



Development of multimodal AI frameworks and foundation models in life sciences

Ryosuke Kojima

Laboratory for Multimodal AI Framework
RIKEN BDR, Japan



Self Introduction: Ryosuke KOJIMA

- 2012: received a B.E. in computer science in 2012 from the Tokyo Institute of technology
- 2014: received an M.E. in information science and engineering in 2014 from the Tokyo Institute of technology
- 2017: received a Ph.D.(engineering) in information science and engineering in 2017 from the Tokyo Institute of technology
- 2017: a program-specific associate professor (AMED) at Kyoto University
- 2021: a lecturer at Kyoto University
- 2024: currently an associate professor at Kyoto University
- 2024: a team director at Laboratory for Multimodal AI Framework, RIKEN BDR, Japan

AI/Machine learning

Machine learning
Biomedical AI
Drug discovery AI

**AI/Machine learning
+ Robotics/time-series analysis**

**AI/Machine learning
+ Biomedical/Drug discovery**

My research question: What is missing for AI to be “truly” useful?

Agenda

- Background
- AI tools: A GNN tool for life science
- Foundation model: Reaction Foundation model
- Neural scaling law in molecules

AI is Becoming Essential in Various Fields of Life Sciences

FY2024, FY2025

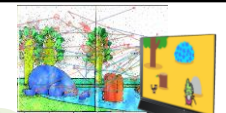
AMED: Elucidation of pathological mechanisms of neurological and psychiatric disorders

Understanding and overcoming psychiatric and neurological disorders through REM sleep approaches (Prof. Hayashi)



JST Moonshot (Goal 2)

「Towards Overcoming Disorders Linked to Dementia based on a Comprehensive Understanding of Multiorgan Network」 (Prof. Takahashi)



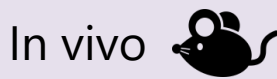
Health care / Preventive Medicine

Sensing data

Nursing, Intervention & Health Guidance

Diagnosis & Treatment

Basic research of life science



Biology/ecology

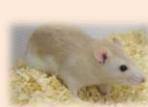
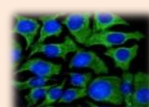
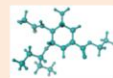
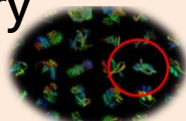


Medicine

Clinical data



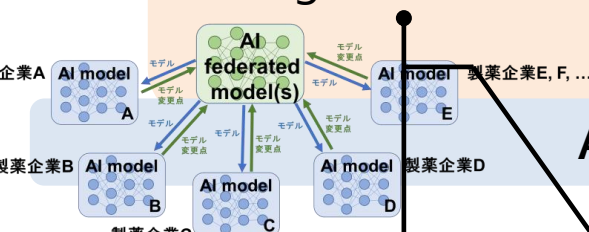
Drug discovery



Target → Lead → Optimization → Assay → Clinical Trial → Approval → Drug Therapy

Medicinal Chemistry

AI platform and databases for life science



AMED DIIA

AI Development for Confidential Data with Federated Learning (Prof. Honma)

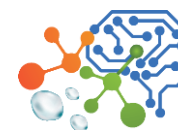
AMED DXPF

AI Development for DX platform based on generative AI and simulation (Prof. Okuno)

Grant-in-Aid for Transformative Research

Area A: Digi TOS

Digitalization-driven transformative organic synthesis (Prof. Ohshima)



JST CREST: Trusted Quality AI Systems

Machine Learning That Connects to Symbolic Reasoning (Prof. Sugiyama)



Background: issues of data science in life science

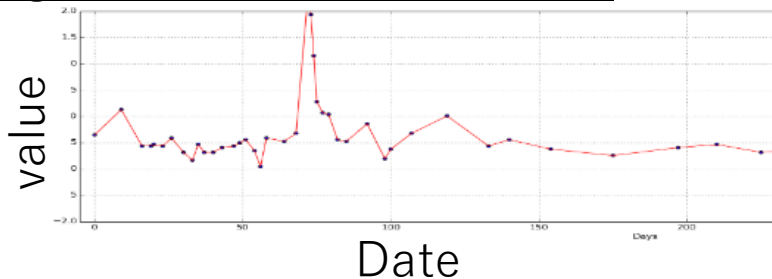
Issue1. Multimodality

Medical data includes multimodal data

Issue2. Real-world data

Missing/biased data due to measurement costs, etc.

e.g. medical health record

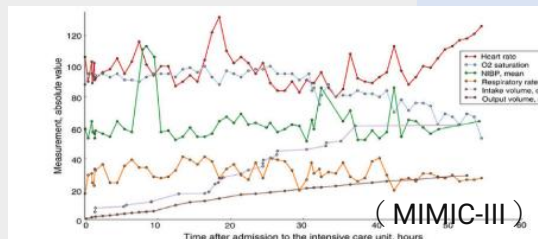


Issue3. Trustworthy AI

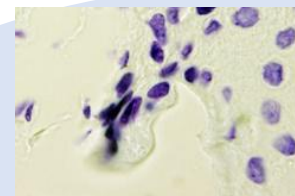
- Utilizing clinical knowledge
- Aligning AI predictions with empirical knowledge



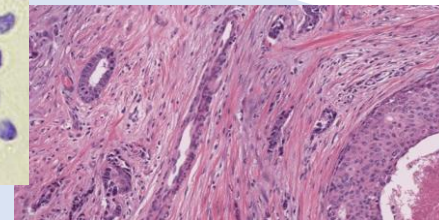
Time series



Vision



Cytology images



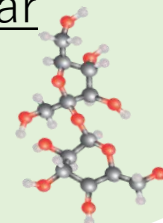
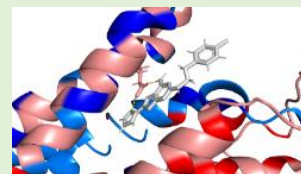
Histopathological images

Language

- Drug package insert
- Medical records

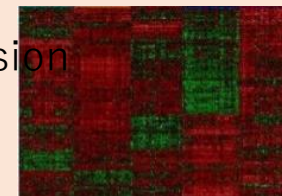


Molecular



Genome

- Gene expression
- Multi-omics
- Variant



Approaches in our researches

Issue1. Multimodality



Deep learning and machine learning for multimodality

Improved performance and discovery of relationships between multiple data

Issue2. Real-world data (missing/biased/multi-time-scale data)



Statistics and probability

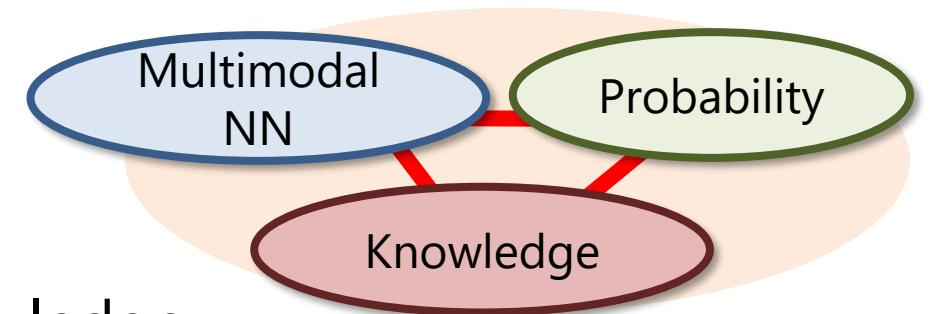
Statistical evaluation for dealing with biased and missing data

Issue3. Trustworthy AI



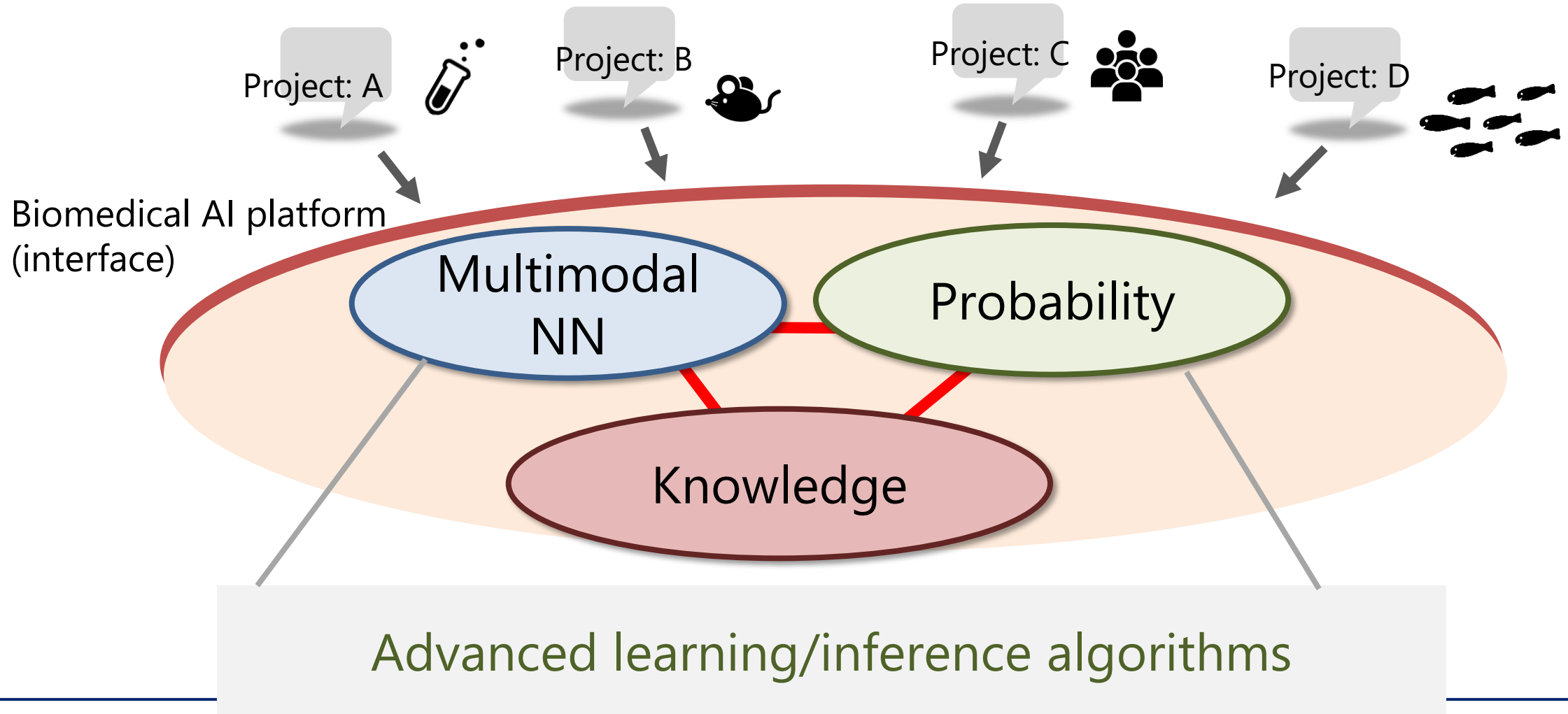
Explainable/interpretable AI

Evaluating consistency/inconsistency with existing knowledge & utilizing empirical knowledge



Our biomedical AI platform approach

The requirements of each biomedical application or project can be used without being aware of behind advanced technology (interface)



Related work

The lack of and need for “interfaces” in AI has also been pointed out in the field of informatics in general by Domingos et al.

	Hardware architecture	OS	Program	Database	Network	AI
High-level	Compilers, OS	Software	Programming	Enterprise applications	Web, email	Vision, NLP, Planning, robotics
Interface	Micro-processors	VM	Language	Relational model	Internet	???
Low-level	ALU, buses (VLSI)	Hardware	Compiler	Query, transaction mgmt.	Protocols, routers	Inference, learning

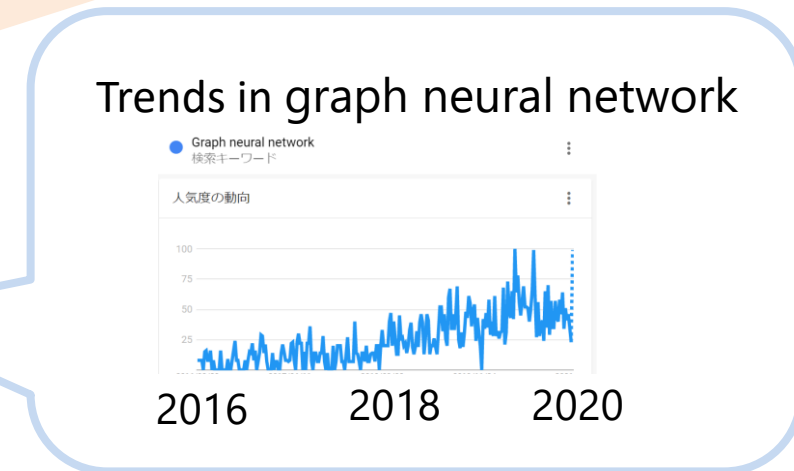
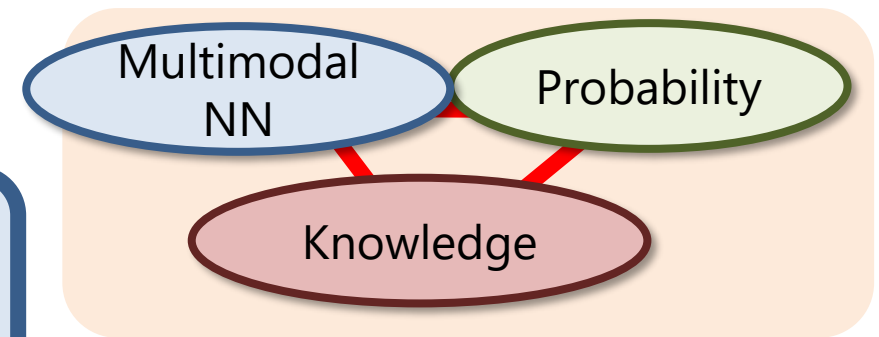
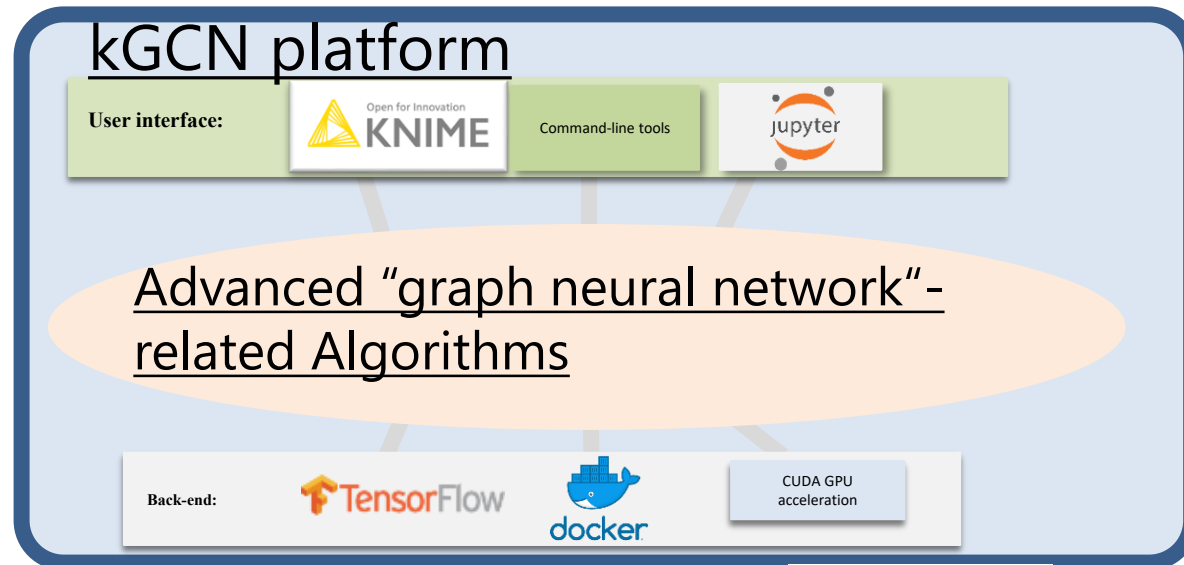
Pedro Domingos et al. 2009: Markov Logic: An Interface Layer for Artificial Intelligence, **What's Missing in AI: The Interface Layer**

Our objective: To develop an AI interface for biomedical practical applications

Graph-based AI platform (1/3)

kGCN: a graph neural network framework for life science

kGCN provides multiple-level interfaces for various users, including scientists specializing in fields other than AI



Prediction of compound-protein interactions: (Master course : M. Ikeguchi)

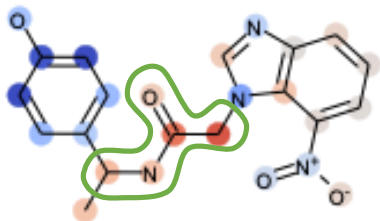
In-silico screening: (Master course : M. Hamatani)

Development of molecular generative models: (Bachelor course : T. Nakai)

Masters and bachelor's students specializing in health and medical sciences can also benefit from this tool.

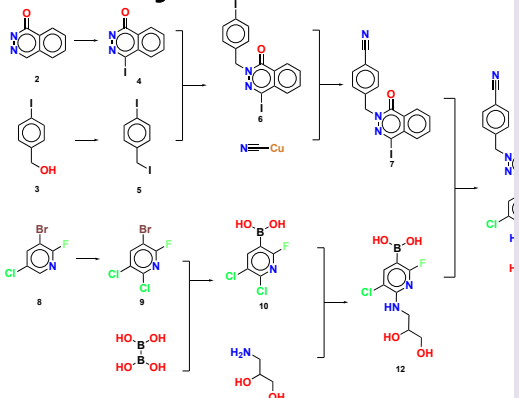
Graph-based AI platform (2/3)

Reaction prediction



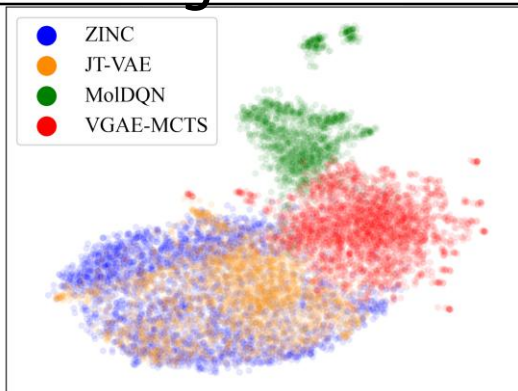
S.Ishida, K.Terayama, R.Kojima, K.Takasu, Y.Okuno:
Networks. In J. of Chem. Info. and Modeling, Vol. 19
No. 12 pp. 5026-5033, 2019.

Retrosynthesis



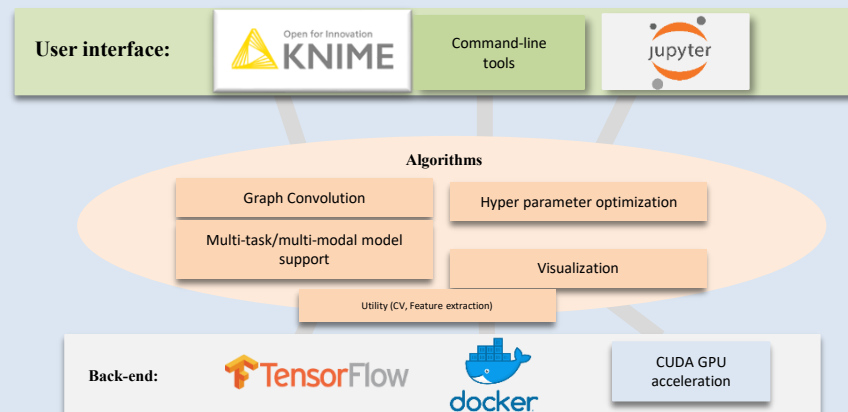
S.Ishida, K.Terayama, R.Kojima, K.Takasu, Y.Okuno:
ChemRxiv, 10.26434/chemrxiv.13386092.v1, 2020

Molecule generative model

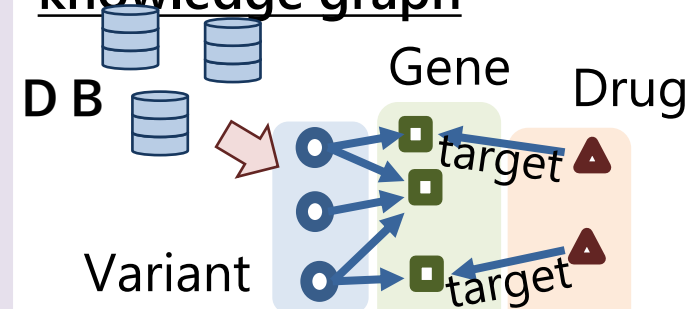


H.Iwata, T.Nakai, T.Koyama, S.Matsumoto, R.Kojima,
Y.Okuno, *ChemRxiv* 2023

kGCN platform

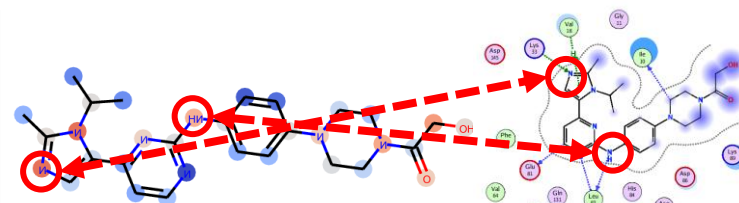


Predicting pathogenicity of mutations from genome-related knowledge graph



M. Kamada, T. Katayama, S. Kawashima, R. Kojima, M. Nakatsui, Y.
Okuno: Applications and Tools for the Life Sciences (SWAT4LS), 2017.

Predicting compound-protein interaction and visualization



R.Kojima, S.Ishida, M.Ohta, H.Iwata, T.Honma, Y.Okuno:
In Journal of Cheminformatics, Springer, Vol. 12 pp. 1-10, 2020.

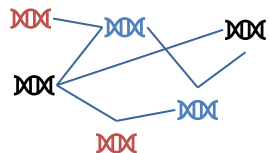
Easy to use, and also applicable to cutting-edge researches

Graph-based AI platform (3/3)

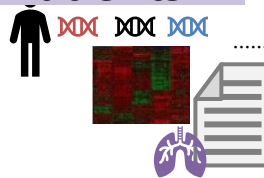
Developing platform software specialized for pathways and molecules based on kGCN

Predicting survival of cancer patients from gene expression

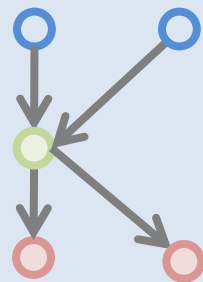
Pathway



Patients



PathwayGCN



K. Inoue, R. Kojima, M. Kamada, Y. Okuno
<https://arxiv.org/abs/2306.17202>

kMoL



<https://github.com/elix-tech/kmol>

R.Cozac+
J. of Chem 2025

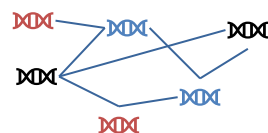
T.Koyama,
S.Matsumoto, H.Iwata,
R.Kojima, Y.Okuno,
JCIM 2023

Self-supervised learning for drug activity prediction

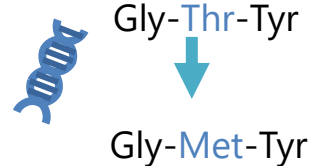


Prediction of cancer driver gene mutations

Pathway

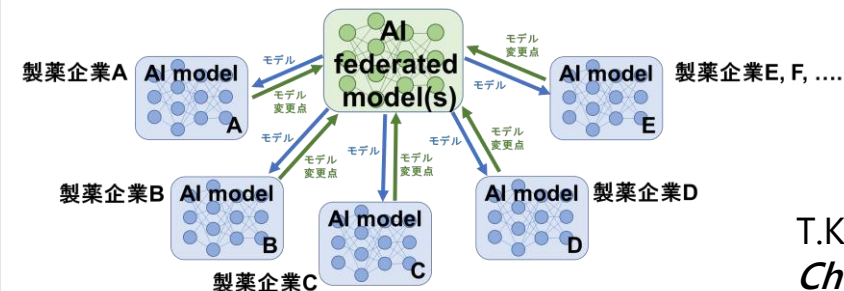


Variants



N.Hatano,
M.Kamada, R.Kojima,
Y.Okuno, *BMC Bioinfo.* 2023

Federated learning



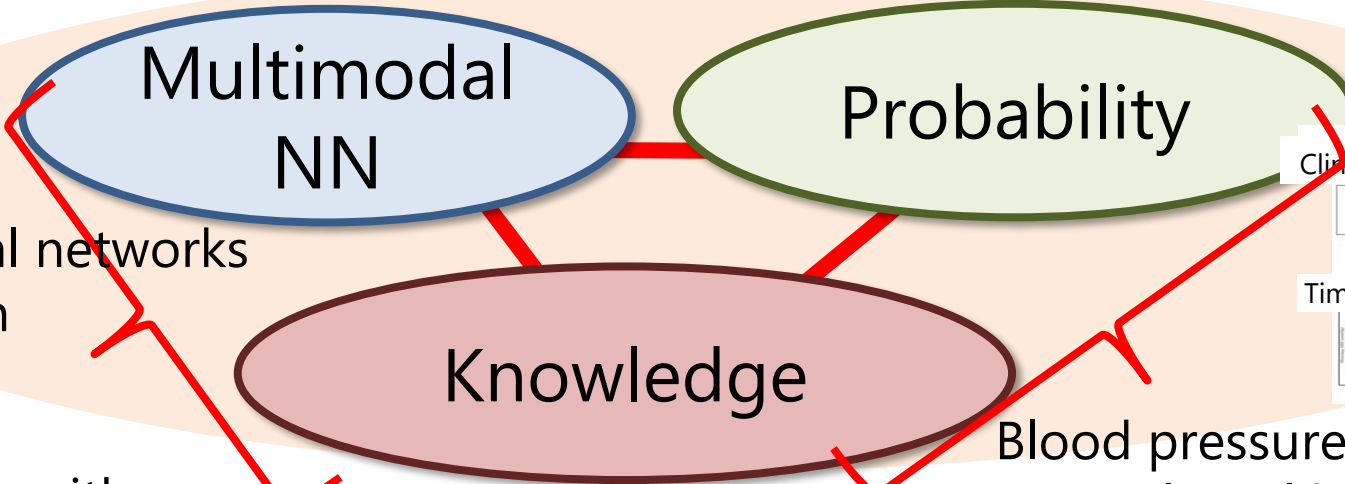
T.Koyama+
ChenRxiv 2025

Recent researches

Probabilistic feasibility assessment of health improvement proposal models

K.Nakamura+:
Nat. Com. 2021.

K.Nakamura+:
In *J. of Bio. Info.*, 2023.



Multimodal graph neural networks for CPI with visualization

R.Kojima+:
J. of Cheminformatics, 2020.

Variant label prediction with Background knowledge

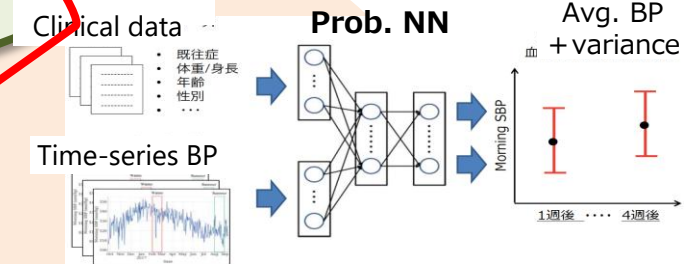
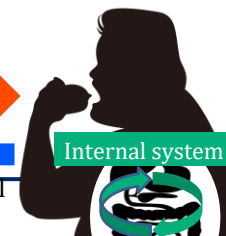
N.Hatano+ :
BMC bioinformatics, 2023.

Learning constraint-based dynamical system

R. Kojima+: *NeurIPS* 2022
Y. Okamoto+: *AAA/2025*

Input u :
Sugar

Output y :
Blood glucose level

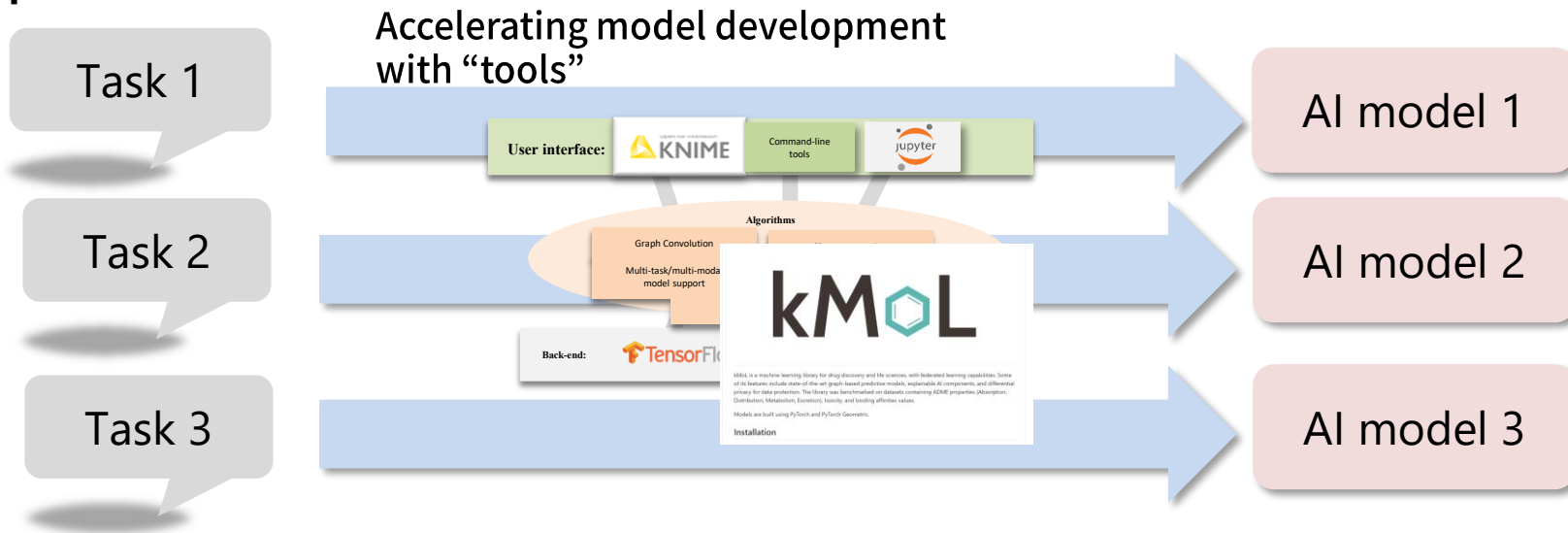


Blood pressure risk prediction using clinical knowledge + probability

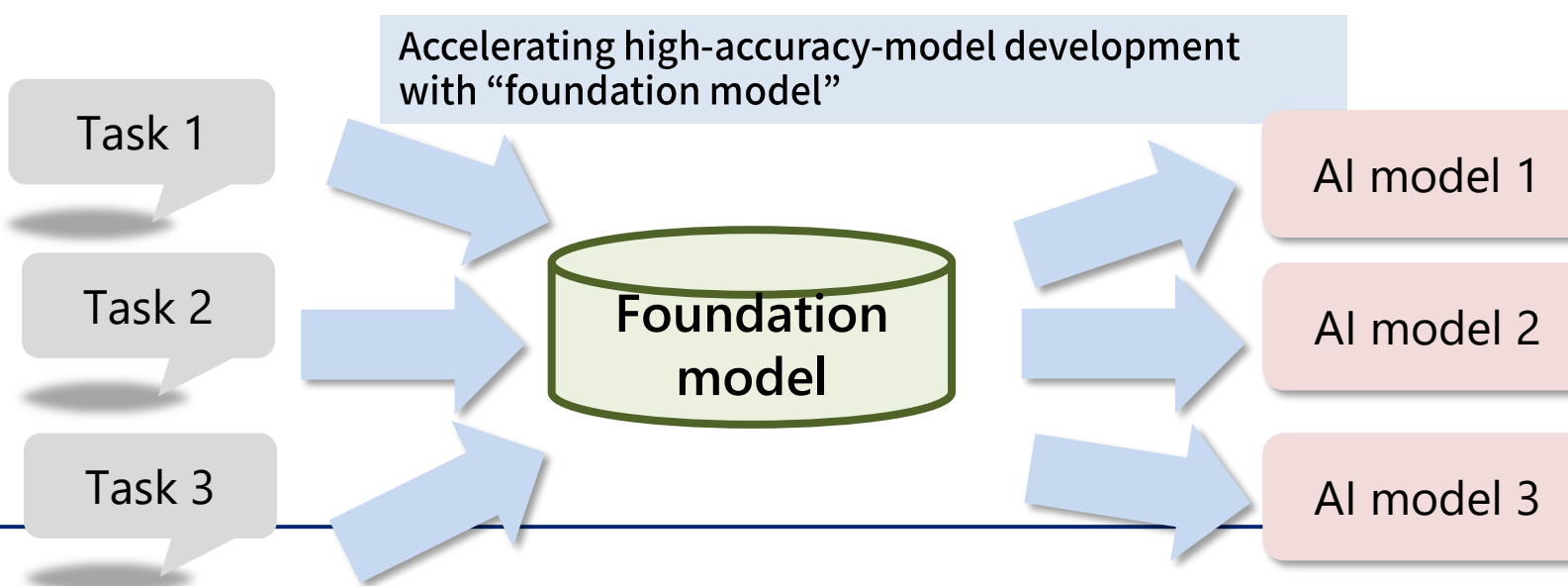
H. Koshimizu+:
Int. J. of Med. Info., 2020.

Ongoing work: foundation modals

Conventional approach

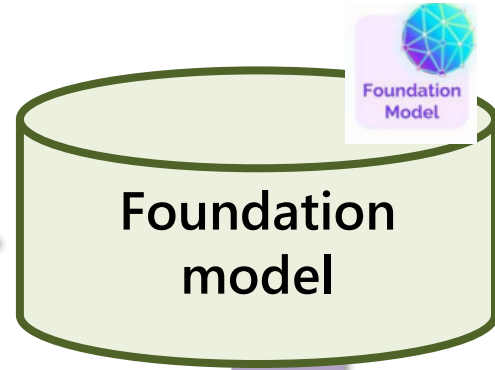
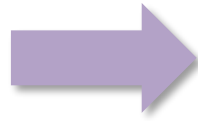
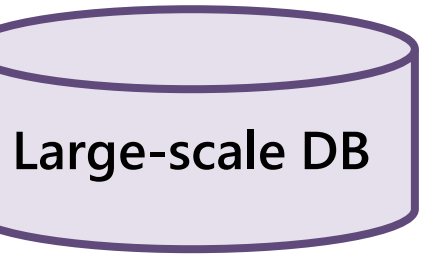


New approach

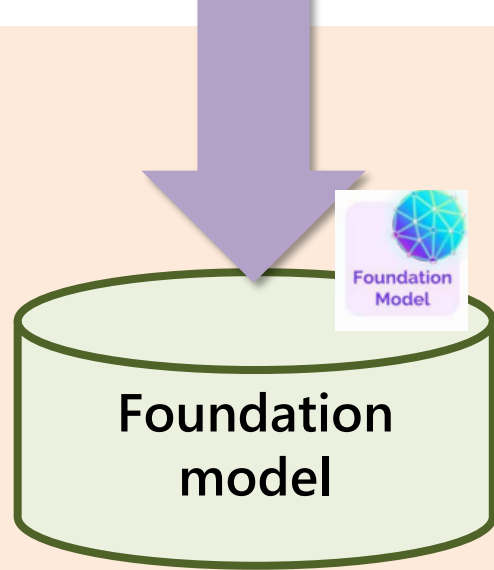
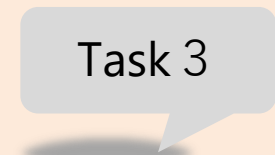
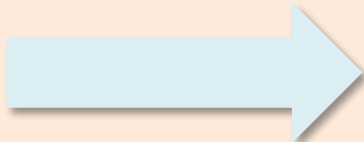
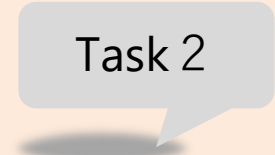
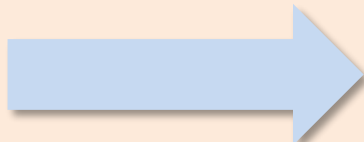
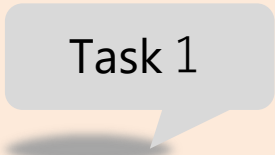


Foundation modal approach

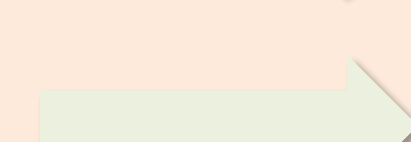
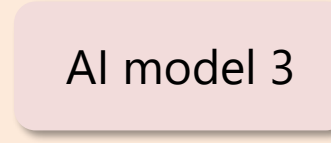
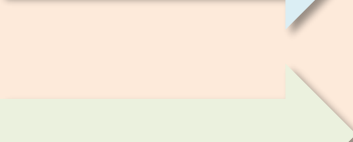
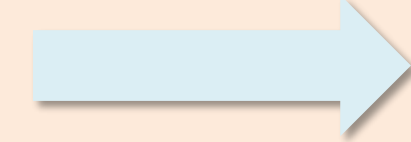
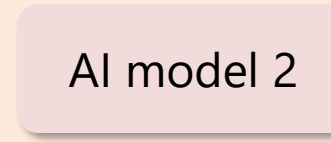
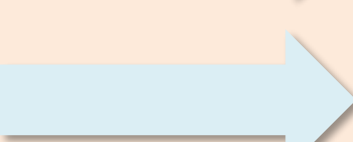
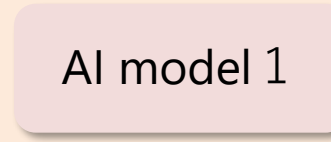
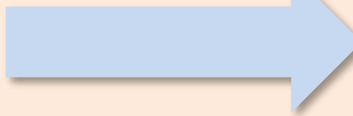
Pretraining



Down stream task



fine tuning



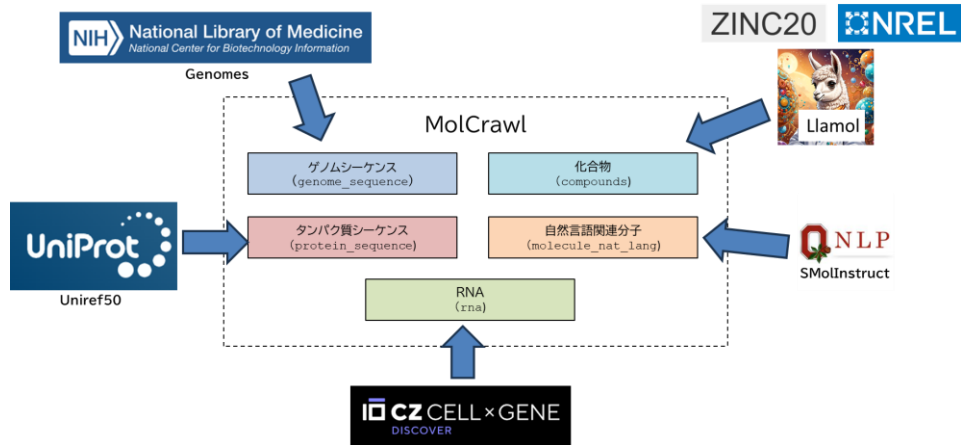
MolCrawl

Development Concept :

Unified pipeline: data download from diverse modality databases
 → preprocessing → tokenization → training trial models

Pretraining databases:

UniRef50、 ZINC 、 Refseq、 SMolInstruct、 Cellxgene



Alpha Models

GPT- and BERT-type models
 Parameter scale: 100M – 10B

Modalities: chemical compounds, genome sequences, protein sequences, RNA expression, molecular text

Outputs

Embedding vectors
 Generation and likelihood estimation

Target Downstream Tasks

Molecular property prediction
 Mutation impact prediction
 Molecular Q&A

Release Plan

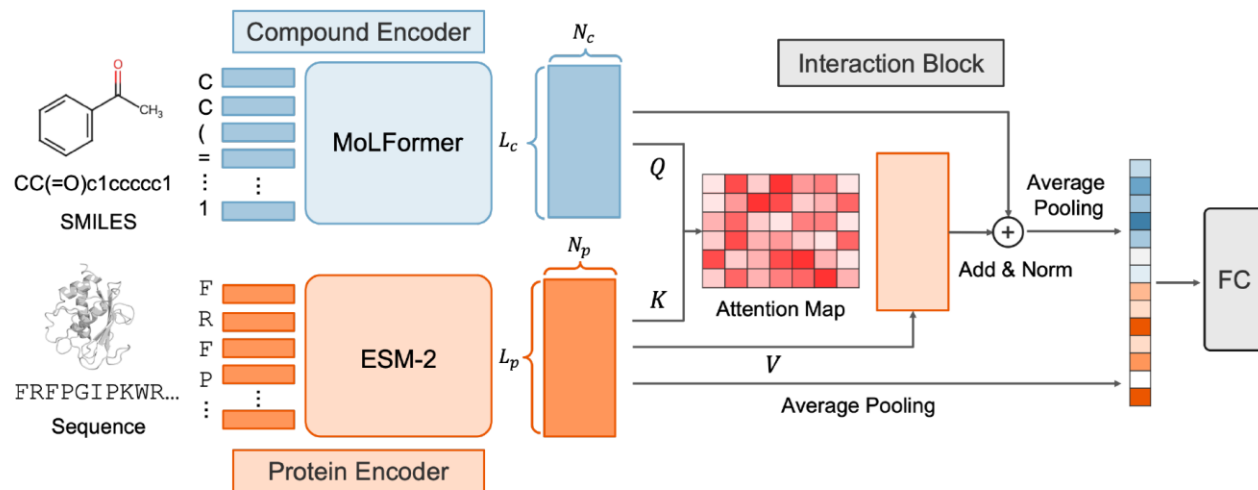
Code: to be open-sourced on GitHub
 Models: to be released on HuggingFace
 Release timing: within this fiscal year

Expected Impact

- Easy construction/reproduction of foundation models for each modality
- Faster development of diverse downstream applications
- Potential global adoption as a basis for molecular foundation models

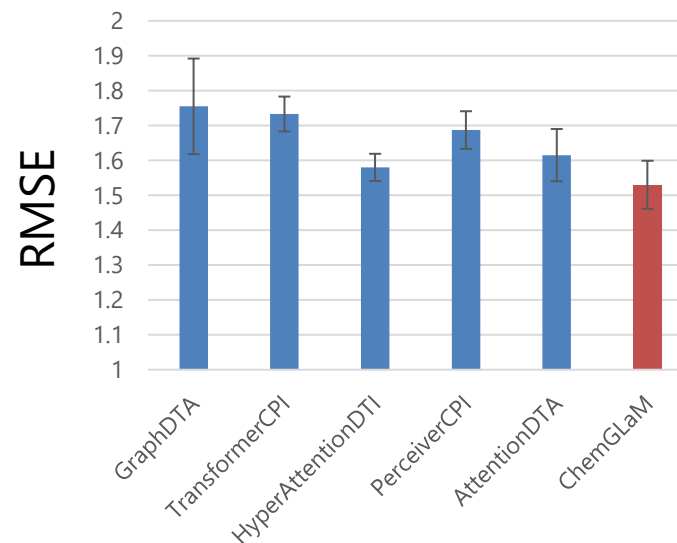
Application Example2 (Downstream Task)

ChemGLaM: Chemical Genomics Language Models for Compound-Protein Interaction Prediction



T. Koyama, H. Tsumura, S. Matsumoto, R. Okita, R. Kojima, Y. Okuno
<https://www.biorxiv.org/content/10.1101/2024.02.13.580100v2>

Figure 1: Schematic illustration of ChemGLaM framework

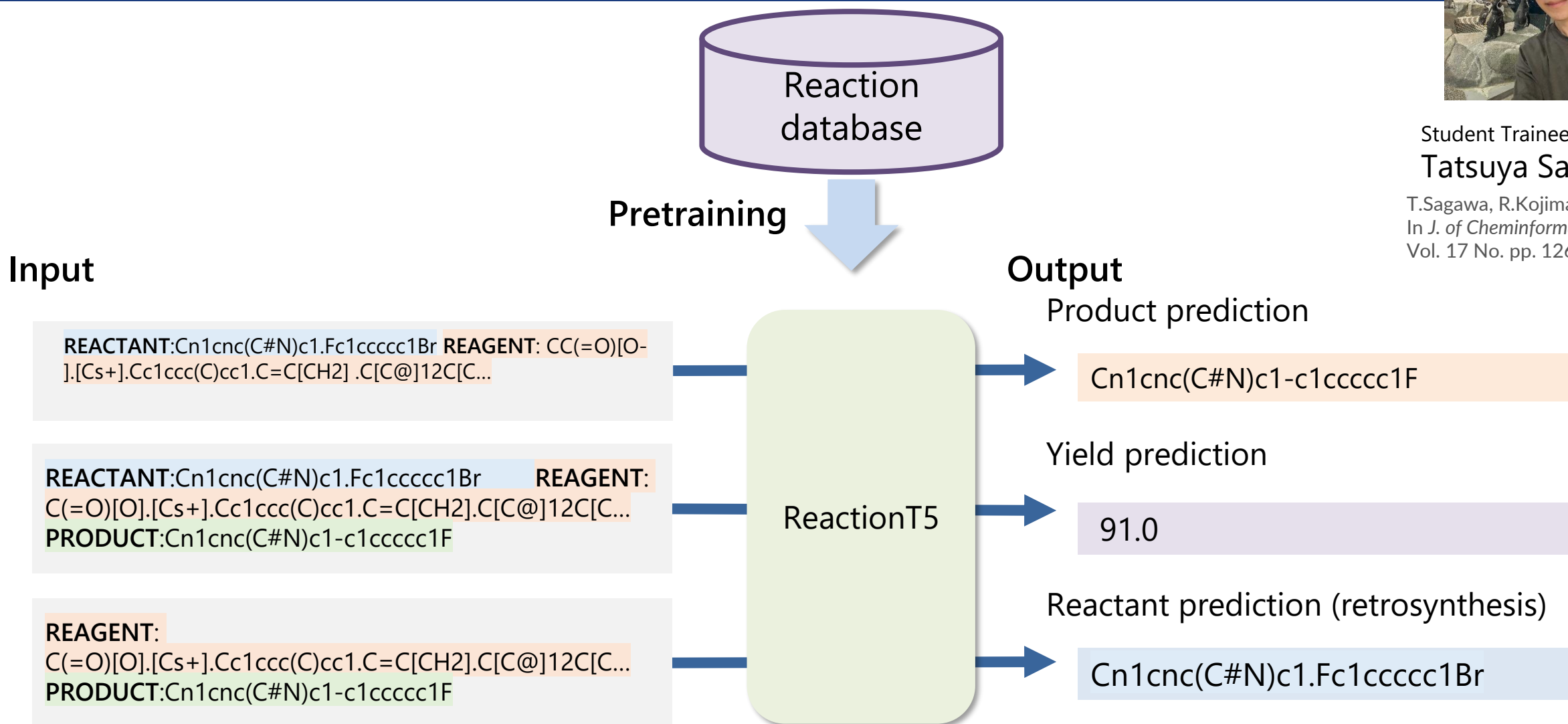


A reaction foundation model: ReactionT5

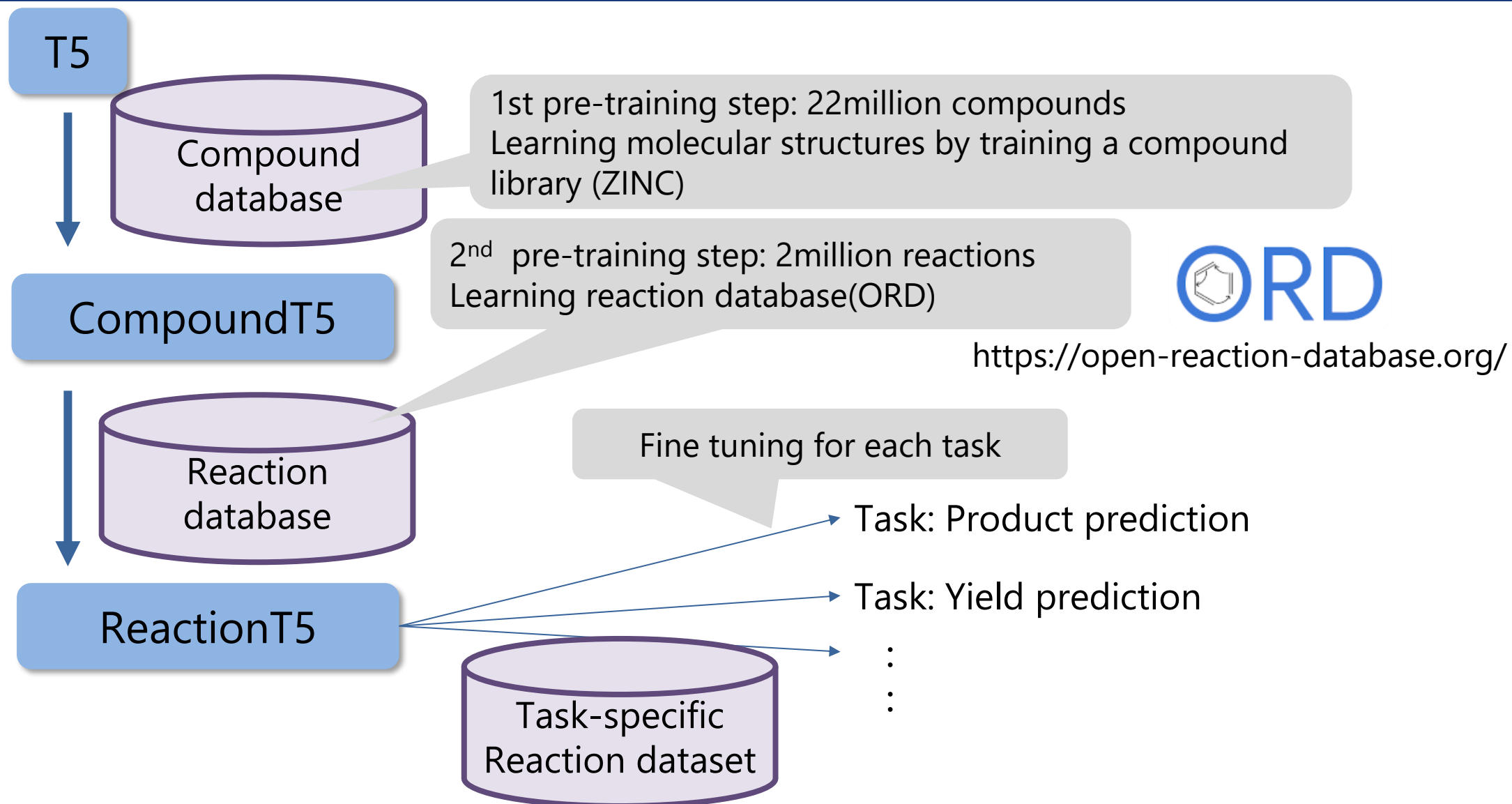


Student Trainee (M2)
Tatsuya Sagawa

T.Sagawa, R.Kojima
In *J. of Cheminformatics*,
Vol. 17 No. pp. 126, 2025.



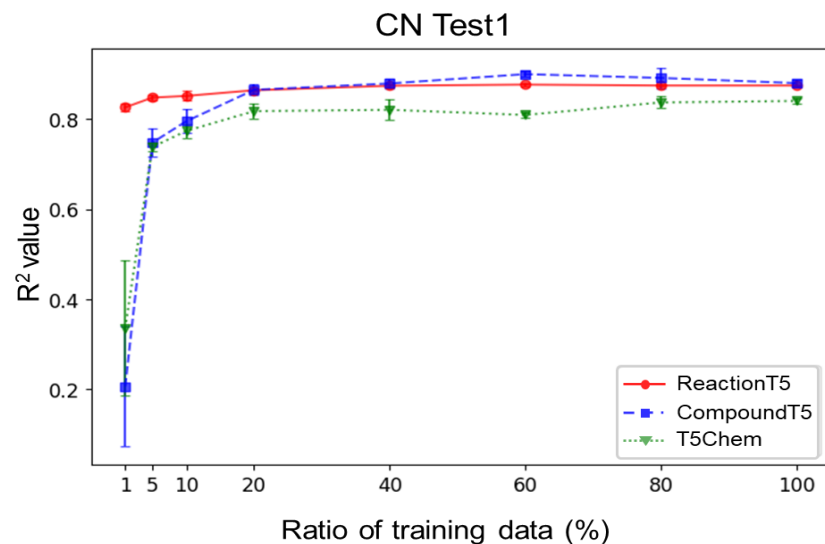
Pre-training / finetuning of ReactionT5



Evaluation: fine-tuning on small amount of data

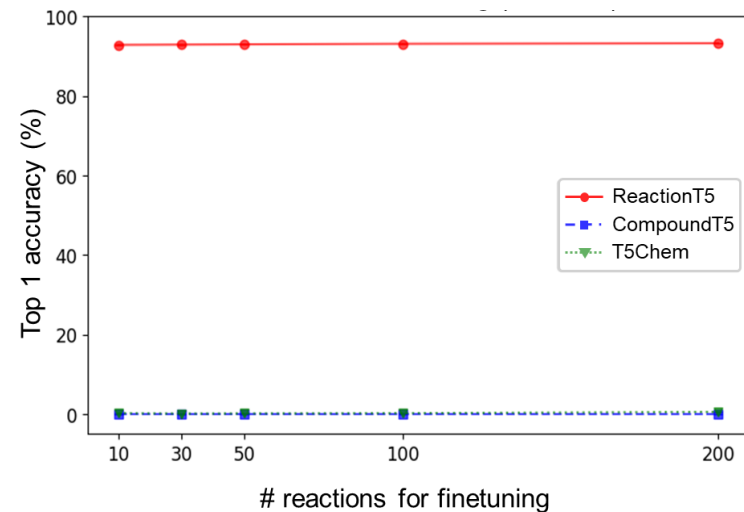
Task: yield prediction

Evaluation: C-N cross-coupling dataset
[Ahneman+]

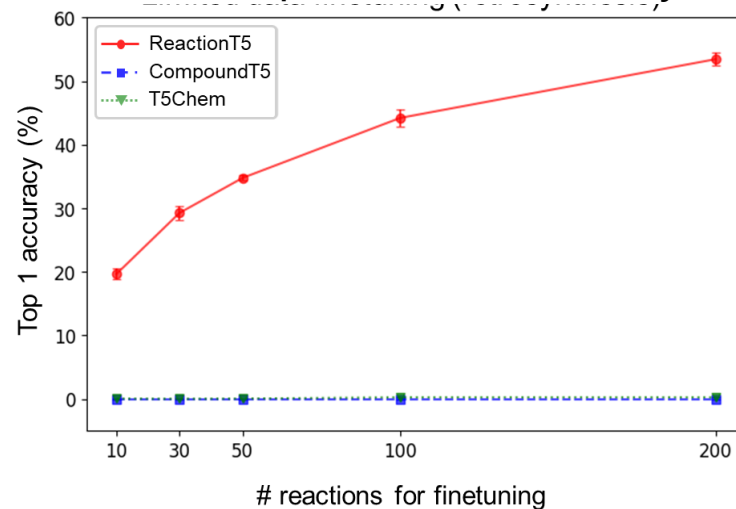


- Significant improvement over T5Chem when data for the downstream task is limited
- Achieves state-of-the-art performance on benchmarks

Product SMILES prediction (Evaluation: USPTO)



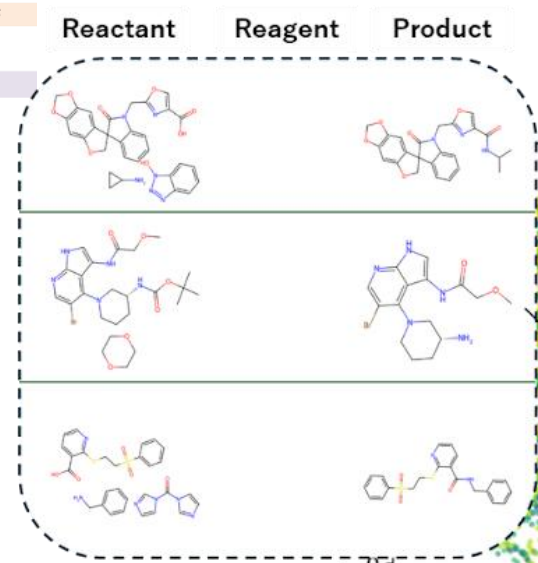
Reactant SMILES prediction/ Retrosynthesis (Evaluation: USPTO)



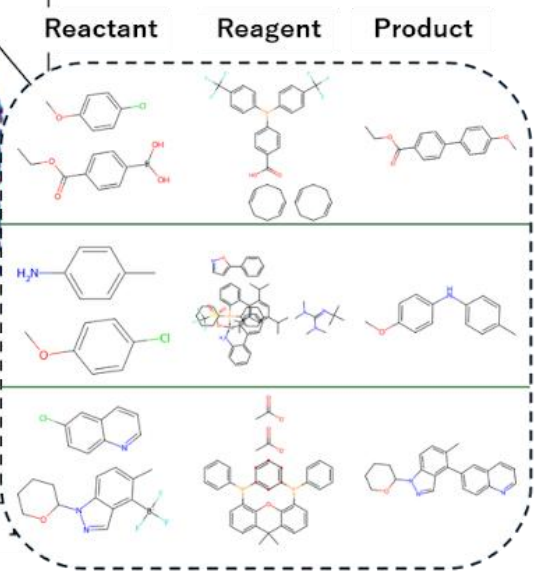
Visualization of latent space

REACTANT: Cn1cnc(C#N)c1.Fc1cccc1Br
 REAGENT: C(=O)[O-].[Cs+].[Cc1ccc(O)cc1.C=C[CH2].[C@H]12C[C@@H]1C2
 REACTANT: Cn1cnc(C#N)c1.Fc1cccc1Br
 REAGENT: C(=O)[O-].[Cs+].[Cc1ccc(O)cc1.C=C[CH2].[C@H]12C[C@@H]1C2
 PRODUCT: Cn1cnc(C#N)c1-c1cccc1F

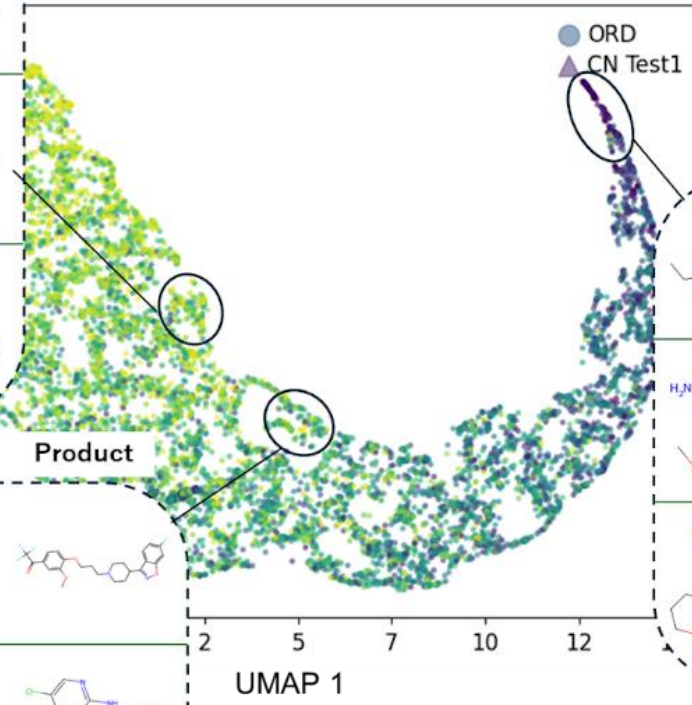
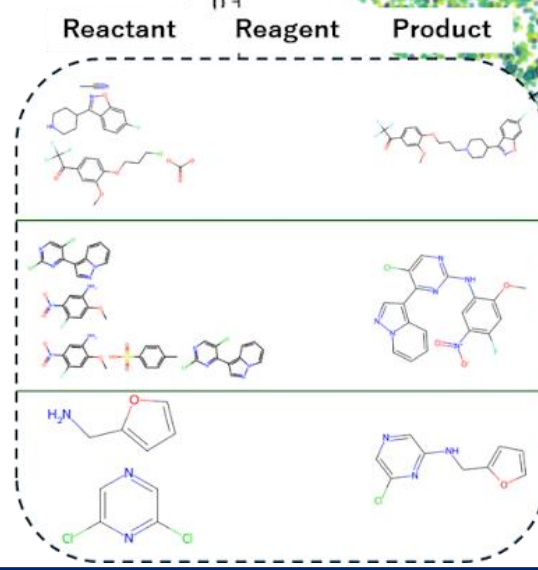
ReactionT5
 Cn1cnc(C#N)c1-c1cccc1F
 91.0



A phosphorus-containing compound as a reagent



Reactions related to an amide bond

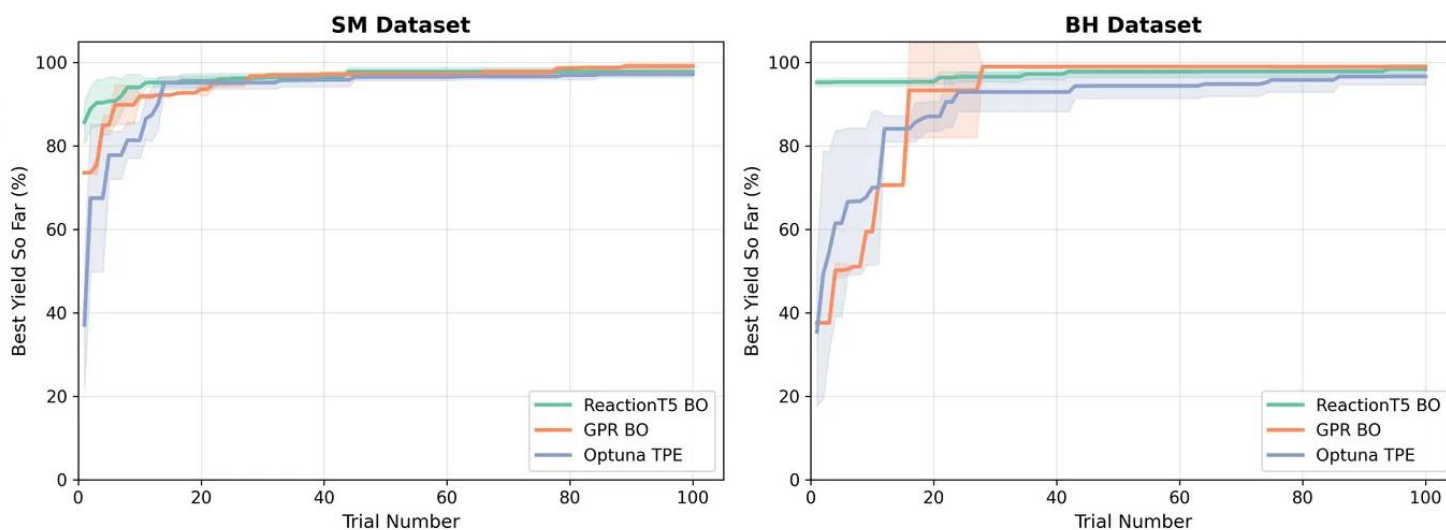


Nucleophilic substitution reactions between a halogenated hydrocarbon and nitrogen

Application: ReactionT5 for Bayesian optimization

Kazumasa Okamoto
Internship student (B2)

Using ReactionT5 as a surrogate model instead of GP for Bayesian optimization to explore yield-maximizing conditions



Benchmark dataset
from rxn4chemistry

Space

Suzuki-Miyaura(SM)
(ORD does not contain
any reactions)

4 reactant1 x
3 catalyst x
11 ligands x
7 reagents x
4 solvents

Buchwald-Hartwig(BH)
(ORD contains all
reactions)

2 ligands x
22 additives x
3 bases x
15 aryl halides

Using ReactionT5 for optimization allows for early optimization.

The closer the reaction space is to the training data, the better the performance can be expected.

ReactionT5 is available

All models and source codes are available:

Demo: yield prediction

<https://huggingface.co/spaces/sagawa/ReactionT5-yield-prediction>

Demo: product prediction

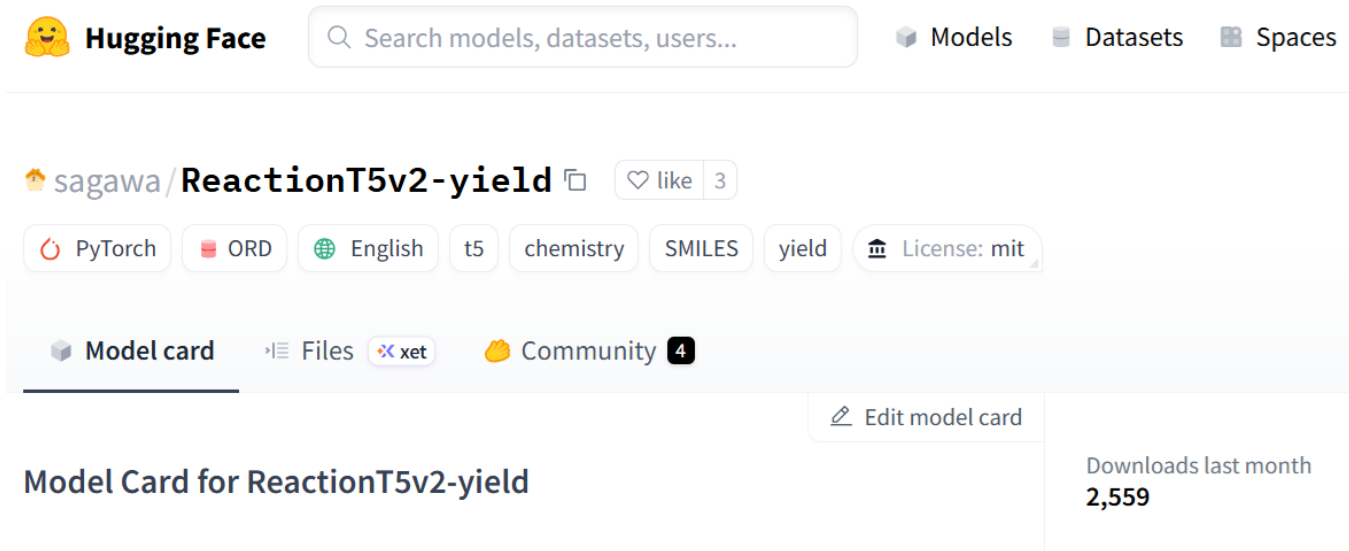
<https://huggingface.co/spaces/sagawa/ReactionT5-product-prediction>

Trained models:

<https://huggingface.co/sagawa/ReactionT5-yield-prediction>

<https://huggingface.co/sagawa/ReactionT5-product-prediction>

<https://huggingface.co/sagawa/ReactionT5-product-prediction>



The screenshot shows the Hugging Face interface for the model 'sagawa/ReactionT5v2-yield'. At the top, there is the Hugging Face logo and a search bar. Below the search bar, the model name 'sagawa/ReactionT5v2-yield' is displayed with a 'like' button showing 3 likes. A row of tags includes 'PyTorch', 'ORD', 'English', 't5', 'chemistry', 'SMILES', 'yield', and 'License: mit'. Below the tags, there are buttons for 'Model card', 'Files', 'xet', and 'Community' (with 4 members). On the right side, there is an 'Edit model card' button and a box showing 'Downloads last month 2,559'. The main title of the model card is 'Model Card for ReactionT5v2-yield'.

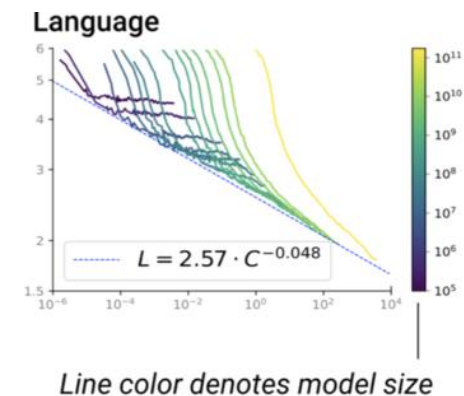
Source codes

<https://github.com/sagawatatsuya/ReactionT5>

Neural scaling raw in chemistry [on going]

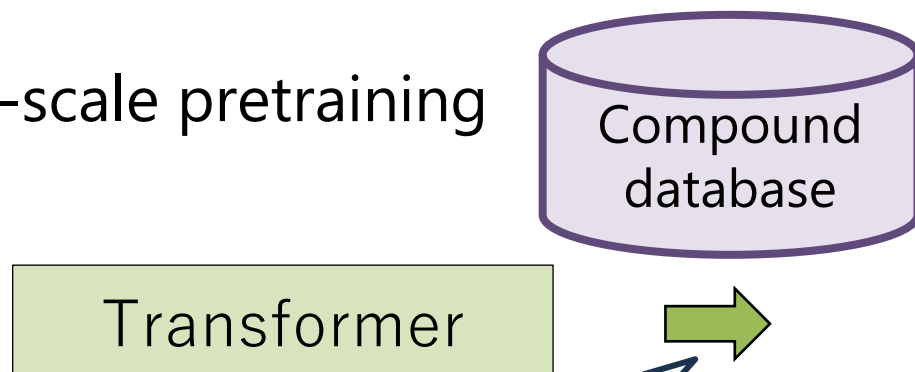
Neural scaling laws [OpenAI 2020]:

- Model performance improves as the amount of data and model size increase



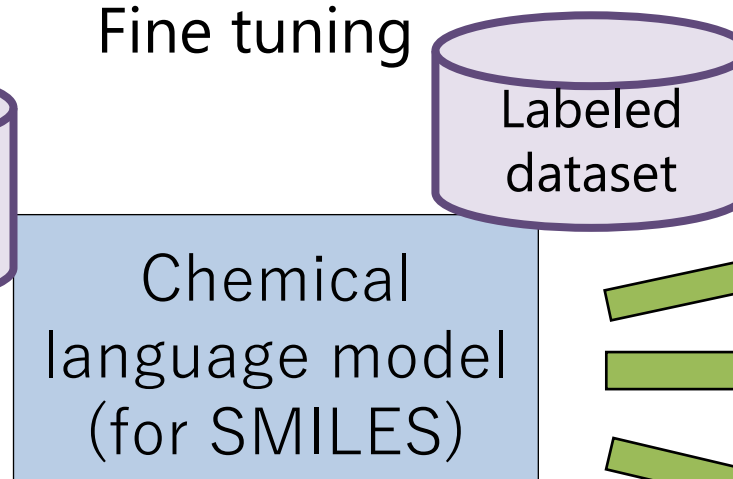
Neural scaling laws in chemistry

Large-scale pretraining



Q1. Do scaling laws exist in molecular pre-training?

Fine tuning

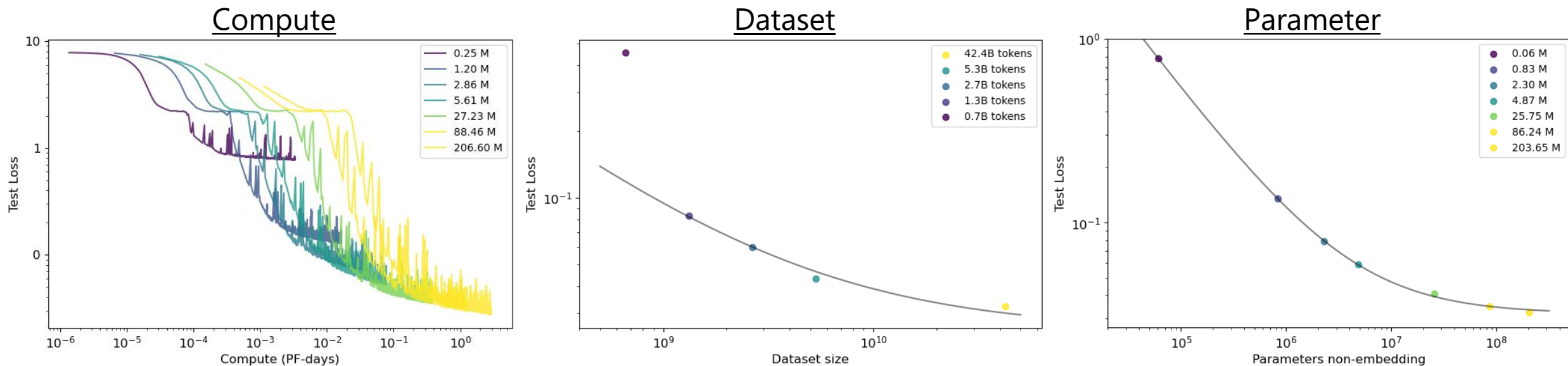


Q2. Do scaling laws exist in molecular fine-tuning?

Molecular property prediction models

- Solubility
 - Toxicity
 - Pharmacokinetic property
 - Quantum chemistry parameters
- etc...

Q1. Do scaling laws exist in molecular pre-training?



Answer: Yes !

Neural scaling law for chemical language model for SMILES:

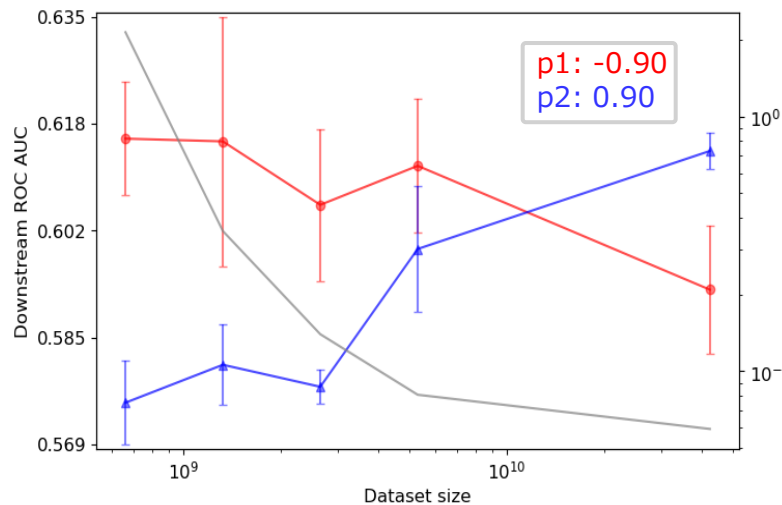
The following power law is observed for dataset size and model size:

$$L(D) \approx 2.53 \times 10^{-2} + 1.09 \times 10^5 \times D^{-0.689}$$

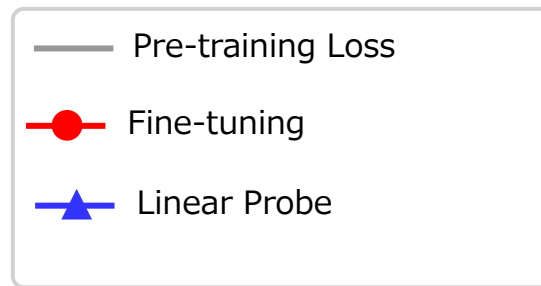
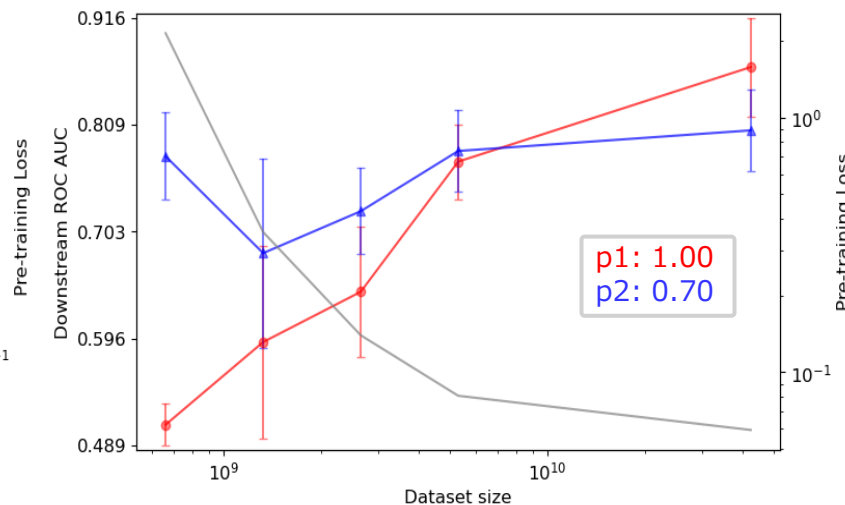
$$L(N) \approx 3.19 \times 10^{-2} + 3.23 \times 10^3 \times N^{-0.760}$$

Q2. Do scaling laws exist in molecular fine-tuning?

SIDER

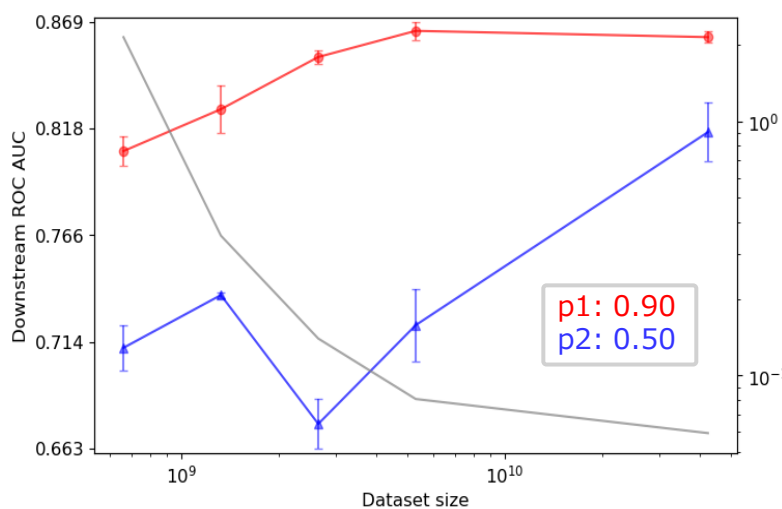


ClinTox

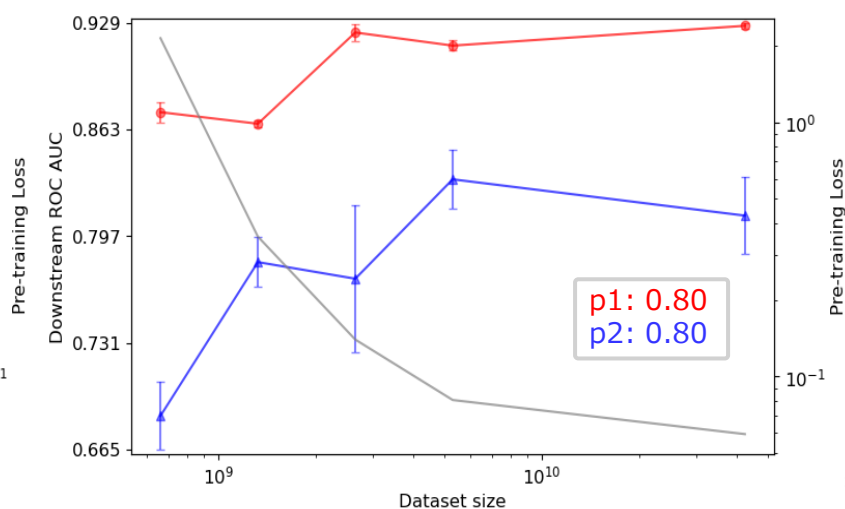


Linear probe is a method to tune only the final layer of NN for each downstream task

BACE



BBBP



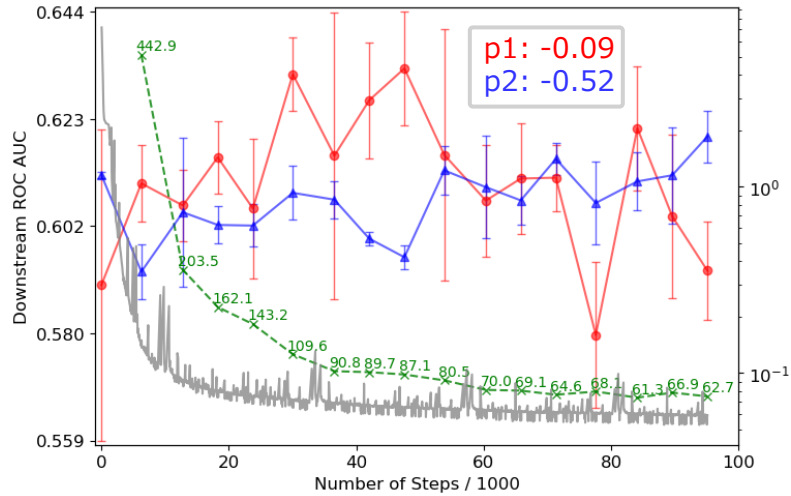
Answer: Almost yes for finetuning dataset size

Finetuning data for downstream tasks is often good, but depends on the task (and quality of the data)

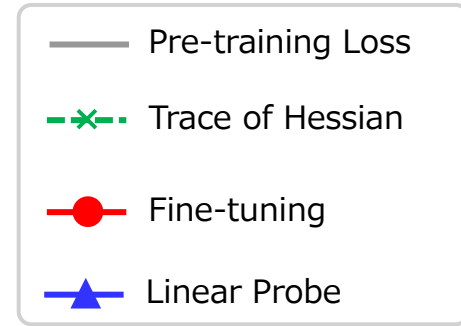
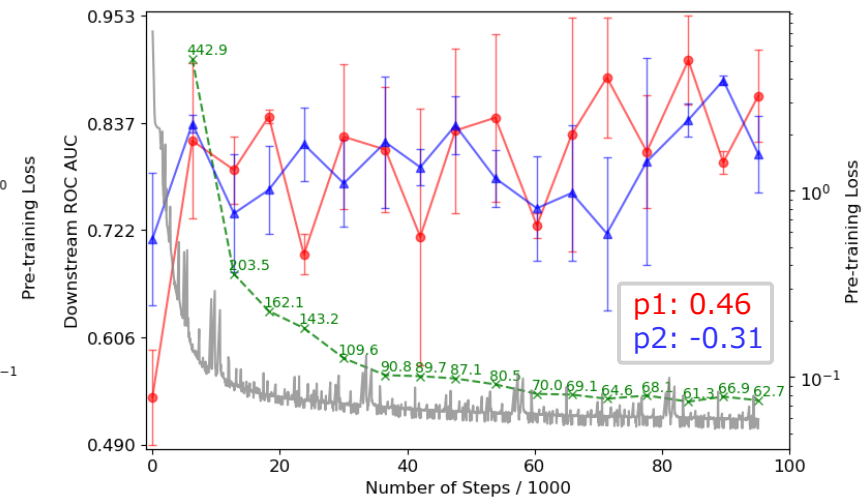
⇒ How about the pre-training data?

Q2'. Does improved pre-training performance lead to improved performance on downstream tasks?

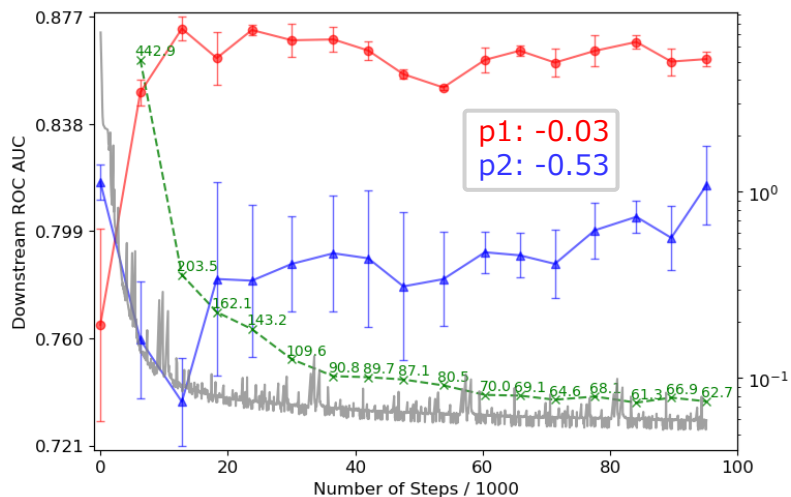
SIDER



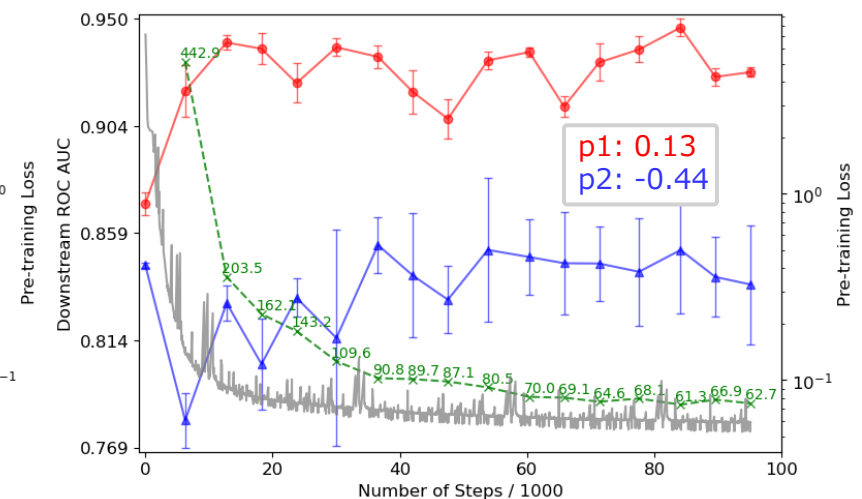
ClinTox



BACE



BBBP



Answer: No!!!

Those are not correlated, and a decrease in pre-training loss does not improve performance in downstream tasks.

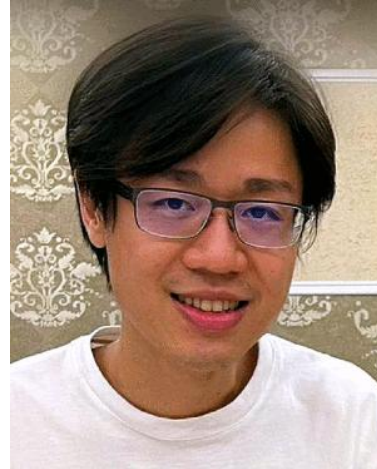
Laboratory for Multimodal AI framework



Team director
Ryosuke Kojima



Senior Scientist
Yoshinobu Igarashi



Postdoctoral Researchers
Yen Benjamin



Technical staff
Kazunobu Matsubara



Student Trainee
Tatsuya Sagawa

We're Looking for Collaborations!

We're open to research and development collaborations.
We also welcome internship students and visiting researchers who want to learn about AI.

Part-time internship students
(including KU intern members)
6 bachelor students
3 master students
3 Ph.D. students