

# ***Can we predict good drugs ?***

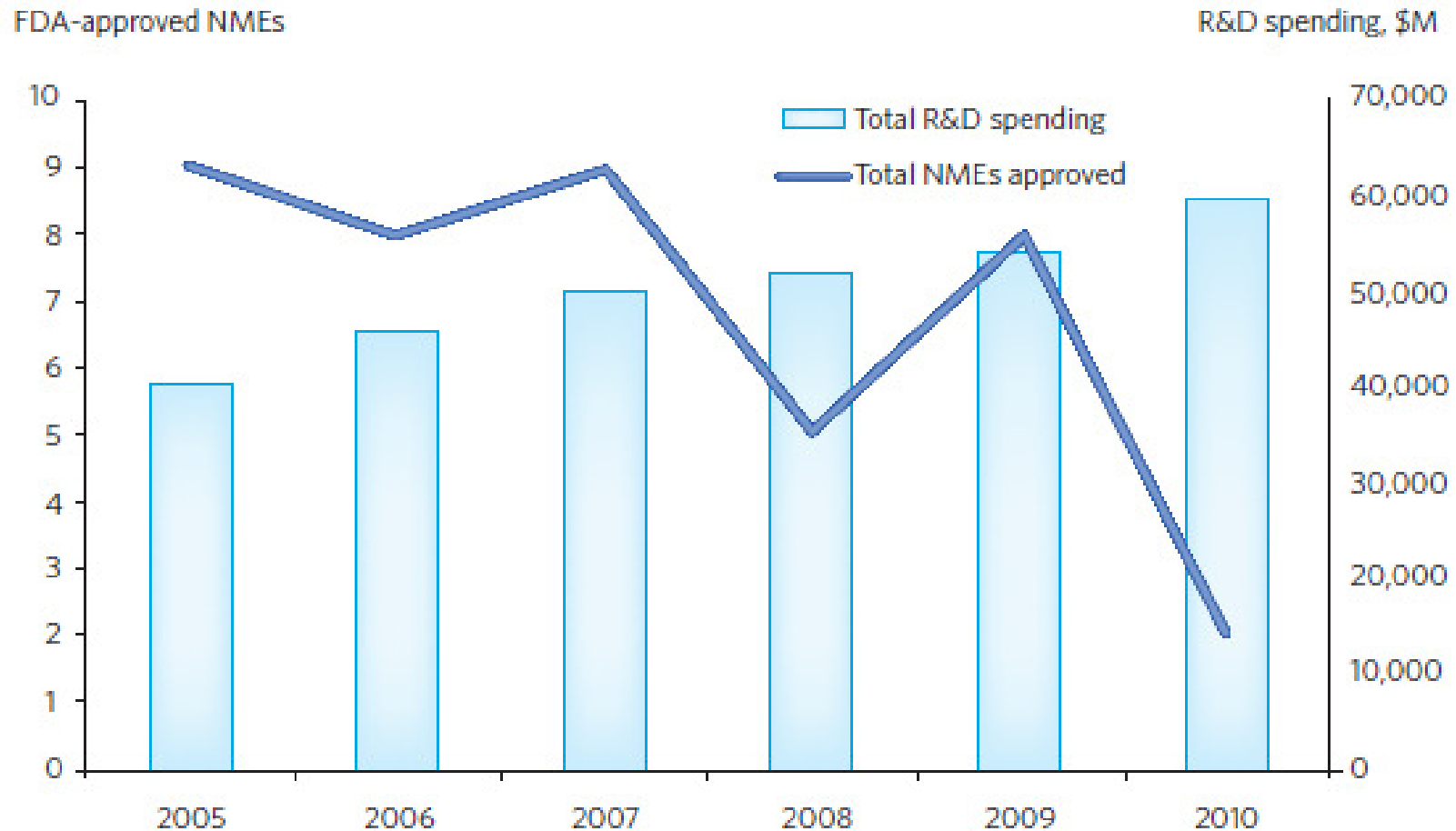
***the application of  
~ big database to R&D of drugs ~***

***Nature 2012, 486, 361.***

10<sup>th</sup> November 2012 (Sat.)

Takuya Matsumoto

# Introduction 1



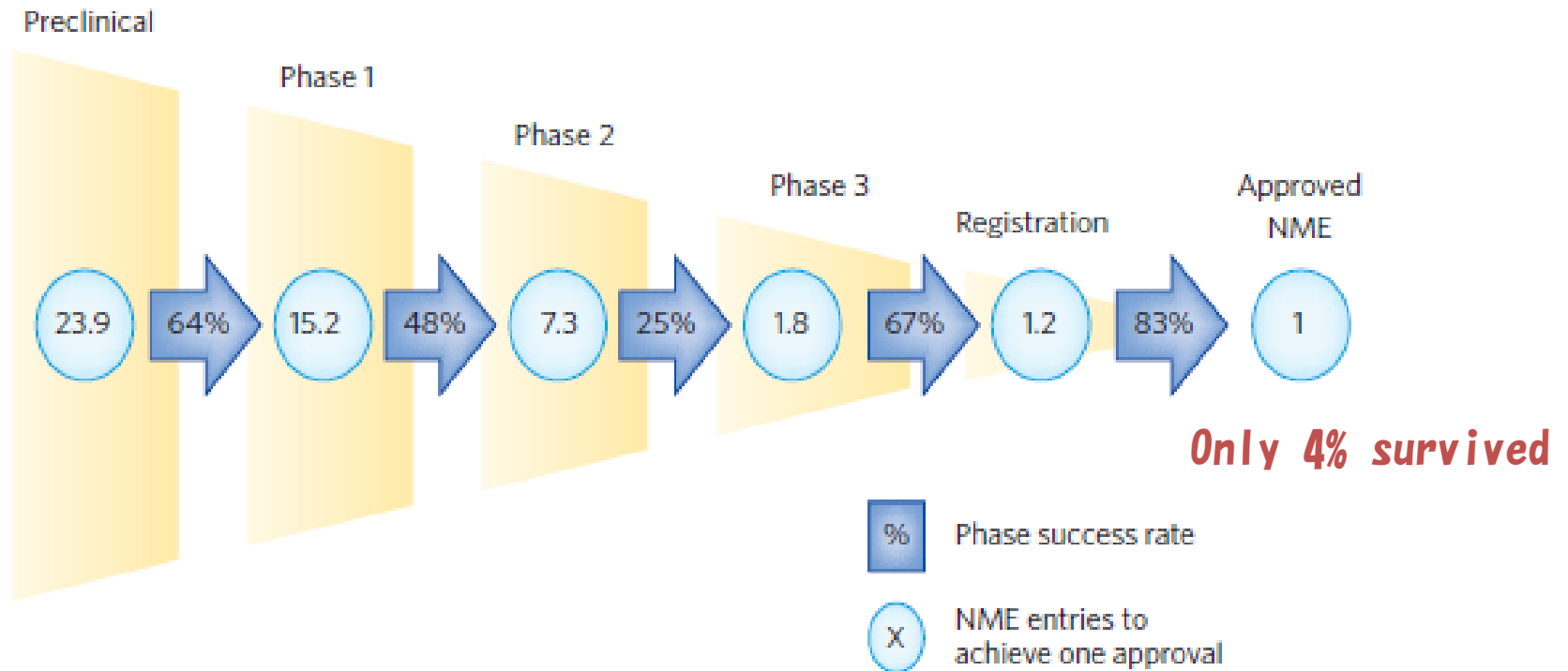
Combined FDA-approved new molecular entities (NMEs) versus R&D spending for the top nine largest pharmaceutical companies.

**NME:** a drug that contains no active moiety that has been approved by the FDA in any other application submitted. NME includes biologicals and vaccines.

*Nature Chem. Biol.* **2011**, 7, 335.

# Introduction 2

Pharmaceutical industry 2005–2009

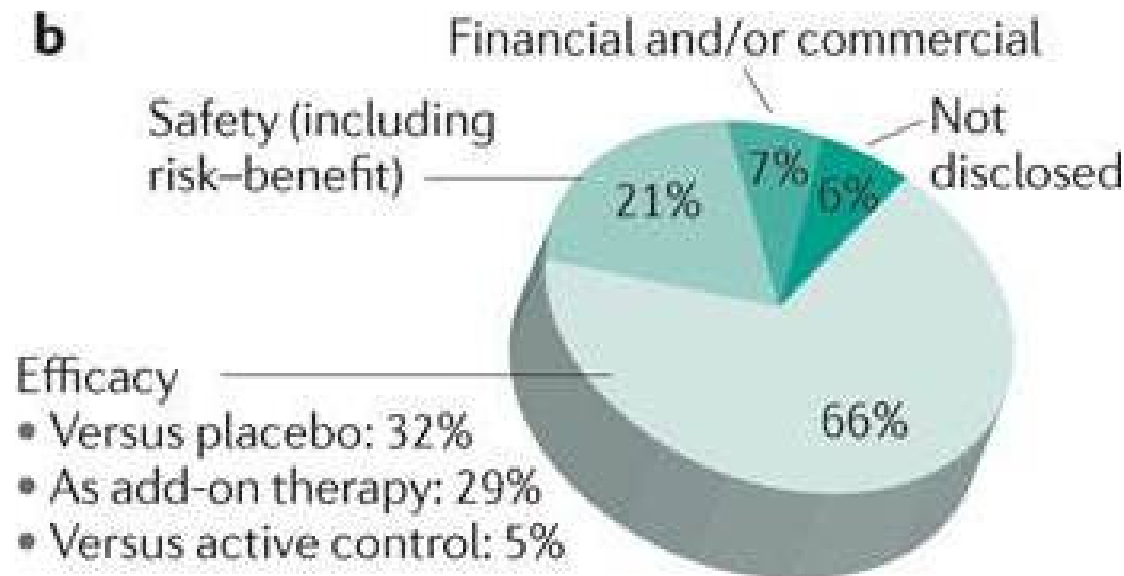


NMEs success rate by phase. Combined R&D survival by development phase for the top fourteen largest pharmaceutical.

Approval data is based on approval of NME by a regulatory authority in a major market (EU, US or Japan).

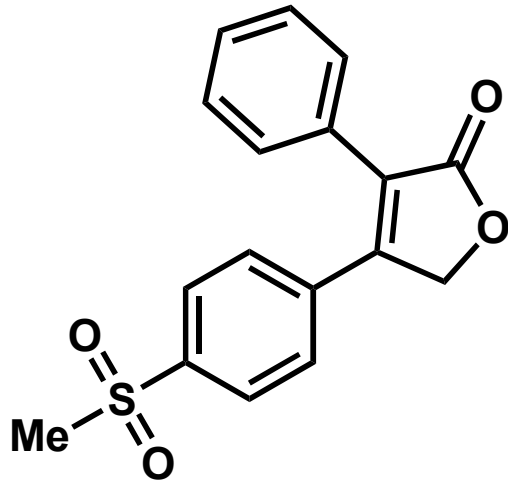
# Introduction 3

## Phase III and submission failures: 2007–2010



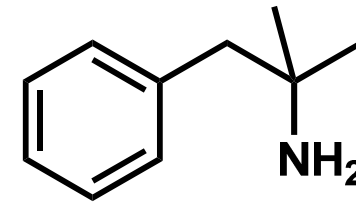
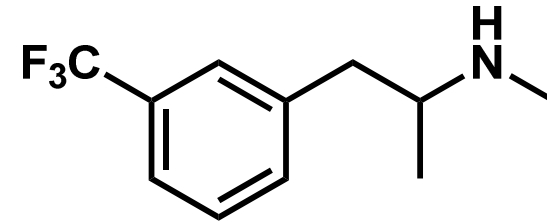
- Next to lack of efficacy, **adverse drug reactions (ADRs)** are the leading cause for attrition in clinical trials of new drugs.
- Some ADRs are caused by modulation of the primary target of a drug, others result from non-specific interactions of reactive metabolites. In many cases, however, ADRs are caused by **unintended activity at off-targets**.

# Introduction 4



VIOXX (refecoxib)

**Selective COX-2 inhibitor**  
cardiovascular event  
(mechanism is still unknown)



fen-phen

**Appetite suppressant**  
One of its metabolites, norfenfluramine activated the 5-HT<sub>2B</sub> receptor, leading to proliferative valvular heart disease.

Cyclooxygenase (COX) has two well-studied isoforms, called COX-1 and COX-2. COX-1 mediates the synthesis of prostaglandins responsible for protection of the stomach lining, while COX-2 mediates the synthesis of prostaglandins responsible for pain and inflammation. By creating "selective" NSAIDs that inhibit COX-2, but not COX-1, the same pain relief as traditional NSAIDs is offered, but with greatly reduced risk of fatal or debilitating peptic ulcers.

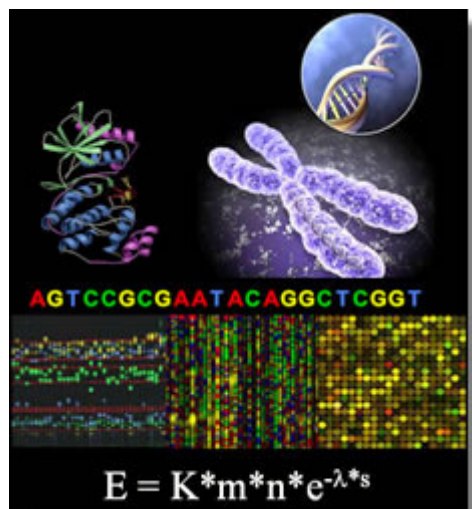
# Today's topic

## ***Predicting unintended off-target toxicity by informatics methods !***

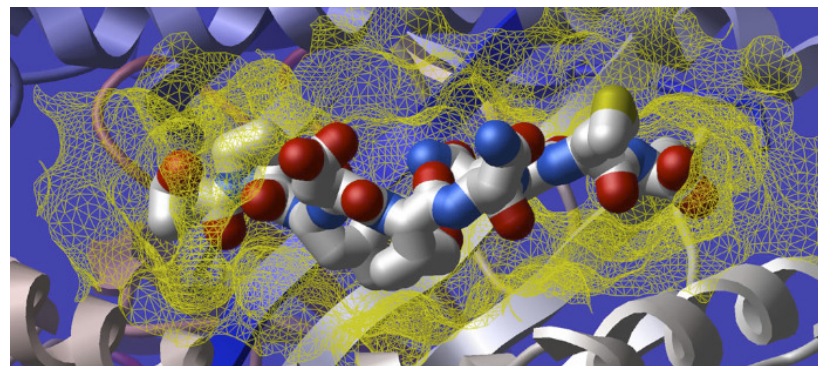
**Informatics:** the science of information

It studies the representation, processing, and communication of information in natural and artificial systems.

**Bioinformatics (DNA/RNA, protein, etc.)**

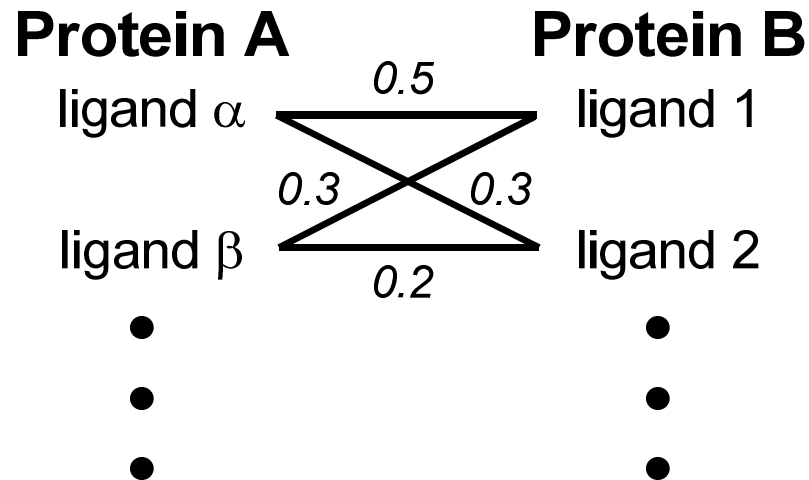


**Chemoinformatics (ligand, drugs)**



# What's SEA ?

**SEA = the Similarity Ensemble Approach**



the total number of  
**Protein: 246**  
**Ligand: 65,241**  
**(Pair:  $5.07 \times 10^9 \doteq 65,241^2$ )**

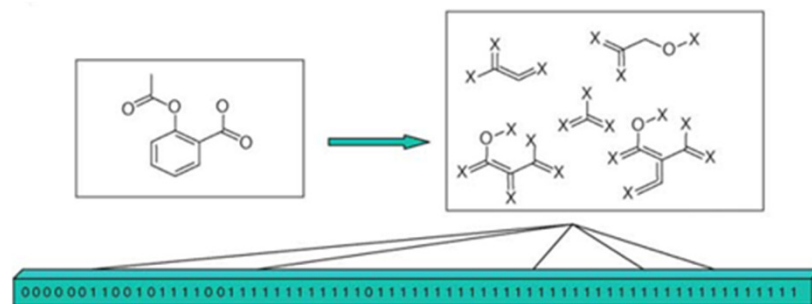
the total similarity score between protein A (ligand  $\alpha$  &  $\beta$ ) & B (ligand 1 & 2)  
 $= 0.2 + 2 \times 0.3 + 0.5 = 1.3$  (the raw similarity score)

1. Score the similarity of all ligands pairs between two proteins with Tanimoto efficient ( $T_c$ ) (0 (complete dissimilarity) <  $T_c$  < 1 (identity))
2. Sum up all  $T_c$  score (the raw similarity score)
3. Calculate expectation values and evaluate the similarity of two proteins

# Tanimoto coefficient (Tc) 1

## Basic idea

$$T_s(X, Y) = \frac{\sum_i (X_i \wedge Y_i)}{\sum_i (X_i \vee Y_i)}$$

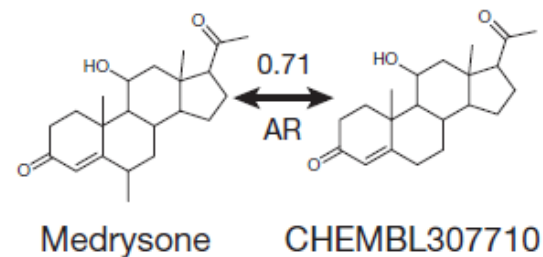
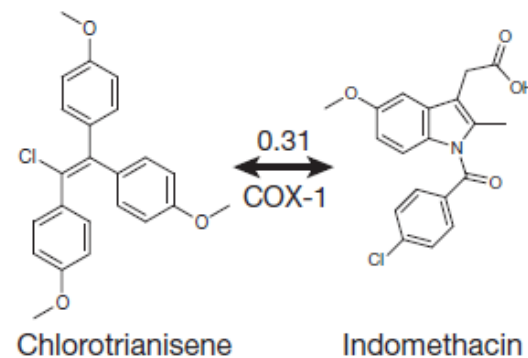
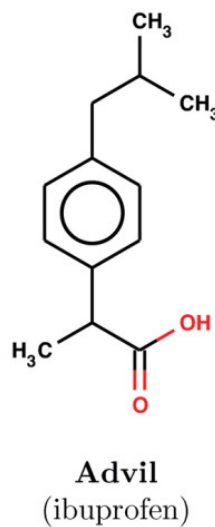
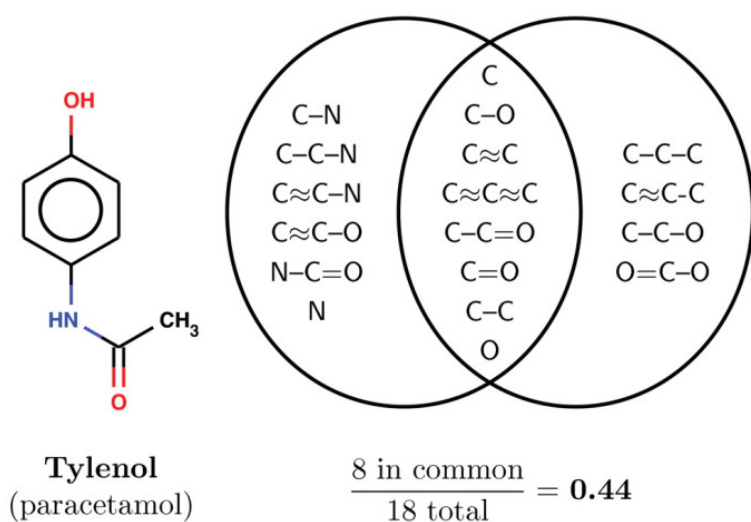


Similarity:

$$TC = \frac{\text{bits in common}}{\text{all bits}}$$

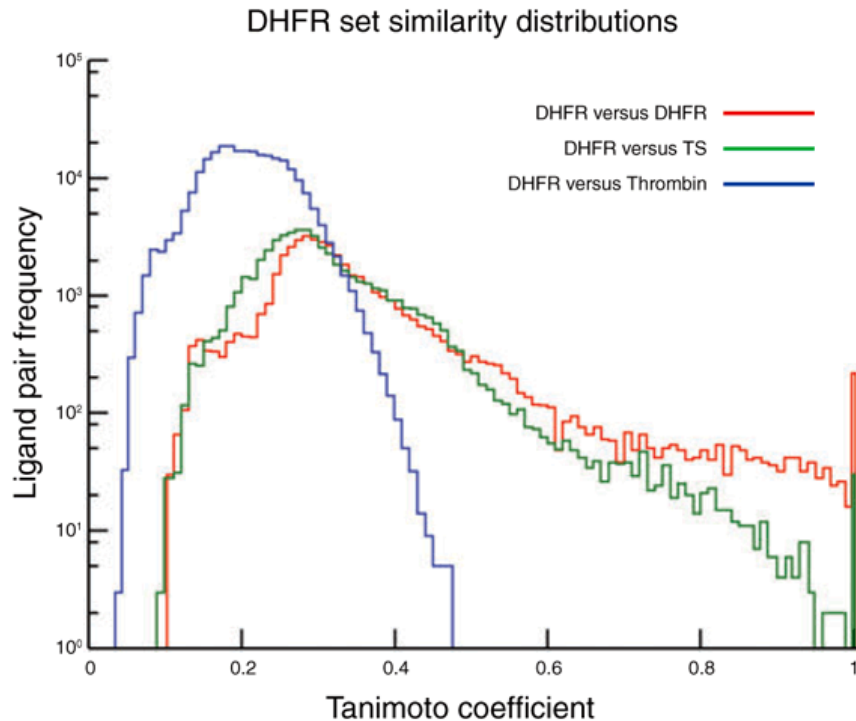
Tanimoto Coefficient TC = 1 for identical compounds

## Examples





# Tanimoto coefficient (Tc) 2



The enzymes **thymidylate synthase (TS)** and **dihydrofolate reductase (DHFR)** both **recognize folic acid derivatives** and are **inhibited by antifolate drugs**. Despite this, two enzymes have **no substantial sequence identity** and are **structurally unrelated**.  
Thrombin: serine protease protein  
(unrelated to DHFR)

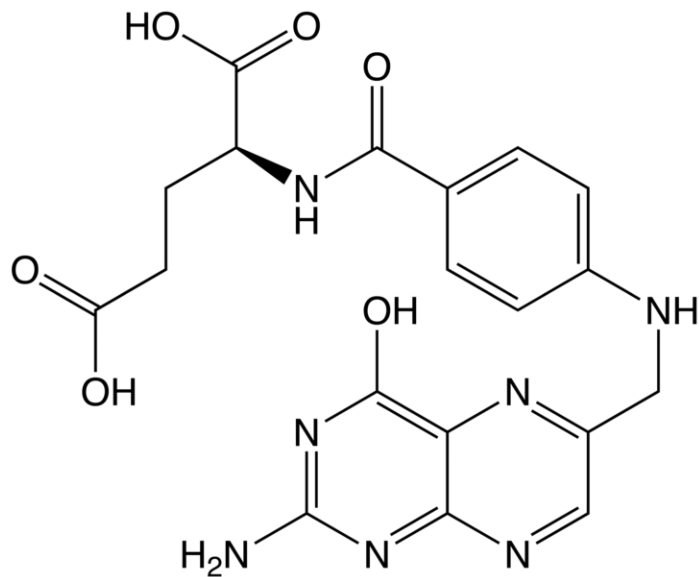
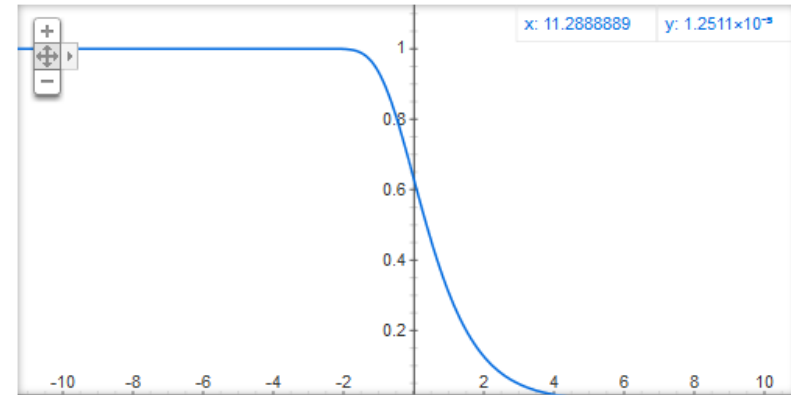
- **For most ligand pairs the Tc was low, in the 0.2 to 0.3 range (insubstantial similarity).**  
This was true even when comparing a set to itself.  
DHFR versus DHFR (red), 80.4% (0.1 to 0.4 range), 4.7% (0.6–1.0 range) and 0.5% (1.0)
- **The raw similarity score:** the sum of ligand pair Tcs over all pairs with  $Tc \geq 0.57$

# Quantifying SEA

$P = 1 - \exp(Kmne^{-\lambda x})$  where  $\chi$  is similarity score  
 $\chi$  is the raw similarity score (?)

The smaller, the more similar

グラフ:  $1 - \exp(-\exp(-x))$

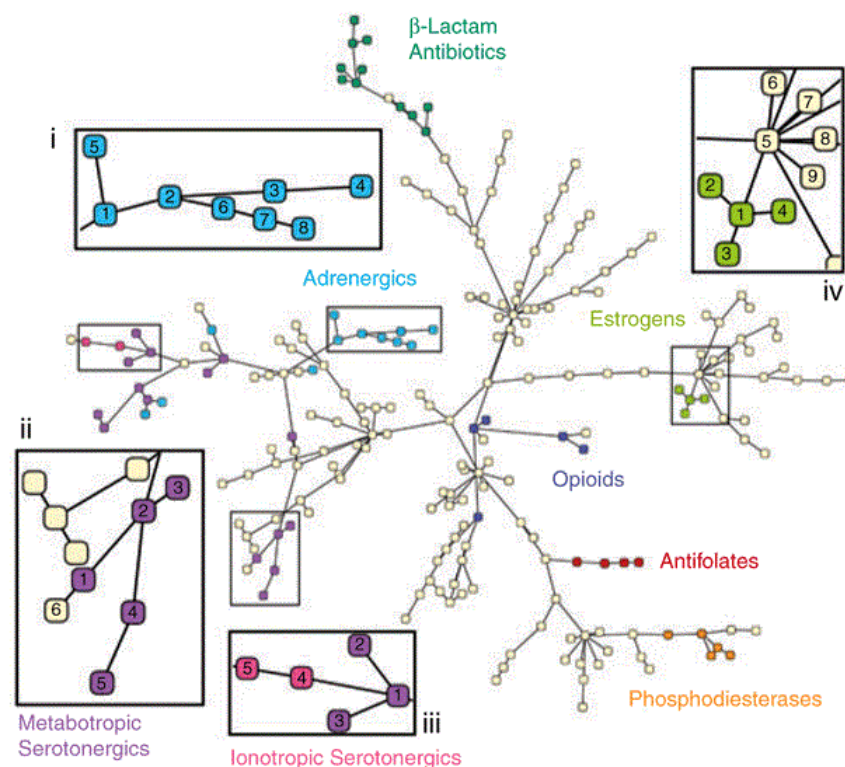


**Folic acid**

Rank	Activity class	E-value	Example molecule
1	DHFR inhibitor	$7.07 \times 10^{-182}$	
2	Glycinamide ribonucleotide formyltransferase inhibitor	$3.97 \times 10^{-100}$	
3	Folypolyglutamate synthetase inhibitor	$4.59 \times 10^{-62}$	
4	TS inhibitor	$1.11 \times 10^{-61}$	

# Patterns of similarity

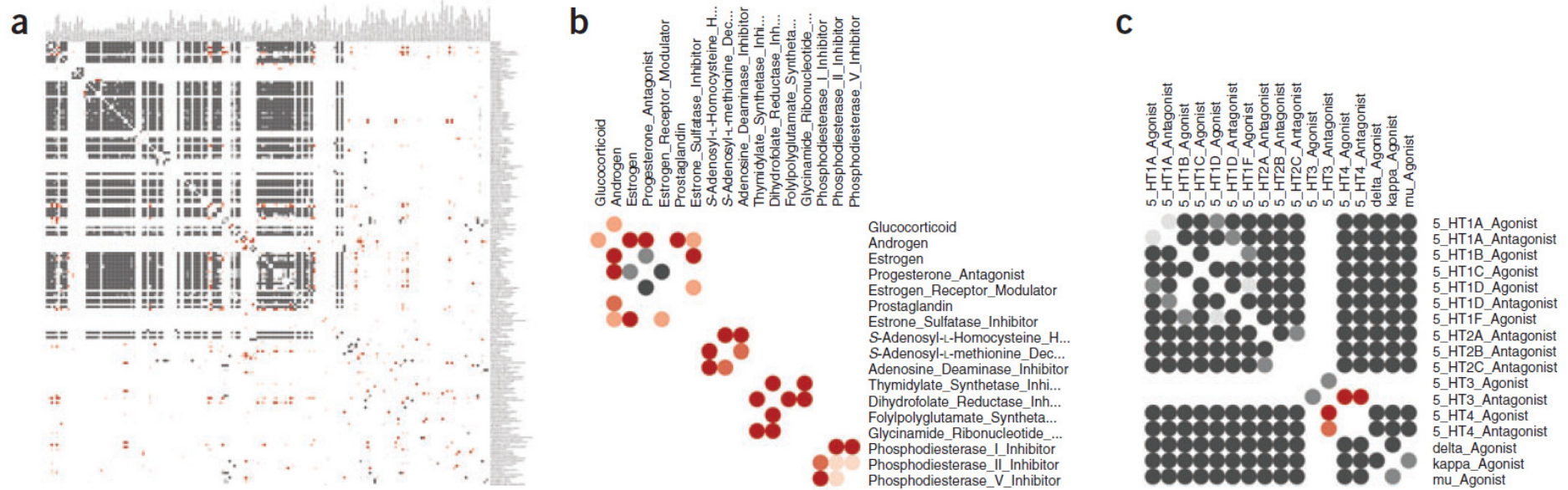
Query	Rank	Size	Similar activity classes	E-value	Tc 1.0	Max Tc
AMPA receptor Antagonist	1	569	AMPA receptor antagonist	$2.45 \times 10^{-219}$	577	1.00
	2	75	Kainic acid receptor antagonist	$5.28 \times 10^{-80}$	74	1.00
	3	1485	NMDA receptor antagonist	$3.08 \times 10^{-63}$	181	1.00
	4	22	Anaphylatoxin receptor antagonist	$3.81 \times 10^{-4}$	0	0.70
	5	130	$\mu$ agonist	$1.69 \times 10^{-3}$	0	0.83
	6	99	Ribonucleotide reductase inhibitor	$1.00 \times 10^{-1}$	0	0.73
Carbacephem	1	98	Carbacephem	0 <sup>a</sup>	106	1.00
	2	1614	Cephalosporin	$1.11 \times 10^{-222}$	14	1.00
	3	35	Isocephem	$2.30 \times 10^{-17}$	0	0.64
	4	257	Penem	$2.43 \times 10^{-4}$	0	0.68
	5	13	Oxacephem	$8.38 \times 10^{-3}$	0	0.69
	6	39	Lactam ( $\beta$ ) antibiotic	$2.62 \times 10^{-2}$	0	0.62
	7	223	Lactamase ( $\beta$ ) inhibitor	$6.58 \times 10^{-1}$	1	1.00
	8	116	Monocyclic $\beta$ -lactam	$3.18 \times 10^2$	0	0.61
Androgen	1	50	Androgen	0 <sup>a</sup>	138	1.00
	2	577	Aromatase inhibitor	$6.87 \times 10^{-307}$	0	0.88
	3	43	Antiglucocorticoid	$2.30 \times 10^{-102}$	0	0.89
	4	6	Cytochrome P450 oxidase inhibitor	$4.01 \times 10^{-93}$	0	0.92
	5	179	Estrogen	$9.97 \times 10^{-89}$	0	0.91
	6	86	Antiestrogen	$2.18 \times 10^{-76}$	0	0.84
	7	936	Steroid ( $5\alpha$ ) reductase inhibitor	$1.58 \times 10^{-72}$	0	0.80
	8	103	Antiandrogen	$1.14 \times 10^{-70}$	0	0.99
	9	86	$17\alpha$ -hydroxylase/C17-20 lyase inhibitor	$7.88 \times 10^{-66}$	0	0.76
	10	164	Progesterone antagonist	$3.26 \times 10^{-44}$	0	0.89
	11	62	Prostaglandin	$1.93 \times 10^{-38}$	0	0.75



- On average, any given receptor was similar to only 5.8 other receptors with an expectation value  $<10^{-10}$ .

- Clusters of biologically related targets may be observed, as no explicit biological information, only ligand information, is used to calculate the cross-target similarity.

# Comparison of sequence and ligand-based protein similarity



**Red** : pairs with strong ligand-set similarity but weaker sequence similarity.

**Dark gray** : pairs with strong sequence similarity but comparatively lower ligand-set similarity

**White** : pairs where pharmacological and sequence similarity approaches agree

(either positively or negatively)

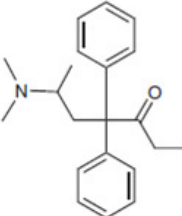
**(a)** overall difference heat map

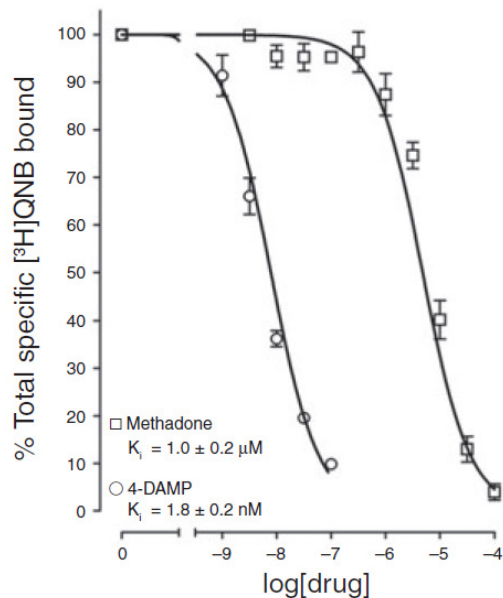
**(b)** folate-recognition enzymes and adenosine-binding enzymes

**(c)** GPCRs, ion channels and nuclear hormone receptors

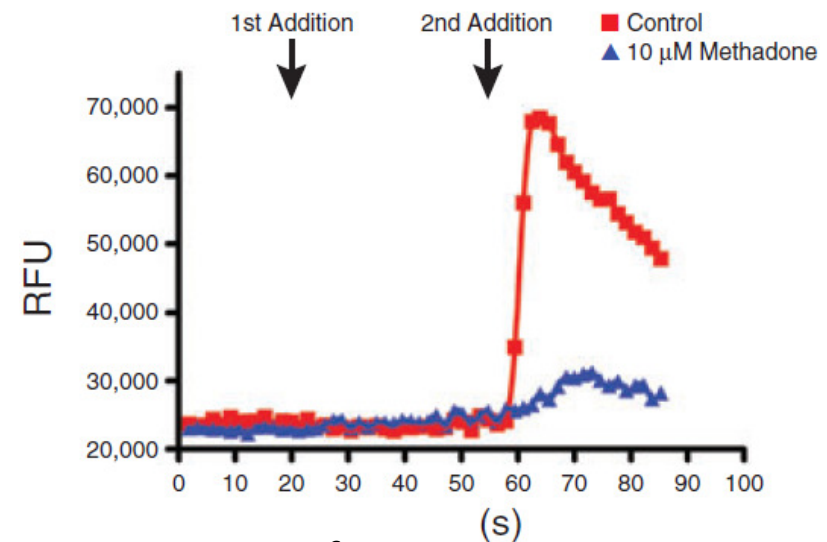
# Predicting and testing drug promiscuity

Methadone is known to have dual specificity for  $\mu$ -opioid receptors and NMDA.

Query	Rank	Activity class	E-value	Max Tc
	1	Antimuscarinic	$4.45 \times 10^{-50}$	0.77
	2	Muscarinic M3 antagonist	$1.22 \times 10^{-11}$	0.67
	3	<u>Opioid agonist</u>	1.84	0.61
	4	<u>NMDA receptor antagonist</u>	9.04	0.67
	5	Muscarinic (M1) agonist	61.9	0.60
	6	Cyclooxygenase inhibitor	12.1	0.61



Antagonism of M3 muscarinic receptors by the  $\mu$ -opioid agonist methadone in a direct binding assay



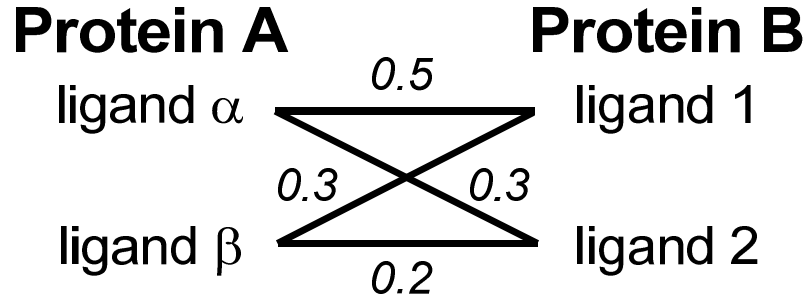
Antagonism of M3 muscarinic receptors by functional assay

*Nature Biotechnology* 2007, 2, 197.



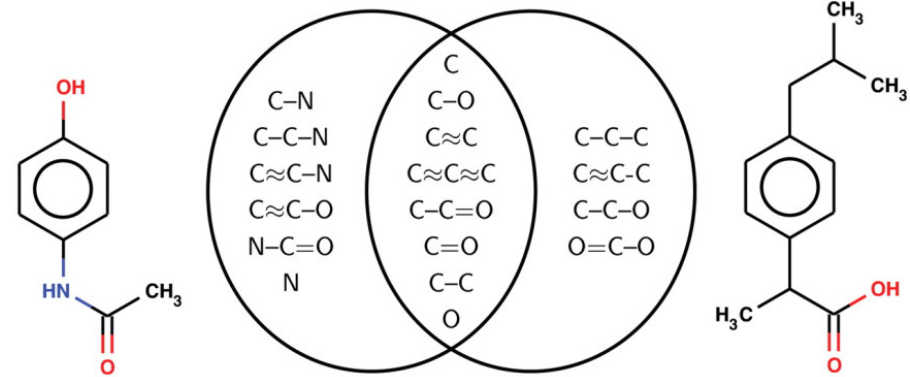
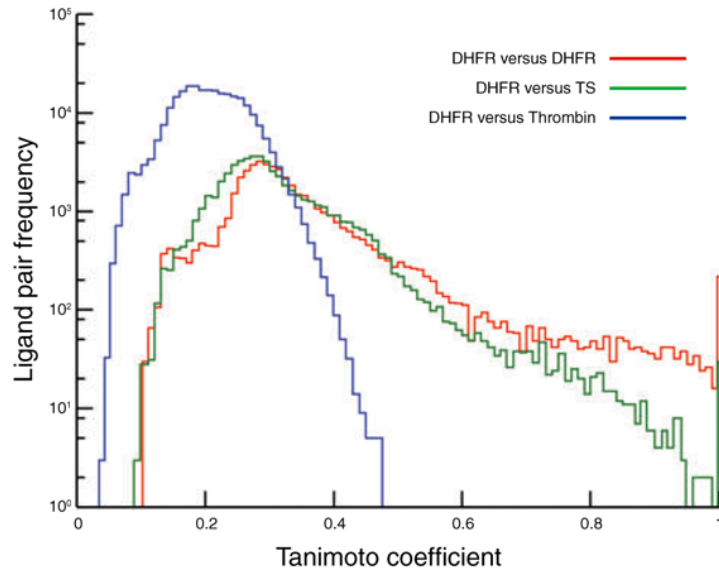
# Small Summary 1

## SEA = the Similarity Ensemble Approach



- 
- 
- 
- 
- 

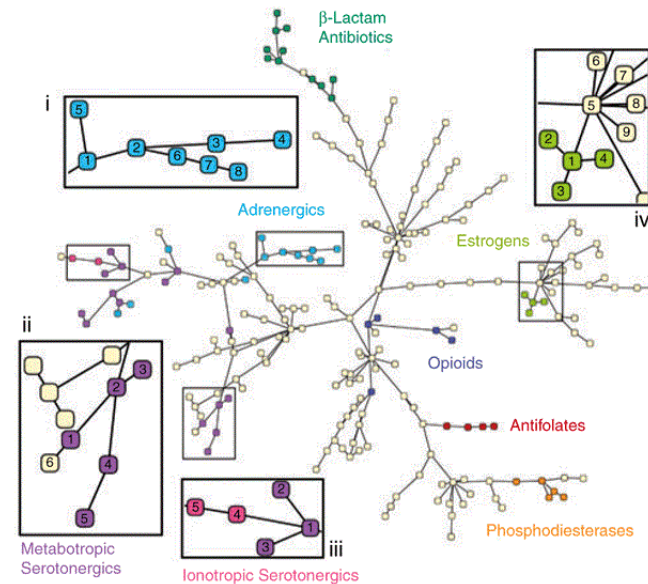
DHFR set similarity distributions



Tylenol  
(paracetamol)

$$\frac{8 \text{ in common}}{18 \text{ total}} = 0.44$$

Advil  
(ibuprofen)



# Predicting new molecular targets for known drugs

## Drugs

878 FDA-approved drugs + 2,787 investigational drugs = 3,665 total

## Data base

same as the previous study (*Nature Biotechnology* 2007, 2, 197.)

246 proteins  
(65,241 ligands)

## Total

901,590 protein-drug pairs (3,665 x 246)

## Result

901,590 protein-drug pairs

↓ SEA

6,928 pairs

with  $E$ -values better than  $1 \times 10^{-10}$



3,832 pairs  
unknown, biologically interesting

sampling ↑

184 pairs

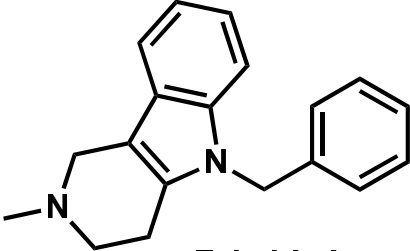
40 pairs  
known

30 pairs  
tested



23 pairs  
 $K_i < 15 \mu\text{M}$

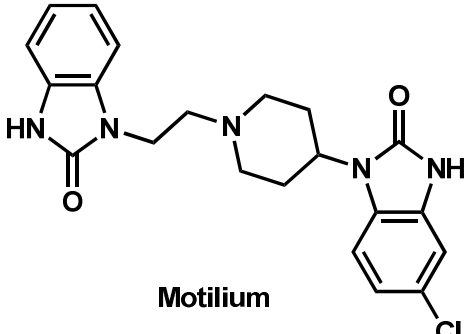
# New targets as primary sites of action

Drug	Pharmacological action	E-value	Predicted target	K <sub>i</sub> (nM)
 Fabahistin	Antihistamine (H <sub>1</sub> receptor)	5.7 x 10 <sup>-57</sup>	5-HT <sub>5A</sub> antagonist (serotonergic receptor)	130

- Used since the 1950s as an antihistamine, Fabahistin is now being investigated for Alzheimer's disease.
- **Fabahistin binds predicted new, off-target(5-HT<sub>5A</sub> receptor) much stronger than its canonical H<sub>1</sub> receptor target.**
- Its activity against 5-HT<sub>5A</sub> and related serotonergic receptors may have implications for Fabahistin's role as an Alzheimer's disease therapeutic.

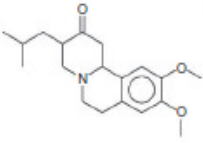
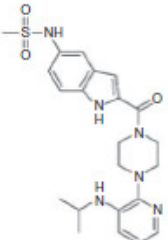
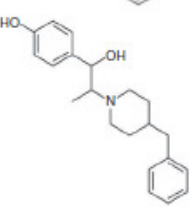
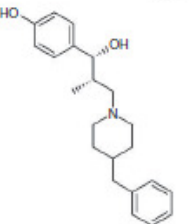


# Off-targets as side-effect mediators

Drug	Pharmacological action	E-value	Predicted target	K <sub>i</sub> (nM)
 <p>Motilium</p>	Antiemetic; peristaltic stimulant (dopamine D <sub>1/2</sub> receptor)	54.8 x 10 <sup>-11</sup>	α <sub>1</sub> adrenergic blocker	α <sub>1A</sub> , 71; α <sub>1B</sub> , 530; α <sub>1D</sub> , 710

- Motilium achieves peak plasma concentrations (C<sub>max</sub>) of 2.8 μM after intravenous administration.
- This formulation was withdrawn owing to **adverse cardiovascular effects**, with the FDA citing cardiac arrest, arrhythmias and sudden death.
- Although Motilium binds the hERG potassium ion channel with a half-maximum inhibitory concentration (IC<sub>50</sub>) of 5 μM, the 71–710nM affinities observed here against α<sub>1A</sub>, α<sub>1B</sub> and α<sub>1D</sub> may also contribute to these cardiovascular effects.

# Drug binding across major protein boundaries

Drug	Canonical target	E-value	Predicted target	K <sub>i</sub> (nM)	
	Xenazine	VMAT2 (transporter)	$1.4 \times 10^{-61}$	$\alpha_2$ adrenergic receptor (GPCR)	$\alpha_{2A}$ , 960; $\alpha_{2C}$ , $1.3 \times 10^3$
				<b>78<sup>th</sup> target</b>	
	Rescriptor	HIV-1 reverse transcriptase (enzyme)	$1.1 \times 10^{-30}$	Histamine H <sub>4</sub> receptor (GPCR)	$5.3 \times 10^3$
				<b>167<sup>th</sup> target</b>	
	Vadilex	NMDAR (ion channel)	$5.1 \times 10^{-13}$ $2.0 \times 10^{-4}$	$\mu$ -opioid receptor (GPCR) 5-HTT; serotonin transporter (transporter)	$1.4 \times 10^3$ 77
	RO-25-6981	NMDAR (ion channel)	$1.5 \times 10^{-8}$ $1.9 \times 10^{-6}$ $3.6 \times 10^{-6}$ $9.1 \times 10^{-5}$	5-HTT; serotonin transporter (transporter) Dopamine D <sub>4</sub> receptor (GPCR) NET; noradrenaline transporter (transporter) $\kappa$ -opioid receptor (GPCR)	$1.4 \times 10^3$ 120 $1.3 \times 10^3$ $3.1 \times 10^3$

- The protein target with the highest sequence similarity to any of a drug's known targets is rarely predicted by the SEA approach.
- Rather, the target predicted by ligand similarity is typically well down in the sequence-similarity ranking.

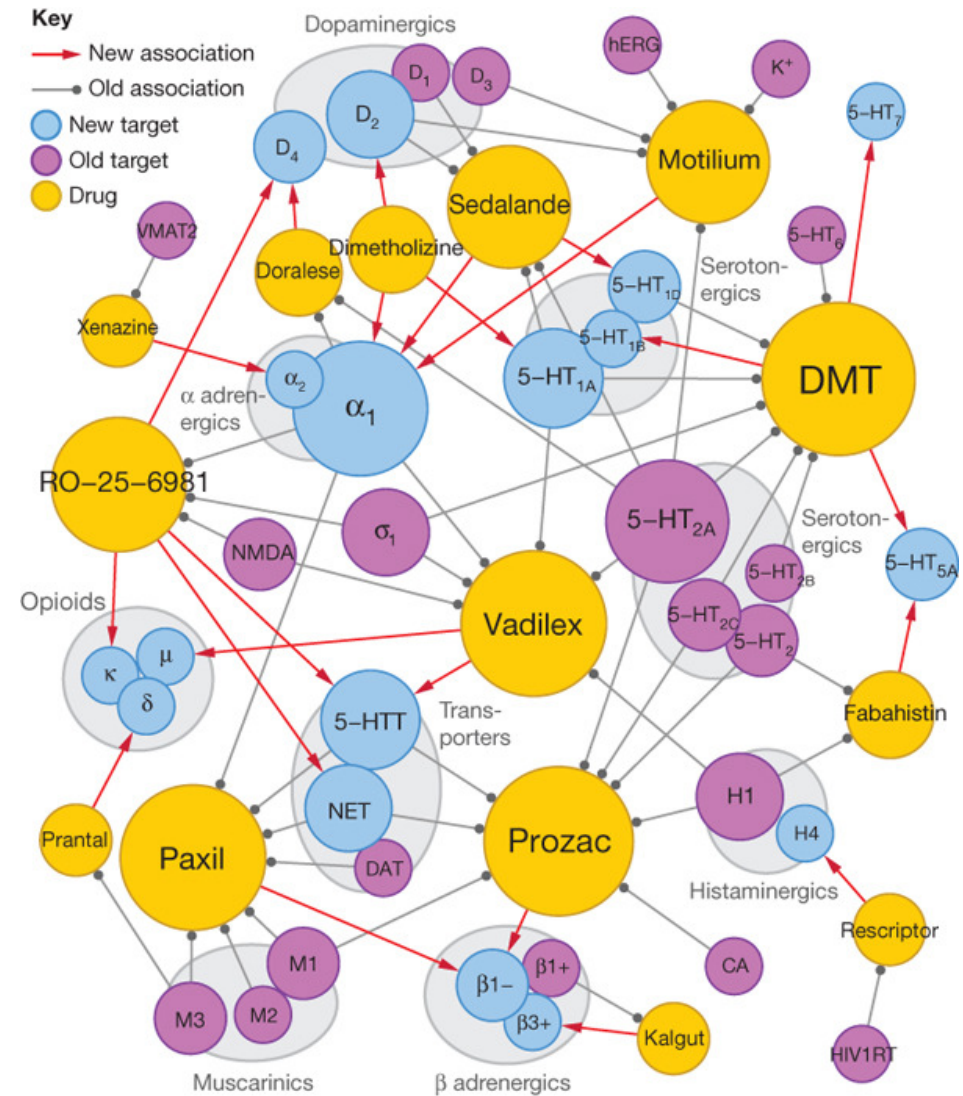
*Nature* **2009**, 462, 175.

# Small Summary 2

3,665 (FDA)-approved and investigational drugs were compared against 246 targets. 364 additional off-targets for 158 drugs are predicted with  $E$ -values better than  $1 \times 10^{-50}$ , whereas 1,853 new off-targets are predicted with  $E$ -values better than  $1 \times 10^{-10}$ .

This compares to the only 972 off-target activities already annotated in the databases. Prediction includes some interesting new off-targets such as;

- (1) the new targets contribute to the primary activity of the drug
- (2) the new targets may mediate drug side effects
- (3) the new targets are unrelated by sequence, structure and function to the canonical targets.



# **Large-scale prediction and testing of drug activity on side-effect targets 1**

- 1. Calculate *E*-value by SEA methods, predict new drug-off-target and confirm by in vitro experiment  
(similar as the previous study (*Nature* 2009, 462, 175.))**
- 2. Quantify the relationships between protein targets and adverse drug reactions (ADRs) by the use of enrichment factor (EF)  
(different point from the previous study)**
- 3. Create a drug–target–ADR network**

# Large-scale prediction and testing of drug activity on side-effect targets <sup>1</sup>

## activity on side-effect targets 2

### Drugs & Targets

656 FDA-approved drugs listed in ChEMBL  
 x 73 with established association of ADRs, for which assays were available at Novartis  
 = 47,888 total

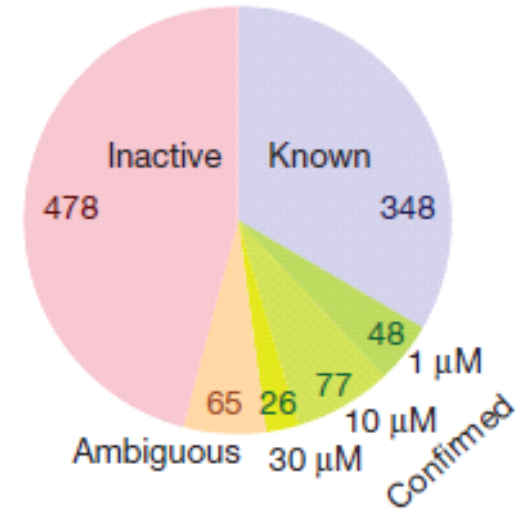
↓ SEA

1,644 pairs with  $E$ -values better than  $1 \times 10^{-4}$

403 pairs listed in ChEMBL ↓

1,241 pairs

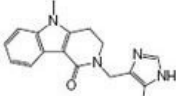
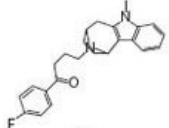
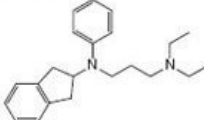
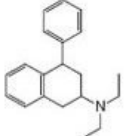
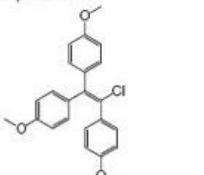
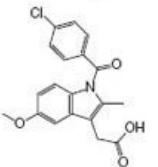
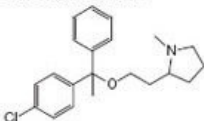
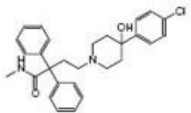
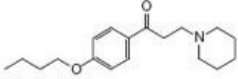
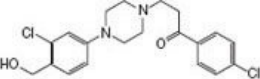
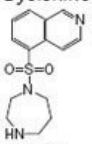
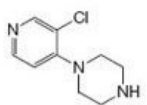
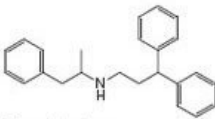
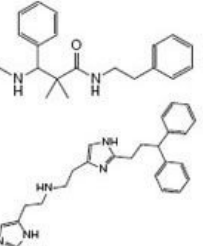
→ 348 pairs listed in other data base



↑ 694 pairs tested at Novartis

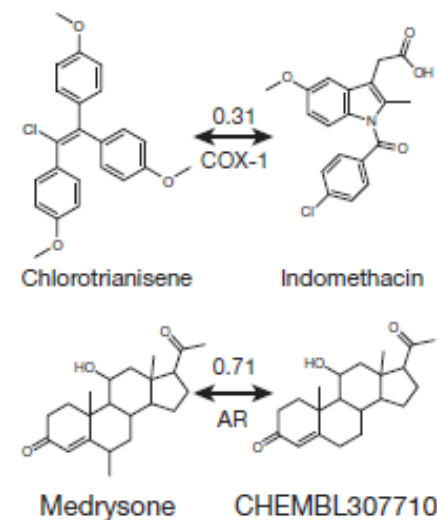
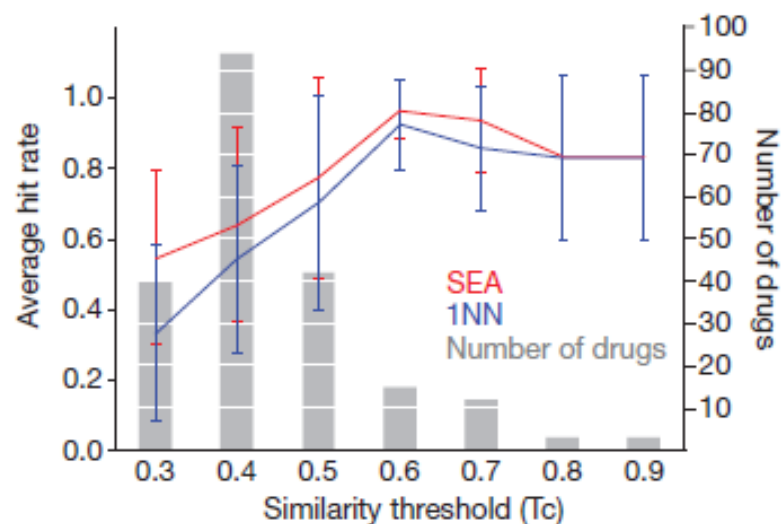
↑ sampling 893 pairs

# New drug-off-target predictions

Drug	Closest chEMBL molecule	Tc value	Target	SEA E value	IC <sub>50</sub> (μM)	Closest known target	BLAST E value
 Alosetron		0.25	HTR2B	$10.6 \times 10^{-17}$	0.02	KCNH7	$3.6 \times 10^2$
 Aprindine		0.38	HRH1	$5.0 \times 10^{-26}$	0.78	SCN5A	$3.3 \times 10^{-1}$
 Chlorotrianisene		0.31	COX-1	$1.9 \times 10^{-17}$	0.16	ESR1	$9.0 \times 10^2$
 Clemastine		0.31	SLC6A4	$1.1 \times 10^{-14}$	0.42	KCNH2	$6.1 \times 10^1$
 Dyclonine		0.36	DRD4	$1.5 \times 10^{-17}$	4.1	SLC6A3	$2.3 \times 10^2$
 Fasudil		0.37	ADRA2A	$1.1 \times 10^{-7}$	4.0	CCR2	$1.5 \times 10^{-9}$
 Prenylamine		0.31	OPRM1	$1.1 \times 10^{-8}$	1.8	CACNA1G	$3.5 \times 10^0$
		0.30	HRH1	$3.2 \times 10^{-66}$	7.9	SCN5A	$3.3 \times 10^{-1}$

# SEA or 1NN

1NN = one-nearest neighbor model



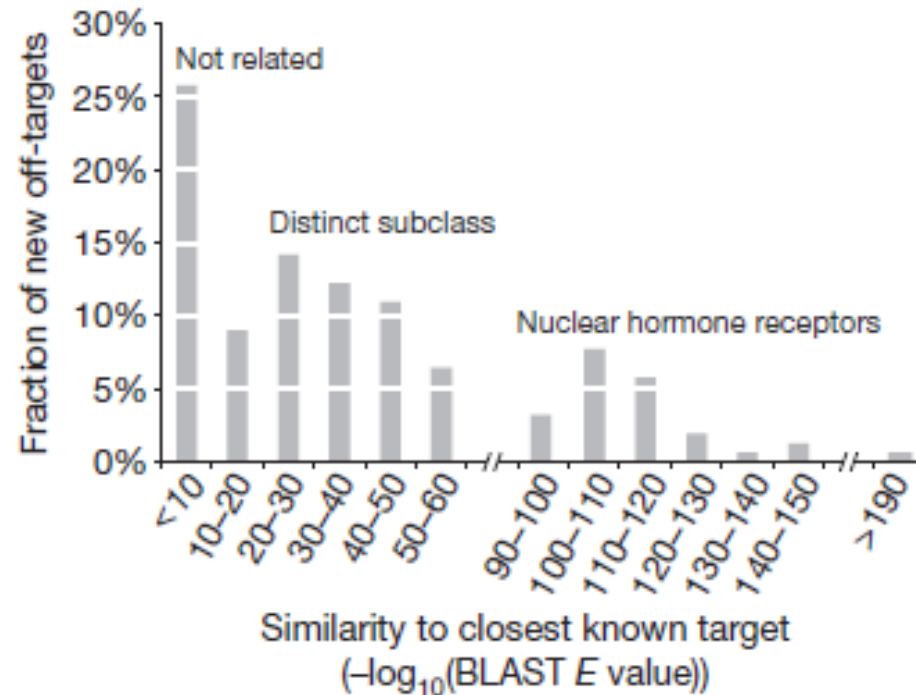
Drug 1

Predicted target	SEA <i>E</i> -value	Tc value of closest ligand	True positive
target A	$2 \times 10^{-15}$	0.31	O
target B	$7 \times 10^{-6}$	0.39	X
----- cut off -----			
target C	$8 \times 10^{-2}$	0.35	X

Model	SEA	1NN
Hit rate	2/3	2/4

**Adjusted hit rate** = (number of true positives+1)/(number of total predictions+1)  
(number of total prediction = number of true positive + number of false positive)

# Are new off-targets predictable ?



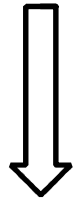
Of the 151 (ref. slide 21) new off-target predictions, 39 (26%) had BLAST E-values greater (worse) than  $10^{-5}$ , suggesting the **previously known targets shared no sequence similarity with the new off-targets.**



# Associating *in vitro* targets with ADRs <sup>2</sup>

## SEA

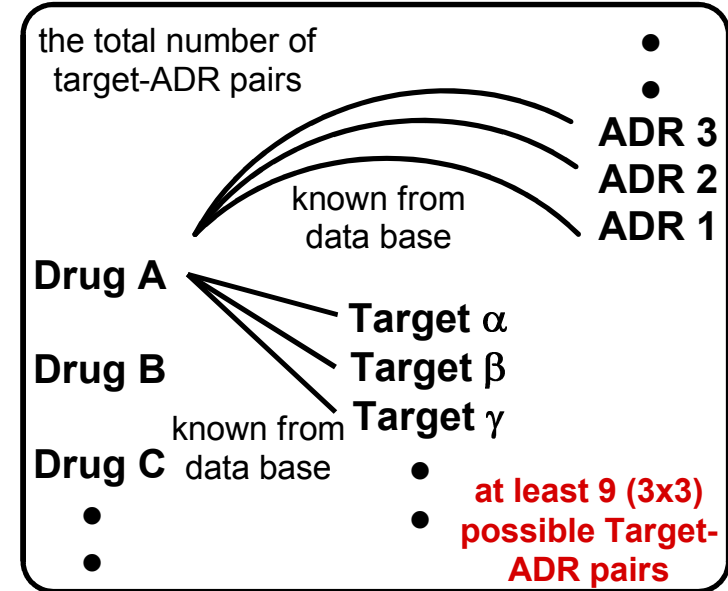
relationships between drugs and targets



to assess the potential clinical relevance of the discovered targets of drugs systematically...

a quantitative score that associated *in vitro* activity with patient ADRs (a score between targets and ADRs)

**solution: Enrichment Factor (EF)**



$$EF = p / (A \times T / P)$$

in which  $p$  is the co-occurrence of target X and ADR Y,  $A$  is the number of times ADR Y was linked to any drug–target pair,  $T$  is the number of times target X was linked with any drug–ADR pair, and  $P$  is the total number of target–ADR pairs.

**45 drugs ( $p$ )** which have the ADR epigastralgia and interact with COX-1

**6,046 drug–target pairs ( $A$ )** where the ADR epigastralgia r was linked with

**2,188 drug–ADR pairs ( $T$ )** where COX-1 was linked with

**681,797 target–ADR pairs overall ( $P$ )**

Thus the pair epigastralgia–COX-1 was enriched **2.3-fold above random**

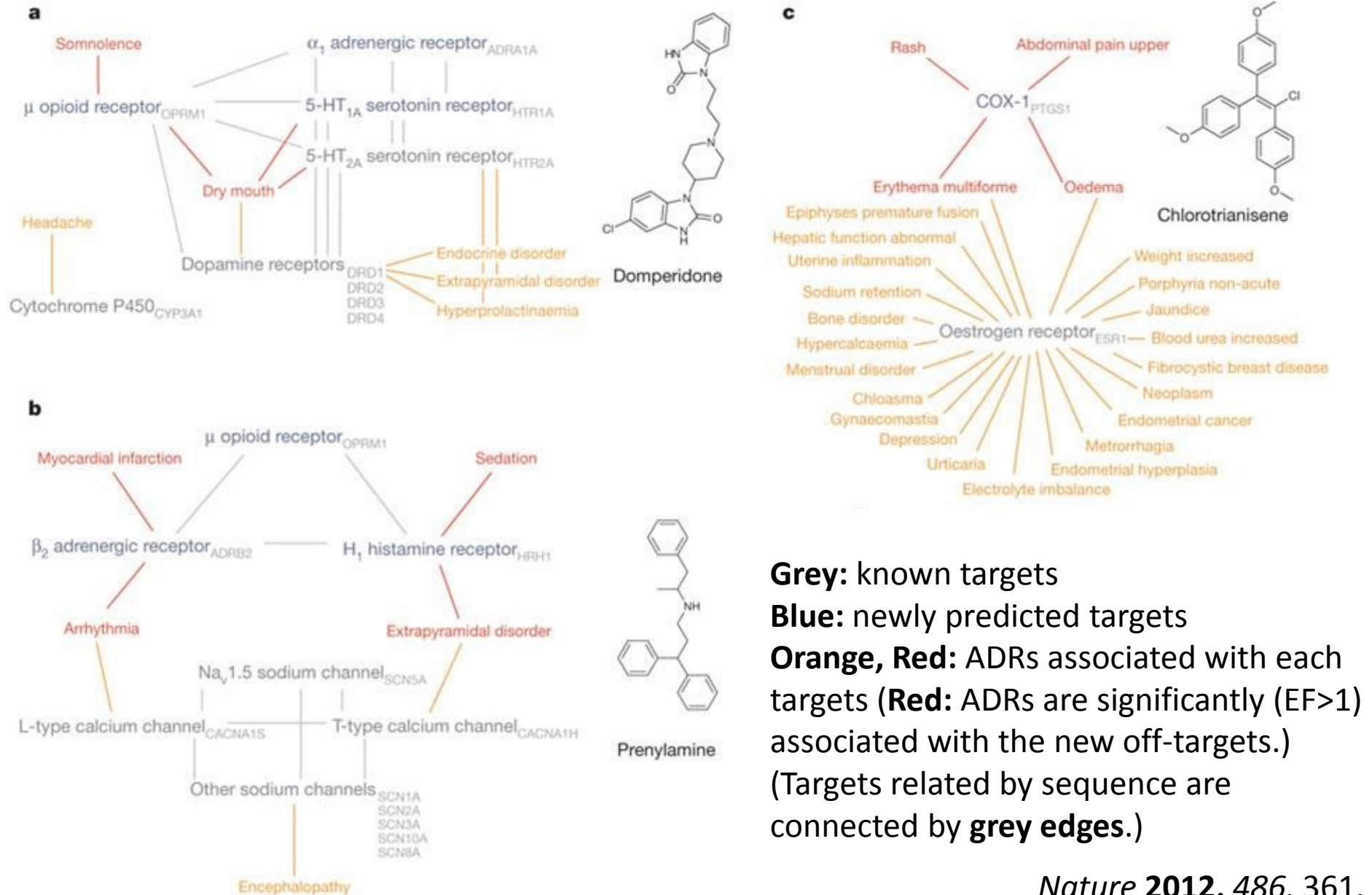
*Nature* 2012, 486, 361.

# New, confirmed targets associated with ADRs<sup>2</sup>

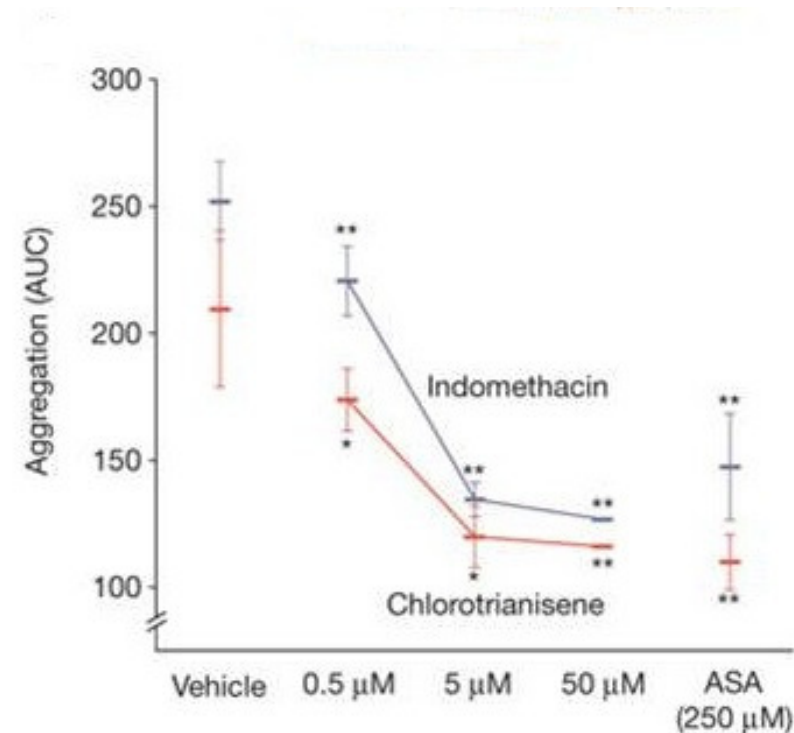
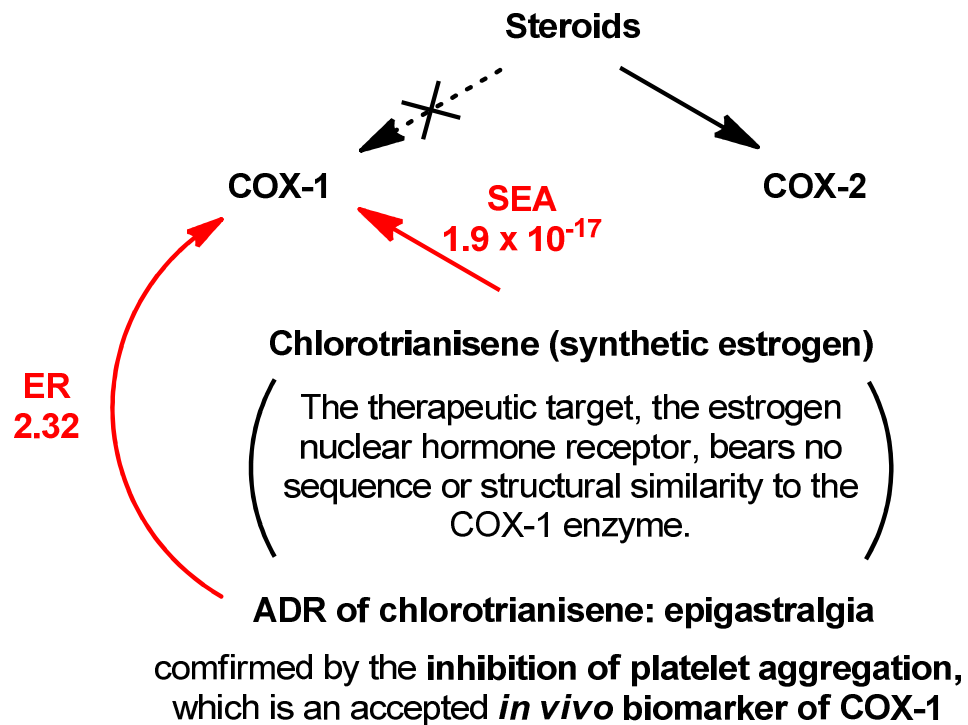
Drug name	Target	Activity ( $\mu$ M) (median)	Adverse event	EF ratio	Alternative target	Comparable drug
Chlorotrianisene	COX-1	0.16	Abdominal pain upper	2.32	None	None
Clemastine	SLC6A4	0.42	Rash	1.79	None	None
Cyclobenzaprine	HRH1	0.02	Sleep disorder	2.15	None	None
			Ataxia	1.73	None	<b>Desipramine</b>
			Somnolence	1.49	None	<b>Aripiprazole</b>
Diphenhydramine	SLC6A3	4.33	Tremor	2.02/1.90	SCN10A	<b>Citalopram</b>
Loxapine	CHRM2	1.12	Tachycardia	2.08/1.97	CHRM1	Sibutramine
Methylprednisolone	PGR	1.30	Depression	3.87/2.49	NR3C1	Flutamide
Prenylamine	HRH1	7.87	Sedation	4.94	None	None
Ranitidine	CHRM2	5.56	Constipation	1.63	None	Haloperidol
Ritodrine	OPRM1	9.18	Hyperhidrosis	3.21	None	<b>Oxycodone</b>

- Of the 151 confirmed (ref. slide 21) new drug–target associations tested at Novartis, 82 were significantly associated with one or more ADR, resulting in a **total of 247 drug–target–ADR links.**
- In **116 cases**, the enrichment factor (EF) of the new drug–target–ADR link was **stronger than that for any previously known target.**

# Drug-Target-ADR network

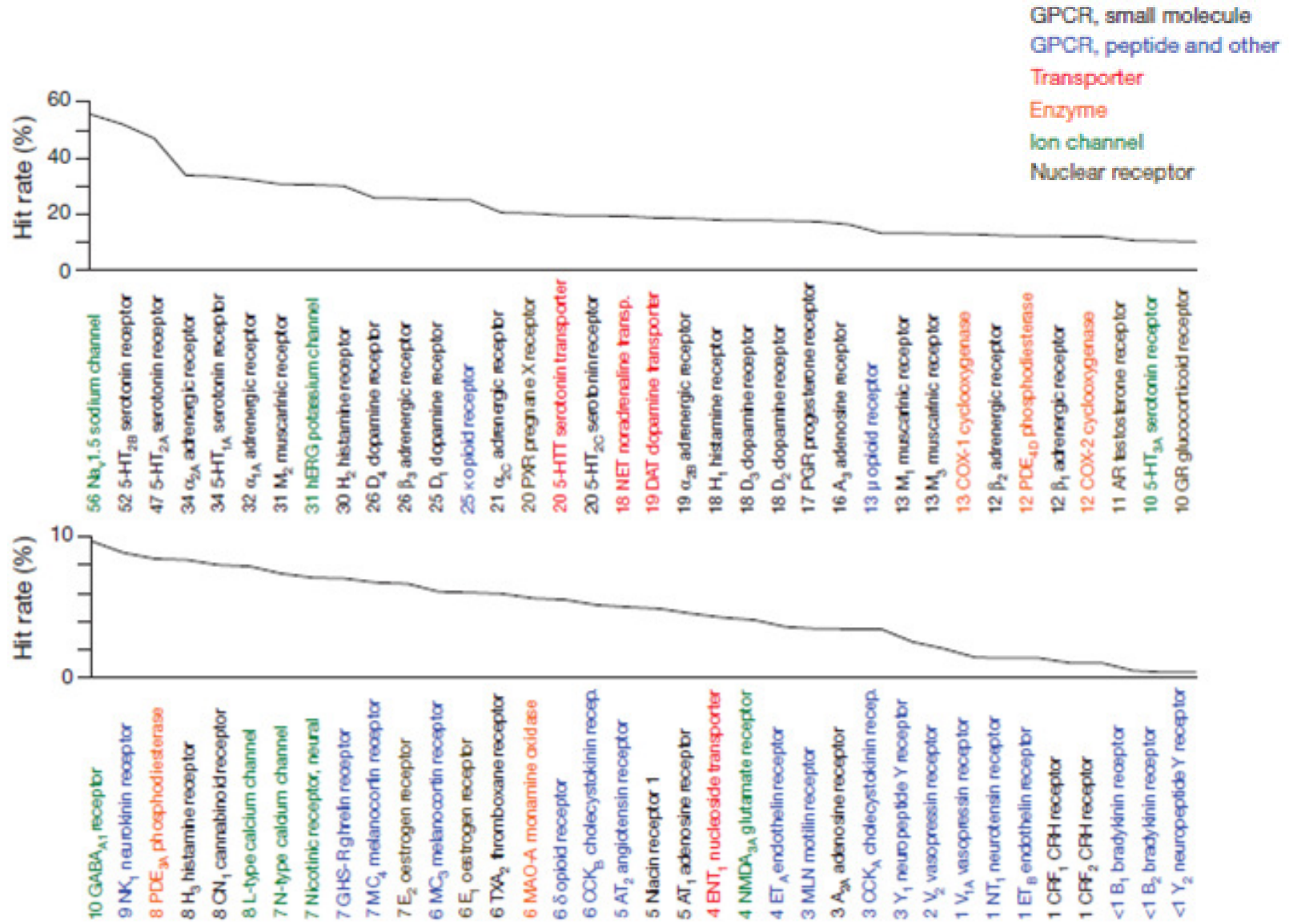


# Demonstration of an association in an accepted *in vivo* biomarker



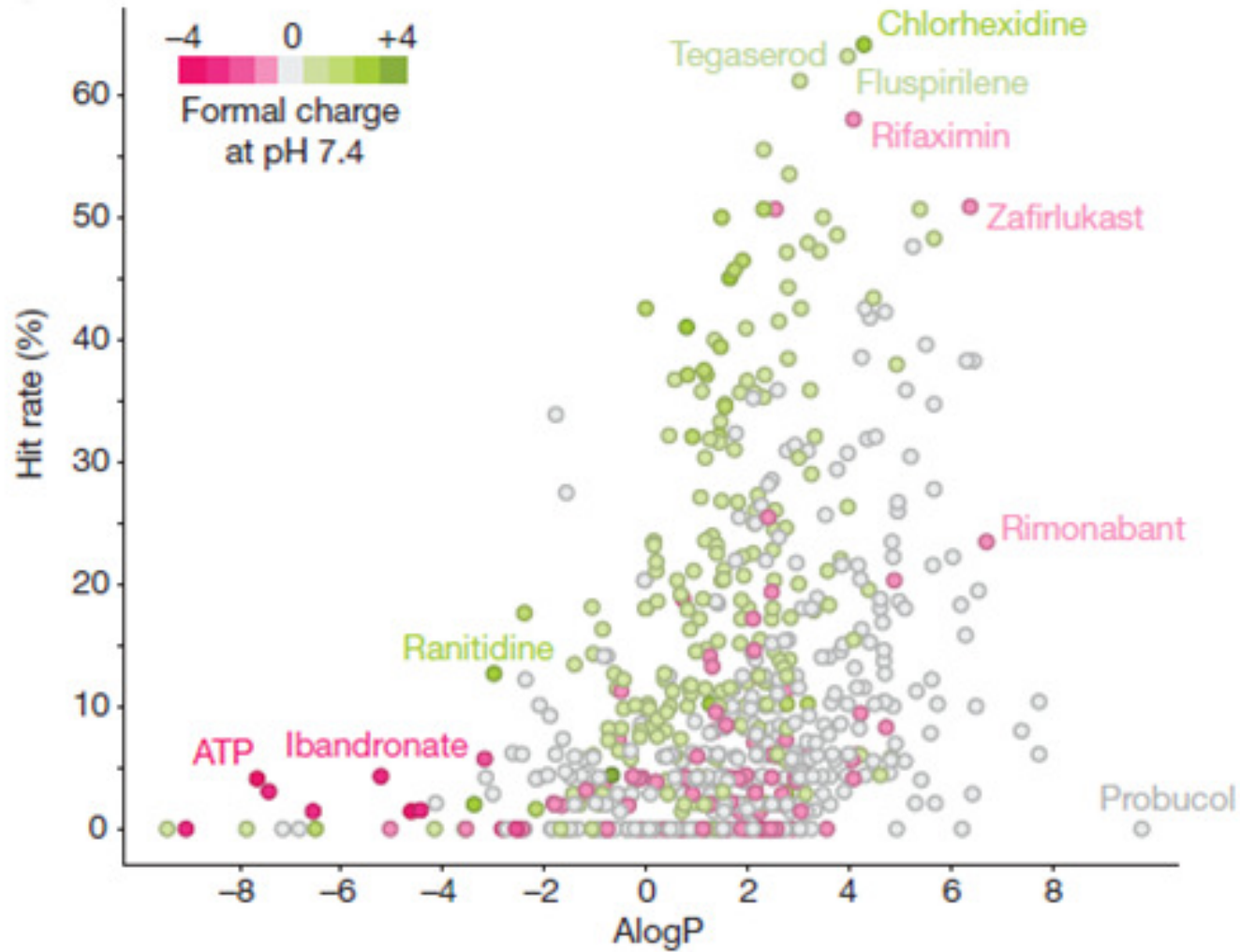
**This was the first example that a synthetic steroid acted on COX-1 enzyme !**

# Target promiscuity





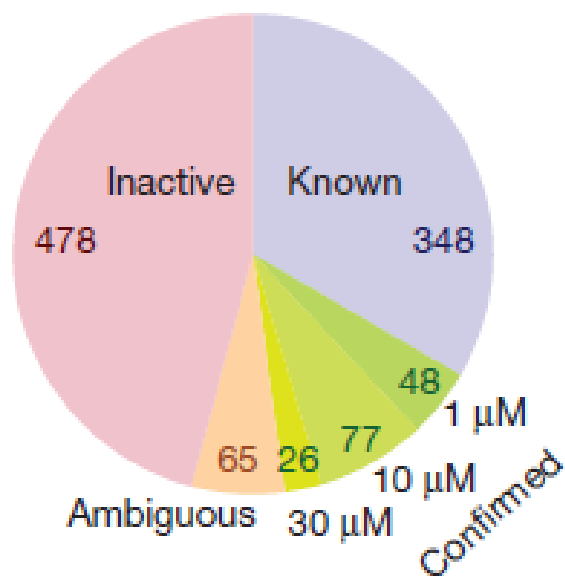
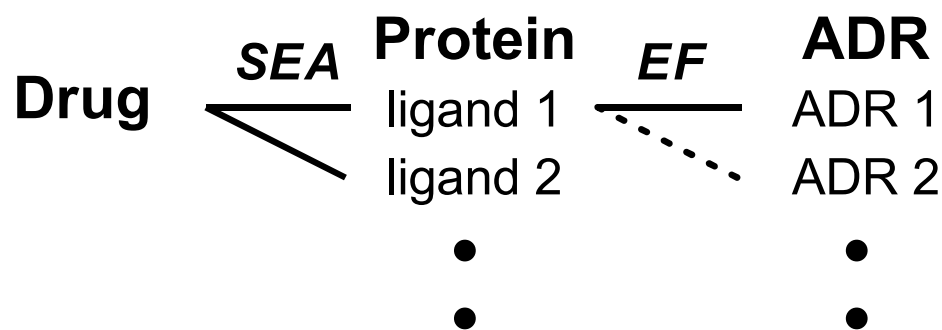
# Drug promiscuity



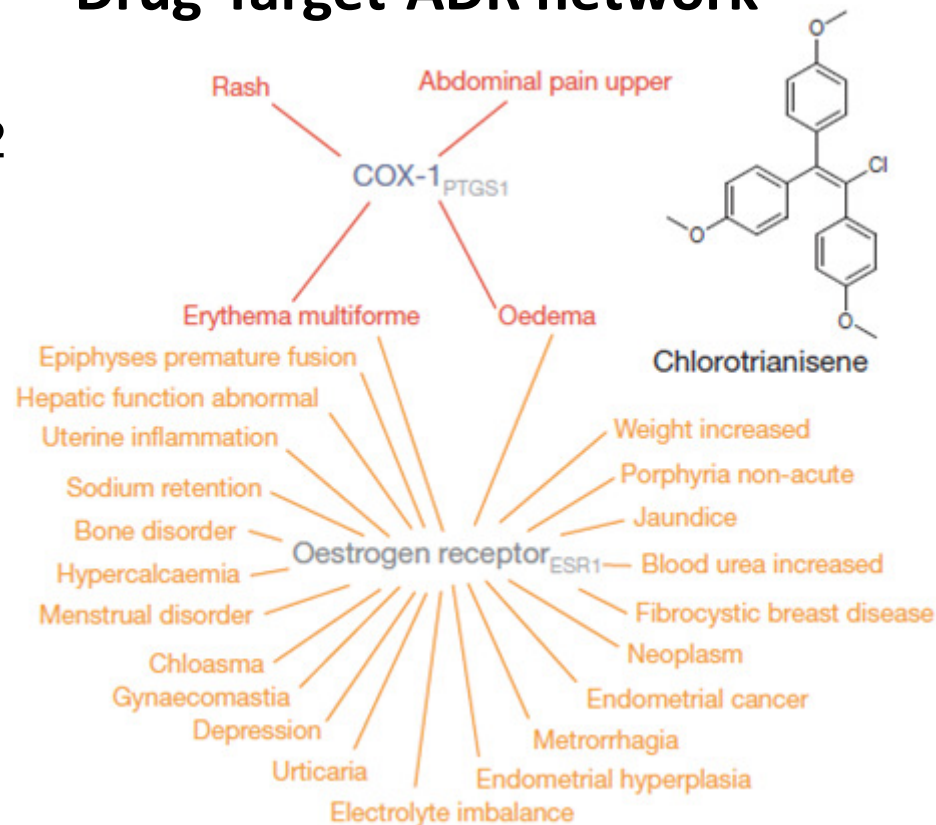
# Summary

SEA = the Similarity Ensemble Approach

EF = Enrichment Factor



## Drug-Target-ADR network



# Comments

- Only some side effects fall into the remit of this approach, which assumes an off-target mechanism.
- Almost 46% of the predicted drug–target associations were disproved, but they were just as often confirmed by experimental ways.
- The method was used automatically at scale, without human intervention. Pragmatically, the ability to calculate drug–target–ADR networks provides a tool to anticipate liabilities among candidate drugs being advanced towards the clinic, or yet earlier, for prioritization of chemotypes in preclinical series.
- The use of Big Data will be dramatically accelerated in almost all fields.



**How to measure “Drug-Likeness” ?  
a new measure taking the place of  
‘the Lipinski’ s Rule of Five’**

*Nature Chem.* 2012, 4, 90.

# Oral drug & Lipinski's rule of five

**Oral drug** is the best, thus the most important way to dose drugs.



*empirical criteria whether a small organic molecule is suitable for a oral drug*

## Lipinski's rule of five

### Lipinski's rule of five

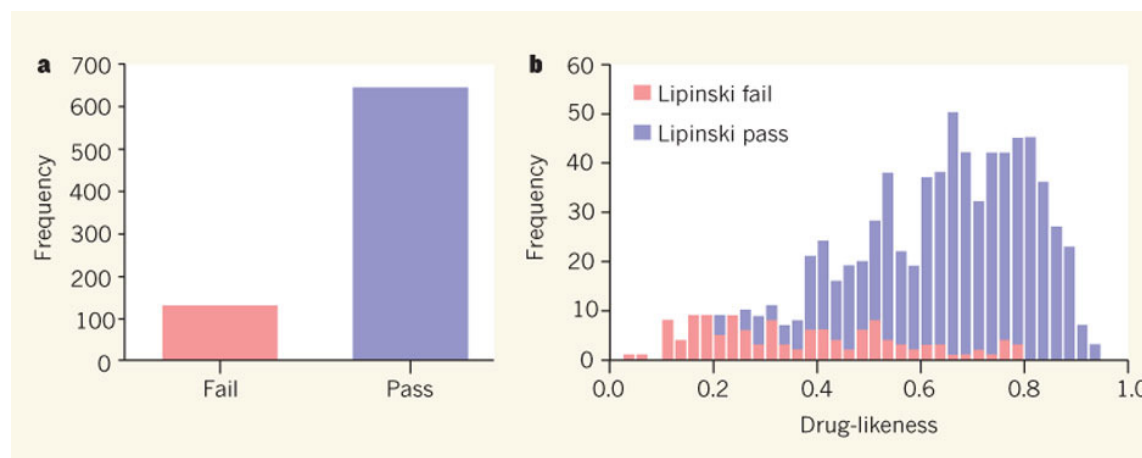
- Its **molecular weight** is **less than 500**.
- The compound's **lipophilicity**, expressed as a quantity known as **logP** (the logarithm of the partition coefficient between water and 1-octanol), is **less than 5**.
- The number of groups in the molecule that can **donate hydrogen atoms to hydrogen bonds** (usually the sum of hydroxyl and amine groups in a drug molecule) is **less than 5**.
- The number of groups that can **accept hydrogen atoms to form hydrogen bonds** (estimated by the sum of oxygen and nitrogen atoms) is **less than 10**.

# The Implementation of Rules

**The rules are only predictive of oral bioavailability** (the absorption by passive diffusion of compounds through cell membranes).

Due to their simplicity, the rules are widely used by medicinal chemists to predict not only the absorption of compounds, as Lipinski originally intended, but also overall **drug-likeness**.

Despite Lipinski's recommendation that the rule be considered as a guideline, in reality **it is used routinely to filter libraries of compounds**. The implementation of rules as filters means that **no discrimination is achieved beyond a qualitative pass or fail**—all compounds that comply with the rules are considered equal, as are all that breach them.



# Quantifying drug-likeness

To quantify compound quality, the concept of desirability was applied to provide a quantitative metric for assessing drug-likeness, which we call **the quantitative estimate of drug-likeness (QED)**. QED values can range from 0 (all properties unfavorable) to 1 (all properties favorable).

$$\text{QED} = \exp\left(\frac{1}{n} \sum_{i=1}^n \ln d_i\right)$$

Desirability takes multiple numerical or categorical parameters measured on different scales and describes each by an individual desirability function. These are then integrated into a single dimensionless score. In the case of compounds, a series of desirability functions  $d$  are derived, each of which corresponds to a different molecular descriptor. Combining the individual desirability functions into the QED is achieved by taking the geometric mean of the individual functions, as shown in equation.

## Weighted QED

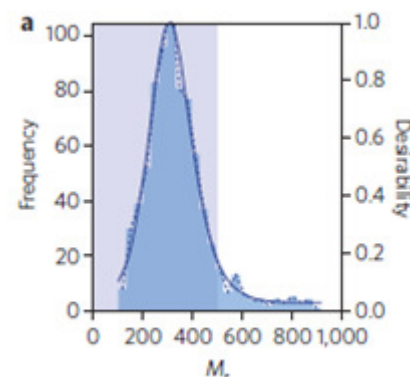
$$\text{QED}_w = \exp\left(\frac{\sum_{i=1}^n w_i \ln d_i}{\sum_{i=1}^n w_i}\right)$$

## Asymmetric Double Sigmoidal (ADS) functions

$$d(x) = a$$

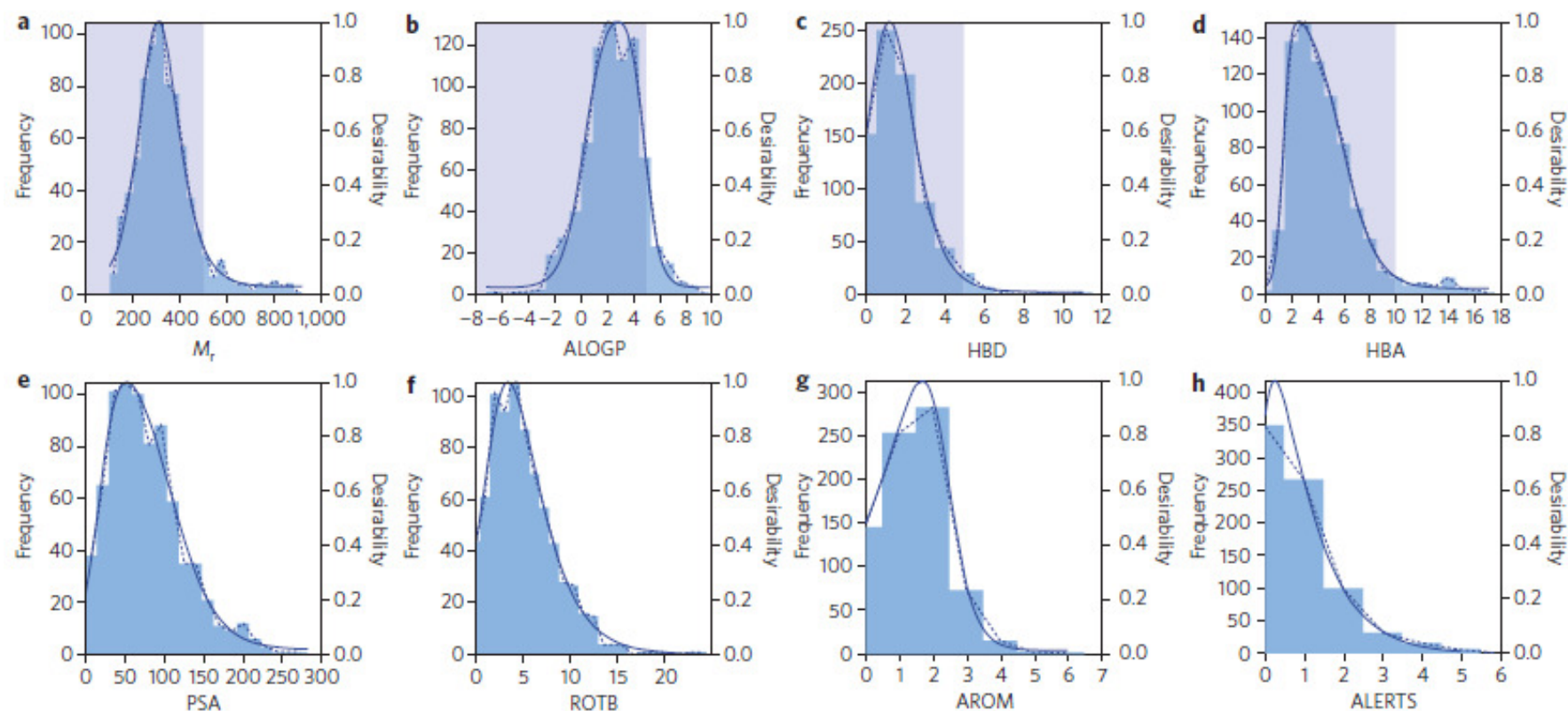
$$+ \frac{b}{\left[1 + \exp\left(-\frac{x-c+\frac{d}{2}}{e}\right)\right]} \left[1 - \frac{1}{\left[1 + \exp\left(-\frac{x-c-\frac{d}{2}}{f}\right)\right]}\right]$$

(a - f: constant values)



Mr: molecular weight

# Histograms of selected properties



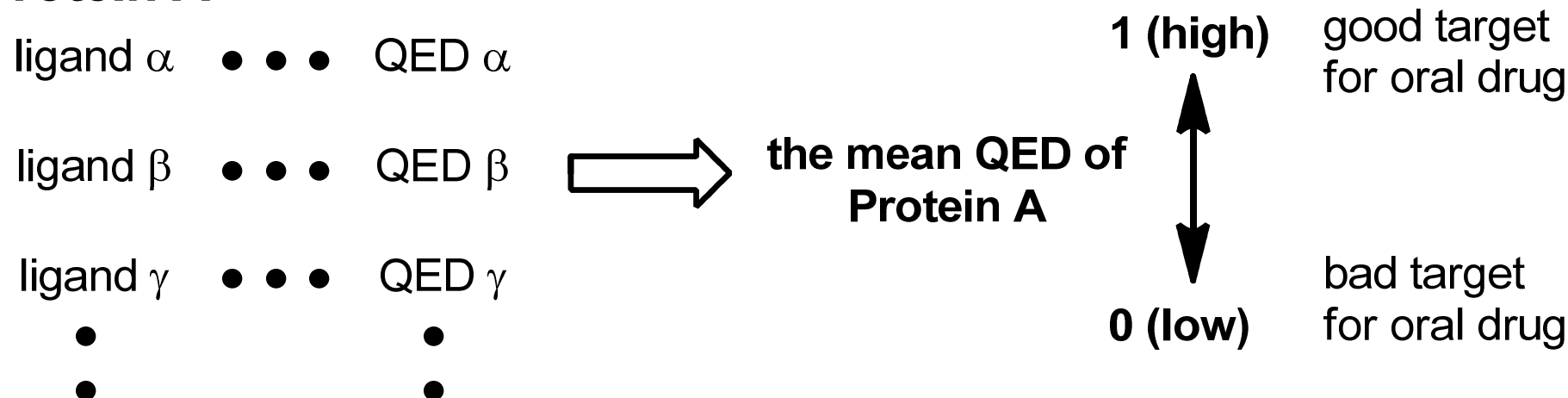
**Histograms of eight selected molecular properties for a set of 771 orally absorbed small molecule drugs.**

molecular weight  $M_r$  (a), lipophilicity estimated by atom-based prediction of ALOGP (b), number of HBDs (c), number of HBAs (d), PSA (e), number of ROTBs (f), number of AROMs (g) and number of ALERTS (h).

The Lipinski-compliant areas are shown in pale blue in (a), (b), (c) and (d).

## Prediction of the drug-likeness of proteins' ligand

### Protein A



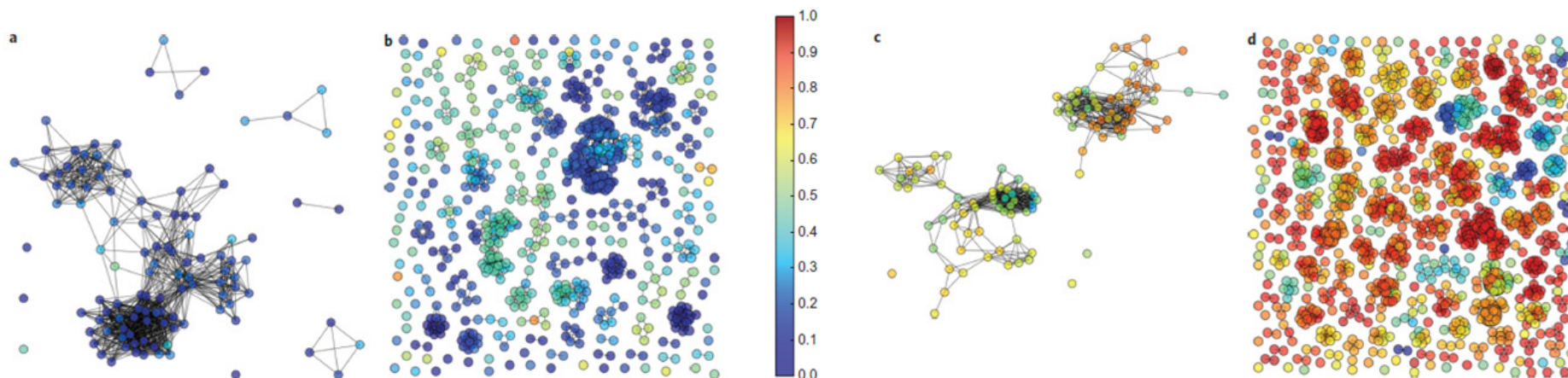
Not all ligand-binding sites have the appropriate physicochemical and topological properties to bind small-molecule drugs non-covalently with sufficient affinity.

Binding sites that do have these characteristics are described as **druggable** (this definition is independent of any wider biological considerations).

QED provides an efficient means to quantify and rank the druggability of targets according to the chemical attractiveness of their associated ligands.

In other words, **proteins whose ligands had the highest QED scores should be the most chemically tractable targets for drug discovery**, because their known ligands are the most drug-like.

# Structural diversity networks



- (a) a target for which the associated bioactive compounds are **neither drug-like nor diverse**
- (b) a target for which the associated bioactive compounds are **diverse, but not drug-like**
- (c) a target for which the associated bioactive compounds are **drug-like, but not diverse**
- (d) a target for which the associated bioactive compounds are **both drug-like and diverse**

In each of the networks compounds are represented as nodes and are coloured by their respective QED values. An edge connects nodes if they are structurally similar (defined by a Tanimoto coefficient  $\geq 0.7$ ).



## Top human targets by three different ranking schemes

Target (UniProt)	Mean QED	Target (UniProt)	Mean QED best cluster	Target (UniProt)	Proportion clusters with mean QED >0.796
1 Free fatty acid receptor 2 (O15552)	0.861	Neuropeptide Y receptor type 5 (Q15761)	0.935	Vesicular acetylcholine transporter (Q16572)	0.714
2 Sodium channel (Q9NY72, O60939, Q8IWTL, Q07699)	0.849	Serotonin transporter (P31645)	0.932	Phosphodiesterase 7A (Q13946)	0.667
3 Voltage-gated potassium channel (P15382, P51787)	0.835	Serotonin 1a (5-HT1a) receptor (P08908)	0.932	Melatonin receptor 1A (P48039)	0.5
4 Phosphodiesterase 9A (O76083)	0.820	Norepinephrine transporter (P23975)	0.932	Melatonin receptor 1B (P49286)	0.462
5 Aldo-keto-reductase family 1 member C3 (P42330)	0.812	Dopamine transporter (Q01959)	0.931	Norepinephrine transporter (P23975)	0.453
6 Cholinergic receptor, nicotinic, beta 1 (muscle) (Q8IZ46)	0.811	Histamine H1 receptor (P35367)	0.928	Histamine H4 receptor (Q9H3N8)	0.444
7 Sorbitol dehydrogenase (Q00796)	0.809	Dopamine D3 receptor (P35462)	0.927	Dopamine transporter (Q01959)	0.409
8 Sodium channel protein type IV alpha subunit (P35499)	0.809	Dopamine D4 receptor (P21917)	0.925	Serotonin 7 (5-HT7) receptor (P34969)	0.4
9 Endothelial lipase (Q9YSX9)	0.804	Thromboxane-A synthase (P24557)	0.921	Neuronal acetylcholine receptor protein beta-4 subunit (P30926)	0.4
10 Vesicular acetylcholine transporter (Q16572)	0.798	Serotonin 2c (5-HT2c) receptor (P28335)	0.917	Neuronal acetylcholine receptor protein alpha-7 subunit (P36544)	0.4

**Left** | ranking targets by the mean QED of their associated ligands

**Center** | ranking targets by the mean of the most drug-like active series (clusters)

**Right** | ranking targets by the degree of enrichment of drug-like series (type (d) target in the previous slide) (targets are ranked by the proportion of active series that have a mean QED above that of the top 10% of the ChEMBL database (0.796)) .

The mean QED for all targets in the list is 0.478. For the targets of approved drugs the mean QED is 0.492 and for the targets of approved oral drugs the mean QED is 0.539 . Drug targets are, indeed, enriched towards the more highly desirable targets, with 70% of the drug targets found in the top 50% of the prioritized target list.