

トーゴーの日シンポジウム2025

# データサイエンティストフレンドリーな 毒性病理画像データベースCuratoxiiの開発

Tadahaya Mizuno<sup>1,2</sup>

<sup>1</sup>Graduate School of Pharmaceutical Sciences, The University of Tokyo

<sup>2</sup>School of Statistical Thinking, The Institute of Statistical Mathematics

tadahaya@gmail.com

2025/10/20



# COI Disclosure Information

Tadahaya Mizuno

The author has the following conflicts of interest related to this presentation:

Cosomil, Inc.: Technical advisor

Japan Tobacco Inc.: Technical advisor

Mitsubishi Tanabe Pharma Corporation: Research collaboration

Shiseido Company, Limited: Research collaboration

Shionogi & Co., Ltd.: Research collaboration

Takeda Pharmaceutical Company Limited: Technical advisor



# 自己紹介



水野 忠快 (みずの ただはや)  
博士(薬学)・生物情報学  
「生命科学データのパターン認識」

- ✓ 2005.3 神奈川県立柏陽高校 卒業
- ✓ 2005.4 東京大学教養学部理科一類 入学
- ✓ 2009.3 東京大学薬学部薬学科 修了
- ✓ 2011.3 東京大学大学院薬学系研究科・修士課程 修了
- ✓ 2014.3 同・博士課程 修了
- ✓ 2014.4 同・特任助教 着任
- ✓ 2015.10 同・助教 着任 (現職)
- ✓ 2025.7 統計数理研究所・特任助教 (兼任)

水野  
忠邦



Was a well-known Japanese Samurai in the 19<sup>th</sup> century.

忠輝

忠典

美智子

忠快

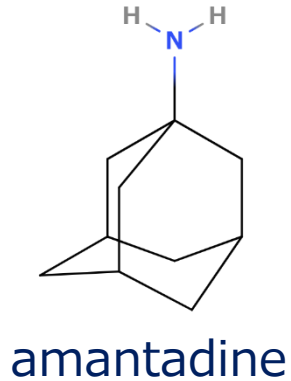
英美

忠玖

美玲

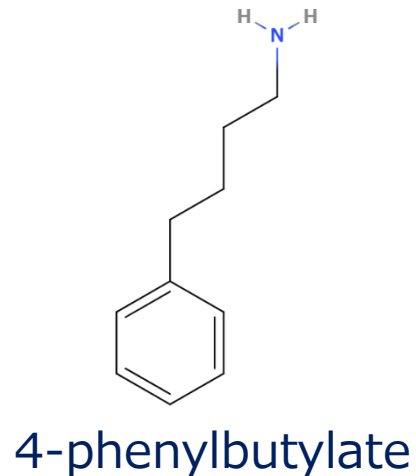


# 医薬品の作用は複合的で**認識されていない作用**も存在する



|                 | Treatment           | Effect                        |
|-----------------|---------------------|-------------------------------|
| Original        | Influenza           | M2 ion channel inhibition     |
| Uncovered later | Parkinson's disease | NMDAR inhibition <sup>1</sup> |

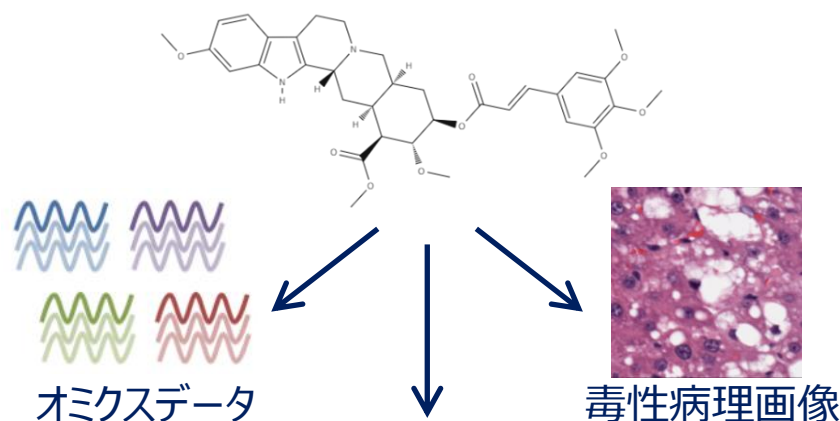
<sup>1</sup> Lipton SA., 2004, *NeuroRX*



|                 | Treatment           | Effect                                      |
|-----------------|---------------------|---------------------------------------------|
| Original        | Urea cycle disorder | Gln scavenger                               |
| Uncovered later | PFIC2               | Inhibition of BSEP degradation <sup>2</sup> |

<sup>2</sup> Hayashi H, et al., 2016, *Hepatol Res*

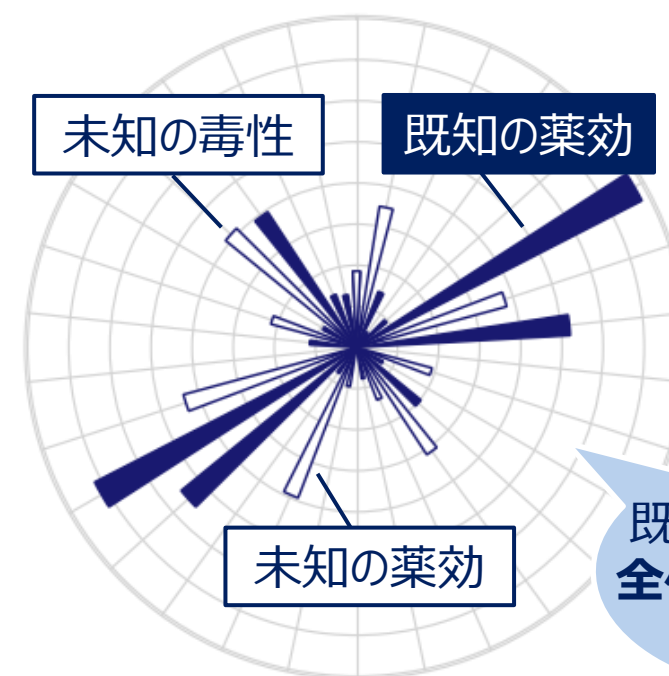
# 生命科学データのパターン認識：生命科学データに潜むパターンを探究し、医薬品の特性や作用を俯瞰的に捉える



COC(=O)C1C2CC3c4[nH]c5cc(OC)ccc5c4CCN3CC2CC(OC(=O)C=Cc2cc(OC)c(OC)c(OC)c2)C1OC

構造文字列表現

パターン認識



多様な形式のデータが存在

## 生命科学データ

- ✓ 一般に**多次元**で冗長
- ✓ ヒトにとって特徴の**認識困難**

## 潜在的なパターン

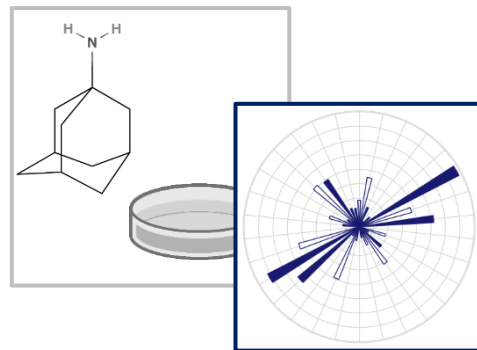
- ✓ **低次元**で本質的な情報
- ✓ ヒトでも**認識が容易**

## オミクスデータのパターン①

**入力：**培養細胞のマイクロアレイ

**手法：**因子分析

**出力：**化合物の複合的作用



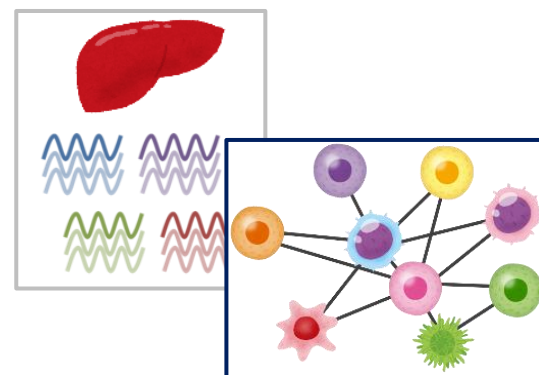
Mizuno T., *Sci Rep.*, 2019  
Nemoto S., *J Nat Prod.*, 2021

## オミクスデータのパターン②

**入力：**組織RNAseq

**手法：**潜在的ディリクレ配分

**出力：**組織中の細胞比率



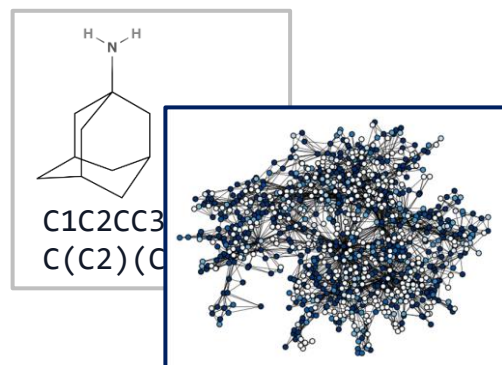
Morita K., *Toxicol Sci.*, 2023  
Azuma I., *Brief Bioinform.*, 2024

## 化合物構造のパターン

**入力：**化合物構造

**手法：**ニューラル機械翻訳

**出力：**化合物構造の数値表現



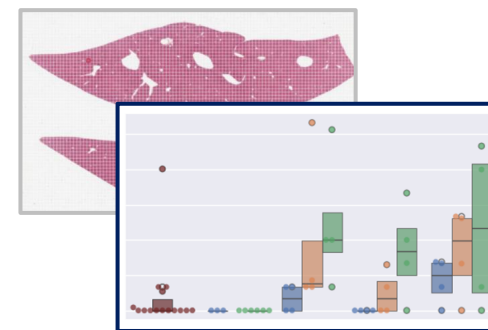
Nemoto S., *J Cheminform.*, 2022  
Yoshikai Y., *Nat Commun.*, 2024

## 毒性病理画像のパターン

**入力：**毒性病理画像

**手法：**画像自己教師あり学習

**出力：**個体応答の潜在表現



Maedera S., *Comput Biol Med.*, 2024  
Iwasaka T., *bioRxiv*, 2024



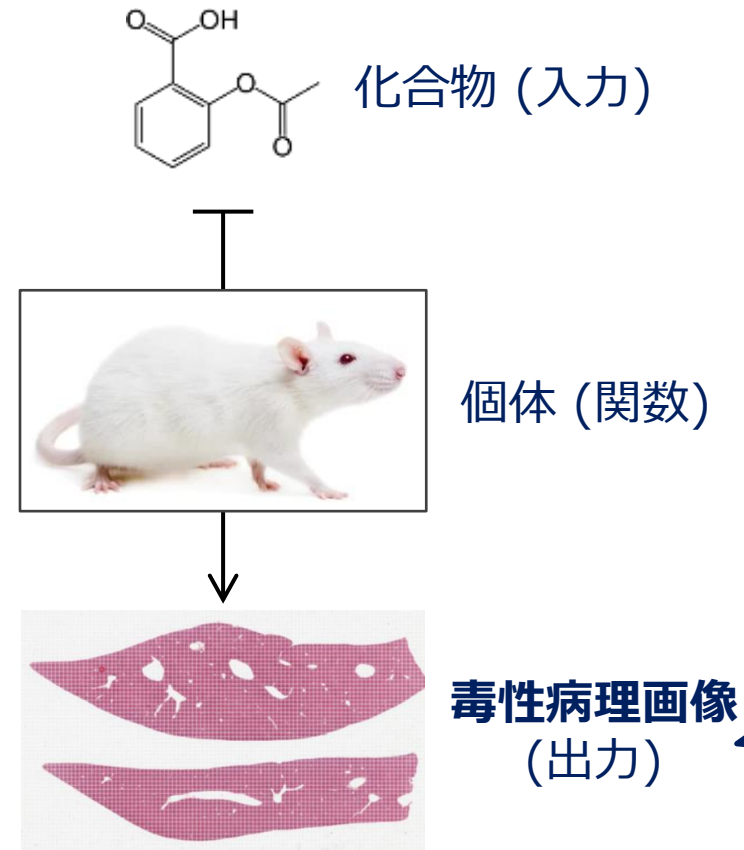
# in vivo応答のソース：毒性病理画像

## 応用

- ✓ 医薬品開発で**必須** Cf. ICH-S4
- ✓ 解析技術の発展が望まれている

## 基礎

- ✓ 化合物の個体での作用を表現
  - ✓ 時系列データ
  - ✓ マルチモーダルデータ
- 28日間反復投与試験
  - 生化学検査値等との同時取得が標準プロトコル



知覚に近い  
恣意性の  
ない数値化

# Latent Representation Predicts MoA Better Than Pathological Findings

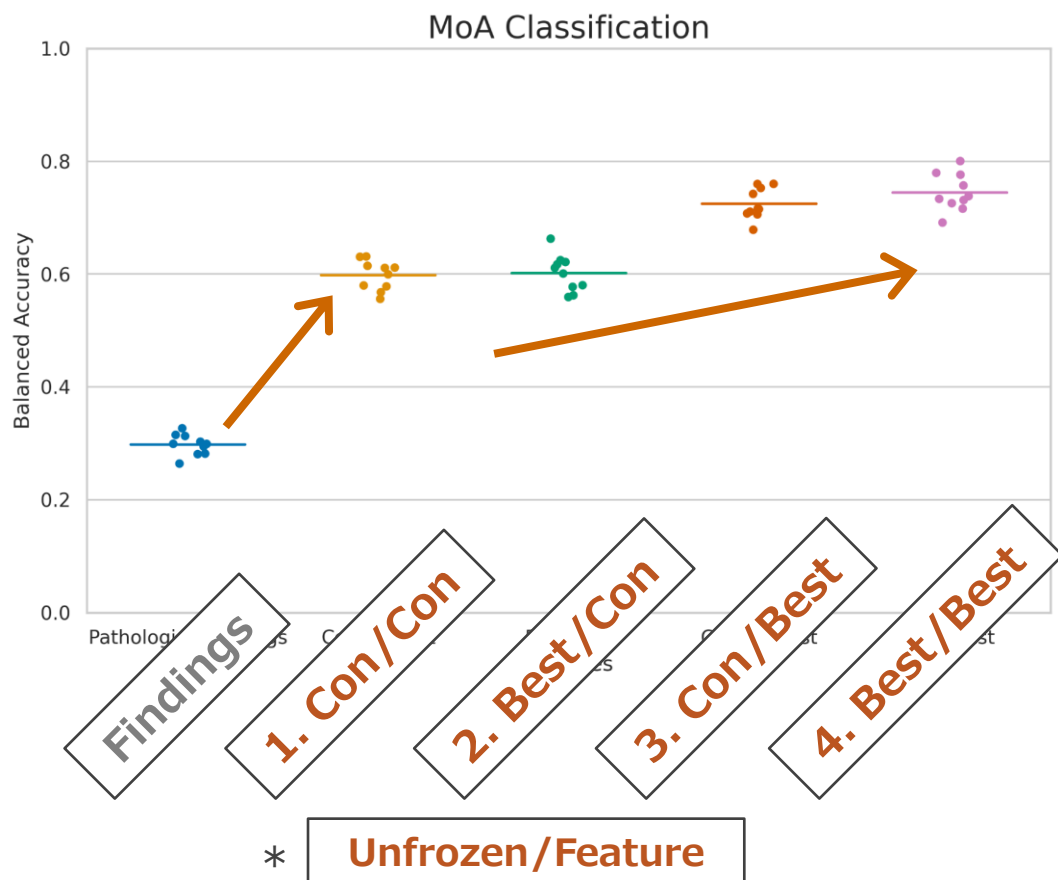
**Input**

Liver pathological image a compound



**Output**

MoA of the treated compound



5-HT<sub>2A</sub>R antagonist

DNA inhibitor

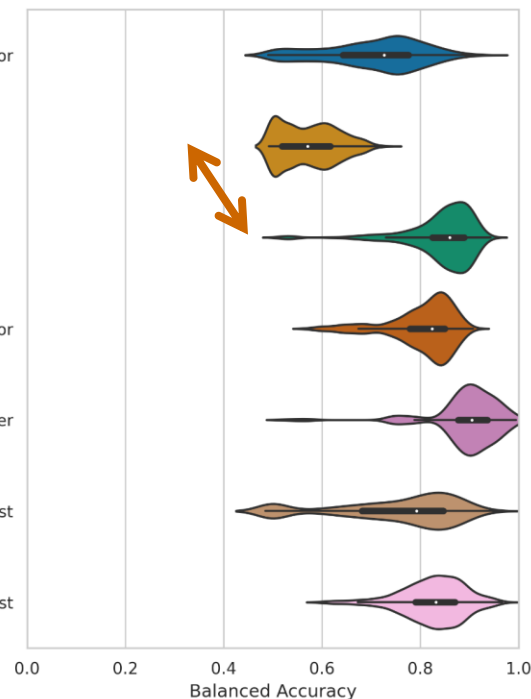
Bacterial 70S ribosome inhibitor

Cyclooxygenase inhibitor

Sulfonylurea receptor 1, Kir6.2 blocker

Histamine H<sub>2</sub> receptor antagonist

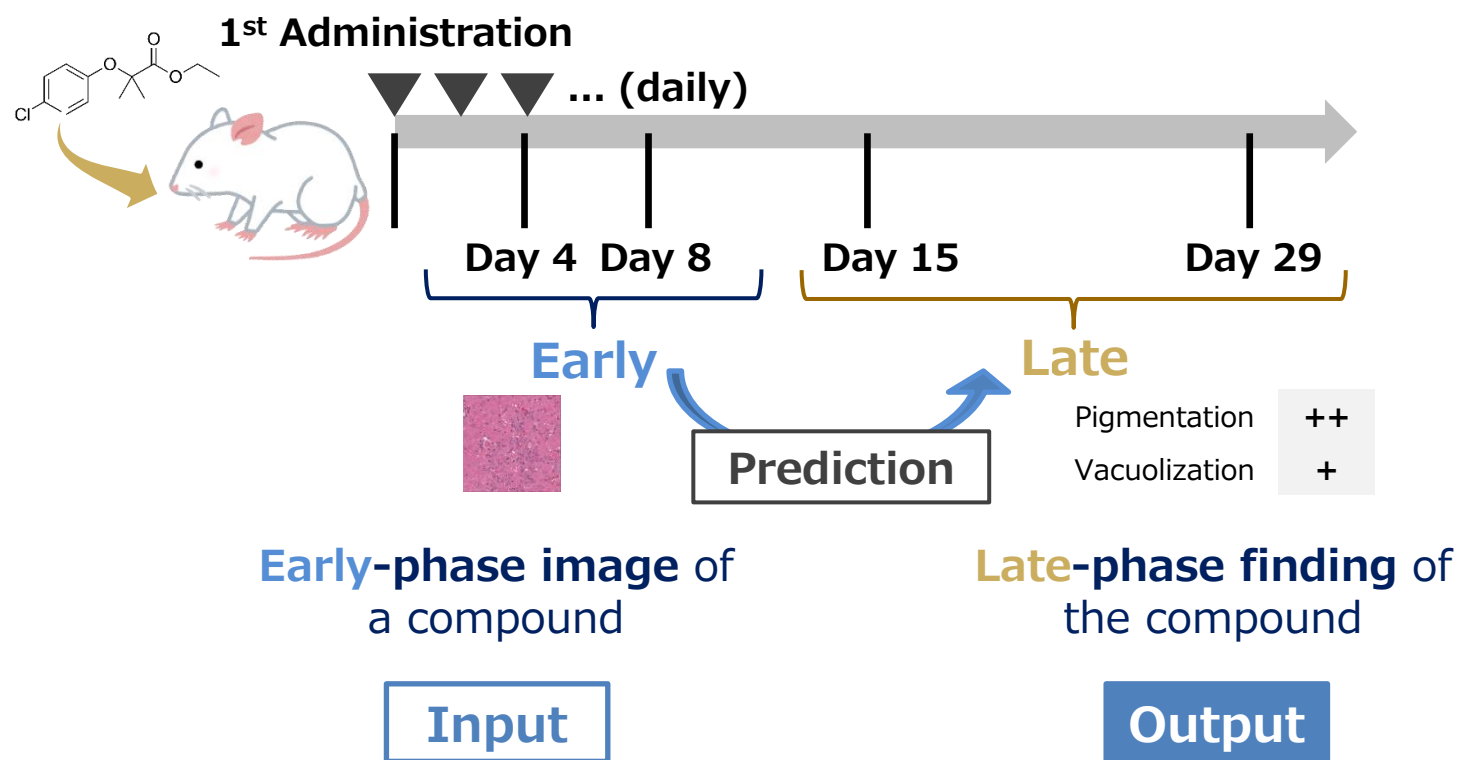
Peroxisome proliferator-activated receptor alpha agonist



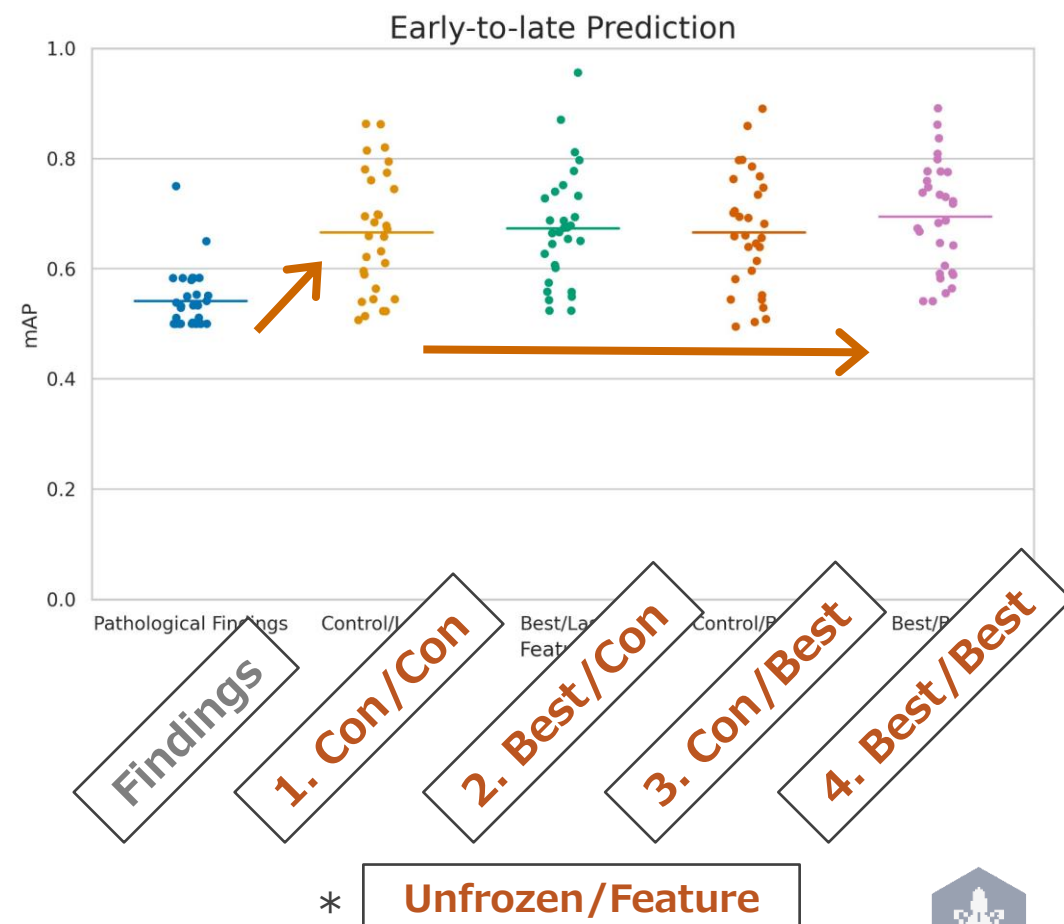


# Latent Representation Predicts Late-phase Finding from Early-phase Images Better Than Finding Itself

## Design

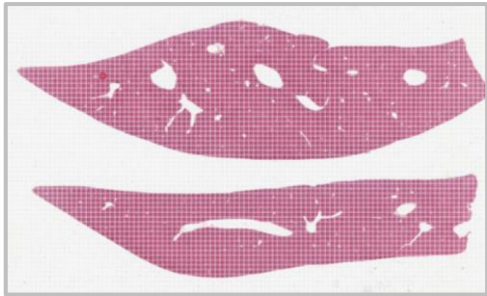


## Result



# 画像解析に向けた毒性病理画像界限の課題：少ない研究者人口

✓ **大容量**のためハンドリングしづらい



**WSI** (whole slide image)

✓ **~40K × 30K** pixels

✓ **~1 GB/image**

Cf. 一般の深層学習が扱う画像: 256 × 256 pixels

✓ 公開データが**整備**されていない

| Find                 | <input type="text"/> |          |               |                                          |       |                                                                                                                               | All | ▼ | Search |
|----------------------|----------------------|----------|---------------|------------------------------------------|-------|-------------------------------------------------------------------------------------------------------------------------------|-----|---|--------|
| Entry                | EXP_ID               | GROUP_ID | INDIVIDUAL_ID | COMPOUND_NAME                            | ORGAN | FILE_LOCATION                                                                                                                 |     |   |        |
| <a href="#">Show</a> | 0698                 | 05       | 2             | 1% cholesterol +<br>0.25% sodium cholate | Liver | <br>[Full-size view /<br>Download (903MB)] |     |   |        |

Open TG-GATEsでは個別DLかバルクDL (**25 TB!**)の二択



製薬企業等とのつながりがないと参入自体が困難

# 課題解決に向けた提案

整備されたDBは分野  
発展に不可欠



扱いやすい公開データを  
提供する**毒性病理画像**  
データベースの構築



ニーズに応じて検索→カスタマイズ  
したデータセットの構築

| COHORT BUILDER                            |                                     |             |
|-------------------------------------------|-------------------------------------|-------------|
| General                                   | Program                             |             |
|                                           | Name                                | Cases       |
|                                           | <input type="checkbox"/> APOLLO     | 87 (0.19%)  |
|                                           | <input type="checkbox"/> BEATAML1.0 | 882 (1.97%) |
| Demographic                               | <input type="checkbox"/> CDDP_EAGLE | 50 (0.11%)  |
| General Diagnosis                         | <input type="checkbox"/> CGCI       | 645 (1.44%) |
| Project                                   |                                     |             |
| Name                                      |                                     |             |
| <input type="checkbox"/> APOLLO-LUAD      |                                     |             |
| <input type="checkbox"/> BEATAML1.0-COHOF |                                     |             |
| <input type="checkbox"/> BEATAML1.0-CRENC |                                     |             |
| <input type="checkbox"/> CDDP_EAGLE-1     |                                     |             |

Cf. TCGAコホートビルダー

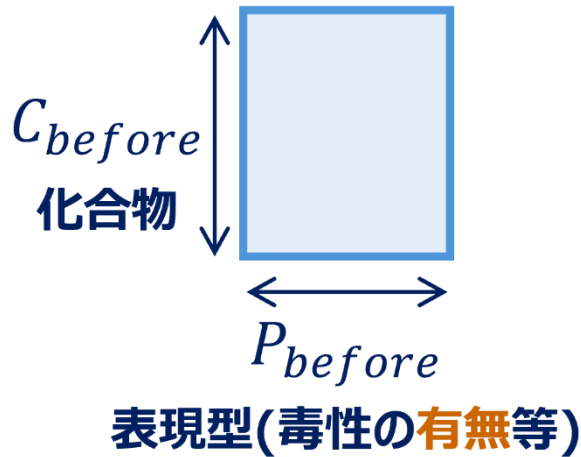
事前定義された  
データセット × リーダーボード

| Open LLM Leaderboard                                                                                                   |      |                             |         |       |       |       |       |       |          |        |
|------------------------------------------------------------------------------------------------------------------------|------|-----------------------------|---------|-------|-------|-------|-------|-------|----------|--------|
| Comparing Large Language Models in an open and reproducible way                                                        |      |                             |         |       |       |       |       |       |          |        |
| Search by model name: Try "meta-llama-3-70b-instruct" (438 / 438) Advanced filters                                     |      |                             |         |       |       |       |       |       |          |        |
| Quick filters: Per Edge Rank: 100 Per Dataset: 100 Max range: 10000 Per Top GPU rank: 100 Only Official Providers: 100 |      |                             |         |       |       |       |       |       |          |        |
| Rank                                                                                                                   | Type | Model                       | Average | 95th  | 99th  | MATH  | GPQA  | MQAR  | MMLU-Pro | CO-Cat |
| 1                                                                                                                      | Open | Meta-Llama-3.1-70B-Instruct | 85.8%   | 86.8% | 86.8% | 86.1% | 85.8% | 85.8% | 85.8%    | 85.8%  |
| 2                                                                                                                      | Open | Meta-Llama-3.1-70B-Instruct | 85.8%   | 86.8% | 86.8% | 86.1% | 85.8% | 85.8% | 85.8%    | 85.8%  |
| 3                                                                                                                      | Open | Meta-Llama-3.1-70B-Instruct | 85.8%   | 86.8% | 86.8% | 86.1% | 85.8% | 85.8% | 85.8%    | 85.8%  |

Cf. Open LLM Leaderboard

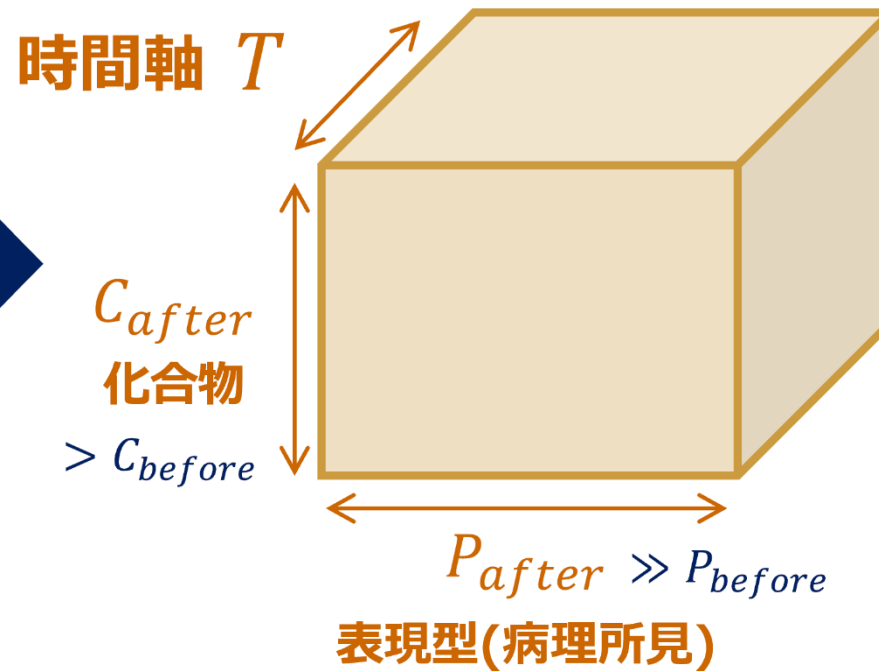
# 化合物－個体の関係性

現在公共より入手可能かつ  
扱いやすい化合物と個体を  
結びつけるデータ



Ex. DrugBank

本研究により公開・共有されるデータ



- ✓ 化合物の増加
  - ✓ 表現型の増加
  - ✓ 時間軸の追加
- 化合物－個体  
を結びつける  
**研究リソースの  
増加**

化合物構造とin vivo  
生体応答の理解が深化

どんな構造がどんな作用を惹起する？

これまでとは異なるより粒度の高いレイヤーを提供

# 利用者像：3タイプの利用者を想定

| 利用者のタイプ                                                         | どのようなクエリか？                                             | どのような結果セットか？                       | 想定される用途は？                                                              |
|-----------------------------------------------------------------|--------------------------------------------------------|------------------------------------|------------------------------------------------------------------------|
| <b>情報系データサイエンティスト</b><br>Ex. 情報学・工学研究者,<br>IT企業等の研究者            | <b>事前定義された<br/>データセットのタイプ</b><br>Ex. 時系列データセット・medium  | <b>比較可能な深層学習用<br/>ベンチマークデータセット</b> | <b>毒性病理画像解析モデルの<br/>開発・性能評価</b>                                        |
| <b>ライフ系データサイエンティスト</b><br>Ex. バイオインフォマティシャン,<br>製薬企業データサイエンティスト | <b>各々が求める実験条件</b><br>Ex. ALT 100 U/L<br>かつTBIL 3 mg/dL | <b>各々の目的に応じた<br/>カスタマイズドデータセット</b> | <b>カスタマイズしたデータセット<br/>での所望の解析</b><br>Ex. 胆汁うっ滞性肝障害を誘導する<br>化合物構造の特徴量抽出 |
| <b>ライフ系研究者</b><br>Ex. 薬学系研究者,<br>企業病理専門家の卵                      | <b>病理所見</b>                                            | <b>病理所見を示す画像</b>                   | <b>病理組織学の<br/>理解・教育</b>                                                |

# 学術的・社会的に期待されるインパクト



参入障壁低減による  
研究者層の拡充

現在限られ  
ている研究  
人口増加

社会的

医薬品安全性評価  
の効率化・高度化

- ✓ 毒性病理画像解析モデルの発達
- ✓ 病理所見判定の自動化
- ✓ 初期画像から後期所見の予測

Mizuno T\*. *Comput Biol Med.* 2024

学術的

in vivoでの化合物の  
作用理解を深化

- ✓ 絶対数が少ないin vivoでの化合物の作用データの蓄積・公開
- ✓ 時系列データからスナップショットでは見られない作用の理解

# Acknowledgement



## Mizuno Team (\*, alumni; \*\*, technical assistant)

|            |                |               |
|------------|----------------|---------------|
| Azuma I    | Kikuchi Y      | Dr. Morita K* |
| Iwasaka T  | Ri T           | Dr. Nemoto S* |
| Nakagawa T | Yoshikawa M    | Fujimoto H**  |
| Yoshikai Y | Honzawa R      | Kuriyama W**  |
| Ohto Y     | Dr. Maedera S* | Nakashima E** |

## Shionogi Team

Dr. Kato Y

Dr. Ishii H

## MST Team

Dr. Takano N

Dr. Miyamoto Y

Dr. Fukumoto Y

Dr. Teshima H

## The Institute of Statistical Mathematics

Prof. Ikeda S

## The University of Tokyo

Dr. Komatsu T

大学院生・ポスドク  
募集中！！



科研費  
KAKENHI

公益財団法人  
武田科学振興財団  
Takeda Science Foundation

公益財団法人 持田記念医学薬学振興財団



Contact: [tadahaya@gmail.com](mailto:tadahaya@gmail.com)