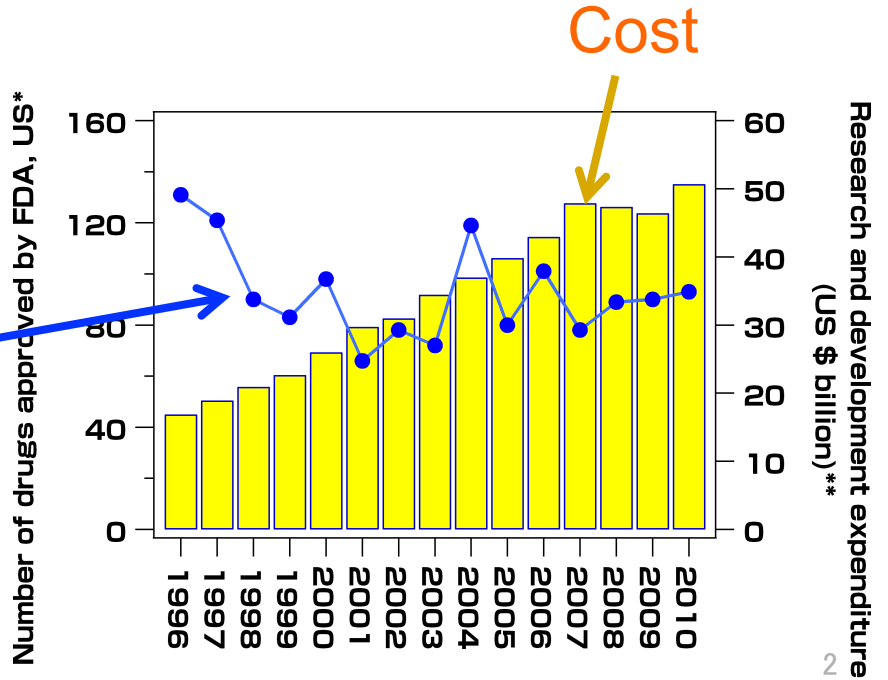
Data-driven drug discovery for a variety of diseases by machine learning
機械学習による様々な疾患に対するデータ駆動型の創薬

Yoshihiro Yamanishi
山西芳裕
Kyushu University Medical Institute of Bioregulation
九州大学　生体防御医学研究所

Drug discovery is very difficult
創薬は難しく、時間も費用もかかる

- Time consuming: 10-15 years
- High cost: about 1 billion $
- High risk: result in failure
  - Insufficient efficacy
  - Difficult production
  - Unexpected toxicity



*http://www.fda.gov, **http://www.phrma.org

Eco-Pharma (Drug repositioning)
エコファーマ（ドラッグリポジショニング）

- Identification of new therapeutic effects (i.e., new applicable diseases) of existing drugs.

  既存薬の新しい効能を発見し、別の疾患の薬として開発

- Rich information on existing drugs available (e.g., safety for human, manufacturing process).

  豊富な情報（人での安全性や製造法など）がある

- Fast development and low risk.

  高速かつ低リスクな創薬が可能

The cost can be reduced in terms of time, risk, and expenditure.
時間、リスク、費用を大幅に削減できる

Process	Traditional approach (10～17 years)	EcoPharma in this study (3～9 years)
1. Screen compounds	○	○
2. Optimize chemical structures	○	-
3. Confirm safety with animals	○	-
4. Confirm efficacy with animals	○	-
5. Confirm safety for human	○	-
6. Confirm efficacy for human	○	○
7. Approve	○	○

Examples 例

- Sildenafil (Viagra)　シルデナフィル（バイアグラ）

  Angina 狭心症

  → Erectile dysfunction 男性機能障害

  → Pulmonary hypertension 肺高血圧症

- Minoxidil (Riup, Rogaine) ミノキシジル

  Hypertension 高血圧

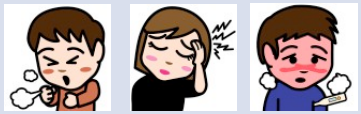
  → Alopecia (hair loss) 脱毛症

Previously, it has been dependent on serendipity.
これまでは偶然の発見に大きく依存していた

Goal of this study　本研究の目標

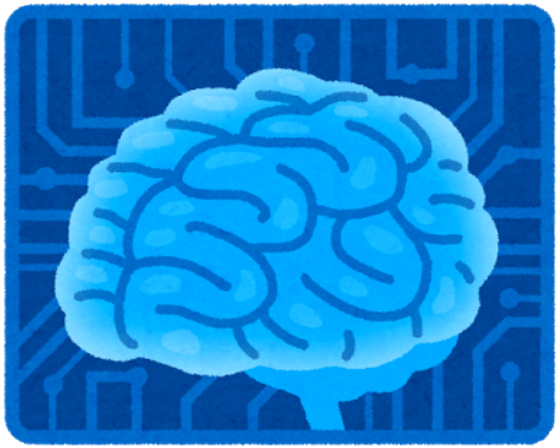
- Automatic prediction of new drug effects from various biomedical big data.

  ビッグデータから薬物の新規効能を自動的に予測

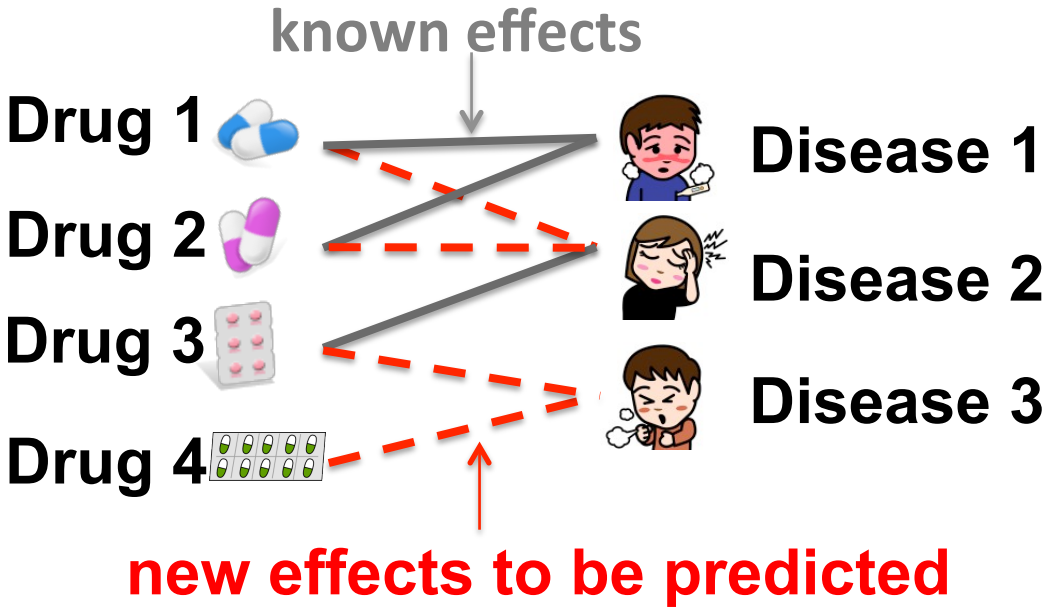
Object	Data
Drugs/compounds 	chemical structures, side effects, clinical reports, drug-induced gene expression profiles, compound-protein interactions
Proteins/genes 	amino acid sequences, pathways, functional motifs, domains, structures, physiological roles, pathological roles
Diseases 	disease-causing genes, disease pathways, environmental factors, biomarkers, gene expression profiles of patients, disease complication

AI-based drug discovery
AI創薬

Machine learning methods to predict new associations between drugs and diseases
薬物と疾患の関係を機械学習で予測する

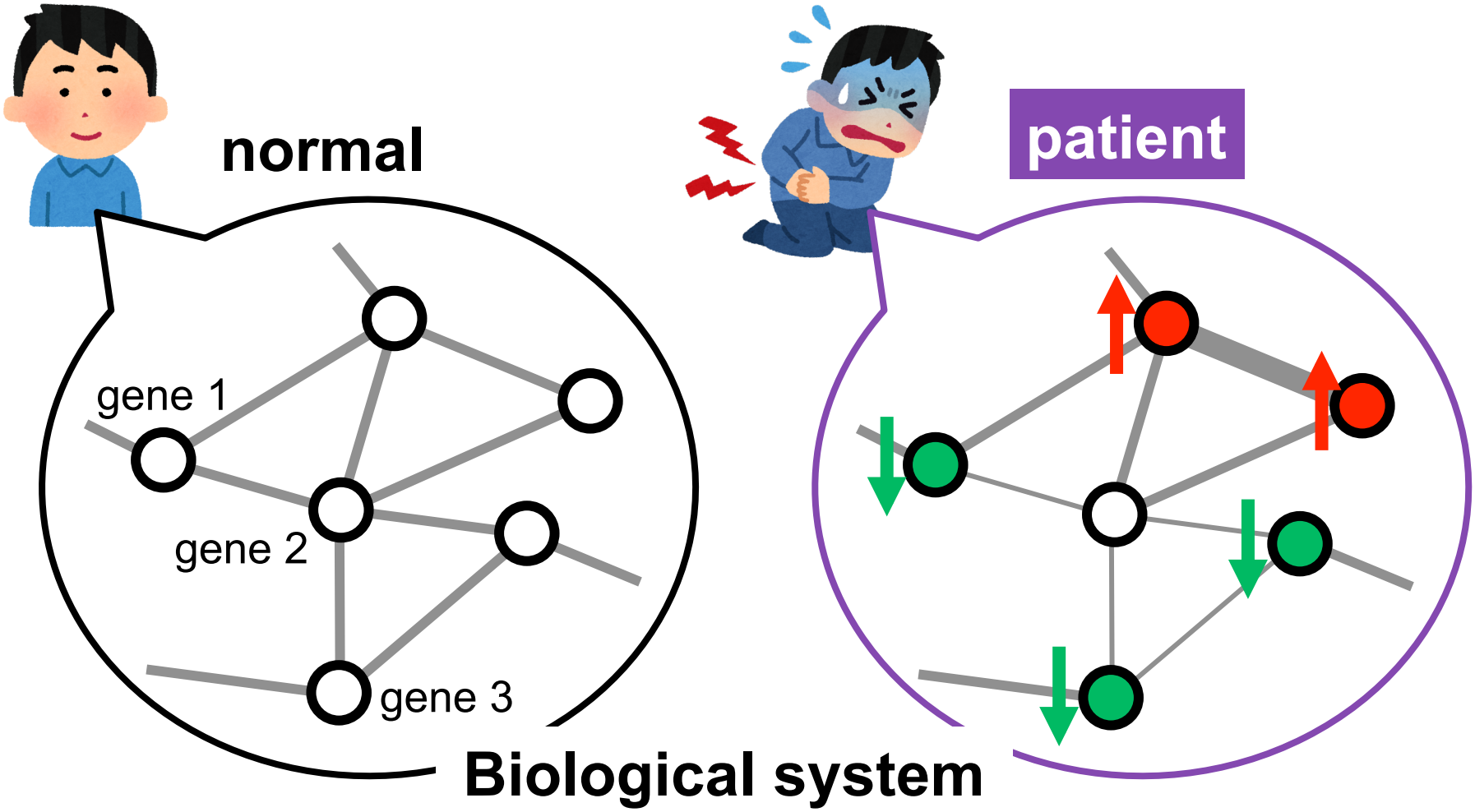


$$f(\mathbf{x},\mathbf{y}) = \mathbf{w}^T \phi(\mathbf{x},\mathbf{y})$$

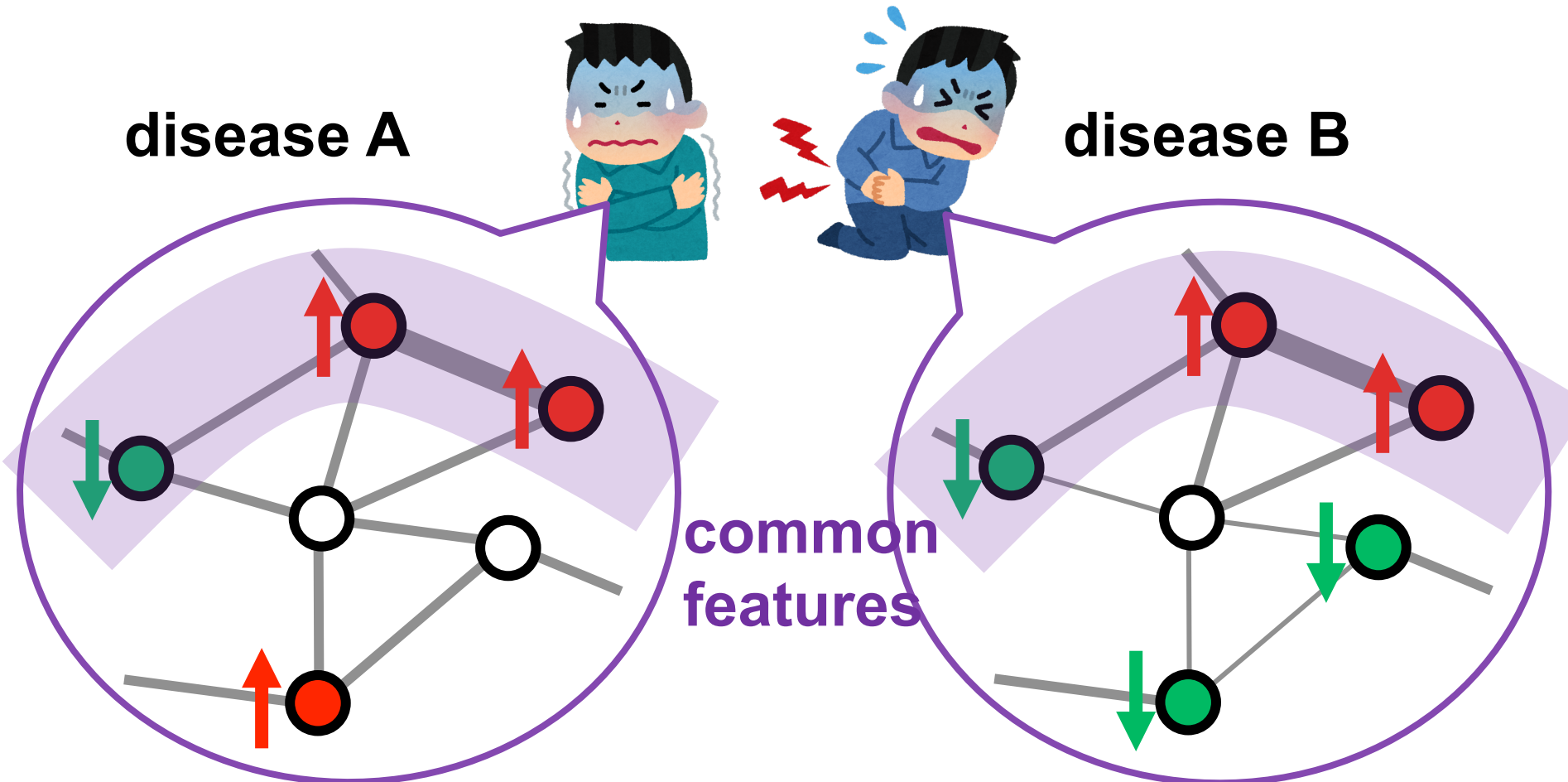


Molecular understanding of a variety of diseases
様々な疾患の分子的理解が進んできた

- disease-causing genes 病因遺伝子
- disordered pathways 異常パスウェイ
- environmental factors 環境因子
- abnormal gene expression 発現異常遺伝子

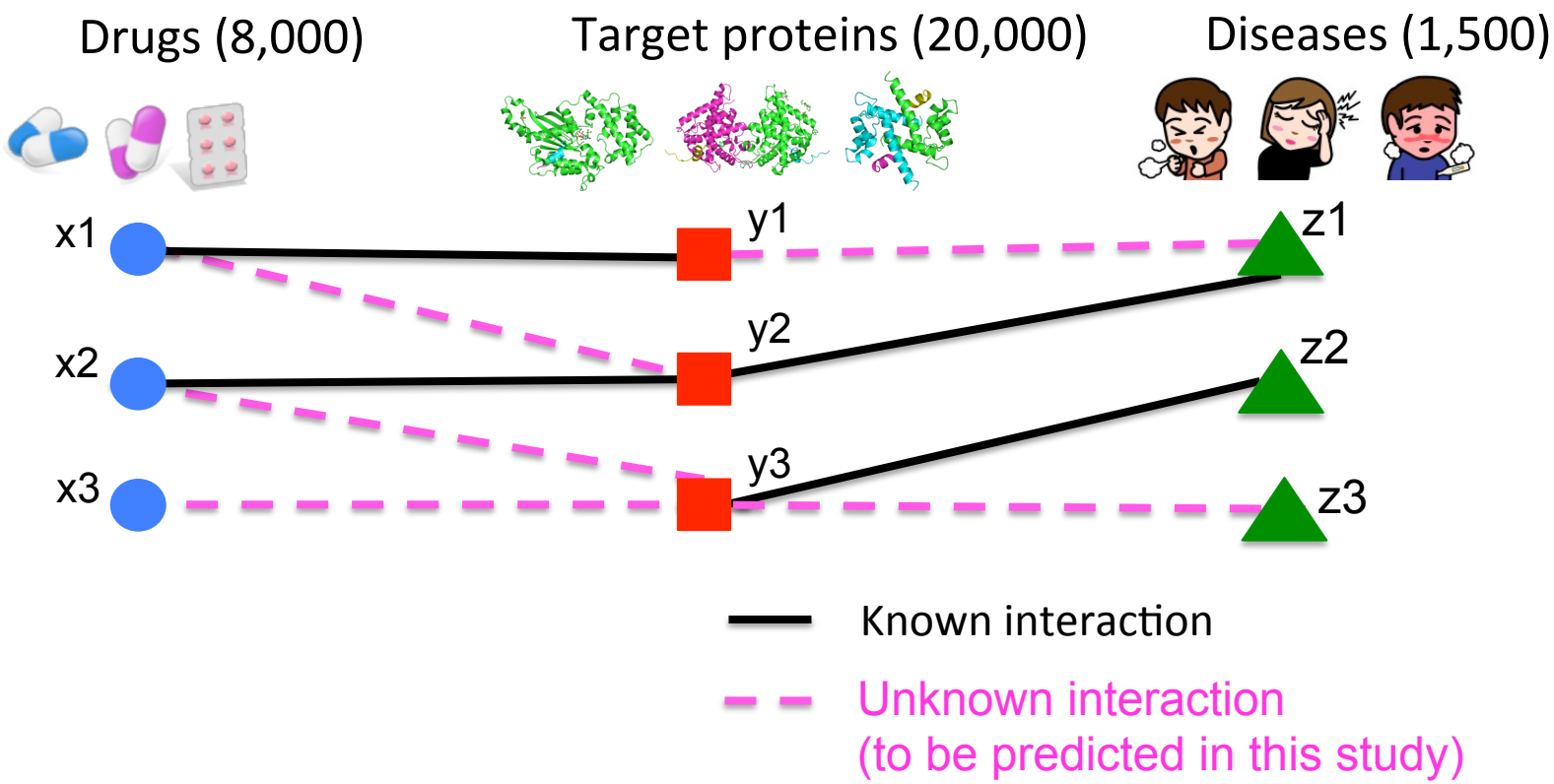


Characteristic molecular features are often shared among different diseases
分子的特徴は疾患間で共通する場合がある



A representation of the drug mechanism

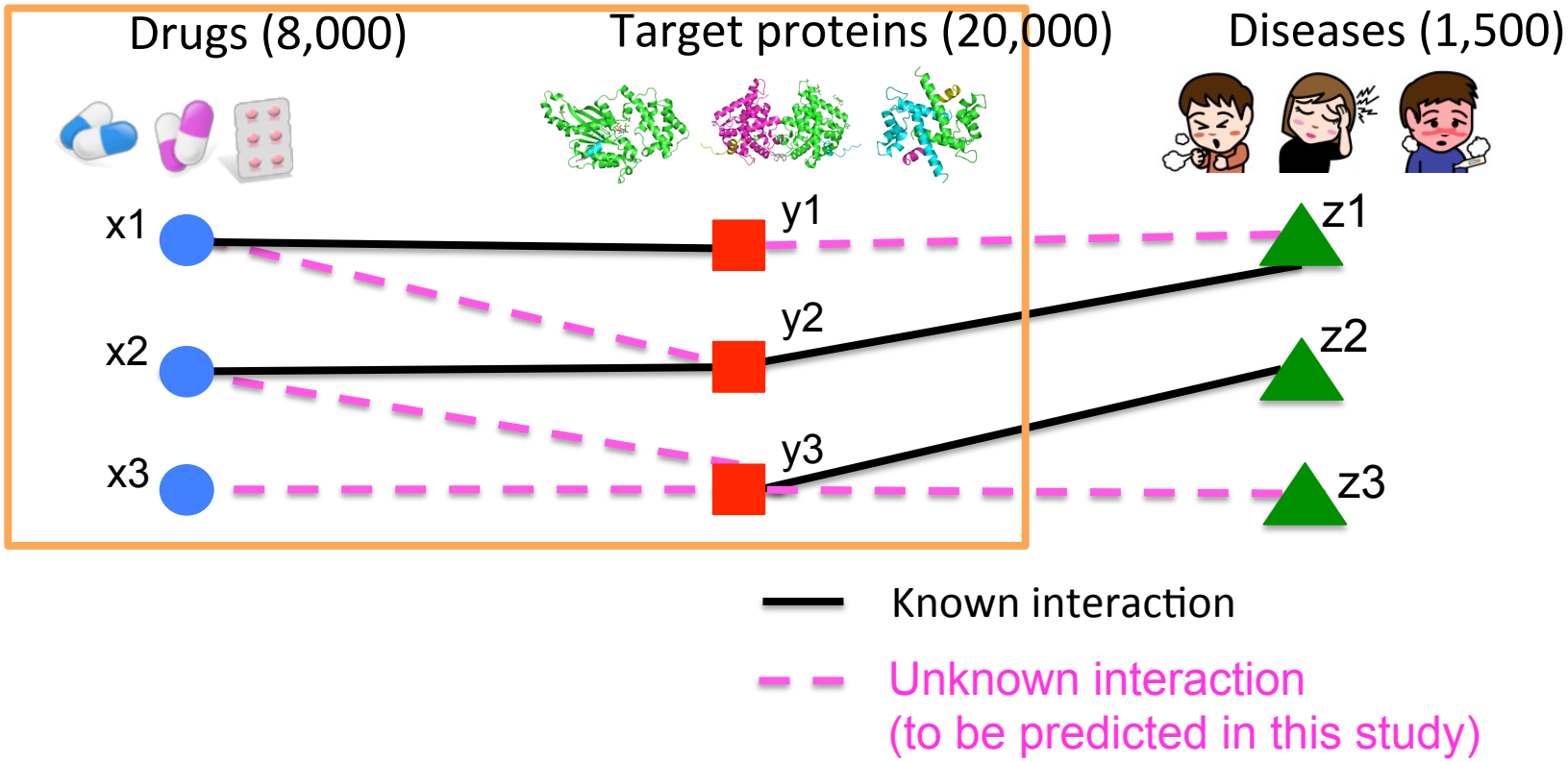
薬物はタンパク質に相互作用し、疾患に対する効能を発揮する



Proposed method　提案手法

Prediction of drug-protein-disease network with machine learning

薬物が、どのタンパク質に相互作用し、どの疾患に効くかを予測

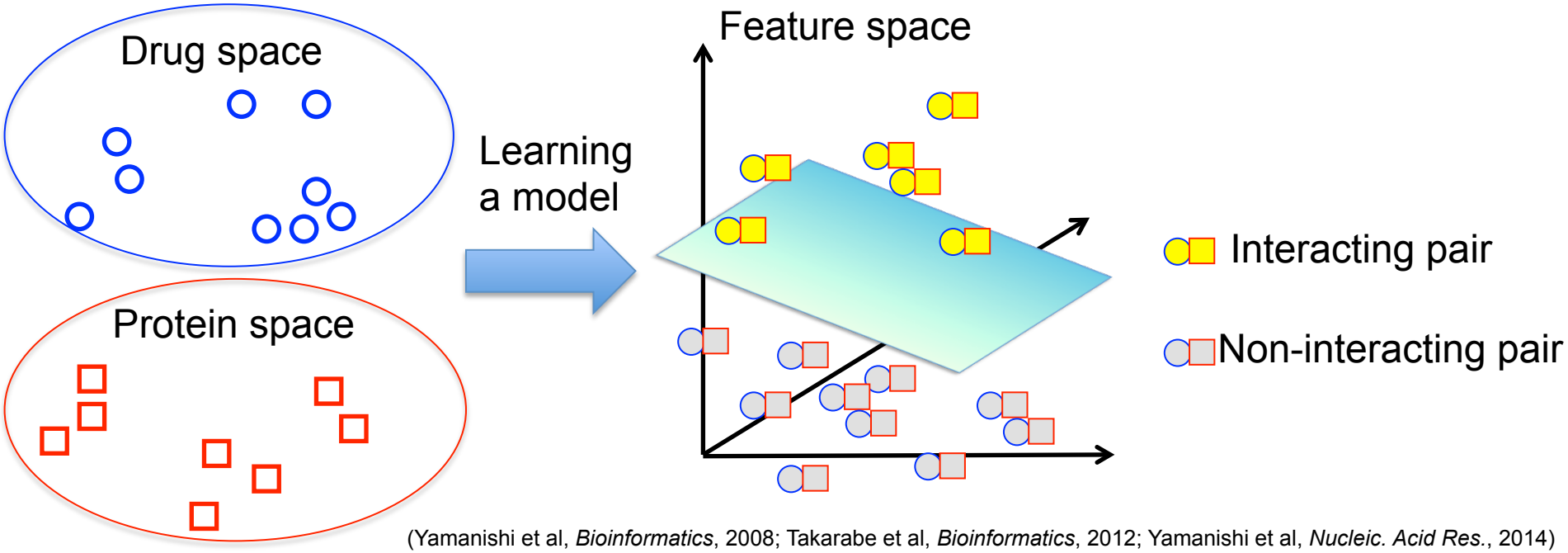


Drug-protein interaction prediction
薬物･タンパク質相互作用の予測

A pairwise model for any drug-protein pair $(\mathbf{x}', \mathbf{z}')$:

$$f(\mathbf{x}', \mathbf{z}') = \sum_{i=1}^{n_x} \sum_{j=1}^{n_z} a_{ij} k((\mathbf{x}_i, \mathbf{z}_j), (\mathbf{x}', \mathbf{z}')) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_z} a_{ij} \underbrace{k_x(\mathbf{x}_i, \mathbf{x}')}_{\text{Drug similarity}} \underbrace{k_z(\mathbf{z}_j, \mathbf{z}')}_{\text{Protein similarity}}$$

Step 1: Pairwise learning



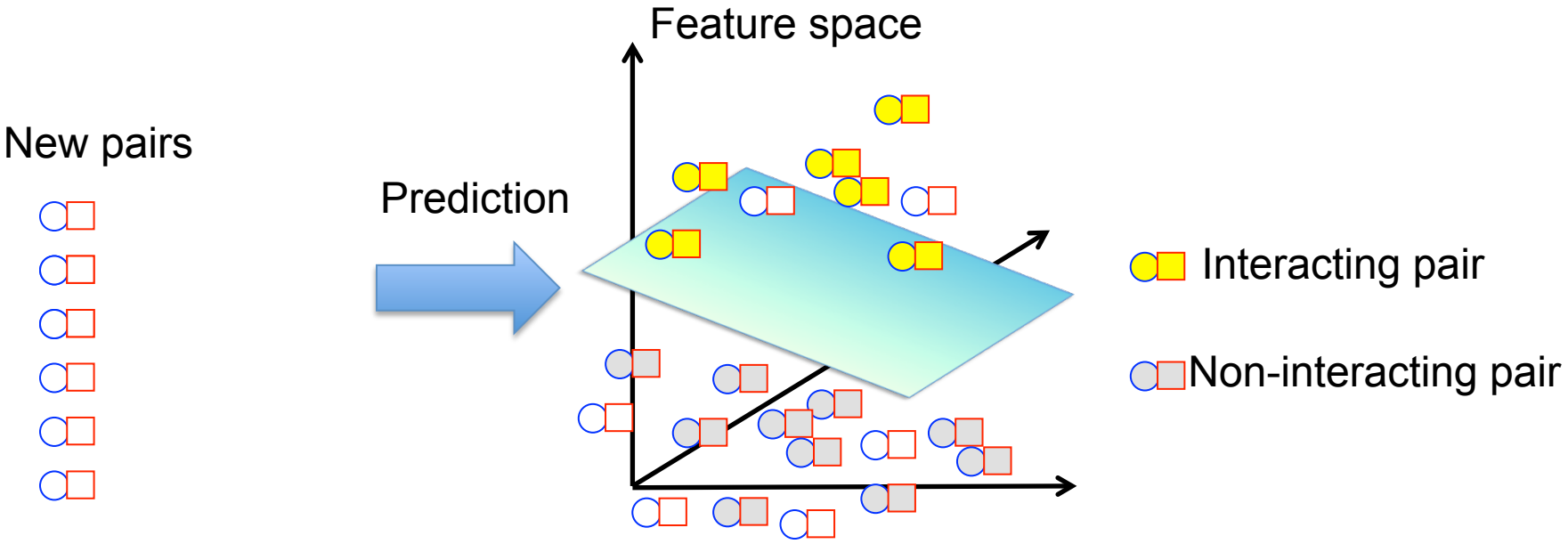
(Yamanishi et al, *Bioinformatics*, 2008; Takarabe et al, *Bioinformatics*, 2012; Yamanishi et al, *Nucleic. Acid Res.*, 2014)

Drug-protein interaction prediction
薬物･タンパク質相互作用の予測

A pairwise model for any drug-protein pair $(\mathbf{x}', \mathbf{z}')$:

$$f(\mathbf{x}', \mathbf{z}') = \sum_{i=1}^{n_x} \sum_{j=1}^{n_z} a_{ij} k((\mathbf{x}_i, \mathbf{z}_j), (\mathbf{x}', \mathbf{z}')) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_z} a_{ij} \underbrace{k_x(\mathbf{x}_i, \mathbf{x}')}_{\text{Drug similarity}} \underbrace{k_z(\mathbf{z}_j, \mathbf{z}')}_{\text{Protein similarity}}$$

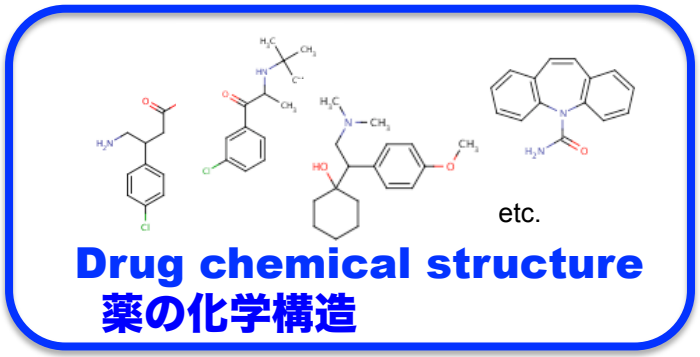
Step 2: Predicting new interactions

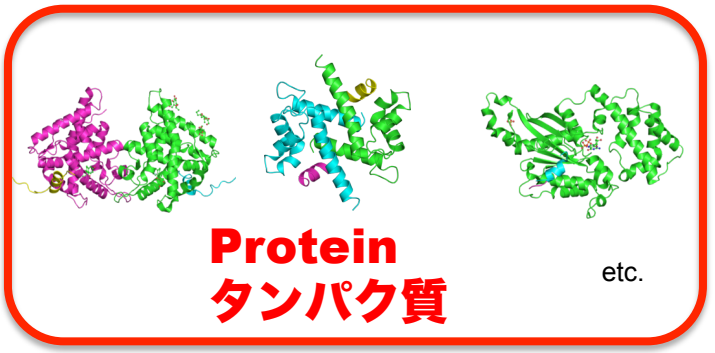


(Yamanishi et al, *Bioinformatics*, 2008; Takarabe et al, *Bioinformatics*, 2012; Yamanishi et al, *Nucleic. Acid Res.*, 2014)

Chemical structure-based approach
化学構造に基づくアプローチ

Strategy: Chemically similar drugs are predicted to interact with similar target proteins



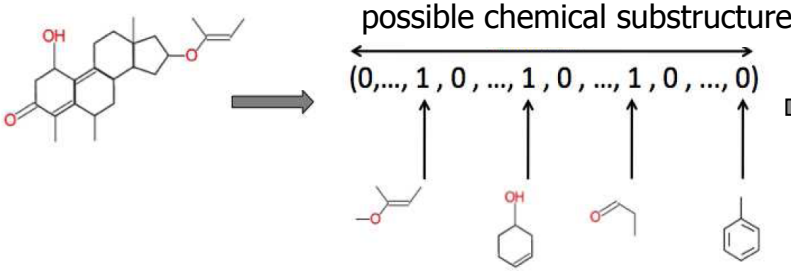


Drug similarity

$$k_x(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{for } i, j = 1, 2, \ldots, n_x$$

475,692 KCF-S substructures
(Kotera et al, *BMC Syst. Biol.*, 2013)

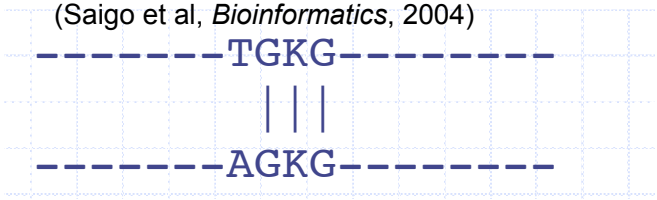


Protein similarity

$$k_z(\mathbf{z}_i, \mathbf{z}_j)$$

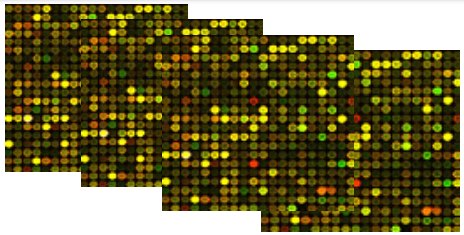
$$\text{for } i, j = 1, 2, \ldots, n_z$$

Local sequence alighnment kernel
(Saigo et al, *Bioinformatics*, 2004)



Gene expression-based approach
遺伝子発現に基づくアプローチ

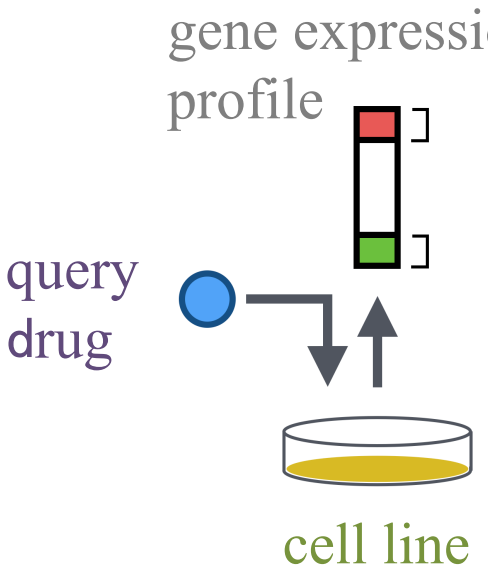
Strategy: Transcriptionally similar drugs are predicted to interact with similar target proteins



Drug-induced gene expression
薬物応答遺伝子発現

Protein
タンパク質

etc.



Each drug is represented by a gene expression profile in which each element is the ratio of drug treatment against control based on LINCS (public database)

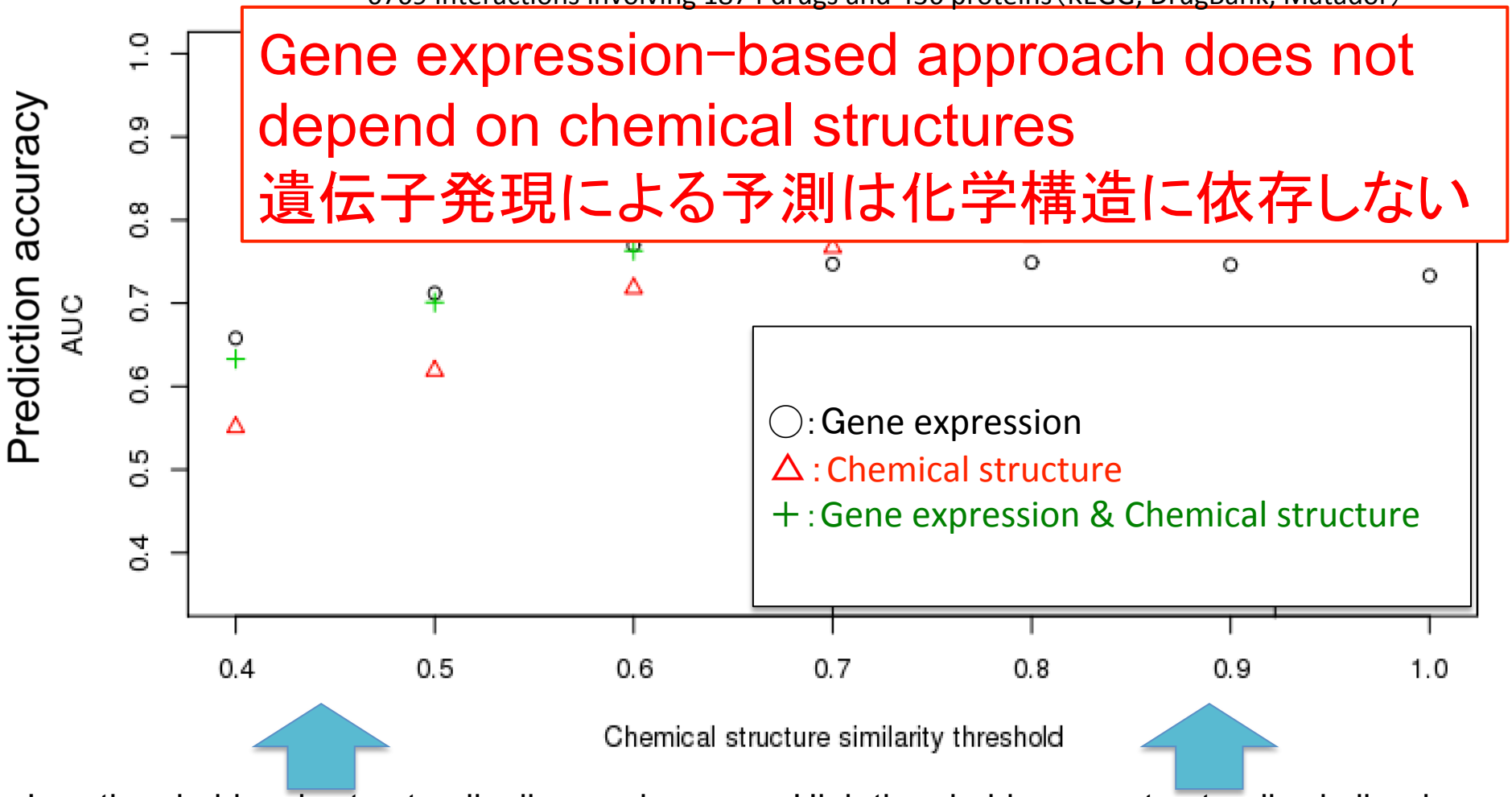
Drug similarity:

$$\mathbf{x} = (x_1, x_2, \cdots, x_{22276})^T$$

→ correlation coefficient

Performance evaluation on several benchmark datasets of different chemical diversities
化学構造の多様性を考慮して性能評価

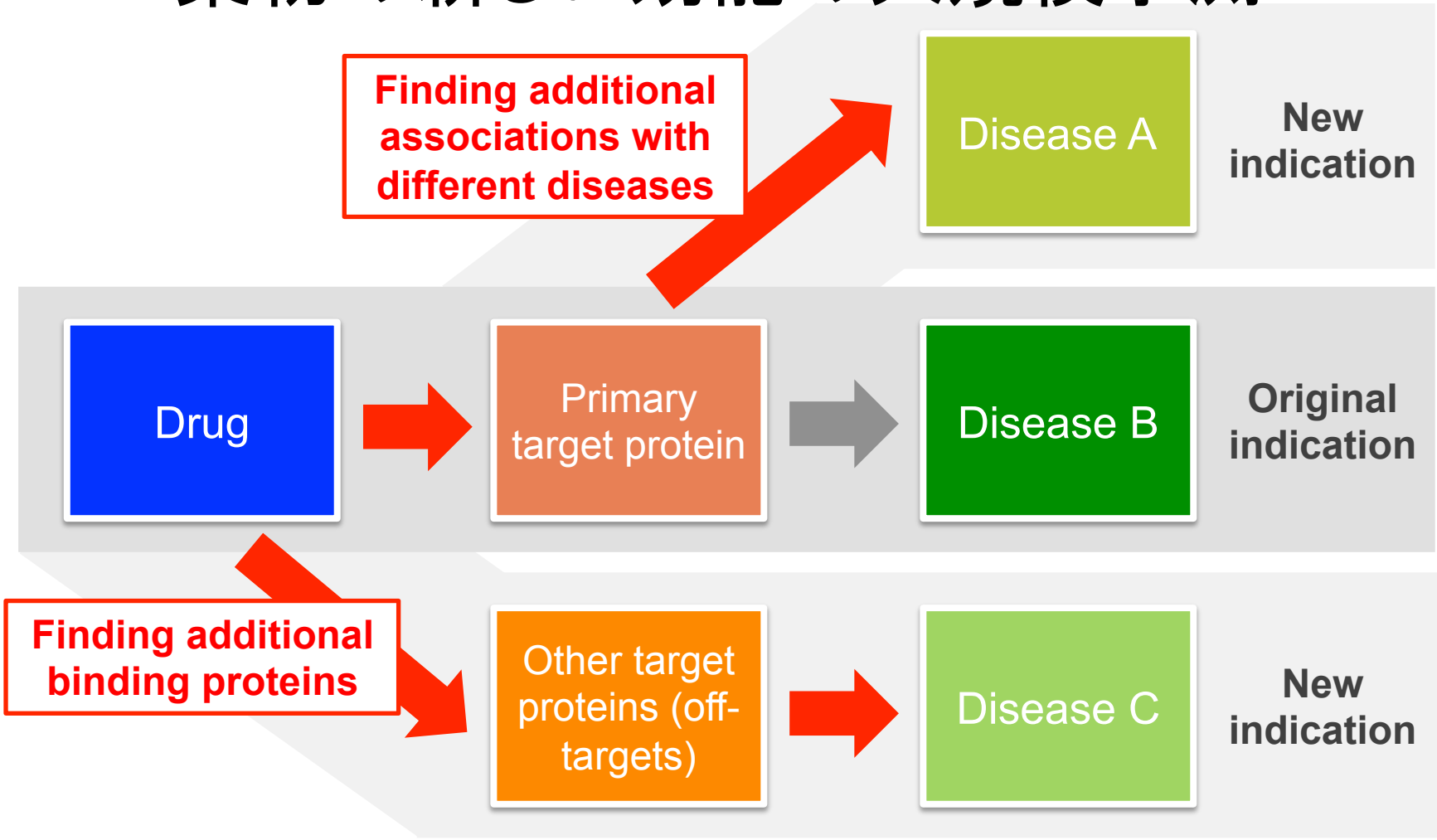
6769 interactions involving 1874 drugs and 436 proteins（KEGG, DrugBank, Matador）



Low threshold: only structurally diverse drugs

High threshold: many structurally similar drugs

Large-scale prediction of new drug indications
薬物の新しい効能の大規模予測



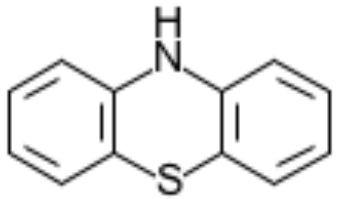
8270 drugs in Japan, US, and EU
1401 diseases

196,048 new drug-disease associations
involving 6301 drugs and 762 diseases

An example of gene expression-based prediction
遺伝子発現情報による予測例

- Phenothiazine (antipsychotic drug)
  フェノチアジン（抗精神病薬）
  - Predicted indication: Prostate cancer 前立腺がん
  - Estimated protein: AR (androgen receptor)

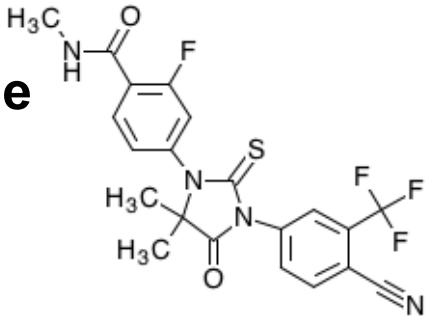
Phenothiazine



D02601

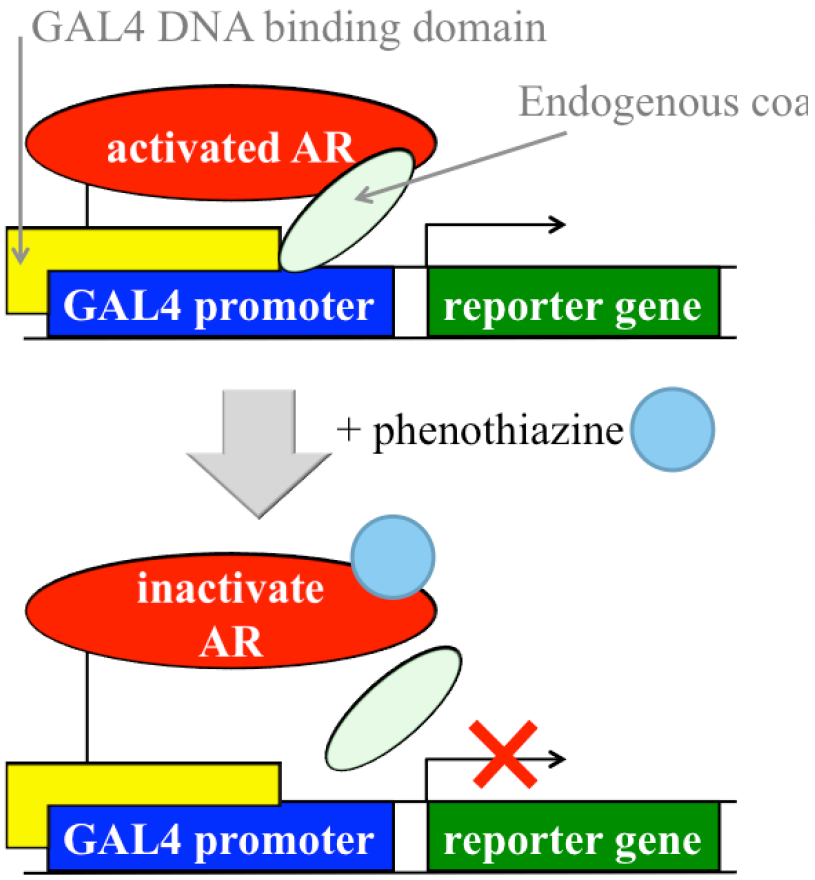
Similar compound in the learning set

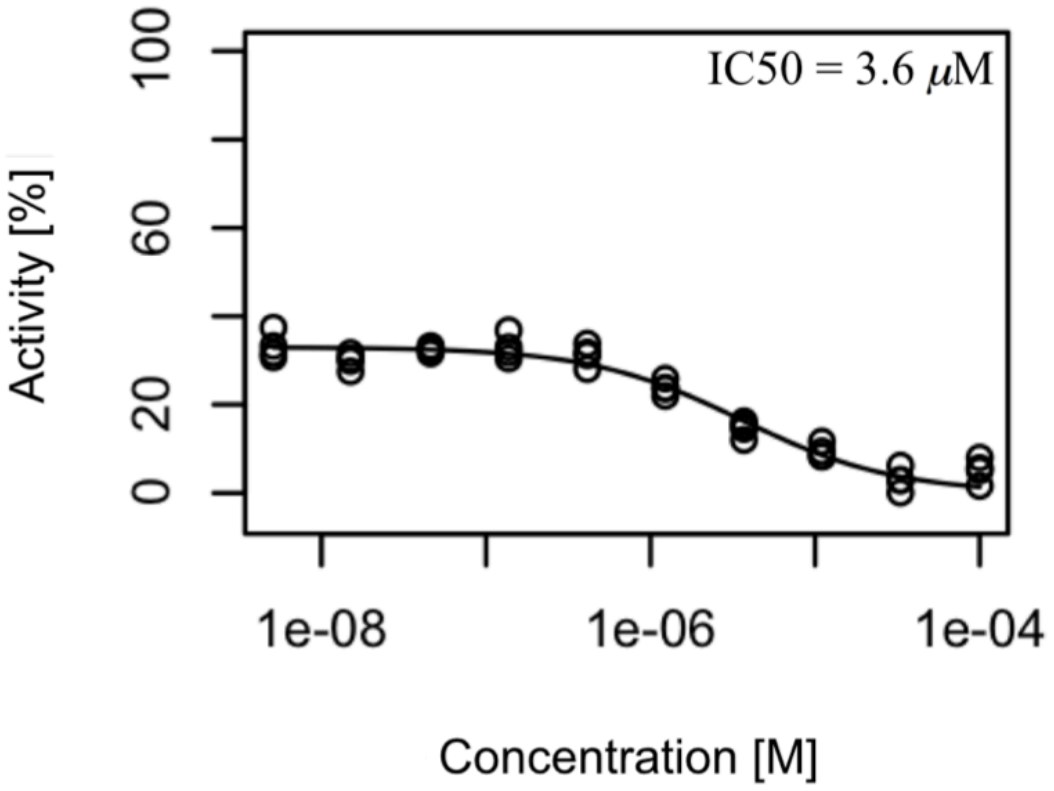
Enzalutamide



D10218

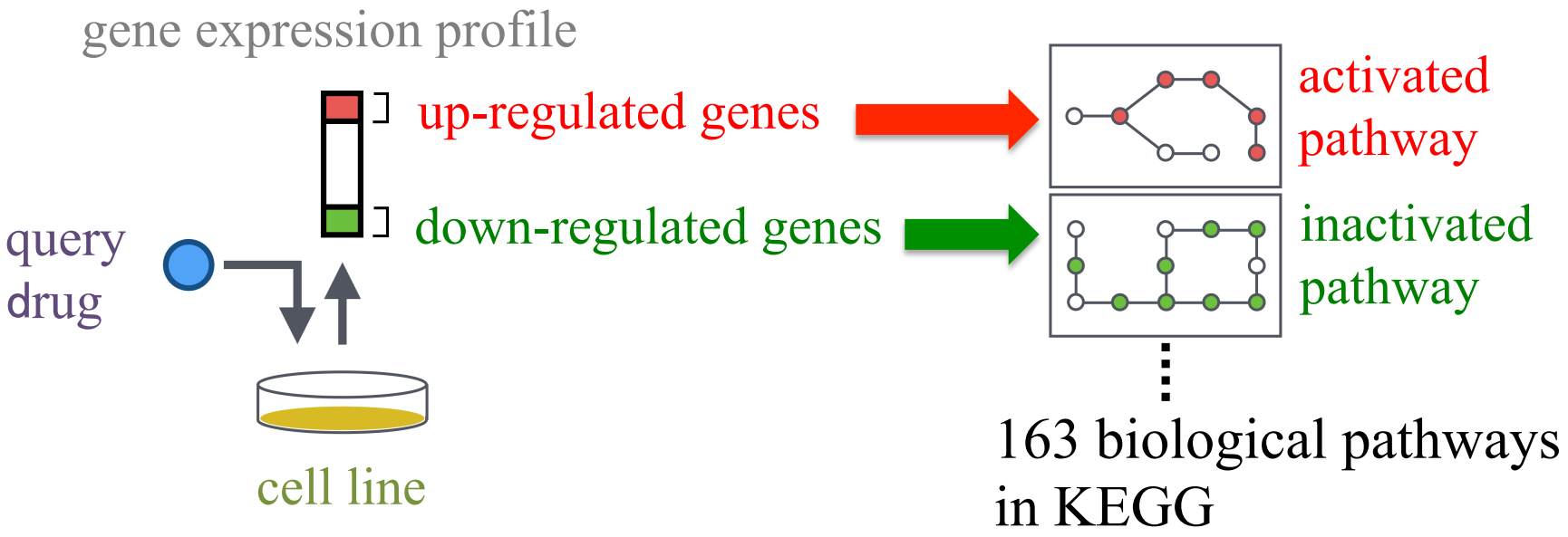
The predicted drug-protein interaction was experimentally confirmed
予測結果はウェット実験で確認できた





(Iwata et al, *Scientific Reports*, 2017)

Elucidating activities of pathways (functional modules)
パスウェイ（遺伝子機能モジュール）の活動を推定できる



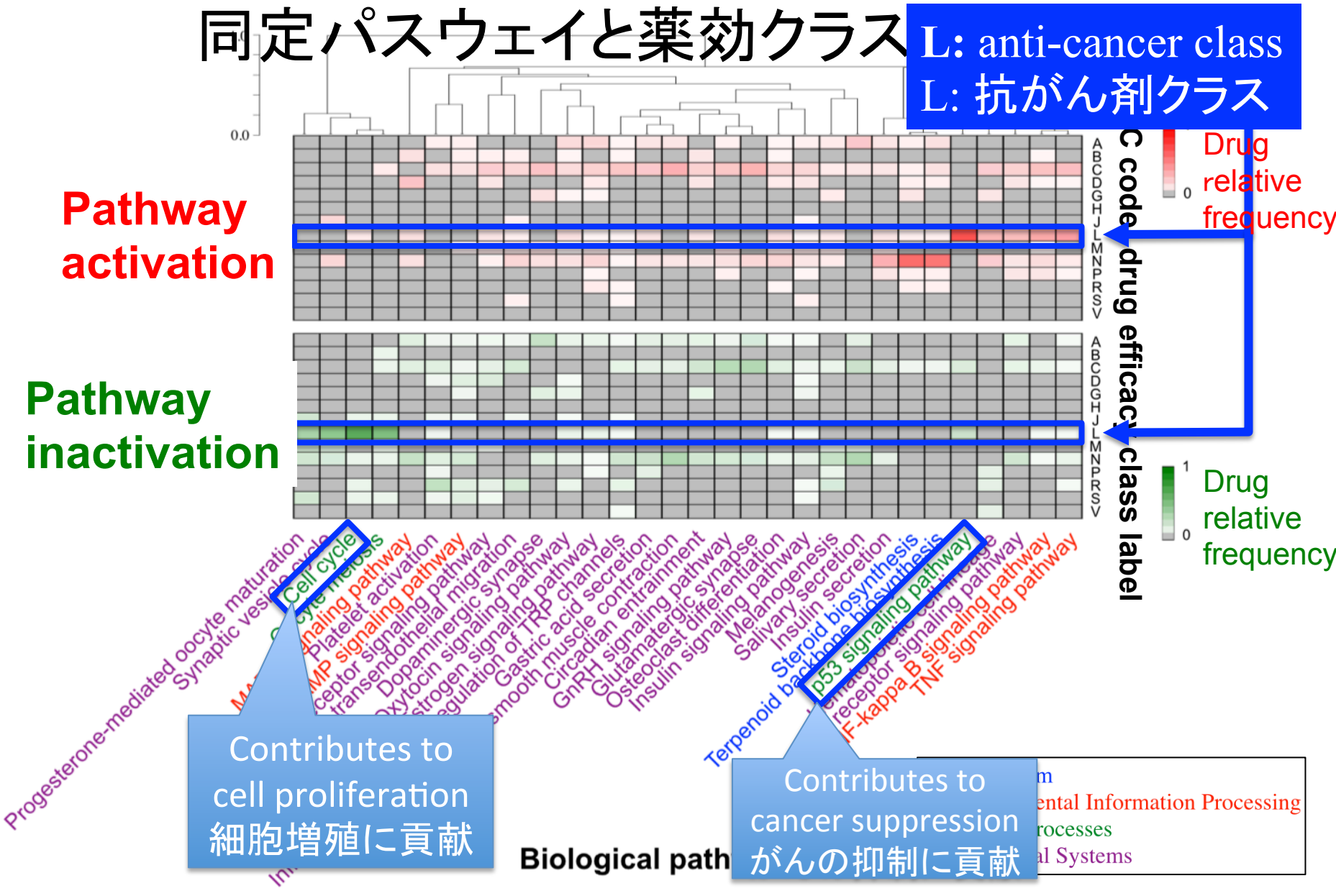
■ hypergeometric test

$$P\text{-value} = \sum_{i=z}^{\min(k,\, r)} \frac{\binom{k}{i}\binom{l-k}{r-i}}{\binom{l}{r}}$$

	Regulated genes	Genes
In a pathway	*i*	*k*
Not in a pathway	*r - i*	*l - k*
Total	*r*	*l*

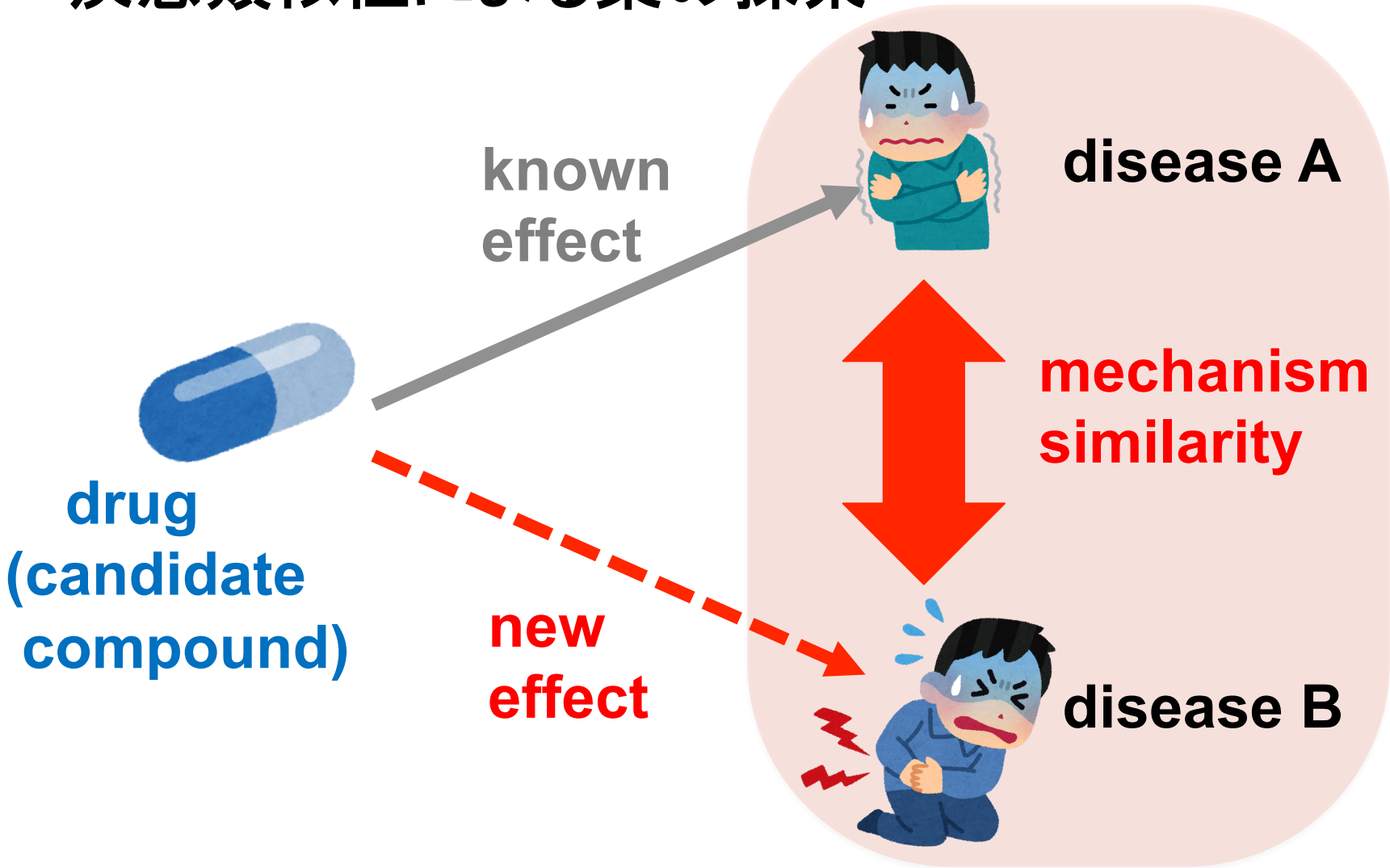
Relationship between identified pathways and drug efficacy classes

同定パスウェイと薬効クラス



Drug discovery based on disease similarity

疾患類似性による薬の探索



Summary まとめ

- The proposed methods can predict potential drug target proteins and new drug effects.
- From organ-based disease classification to mechanism-based disease classification.
- It is possible to deliver necessary drugs to patients quickly.

- 薬物の潜在的な標的タンパク質や新しい効能をデータ駆動で予測。
- 臓器別ではなく分子機序で疾患を分類。
- 様々な疾患の患者に、早く、安く、必要な薬を届けることができる。

Paradigm shift パラダイムシフト

- Traditional drug discovery with mass consumption
  従来の大量消費型の創薬

- Data-driven drug discovery with energy saving-mode

エコノミカル（安く効率的）かつエコロジカル（省エネで環境に優しい）なデータ駆動型の創薬

Thank you for your attention.

ご清聴ありがとうございました。