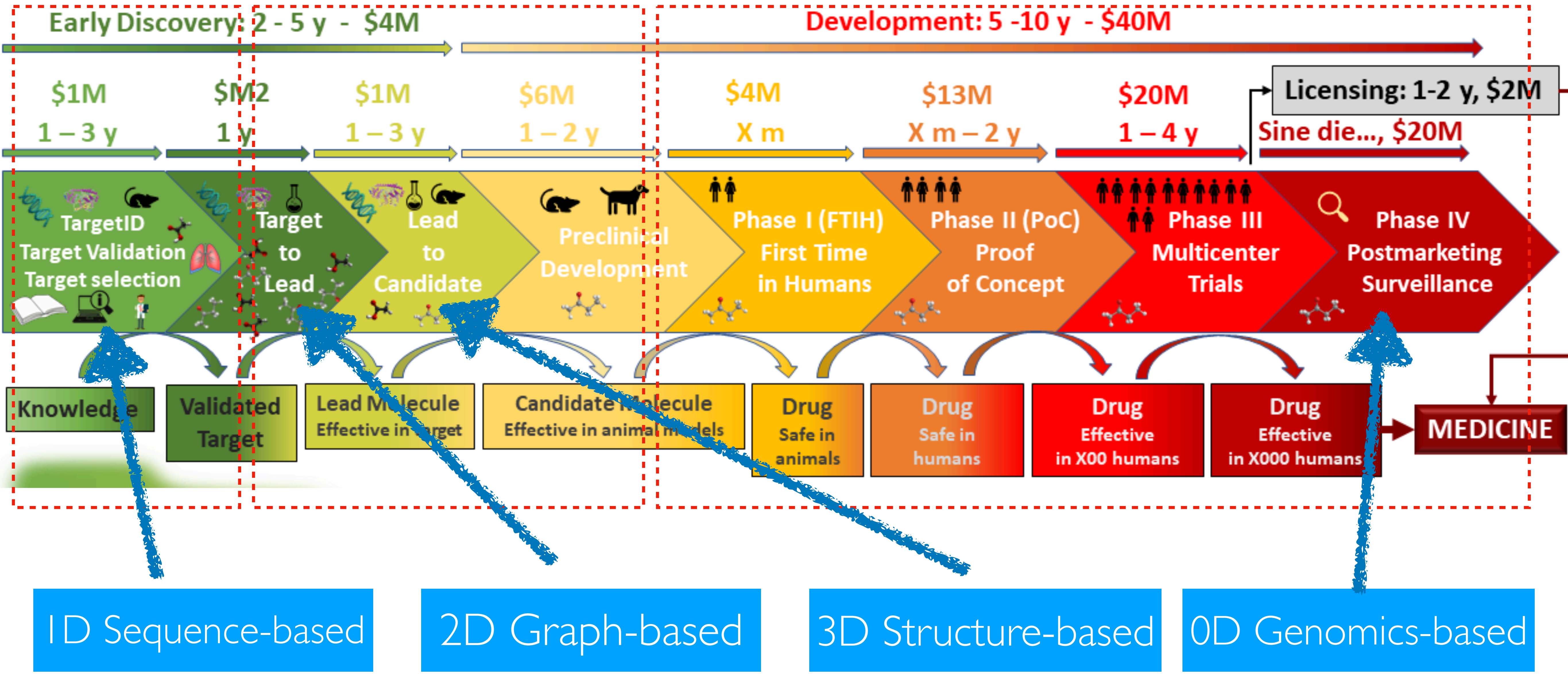
Precision medicine

Sheng Wang

Recap of previous lectures



Early Discovery: 2 - 5 y - $4M

Development: 5 -10 y - $40M

$1M 1 – 3 y

$M2 1 y

$1M 1 – 3 y

$6M 1 – 2 y

$4M X m

$13M X m – 2 y

$20M 1 – 4 y

Licensing: 1-2 y, $2M

Sine die..., $20M

TargetID Target Validation Target selection

Target to Lead

Lead to Candidate

Preclinical Development

Phase I (FTIH) First Time in Humans

Phase II (PoC) Proof of Concept

Phase III Multicenter Trials

Phase IV Postmarketing Surveillance

Knowledge

Validated Target

Lead Molecule Effective in target

Candidate Molecule Effective in animal models

Drug Safe in animals

Drug Safe in humans

Drug Effective in X00 humans

Drug Effective in X000 humans

MEDICINE

1D Sequence-based

2D Graph-based

3D Structure-based

0D Genomics-based

Sequence: understand target function using protein sequence. NLP to find targets (word sequence).

Graph: generate compound graph 2D structure (deep generative model)

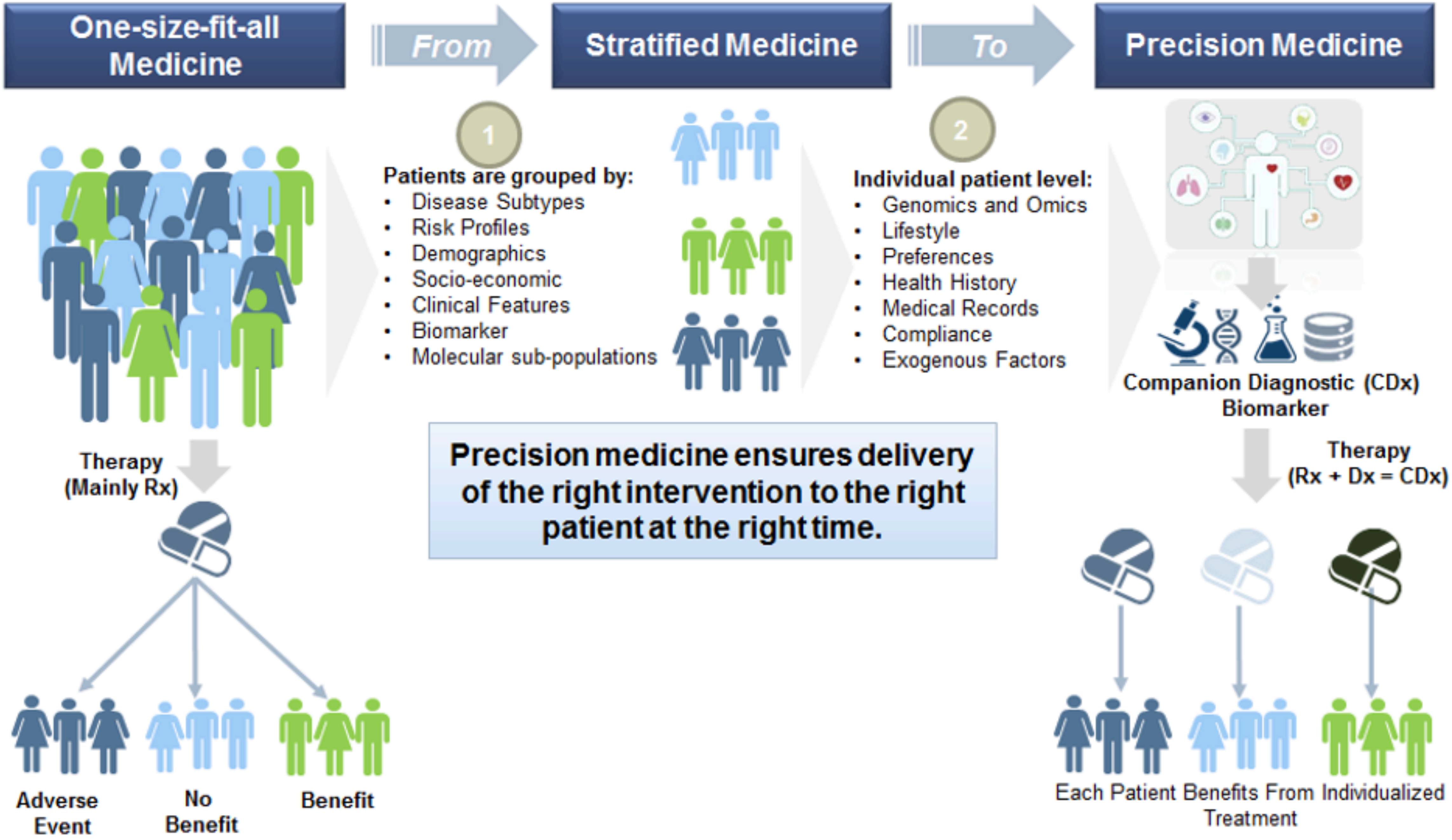
Structure: modify structure according to 3D structure (geometric deep learning)

Genomics: side effects, personalized efficacy, repurposing, etc. (multi-modality)

- how to reuse an old drug

Precision medicine:

the right patient, the right drug, the right time, the right dose



Patients are grouped by:

- Disease Subtypes
- Risk Profiles
- Demographics
- Socio-economic
- Clinical Features
- Biomarker
- Molecular sub-populations

Individual patient level:

- Genomics and Omics
- Lifestyle
- Preferences
- Health History
- Medical Records
- Compliance
- Exogenous Factors

Precision medicine ensures delivery of the right intervention to the right patient at the right time.

Frost and Sullivan: new paradigm shift in treatment.

We don’t have so many “drugs”

- Discovery new drug?
  - Often not in the scope of precision medicine
  - New patient cannot wait for a new drug
- Drug repurposing
  - Drug A, which is used to treat disease X, is later used to treat disease Y
  - Well-documented side effects and less restriction from FDA
- Drug combination
  - Drug A is not effective. Drug B is not effective. Durg A and B used together is effective.
- Personalized dosage
  - Widely used in clinics. Use genomics data to determine dosage (regression).

Drug repurposing

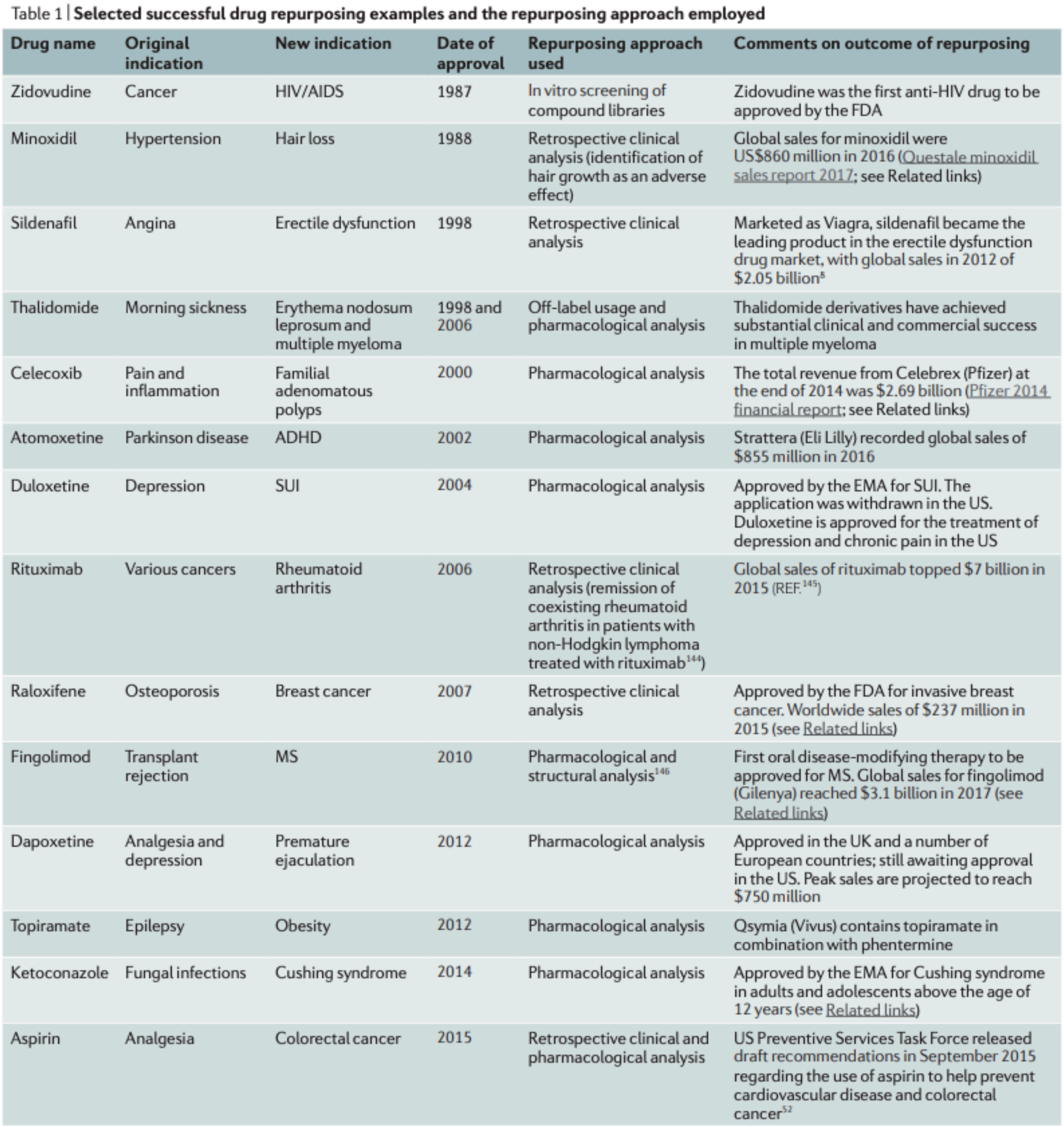
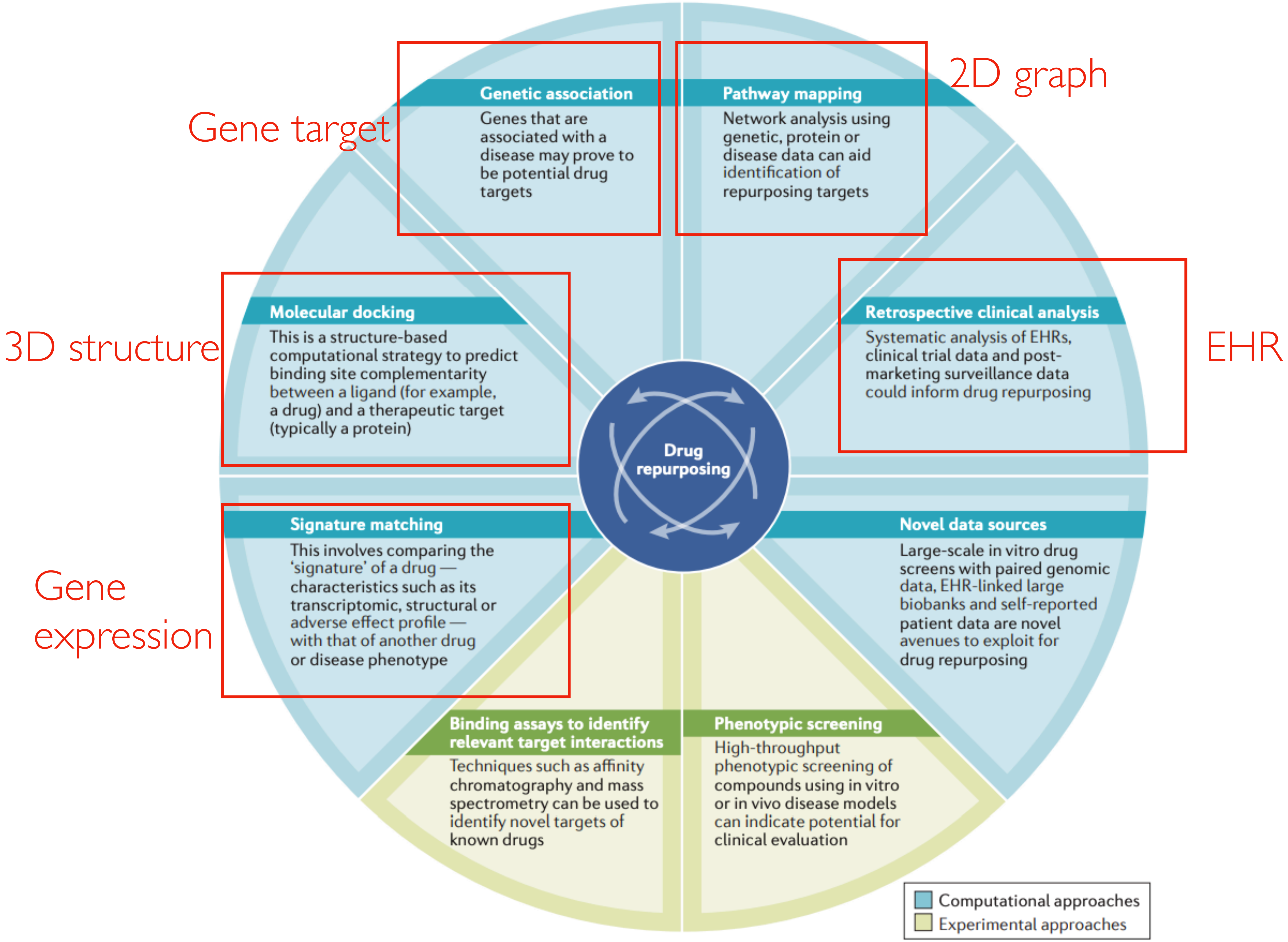


Table 1 | **Selected successful drug repurposing examples and the repurposing approach employed**

Drug name	Original indication	New indication	Date of approval	Repurposing approach used	Comments on outcome of repurposing
Zidovudine	Cancer	HIV/AIDS	1987	In vitro screening of compound libraries	Zidovudine was the first anti-HIV drug to be approved by the FDA
Minoxidil	Hypertension	Hair loss	1988	Retrospective clinical analysis (identification of hair growth as an adverse effect)	Global sales for minoxidil were US$860 million in 2016 (Questale minoxidil sales report 2017; see Related links)
Sildenafil	Angina	Erectile dysfunction	1998	Retrospective clinical analysis	Marketed as Viagra, sildenafil became the leading product in the erectile dysfunction drug market, with global sales in 2012 of $2.05 billion[8]
Thalidomide	Morning sickness	Erythema nodosum leprosum and multiple myeloma	1998 and 2006	Off-label usage and pharmacological analysis	Thalidomide derivatives have achieved substantial clinical and commercial success in multiple myeloma
Celecoxib	Pain and inflammation	Familial adenomatous polyps	2000	Pharmacological analysis	The total revenue from Celebrex (Pfizer) at the end of 2014 was $2.69 billion (Pfizer 2014 financial report; see Related links)
Atomoxetine	Parkinson disease	ADHD	2002	Pharmacological analysis	Strattera (Eli Lilly) recorded global sales of $855 million in 2016
Duloxetine	Depression	SUI	2004	Pharmacological analysis	Approved by the EMA for SUI. The application was withdrawn in the US. Duloxetine is approved for the treatment of depression and chronic pain in the US
Rituximab	Various cancers	Rheumatoid arthritis	2006	Retrospective clinical analysis (remission of coexisting rheumatoid arthritis in patients with non-Hodgkin lymphoma treated with rituximab[144])	Global sales of rituximab topped $7 billion in 2015 (REF.[145])
Raloxifene	Osteoporosis	Breast cancer	2007	Retrospective clinical analysis	Approved by the FDA for invasive breast cancer. Worldwide sales of $237 million in 2015 (see Related links)
Fingolimod	Transplant rejection	MS	2010	Pharmacological and structural analysis[146]	First oral disease-modifying therapy to be approved for MS. Global sales for fingolimod (Gilenya) reached $3.1 billion in 2017 (see Related links)
Dapoxetine	Analgesia and depression	Premature ejaculation	2012	Pharmacological analysis	Approved in the UK and a number of European countries; still awaiting approval in the US. Peak sales are projected to reach $750 million
Topiramate	Epilepsy	Obesity	2012	Pharmacological analysis	Qsymia (Vivus) contains topiramate in combination with phentermine
Ketoconazole	Fungal infections	Cushing syndrome	2014	Pharmacological analysis	Approved by the EMA for Cushing syndrome in adults and adolescents above the age of 12 years (see Related links)
Aspirin	Analgesia	Colorectal cancer	2015	Retrospective clinical and pharmacological analysis	US Preventive Services Task Force released draft recommendations in September 2015 regarding the use of aspirin to help prevent cardiovascular disease and colorectal cancer[52]

Pushpakom et al. Drug repurposing: progress, challenges and recommendations

Approaches used in drug repurposing



Gene target

Genetic association

Genes that are associated with a disease may prove to be potential drug targets

Pathway mapping

Network analysis using genetic, protein or disease data can aid identification of repurposing targets

2D graph

3D structure

Molecular docking

This is a structure-based computational strategy to predict binding site complementarity between a ligand (for example, a drug) and a therapeutic target (typically a protein)

Retrospective clinical analysis

Systematic analysis of EHRs, clinical trial data and post-marketing surveillance data could inform drug repurposing

EHR

Drug repurposing

Gene expression

Signature matching

This involves comparing the 'signature' of a drug — characteristics such as its transcriptomic, structural or adverse effect profile — with that of another drug or disease phenotype

Novel data sources

Large-scale in vitro drug screens with paired genomic data, EHR-linked large biobanks and self-reported patient data are novel avenues to exploit for drug repurposing

Binding assays to identify relevant target interactions

Techniques such as affinity chromatography and mass spectrometry can be used to identify novel targets of known drugs

Phenotypic screening

High-throughput phenotypic screening of compounds using in vitro or in vivo disease models can indicate potential for clinical evaluation

🟦 Computational approaches
🟨 Experimental approaches

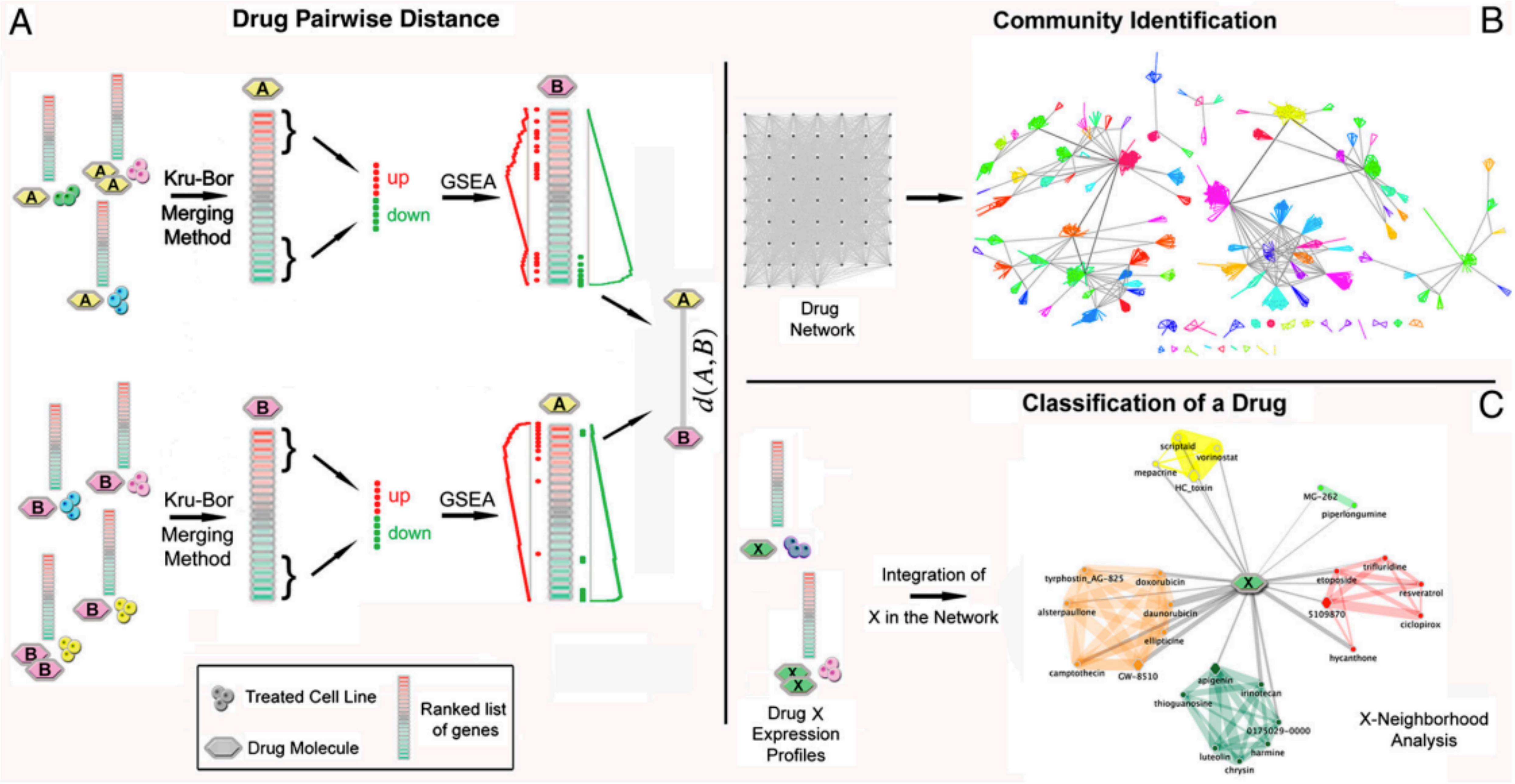
Drug repurposing strategy

- Drug-based
  - If drug A can cure disease X and is similar to drug B, then B might be also treat X
- Disease-based
  - If disease X and Y have similar profiles and indications, and drug R can cure X, then R can also cure Y.

Jarada et al. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions

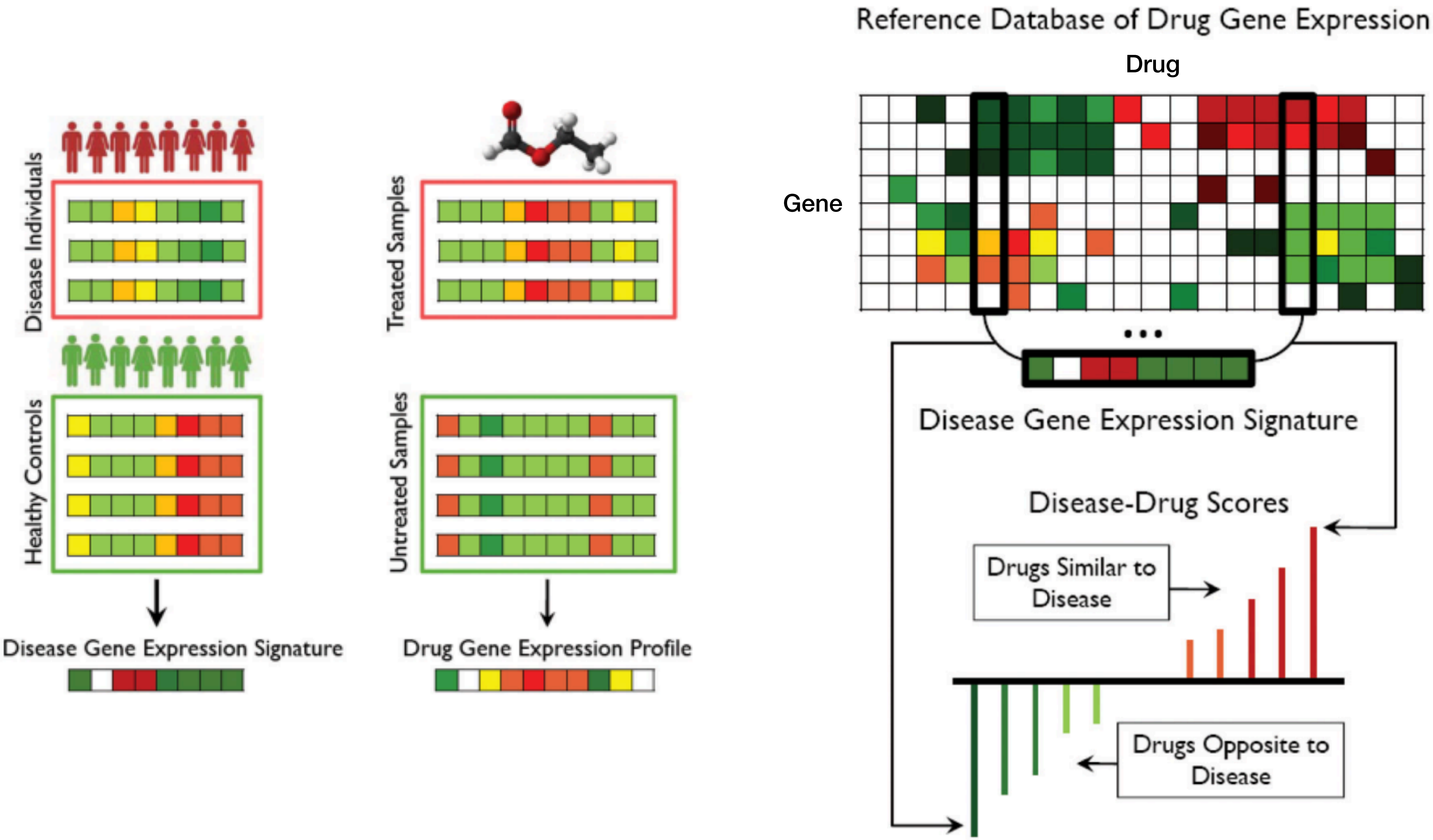
Use gene expression after treatment

Drugs target on similar proteins or have similar Mode of Actions have similar (after treatment) expression.



Iorio et al. Discovery of drug mode of action and drug repositioning from transcriptional responses

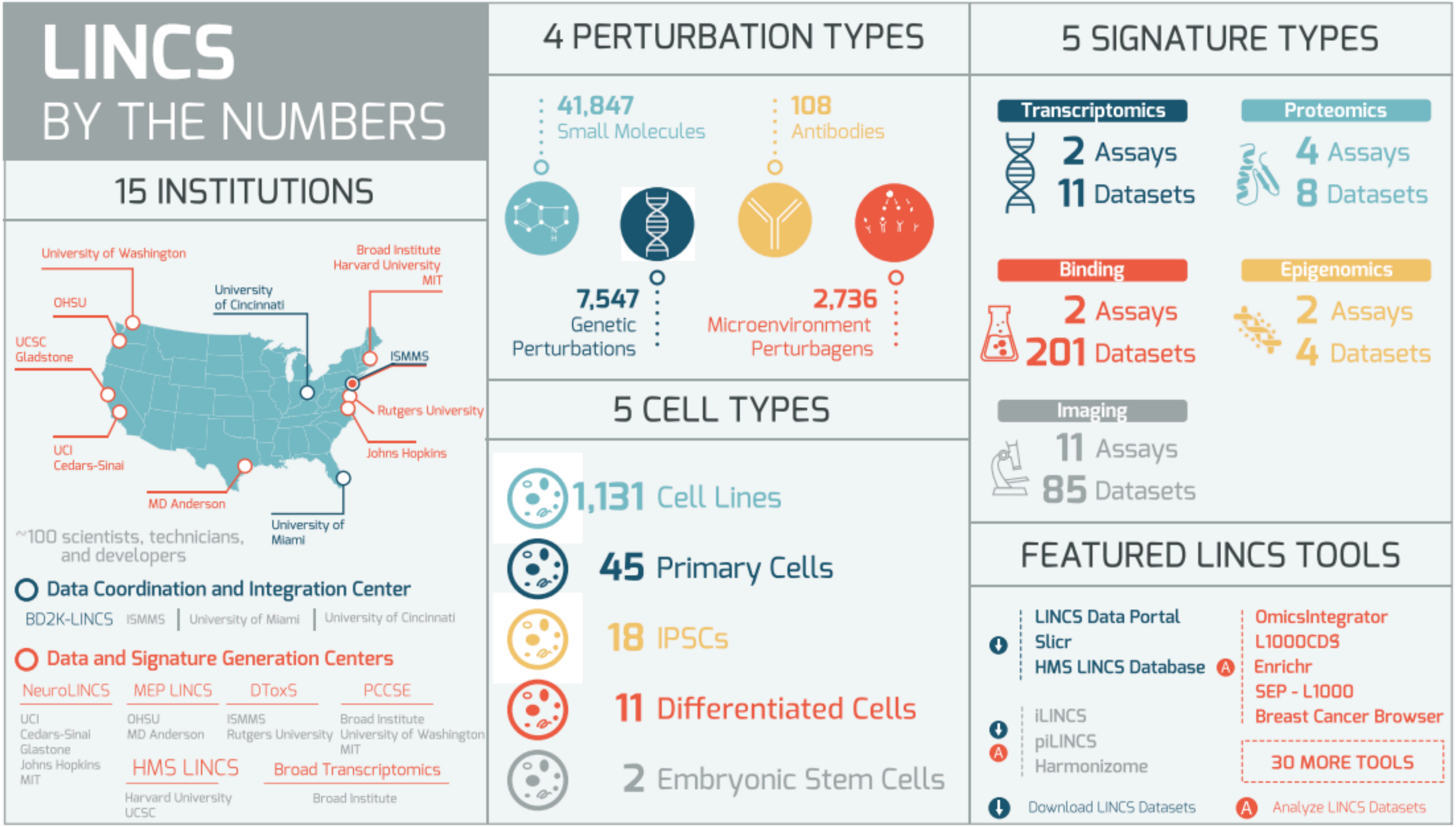
Compare disease expression and drug expression


Disease Individuals
Healthy Controls
Disease Gene Expression Signature
Treated Samples
Untreated Samples
Drug Gene Expression Profile
Reference Database of Drug Gene Expression
Drug
Gene
...
Disease Gene Expression Signature
Disease-Drug Scores
Drugs Similar to Disease
Drugs Opposite to Disease

Sirota et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data

Expression-based drug repurposing

- People realized that the performance (accuracy, coverage) depends on the data, rather than the model
- How about we just generate the expression of X drugs on Y tissues
  - LINCS: Library of Integrated Network-based Cellular Signatures
  - 15 institutions, >1000 cell lines, >5000 drugs, 1000 genes
  - 1.3 million after treatment gene expression vectors
  - cMAP: 3 cell lines, but 20k genes



LINCS BY THE NUMBERS

15 INSTITUTIONS

~100 scientists, technicians, and developers

○ Data Coordination and Integration Center

BD2K-LINCS: ISMMS | University of Miami | University of Cincinnati

○ Data and Signature Generation Centers

NeuroLINCS	MEP LINCS	DToxS	PCCSE
UCI Cedars-Sinai Glastone Johns Hopkins MIT	OHSU MD Anderson	ISMMS Rutgers University	Broad Institute University of Washington MIT

HMS LINCS	Broad Transcriptomics
Harvard University UCSC	Broad Institute

4 PERTURBATION TYPES

41,847 Small Molecules

108 Antibodies

7,547 Genetic Perturbations

2,736 Microenvironment Perturbagens

5 CELL TYPES

1,131 Cell Lines

45 Primary Cells

18 IPSCs

11 Differentiated Cells

2 Embryonic Stem Cells

5 SIGNATURE TYPES

	Assays	Datasets
Transcriptomics	2	11
Proteomics	4	8
Binding	2	201
Epigenomics	2	4
Imaging	11	85

FEATURED LINCS TOOLS

- LINCS Data Portal
- Slicr
- HMS LINCS Database Ⓐ
- iLINCS
- piLINCS
- Harmonizome
- OmicsIntegrator
- L1000CDS
- Enrichr
- SEP - L1000
- Breast Cancer Browser

30 MORE TOOLS

⬇ Download LINCS Datasets Ⓐ Analyze LINCS Datasets

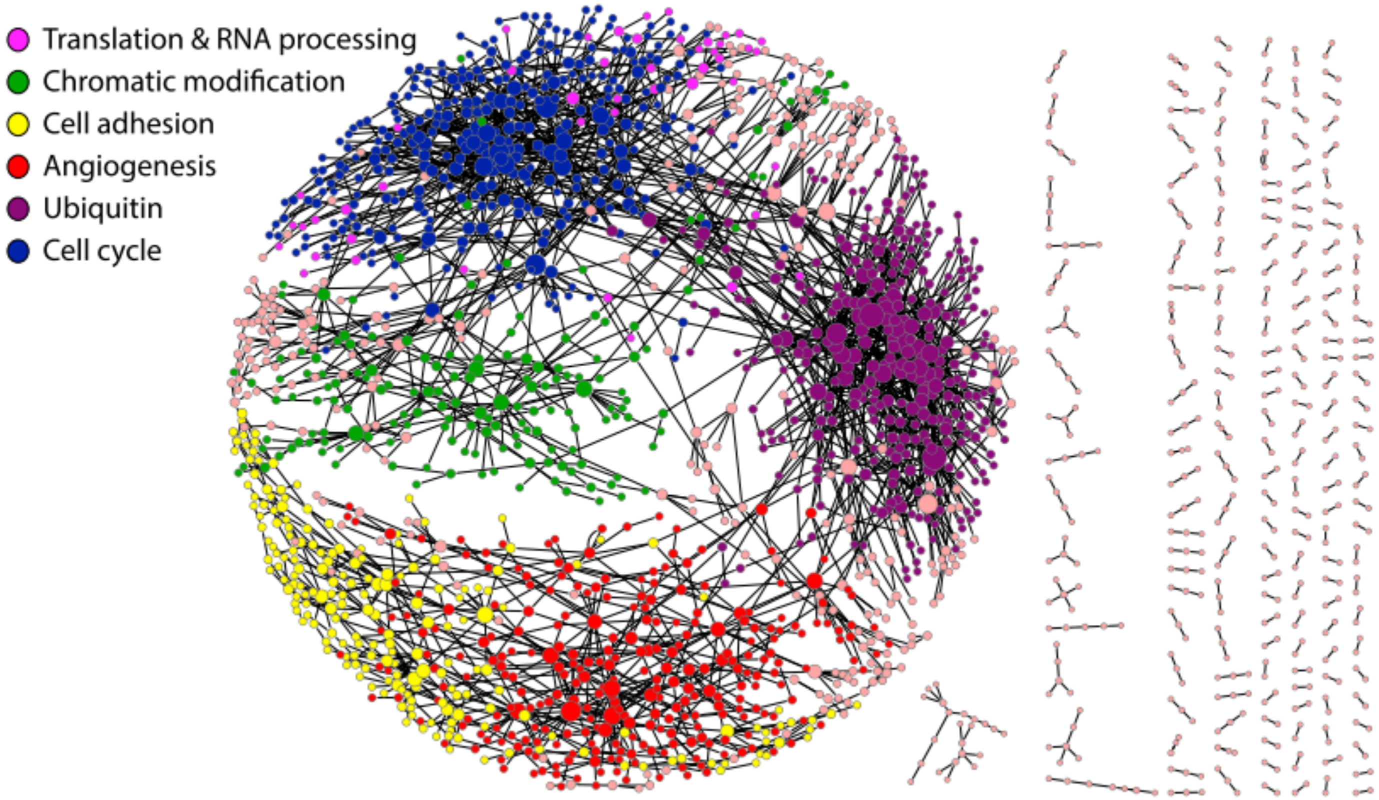
- **Adverse drug reaction prediction** (Wang et al. Drug-induced adverse events prediction with the LINCS L1000 data)
- **Drug target identification** (Xia et al. Target Predictions using LINCS Data)
- **Expression signature comparison** (Xiao et al. SigMat: A Classification Scheme for Gene Signature Matching)
- **Drug response prediction** (Lu et al. Drug-induced cell viability prediction from LINCS-L1000 through WRFEN-XGBoost algorithm)

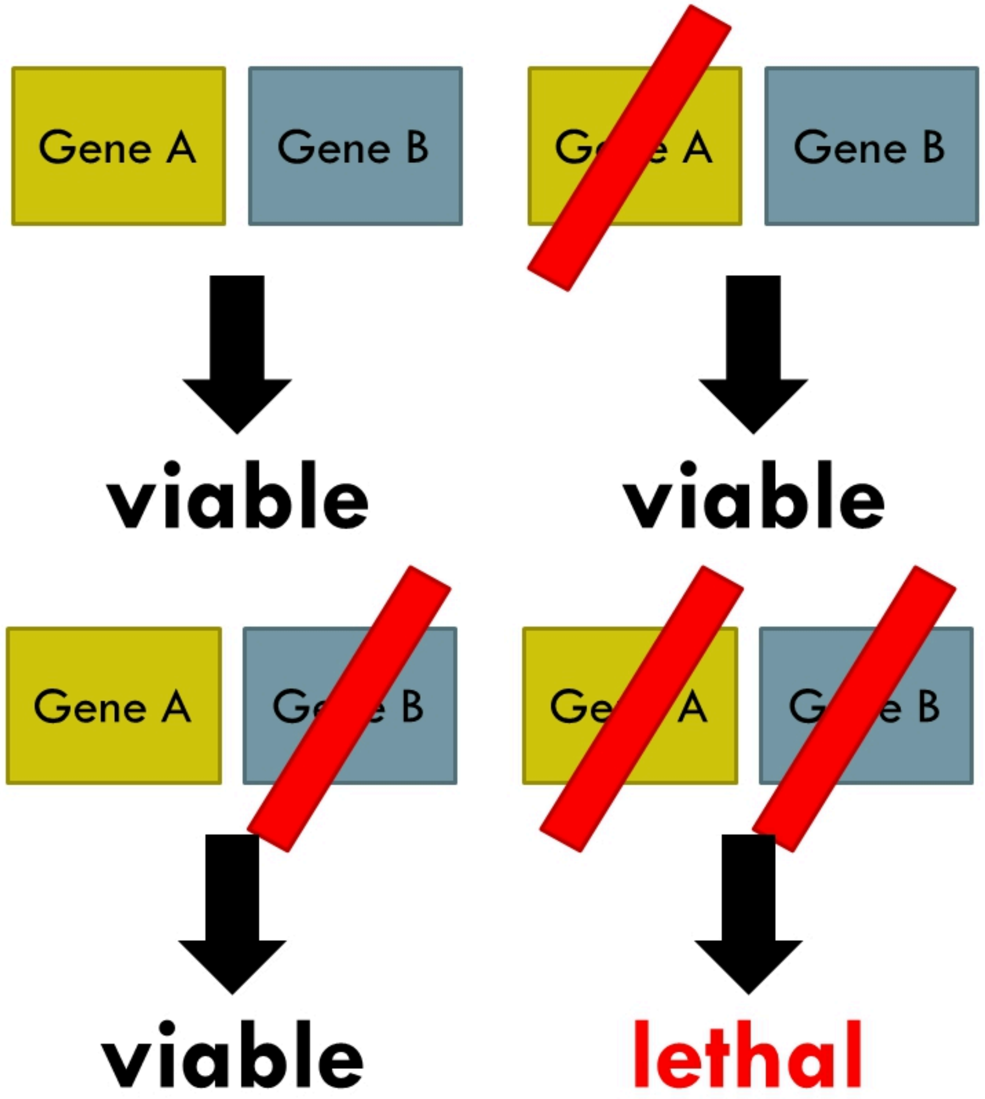
Keenan et al. The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations

We don't have so many "drugs"

- Discovery new drug?
  - Often not in the scope of precision medicine
  - New patient cannot wait for a new drug
- Drug repurposing
  - Drug A, which is used to treat disease X, is later used to treat disease Y
  - Well-documented side effects and less restriction from FDA
- Drug combination
  - Drug A is not effective. Drug B is not effective. Durg A and B used together is effective.
- Personalized dosage
  - Widely used in clinics. Use genomics data to determine dosage (regression).

Synthetic lethality: Gene A **OR** Gene B





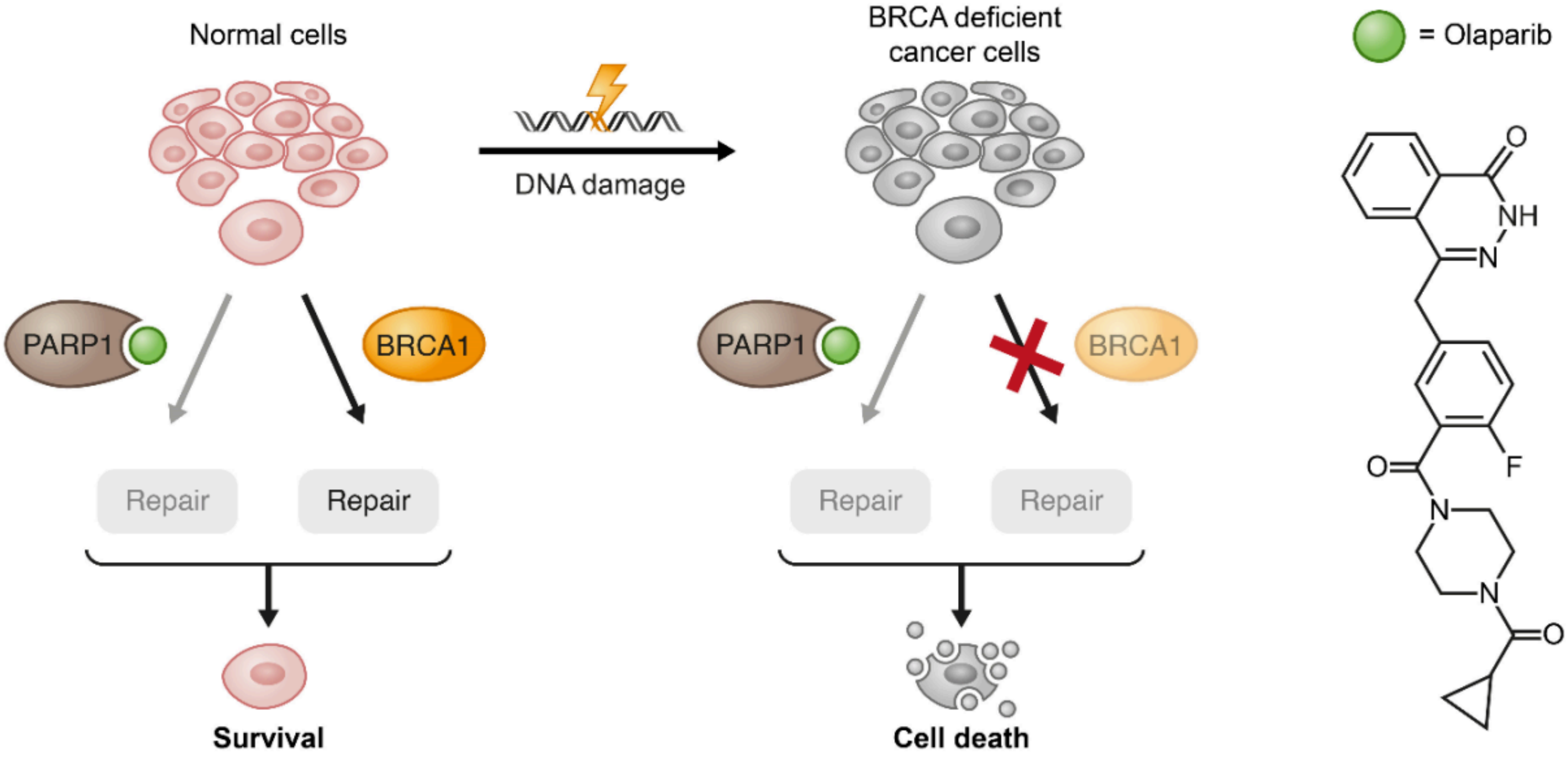
Question: how to leverage SL in drug combination discovery?

Source: wikipedia, Jerby-Arnon et al. Predicting Cancer-Specific Vulnerability via Data-Driven Detection of Synthetic Lethality

Drug combination therapy

- **Breast cancer**
  - an alkylating agent (cyclophosphamide) and antimetabolites (methotrexate and 5-fluorouracil)
- **Anti-HIV cocktail**
  - Use of three or more antiretroviral medicines
- We don’t have so many single drug candidate
- Drug combinations (k>=2) offer us more treatment plans

Drug treatment based on synthetic lethality



Goal: We want to make normal cells survive and kill cancer cells (BRCA deficient cancer cells)

Prior knowledge: PARP1 (off) + BRCA1 (off) -> cell death

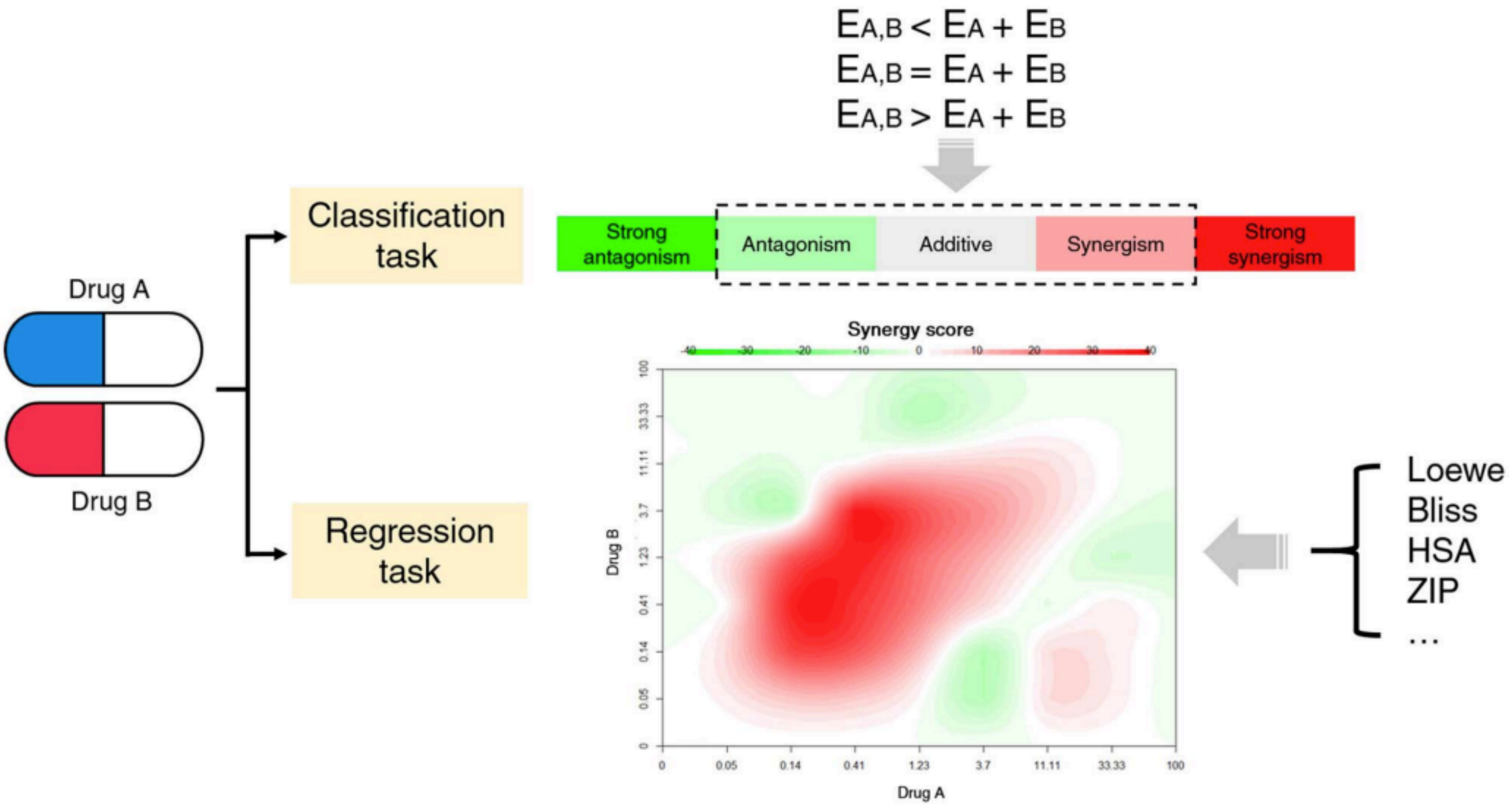
Solution: Turn off PARP1 using Olaparib

Results:

- Normal cells: PARP1 (off) + BRCA1 (on) -> cell survive
- Cancer cells: PARP1 (off) + BRCA1 (off) -> cell death

Gilad et al. Drug Combination in Cancer Treatment—From Cocktails to Conjugated Combinations

Drug combination prediction



$E_{A,B} < E_A + E_B$

$E_{A,B} = E_A + E_B$

$E_{A,B} > E_A + E_B$

Drug A

Drug B

Classification task

Regression task

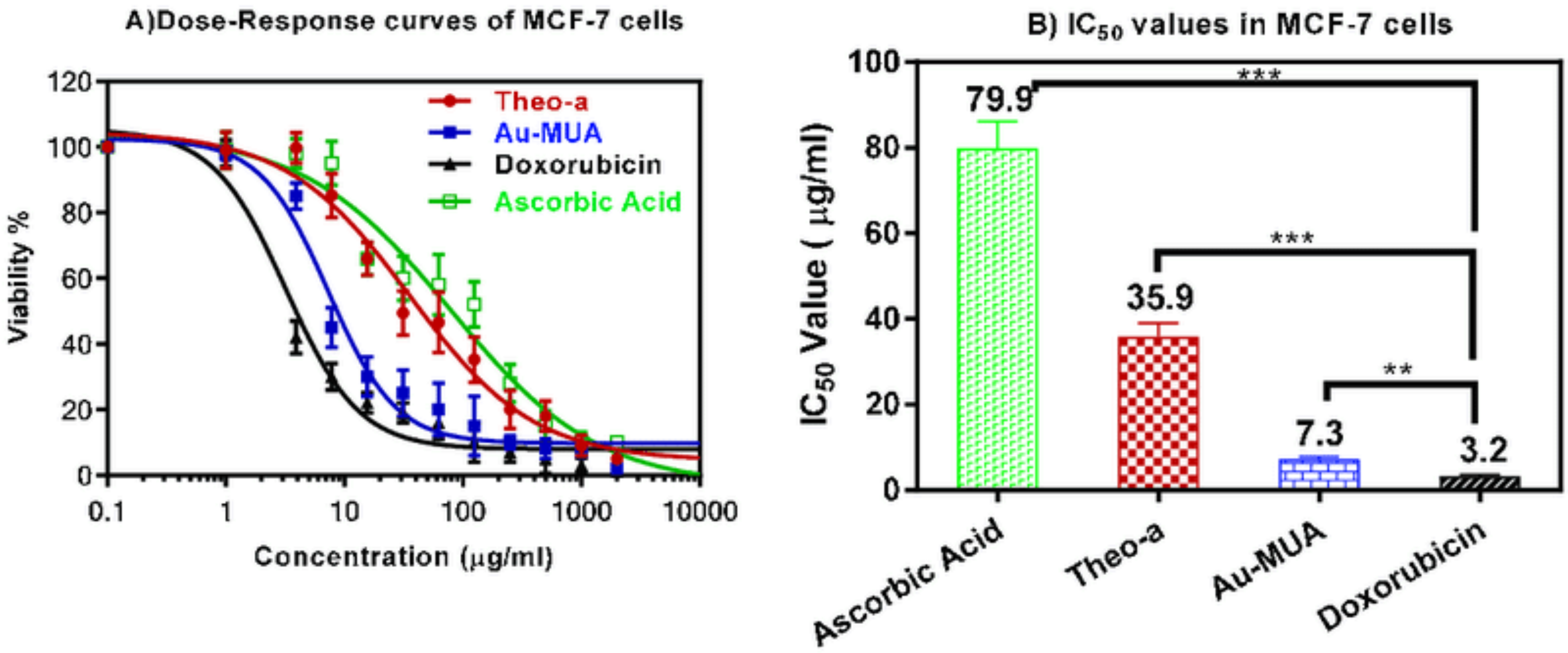
Strong antagonism	Antagonism	Additive	Synergism	Strong synergism

Loewe
Bliss
HSA
ZIP
…

E(A) is the efficacy of using drug A (e.g., IC50)

Wu et al. Machine learning methods, databases and tools for drug combination prediction

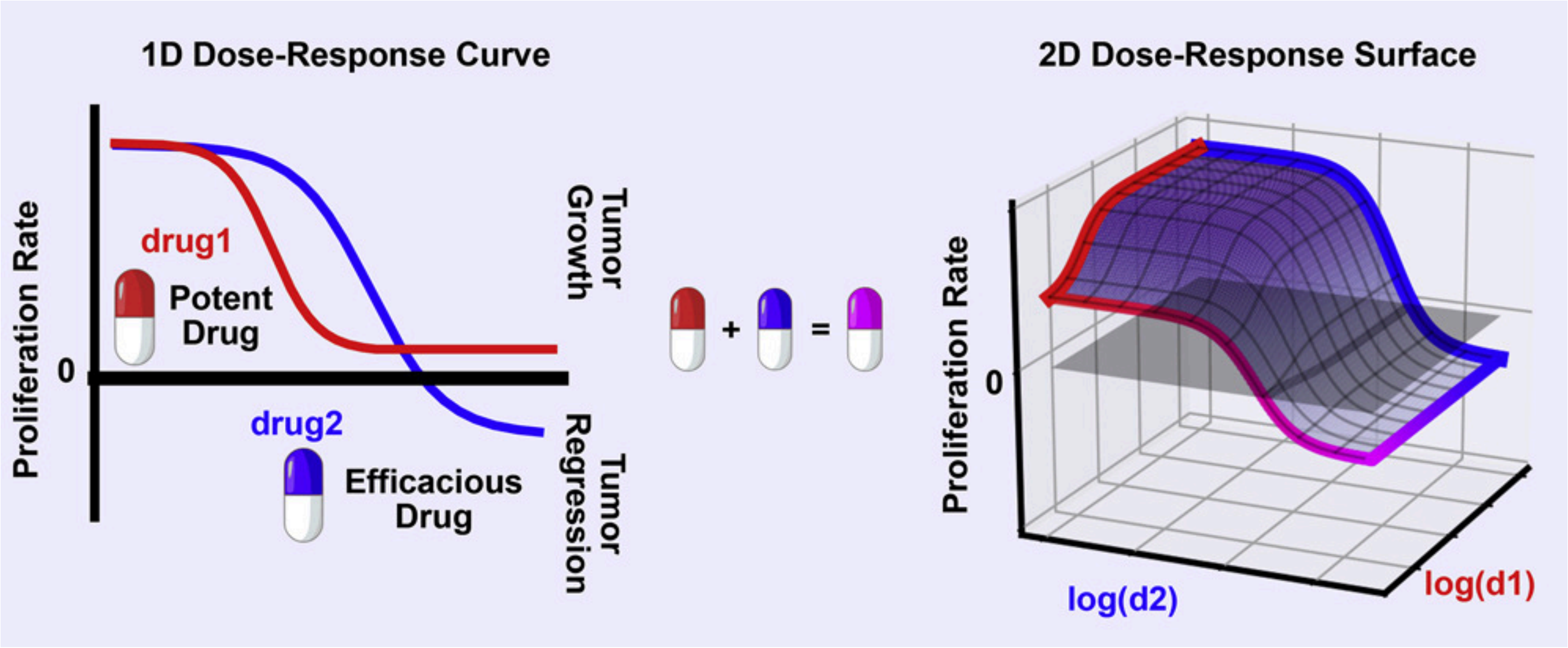
Dose-response curve



IC(50): concentration of 50% viability

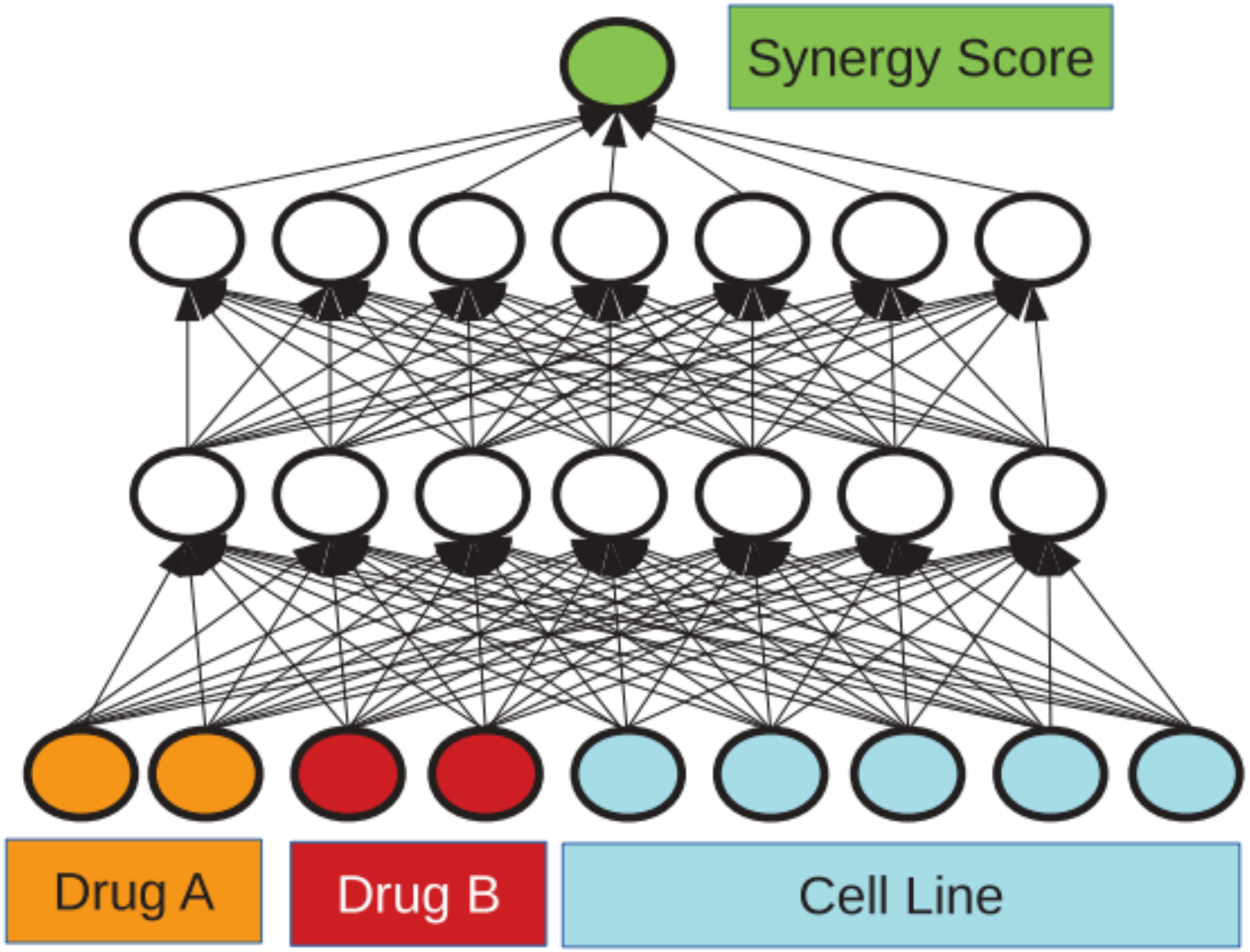
Mohammed et al. Formulation of Ethyl Cellulose Microparticles Incorporated Pheophytin A Isolated from Suaeda vermiculata for Antioxidant and Cytotoxic Activities

Drug combination dose-response curve



Meyer et al. Quantifying Drug Combination Synergy along Potency and Efficacy Axes

DeepSynergy: deep learning-based drug synergistic prediction



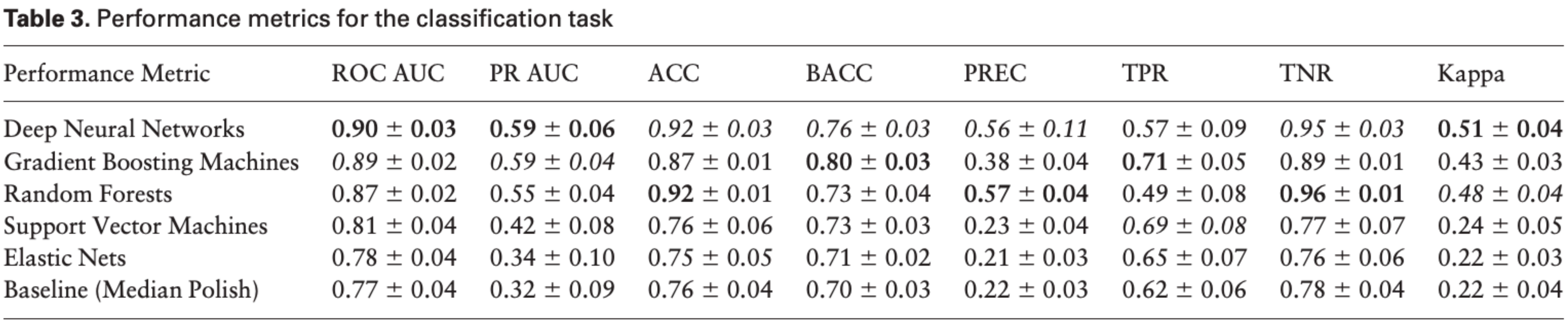
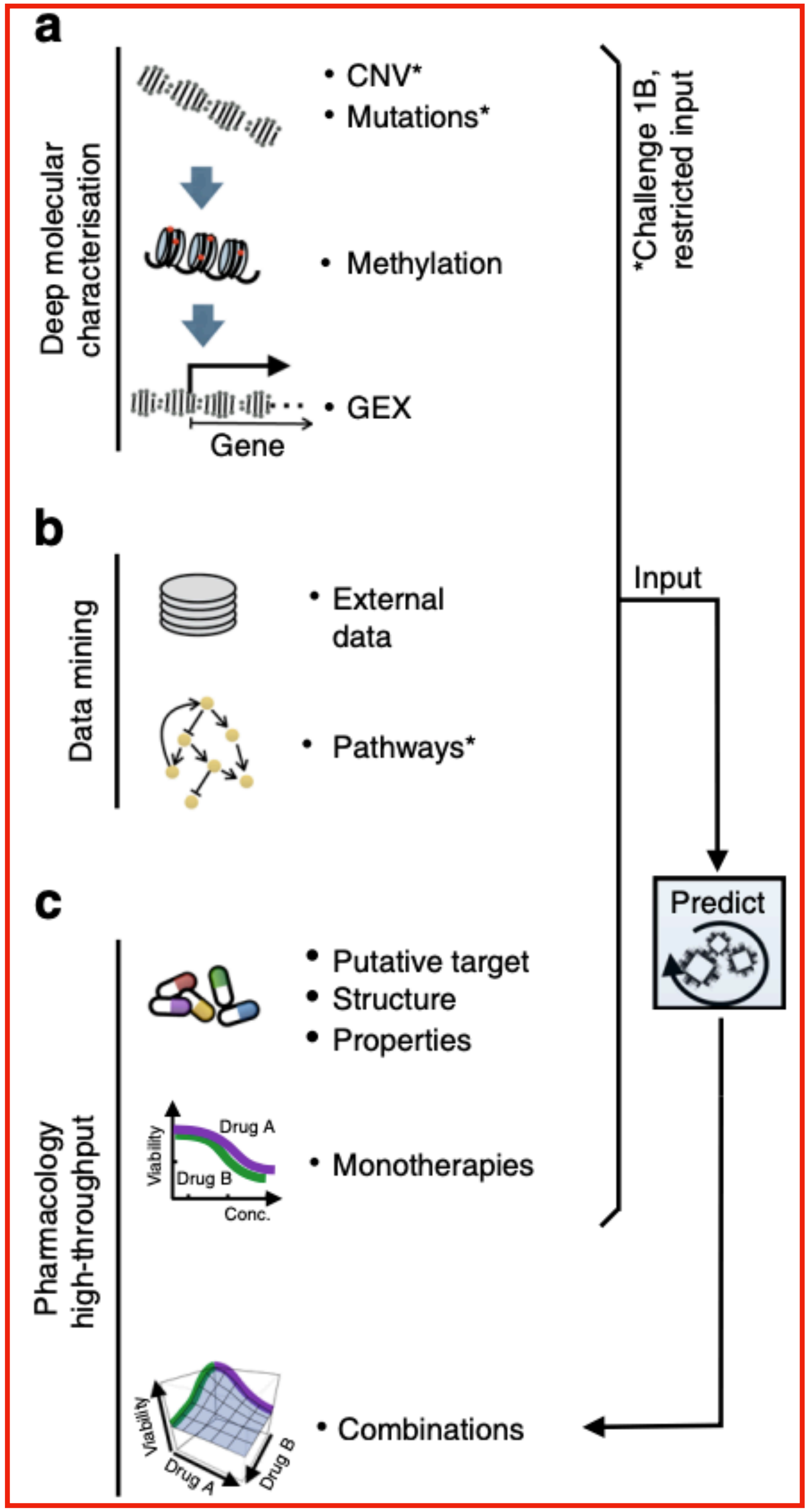


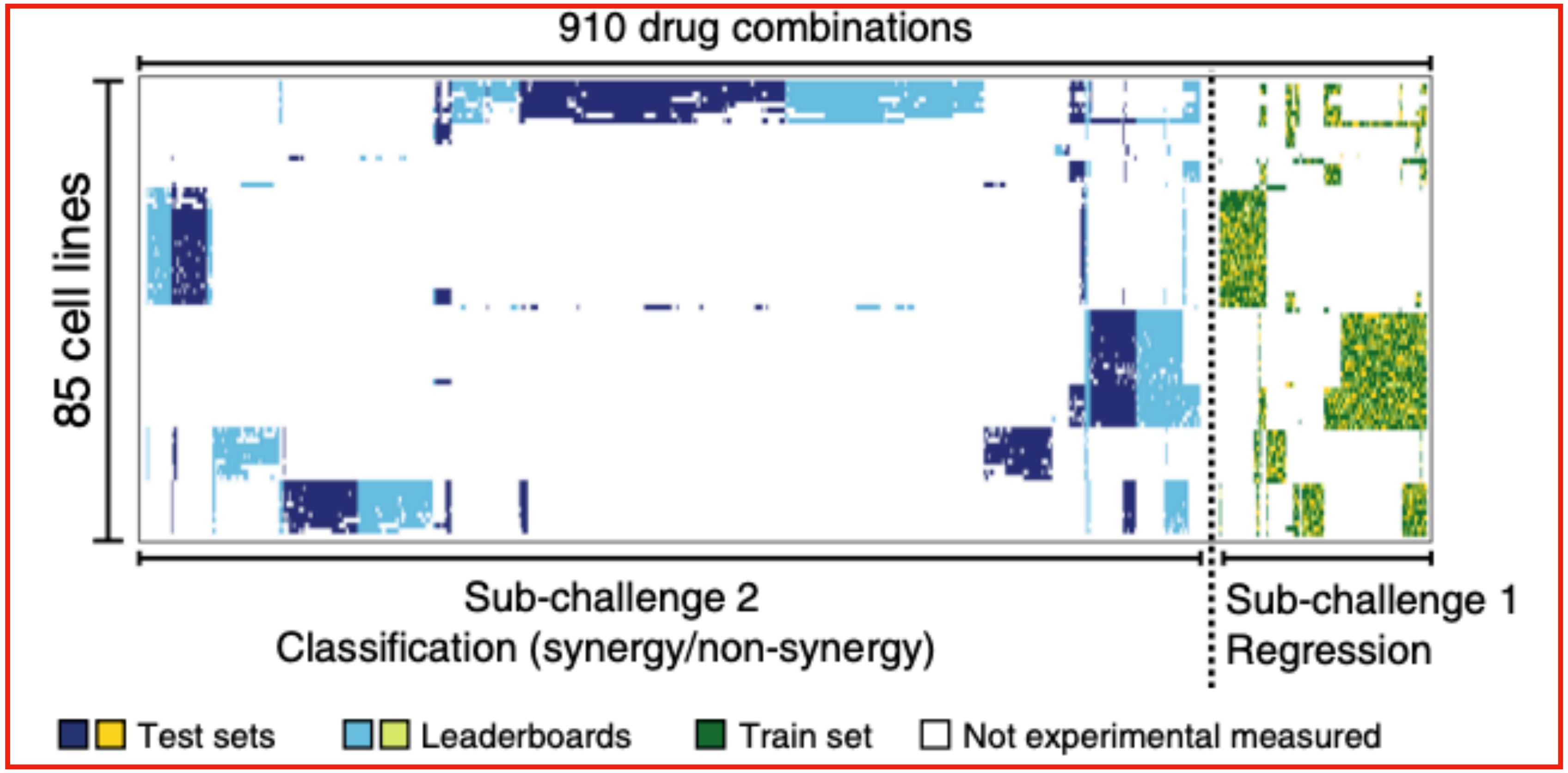
Table 3. Performance metrics for the classification task

Performance Metric	ROC AUC	PR AUC	ACC	BACC	PREC	TPR	TNR	Kappa
Deep Neural Networks	**0.90 ± 0.03**	**0.59 ± 0.06**	*0.92 ± 0.03*	*0.76 ± 0.03*	*0.56 ± 0.11*	0.57 ± 0.09	*0.95 ± 0.03*	**0.51 ± 0.04**
Gradient Boosting Machines	*0.89* ± 0.02	*0.59 ± 0.04*	0.87 ± 0.01	**0.80 ± 0.03**	0.38 ± 0.04	**0.71** ± 0.05	0.89 ± 0.01	0.43 ± 0.03
Random Forests	0.87 ± 0.02	0.55 ± 0.04	**0.92** ± 0.01	0.73 ± 0.04	**0.57 ± 0.04**	0.49 ± 0.08	**0.96 ± 0.01**	*0.48 ± 0.04*
Support Vector Machines	0.81 ± 0.04	0.42 ± 0.08	0.76 ± 0.06	0.73 ± 0.03	0.23 ± 0.04	*0.69 ± 0.08*	0.77 ± 0.07	0.24 ± 0.05
Elastic Nets	0.78 ± 0.04	0.34 ± 0.10	0.75 ± 0.05	0.71 ± 0.02	0.21 ± 0.03	0.65 ± 0.07	0.76 ± 0.06	0.22 ± 0.03
Baseline (Median Polish)	0.77 ± 0.04	0.32 ± 0.09	0.76 ± 0.04	0.70 ± 0.03	0.22 ± 0.03	0.62 ± 0.06	0.78 ± 0.04	0.22 ± 0.04

Preuer et al. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning.

Problem setting

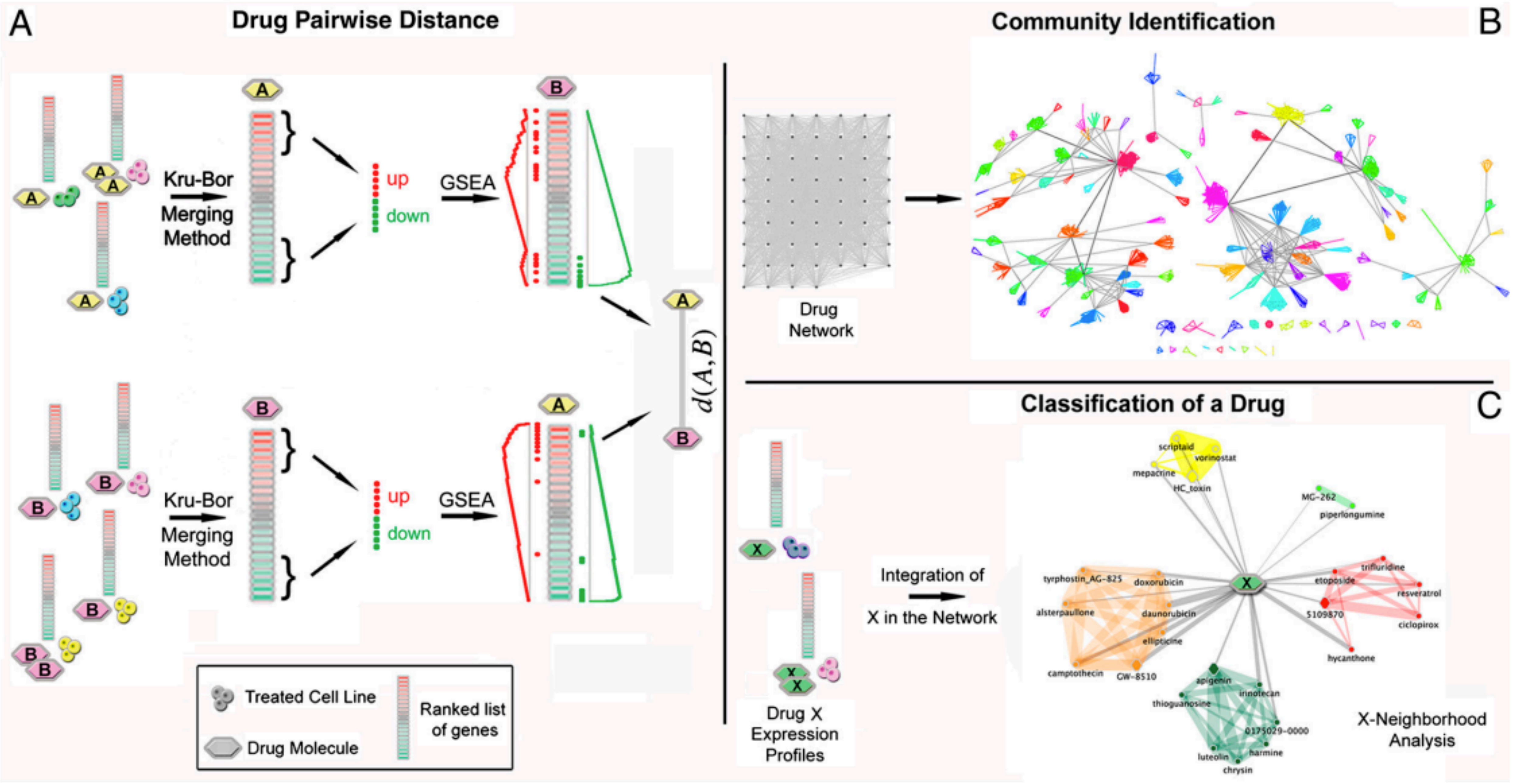




Menden et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen.

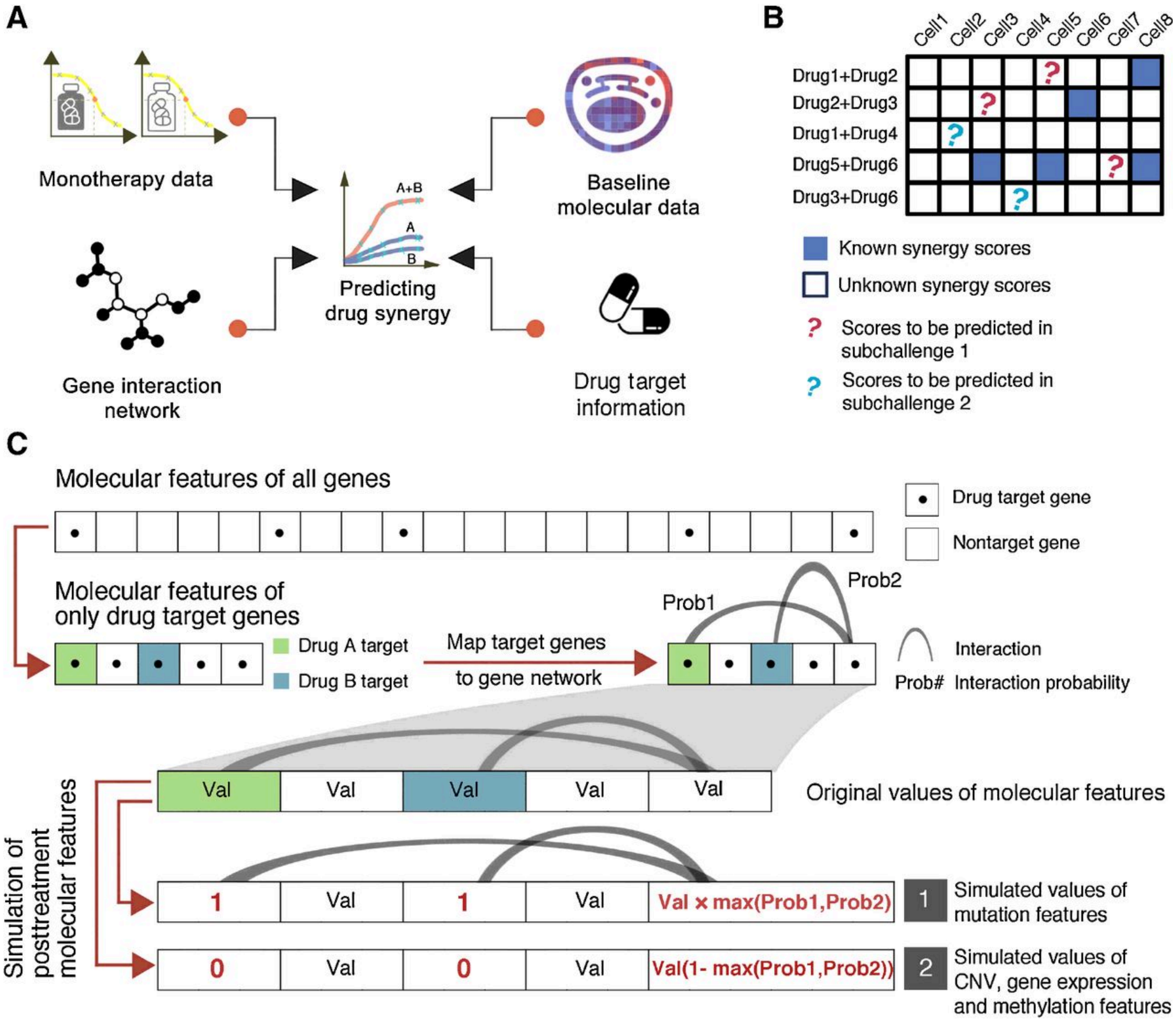
Use gene expression after treatment

Drugs target on similar proteins or have similar Mode of Actions have similar (after treatment) expression.



Iorio et al. Discovery of drug mode of action and drug repositioning from transcriptional responses

TAIJI: simulate post treatment expression


A
Monotherapy data
A+B
A
B
Predicting drug synergy
Baseline molecular data
Gene interaction network
Drug target information
B
C
Molecular features of all genes
Drug target gene
Nontarget gene
Molecular features of only drug target genes
Drug A target
Drug B target
Map target genes to gene network
Prob1
Prob2
Interaction
Prob# Interaction probability
Val
Val
Val
Val
Val
Original values of molecular features
1
Val
1
Val
Val × max(Prob1,Prob2)
1
Simulated values of mutation features
0
Val
0
Val
Val(1- max(Prob1,Prob2))
2
Simulated values of CNV, gene expression and methylation features
Simulation of posttreatment molecular features

	Cell1	Cell2	Cell3	Cell4	Cell5	Cell6	Cell7	Cell8
Drug1+Drug2					?			■
Drug2+Drug3			?			■		
Drug1+Drug4		?						
Drug5+Drug6			■		■		?	■
Drug3+Drug6				?				

■ Known synergy scores

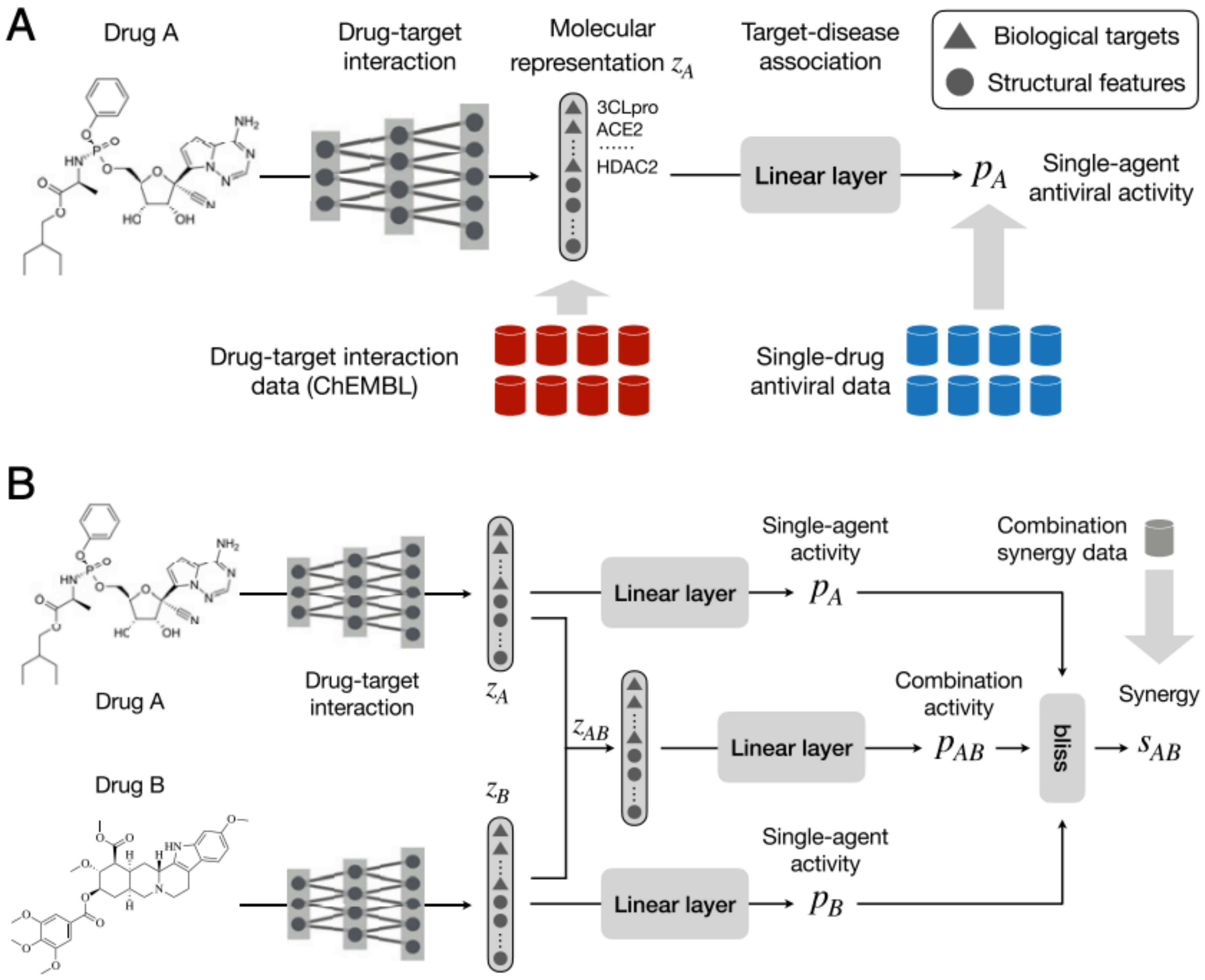
□ Unknown synergy scores

? Scores to be predicted in subchallenge 1

? Scores to be predicted in subchallenge 2

Li et al. Network Propagation Predicts Drug Synergy in Cancers

Drug combinations for treating COVID-19



Jin et al. Deep learning identifies synergistic drug combinations for treating COVID-19.

Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data

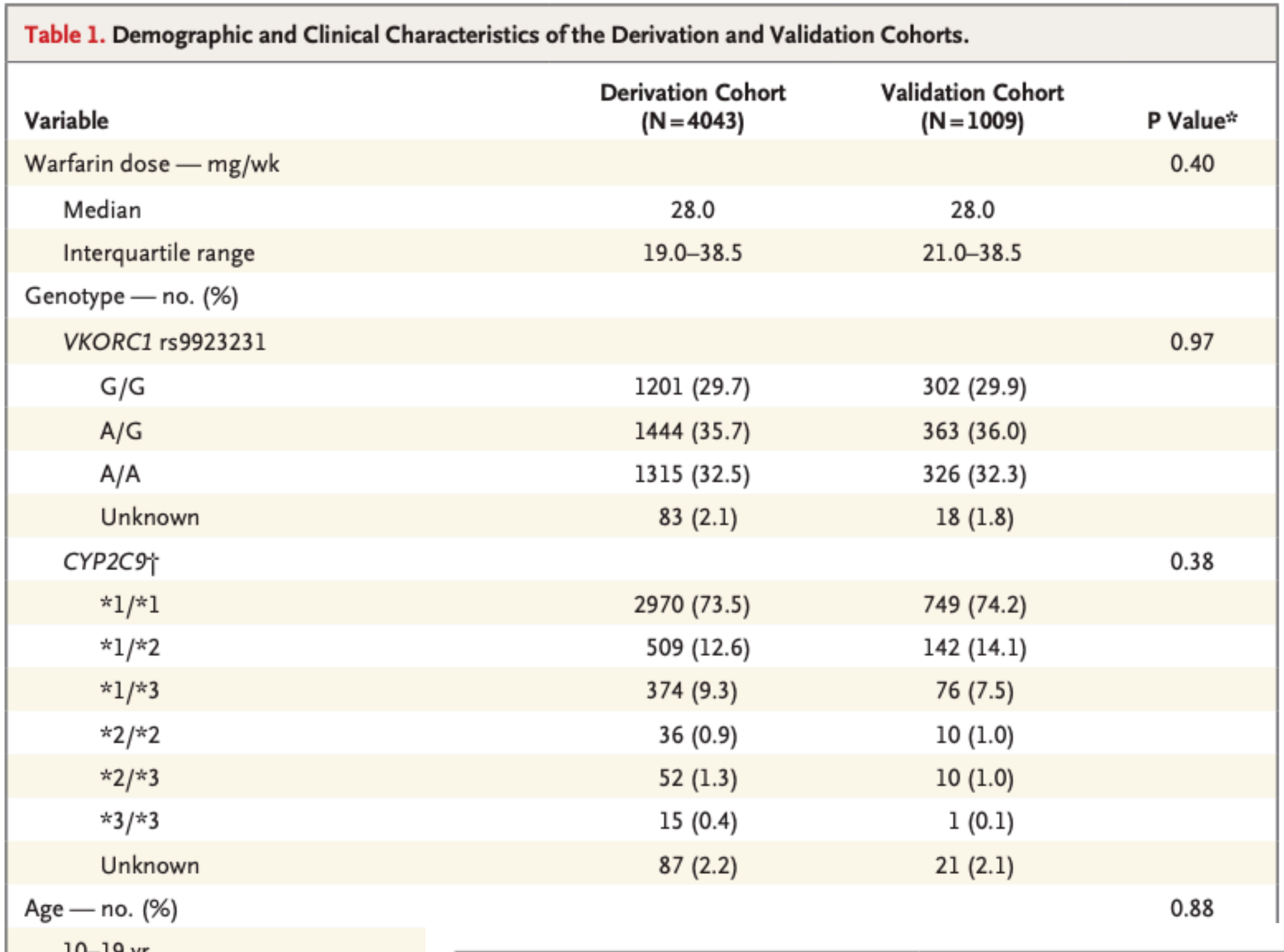


Table 1. Demographic and Clinical Characteristics of the Derivation and Validation Cohorts.

Variable	Derivation Cohort (N=4043)	Validation Cohort (N=1009)	P Value*
Warfarin dose — mg/wk			0.40
Median	28.0	28.0	
Interquartile range	19.0–38.5	21.0–38.5	
Genotype — no. (%)			
VKORC1 rs9923231			0.97
G/G	1201 (29.7)	302 (29.9)	
A/G	1444 (35.7)	363 (36.0)	
A/A	1315 (32.5)	326 (32.3)	
Unknown	83 (2.1)	18 (1.8)	
CYP2C9†			0.38
*1/*1	2970 (73.5)	749 (74.2)	
*1/*2	509 (12.6)	142 (14.1)	
*1/*3	374 (9.3)	76 (7.5)	
*2/*2	36 (0.9)	10 (1.0)	
*2/*3	52 (1.3)	10 (1.0)	
*3/*3	15 (0.4)	1 (0.1)	
Unknown	87 (2.2)	21 (2.1)	
Age — no. (%)			0.88
10–19 yr			
20–29 yr			
30–39 yr			
40–49 yr			
50–59 yr			
60–69 yr			
70–79 yr			
80–89 yr			
≥90 yr			

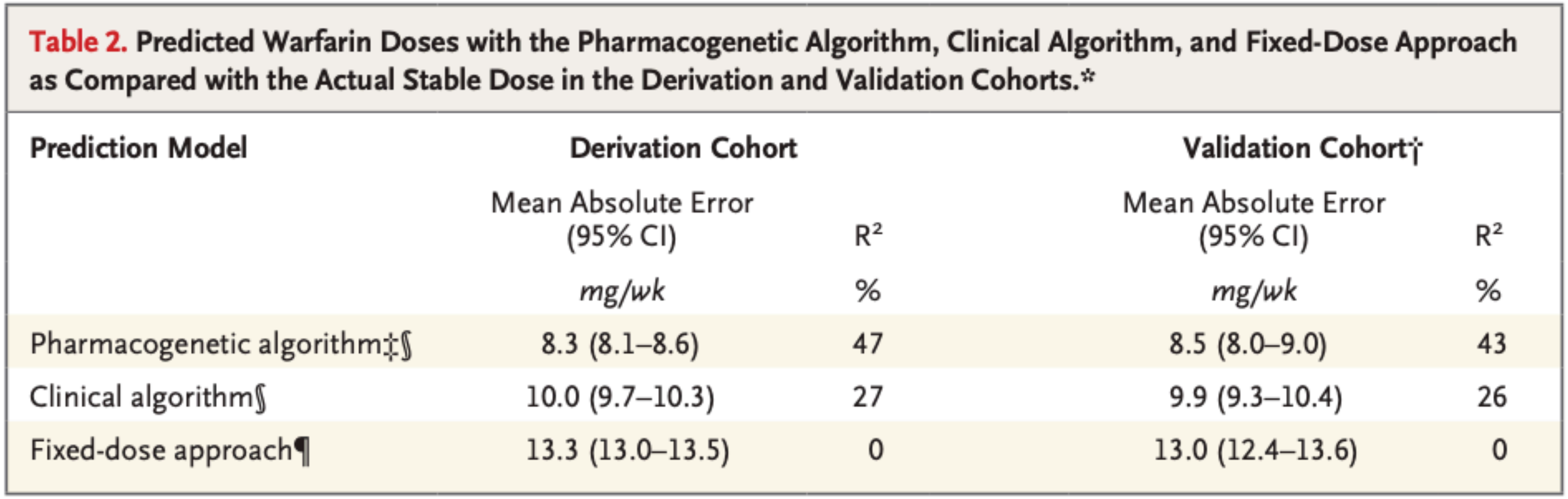
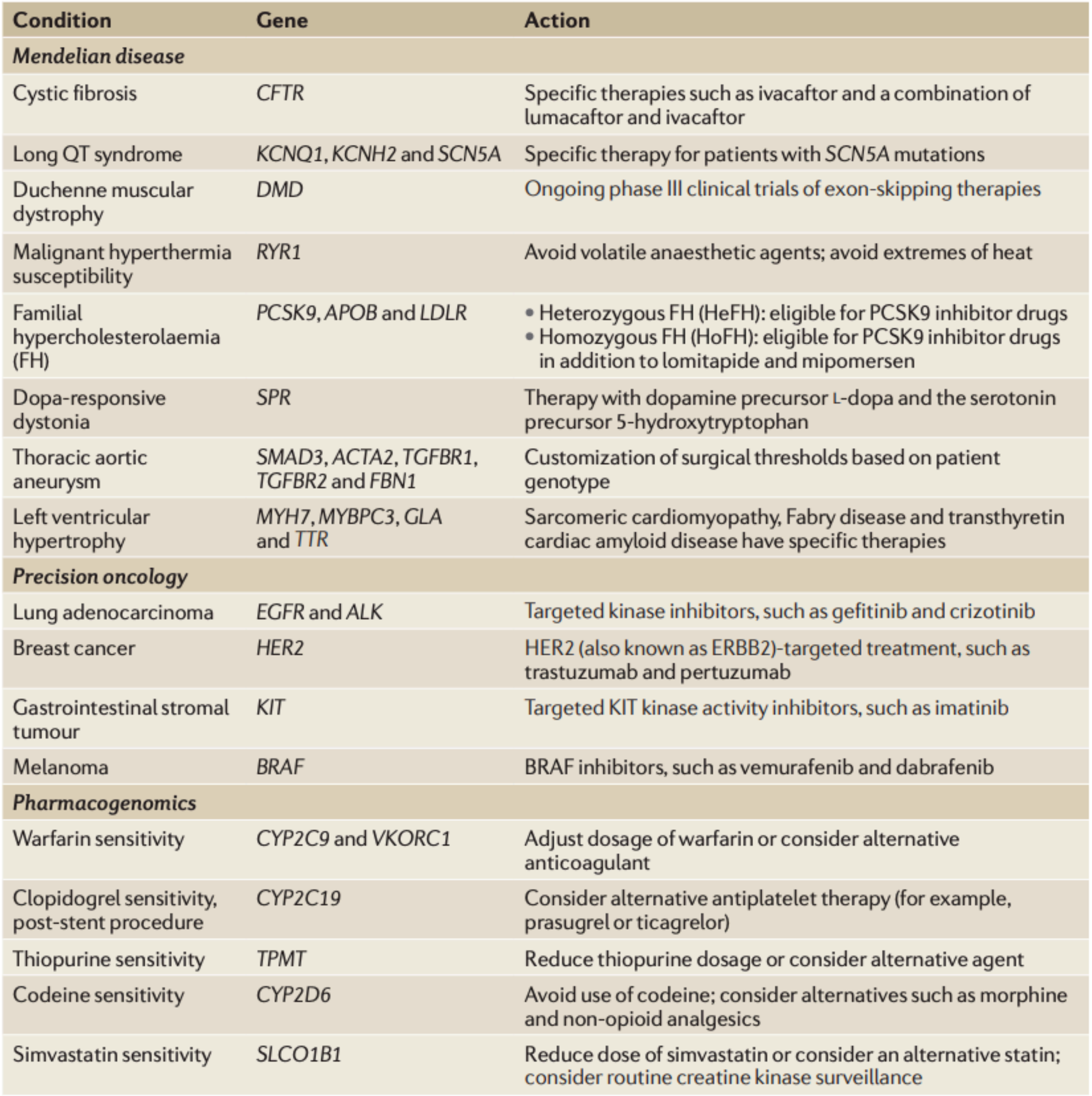


Table 2. Predicted Warfarin Doses with the Pharmacogenetic Algorithm, Clinical Algorithm, and Fixed-Dose Approach as Compared with the Actual Stable Dose in the Derivation and Validation Cohorts.*

Prediction Model	Derivation Cohort		Validation Cohort†	
	Mean Absolute Error (95% CI)	R^2	Mean Absolute Error (95% CI)	R^2
	mg/wk	%	*mg/wk*	%
Pharmacogenetic algorithm‡§	8.3 (8.1–8.6)	47	8.5 (8.0–9.0)	43
Clinical algorithm§	10.0 (9.7–10.3)	27	9.9 (9.3–10.4)	26
Fixed-dose approach¶	13.3 (13.0–13.5)	0	13.0 (12.4–13.6)	0

The International Warfarin Pharmacogenetics Consortium. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data

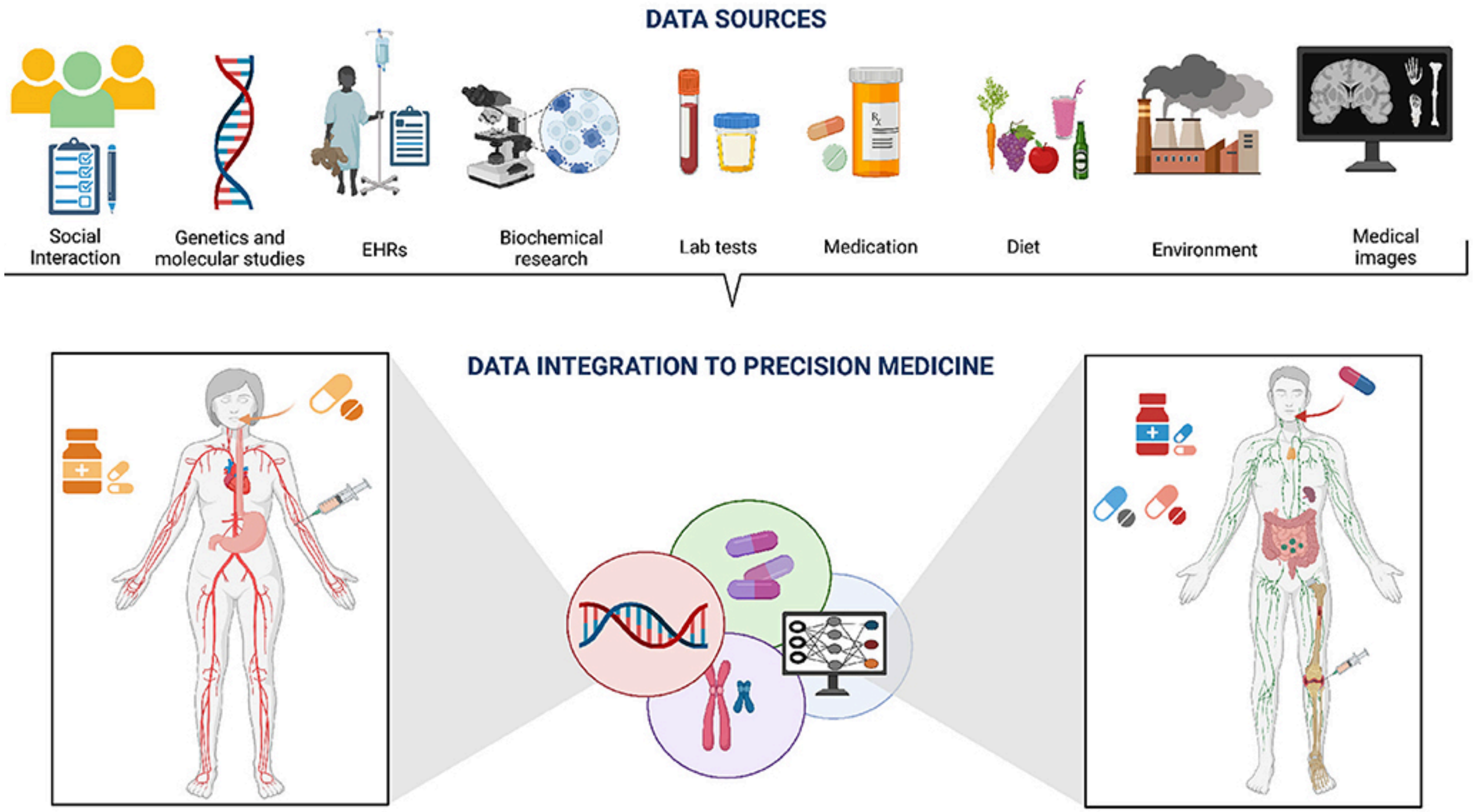
Example of precision medicine



Condition	Gene	Action
Mendelian disease		
Cystic fibrosis	*CFTR*	Specific therapies such as ivacaftor and a combination of lumacaftor and ivacaftor
Long QT syndrome	*KCNQ1*, *KCNH2* and *SCN5A*	Specific therapy for patients with *SCN5A* mutations
Duchenne muscular dystrophy	*DMD*	Ongoing phase III clinical trials of exon-skipping therapies
Malignant hyperthermia susceptibility	*RYR1*	Avoid volatile anaesthetic agents; avoid extremes of heat
Familial hypercholesterolaemia (FH)	*PCSK9*, *APOB* and *LDLR*	• Heterozygous FH (HeFH): eligible for PCSK9 inhibitor drugs • Homozygous FH (HoFH): eligible for PCSK9 inhibitor drugs in addition to lomitapide and mipomersen
Dopa-responsive dystonia	*SPR*	Therapy with dopamine precursor L-dopa and the serotonin precursor 5-hydroxytryptophan
Thoracic aortic aneurysm	*SMAD3*, *ACTA2*, *TGFBR1*, *TGFBR2* and *FBN1*	Customization of surgical thresholds based on patient genotype
Left ventricular hypertrophy	*MYH7*, *MYBPC3*, *GLA* and *TTR*	Sarcomeric cardiomyopathy, Fabry disease and transthyretin cardiac amyloid disease have specific therapies
Precision oncology		
Lung adenocarcinoma	*EGFR* and *ALK*	Targeted kinase inhibitors, such as gefitinib and crizotinib
Breast cancer	*HER2*	HER2 (also known as ERBB2)-targeted treatment, such as trastuzumab and pertuzumab
Gastrointestinal stromal tumour	*KIT*	Targeted KIT kinase activity inhibitors, such as imatinib
Melanoma	*BRAF*	BRAF inhibitors, such as vemurafenib and dabrafenib
Pharmacogenomics		
Warfarin sensitivity	*CYP2C9* and *VKORC1*	Adjust dosage of warfarin or consider alternative anticoagulant
Clopidogrel sensitivity, post-stent procedure	*CYP2C19*	Consider alternative antiplatelet therapy (for example, prasugrel or ticagrelor)
Thiopurine sensitivity	*TPMT*	Reduce thiopurine dosage or consider alternative agent
Codeine sensitivity	*CYP2D6*	Avoid use of codeine; consider alternatives such as morphine and non-opioid analgesics
Simvastatin sensitivity	*SLCO1B1*	Reduce dose of simvastatin or consider an alternative statin; consider routine creatine kinase surveillance

Euan A. Ashley. Towards precision medicine.

Two key problems



- How to cluster patients
- We don’t have so many “drugs”

Martinez-Garcia et al. Data Integration Challenges for Machine Learning in Precision Medicine

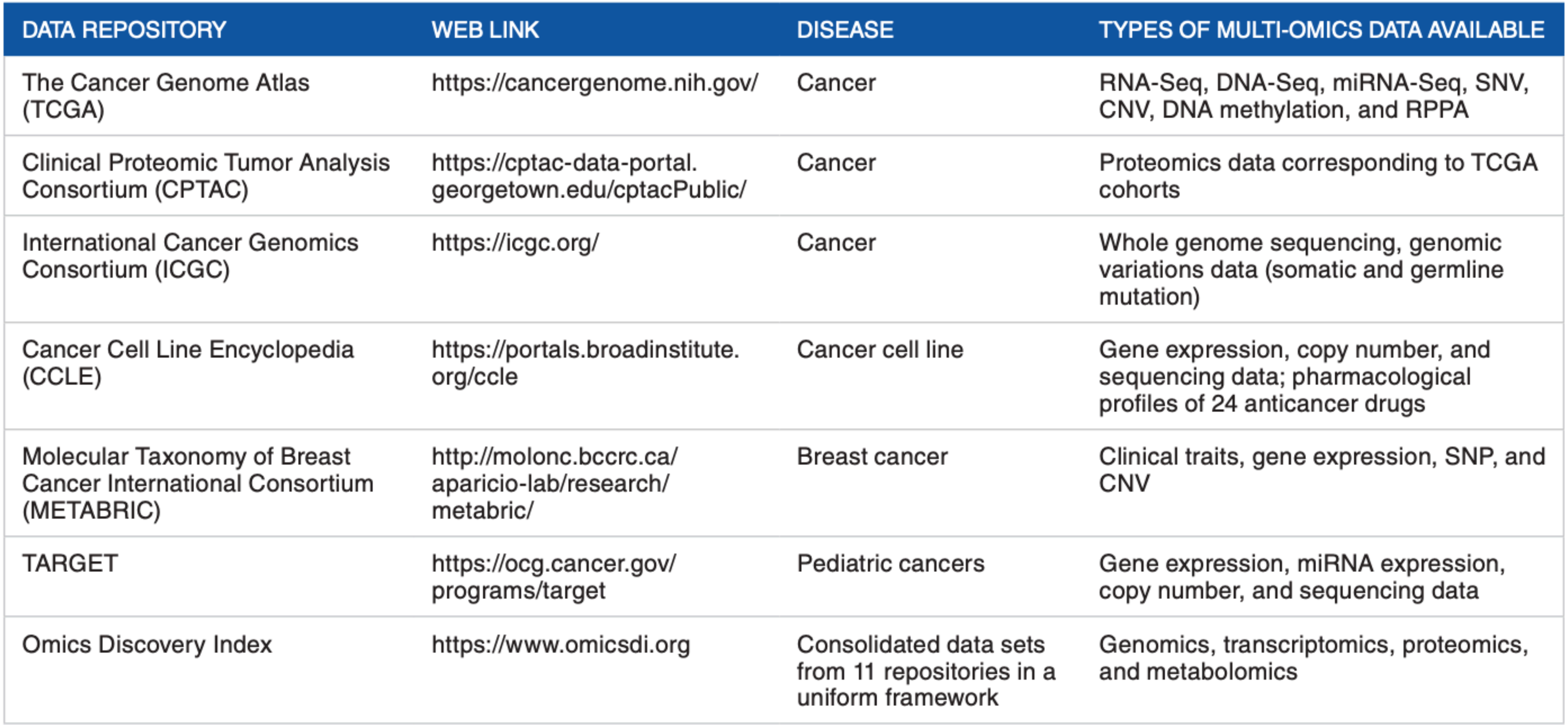
How to cluster patients

DATA SOURCES
Social Interaction
Genetics and molecular studies
EHRs
Biochemical research
Lab tests
Medication
Diet
Environment
Medical images

- **Patient clustering = data integration**
  - Find a "signature" vector for each patient
  - Signature is integrated from different data sources
- **Heterogeneous data integration**
  - General challenges: Heterogenous, missing values, noise, privacy
- **Precision medicine specific data integration challenges:**
  - Batch effects (different preprocessing pipelines, sequencing techniques, reference ranges)
  - Unpaired data (some patients only have genomics data, some patients only have EHR data, very few patients have both)

Grapov et al. Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine

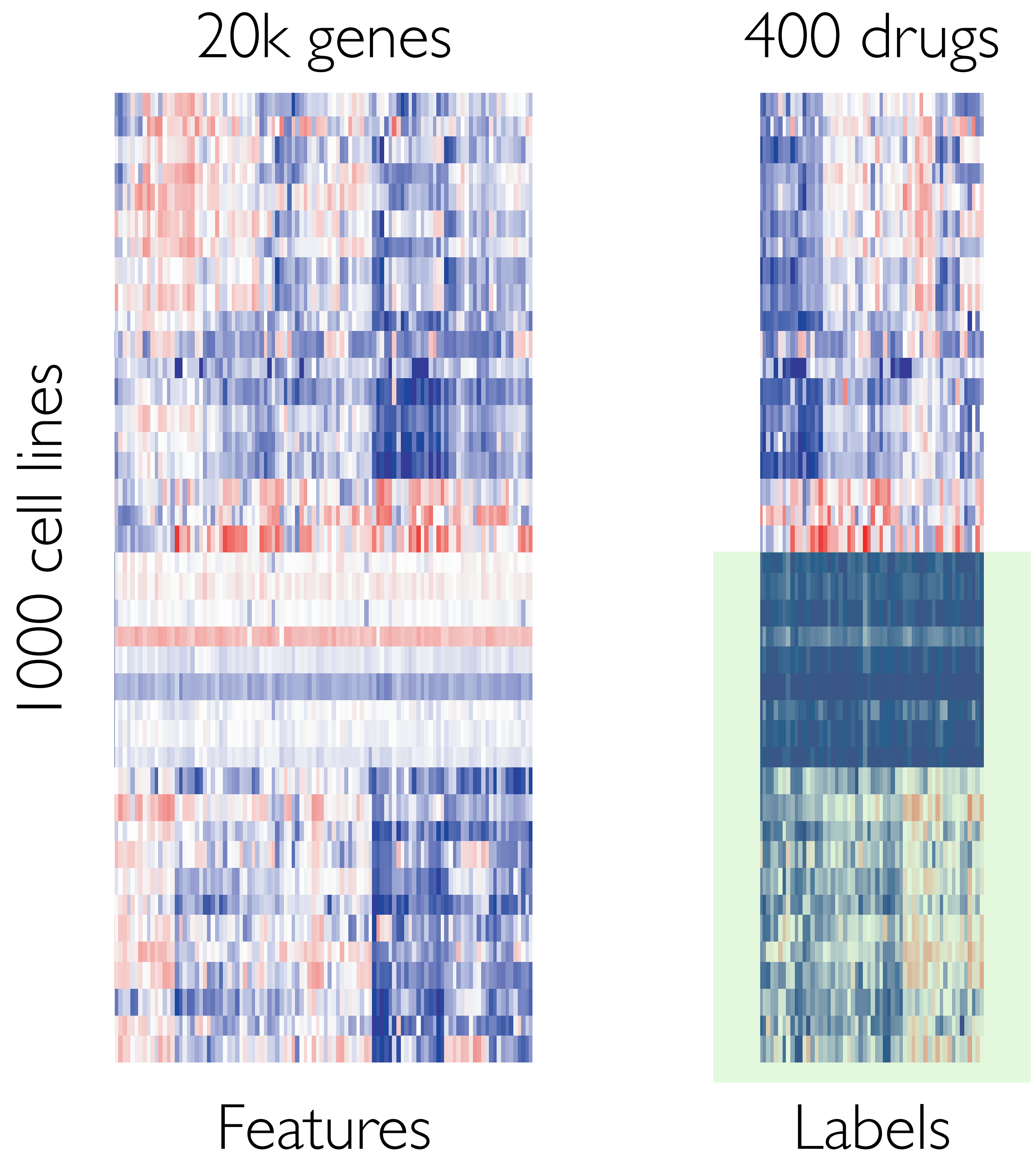
Public biomedical databases



DATA REPOSITORY	WEB LINK	DISEASE	TYPES OF MULTI-OMICS DATA AVAILABLE
The Cancer Genome Atlas (TCGA)	https://cancergenome.nih.gov/	Cancer	RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, and RPPA
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	https://cptac-data-portal.georgetown.edu/cptacPublic/	Cancer	Proteomics data corresponding to TCGA cohorts
International Cancer Genomics Consortium (ICGC)	https://icgc.org/	Cancer	Whole genome sequencing, genomic variations data (somatic and germline mutation)
Cancer Cell Line Encyclopedia (CCLE)	https://portals.broadinstitute.org/ccle	Cancer cell line	Gene expression, copy number, and sequencing data; pharmacological profiles of 24 anticancer drugs
Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)	http://molonc.bccrc.ca/aparicio-lab/research/metabric/	Breast cancer	Clinical traits, gene expression, SNP, and CNV
TARGET	https://ocg.cancer.gov/programs/target	Pediatric cancers	Gene expression, miRNA expression, copy number, and sequencing data
Omics Discovery Index	https://www.omicsdi.org	Consolidated data sets from 11 repositories in a uniform framework	Genomics, transcriptomics, proteomics, and metabolomics

Subramanian et al. Multi-omics Data Integration, Interpretation, and Its Application

Personalized drug response prediction: multi-label regression problem

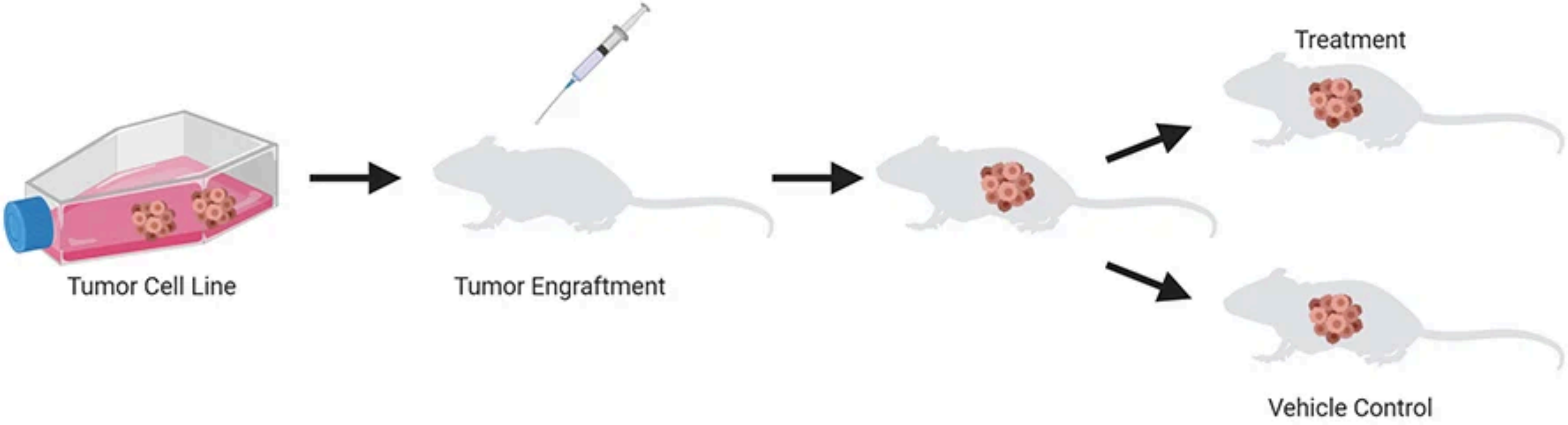


CCLE data: ~1000 cell lines, 20k genes, 400 drugs

Three settings

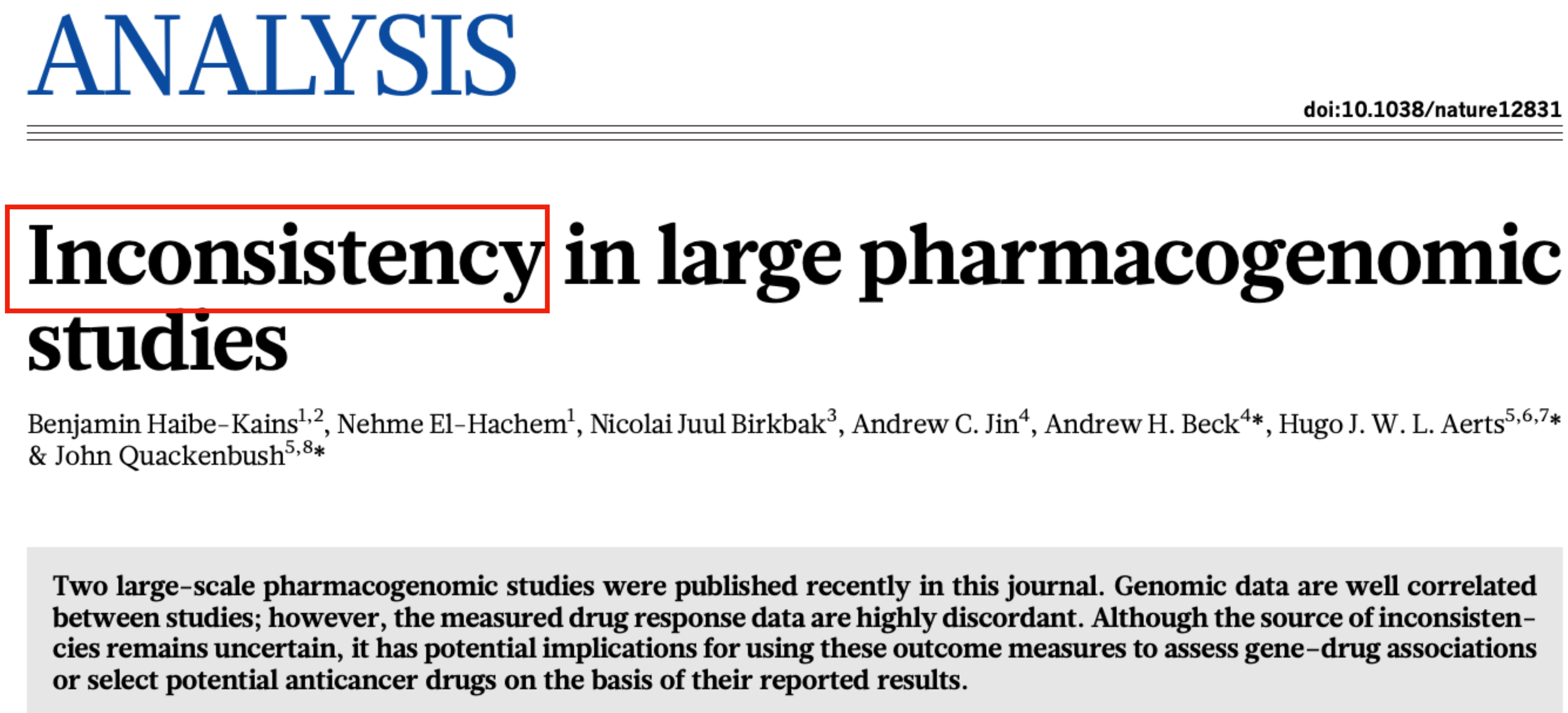
- Test patients: no drugs are observed for this patient
- Test drugs: no patients are observed for this drug
- Test <patient, drug> pairs

Cell line, xenograft, tumor, patient



- Cell line is a "copy" of a patient. We cannot test one patient with many drugs. But we can copy a cell line many times.
- Cell line is cheaper than xenograft. Xenograft is cheaper than patient data
- Xenograft data: Gao et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response
- TCGA has some patient data
- ML question: how to integrate cell line data, xenograft data and patient data

Batch effects: inconsistency or consistency?



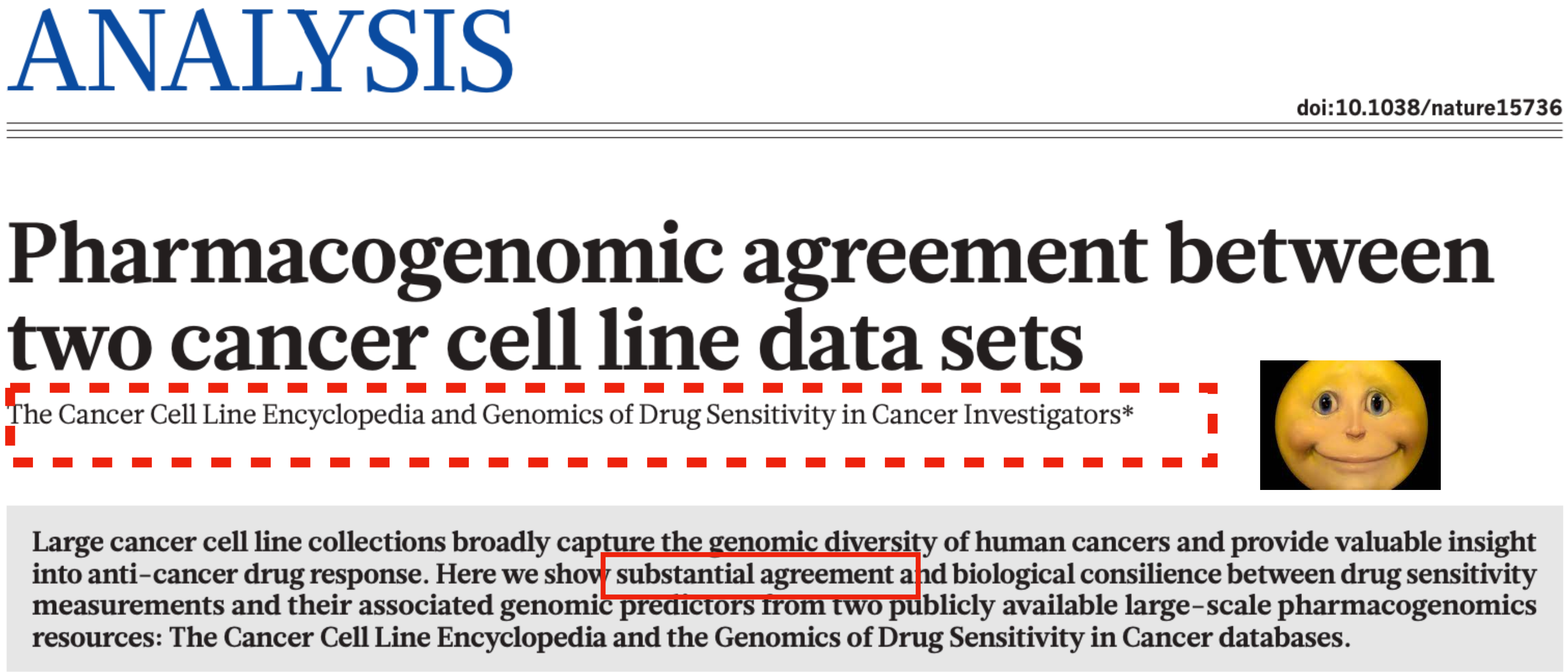
ANALYSIS

doi:10.1038/nature12831

Inconsistency in large pharmacogenomic studies

Benjamin Haibe-Kains[1,2], Nehme El-Hachem[1], Nicolai Juul Birkbak[3], Andrew C. Jin[4], Andrew H. Beck[4]*, Hugo J. W. L. Aerts[5,6,7]* & John Quackenbush[5,8]*

Two large-scale pharmacogenomic studies were published recently in this journal. Genomic data are well correlated between studies; however, the measured drug response data are highly discordant. Although the source of inconsistencies remains uncertain, it has potential implications for using these outcome measures to assess gene-drug associations or select potential anticancer drugs on the basis of their reported results.



ANALYSIS

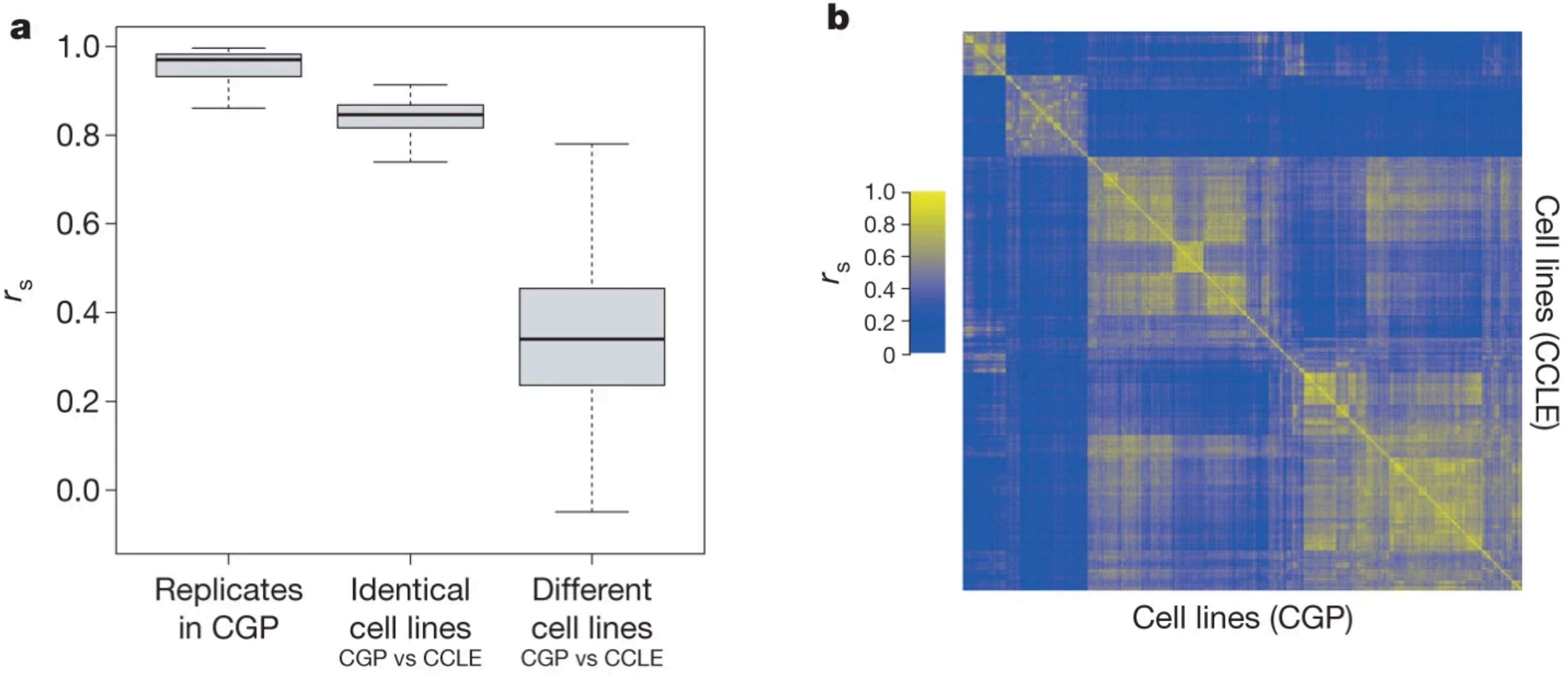
doi:10.1038/nature15736

Pharmacogenomic agreement between two cancer cell line data sets

The Cancer Cell Line Encyclopedia and Genomics of Drug Sensitivity in Cancer Investigators*

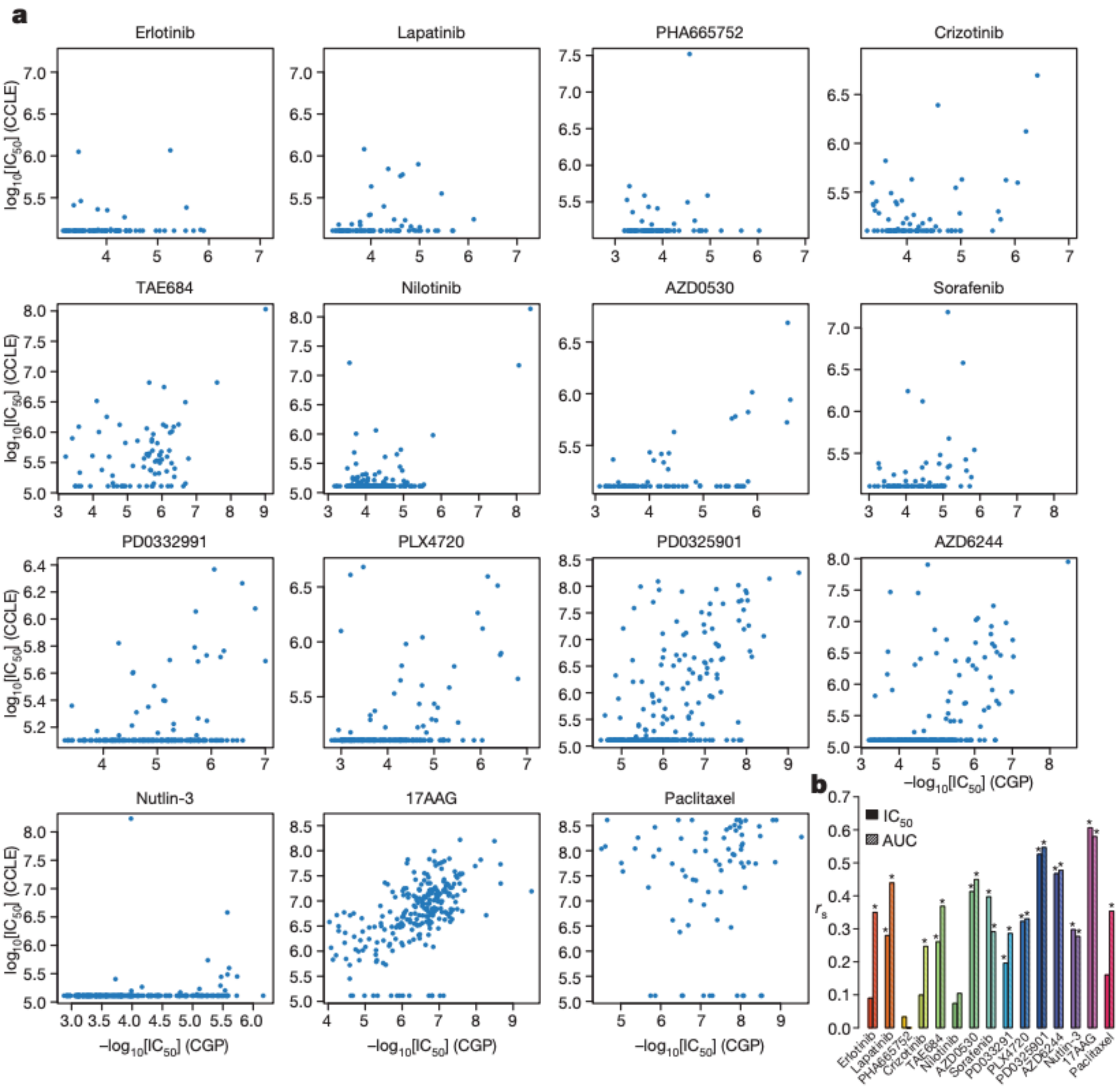
Large cancer cell line collections broadly capture the genomic diversity of human cancers and provide valuable insight into anti-cancer drug response. Here we show substantial agreement and biological consilience between drug sensitivity measurements and their associated genomic predictors from two publicly available large-scale pharmacogenomics resources: The Cancer Cell Line Encyclopedia and the Genomics of Drug Sensitivity in Cancer databases.

Low correlation between drug response data

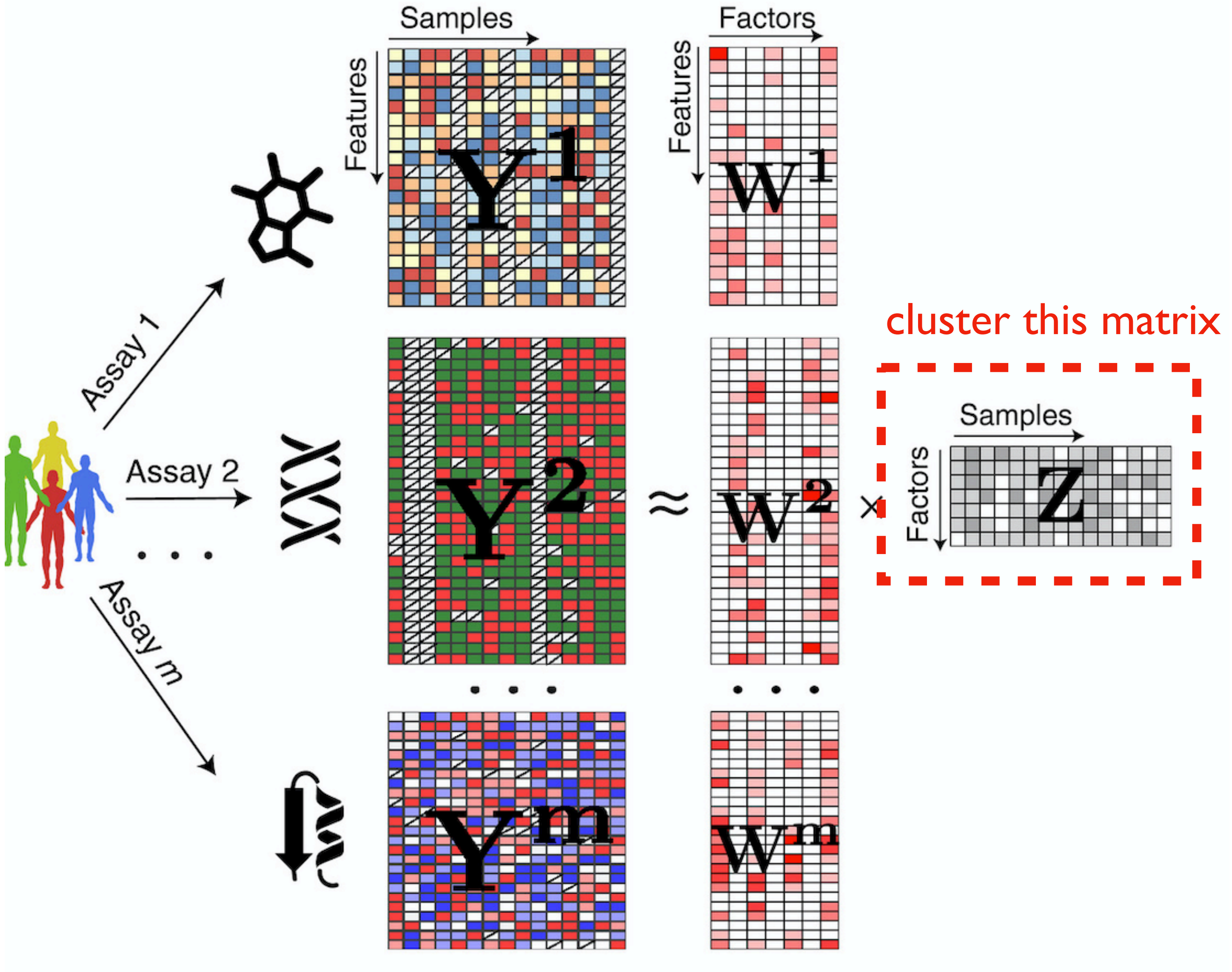


Integrating two datasets

High correlation between gene expression (features)



General framework: jointly decompose multiple data matrices

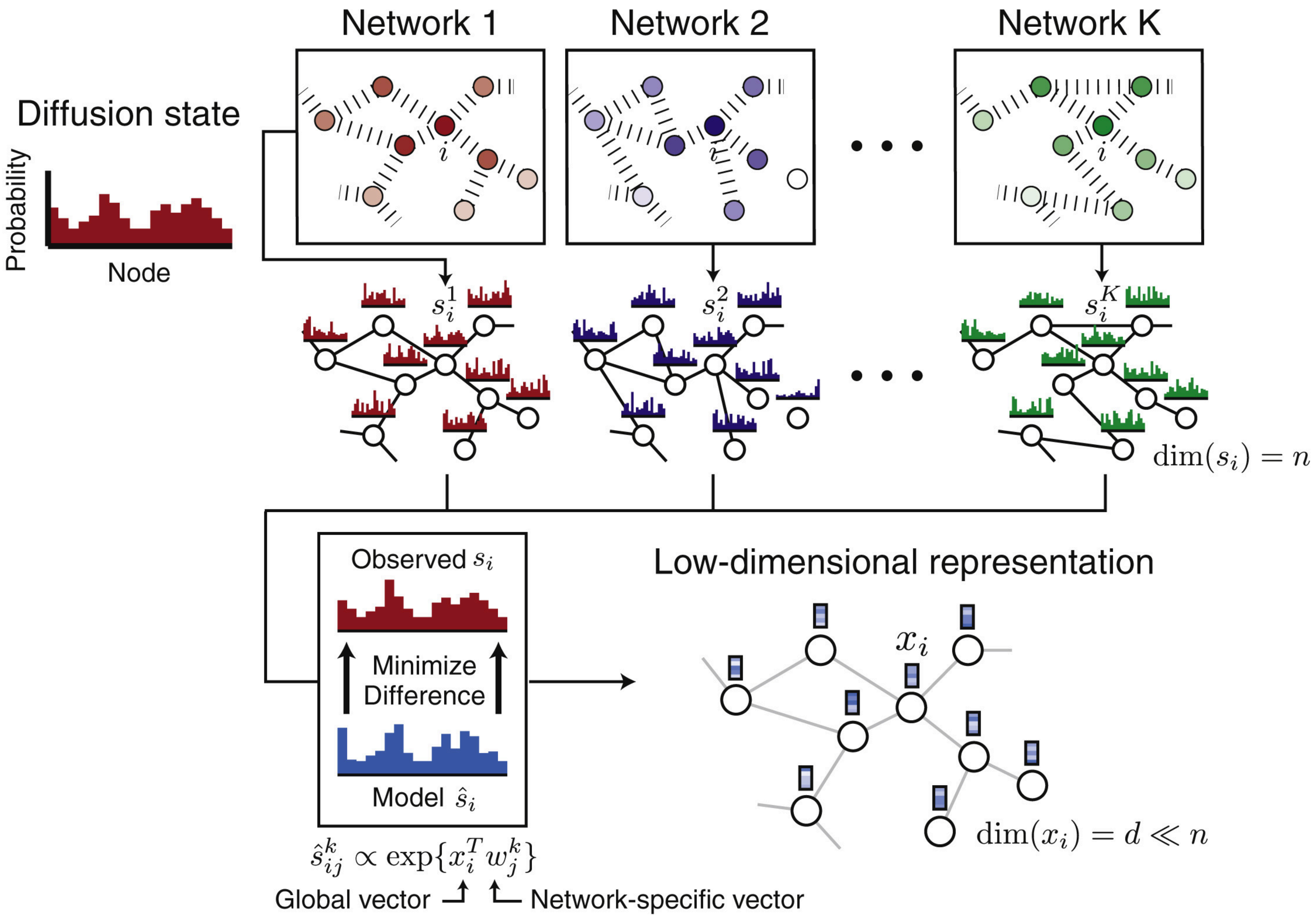


cluster this matrix

- Key ideas
  - One matrix capture batch effects
  - One matrix capture common patterns
- Detail implementations
  - What distribution?
    - Mutation (Bernoulli)
    - Expression count (Poisson)
  - How to decompose?
    - NMF, SVD, NN, MF

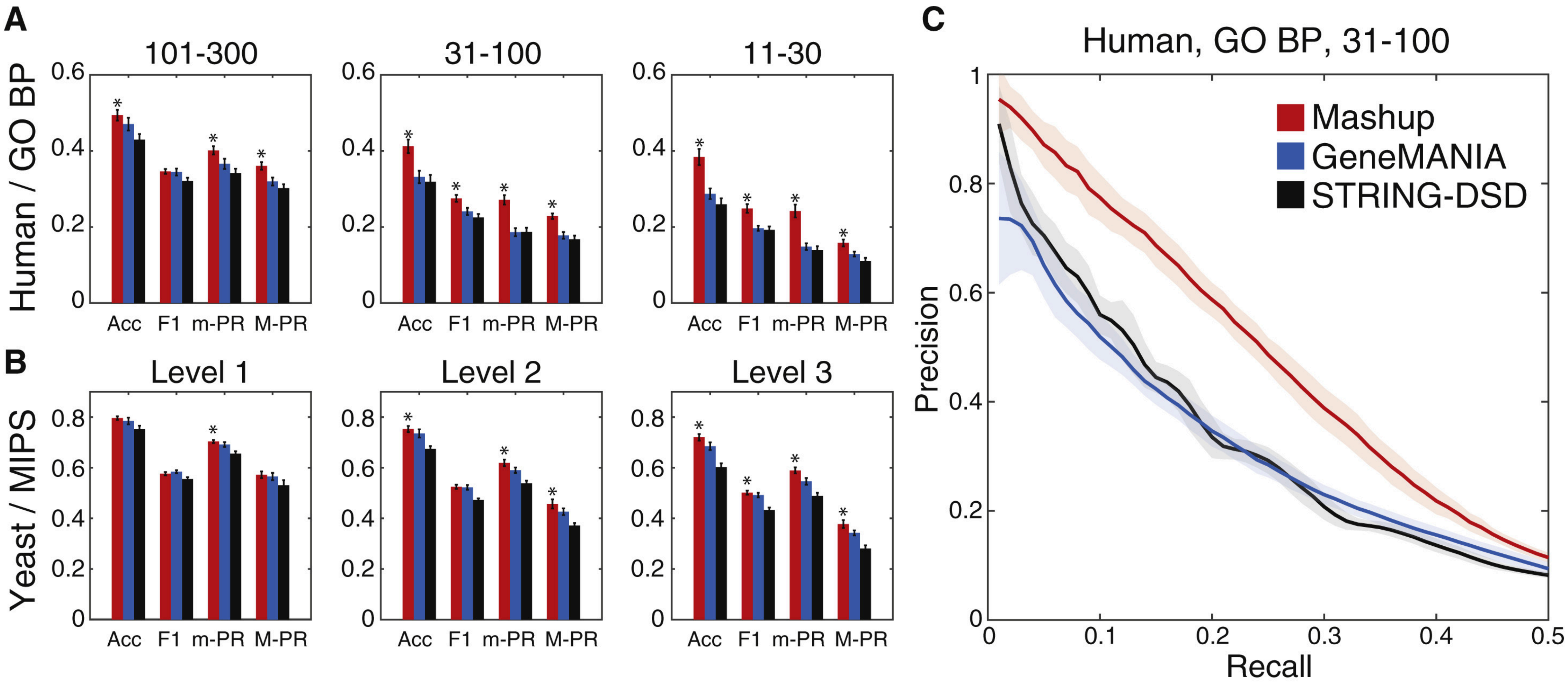
Argelaguet et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets

Mashup: integrating multiple networks



Cho et al. Compact Integration of Multi-Network Topology for Functional Analysis of Genes

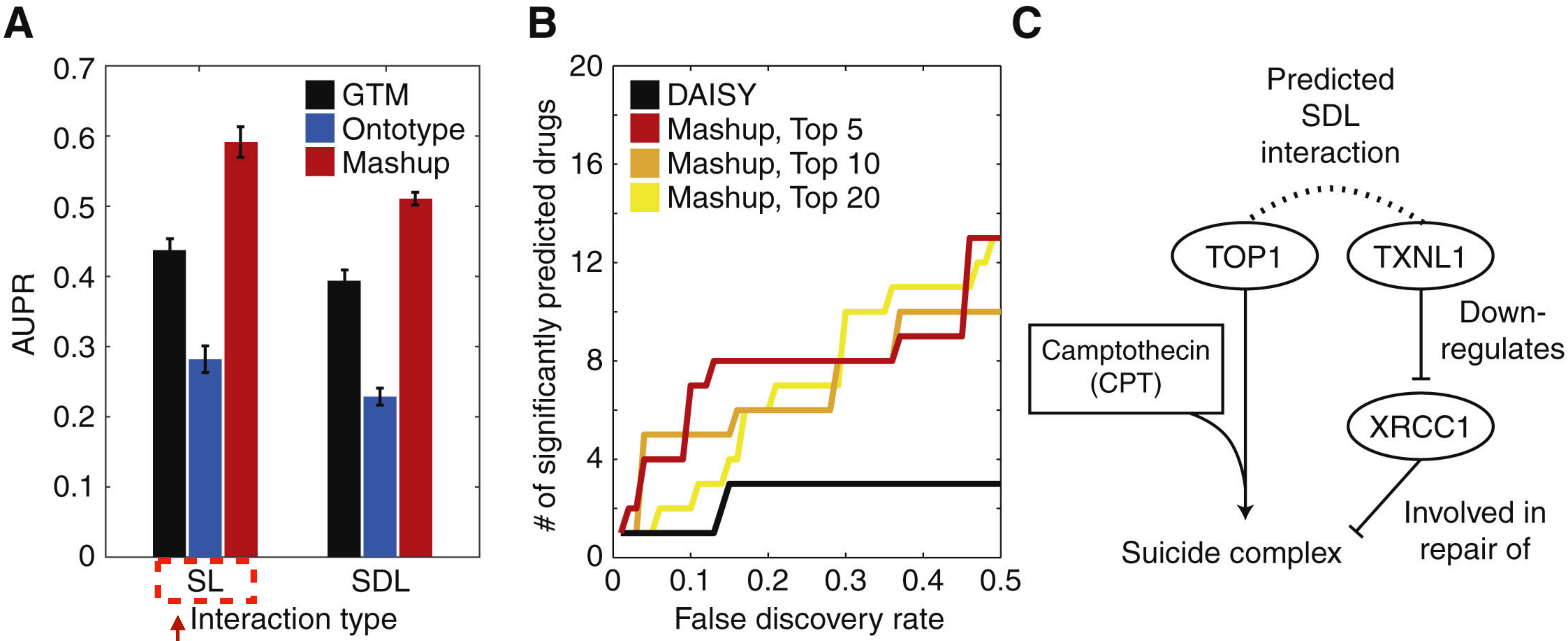
Mashup improves protein function prediction



Protein function prediction is a good benchmark for machine learning algorithms because of it is high-quality and has many annotations. It can be used to evaluate:

- Network-based approach
- Sequence-based approach
- Few-shot/zero-shot learning

Mashup enables genetic interaction prediction

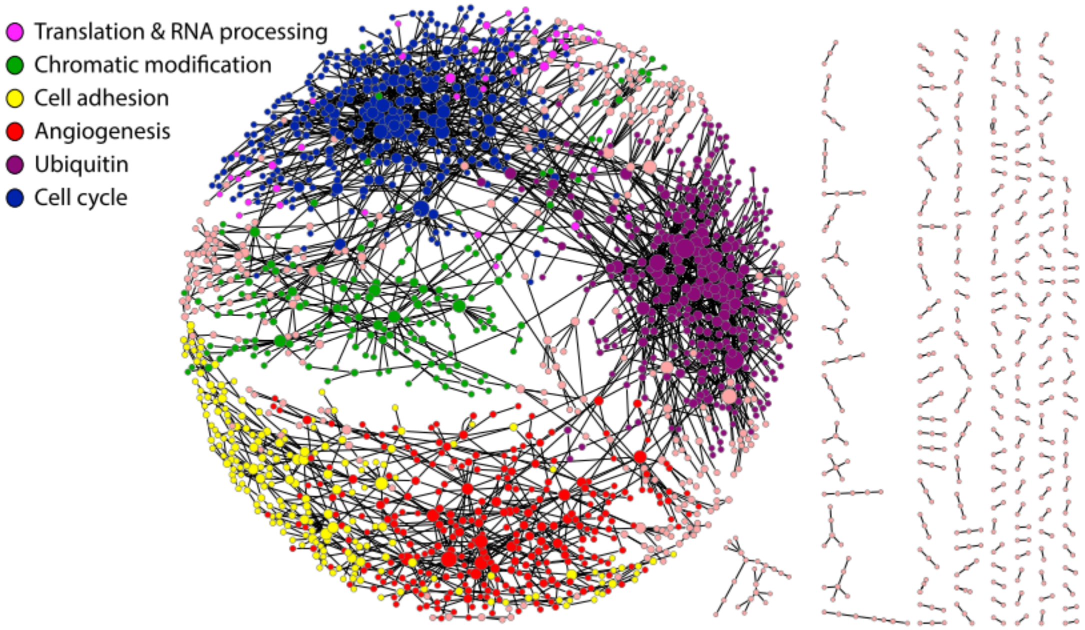


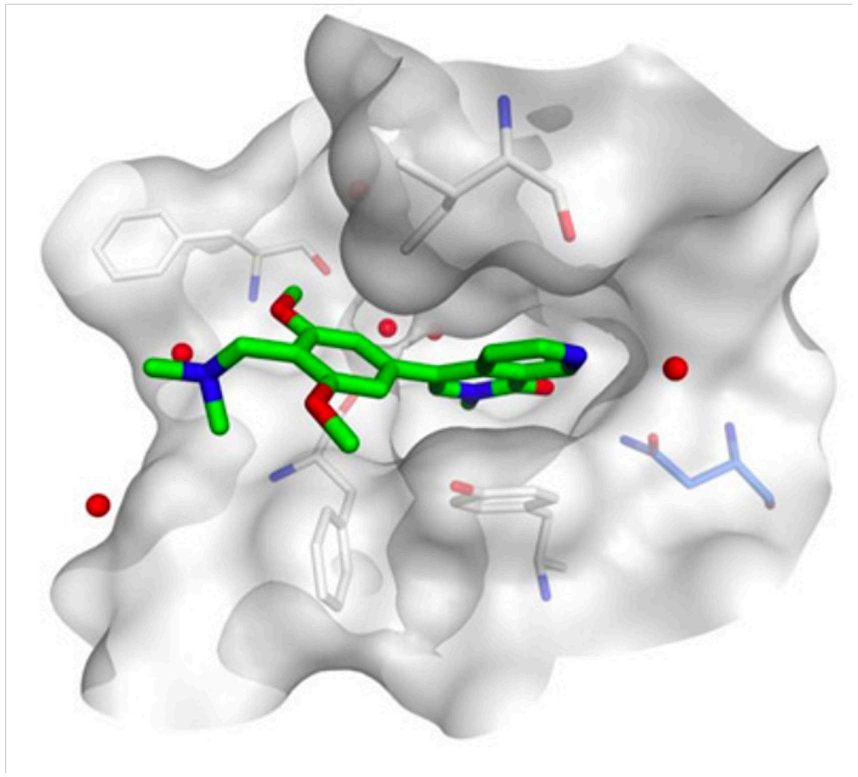
Synthetic Lethality

How to leverage SL to develop (personalized) drug (combination) therapy?

Integrate three sources:

- Mutation data of the patient (mutation A)
- SL network (Gene B has SL effect with Gene A)
- Drug target information (Drug X inhibits Gene B)

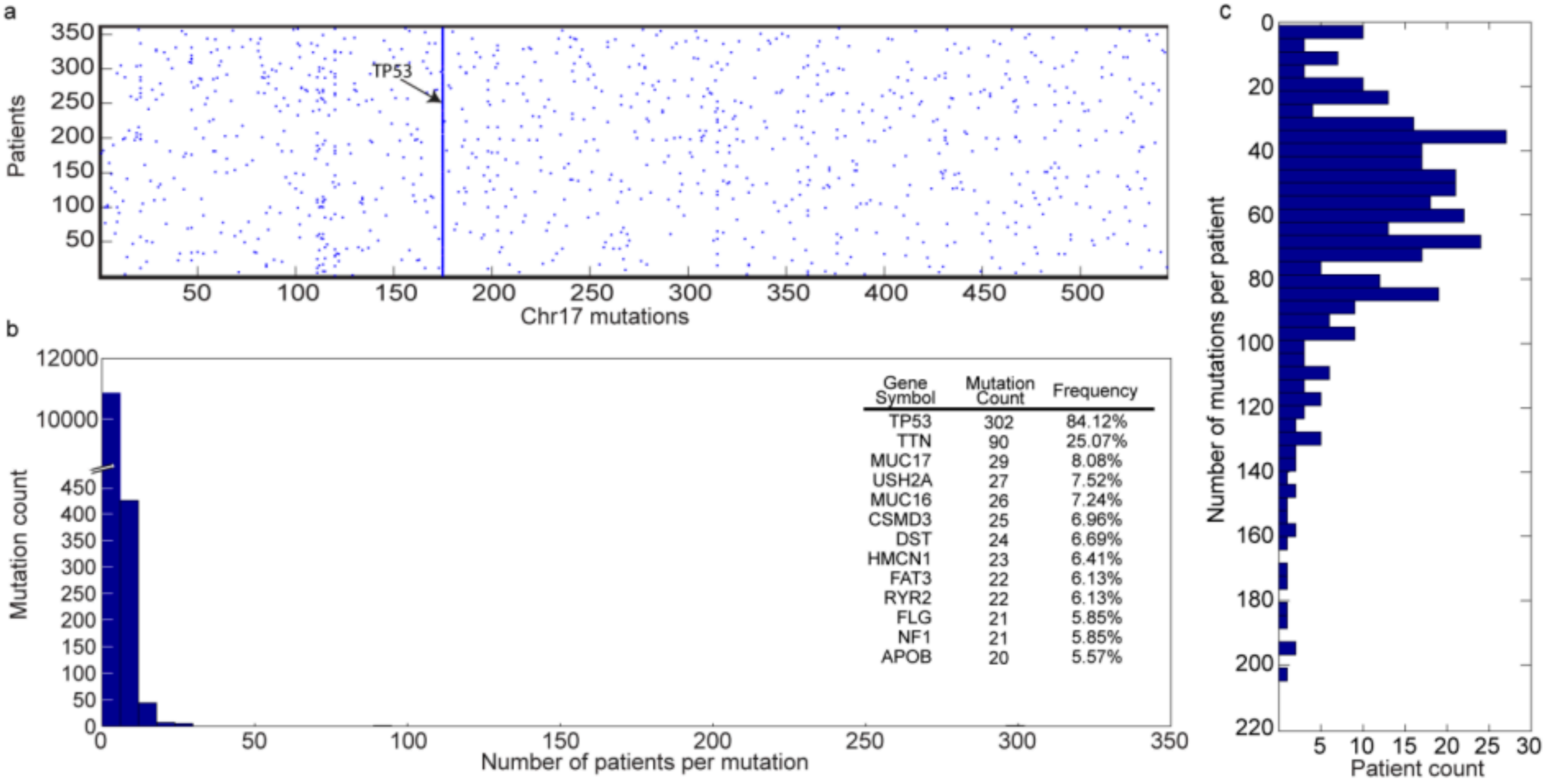




Source: Google image

How to integrate network with a patient matrix

Patient matrix is very sparse
Use network to smooth it



Gene Symbol	Mutation Count	Frequency
TP53	302	84.12%
TTN	90	25.07%
MUC17	29	8.08%
USH2A	27	7.52%
MUC16	26	7.24%
CSMD3	25	6.96%
DST	24	6.69%
HMCN1	23	6.41%
FAT3	22	6.13%
RYR2	22	6.13%
FLG	21	5.85%
NF1	21	5.85%
APOB	20	5.57%

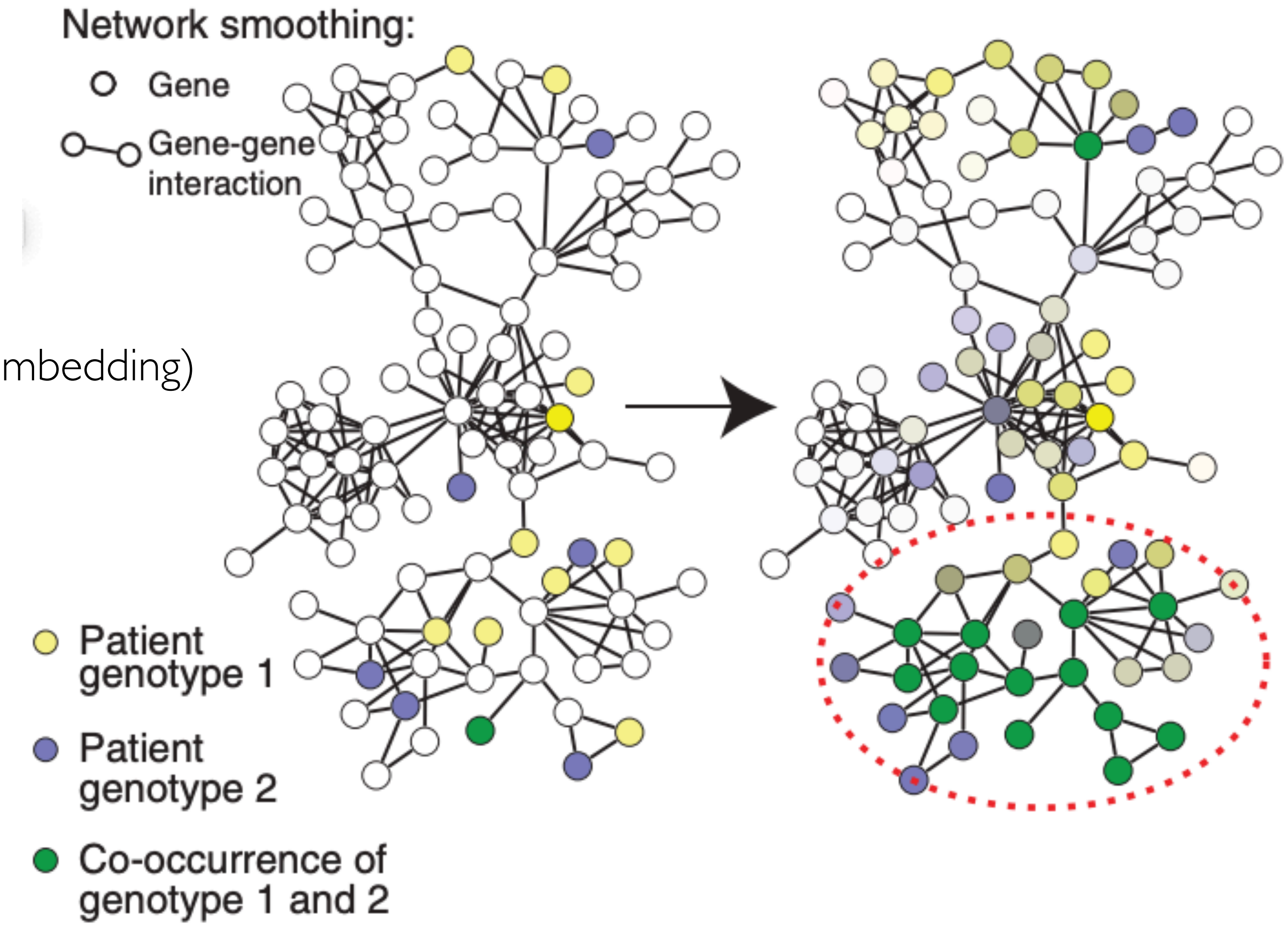
Hofree et al. Network-based stratification of tumor mutations

How to integrate network with a patient matrix

Project each patient's vector to the network
Use diffusion model to find overlapping regions

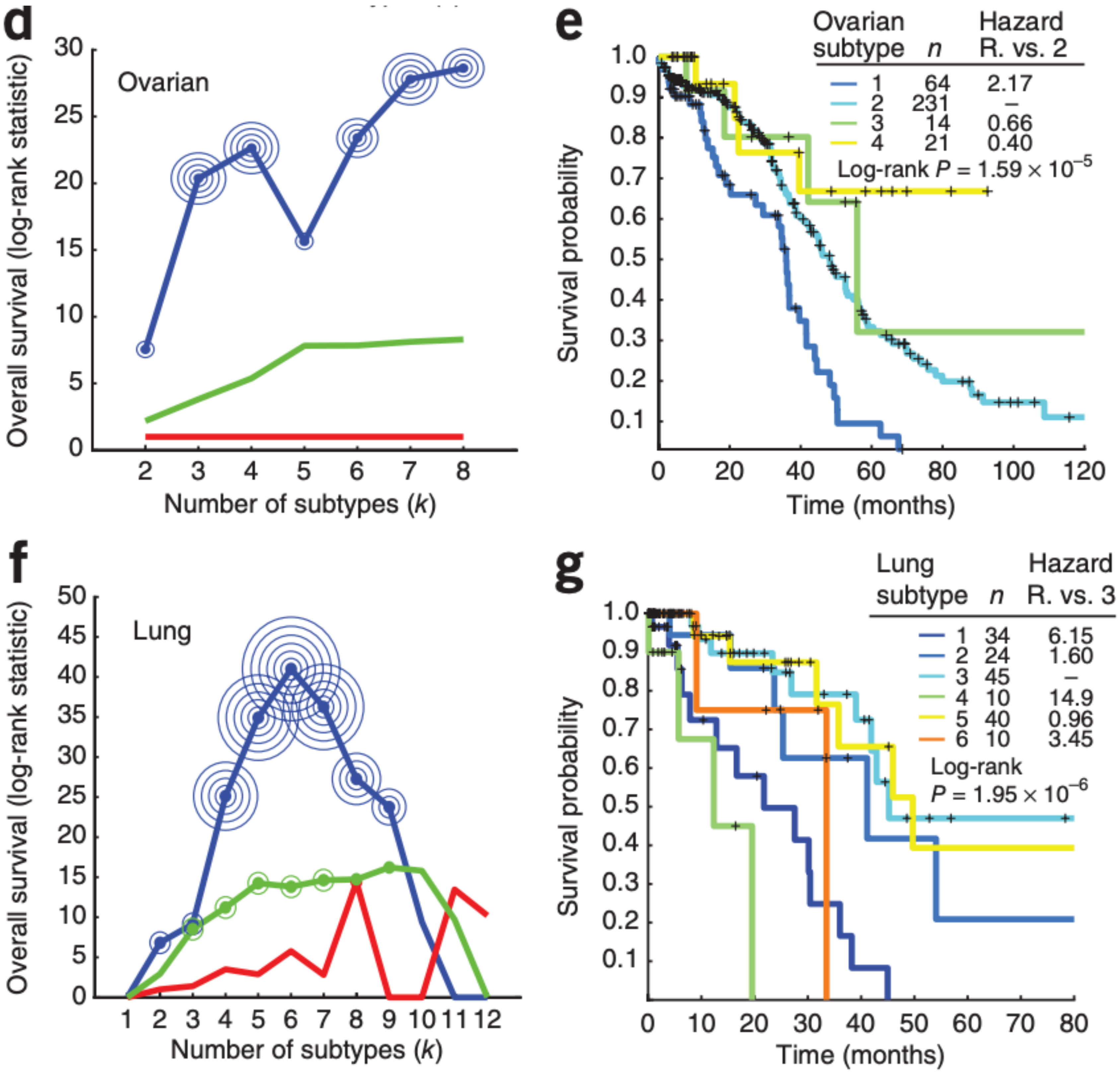
What diffusion models to use?

- Random walk with restart
- Heat diffusion
- Graph neural network (GCN, GAT)
- Network embedding (no node features)
- Other non-euclidean geometry (hyperbolic embedding)



Hofree et al. Network-based stratification of tumor mutations

Results on TCGA: use patient survival data as a benchmark to evaluate patient clustering model



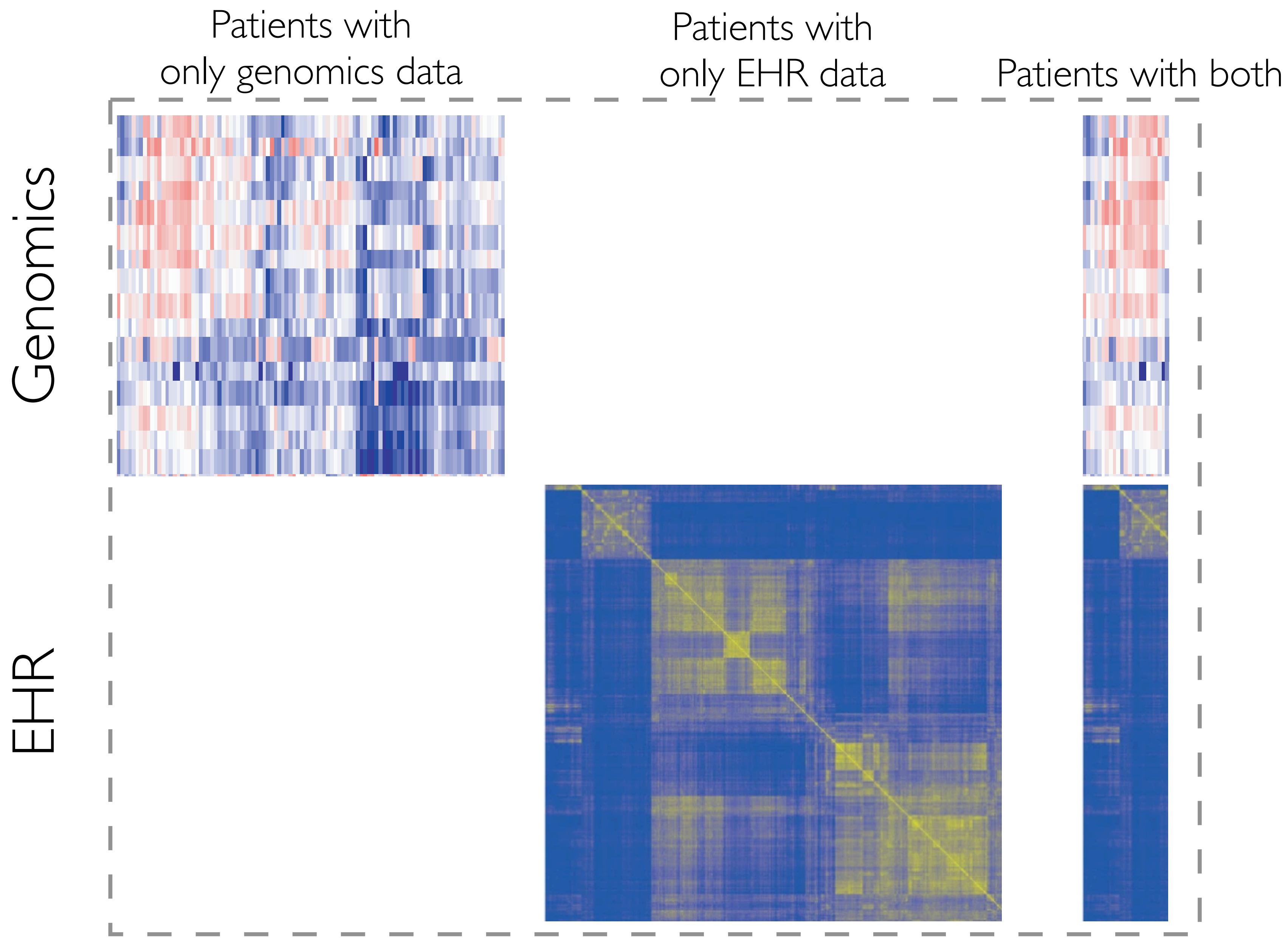
How to cluster patients

DATA SOURCES
Social Interaction
Genetics and molecular studies
EHRs
Biochemical research
Lab tests
Medication
Diet
Environment
Medical images

- **Patient clustering = data integration**
  - Find a "signature" vector for each patient
  - Signature is integrated from different data sources
- **Heterogeneous data integration**
  - General challenges: Heterogenous, missing values, noise, privacy
- **Precision medicine specific data integration challenges:**
  - Batch effects (different preprocessing pipelines, sequencing techniques, reference ranges)
  - Unpaired data (some patients only have genomics data, some patients only have EHR data, very few patients have both)

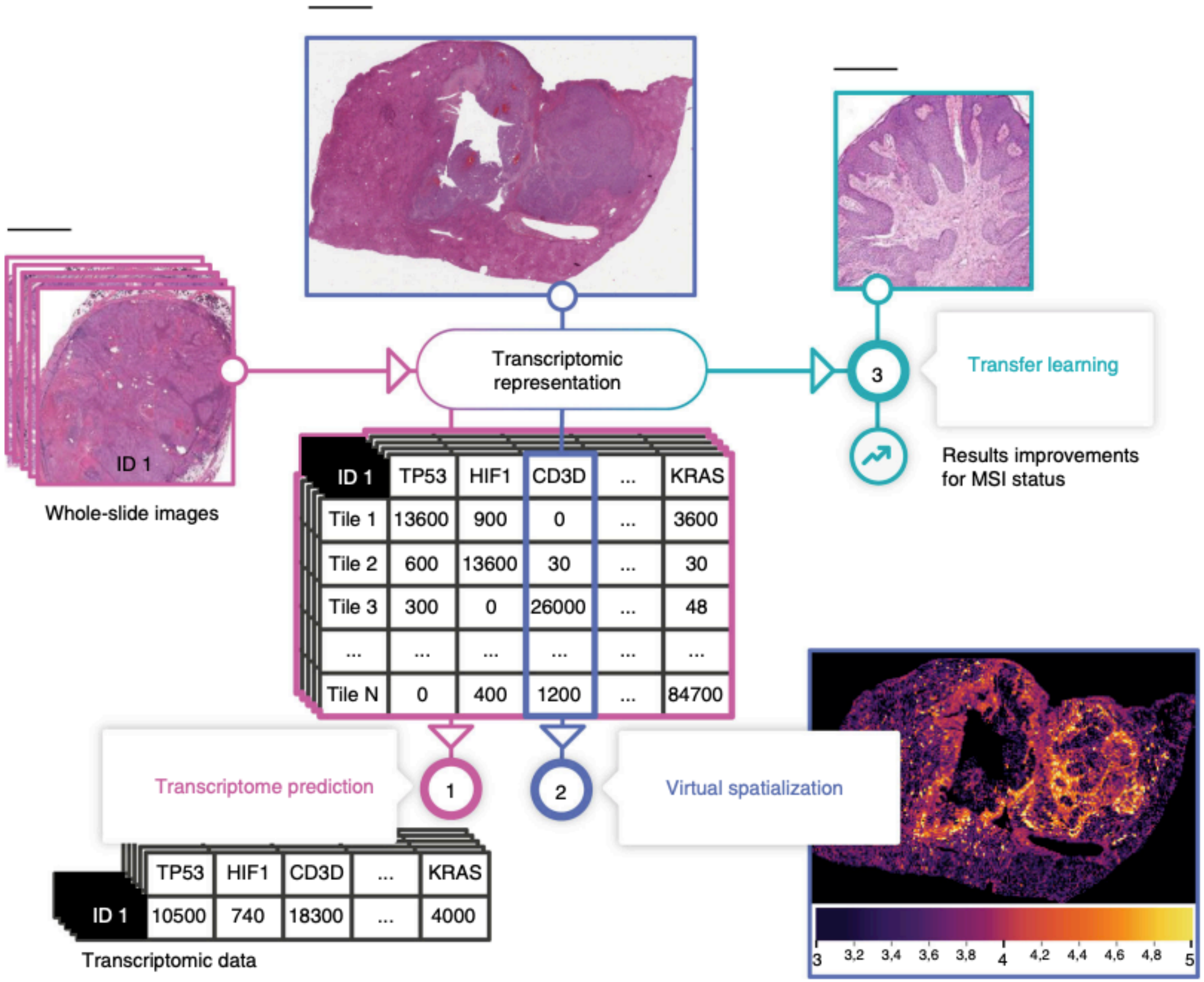
Grapov et al. Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine

How to handle unpaired data



ML question: How to integrate all these patients?

Translation between features: generate expression from image

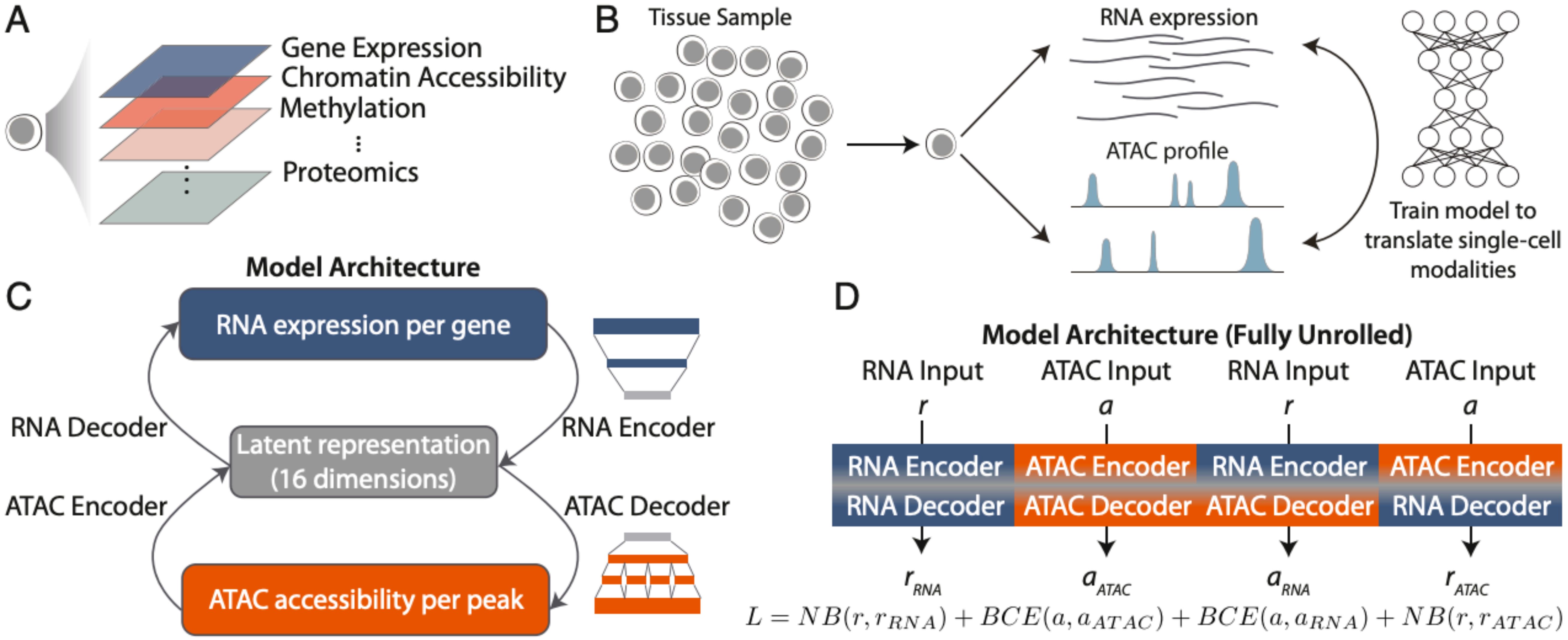


ID 1	TP53	HIF1	CD3D	...	KRAS
Tile 1	13600	900	0	...	3600
Tile 2	600	13600	30	...	30
Tile 3	300	0	26000	...	48
...	...	...	...	...	...
Tile N	0	400	1200	...	84700

	TP53	HIF1	CD3D	...	KRAS
ID 1	10500	740	18300	...	4000

Schmauch et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images

Translation between features: RNA to ATAC translation



$$L = NB(r, r_{RNA}) + BCE(a, a_{ATAC}) + BCE(a, a_{RNA}) + NB(r, r_{ATAC})$$

Wu et al. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution