Ping Zhang
pzhang@us.ibm.com
IBM T.J. Watson Research Center
USA

Lun Yang
Lun.Yang@gmail.com
GlaxoSmithKline
USA

Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Data Sources for Data Mining Applications
- Case study: Personalized Medicine
- Case Study: Drug Repositioning
- Future Challenges and Summary

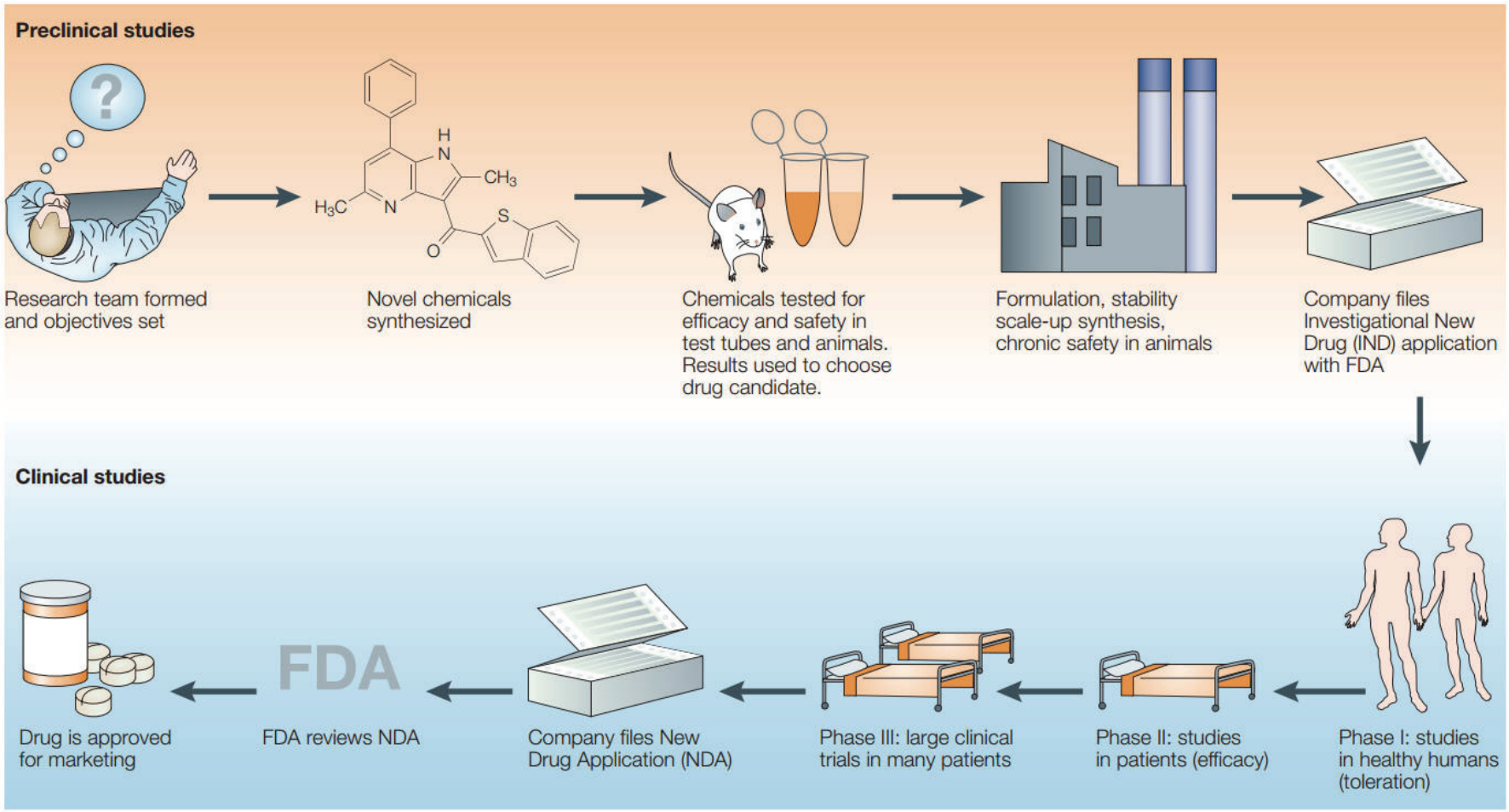
Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Data Sources for Data Mining Applications
- Case study: Personalized Medicine
- Case Study: Drug Repositioning
- Future Challenges and Summary

Brief history of drug discovery and development

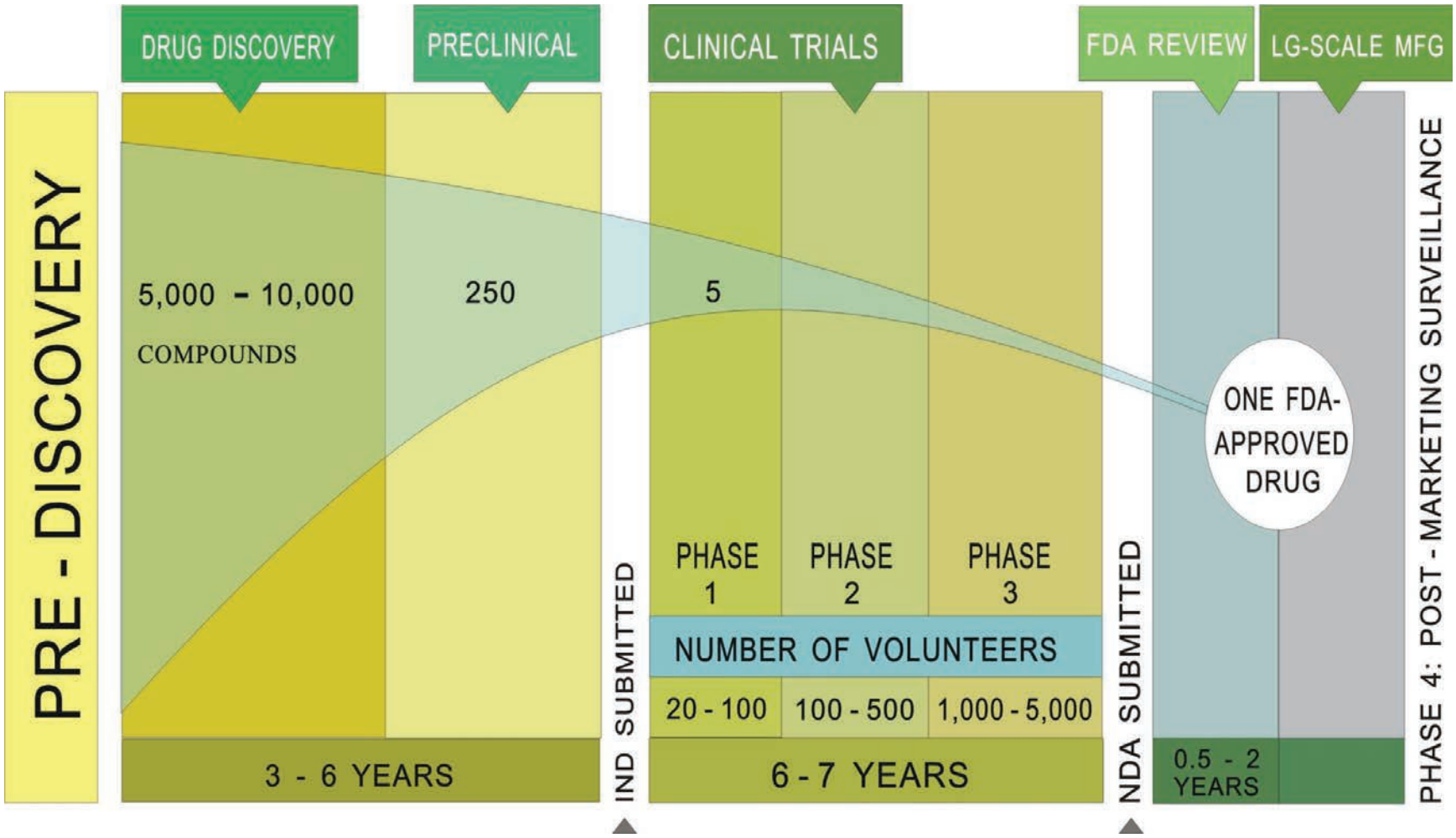
- Empirical – up until 1960's
  - 14th–11th centuries BCE: herbal drugs, serendipitous discoveries
  - Late 1800's: major pharmaceutical companies, mass production
  - 1920's, 30's: vitamins, vaccines
  - 1930-1960: major discoveries (insulin, penicillin, …)
- Rational – 1960's to 1990's
  - Designing molecules to target protein active sites – "lock and key"
  - Computational drug discovery
  - Biggest success HIV (RT, protease inhibitors)
- Big Experiment – 1990's to 2000's
  - High throughput screening
  - Microarray assays
  - Gene sequencing and human genome project
- Big Data – 2010's onwards
  - Informatics-driven drug discovery
  - Everything is connected

Stages in the drug discovery process



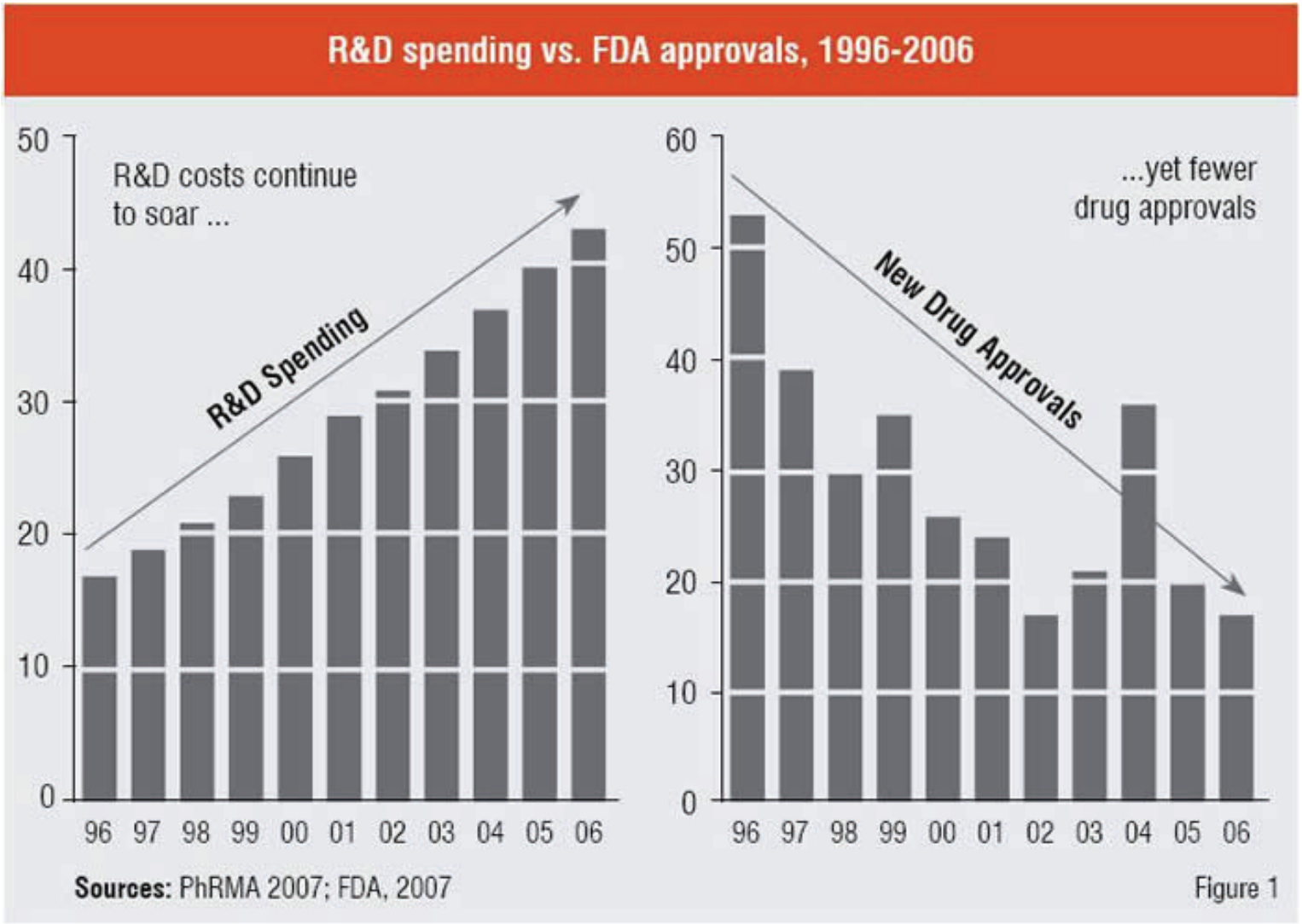
Lombardino JG, Lowe JA 3rd. Nat Rev Drug Discov. 2004 Oct;3(10):853-62.

Timescale in the drug discovery process



Available at http://www.innovation.org/drug_discovery/objects/pdf/RD_Brochure.pdf

Bottleneck in drug discovery



Available at http://www.pharmafocusasia.com/strategy/drug_discovery_india_force_to_reckon.htm

Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Data Sources for Data Mining Applications
- Case study: Personalized Medicine
- Case Study: Drug Repositioning
- Future Challenges and Summary

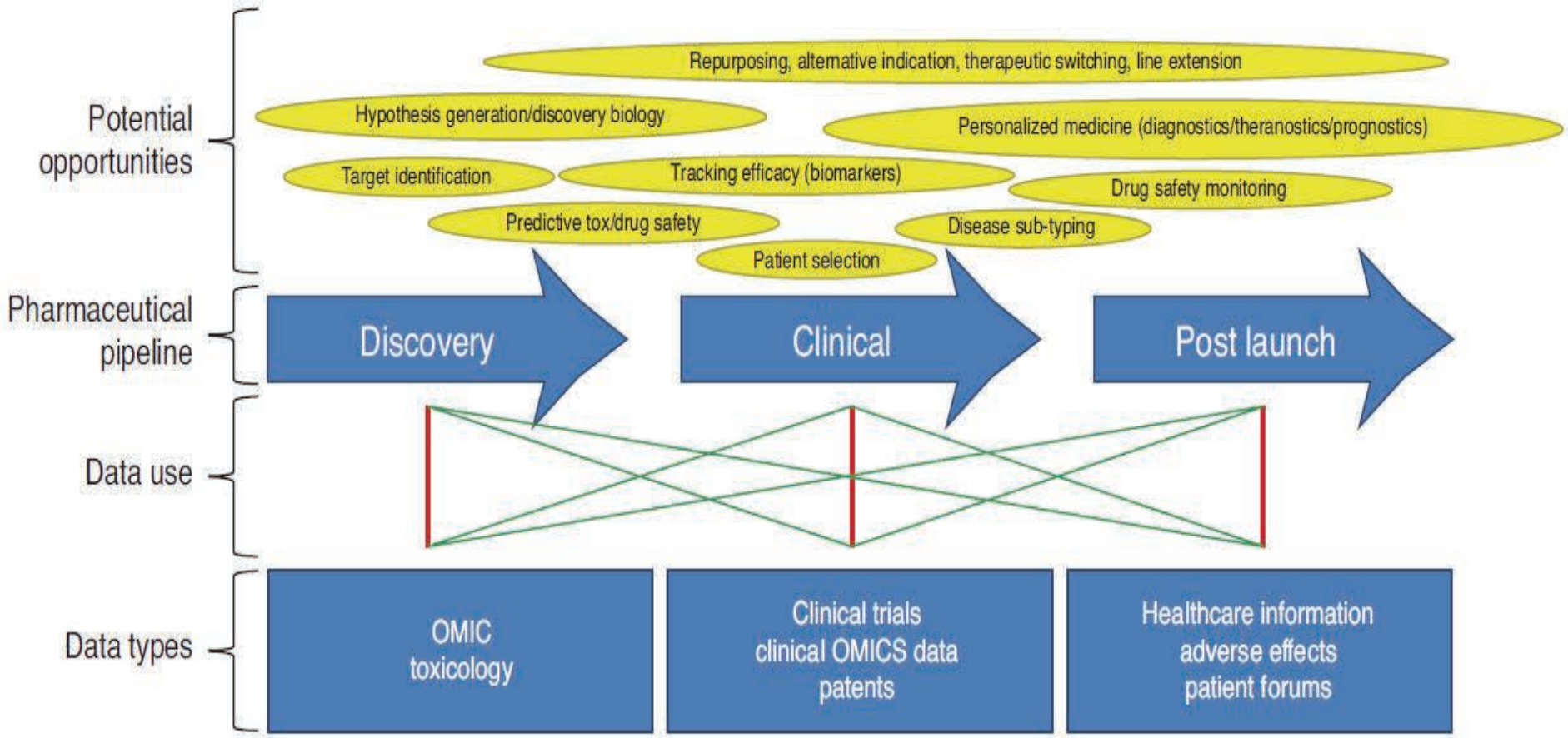
Big Data in the public domain

- There is now an incredibly rich resource of public information relating compounds, targets, genes, pathways, and diseases. Just for starters there is in the public domain information on*:
  - 48,777,362 compounds, 127,906,628 substances, 739,657 bioassays (PubChem)
  - 1552 FDA-approved small molecule drugs, 284 biotech drugs, 6009 experimental drugs (DrugBank)
  - 542,258 manually reviewed protein sequences, 51,616,950 un-reviewed protein sequences (Swiss-Prot/UniProtKB), 95,968 3D structures (PDB)
  - 22 million life science publications – 1 million new each year (PubMed)
  - 160,781 clinical studies with locations in all 50 states and in 185 countries (ClinicalTrials.gov)

- Even more important are the relationships between these entities. For example a chemical compound can be linked to a gene or a protein target in a multitude of ways:
  - Crystal structure of ligand/protein complex
  - Co-occurrence in a paper abstract
  - Computational experiment (docking, predictive model)
  - System association (e.g. involved in same pathways cellular processes)
  - Statistical relationship

* All databases were accessed on 02/08/2014

Why Data Mining is appealing



Buchan NS et al. Drug Discov Today. 2011 May;16(9-10):426-34.

Why Drug Discovery and Development is appealing

- Drug discovery is highly data driven and data are increasingly becoming public available
  - NIH has started ambitious extramural funding programs to support academic-based drug discovery programs recently
  - Pharms begin to make the trove of detailed raw data underlying its clinical trials systematically available to researchers
- Having ample data, bring challenging problems, demanding more knowledge
- Spans full data analytics cycles
  - Data collection, data cleansing, data semantics, data integration, data representation
  - Model inference, model selection, modal average, model interpretation
- We see many different data types
  - Vector, semi-structured, time-series, spatial-temporal, images, video, hypertext, literature
- Data analytics and data management challenges are from all aspects
  - Large volume, high dimensional, high noise, large amount of missing values, non iid data, structured input and output, unlabeled data
  - Multi-instance (label, class, task)

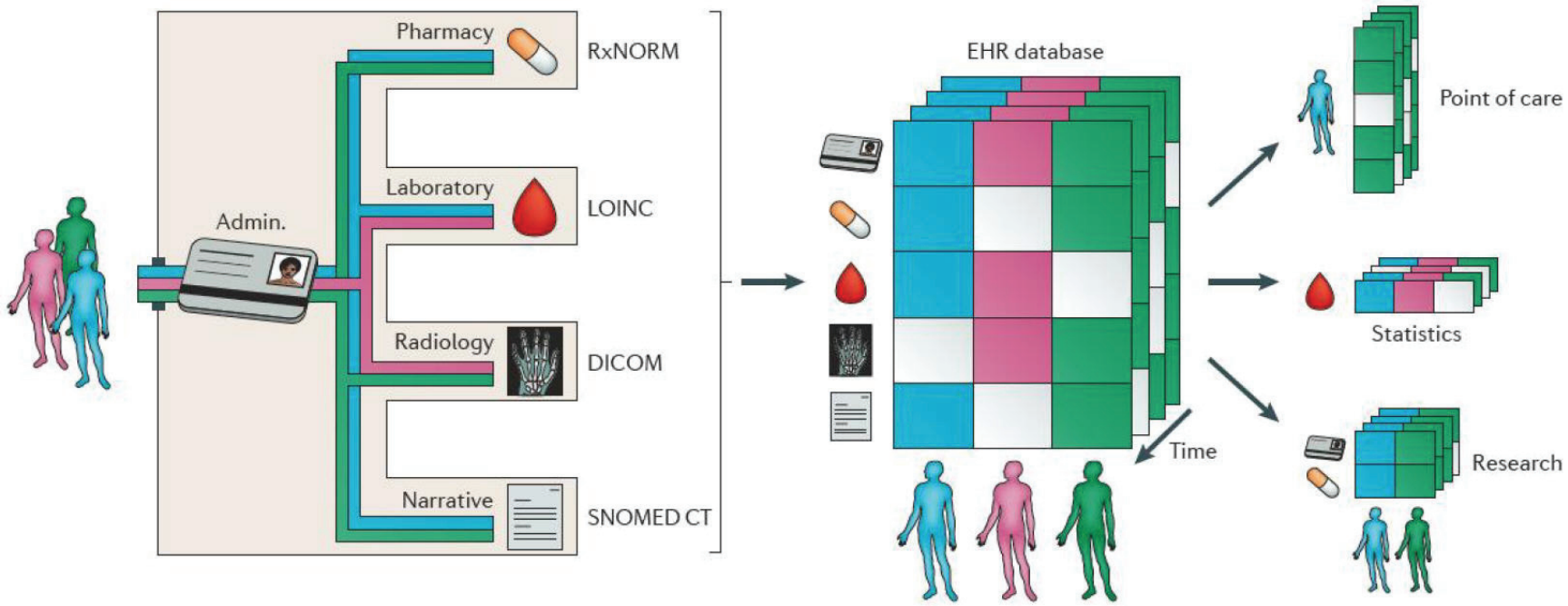
Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Data Sources for Data Mining Applications
- Case study: Personalized Medicine
- Case Study: Drug Repositioning
- Future Challenges and Summary

Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Data Sources for Data Mining Applications
- Case study: Personalized Medicine
- Case Study: Drug Repositioning
- Future Challenges and Summary

EHR data collection and analysis



Effectively integrating and efficiently analyzing various forms of healthcare data over a period of time can answer many of the impending healthcare problems.

Jensen PB, Jensen LJ, Brunak S. Nat Rev Genet. 2012 May 2;13(6):395-405.

Diagnosis data - ICD codes

- ICD stands for International Classification of Diseases
- ICD is a hierarchical terminology of diseases, signs, symptoms, and procedure codes maintained by the World Health Organization (WHO)
- In US, most people use ICD-9, and the rest of world use ICD-10
- Pros: Universally available
- Cons: medium recall and medium precision for characterizing patients

Hypertensive disease (401 - 405)

- (401) Essential hypertension
  - (401.0) Hypertension, malignant
  - (401.1) Hypertension, benign
  - (401.9) Hypertension, Unspecified
- (402) Hypertensive heart disease
- (403) Hypertensive renal disease
  - (403.0) Malignant hypertensive renal disease
  - (403.1) Benign hypertensive renal disease
- (404) Hypertensive heart and renal disease
- (405) Secondary hypertension
  - (405.0) Malignant secondary hypertension
    - (405.01) Hypertension, renovascular, malignant
  - (405.1) Benign secondary hypertension
    - (405.11) Hypertension, renovascular benign

Procedure data - CPT codes

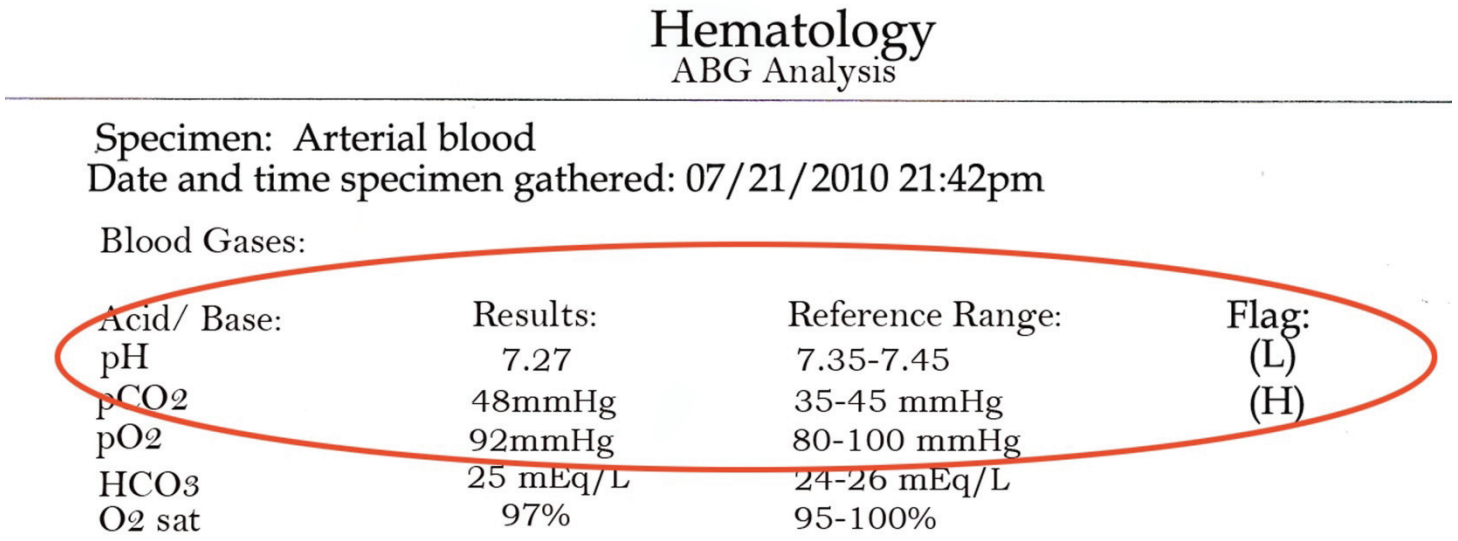
- CPT stands for Current Procedural Terminology created by the American Medical Association
- CPT is used for billing purposes for clinical services
- Pros: High precision
- Cons: Low recall

Codes for surgery: 10021 - 69990

- (10021 - 10022) general
- (10040 - 19499) integumentary system
- (20000 - 29999) musculoskeletal system
- (30000 - 32999) respiratory system
- (33010 - 37799) cardiovascular system
- (38100 - 38999) hemic and lymphatic systems
- (39000 - 39599) mediastinum and diaphragm
- (40490 - 49999) digestive system
- (50010 - 53899) urinary system
- (54000 - 55899) male genital system
- (55920 - 55980) reproductive system and intersex
- (56405 - 58999) female genital system
- (59000 - 59899) maternity care and delivery
- (60000 - 60699) endocrine system
- (61000 - 64999) nervous system
- (65091 - 68899) eye and ocular adnexa
- (69000 - 69979) auditory system

Lab results

- The standard code for lab is Logical Observation Identifiers Names and Codes (LOINC®)
- Challenges for lab
  - Many lab systems still use local dictionaries to encode labs
  - Diverse numeric scales on different labs
    - Often need to map to normal, low or high ranges in order to be useful for analytics
  - Missing data
    - not all patients have all labs
    - The order of a lab test can be predictive, for example, BNP indicates high likelihood of heart failure



Hematology
ABG Analysis

Specimen: Arterial blood
Date and time specimen gathered: 07/21/2010 21:42pm

Blood Gases:

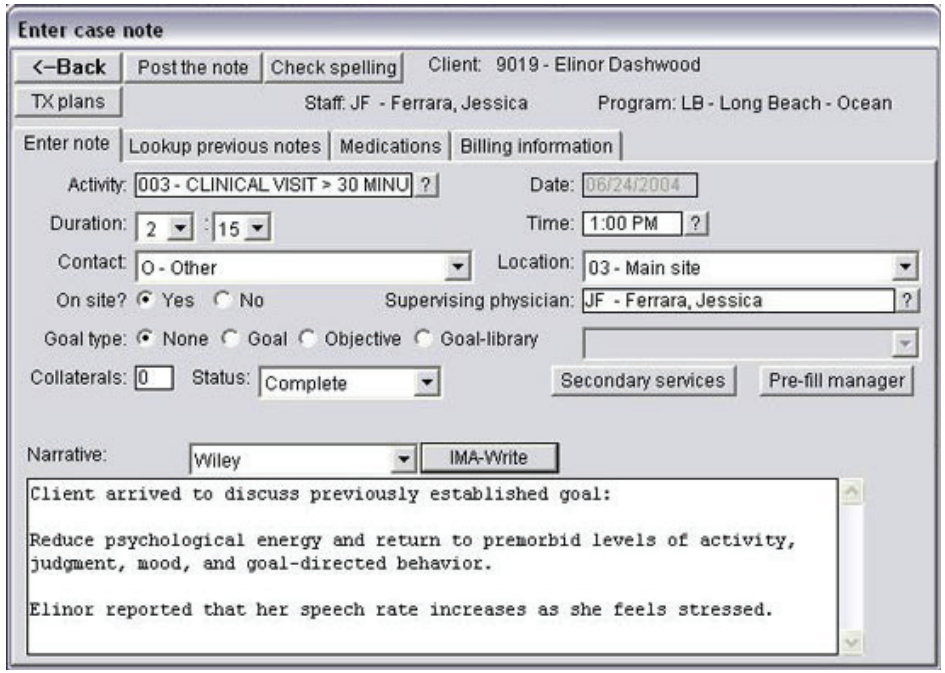
Acid/ Base:	Results:	Reference Range:	Flag:
pH	7.27	7.35-7.45	(L)
pCO2	48mmHg	35-45 mmHg	(H)
pO2	92mmHg	80-100 mmHg	
HCO3	25 mEq/L	24-26 mEq/L	
O2 sat	97%	95-100%	

Medication

- Standard code is National Drug Code (NDC) by Food and Drug Administration (FDA), which gives a unique identifier for each drug
  - Not used universally by EHR systems
  - Too specific, drugs with the same ingredients but different brands have different NDC
- RxNorm: a normalized naming system for generic and branded drugs by National Library of Medicine
- Medication data can vary in EHR systems
  - can be in both structured or unstructured forms
- Availability and completeness of medication data vary
  - Inpatient medication data are complete, but outpatient medication data are not
  - Medication usually only store prescriptions but we are not sure whether patients actually filled those prescriptions

Clinical notes

- Clinical notes contain rich and diverse source of information
- Challenges for handling clinical notes
  - Ungrammatical, short phrases
  - Abbreviations
  - Misspellings
  - Semi-structured information
    - Copy-paste from other structure source
      - Lab results, vital signs
    - Structured template:
      - SOAP notes: Subjective, Objective, Assessment, Plan

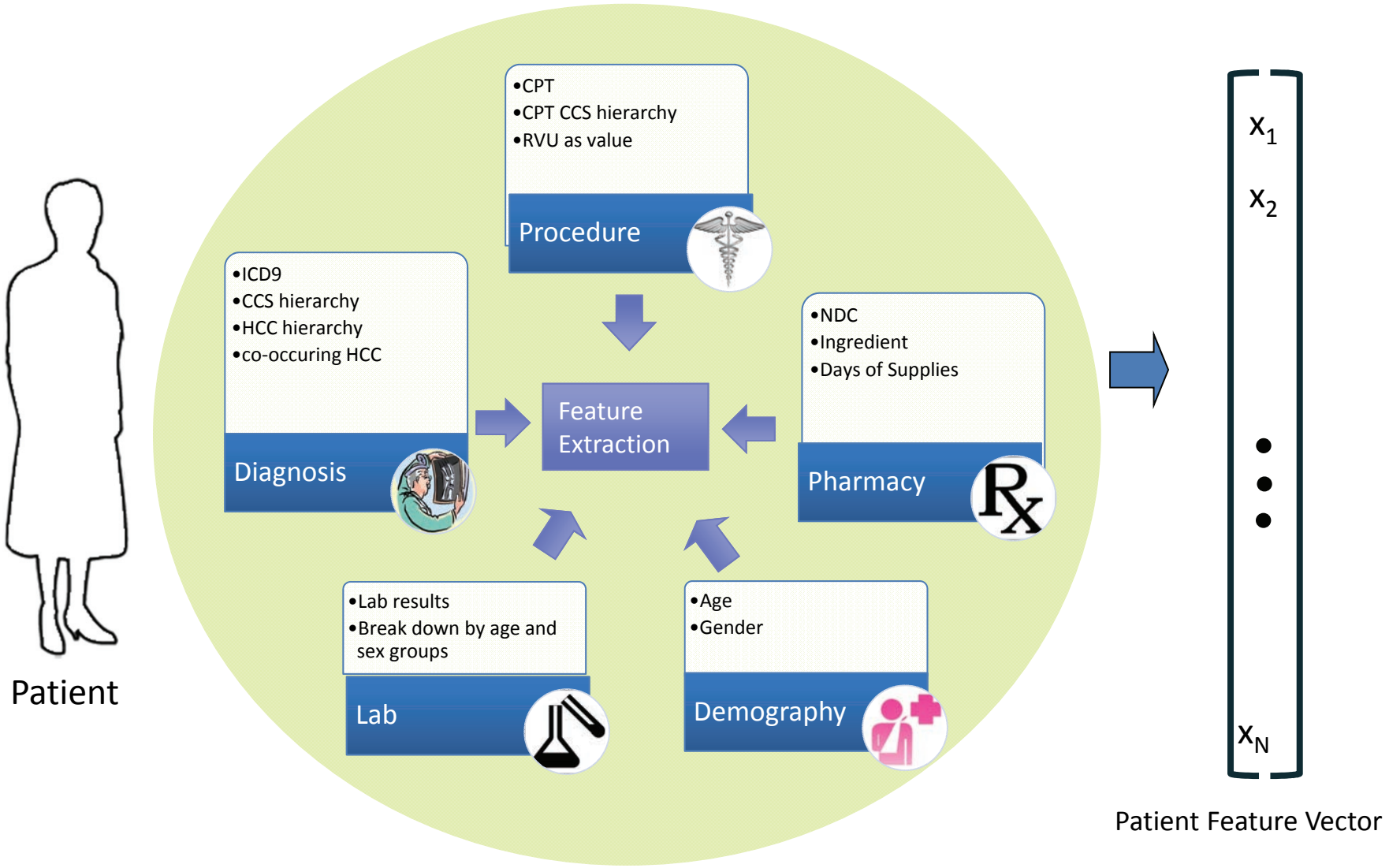


Strengths and weakness of data classes within EHRs

	ICD codes	CPT codes	Laboratory Data	Medication records	Clinical Documentation
Availability in EHR systems	Near-universal	Near-universal	Near-universal	Variable	Variable
Recall	Medium	Poor	Medium	Inpatient: High Outpatient: Variable	Medium
Precision	Medium	High	High	Inpatient: High Outpatient: Variable	Medium-High
Fragmentation effect	Medium	High	Medium-High	Medium	Low-Medium
Query method	Structured	Structured	Mostly structured	Structured, text queries, and NLP	NLP, text queries, and rarely structured
Strengths	-Easy to query -Serves as a good first pass of disease status	-Easy to query -High precision	-Value depends on test -High data validity	Can have high validity	Best record of what providers thought
Weaknesses	-Disease codes often used for screening when disease not a ctually present -Accuracy hindered by billing realities and clinic workflow	-Most susceptible to missing data errors (e.g., performed at another hospital) -Procedure receipt influenced by patient and payer factors external to disease process	-May need to aggregate different variations of the same data elements -Normal ranges and units may change over time	-Often need to interface inpatient and outpatient records -Medication records from outside providers not present -Medications prescribed not necessary taken	-Difficult to process automatically -Interpretation accuracy depends on assessment method -May suffer from significant "cut and paste" -Not universally available in EHRs -May be self-contradictory
Summary	Essential first element for electronic phenotyping	Helpful addition if relevant	Helpful addition if relevant	Useful for confirmation and a marker of severity	Useful for confirming common diagnoses or for finding rare ones

Denny JC. PLoS Comput Biol. 2012;8(12):e1002823.

EHR data description and patient vector represent



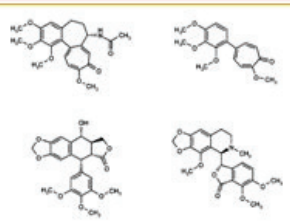
Moving towards to Personalized Medicine

- Personalized Medicine: the right patient with the right drug at the right dose at the right time.
- Safer, more effective drugs: end of one-size fits-all drugs. Target discoveries enable development of drugs that will be safer and effective for specific populations.
- Faster time to market: using genomic and real-world data to find disease targets. Speedier clinical trials based on high responder population.
- Cost-effective healthcare: reduced costs, due to avoidance of futile treatments and improved clinical outcomes. Better treatment adherence = increased profitability.

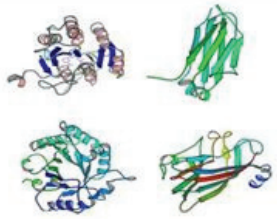
Intuition of Personalized Medicine methods

- weight loss
- impotence
- dizziness
- blurred vision
- ……

Side-effects Similarity



Drug structures Similarity



Target proteins Similarity

- Combine patient similarity with drug similarity analysis.
- Leverages large amount of real-world data available for "mature" drugs to derive information relevant for a new drug.

- ICD9
- CCS hierarchy
- HCC hierarchy
- co-occuring HCC

Diagnosis

- CPT
- CPT CCS hierarchy
- RVU as value

Procedure

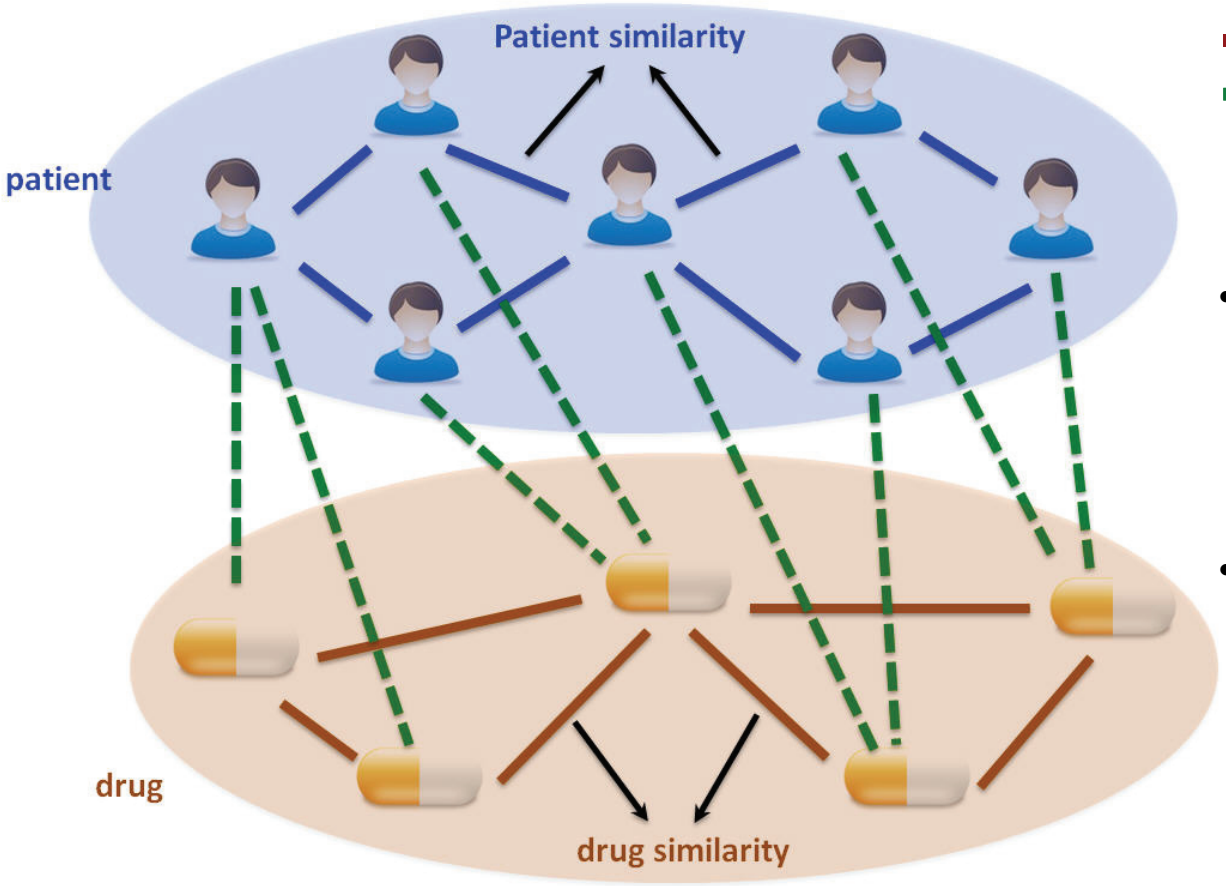
- Age
- Gender

Demography

- Lab results
- Break down by age and sex groups

Lab

Network based approach combining drug similarity with patient similarity



Patient similarity (S_p)

Drug similarity (S_d)

Patient-drug prior association (R): measured by drug's therapeutic indications

- Construct a heterogeneous patient-drug graph A:

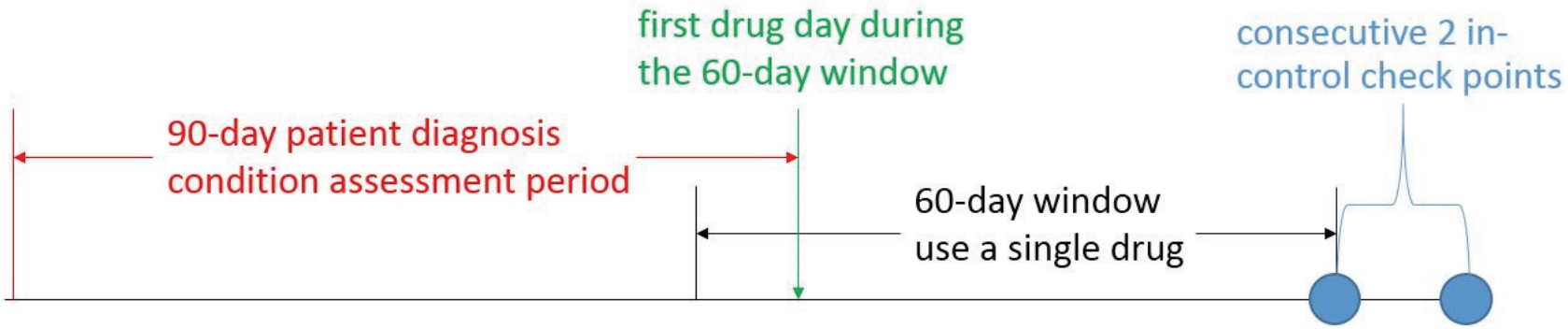
$$A=\begin{bmatrix} S_p & R \\ R^T & S_d \end{bmatrix}$$

- Spread the information representing the effectiveness of different drugs for different patients (vector Y) by a label propagation algorithm:

$$F=(1-\mu)(I-\mu W)^{-1}Y$$

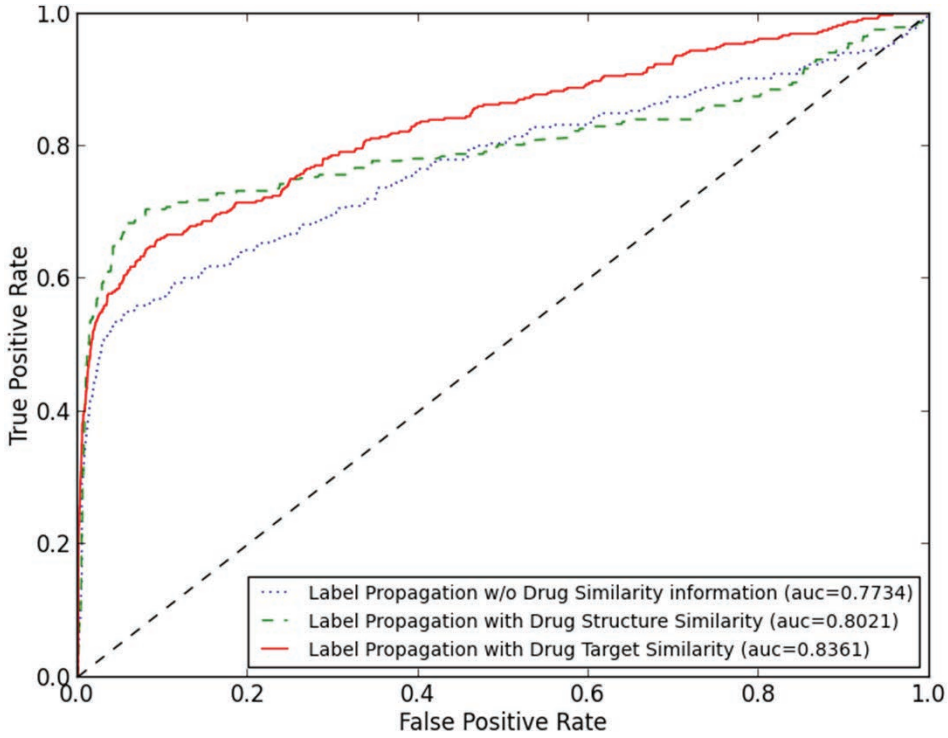
Zhang P et al. Summit on Translational Bioinformatics (TBI), 2014.

Application: personalized treatments for hypercholesterolemia



Data: 1219 distinct patients and 4 statin cholesterol-lowering drugs from a real-world EHR

Drug	Patient #
Atorvastatin	97
Lovastatin	221
Pravastatin	24
Simvastatin	877



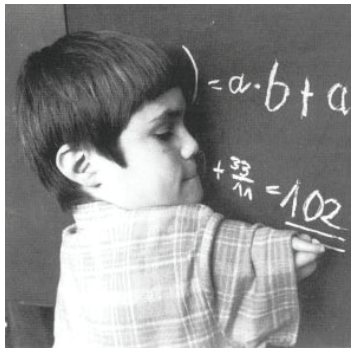
Outline

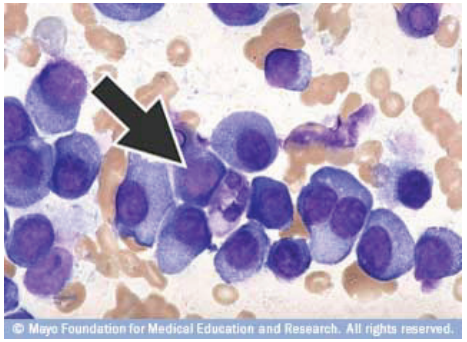
- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Data Sources for Data Mining Applications
- Case study: Personalized Medicine
- Case Study: Drug Repositioning
- Future Challenges and Summary

Examples of drug repositioning

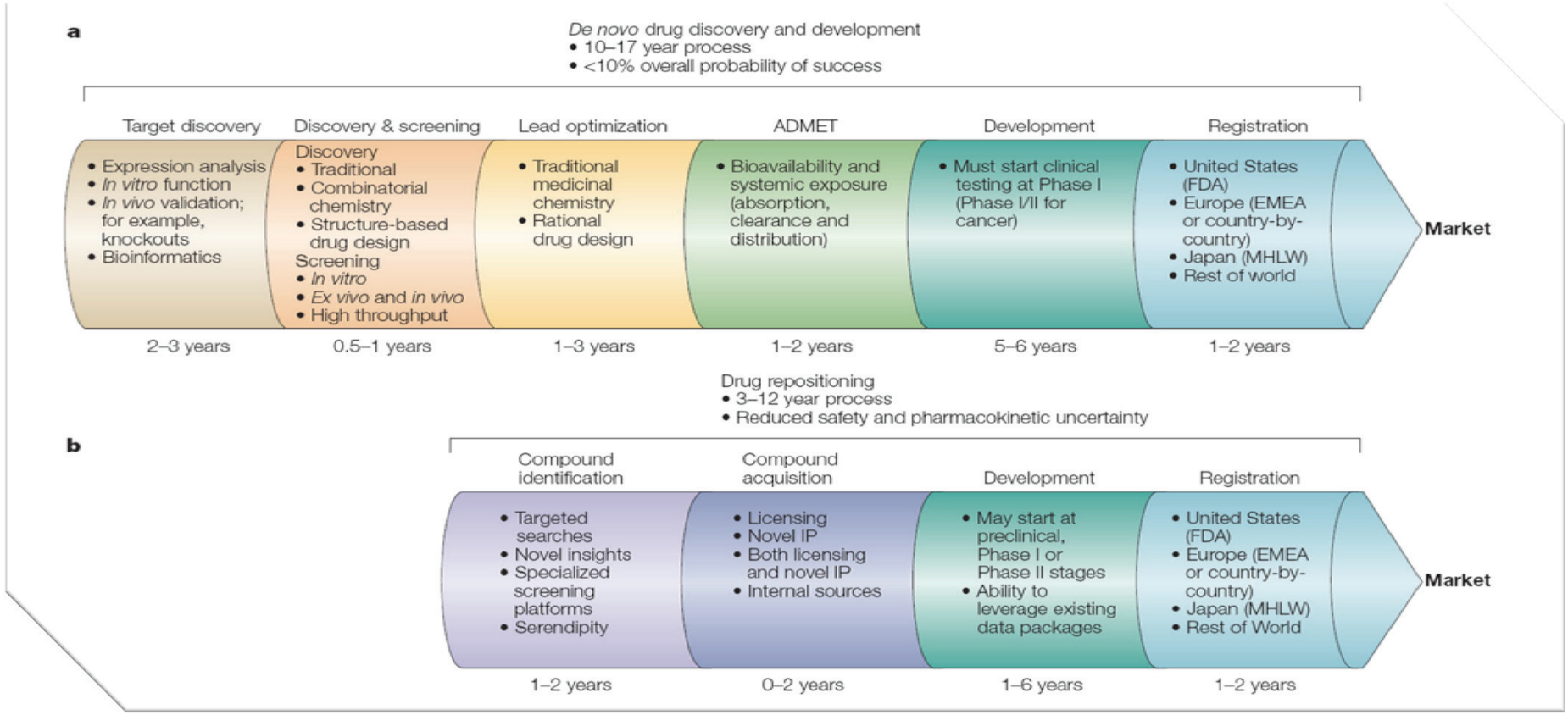
New uses for old drugs

Drug	Original indication	New indication
Viagra	Hypertension	Erectile dysfunction
Wellbutrin	Depression	Smoking cessation
Thalidomide	Antiemetic	Multiple Myeloma



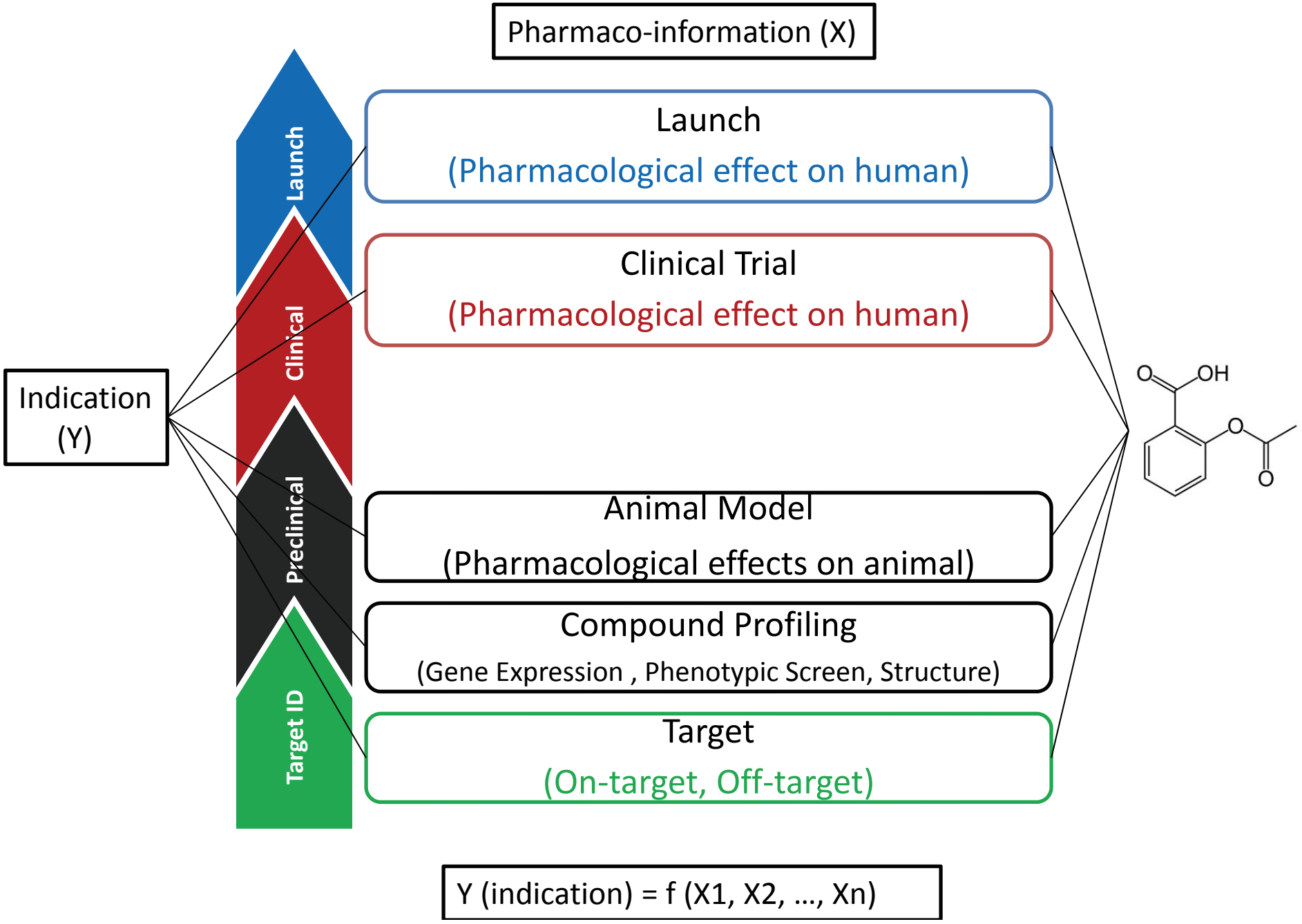


Meet the unmet medical needs efficiently



Ashburn & Thor (2004). Nature Rev Drug Disc 3, 673.

Dependent and Independent Variables in Drug Repositioning



Pharmaco-information (X)

Target

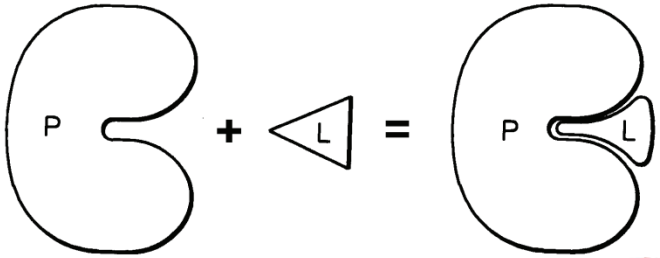
(On-target, Off-target)

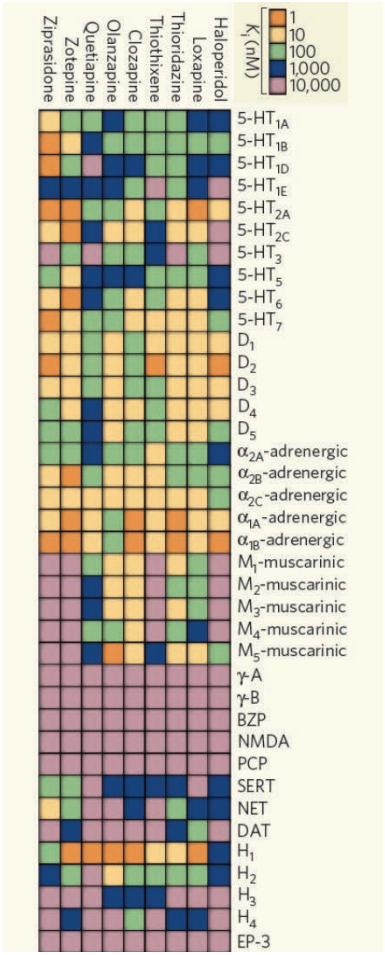
Identify the off-targets via Chemical-Protein Interactome (CPI)

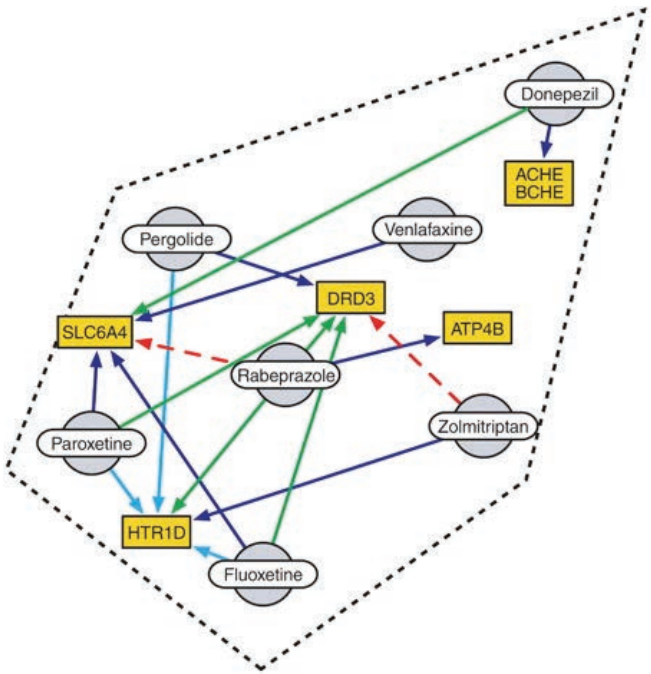
- Introduction of the CPI

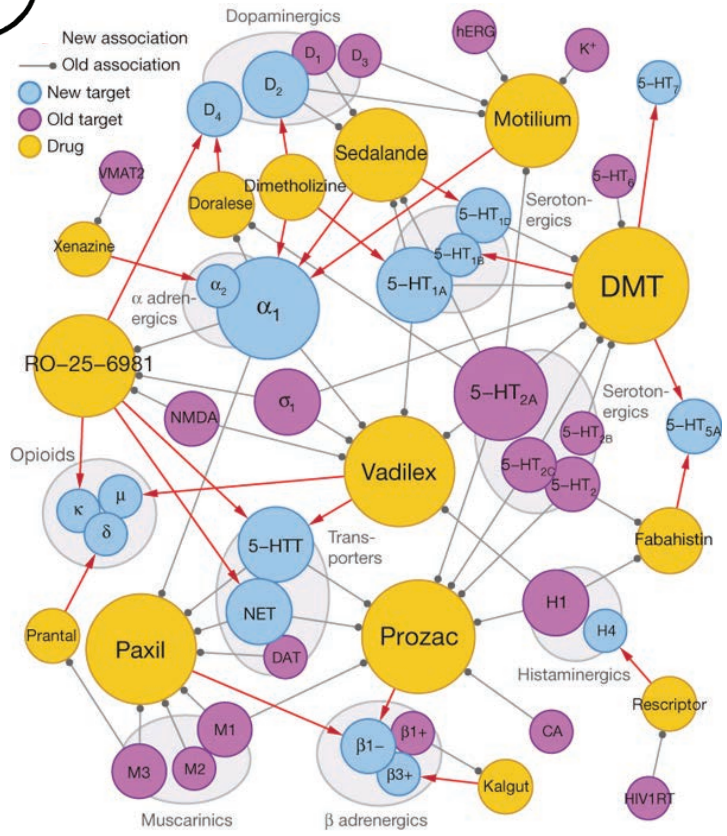
- Case study
  - Clozapine induced agranulocytosis (CIA)
    - Although agranulocytosis is a side effect, the methodology is applicable to the identification of the therapeutic effect

Chemical-protein interactions







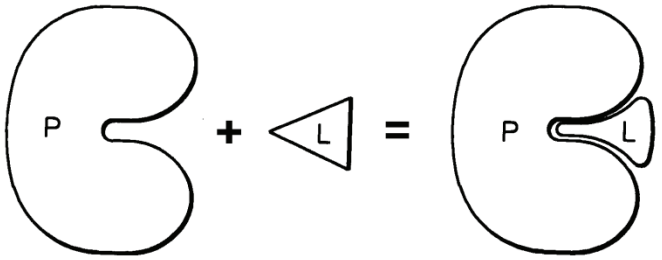


Nature Reviews Drug Discovery 3, 353-359 (2004)

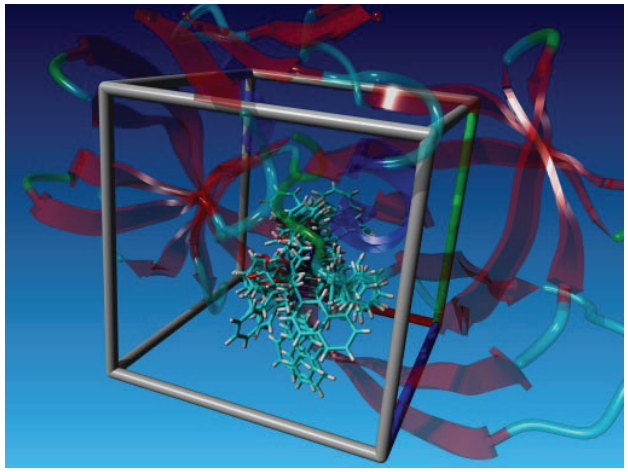
Bork, P. et al. Science 321(5886):263-266 (2008)

Keiser, M. et al. Nature 462, 175-181 (2009)

The DOCK



- A program used to simulate the chemical-protein interactions and to measure the interaction strength



- Provide the theoretical binding conformation of the drug's binding to protein

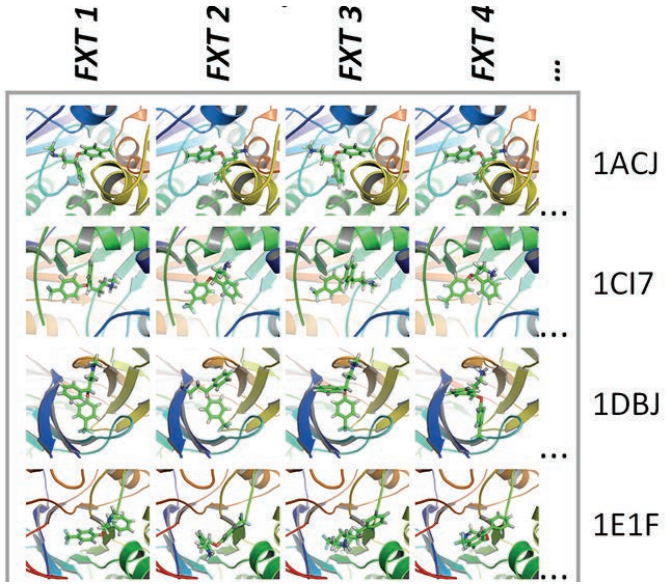
$$E_{\text{inter}} = \sum_{i=1}^{lig}\sum_{j=1}^{rec}\left(\frac{A_{ij}}{r_{ij}^{a}} - \frac{B_{ij}}{r_{ij}^{b}} + 332.0\frac{q_i q_j}{Dr_{ij}}\right),$$

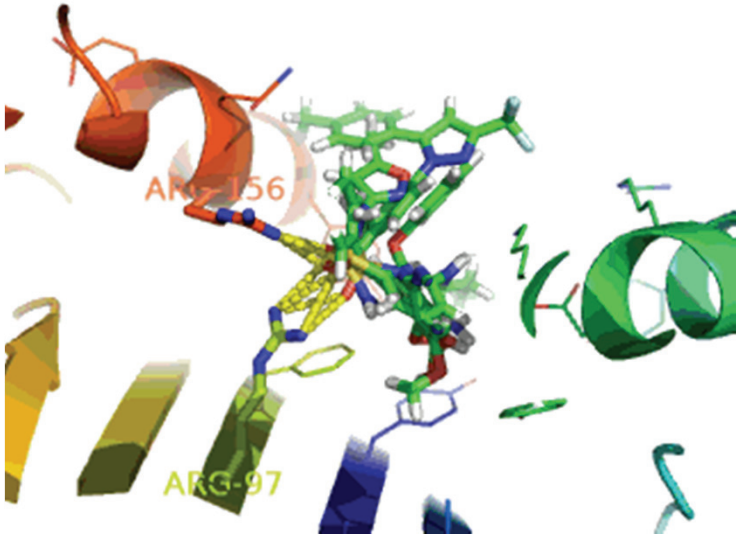
van der Waals and electrostatic interaction

- A lower docking score means a higher binding strength

http://dock.compbio.ucsf.edu/

Binding conformation in Chemical-Protein Interactome (CPI)

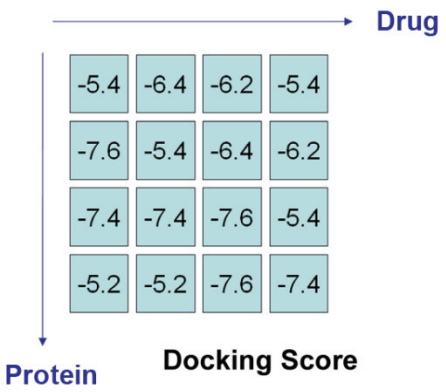




Direct binding model of sulfonamides - MHC I (Cw*4) interactions

Yang, L. et al. PLoS Comput Biol. 2009

Binding strength in CPI



Two Directional Z-transformation (2DIZ) of Docking Scores X_{ij}

$$Z_{ij} = \frac{X_{ij} - \overline{X_j}}{SD_{X_j}} \qquad Z'_{ij} = \frac{Z_{ij} - \overline{Z_i}}{SD_{Z_i}}$$

Linear Model of the Docking Scores

$$X_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \qquad b = \frac{\sum_{q=1}^{n}\sum_{k=1}^{m}(\alpha\beta)_{kq}}{mn}$$

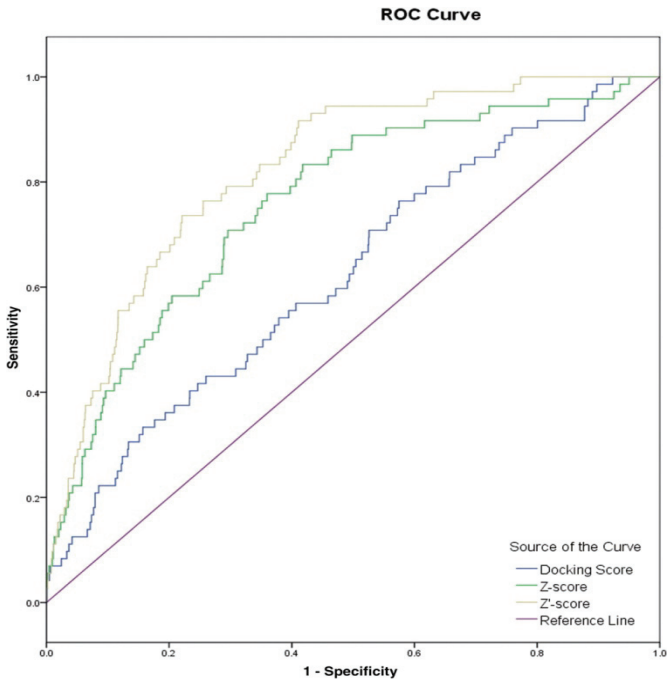
$$Z'_{ij} = \frac{-b\sqrt{n-1}}{\sqrt{(n-1)b^2 + [(\alpha\beta)_{ii} - b]^2}} \ (i \neq j), \text{ when } (m \to +\infty, n \to +\infty)$$

$$Z'_{ij} = \left[(\alpha\beta)_{ij} - b\right]\sqrt{\frac{(n-1)}{(n-1)b^2 + \left[(\alpha\beta)_{ij} - b\right]^2}} \ (i = j),$$

$$\text{when } (m \to +\infty, n \to +\infty),$$

Yang, L. et al. PLoS ONE. 2010

Improve the performance of the docking scores via using 2DIZ

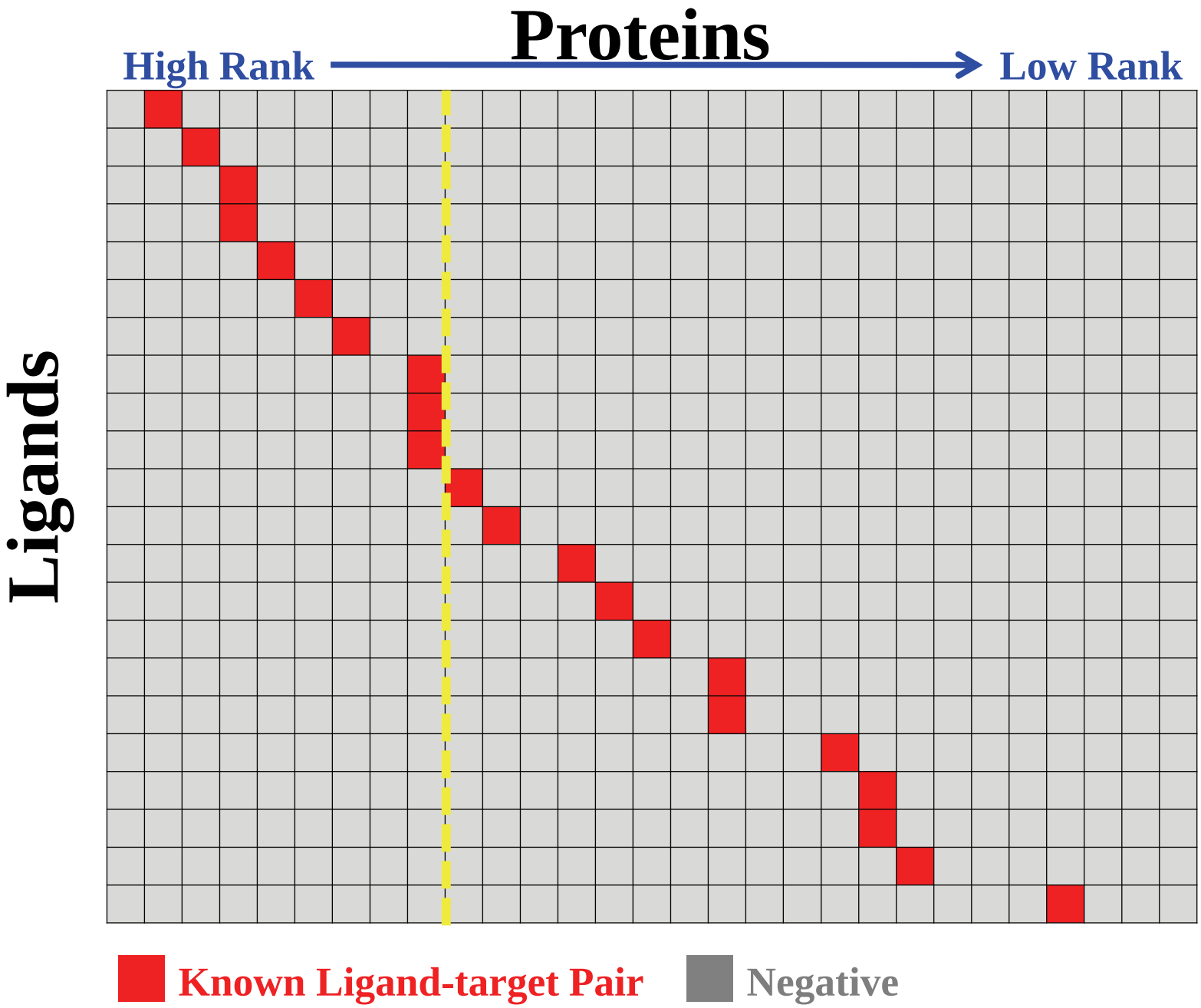


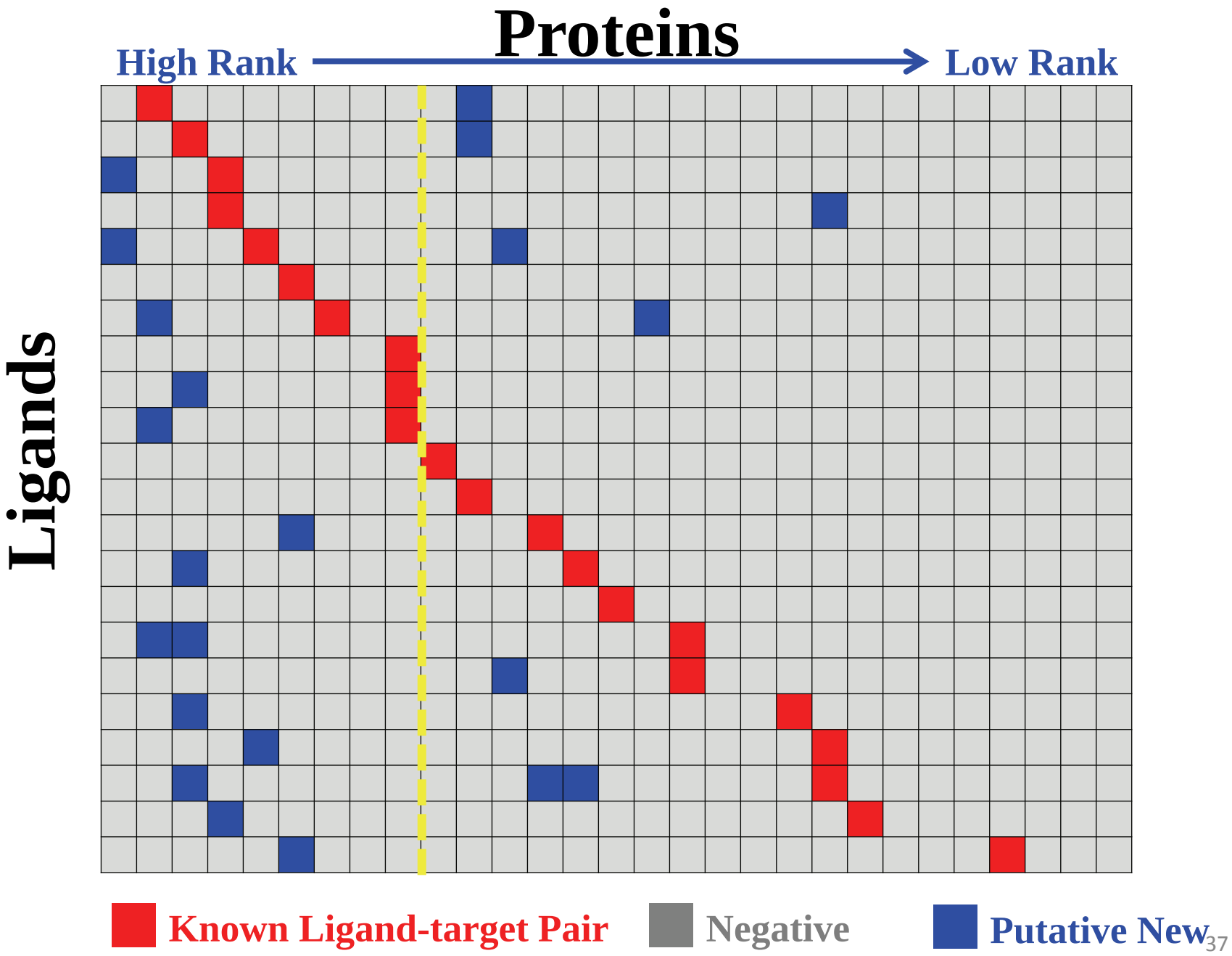
Benchmark structural model set:
100 pockets with their embedded ligands

High variability in ligand structures

Test Result Variable(s)	AUC	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Docking Score	**.623**	.033	.000	.558	.687
Z-score	**.759**	.028	.000	.703	.815
Z'-score	**.823**	.021	.000	.781	.865

Identify the True Chemical-Protein Interactions




Proteins
High Rank
Low Rank
Ligands
Known Ligand-target Pair
Negative
Putative New

False Positive - Tolerant MCC (FPT-MCC)

	p' (Predicted)	n' (Predicted)
P (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad |\mathrm{MCC}| = \sqrt{\frac{\chi^2}{n}}$$

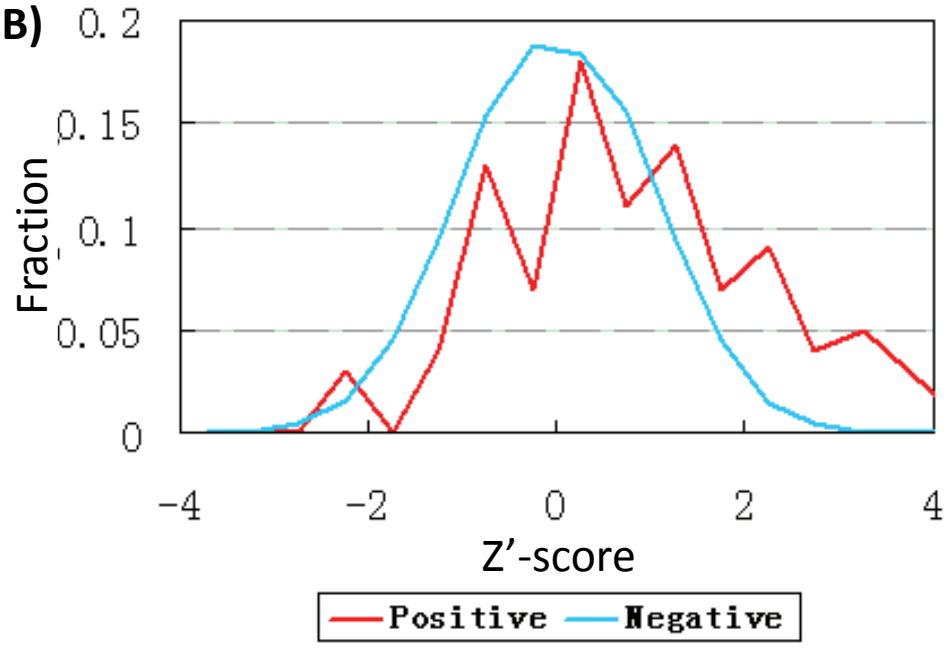
$$\textbf{(FPT-MCC)} = \frac{\boldsymbol{TP'} \times TN - \boldsymbol{FP'} \times FN}{\sqrt{(\boldsymbol{TP'}+\boldsymbol{FP'})(\boldsymbol{TP'}+FN)(TN+\boldsymbol{FP'})(TN+FN)}}$$

$$\boldsymbol{TP' = TP + \alpha FP}$$
$$\boldsymbol{FP' = (1-\alpha)FP}$$

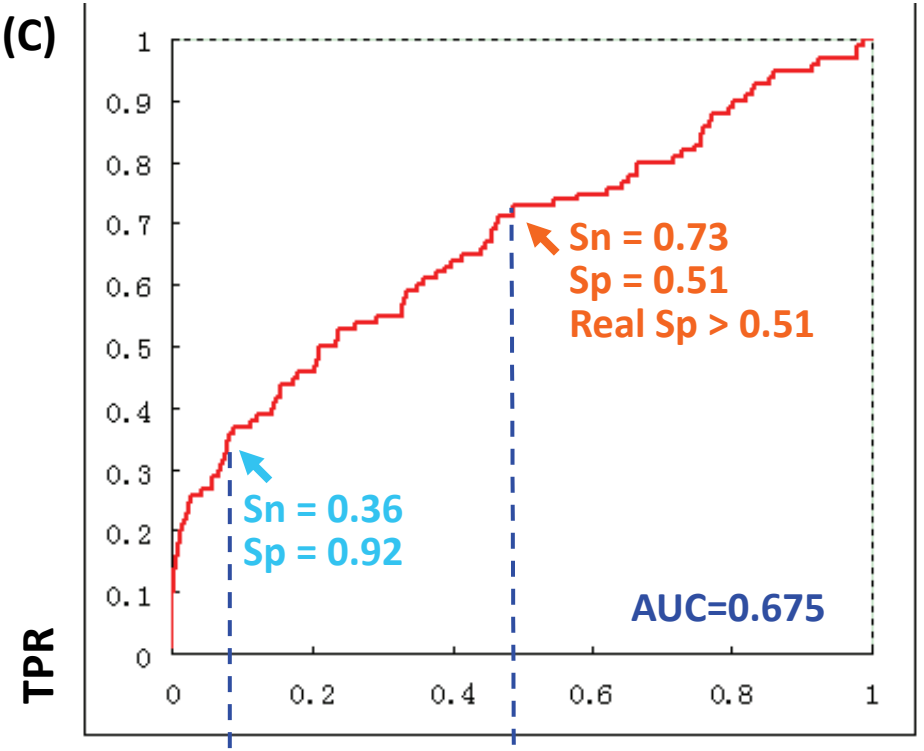
(A)

Class	Positive	Negative
Volume	100	10,000
Mean	1	0
St. Dev	1.5	1

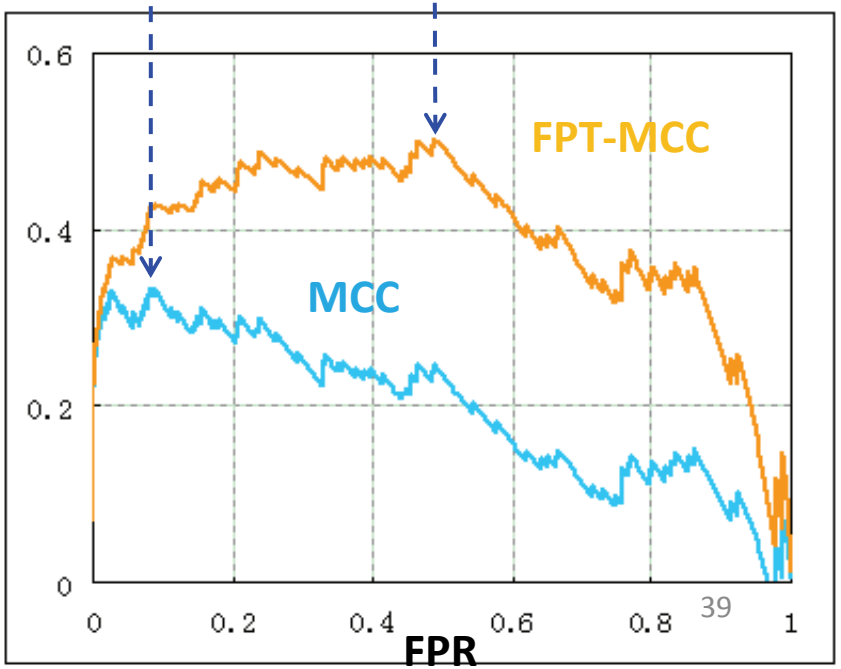
(B)



(C)



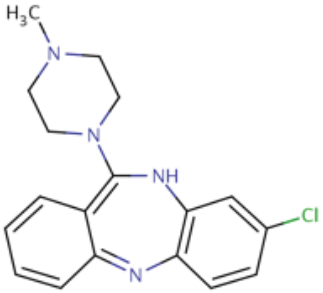
(D)



Case Study:

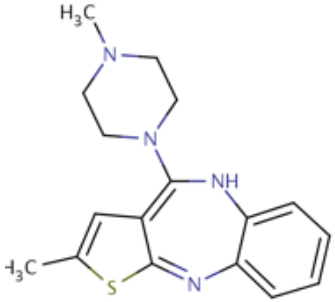
Identify Off-targets for Clozapine induced agranulocytosis (CIA)

Clozapine (CLZ)



Causing fatal agranulocytosis -- A deficiency of granulocytes in the blood

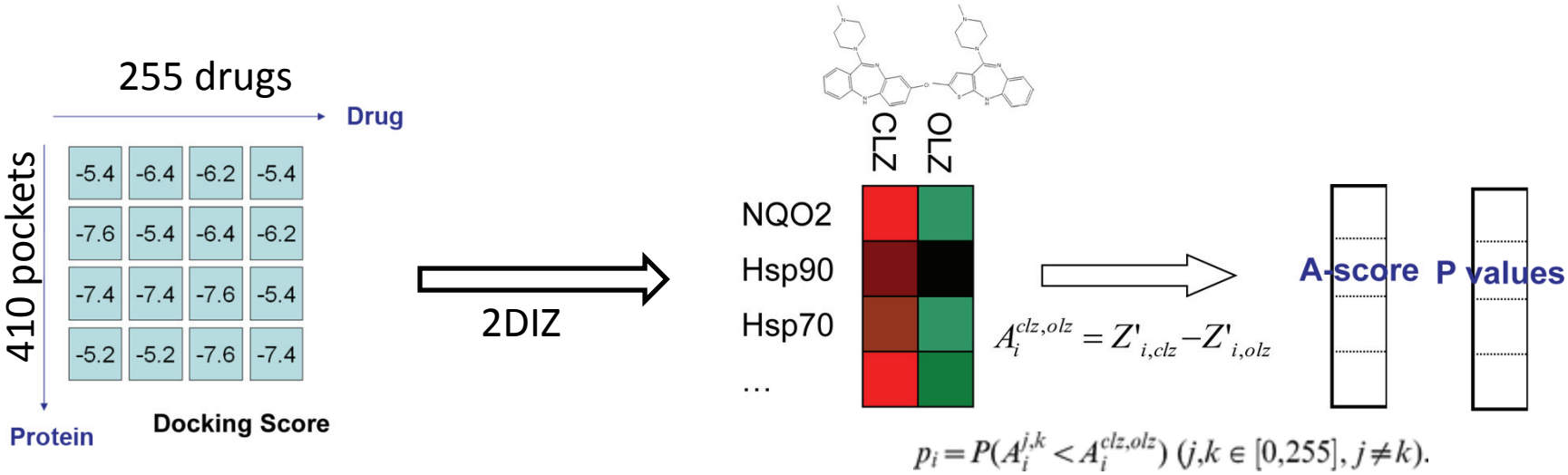
Olanzapine (OLZ)



The difference of the agranulocytosis report rate between clozapine and olanzapine in the FDA AERS

	Clozapine	Olanzapine
Agranulocytosis Reports	185	16
Total Reports	16813	11304
Ratio of Agranulocytosis Report (%)	1.1	0.14
$p_{CLZ\text{-}OLZ}$*		8.2E-21

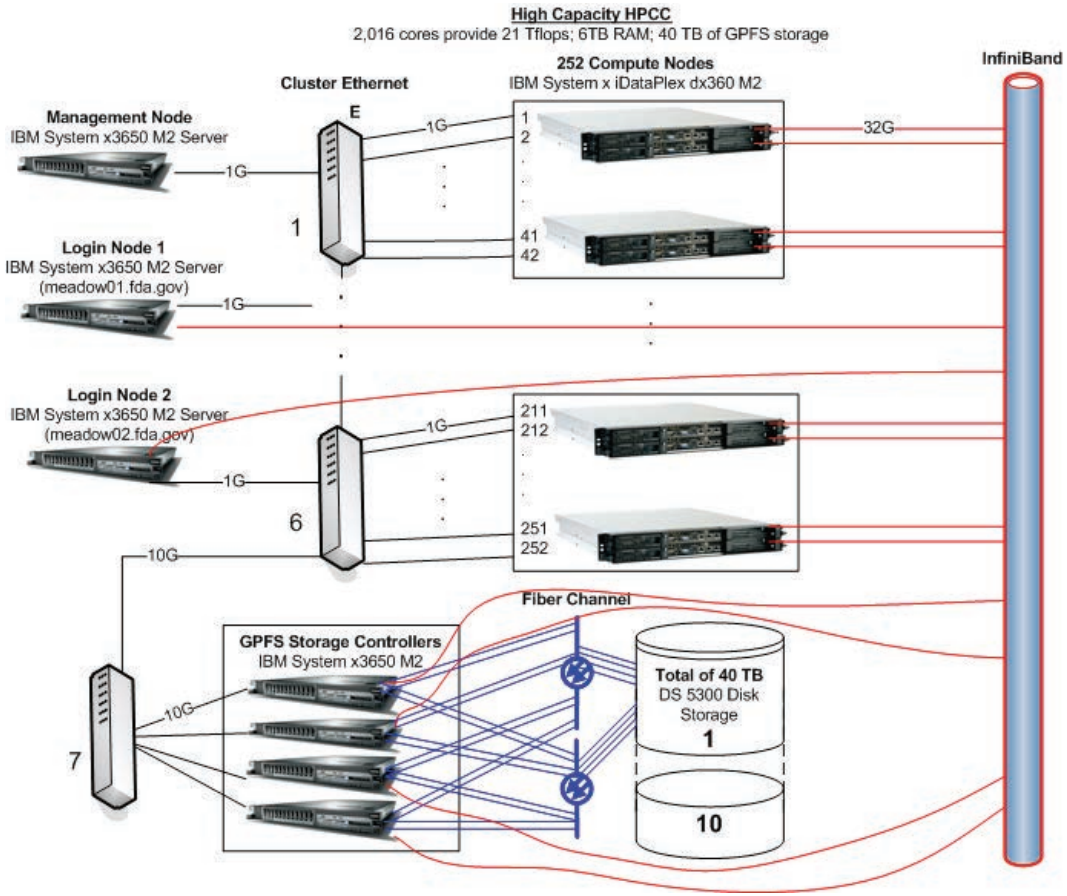
Identifying off-targets for CIA



Yang,L. et al. PLoS Comput. Biol. 2011

Resource Specifications for Docking

- Blue Meadow cluster
  - Located at Ashburn contractor operated data center
  - IBM iDataPlex dx360 M2 Server machines & Sun Grid Engine, PBS
  - 252 nodes x 8 cores = 2016 cores
  - 6TB RAM, or 24 GB per node
  - Memory distributed between nodes & shared within nodes



ANOVA of the chemical-protein interactive effect before and after 2DIZ

Normalize the chemical effect first

Before 2DIZ				
	Df	Sum Sq	F	p value
Protein	409	2332527	111.22	<2.2e-16
Chemical	254	10330585	793.27	<2.2e-16
Interactive	95344	4888387		
After 2DIZ				
Protein	409	0	1.37E-19	1
Chemical	254	1052	4.1776	<2.2e-16
Interactive	95344	94546		

Protein Effect has been excluded. May use drug to fish proteins.

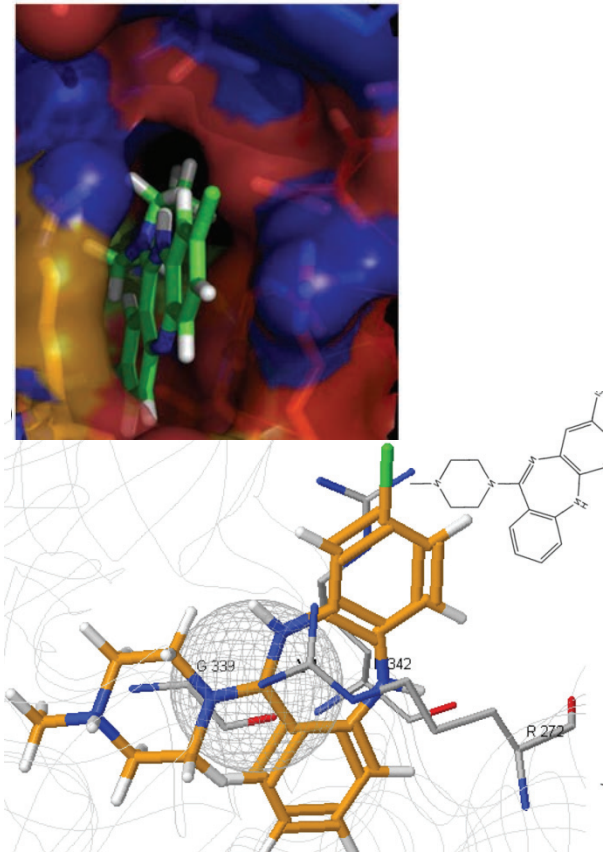
Candidate off-targets from Binomial Antithesis CPI

PDB ID[#]	Target Name	Gene Name	Z′ (CLZ)*	Z′ (OLZ)	A-score	p value for CPI	Role
1CBS	Cellular retinoic acid-binding protein 2	CRABP2	−0.922	1.653	−2.575	0.000	
1D1T	Alcohol dehydrogenase class 4 mu/sigma chain	ADH7	−1.191	1.525	−2.716	0.000	OR
1IHI_1	Aldo-keto reductase family 1 member C2	AKR1C2	−0.781	2.545	−3.326	0.000	OR
1IHI_2	Aldo-keto reductase family 1 member C2	AKR1C2	−1.605	1.023	−2.628	0.000	OR
1OIZ	Alpha-tocopherol transfer protein	TTPA	−1.269	1.171	−2.440	0.000	
2E8A	Heat shock 70 kDa protein 1	HSPA1A/HSPA1B	−1.381	0.150	−1.531	0.001	
1D2V	Myeloperoxidase	MPO	−2.753	−0.646	−2.107	0.005	OR
1DB1	Vitamin D3 receptor	VDR	−0.660	0.748	−1.409	0.012	
1MRQ_2	Aldo-keto reductase family 1 member C1	AKR1C1	−2.034	0.123	−2.158	0.016	OR
1MRQ_1	Aldo-keto reductase family 1 member C1	AKR1C1	−1.036	0.601	−1.637	0.021	OR
1DHT	Estradiol 17-beta-dehydrogenase 1	HSD17B1	−1.822	0.158	−1.980	0.021	OR
1MUO	Serine/threonine-protein kinase 6	AURKA	−1.136	0.529	−1.665	0.027	
1VJ5	Epoxide hydrolase 2	EPHX2	−1.088	0.228	−1.315	0.027	
4GTU	Glutathione S-transferase Mu 4	GSTM4	−0.749	1.060	−1.809	0.036	GT
1HDR	Dihydropteridine reductase	QDPR	−1.469	0.561	−2.030	0.038	OR
1YB5	Quinone oxidoreductase	CRYZ	−1.212	0.284	−1.496	0.039	OR
1CM8	Mitogen-activated protein kinase 12	MAPK12	−1.202	0.301	−1.503	0.039	
1XF0_2	Aldo-keto reductase family 1 member C3	AKR1C3	−0.865	0.441	−1.306	0.041	OR
1HMR	Fatty acid-binding protein, heart	FABP3	−0.826	0.270	−1.095	0.046	

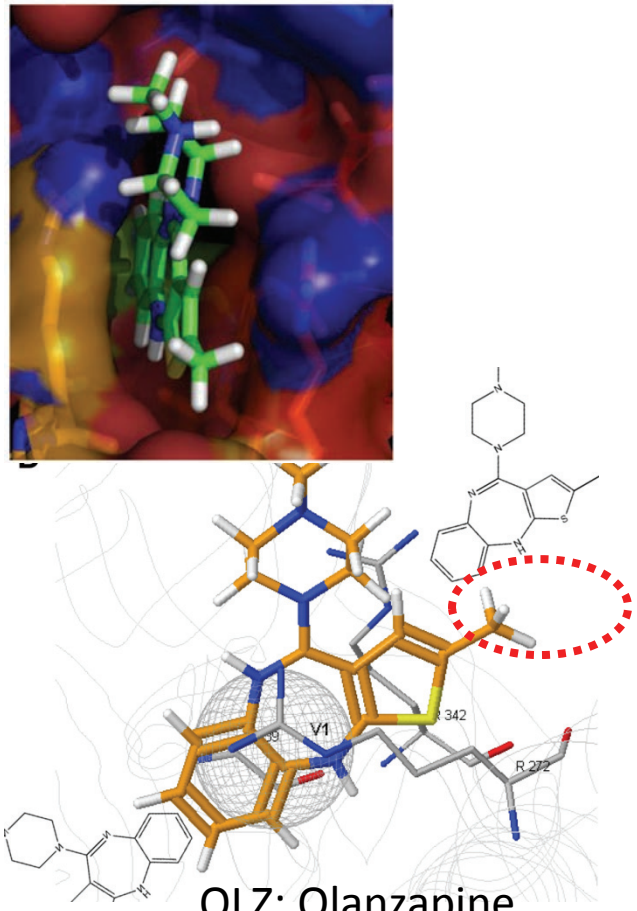
[#]An entry name that ends with a number represents the pocket number of its PDB structure.
*The smaller Z′-score represents a higher theoretical interaction strength.
In the "Role" column, OR and GT indicate oxidoreductases and gluthathione metabolism related proteins, respectively.

Hsp70 protein is the candidate off-target of CLZ not OLZ

PDB ID#	Target Name	Gene Name	Z′ (CLZ)*	Z′ (OLZ)	A-score	p value for CPI
2E8A	Heat shock 70 kDa protein 1	HSPA1A/HSPA1B	−1.381	0.150	−1.531	0.001



CLZ: Clozapine



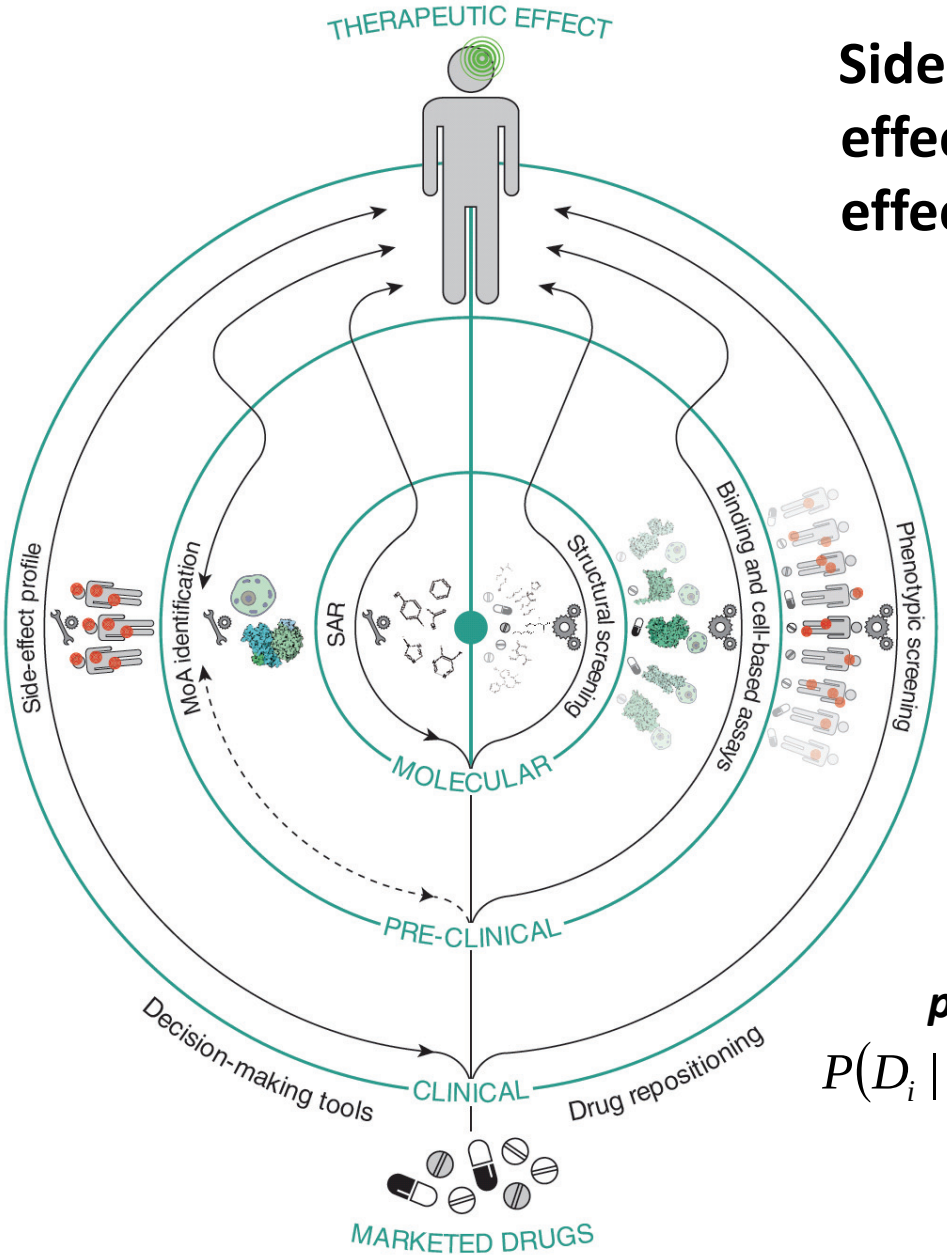
OLZ: Olanzapine

Candidate off-targets prioritized from Multiple Antithesis CPI

PDB ID[a]	Target Name	Gene Name	a	b	c	d	RR	p value	Role
1I10	L-lactate dehydrogenase A chain	LDHA	21	1	18	14	1.697	0.002	OR
2HGS_2	Glutathione synthetase	GSS	24	2	15	13	1.723	0.002	GT
2HRB	Carbonyl reductase NADPH 3	CBR3	17	0	22	15	1.682	0.002	OR
1KBQ	NAD(P)H dehydrogenase quinone 1	NQO1	16	0	23	15	1.652	0.002	OR
1EEM	Glutathione S-transferase omega-1	GSTO1	19	1	20	14	1.615	0.004	GT
1SG0_2	Ribosyldihydronicotinamide dehydrogenase quinone	NQO2	14	0	22	15	1.682	0.005	OR
1G0X	Leukocyte immunoglobulin-like receptor subfamily B member 1	LILRB1	15	0	24	15	1.625	0.005	
2AHE	Chloride intracellular channel protein 4	CLIC4	14	0	24	15	1.625	0.005	
1DIA	Formyltetrahydrofolate synthetase	MTHFD1	14	0	25	15	1.600	0.006	OR
11GS	Glutathione S-transferase P	GSTP1	18	1	21	14	1.579	0.009	GT
1FIE	Coagulation factor XIII A chain	F13A1	12	0	23	15	1.652	0.010	
1Q4O	Serine/threonine-protein kinase PLK1	PLK1	13	0	25	15	1.600	0.011	
1LJR	Glutathione S-transferase theta-2	GSTT2B	8	0	12	14	2.167	0.011	GT
1FPR	Tyrosine-protein phosphatase non-receptor type 6	PTPN6	13	0	26	15	1.577	0.011	
1HSO	Alcohol dehydrogenase 1A	ADH1A	13	0	26	15	1.577	0.011	OR
1IHI_1	Aldo-keto reductase family 1 member C2	AKR1C2	20	2	19	13	1.531	0.014	OR
1SG0_1	Ribosyldihydronicotinamide dehydrogenase quinone	NQO2	16	1	22	14	1.540	0.020	OR
1IHI_2	Aldo-keto reductase family 1 member C2	AKR1C2	16	1	23	14	1.514	0.021	OR

	Agranulocytosis +	Agranulocytosis-
Binding	a	b
Not Binding	c	d

Rationale of Using Pharmacological Effects in Drug Repositioning



Duran-Frigola M. & Aloy P. Genome Med. 2012 4(1):3

Side-effects (SE) and therapeutic effects are clinical phenotypic effects of drug treatment

- They may associate with each other via underlying mechanism

Clinical phenotypic information comes from patients, not animals

Mimics a human phenotypic 'assay'
May have less translational issue

Quantitative Rational

$$\max(P(D_i \mid se_1, se_2, ..., se_m)),\ i \in (|D|)$$

posterior ***prior***

$$P(D_i \mid se_1, se_2, ..., se_m) = \frac{P(se_1, se_2, ..., se_m \mid D_i) P(D_i)}{P(se_1, se_2, ..., se_m)}$$

$$P(se_1, se_2, ..., se_m \mid D_i) = \prod_{j=1}^{m} P(se_j \mid D_i)$$

- *Identification of the disease-side effect associations*

Retrieving side-effect/disease information from drug label and PharmGKB

SIDE EFFECT

Skin:

Allergic skin reactions, e.g., pruritus, erythema, urticaria, vasculitis, and morbilliform or maculopapular eruptions, occur in less than 1% of treated patients. Such mild reactions may develop into serious reactions sometimes progressing to shock. These may be transient and may disappear despite continued use of glimepiride if skin reactions persist, the drug should be discontinued. Although there have been no reports for glimepiride, porphyria cutanea tarda

SE –> drug –> Disease

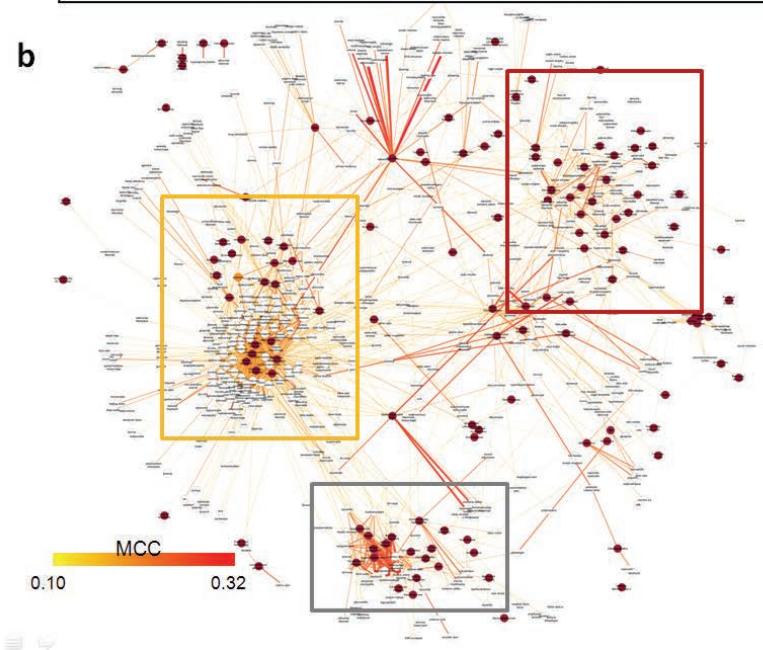
Bork, et al. 2008 – 2010; Altman, et al. 2001-2012

Identification of the disease-side effect associations

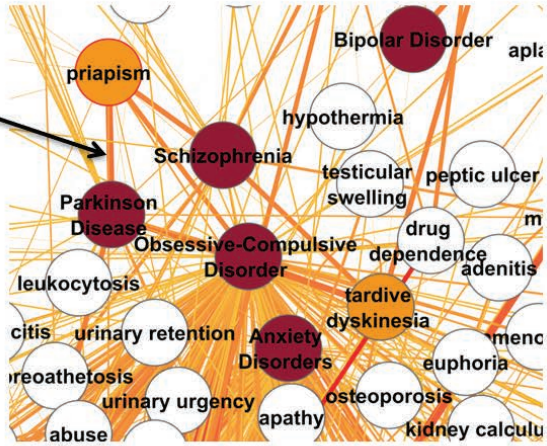
a

	Case: Parkinson +	Ctrl: Parkinson -	
Priapism +	10 (TP)	18 (FP, drugs listing priapism not used to treat PD)	MCC 0.47
Priapism -	4 (FN, drugs treating PD not inducing priapism)	271 (TN)	Two sided p 6.74E-09
	Sensitivity = 10 / (10 + 4) =0.71	Specificity = 271 / (271 + 18)=0.94	

b



c



d

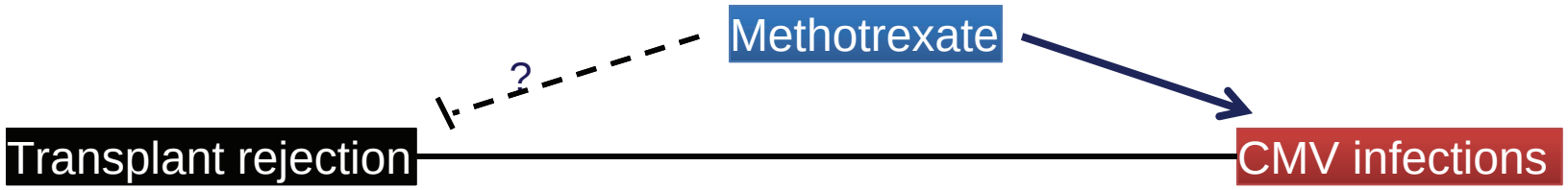
Disease	MCC	sn	sp
Parkinson Disease	0.47	0.71	0.94
Schizophrenia	0.42	0.59	0.94
Obsessive-Compulsive Disorder	0.41	0.86	0.93
Anxiety Disorders	0.30	0.63	0.92

584 side effects; 145 diseases; 3175 informative drug-SE associations

Yang, L. Agarwal, P. PLoS ONE, 2011

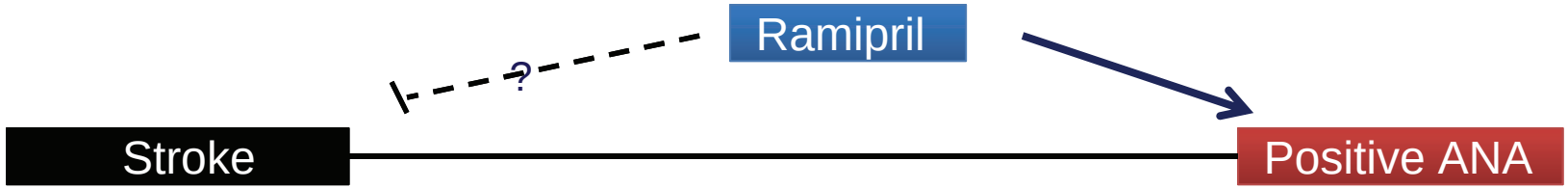
Examples of disease-side effect associations

Disease Class	Disease	Side Effect	MCC	sn	sp	p value	Predictions
Circulation System	Stroke	Positive ANA	0.46	0.47	0.98	1.8E-15	statins, ramipril
Immune System	Transplant rejection	Cytomegalovirus infection	0.75	0.75	0.99	3.5E-06	methotrexate
Metabolite disease	Diabetes Mellitus	Porphyria	0.44	0.50	0.98	8.8E-06	valproic acid, pyrazinamide, naproxen, estradiol
Psychiatric disease	Depressive Disorder	Delusions	0.46	1.00	0.91	1.1E-08	cabergoline, memantine, pergolide
Psychiatric disease	Depressive Disorder	Hyperacusis	0.55	0.88	0.96	9.0E-09	phenytoin, modafinil
Neoplasms	Neoplasms	Constitutional symptoms	0.50	0.56	0.94	2.6E-18	nevirapine



Stroke - *positive Antinuclear Antibodies (ANA)*

Disease Class	Disease	Side Effect	MCC	sn	sp	p value	Predictions
Circulation System	Stroke	Positive ANA	0.46	0.47	0.98	1.8E-15	statins, ramipril



- The SE *positive ANA* is shared by drugs treating Stroke
  - mainly ticlopidine and several angiotensin-converting enzyme (ACE) inhibitors
- Stroke is associated with severe immune suppression
- Drugs that are associated with increasing immune response in terms of *positive ANA* may help stroke patients
- Ramipril lists *positive ANA* as a SE
  - showed a 32% risk reduction for stroke

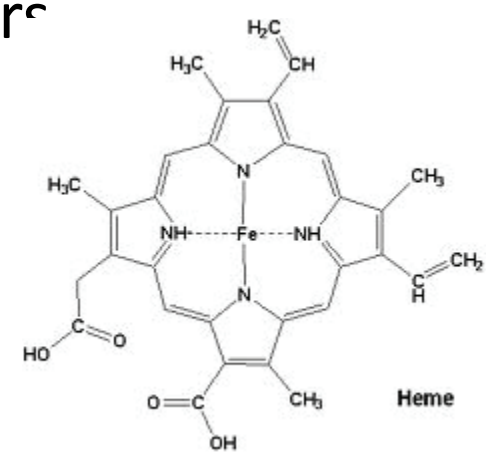
Bosch, J., *et al*. *BMJ* **324**, 699-702 (2002).

Vogelgesang, A. & Dressel, A. *J Neuroimmunol* **231**, 105-110 (2011).

Diabetes - *porphyria*

Disease Class	Disease	Side Effect	MCC	sn	sp	p value	Predictions
Metabolite disease	Diabetes Mellitus	Porphyria	0.44	0.50	0.98	8.8E-06	valproic acid, pyrazinamide, naproxen, estradiol

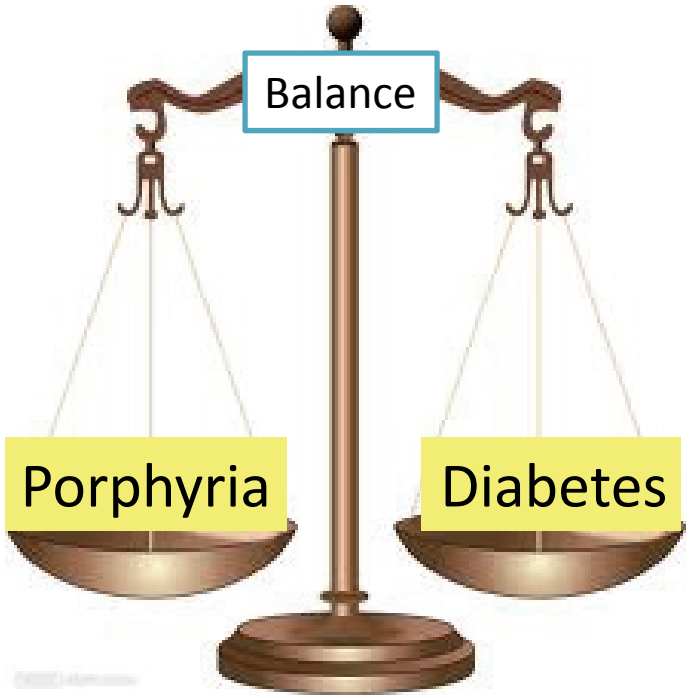
- A metabolic disease characterized by errors in the biosynthetic pathway of HEME
- A partial defect in the activity of the porphobilinogen deaminase during the HEME synthesis
- Carbohydrates modulate the HEME synthesis
- *Porphyria* is hereditary – this 'biomarker' only works on these knock-out people



Folke Lithner. *Diabetes Care* **25,** 797-798 (2002)

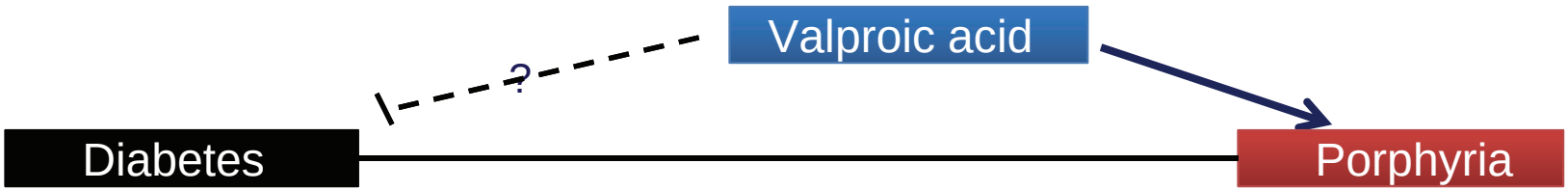
The balancing between SE and disease

- Diabetic drug worsen *porphyria*
- *Porphyria* resolved after developing diabetes
  - 328 patients with *porphyria*, the 16 patients that developed diabetes all had their *porphyria* symptoms resolved
  - 16 “knock-out” people mimic **a phenotypic screening for diabetic drug**



Andersson, C., Bylesjo, I. & Lithner, F. *J Intern Med* **245**, 193-197 (1999).

Diabetes - *porphyria*



- Drugs list *porphyria* as a SE but are not indicated for diabetes could be tested for treating diabetes
  - Valproic acid is an anticonvulsant and a recent study found it effective in lowering blood glucose levels in mice
  - In mice, pain killer naproxen is used as a tool to delay or prevent the development of type II diabetes from a pre-diabetic condition

Kendig, E.L., Schneider, S.N., Clegg, D.J., Genter, M.B. & Shertzer, H.G.. *Biochem Pharmacol* **76**, 216-224 (2008).
Terasmaa, A., *et al.*. *J Physiol Biochem* (2011)

- *Drug Repositioning based on Side Effects* (***DRoSEf***) *for marketed drugs*

Repositioning using single SE feature

	Case: Parkinson +	Ctrl: Parkinson -
Priapism +	10 (TP)	18 (FP, drugs listing priapism not used to treat PD)
Priapism -	4 (FN, drugs treating PD not inducing priapism)	271 (TN)
	Sensitivity = 10 / (10 + 4) =0.71	Specificity = 271 / (271 + 18)=0.94

- 27% (16346) of the new drugs-disease association suggested by DRoSEf have at least one entry in PubMed
- 44194 repositioning opportunities for marketed drugs

AUCs of 10-fold cross validations across 145 diseases using multiple SE features

Disease	AUC	Disease	AUC
Amyotrophic Lateral Sclerosis	1	Influenza, Human	0.997
Anemia	1	Leukemia, Lymphocytic, Chronic, B-Cell	1
Arthritis	1	Liver Neoplasms	1
Asthma	0.959	Migraine without Aura	1
Cough	0.998	Myopathy, Central Core	1
Dementia	1	Non-small cell lung cancer	0.986
Diabetic Nephropathies	1	normal tension glaucoma	1
Diarrhea	0.982	Osteonecrosis	0.993
Esophogeal Neoplasms	0.983	Osteoporosis, Postmenopausal	1
estrogen-dependent carcinogenesis	1	Pain	0.983
Gastroesophageal Reflux	0.997	Parkinson Disease	0.959
Glioblastoma	1	Peripheral Nervous System Diseases	0.957
Glomerulosclerosis	0.997	Psoriasis	0.962
Heart Diseases	1	Rectal Neoplasms	0.983
Hyperlipidemias	0.981	Rheumatic Diseases	0.994

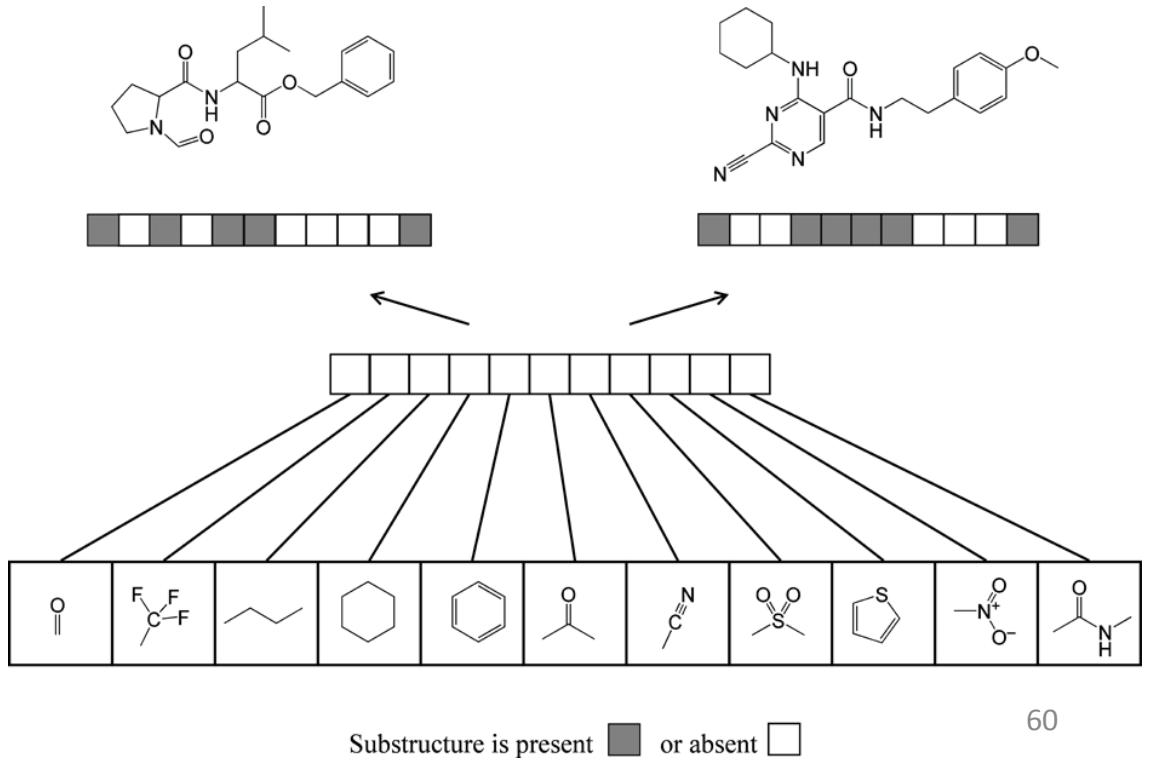
…

92% of the AUCs were above 0.8

- *DRoSEf for clinical molecules*

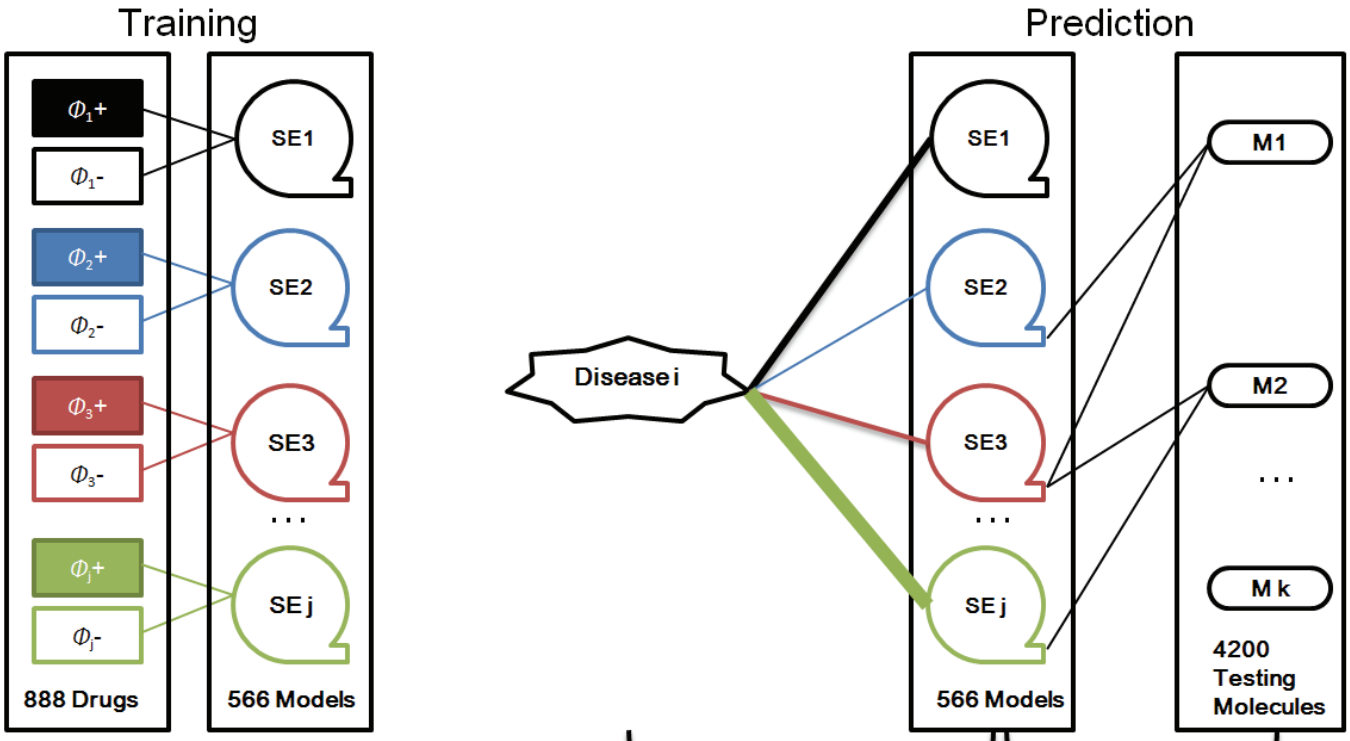
Quantitative Structure Activity Relationship (QSAR) Modeling

- Drug-like properties
  - Octanol-water partition coefficient (logP)
  - Hydrogen bond donors
  - Hydrogen bond acceptors
  - Molecular Mass
- Structural Signature



DRoSEf for clinical molecules

- 4,200 clinical molecules that are indicated for at least one of 101 diseases



$$DS_i = [ds_{i1}, ds_{i2}, \ldots, ds_{ij}], \quad j \in [1,566], \quad i \in [1,101]$$

$$SM_k = [sm_{1k}, sm_{2k}, \ldots, sm_{jk}], \quad j \in [1,566], \quad k \in [1,4200]$$

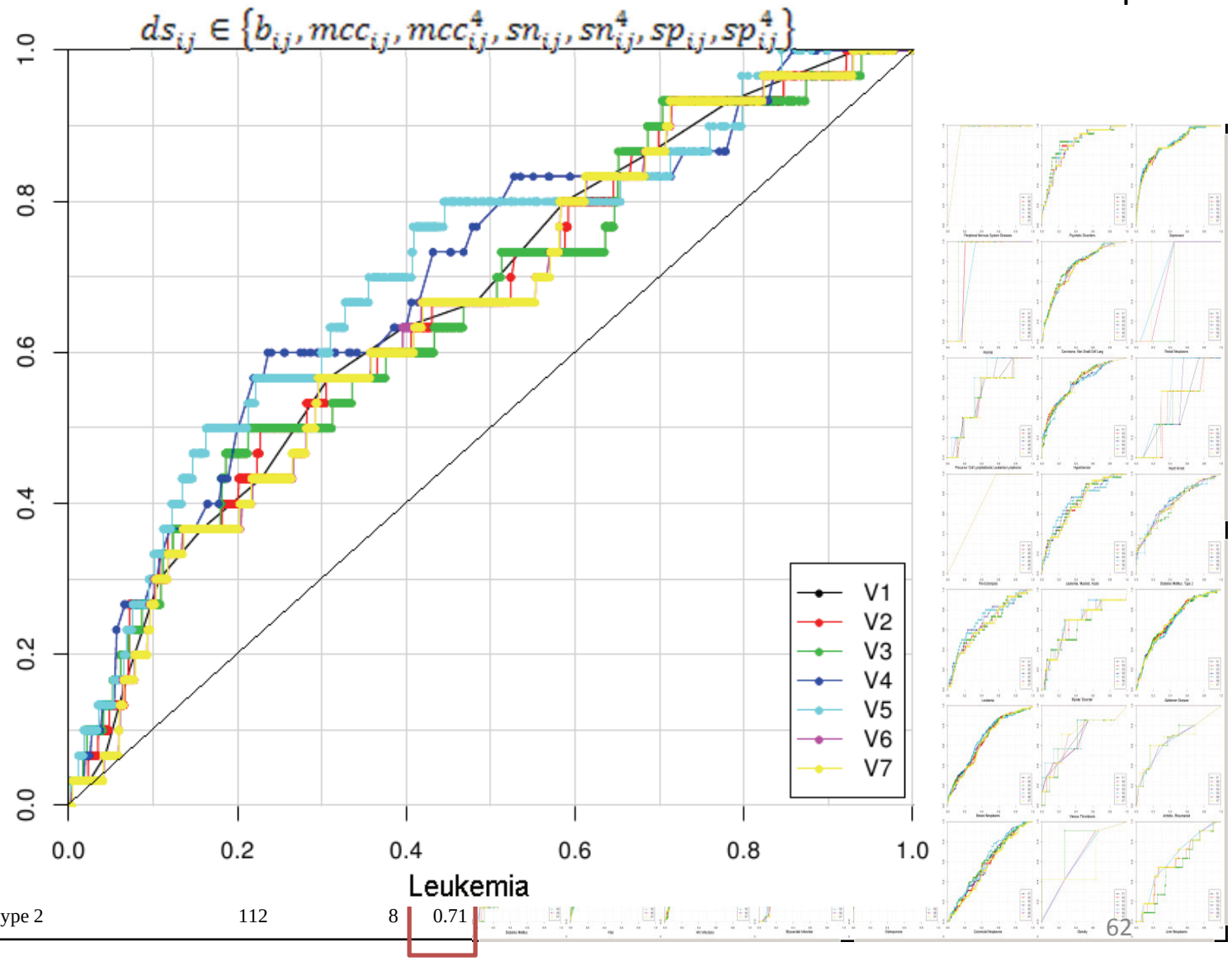
$$ds_{ij} \in \{b_{ij}, mcc_{ij}, mcc_{ij}^4, sn_{ij}, sn_{ij}^4, sp_{ij}, sp_{ij}^4\}$$

$$\Theta_{ik} = \sum_{j=1}^{566} ds_{ij}\, sm_{jk} \rightarrow \Theta_{i2} > \Theta_{i1}$$

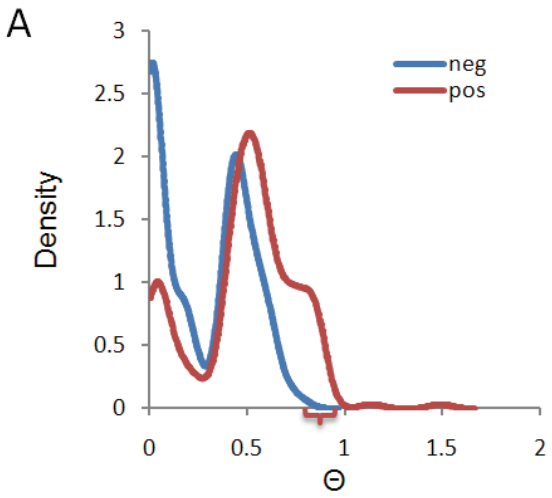
Prediction results for clinical molecules

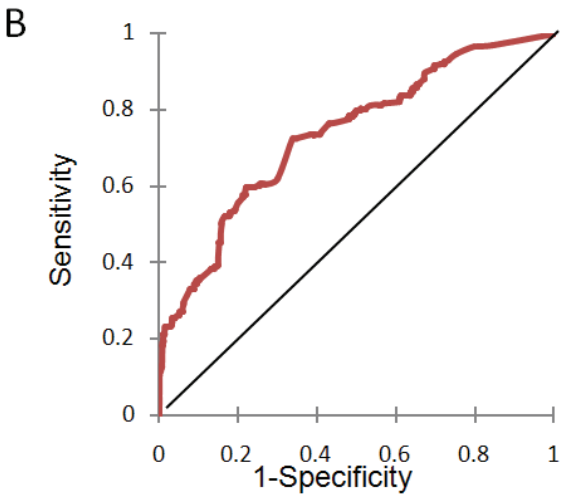
4200 Molecules * 101 Disease Endpoints

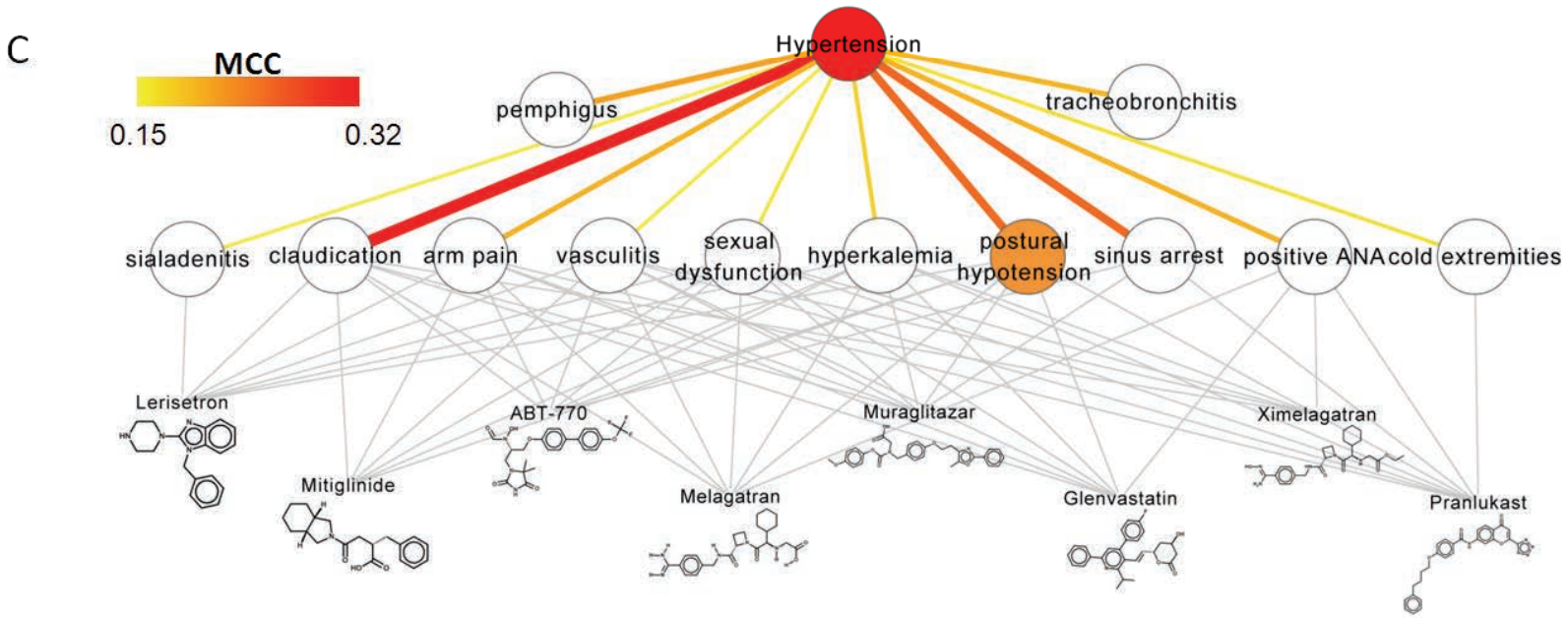
Disease category	Disease[a]			
Neuropsychiatric	Depression			
	Depressive Disorder			
	Schizophrenia			
	Depressive Disorder			
	Anxiety Disorders			
Neoplasms	Stomach Neoplasms			
	Carcinoma, Non-Sm			
	Lung Neoplasms			
	Neoplasms			
	Lymphoma			
	Leukemia			
	Head and Neck Neo			
Others	Hypertension			
	Diabetes Mellitus, Type 2	112	8	0.71



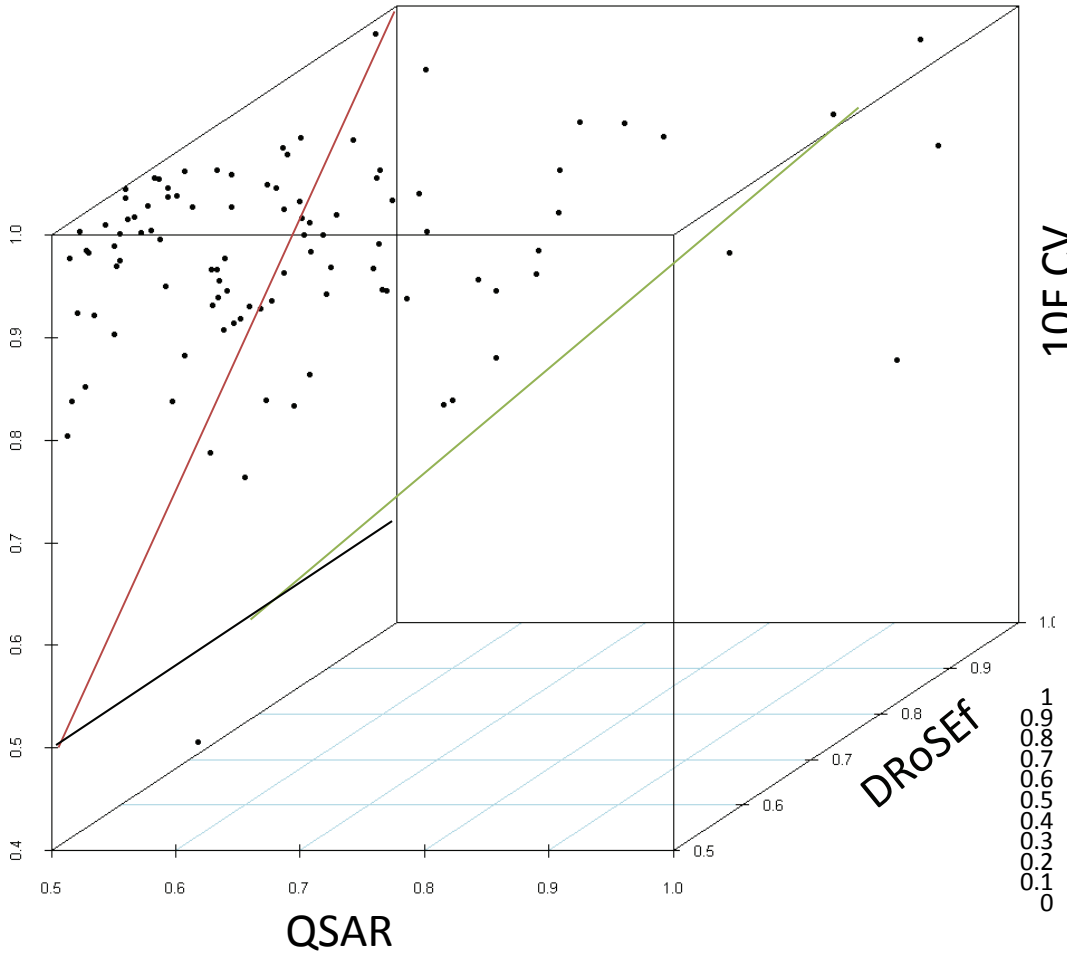
Case Study: Predict drugs' repositioning potential for hypertension

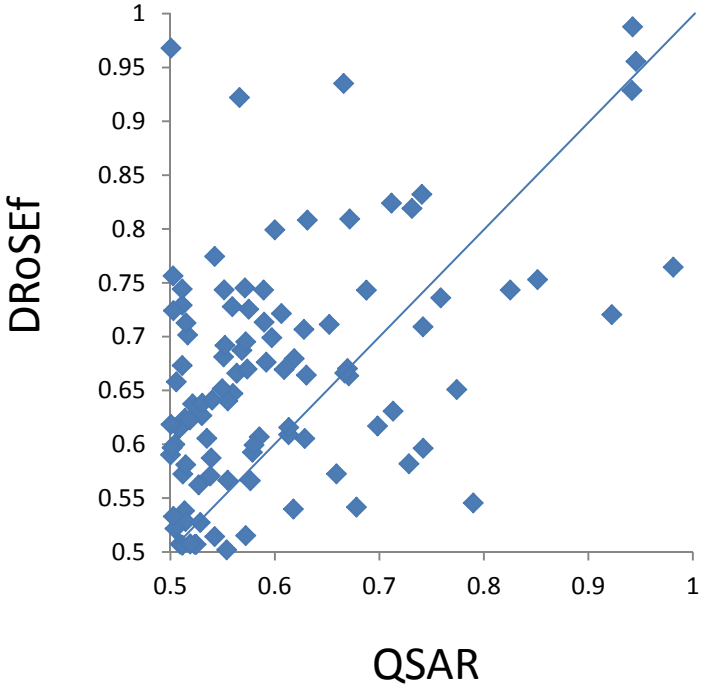


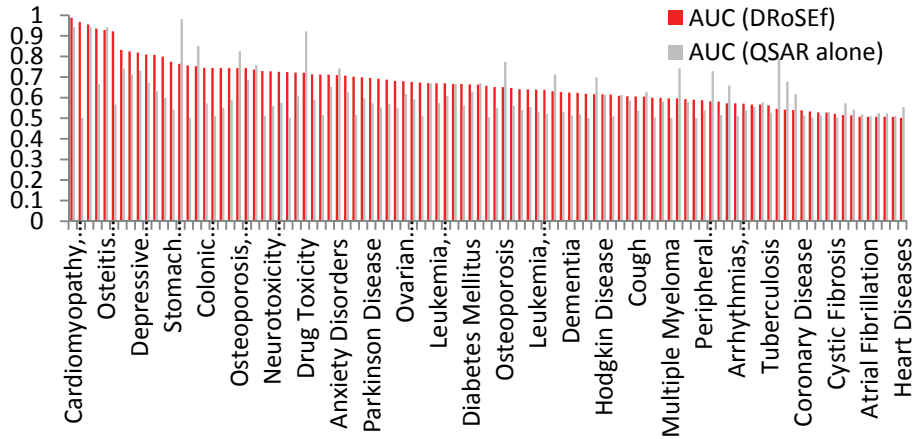




DRoSEf vs. QSAR alone







AUC comparison of different methods

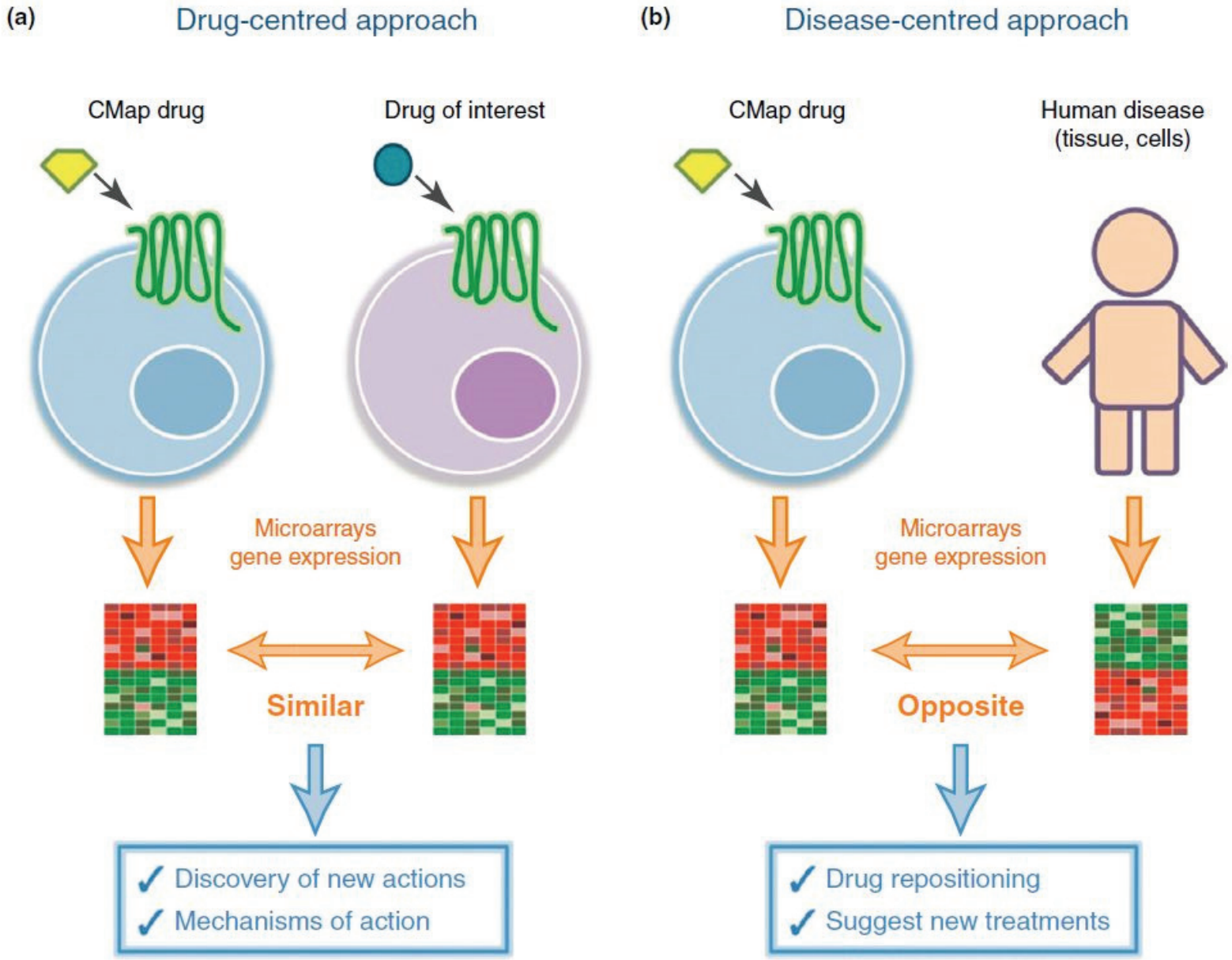
Discussion

- Frequency information of the SE is not considered
- High frequency may not be informative
  - diarrhea, dizziness, Vomiting, Nausea
- Rare SE may be informative
  - Although some side effects are rare, but they can also
    - shed lights on the association between side effects and new indication
    - drugs are screened on the models using ‘knock-out’ animals
  - Just like rare mutations identified from the GWAS of common diseases could
    - shed lights on the pathogenesis of common diseases

Take-home message

- Drug indication can be suggested based on clinical side-effects
  - Mimicking a phenotypic efficacy screen on some "knock-out" human
- DRoSEf may also suggest the neglected pathogenesis of disease
  - For example, studying *porphyria* may help discover potential new mode of action for diabetes therapy

The Connectivity Map concept

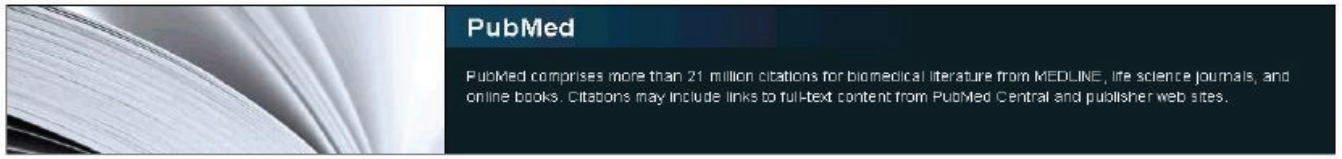


Montero-Melendez T and Perretti M. Drug Discov Today. 2013 Dec 6. pii: S1359-6446(13)00418-2.

What we are interested in

- Identify Associations among diseases (Y) and:
  - Chemical-protein interactome profile of drug
  - Gene expression profile after drug treatment
  - Side Effect of drug [e.g., Parkinson Disease and Priapism]
- Prediction
  - How to use the Pharmaco-information (X_1, X_2, …, X_n) to predict indication of it
  - Prediction results should be understood by biologists

Drug-disease Associations from various resources



facebook

twitter

PLOS one

PLOS BIOLOGY

PLOS MEDICINE

PLOS COMPUTATIONAL BIOLOGY

PLOS GENETICS

Challenges

- Many factors need to be considered for practical use of drug repositioining
  - the unmet medical need for the disease
  - the CNS penetration of the molecule
  - therapeutic effect is significant compared with the active comparators
  - the previous therapeutic effect could now become a side effect
  - …
- These issues may be controlled via choosing a suitable formulation, dose, and the sub population

Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Data Sources for Data Mining Applications
- Case study: Personalized Medicine
- Case Study: Drug Repositioning
- Future Challenges and Summary

Thank you! | Questions?

Ping Zhang: pzhang@us.ibm.com

Lun Yang: Lun.Yang@gmail.com