2023/8/31 近化若手フォーラム【第６回】

有機化学のための深層学習ベースの AI基盤と基盤モデル

小島　諒介
（京都大学　医学研究科）

自己紹介

名前：小島諒介

略歴：

2014年　東京工業大学　情報理工学研究科 学士・修士課程

2017年　東京工業大学　博士（工学）　機械学習

2017年～　京都大学　医学研究科　機械学習＋ロボット
人間健康科学系専攻　ビッグデータ医科学分野
奥野恭史研究室

2021年～　京都大学　医学研究科
DXセンター医療データ活用部門
講師　創薬・医療＋機械学習

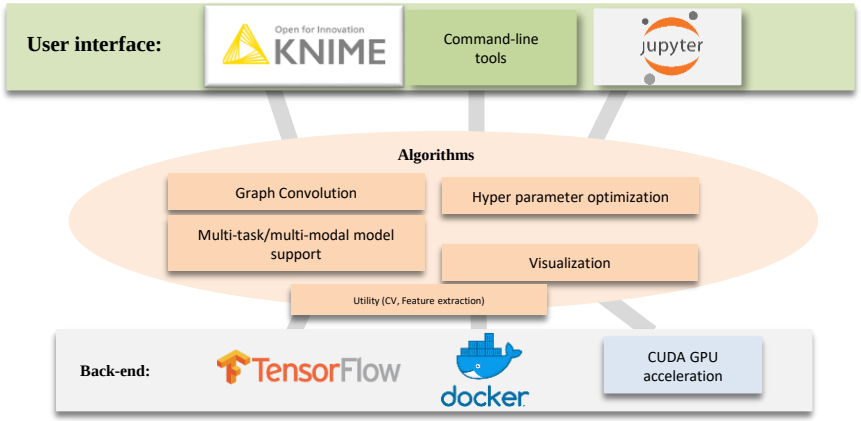
興味：
創薬/医療のための
“実際に使える” 人工知能の手法開発と応用(AI基盤)

ＡＩ基盤とその周辺の研究(1/2)

ライフサイエンスのためのAI基盤：kGCN

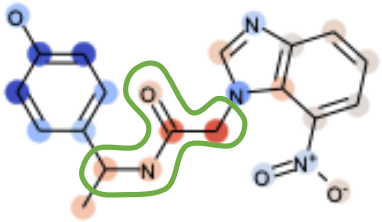
・有機化学でもよく利用されるグラフニューラルネットワークに関するツール群

・さまざまなレベルの人が使えるように複数のインターフェースを備える



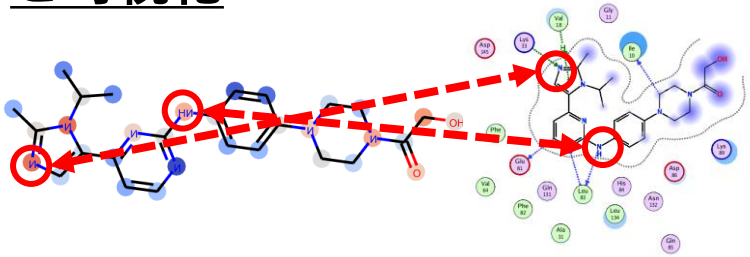
https://github.com/clinfo/kGCN

化合物の反応予測



S.Ishida, K.Taerayama, R.Kojima, K.Takasu, Y.Okuno: **JCIM 2019**.

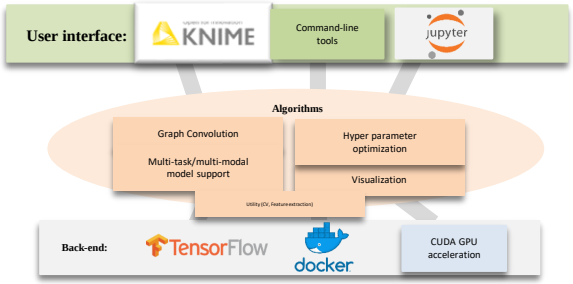
薬剤の活性予測と可視化



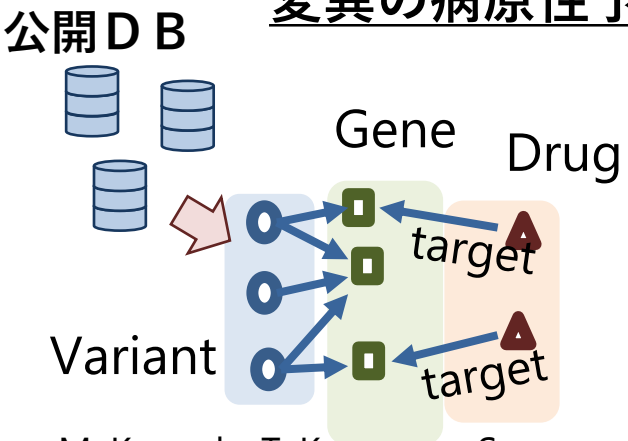
R.Kojima, S.Ishida, M.Ohta, H.Iwata, T.Honma, Y.Okuno: In **JChemInfo**, **2020**.

簡単に利用できる一方、習得すれば最新の研究にも利用可能

ＡＩ基盤とその周辺の研究(2/2)

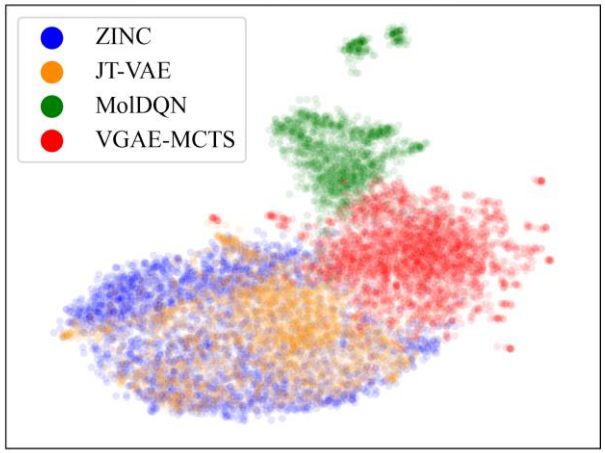


ゲノム関連知識グラフから変異の病原性予測



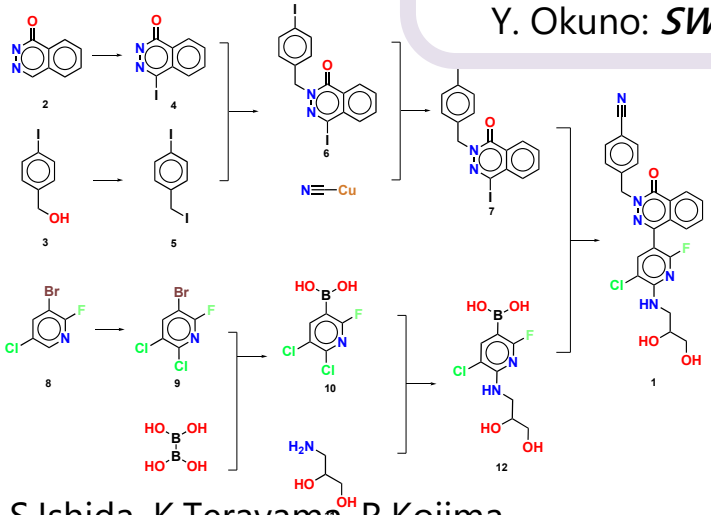
M. Kamada, T. Katayama, S. Kawashima, R. Kojima, M. Nakatsui, Y. Okuno: ***SWAT4LS*, 2017.**

生成モデル



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

逆合成経路解析



S.Ishida, K.Terayama, R.Kojima, K.Takasu, Y.Okuno: ***JCIM* 2022.**

グラフベース予測モデル＋ Federated Learning

kMoL is a machine learning library for drug discovery and life sciences, with federated learning capabilities. Some of its features include state-of-the-art graph-based predictive models, explainable AI components, and differential privacy for data protection. The library was benchmarked on datasets containing ADME properties (Absorption, Distribution, Metabolism, Excretion), toxicity, and binding affinities values.

Models are built using PyTorch and PyTorch Geometric.

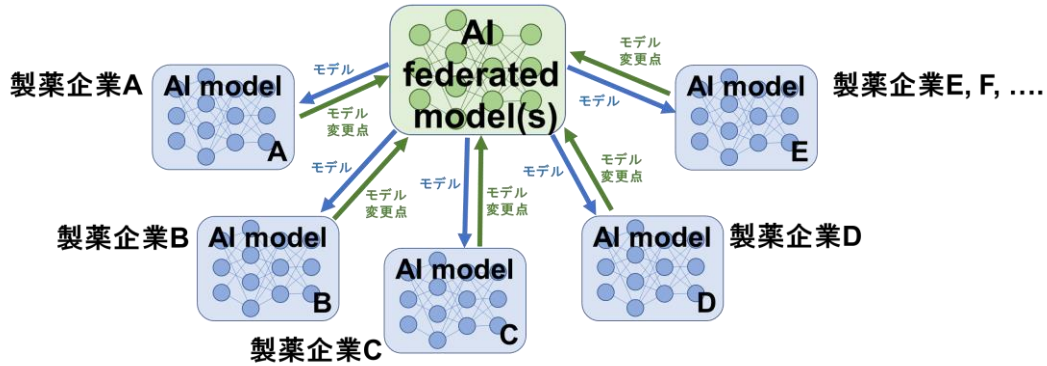
Installation

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

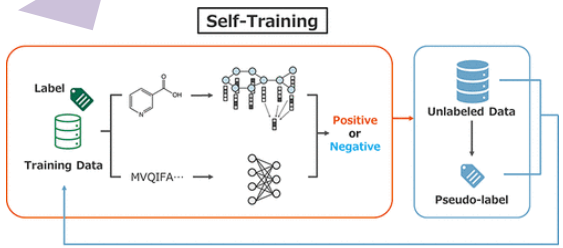
kGCNでの知見をもとに分子に特化して再開発した新規ライブラリ

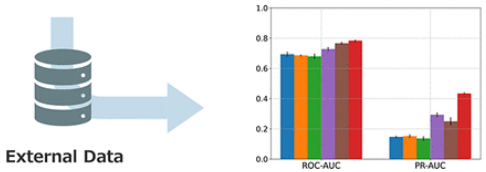
グラフベース予測モデル

＋**Federated Learning（連合学習）**



薬剤の活性予測の自己教師あり学習



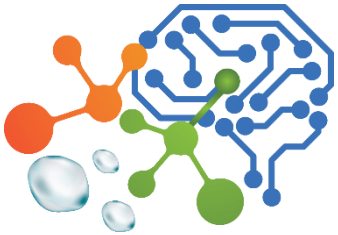


T.Koyama, S.Matsumoto, H.Iwata, R.Kojima, Y.Okuno, *JCIM* 2023

背景: 有機化学のための基盤モデル

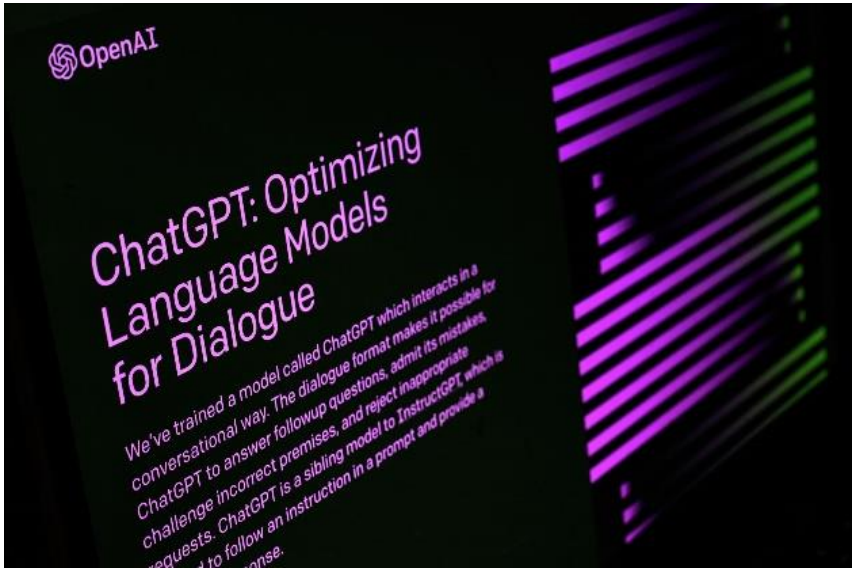
デジタル有機合成

文部科学省科学研究費助成事業
「学術変革研究A:
　　デジタル化による高度精密有機合成の新展開」

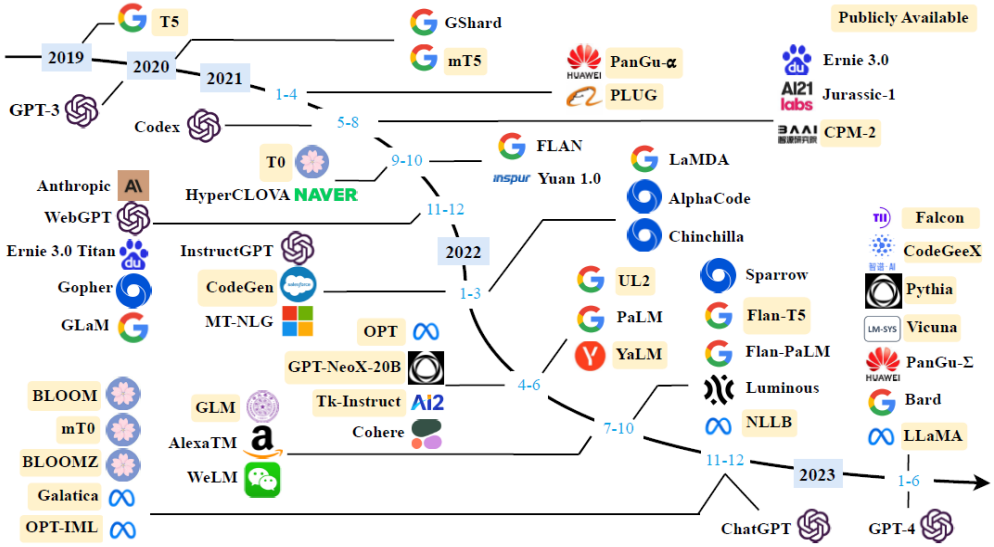


A03-2「データ駆動有機合成経路・反応予測モデルの構築」
　実際に実験現場で役に立つ
　有機化学に関するモデルを開発することを目指す

背景: 基盤モデル



大規模なデータベースを用いた事前学習
⇒きわめて少数のデータでチューニング
（or プロンプトと呼ばれる入力を工夫）
することで高い性能を達成する
（or 様々なタスクを解く）
ことができる

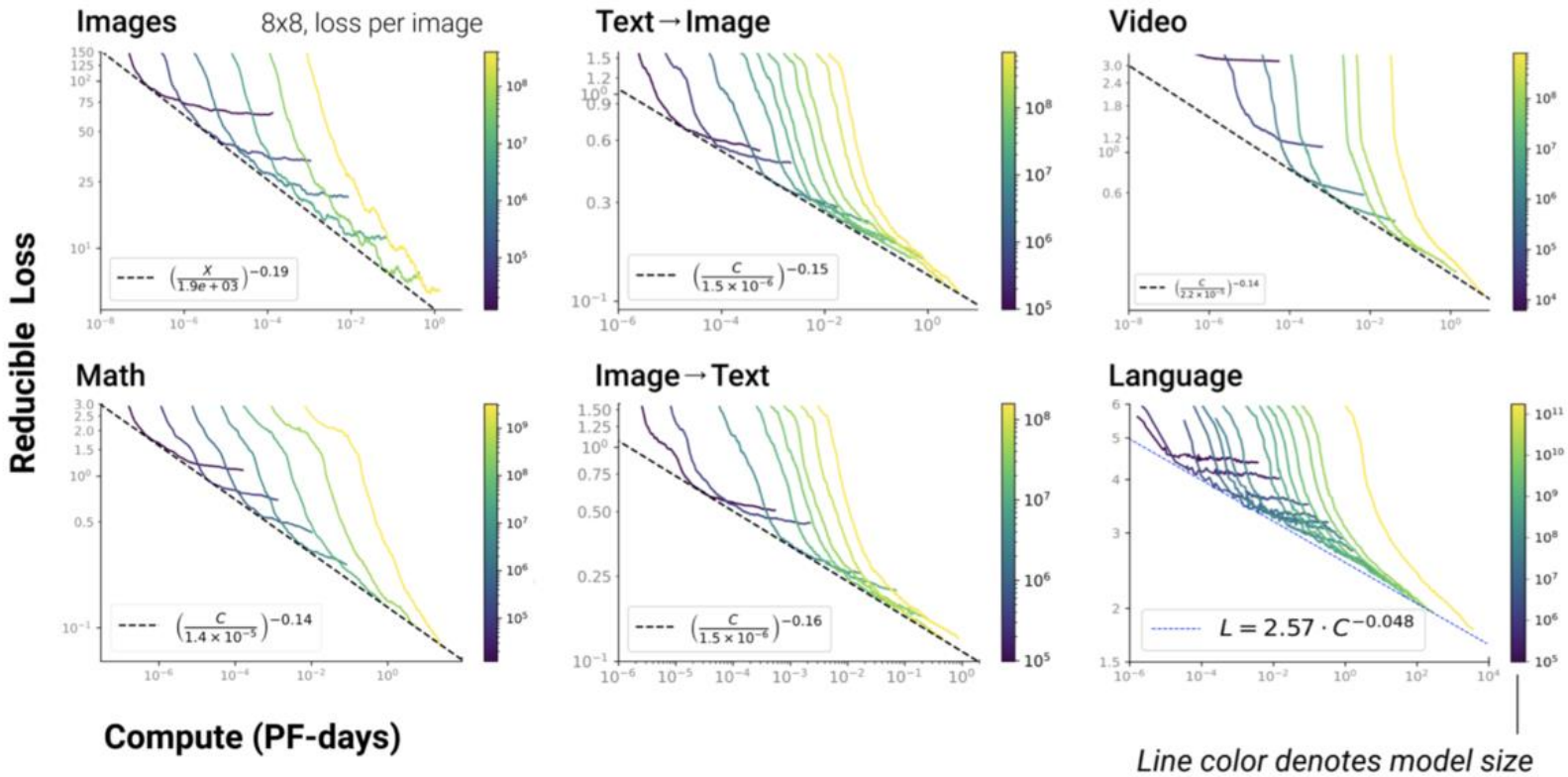


A Survey of Large Language Models
https://arxiv.org/abs/2303.18223

特にChatGPTを中心に，自然言語において大規模言語モデルが注目されている

大規模モデルへの期待の高まり

Scaling law: 様々なタスクでパラメータを増やすと性能がべき乗でよくなっていく [OpenAI 2020]

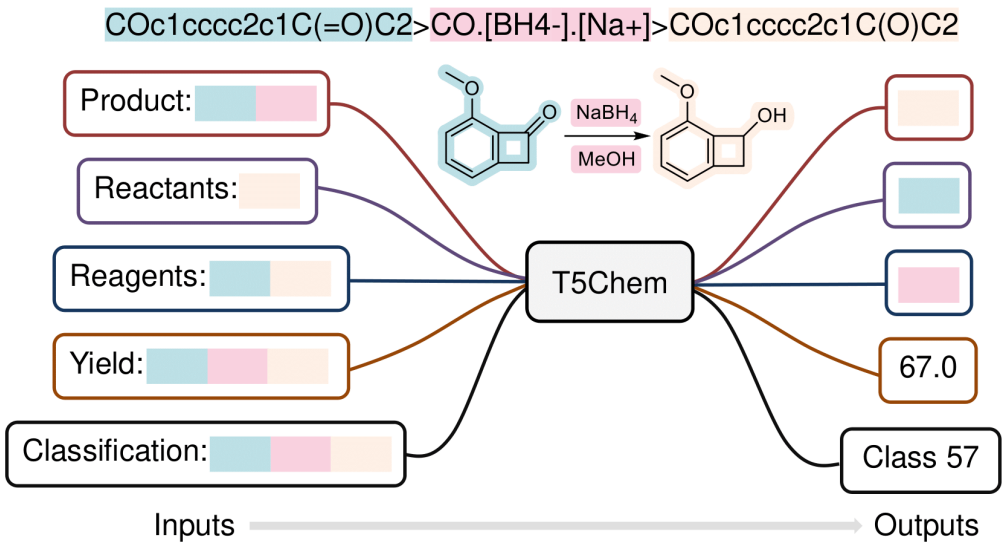


Chemistry？

既存研究：化学に関するタスクに応用した例

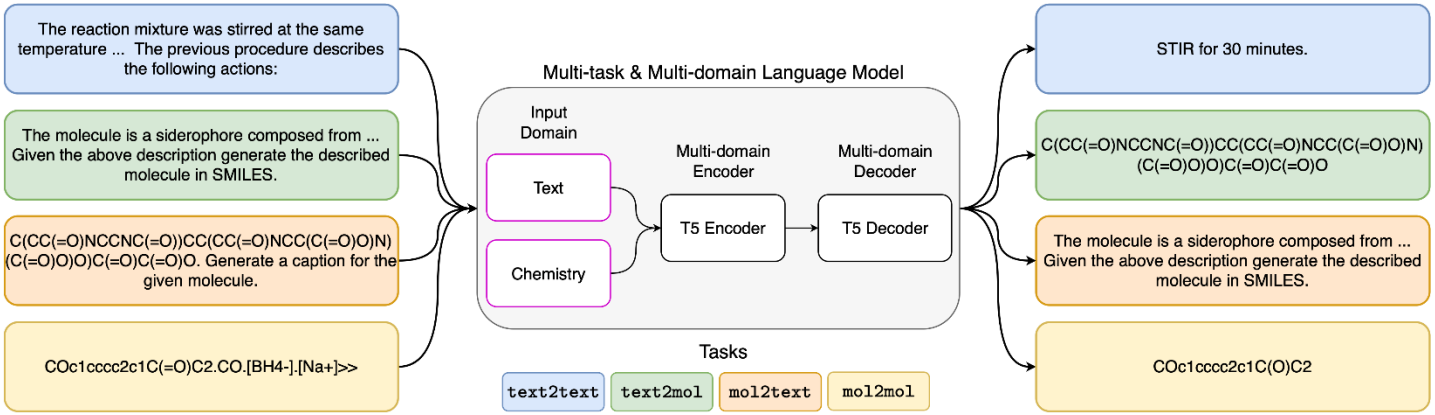
T5Chem [Lu+ 2022]

主に特許のデータで学習し、いくつかのハイスループットデータで評価



Multitask Text and Chemistry T5 [Christofidellis+ 2022]

Q＆A形式で化学に関する自然言語の問いに回答するモデル



課題とアプローチ

- 反応に特化した大規模事前学習モデルは少ない

  単一分子に関するモデルがほとんど

  ➡ 反応をテキスト化し，テキストで入力と出力を表現したモデルを構築する

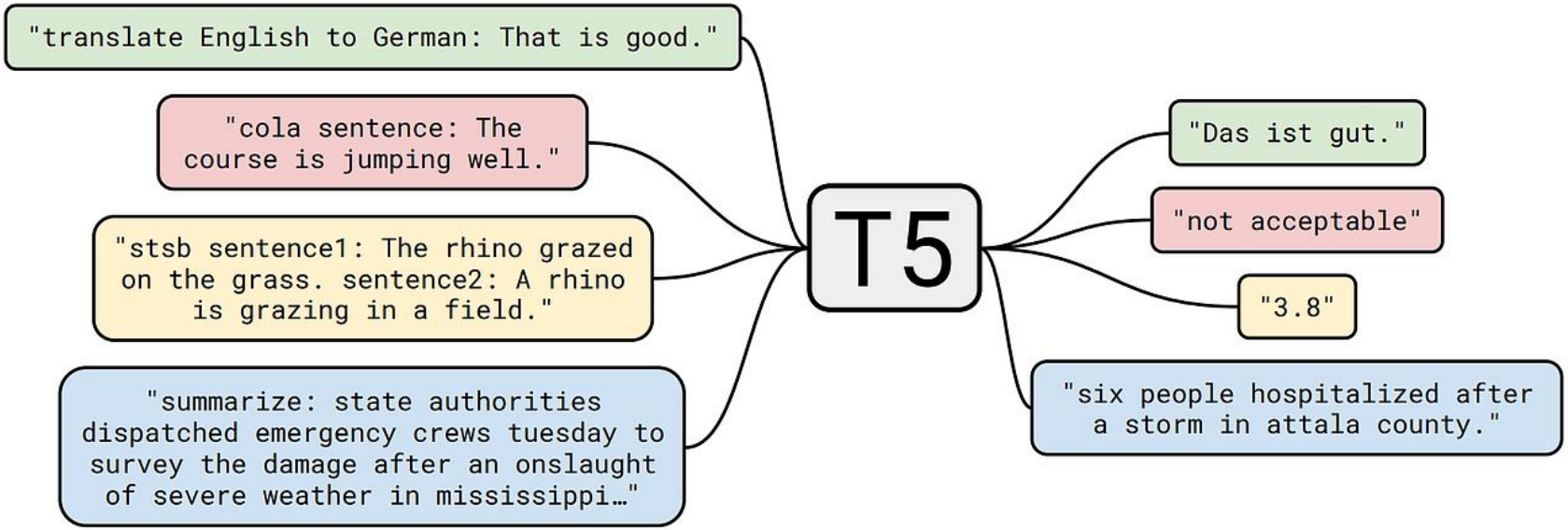
- 反応が特許に依存している

  ドメイン外の反応に関して非常に性能が劣化する

  ➡ 大量の反応を含むOpen Reaction Database(後述)を利用して学習，適宜ファインチューニングも併用

Text-to-Text Transfer Transformer

T5(Text-to-Text Transfer Transformer)
テキスト入力テキスト出力の問題を解くためのモデル



https://arxiv.org/abs/1910.10683

二つの学習方法を組み合わせて学習する
・テキストの一部を隠して再構築する学習
・入出力ペアからの学習

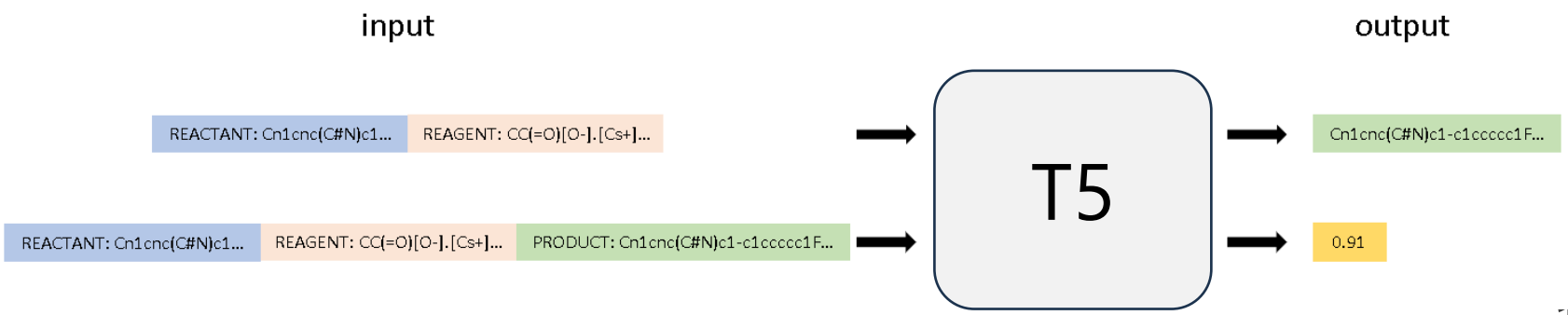
反応のテキスト化

テキスト化

reactant	reagent	solvent
Cn1cnc(C#N)c1.Fc1ccccc1Br	CC(=O)[O-].[Cs+]	Cc1ccc(C)cc1

catalyst	product	yield
C=C[CH2-].C[C@]12C[C@@]3(C)O[C@](C)(C[C@@](C)(O1)O3)P2c1ccccc1.Cl[Pd+]	Cn1cnc(C#N)c1-c1ccccc1F.Cn1cnc(C#N)c1-c1ccccc1F	91.0

予測したいテキストを出力,
それ以外を入力にして学習



大規模反応データセット(ORD)

RETURN TO ISSUE | < PREV **PERSPECTIVE** NEXT >

The Open Reaction Database

Steven M. Kearnes*, Michael R. Maser, Michael Wleklinski, Anton Kast, Abigail G. Doyle, Spencer D. Dreher, Joel M. Hawkins, Klavs F. Jensen, and Connor W. Coley*

Cite this: *J. Am. Chem. Soc.* 2021, 143, 45, 18820−18826
Publication Date: November 2, 2021
https://doi.org/10.1021/jacs.1c09820
Copyright © 2021 The Authors. Published by American Chemical Society
RIGHTS & PERMISSIONS

Article Views	Altmetric	Citations
19242	127	2

LEARN ABOUT THESE METRICS

Share Add to Export RIS

Journal of the American Chemical Society

2022/01月段階で 2,268,981反応

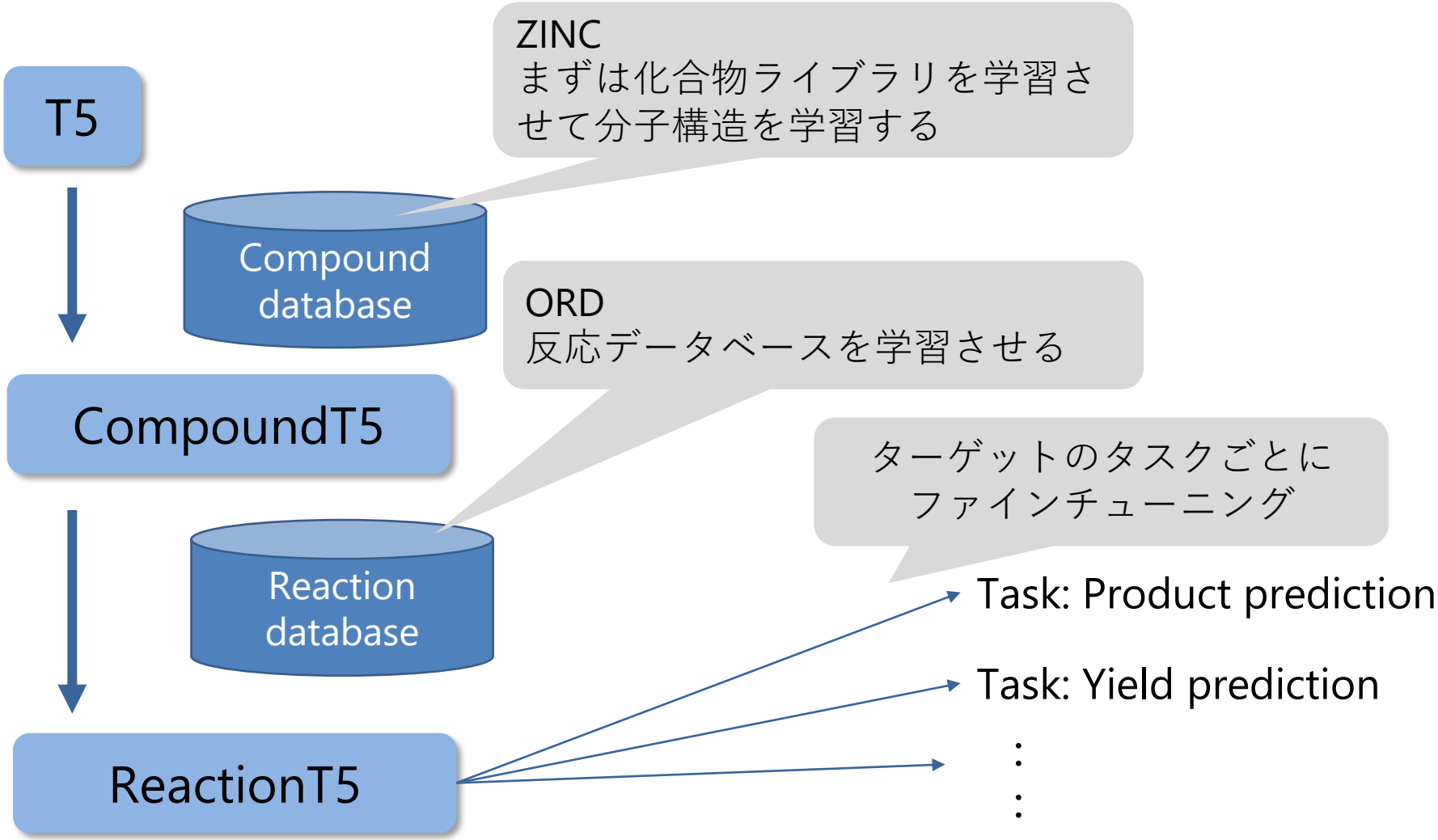
ORD

⇒これを用いて大規模反応モデルを構築できないか？

ORDを用いたT5モデルを構築
（T5の構成はT5Chemを参考にしつつ一部改良）

提案モデル：ReactionT5

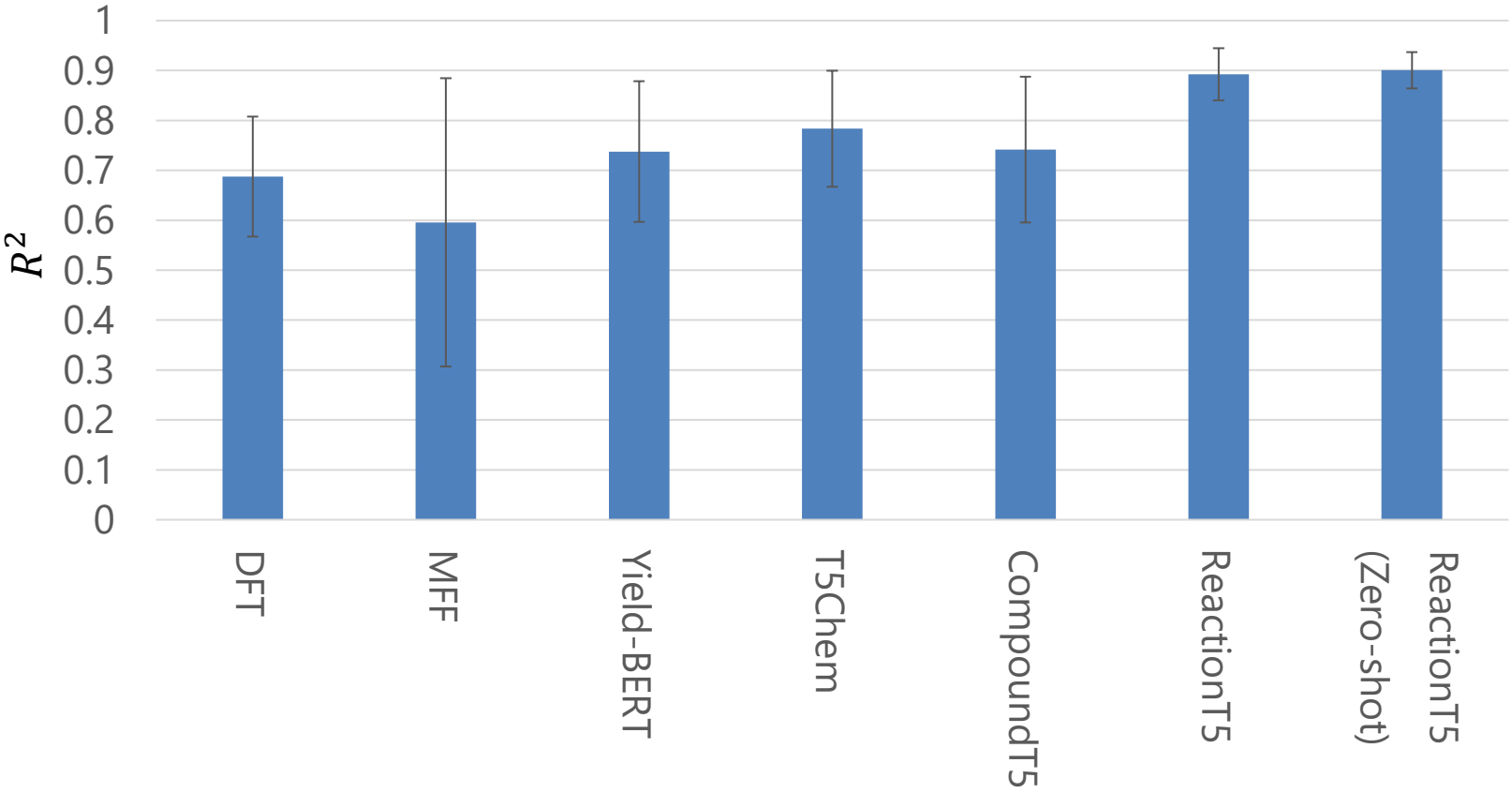
https://github.com/sagawatatsuya/ReactionT5



結果：収率予測(ベンチマーク)

C-N クロスカップリング反応データセット[Ahneman+] を用いて評価
（４つのテストセット）

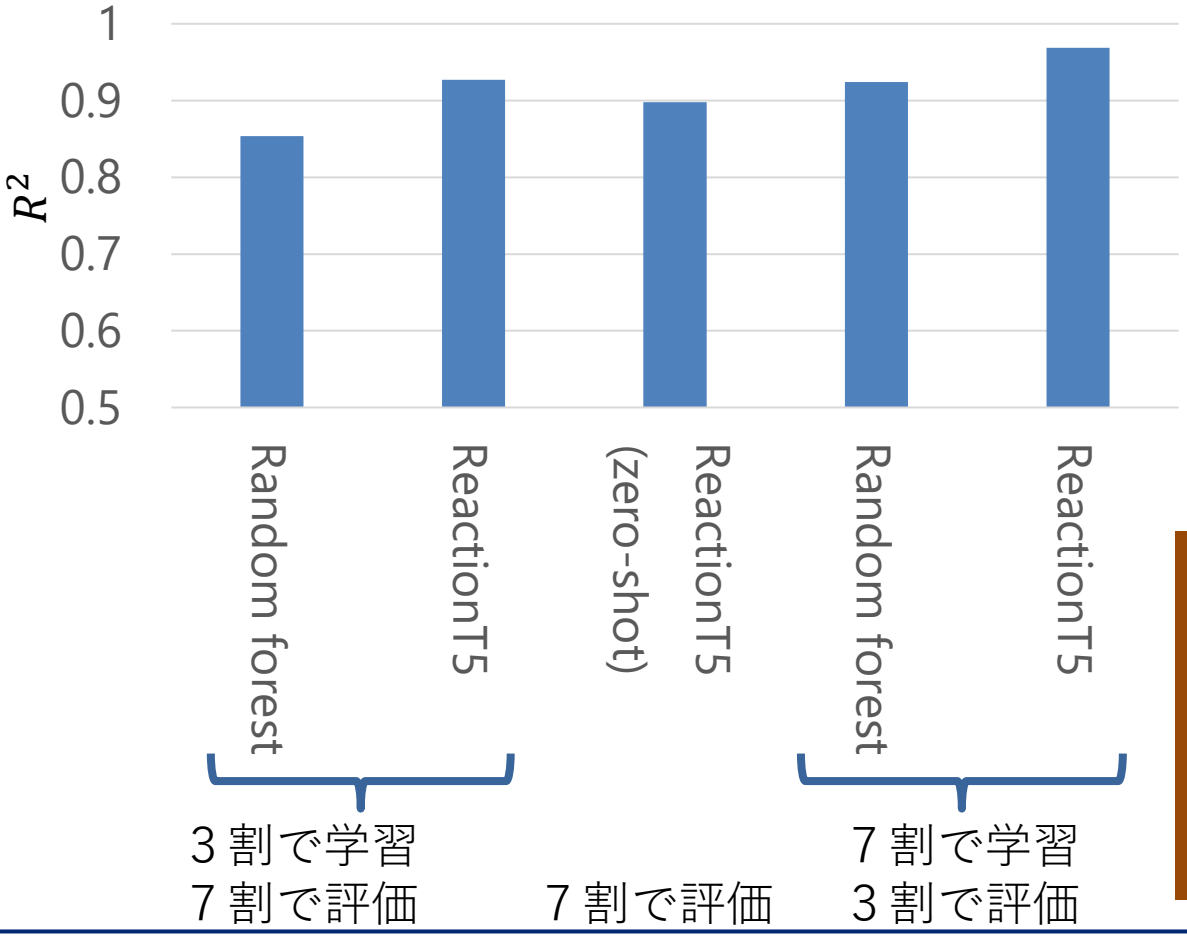
事前学習済みのReactionT5をさらに3,955の反応で
ファインチューニングして収率を予測



結果：収率予測（学習データ量との対応）

C-N クロスカップリング反応データセット[Ahneman+] を用いて評価
（４つのテストセット）

3,955の反応と収率

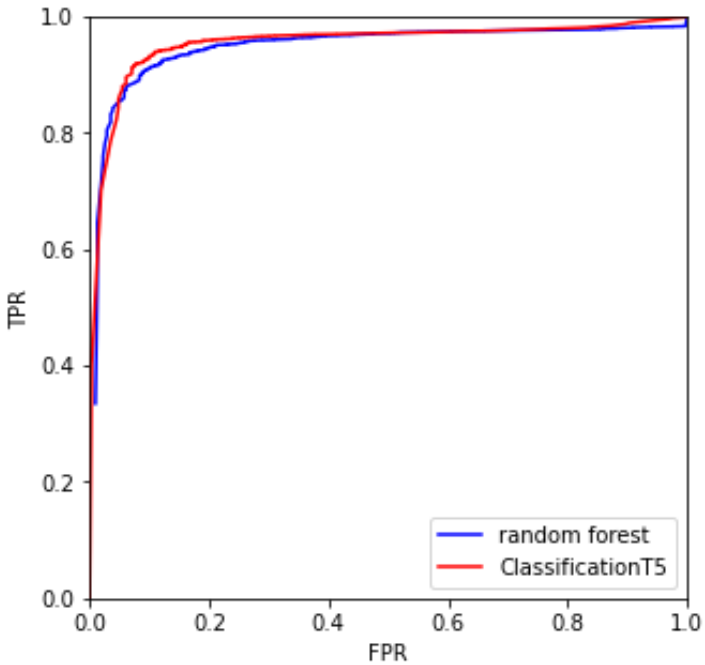


データ少+RFの性能より
Zero-shot でのReactionT5の性能の方がよい

少数でもファインチューニングすればさらに良くなる

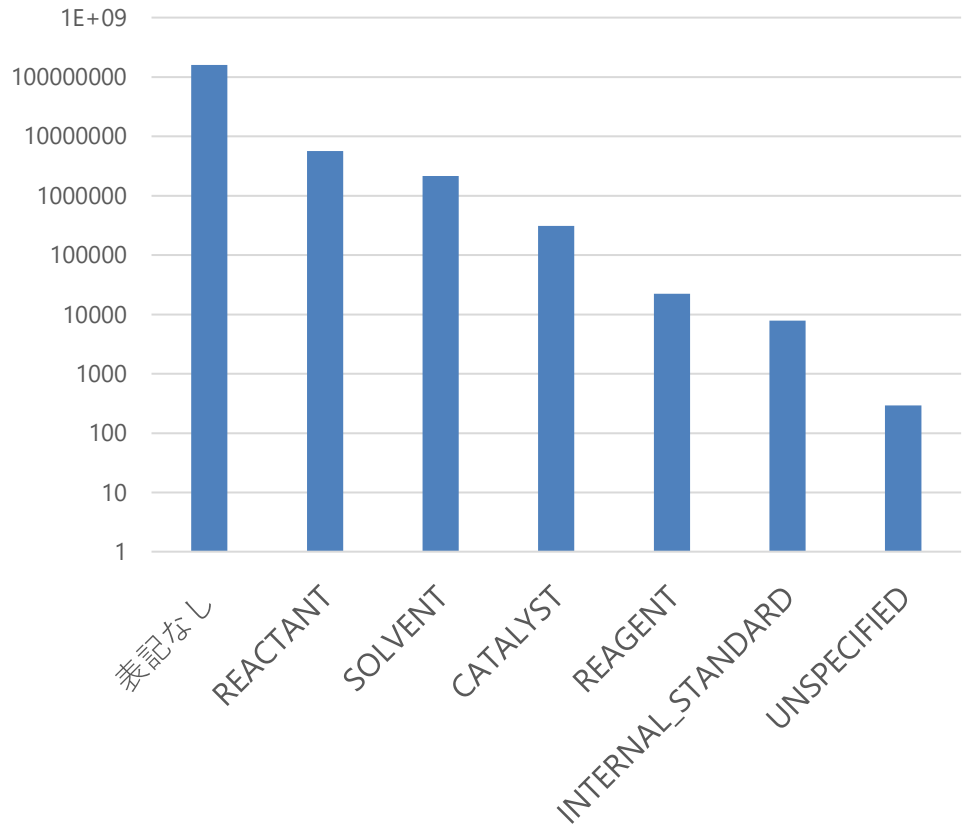
生成物予測の前に：データベースの修復

表記されてるものから
「表記なし」がReactant
かそれ以外かを予測



ORD

Inputの分子が持つ役割(role)の分布

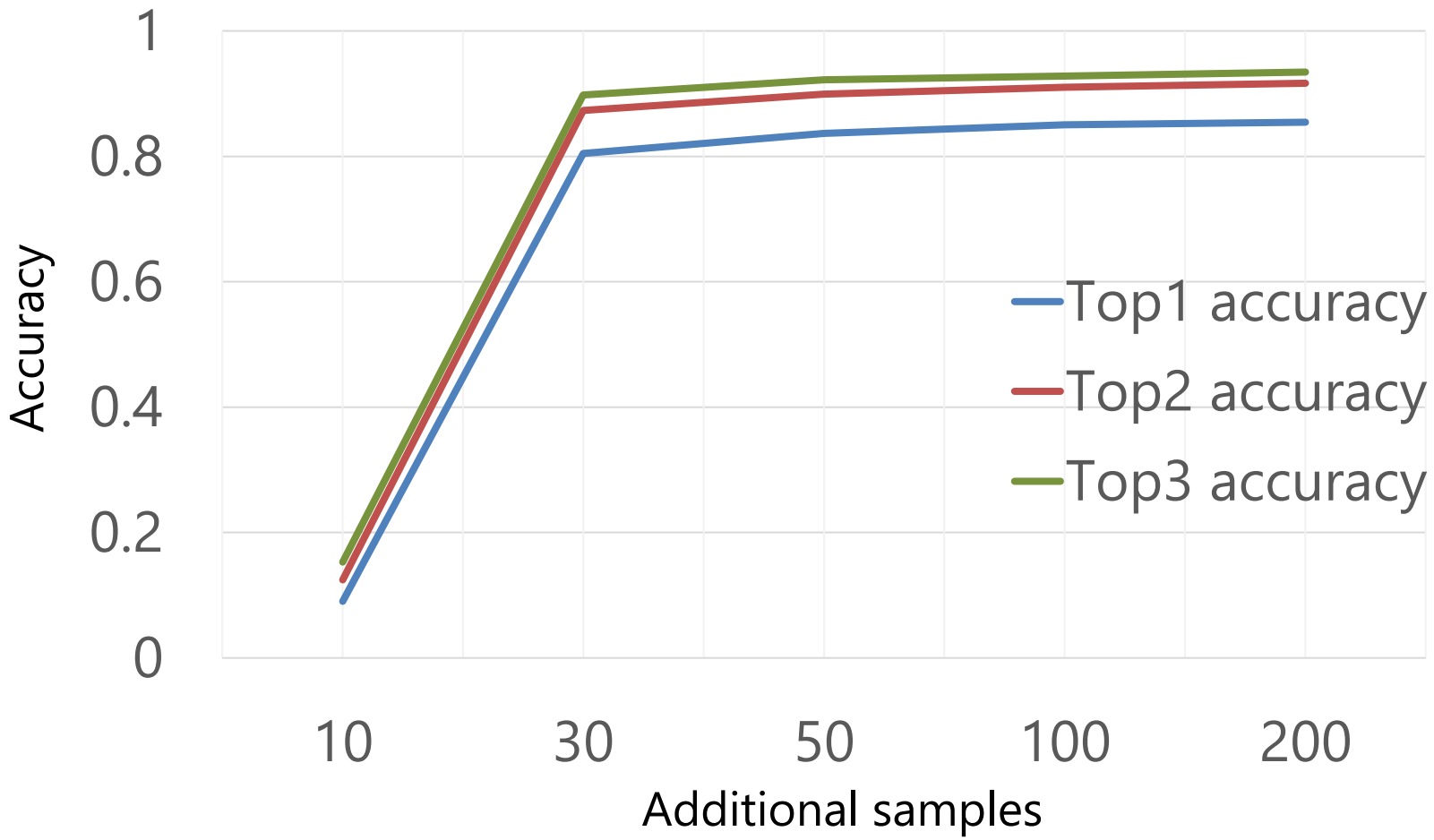


結果：生成物予測

入力：反応前の情報
出力：反応後生成物SMILES
評価：厳密一致
Accuracy[%]
評価データ：USPTO

	Train	Top1	Top2	Top3
Seq-to-seq	USPTO	80.3	84.7	86.2
WLDN	USPTO	80.6	90.5	92.8
Mol Transformer	USPTO	88.8	92.6	–
T5Chem	USPTO	90.4	94.2	–
CompoundT5	USPTO	88.0	92.4	93.9
ReactionT5	ORD	0.0	0.0	0.0
ReactionT5	ORD +USPTO200	0.0	0.0	0.0
ReactionT5	ORD (restored)	0.0	0.0	0.0
ReactionT5	ORD (restored) +USPTO200	85.0	90.6	92.3

追加反応データ数と精度



ドメイン外でも（大きく外れていなければ）数十反応のデータで適応できる可能性がある

まとめ

ORDを用いた大規模事前学習モデルReactionT5を構築・公開

収率予測：zero-shotの予測でもそれなりの性能が出ることを示した
生成物予測：ドメイン外のデータで性能が急激に劣化するが
数十程度のデータでファインチューニングすれば高性能

収率予測のデモ
https://huggingface.co/spaces/sagawa/ReactionT5-yield-prediction

生成物予測のデモ
https://huggingface.co/spaces/sagawa/predictproduct-t5

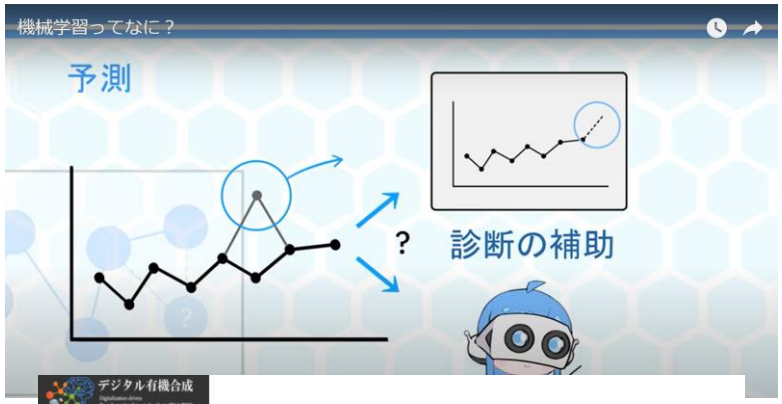
学習済みモデル
https://huggingface.co/sagawa/ReactionT5-yield-prediction
https://huggingface.co/sagawa/ReactionT5-product-prediction

ソースコード
https://github.com/sagawatatsuya/ReactionT5

宣伝：AIの学習

製薬企業・医療従事者向け

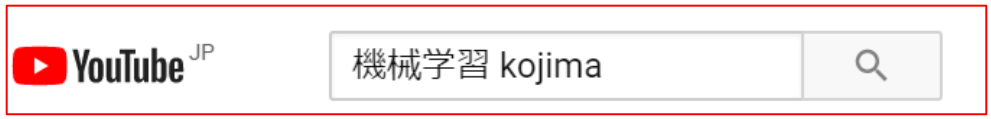
tsucom塾
https://lp.tsucom.ai/contents/kojima-data-analysis



有機化学者向け

学術変革研究A: デジタル化による高度精密有機合成の新展開
https://digi-tos.jp/news/movie1/

ちゃんと学びたいという人向け



https://www.youtube.com/channel/UCo33DtDOIttjiKKqn-YOHSw