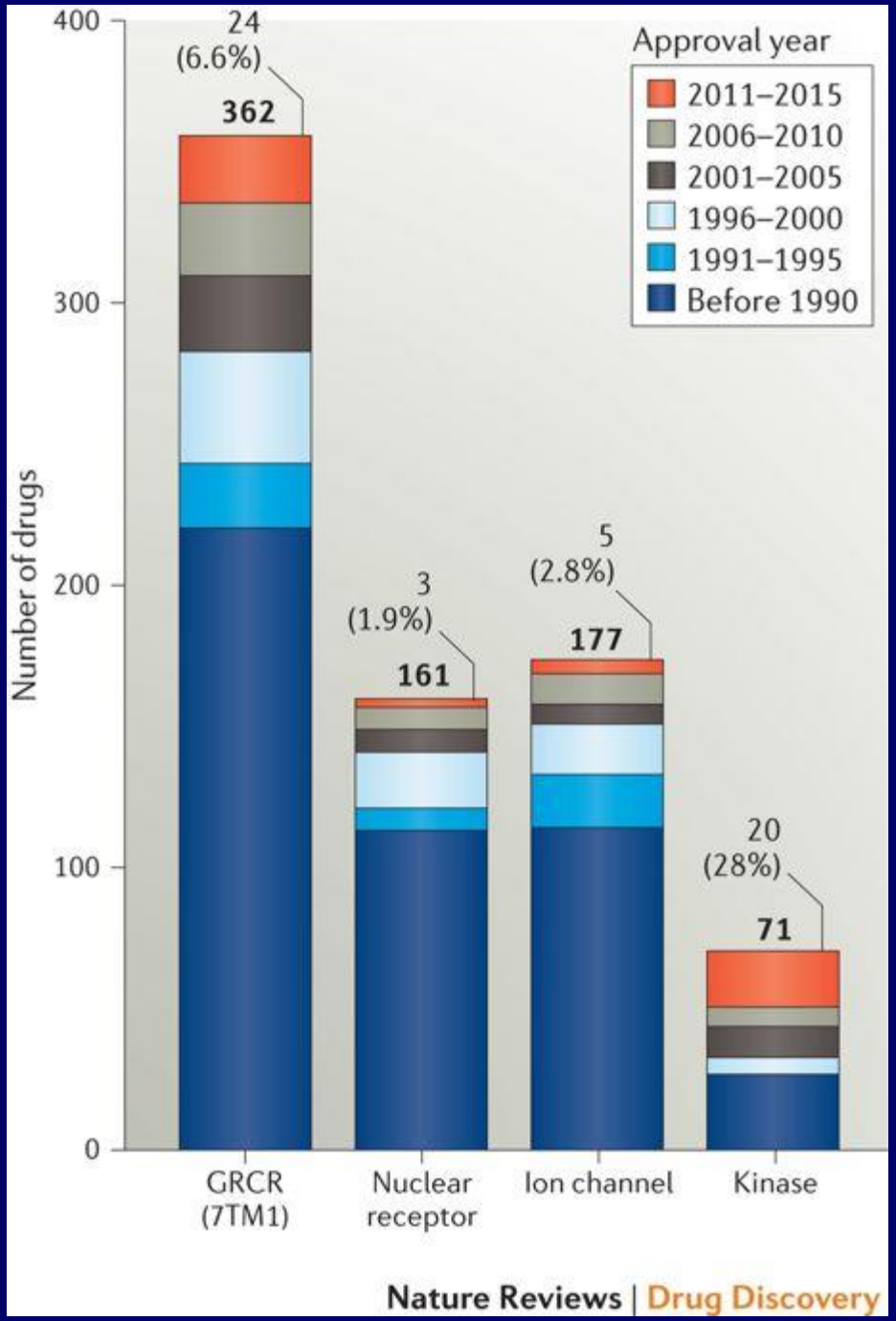
Drug target class	Targets			Drugs		
	Total targets	Small-molecule drug targets	Biologic drug targets	Total drugs	Small molecules	Biologics
Human protein	667	549	146	1,194	999	195
Pathogen protein	189	184	7	220	215	5
Other human biomolecules	28	9	22	98	63	35
Other pathogen biomolecules	9	7	4	79	71	8

The list also includes antimalarial drugs approved elsewhere in the world.

- Santos et. al. NRDD 2017

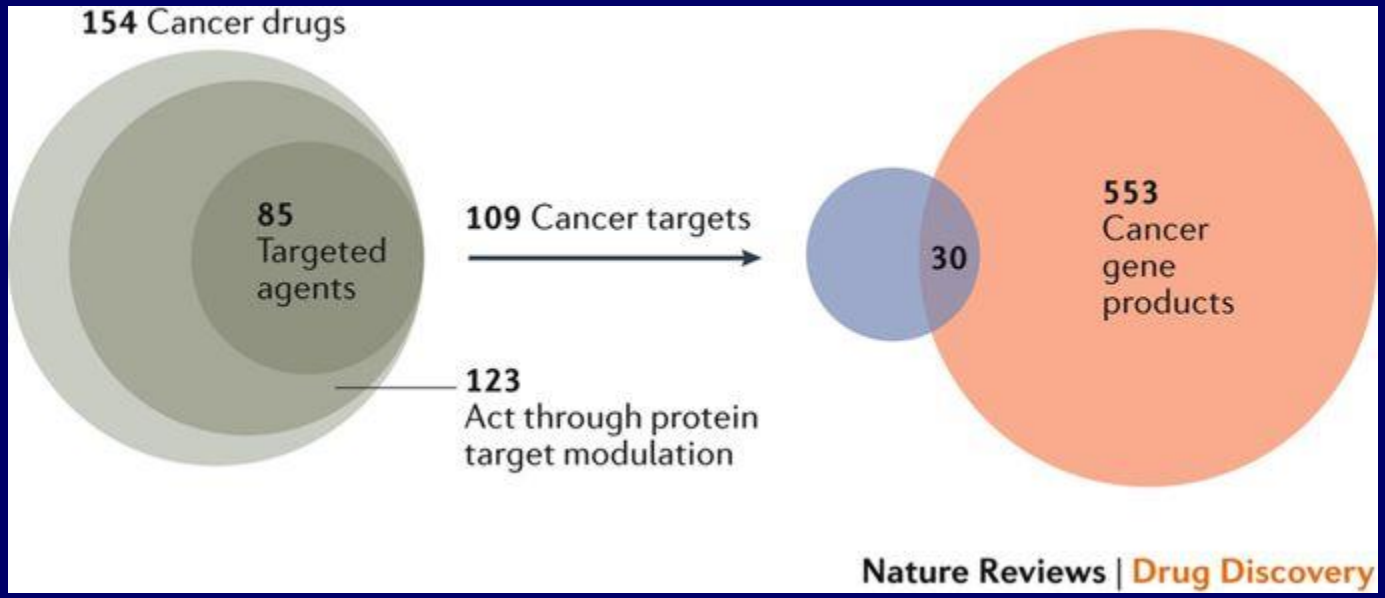
Targets change over time

- GPCRs (from 70s and earlier), then kinases (from 2000s), then epigenetics targets (this decade)…
- … changes ‘desired’ chemical space as well!



Disease drivers and drug targets are quite different beasts!

- Only minority of gene products driving cancer are targeted by drugs!



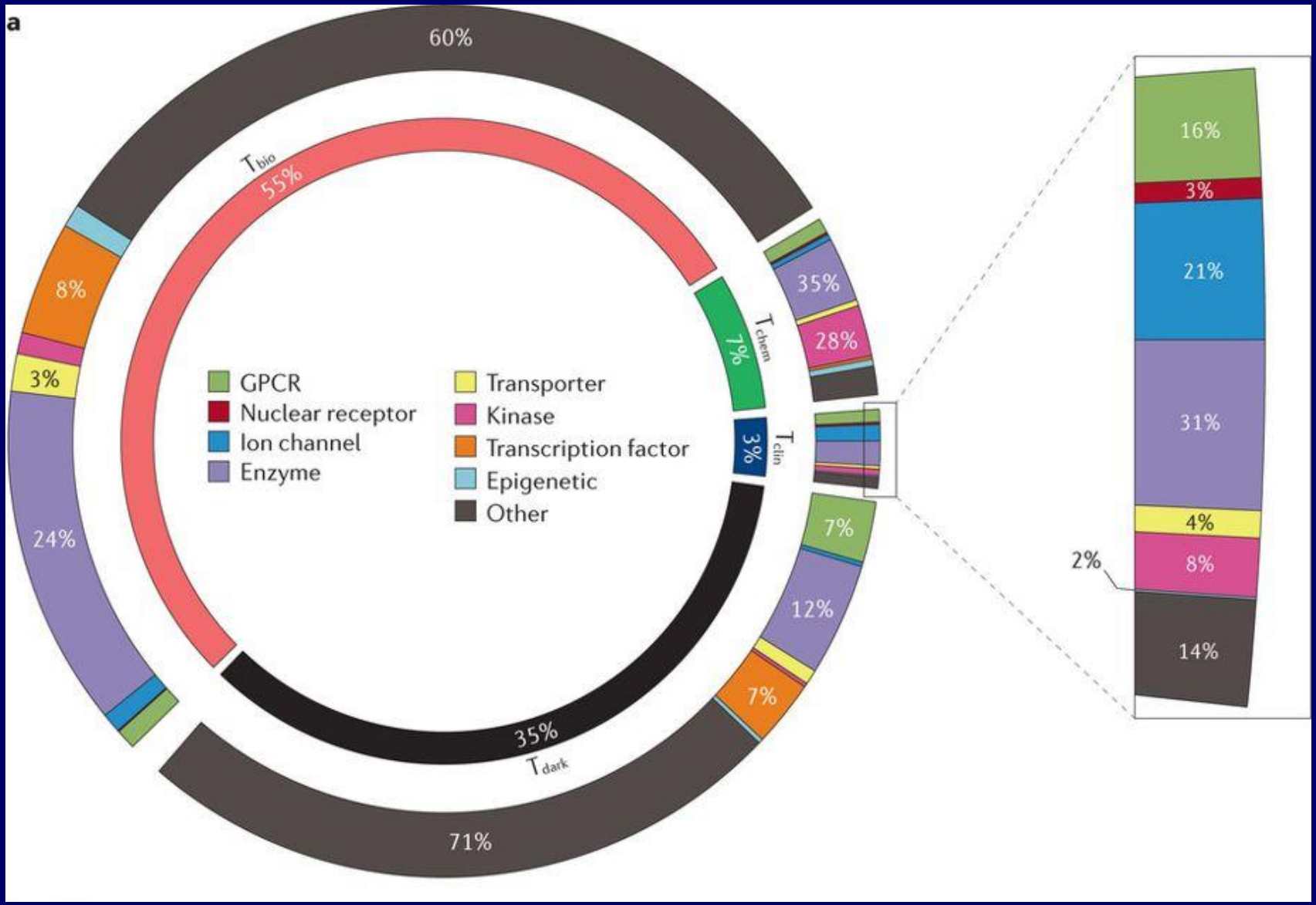
- Don’t mix up genetic drivers causing disease and ability to find small molecule modulator to reverse disease!

These were current drugs… how about ligandable proteins, those involved in disease?

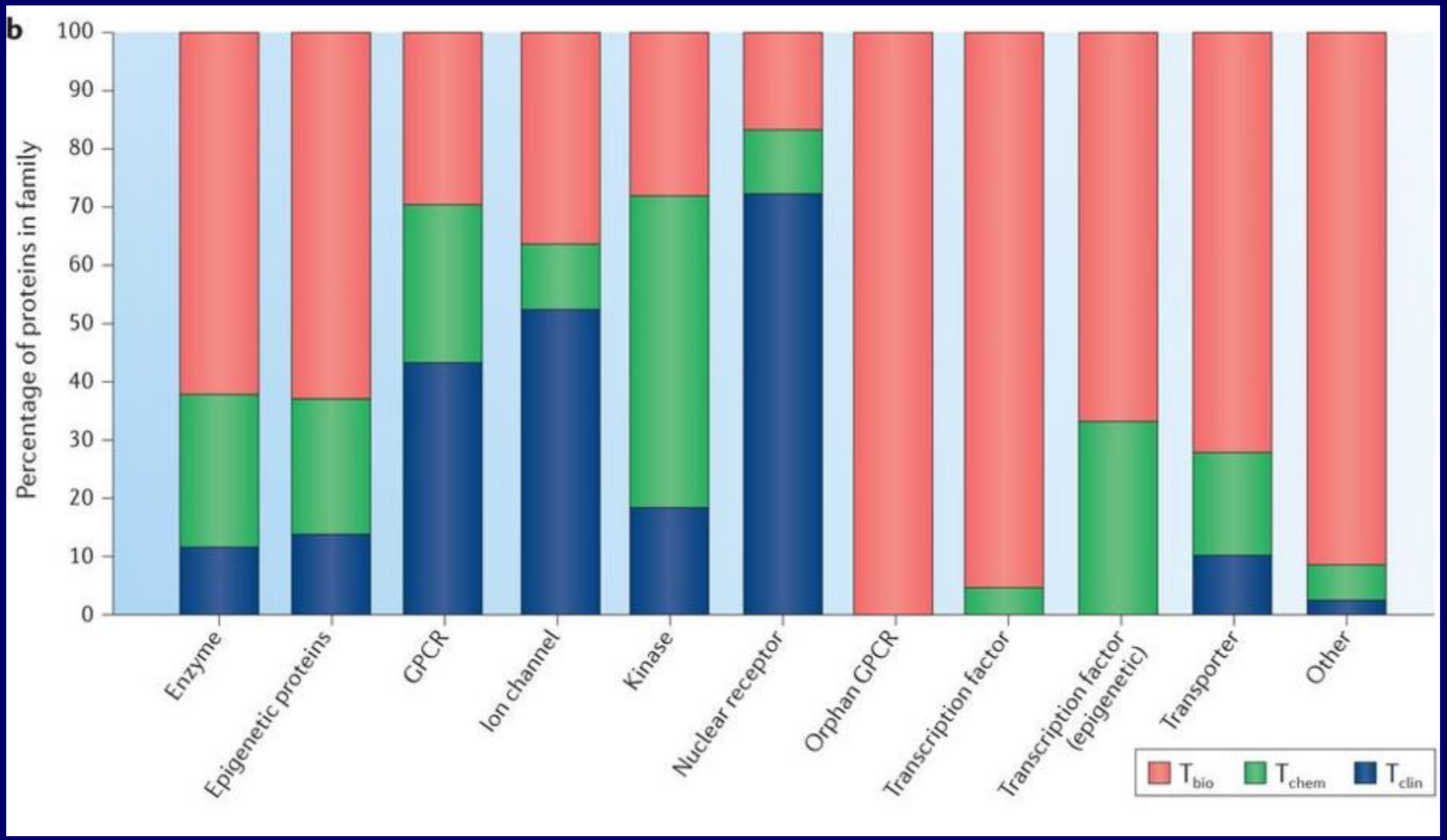
- Tclin (clinical) proteins – linked to at least one drug
- Tchem (chemistry) proteins – not linked mechanistically to current drug effect, but has potent ligand (≤30 nM for kinases, ≤100 nM for GPCRs and nuclear receptors, ≤10 μM for ion channels, ≤1 μM for other target families)
- Tbio (biology) - confirmed Mendelian disease phenotype in OMIM, or GO leaf term annotations based on experimental evidence, or >5 publications/>3 RIF annotations/>50 antibodies
- Tdark (dark genome) – remaining proteins (no high-affinity ligands, *insufficient* (but not necessarily no!) information about function

- Oprea et al NRDD 2018

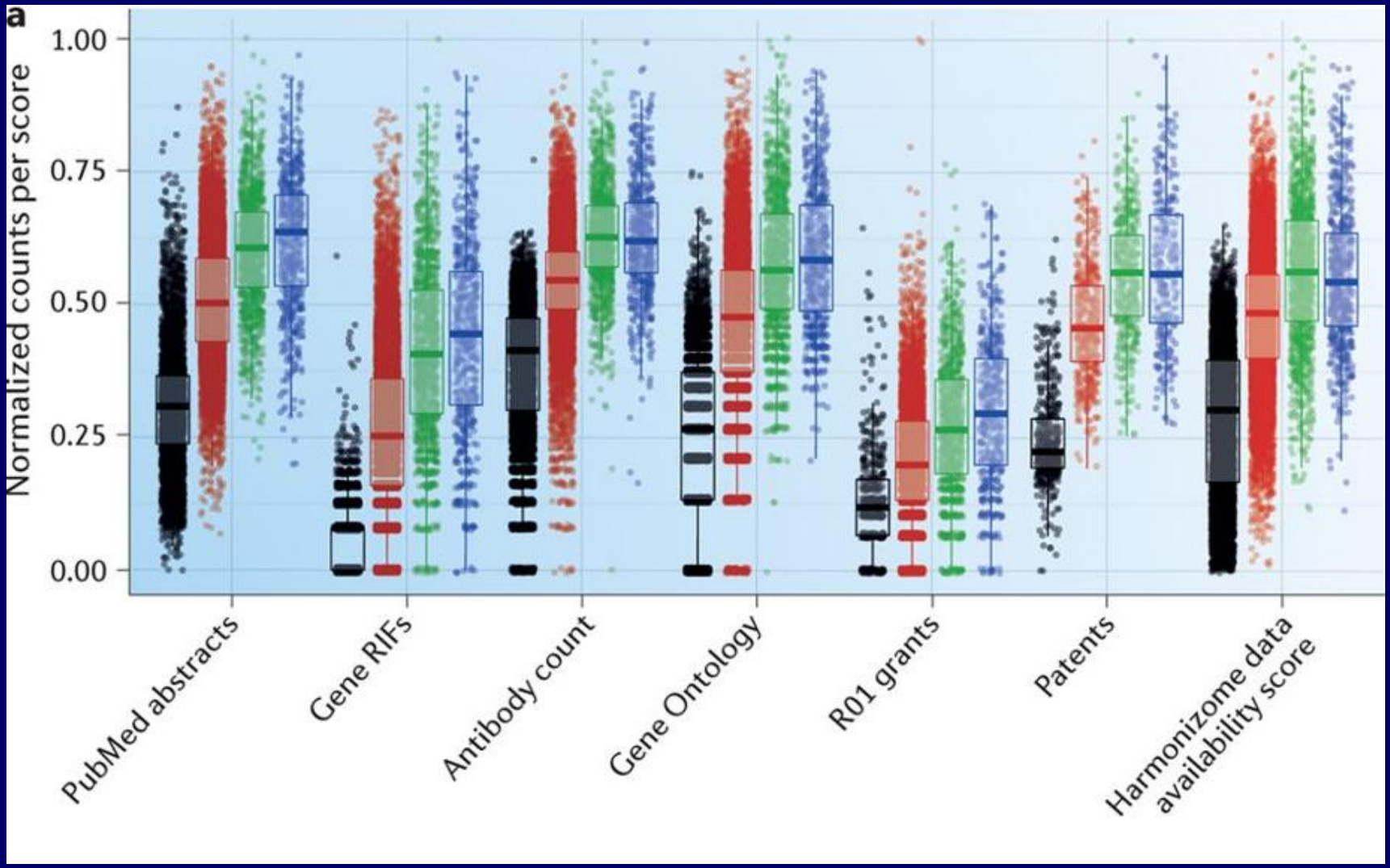
Current drugs focus on tiny amount of proteome (3%)



Proteins with disease associations in OMIM (n=3,644) – MoA diversity depends on target class



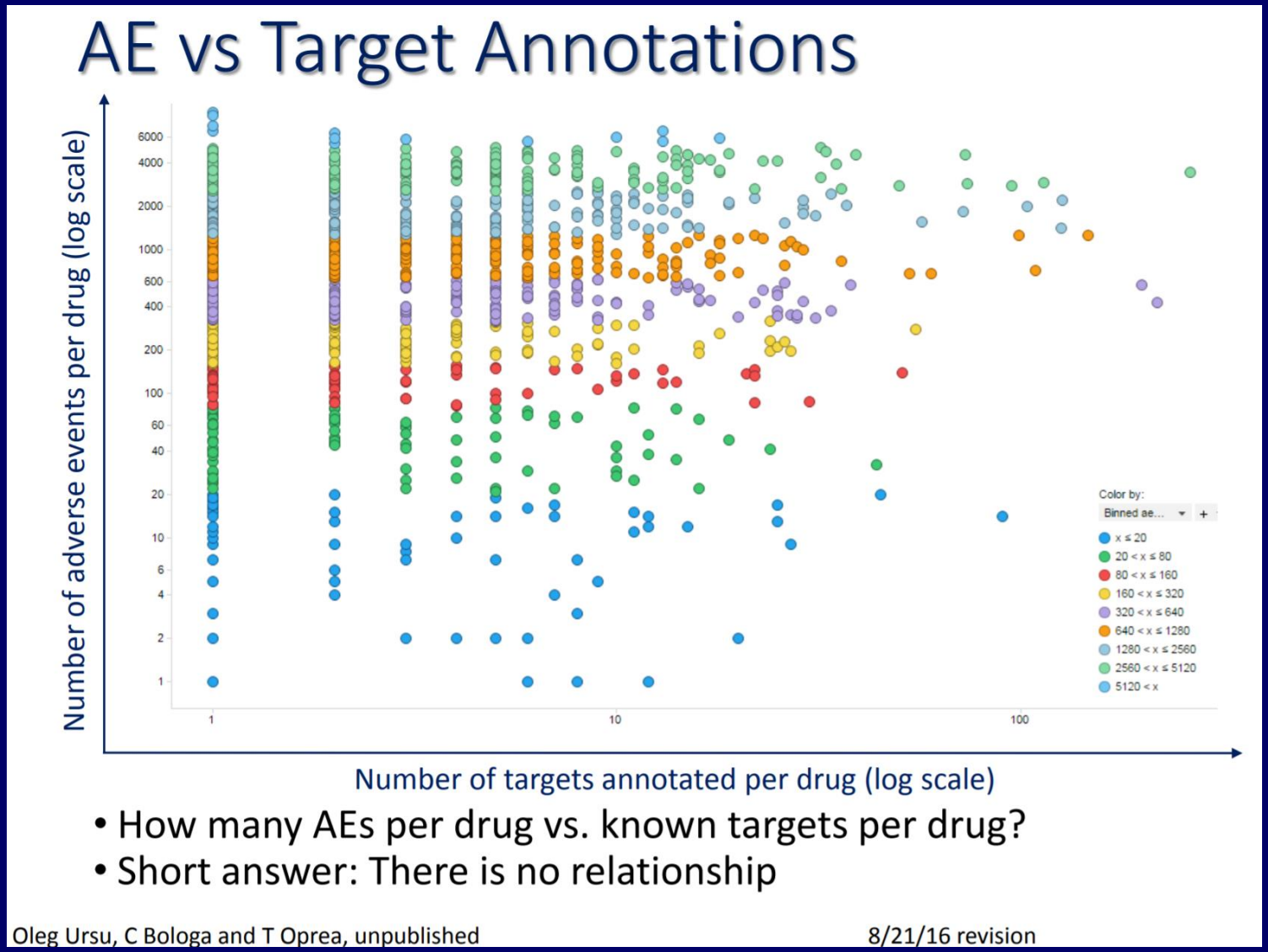
‘The rich get richer’… also current work (eg grants) is focused on few proteins



So what is current ‘drugged’ and ‘druggable’ protein space?

- Drug discovery *tends to follow past successes* (established chemistry, proteins); focuses on small number of proteins
- In many cases disease connection *and* small molecule ligand are known… can we match the two better?
- .. Do we need novel chemistry/new modalities? Likely – but even the current proteome which we have ligands for is underexplored!

Aside: Do selective drugs cause fewer side effects? Not really…



- How many AEs per drug vs. known targets per drug?
- Short answer: There is no relationship

Oleg Ursu, C Bologa and T Oprea, unpublished

8/21/16 revision

Are we looking for any particular physicochemical properties of drugs?

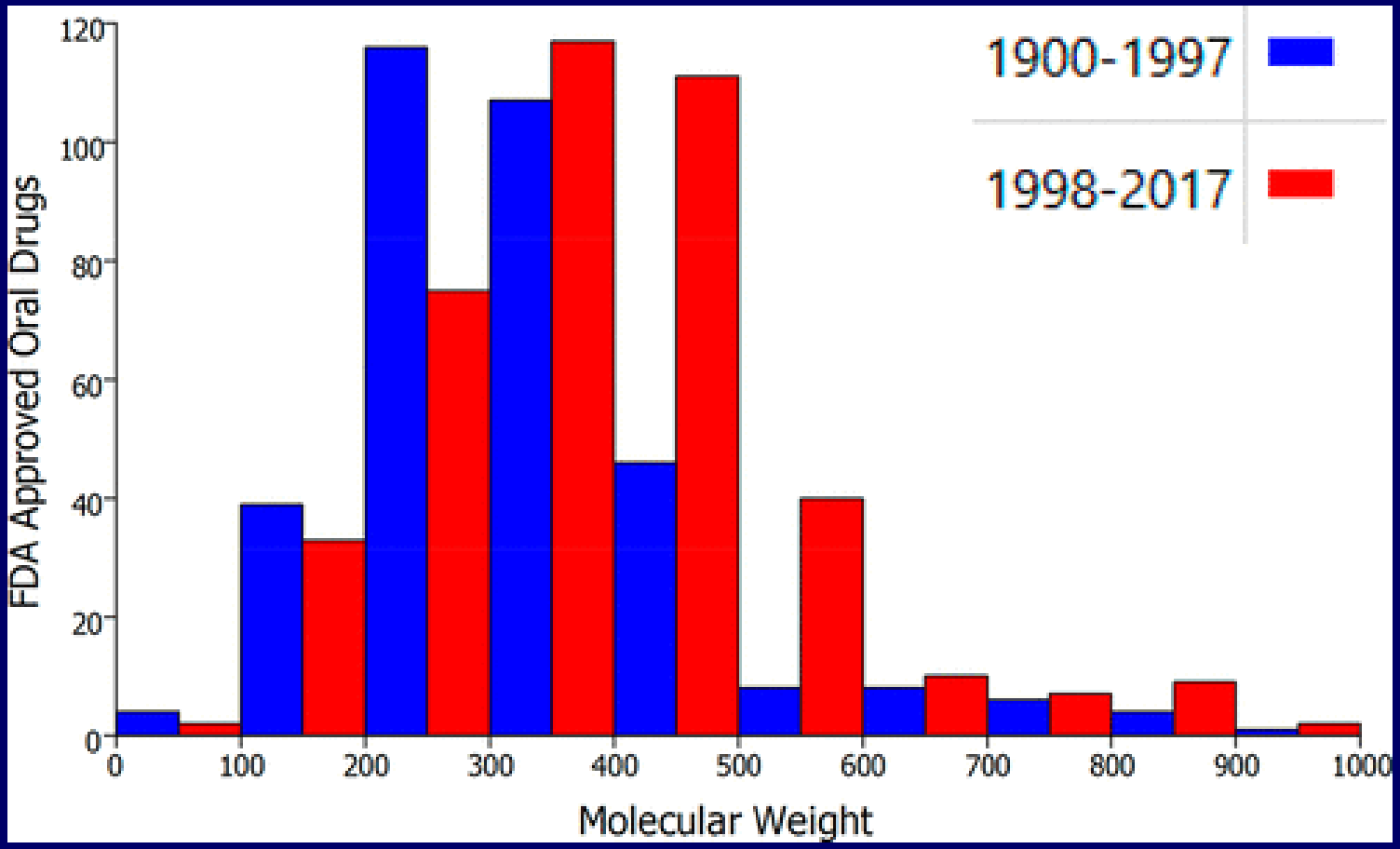
- Original Lipinski rules: “90% of orally available drugs have clogP<5, MW<500, HBD<5, HBA<10”

Table 1. Analysis of FDA Approved Oral NCEs from 1900 to 1997

	clogP	MW	HBD	HBA	TPSA	RotB	Fsp^3	#ArRNG
Ro5	5	500	5	10	NA	NA	NA	NA
90th percentile	4.7	470.3	4.0	10	139.8	10.0	0.83	3
median	2.3	308	1	4	67.51	4	0.40	1
mean	2.1	332.0	1.9	5.5	78.8	5.0	0.43	1.42
10th percentile	–0.6	171.2	0.0	2.0	21.3	1.0	0.08	0

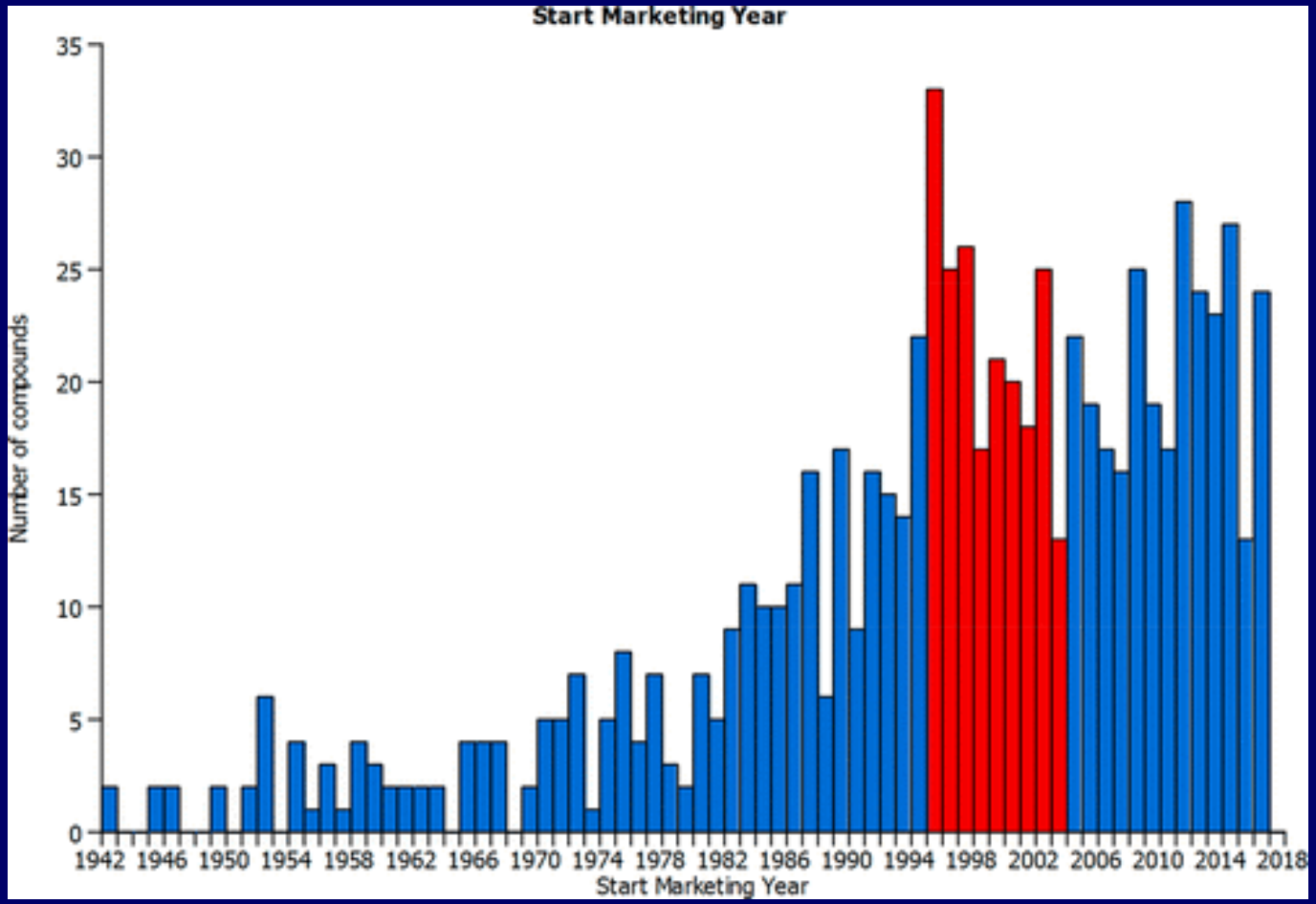
- Michael Shultz, J Med Chem 2019

‘Drug-likeness’ is a very time-dependent property!

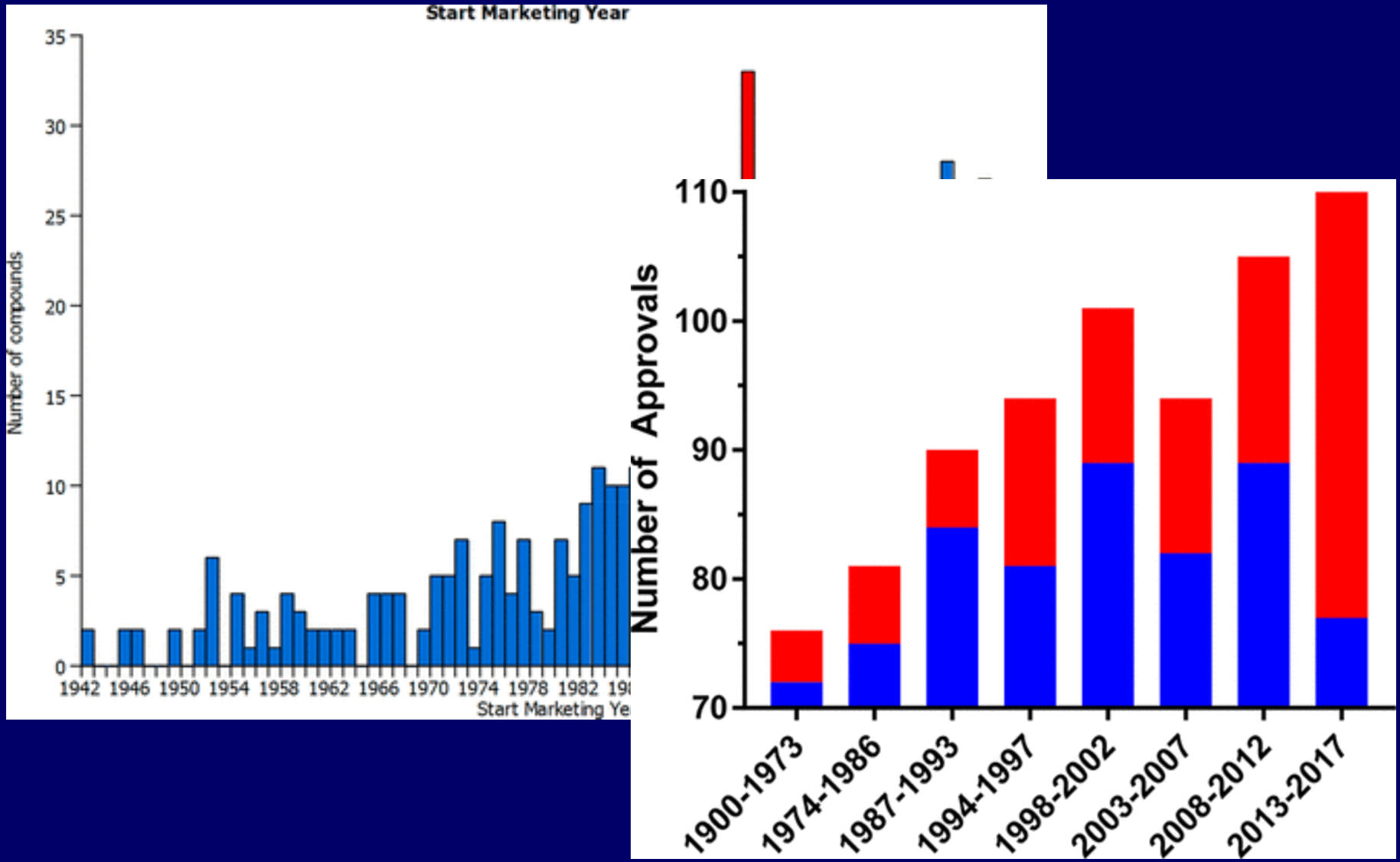


- So does it mean anything…?

Number of new drug approvals shown to correlate with *increase* in molecular weight!



Number of new drug approvals shown to correlate with *increase* in molecular weight!



Currently drugged proteome and requirements for chemistry

- Huge bias of research efforts on minority of proteins
- We know ligands for many more!

- Targets of interest, chemical space deemed to be ‘relevant’ changing, so concepts like ‘Lipinski’ rules need to be applied carefully

Outline

- Finding ligands if a starting point (active molecule) is know… virtual screening (descriptors, similarities)

Virtual Screening and Similarity Searching

- Virtual screening describes any *in silico* technique to select bioactive compounds

- Similarity searching is also called ‘ligand-based virtual screening’ (for individual compounds)

- Most often employed when no target structure is known, but even if it often gives ‘superior’ results to target-based virtual screening (depends on evaluation though!)

‘Similar Molecules Have Similar Properties’

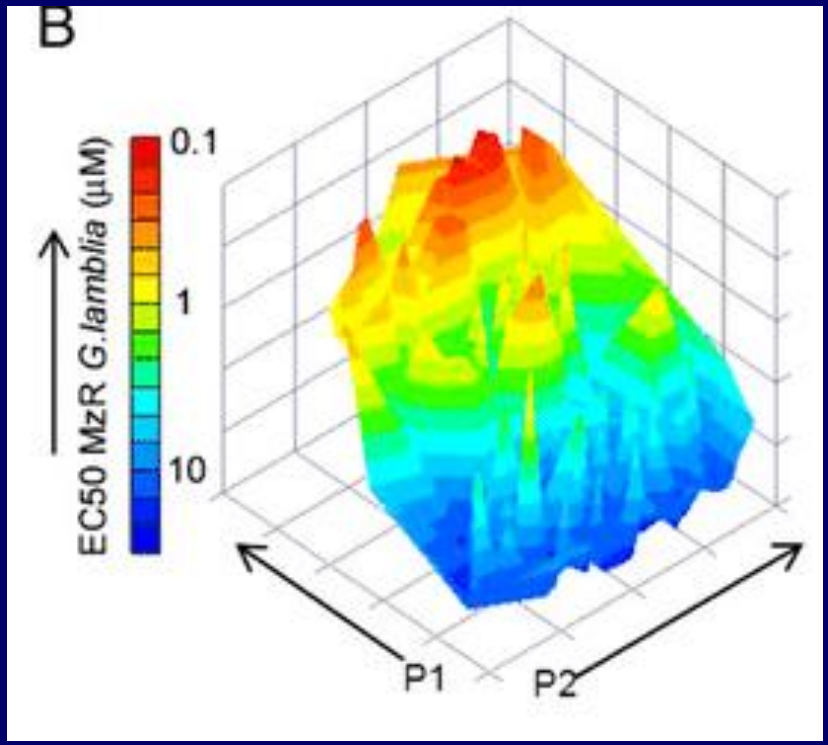
- Core assumption of virtually all computer-aided drug design

- The question is now how to define ‘similar’, and choose property of interest

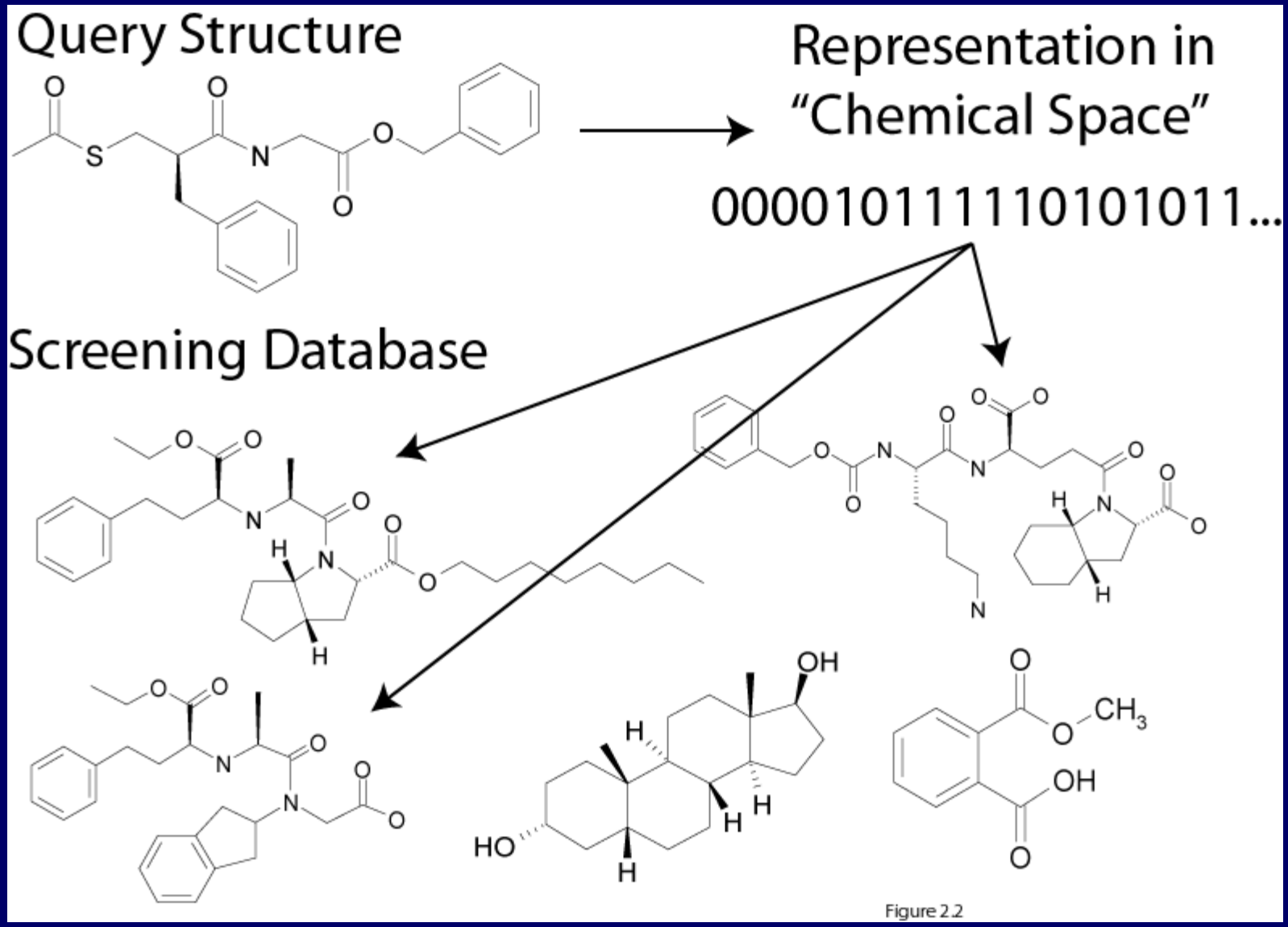
- To define ‘similarity’ involves usually two steps
  - Firstly, a molecular representation
  - Secondly, a similarity/distance measure

(Ligand-based) virtual screening

- We hope property of interest remains, but other properties change
- Assumes somewhat smooth activity landscape



Similarity Searching Requires an "Abstract" Representation (Descriptor)



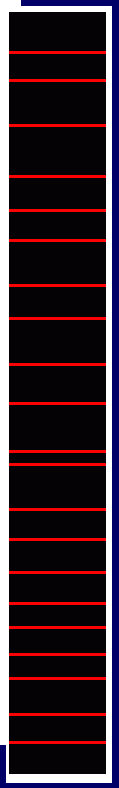
Similarity Searching in Practice

- Usually, one compound is used as a ‘query’ to rank the whole database available

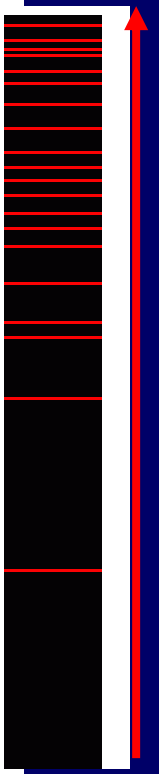
- Assuming that the query is active against a certain enzyme (receptor,…), the assumption is that the most similar compounds are *also* active against that enzyme (more generally, show similar properties overall)

The effect of sorting a database according to similarity to a query

- Without sorting, actives (=red) are distributed randomly across the database (black = inactives)



- With sorting, actives (=red) are more frequent among the compounds similar to the query

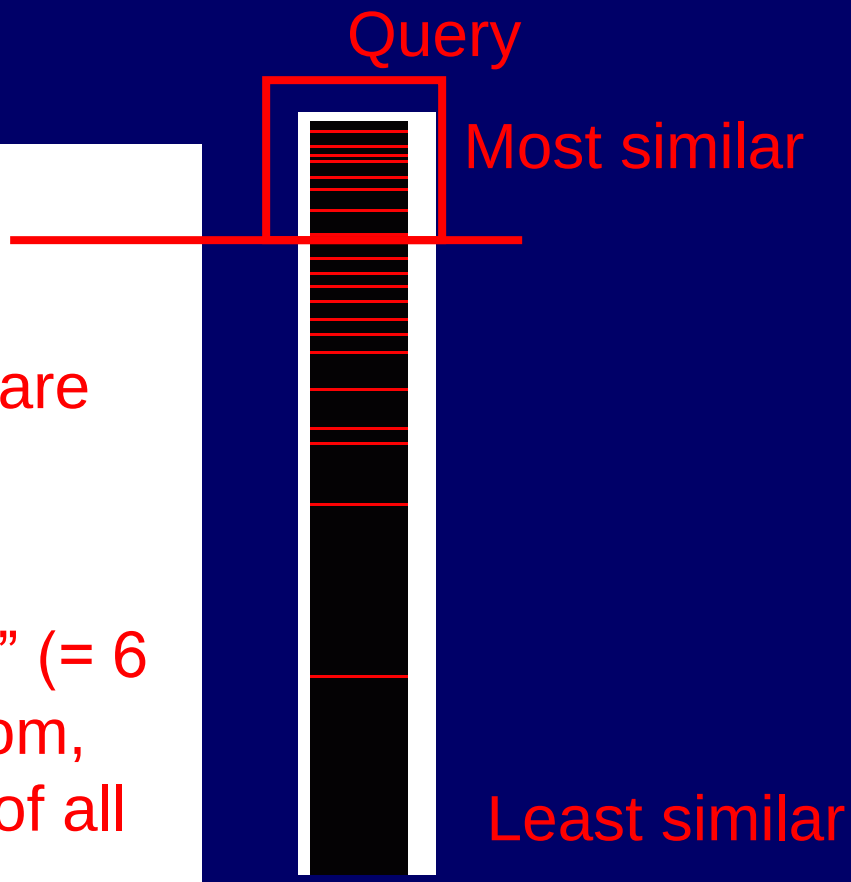


Numerical Performance Measure: Fraction of Actives in Top n% of Ranked Database

Cutoff at e.g. 10% of database –

e.g. 60% of all actives are found

Often written as an “enrichment factor of 6” (= 6 times better than random, which would find 10% of all active compounds



Problem: Evaluation!

- You don’t want to find ‘just the same thing’ again – performance measure needs to account for ‘novelty’

- For practical purposes, parallel prospective validation difficult
- Needs to be done on historical data
- … has biases

- Eg clustering, time-split validation possible

Difficulties with validations (2)

- Not learning the actual features relating to class, but correlated features (in both input features, and class labels)

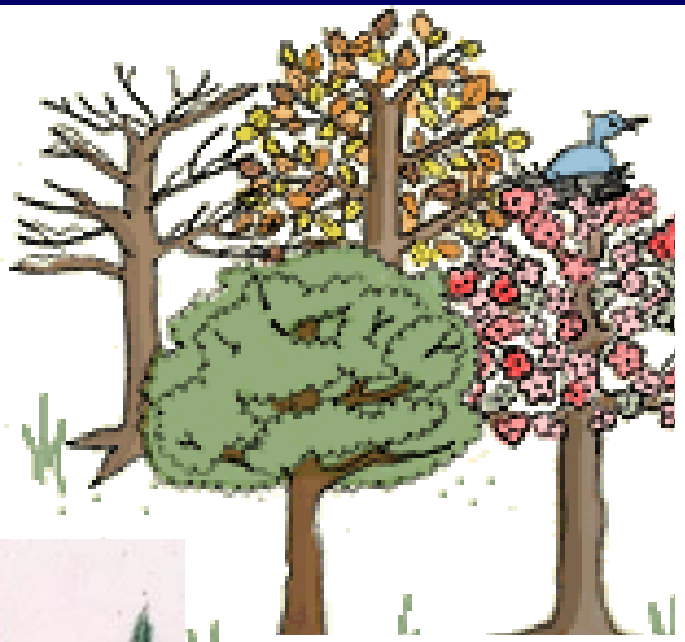
- Eg modelling activity in cellular assay – does your model simply capture permeability?
- Classifying compounds for DILI – are you modelling a different distribution of ATC codes/physicochemical properties?
- Etc

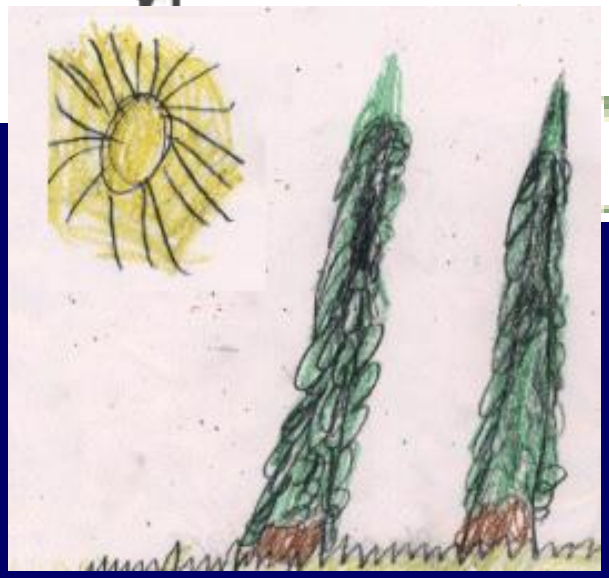
Descriptors: No ‘natural’ way to describe chemistry

- Sometimes reactivity (the particular functional group) matters (eg for toxicity), sometimes the surface (eg for ligand-protein interaction), sometimes the physicochemical properties (eg for permeability and solubility), …

- No ‘start’, no ‘end’ of molecule, difficult to encode 3D information (plus flexibility!)

Descriptor Choice – What Is A ‘Tree’?





1D/2D/3D (4D/5D/6D/…) Descriptor Classification

- Often, molecular descriptors are classified according to their ‘dimensionality’
- 1D descriptors: MW, polar surface area, …
- 2D descriptors: Fragments, fingerprints, …
- 3D descriptors: 3-point pharmacophores, shapes/surfaces, …

- 4D could include conformations

- Also image-based representations (2D picture of Kekule structure in pixel format) have been explored, etc.

Standard Descriptors – 1D

- 1-dimensional (1D) descriptors assign a single number to a molecule
- This number should be correlated with the activity one is interested in
- Examples: Physicochemical properties (logP, MW)

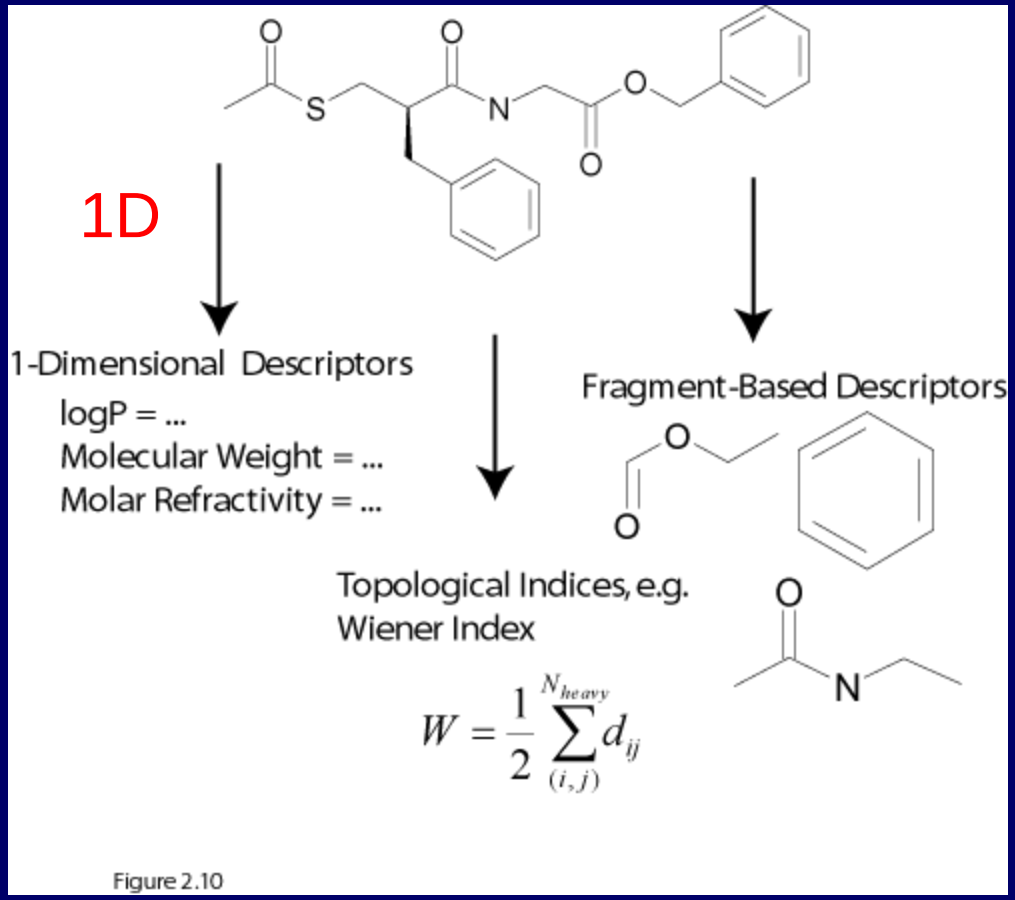
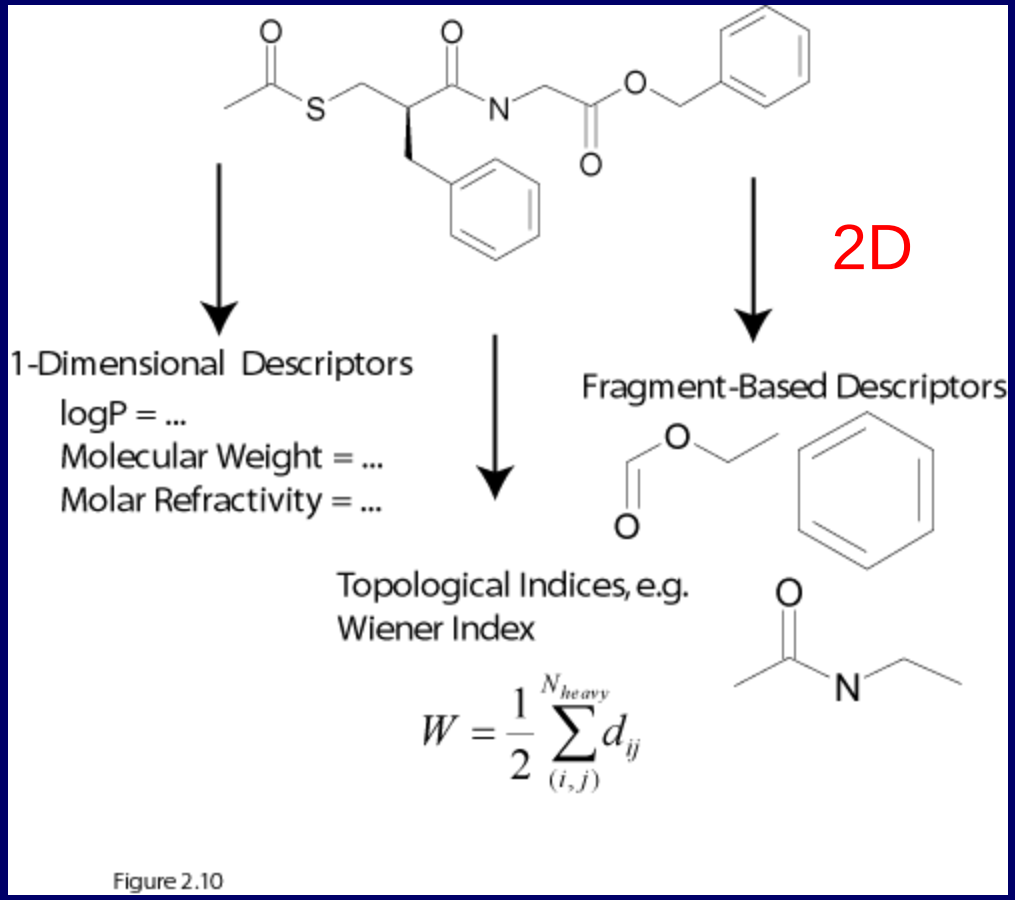


Figure 2.10

MW for example is a major determinant of absorption!

Standard Descriptors – 2D

- "2D" descriptors are based on the graph of a molecule
- Examples are substructures/fragments
- *Very* often used in practice – work well, quick to calculate!



Example of a well-performing 2D descriptor: “Circular Fingerprints”

- Describes a molecule by the arrangement of heavy atoms around the central atom
- Calculated for each heavy atom of the molecule
- Molecule is represented by short vectors (here 10 = number of heavy atoms)

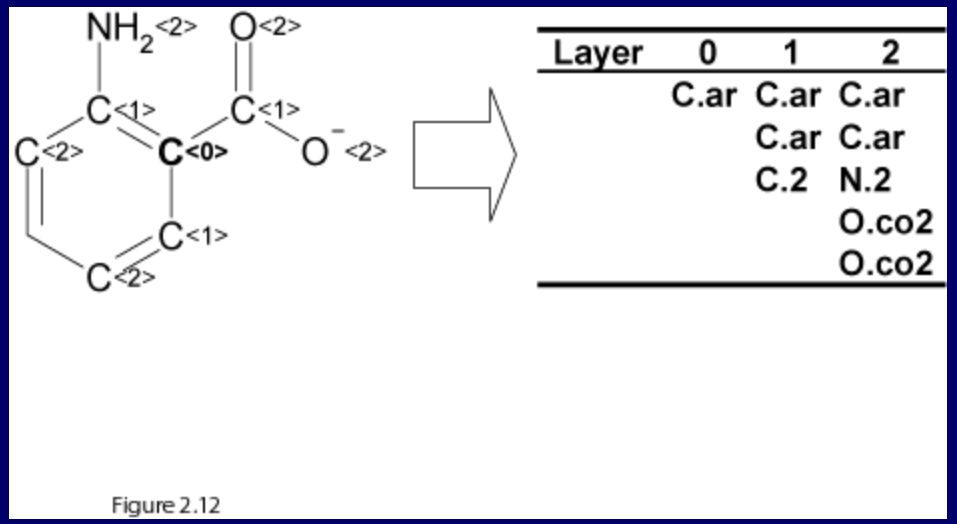


Figure 2.12

MDL keys… sometimes unintended uses work fine!

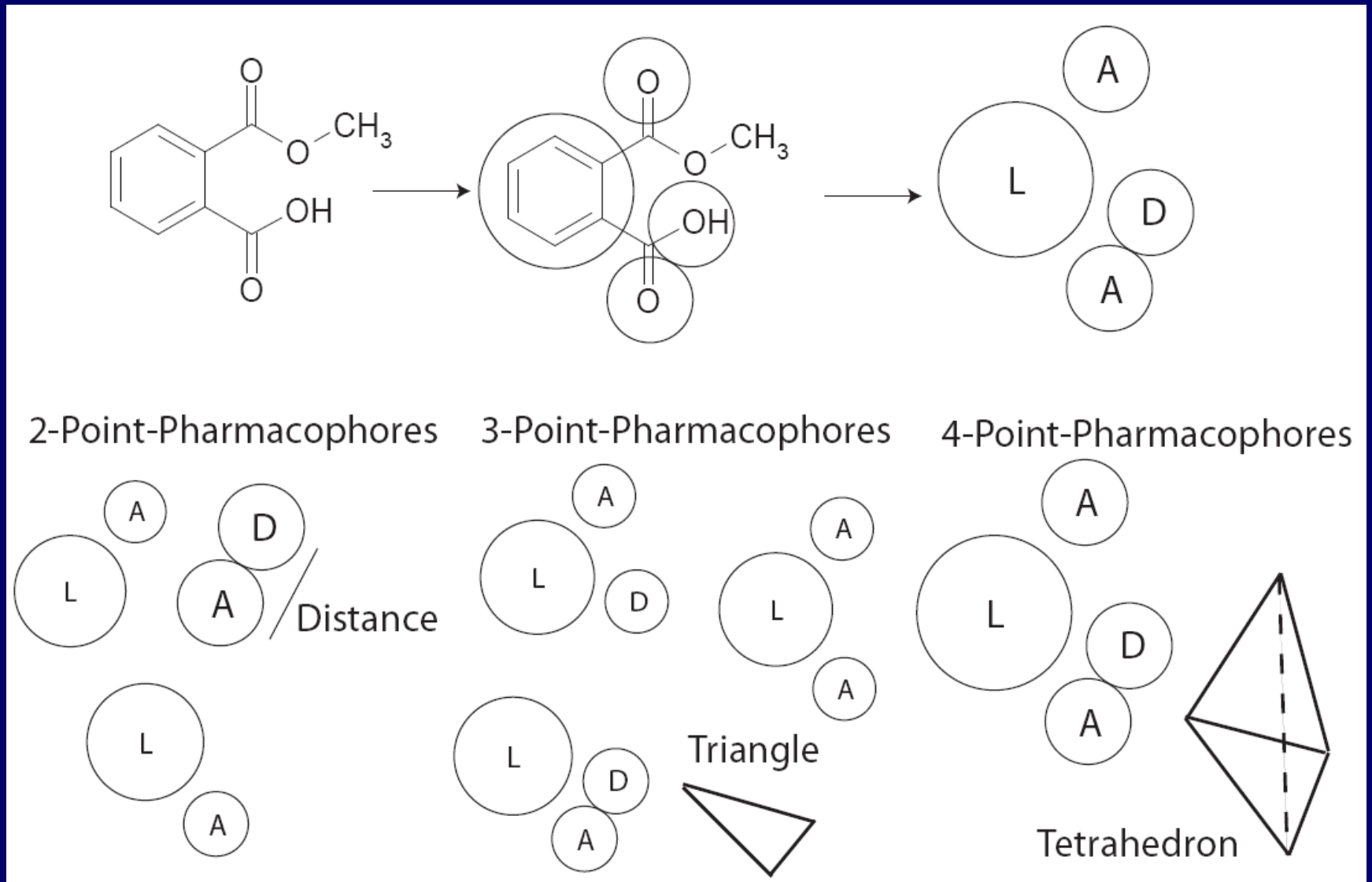
- 166 (public) key set defined by MDL in 1980s, originally to speed up database searches
- Have been ‘repurposed’ as descriptors, worked remarkably well in some areas
- Still have little resolution

n	A
0	null
1	atom with at least three neighbors
2	heteroatom
3	atom involved in some multiple bonds, not aromatic
4	atom with at least four neighbors
5	atom with at least two heteroatom neighbors
6	atom with at least three heteroatom neighbors
7	heteroatom with at least one hydrogen attached
8	carbon with at least two single bonds and at least two hydrogens attached
9	carbon atom in a C=C double bond
10	atom has at least two single bonds
11	atom has at least three single bonds
12	atom is in at least two different six-membered rings

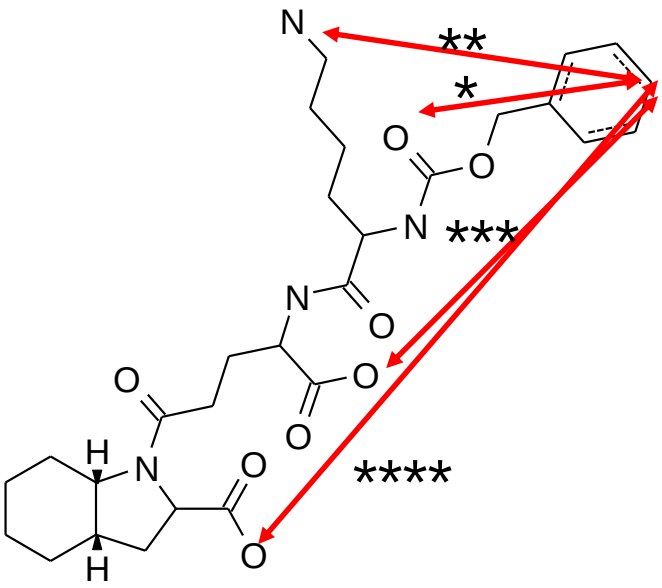
3D Descriptors: Invariance w.r.t. *translations, rotations and (sometimes) conformations is important*

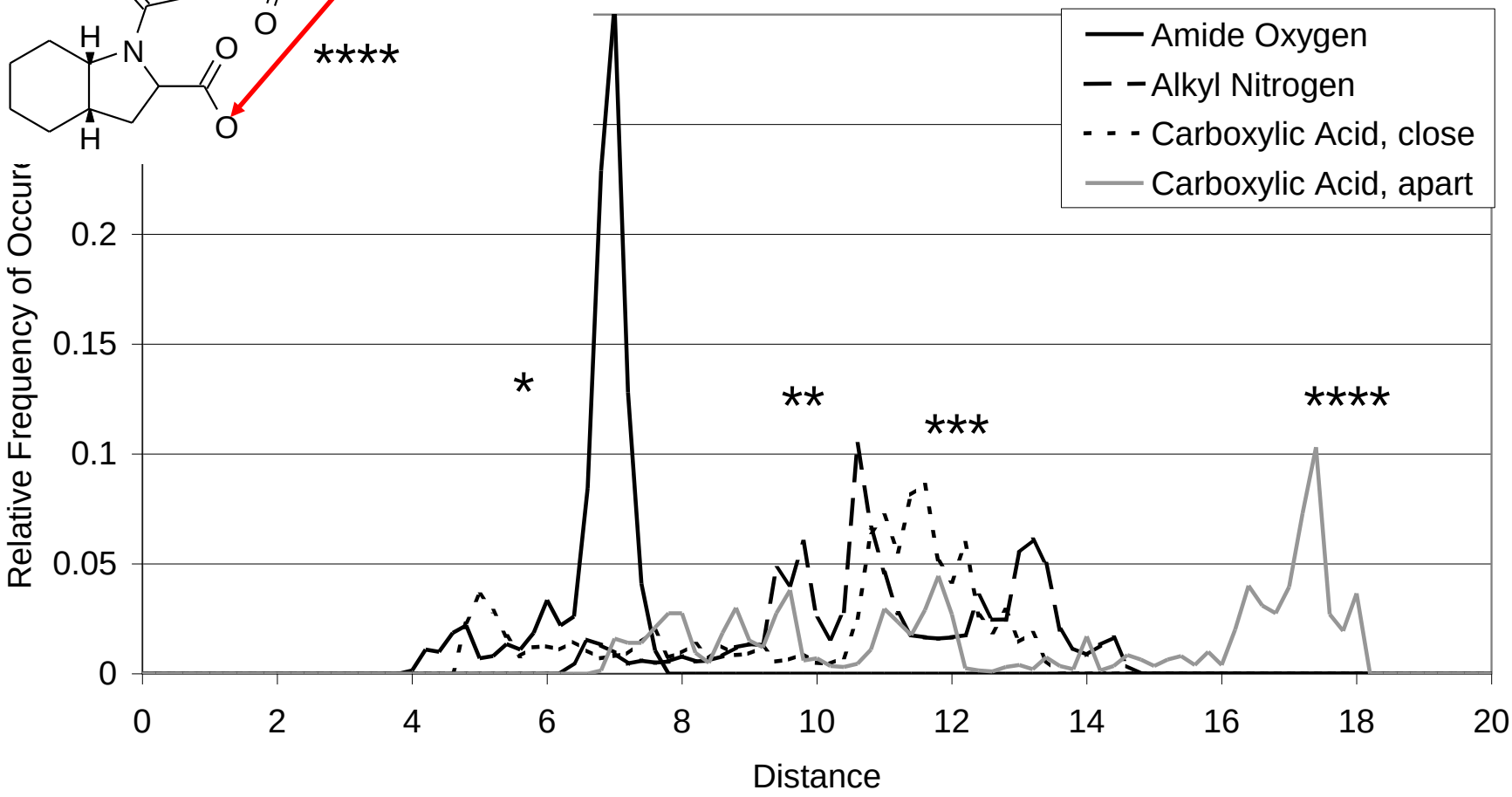
- In 3D, location and rotation of molecules are arbitrary (in absence of receptor)
- Thus, 3D descriptors need to be translationally and rotationally invariant
- Often, internal coordinates (distances between features) are used to achieve this
- Problem: Consideration of conformational flexibility; increases computational demand *and* can introduce noise!

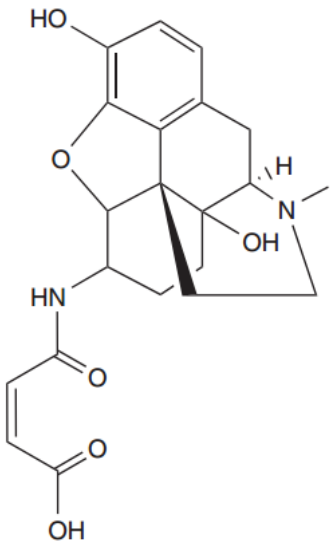
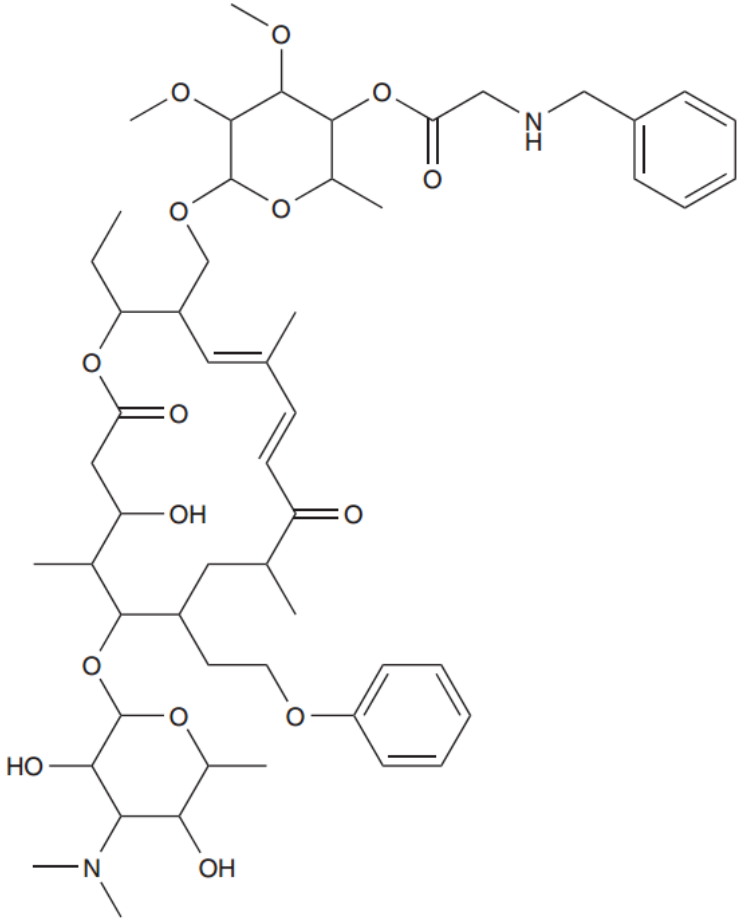
Pharmacophoric Descriptors – Going 3D



The Conformational Problem – Why 3D Can Be Tricky





Query structure	Target structure with most variance in ranking positions	Descriptors used and ranking position (out of 1000 compounds)	
		TAT	7
		TGT	23
		Similog	24
		piDAPH3	32
		FCFP2	38
		MACCS	42
		Unity	52
		MDL	66
		FPFP4	201
		piDAPH4	289
		ECFP4	395
		TGD	644
		TAD	710
		FCFC2	849
		FEPOPS	927
		SCINS	999
		ESShape3D	1000
		AtomCounts	1000

Two query structures and the corresponding targets structures are shown which, out of 1000 diverse drug-like compounds, showed most variance in ranking positions. Count-based fingerprints are able to perceive the considerable different in molecular size while presence/absence descriptors are not (for a detailed discussion see main text).

A. Bender “How similar are those molecules after all? Use two descriptors and you will have three different Answers” *Exp. Op. Drug Discov.*2018

Representation is one step – comparison of molecules the next

- Similarity vs
- Distance vs
- Dissimilarity (Diversity)

Comparing Molecules: Similarity Coefficients

- Common representation of molecules: Bitstrings, giving presence and absence of features

- Comparison *via* “Similarity Coefficients”, which assign one number (similarity index) to two fingerprints

- Most common: Tanimoto Coefficient: Tc = AND / OR
- AND = bits in common; OR = bits set in either string
- Example: Tc (111000,101101) = 2 / 5 = 0.4

- Tc in [0;1] – 0 means no features overlap, 1 means all *features* overlap (but doesn’t mean it’s identical molecules!)

Similarity vs Distance Coefficients

- “A table and an elephant are ‘similar’ (‘zero distance’) because both cannot fly”

- Distance focuses on differences (in descriptor space) – if this space is not well-chosen ‘no distance’ is meaningless!
- Similarity focuses on both common and absent features

- [1,1,1,1,0] and [1,1,1,1,1] Tc = 4/5; Distance = 1
- [0,0,0,0,0] and [0,0,0,0,1] Tc = 0/5; Distance = 1

- Often, similarity coefficients behave more as ‘what you expect’ psychologically

Is dissimilarity the opposite of similarity?

- Formally, dissimilarity *could* be seen as 1/similarity, or 1-similarity

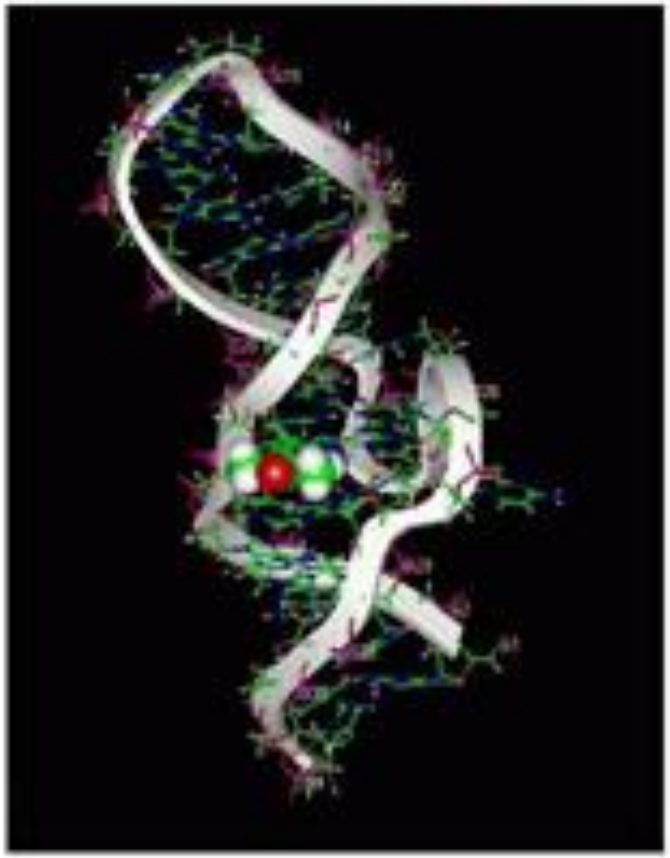
- *But* similarity measures (and features) have resolution in *close proximity*… and much less resolution if far apart

- Application of similarity measures for diversity applications can be tricky

The importance of shape – overall similarity can be very misleading!

University of Heidelberg BASF

Selective Recognition of Theophylline by RNA



A theophyllin-binding aptamer binds theophylline (R = H)

10,000-times better than caffeine (R = Me)

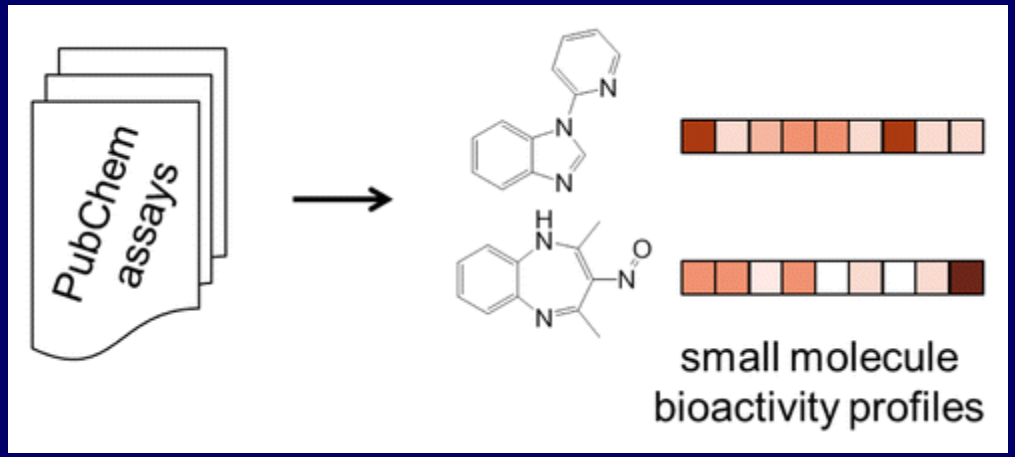
G. R. Zimmermann et al., Nat. Struct. Biol. 4, 644-649 (1997)

Similarity is not intrinsic property of objects

- Similarity depends on context
- Ideally represented in features chosen, similarity measures
- But difficult to define… often empirical

- Using *external reference frame* for similarity has become more popular recently

Representation using biological information – HTS, docking, predicted targets, GE, images, …



Query Fingerprint	1	1	1	0	0	0		
Evaluated in *Panel* Of Bayes Models	Target1	Target2	..	Target 1000				
	-0.39	43.21	..	3.4				
Evaluated in *Panel* Of Bayes Models	Target1	Target2	..	Target 1000				
	-0.24	33.21	..	3.1				
	13.4	-3.3	..	13.9				
Library Structure Fingerprints	1	1	1	0	0	0		
	1	1	0	0	0	1		

Pearson Correlation between Query and Library Bayes Scores
0.87
0.21

Is it possible and sensible to define "molecular similarity"?

- YES, but one needs to be careful …

- Similarity depends on the *context* (e.g. the particular receptor – different in case of non-directional properties, e.g. logP, solubility etc.!)
- Chose a meaningful representation, and similarity measure, depending on the *purpose*

Library design, eg for high-content screens

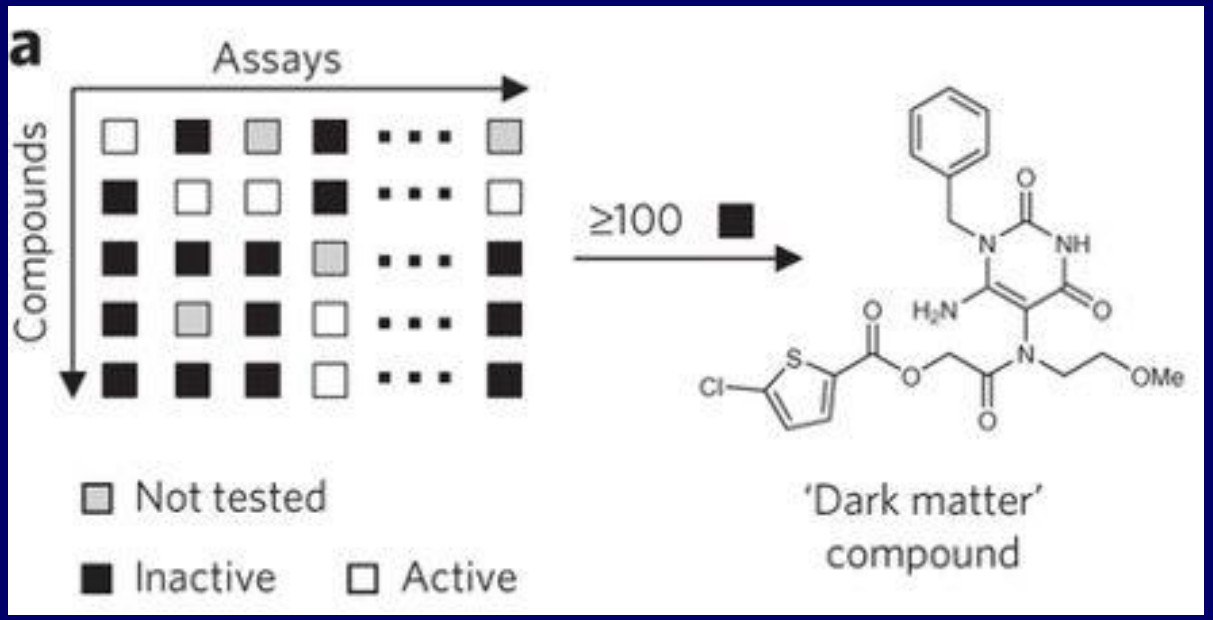
Aim: To have a manageable selection of small molecules for testing – reasonable starting point for drug discovery (‘good’ hit rates, suitable chemistry, physchem properties etc.)

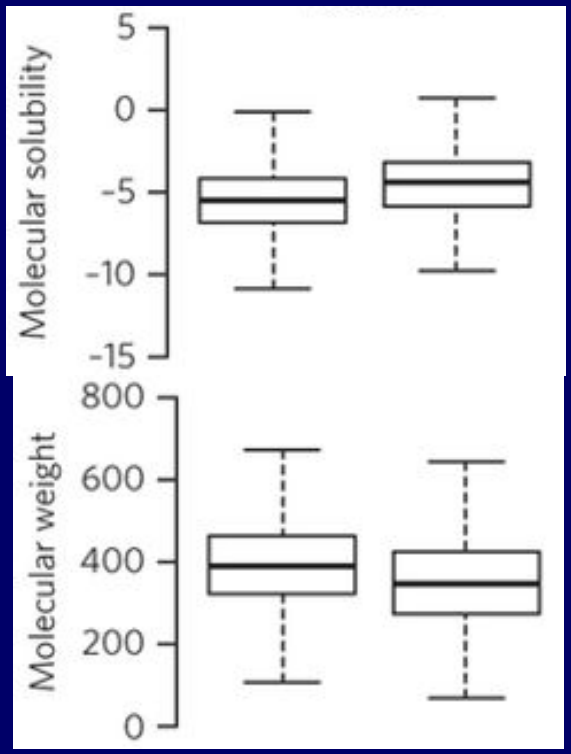
Two aspects:

- Which space to sample from
- How to sample

Chemical space is not equal

- Eg promiscuous compounds (kinase inhibitors, GPCR ligands, …)
- Opposite end of spectrum: “Dark chemical matter”, Wassermann et al., NCB 2015





Quite a number of different properties relevant for ‘drugs’

- Active, soluble, not toxic

- Tolerable off-target/side effects (multitude of factors!)
- Favourable PK/PD properties (multitude of factors!)

- Practical factors: synthesizability, cost, patentability, etc.

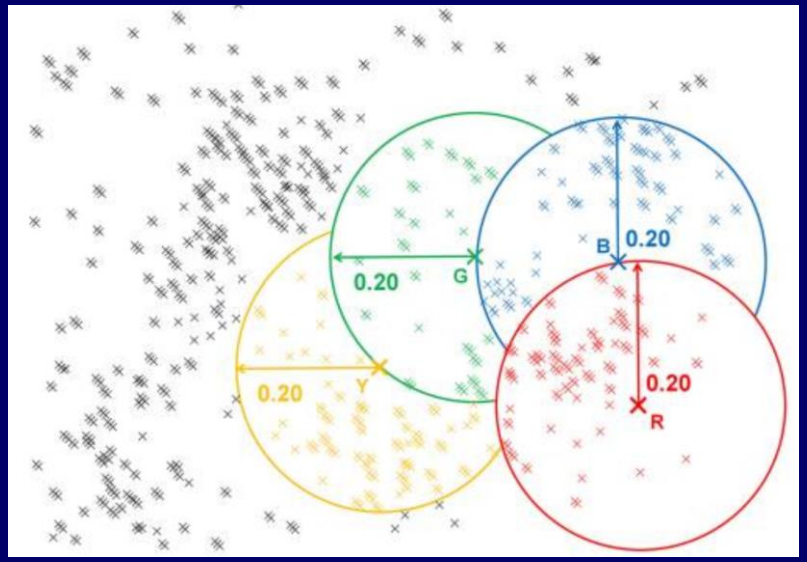
Representation matters

- Chemical (eg fingerprint) representation
- Shape diversity (eg sampling the ‘pocketome’)
- Substructural diversity (eg different ring systems, scaffolds)

- Functional diversity – eg based on HTS-FPs, predicted bioactivities, image-based/GE readouts, …

‘Maximum Diversity’ can give you unsuitable compounds!

- Eg clustering-based (difficult on large collections)
- Maximizing overall diversity (dissimilarity) *not* generally suitable!
- Max-min solutions (maximizing minimum distance between compounds) more suitable



Problem with diversity… can we actually pick a ‘representative subset’?

- We operate in a very high-dimensional space (say, 100-dimensional)

- *If* the bioactivity landscape is very smooth we could pick ‘representative’ examples (if dimensionality is not too high)

- *However*, if impact of bioactivity cliffs is profound we can *never* expect to sample bioactivity space, even conceptually…

Practical aspects of diversity selection for screening

- Screens are not perfect (noisy)
- If we want to learn about the activity of a compounds, quite often we also would like to
  - Know about confidence (do similar molecules have similar effect)?
  - Know about the SAR around the molecule, to have starting points for optimization (and avoid shallow SAR!)

- So practical selection can differ from theory!

Empirically frequently used solution

- Use set of well-annotated compounds
- MoA known, diverse MoAs
- Multiple representatives per MoA class, different scaffolds
- Leads to ~1,000-10,000 (or so) compounds for which this is possible

- Avoids more theoretical considerations of designing screening libraries
- *But* stays within well-annotated chemical/bioactivity space

Diversity library design conclusion

We need to

- Define relevant chemical space of interest (difficult to do looking forward!)
- Have some kind of descriptors and similarity measure (based on chemistry or effects)
- Diversity selection in high-dimensional chemical space is not trivial (dimensionality and number of data points)
- Eg annotated libraries good starting point

Summary

- Drug discovery, biological and chemical space we currently ‘know’ is hugely biases by the past

- If a starting point is know we can perform virtual screening (descriptors, similarities)

- If no starting point is known we can design libraries for prospective screening