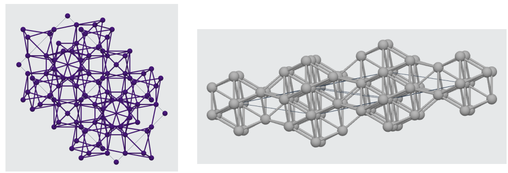
グラフニューラルネットワークと
マルチタスク学習による
汎用的物性予測モデルの構築

JHPCN 2022 課題番号: jh221005

華井雅俊 (東京大学 情報基盤センター)、 鈴村豊太郎 (東京大学 情報基盤センター)
大西 正人 (東京大学工学部)、塩見 淳一郎 (東京大学工学部)

Materials Informatics (MI)

- 材料開発と計算科学の融合分野。 **本研究では特に物性計算を対象**
- **本研究の大きな目的:** 特殊な物性を持つ物質を既存データベースから発見
  - 例えば、金属全般において
    ”電気抵抗が低い => (自由活動する電子が多い) => 熱伝導が良い”
    が知られているが、 ダイアモンドなどは、
    ”電気抵抗が高く (自由電子が皆無) かつ熱伝導が良い”
    ことが知られ、これは電子の代わりにフォノン量子が熱を伝搬するからであり、半導体の熱管理などで需要がある。
    “安価かつダイヤモンドのような物質は無いのか？"などの問いに対し、**既存物質の総当りの物性計算 (screening)** で発見を目指す
  - 各物質の物性値の計算はスパコンなどの大規模計算が必要になることも多く、**総当たり計算は困難**がともなう



分子構造データ*

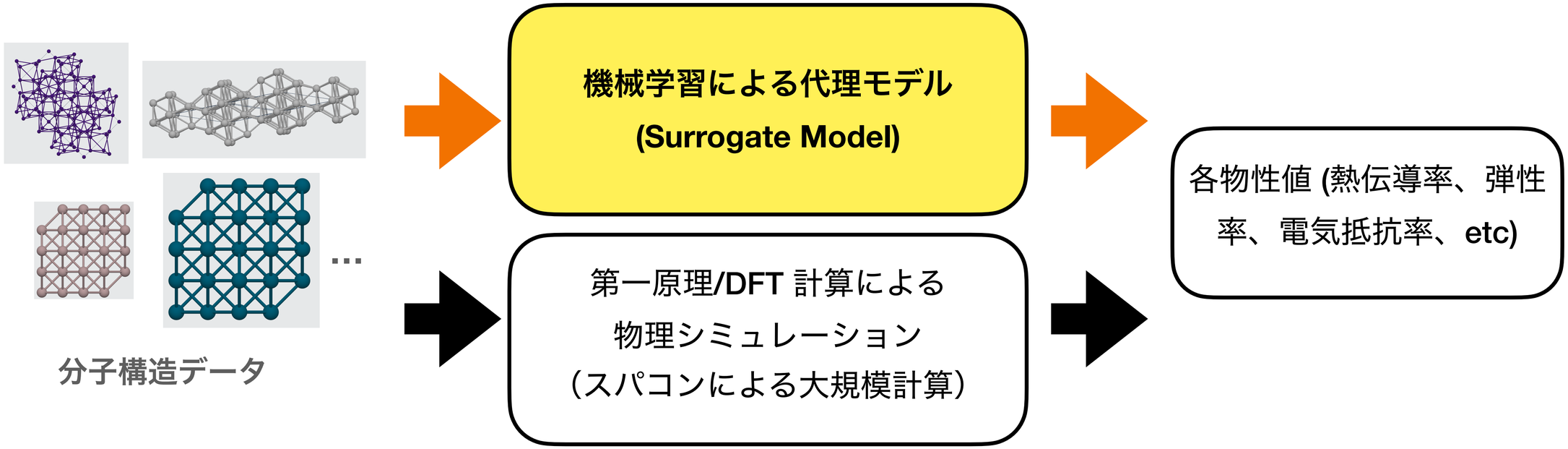
第一原理/DFT 計算による
物理シミュレーション
（スパコンによる大規模計算）

各物性値 (熱伝導率、弾性率、電気抵抗率、etc)

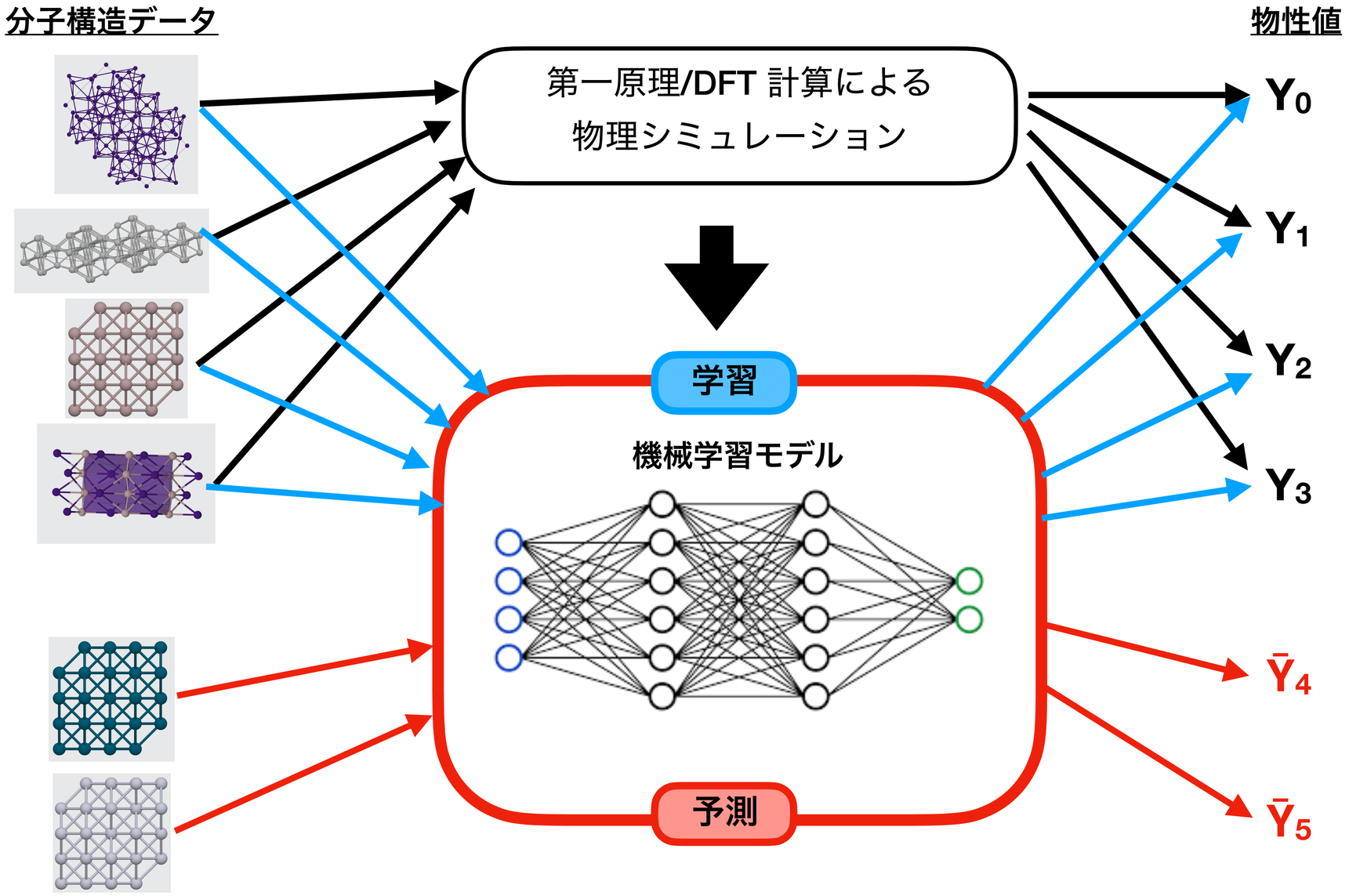
*分子構造図はMaterial Project より取得 https://materialsproject.org

機械学習による物性値予測

- **物性値計算:** 分子構造データからターゲットの物性値をシミュレーションにて計算
  - 既存の分子構造データはこれまでの研究によって豊富かつ網羅的なデータが存在
  - 一方で、網羅的に得られる物性値は少なく、各物性値を分子構造からシミュレーションによって網羅的に計算することは非現実
- **物性値予測: 既知のシミュレーション結果から他の物質での結果を予測**
  - **機械学習による物性値シミュレーション計算の代理モデル (Surrogate Model)**
  - **既知の結果を学習データとし予測モデルを構築**

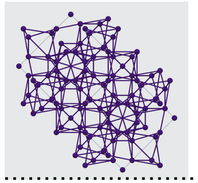
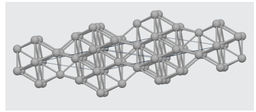
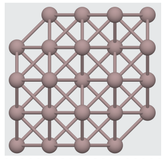
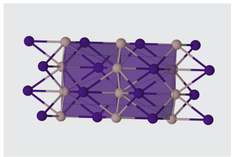
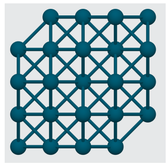
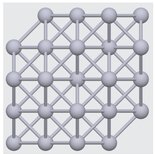


機械学習による物性値予測



研究課題: Multi-task Learning の物性予測への応用

- 予測モデルの構築には大規模なデータが必要だが、**多くの物性値で取得できるデータは小規模かつ限定的**
  - 物性値データセット1つ1つは小規模だが**多種類存在**
- 本研究では**Multi-task Learningを利用し小規模な物性値データを有効的に組み合わせる**方法を探求
  - それぞれの未知な物性値を相互補完的に予測する手法の確立を目指す

	物性値 A	物性値 B	物性値 C	物性値 D
	A_0	?	?	?
	A_1	B_1	?	?
	A_2	B_2	C_2	?
	A_3	B_3	?	?
	?	B_4	C_4	?
	?	?	?	D_5

オープンデータセットの取得

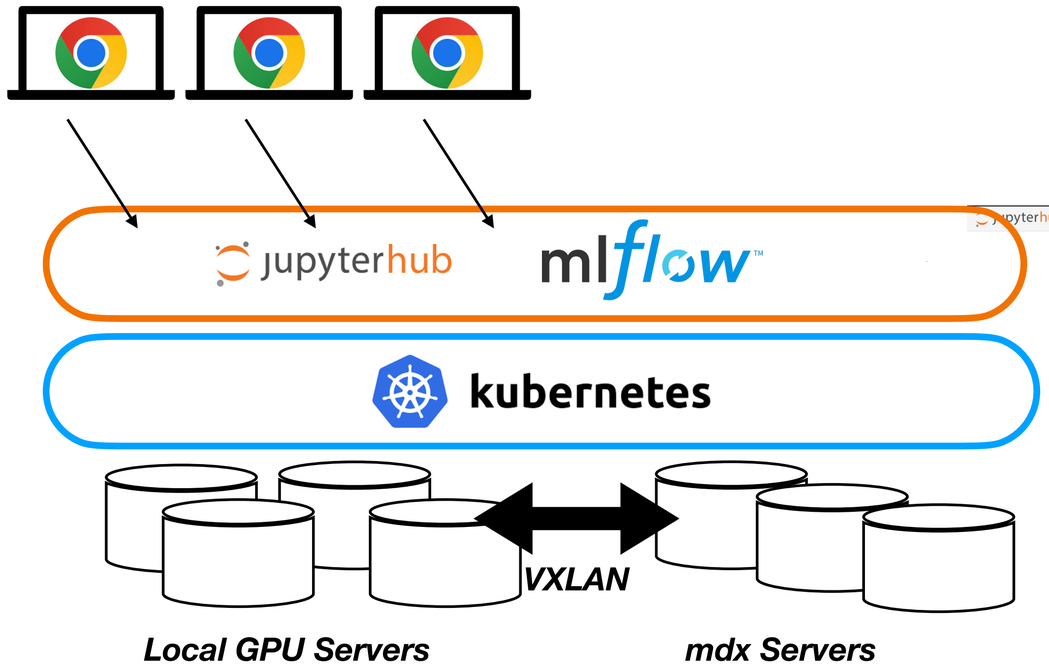
	Materials Project	AFLOW	Open Catalyst Project	COD	PCOD	QMOD
データエントリー数 (物質数)	468K	3.5M	1.3 M	489K	1.0 M	1.0 M
URL	https://materialsproject.org	https://aflowlib.org	https://opencatalystproject.org	http://www.crystallography.net/cod/	http://www.crystallography.net/pcod/	https://oqmd.org
備考	無機物質全般 電池系	無機物質全般	触媒系・無機物質系	固体構造データ	固体構造データ (計算値)	固体・熱力学系データ

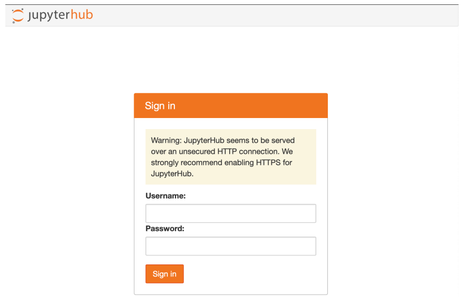
研究タスク

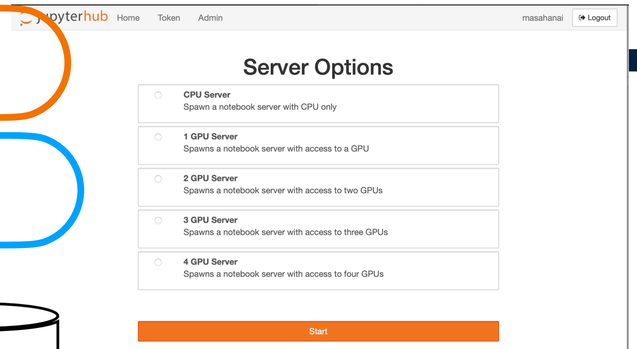
- 異なる物性値データを組み合わせる方法は、Transfer Learningの利用によって多く発表されているが、**既存の研究では最大で3物性間**にとどまっていることがわかった。
  - ランダムに組み合わせた際、ほとんどの場合で Negative Transfer による精度劣化がみられることを確認。
  - 組み合わせ方の候補は膨大となるが、効果的な組み合わせ方を発見する方法は確立されていない。
- 一方で、画像認識・NLPなど機械学習全般分野においてMulti-task Learningは以下の進展が見られることがわかった。
  - **アーキテクチャ・組み合わせ方法の自動取得手法 (Neural Architecture Search NAS)**
    - ***“Adashare: Learning What to Share for Efficient Deep Multi-Task Learning” (NeurIPS 2020)***
  - **生成モデルによる小規模データ群のMulti-Task化**
    - ***“Variational Multi-Task learning with Gumbel-Softmax Pariors” (NeurIPS 2021)***
  - **Transformer によるMulti-Task Learningのタスク数の大規模化・高精度化**
    - ***“UniT: Multimodel Multitask learning with a Unified transformer” (ICCV 2021)***
- 本研究では以下の3つのタスクに注力する
  - **ニューラルネットワークの構築方法として、候補アーキテクチャの膨大化に対処すべく自動化手法を構築**
    => Q1ではシステムの整備と予備実験によるシステム確認までを行なった
  - **生成モデルを利用し、小規模なデータから最大限の情報を抽出する方法の追求**
  - **最新の高精度モデル (Transformer) をベースとした精度追求**

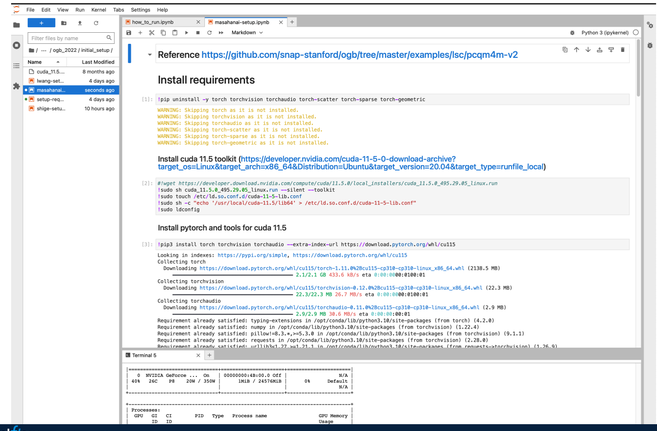
mdx 上での機械学習モデル構築

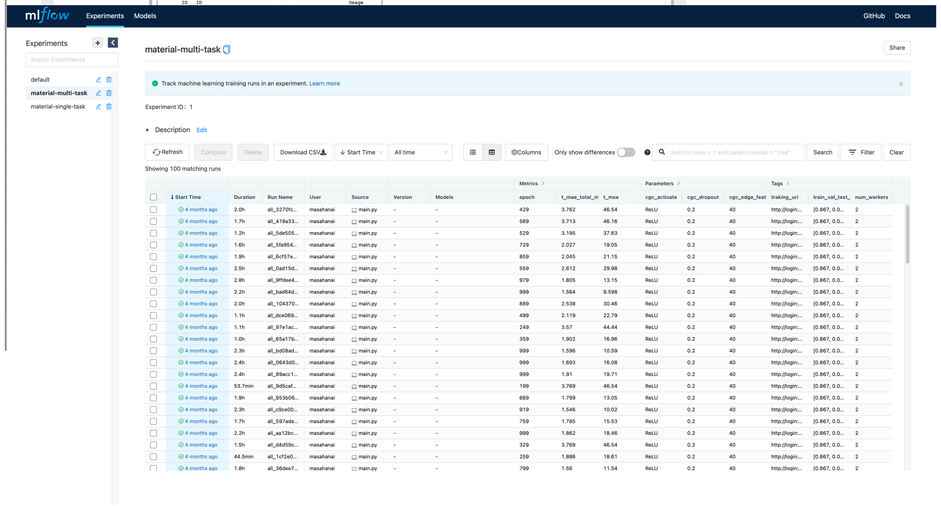
- ローカルGPUサーバー上のKubernetes クラスタをmdx仮想ノード群へ拡張しリソースの一括管理
- JupyterHub (https://jupyter.org/hub) によるブラウザベースでの開発環境、実行環境・ユーザー管理、ファイル共有
- mlflow (https://mlflow.org) による実験結果のデータベース化。ハイパーパラメータ実験の効率化。











予備実験 (物性予測 vs 画像認識・NLPのMulti-task Learningの違い)

- 画像認識やNLPのMulti-task learningではベースとなるSingle-taskモデルを拡張し出力部分を複数分岐させることでMulti-taskモデルを構築する。
- 予備実験では、そのようなベースとなるSingle-taskモデルを作るために、さまざまな物性値に対して、その最適なモデルサイズ (パラメータ数 + レイア数) を計算
  - 代表的なGNNモデル (CGCNN) を利用し、以下をGrid searchにてハイパーパラメータサーチを実行。合計 4 * 5 * 2 * 4 = 160 パターンの学習を実行
    1. GNNレイア数 (Message passingの回数) [1, 2, 4, 8]
    2. Decoder MLPレイア数 / Hidden Unit 数 (GNN後から OutputまでのLayer) [1,2,3,4,5] / [64, 128]
- **結果: ベースとなるような典型的なモデルサイズはなく、物性値によって基本的にバラバラ**
  - **モデルの構築自体が複雑であり、その自動化が必要であることがわかった。**

物性	Best GNN Size	Best Decoder Size
Band gap	4 Layers	1 Layer / 128 Hidden Unit
Elasticity	2 Layers	4 Layers / 64 Hidden Unit
Dielectric Properties	1 Layer	1 Layer / 64 Hidden Unit
magnetization	4 Layers	5 Layers / 64 Hidden Unit

まとめ

- **Q1にて以下に着手**
  - オープンデータセットの大規模クローリング
  - 関連研究の調査により具体的なタスクを設計
  - ハイパーパラメータサーチ・アーキテクチャサーチのための機械学習基盤環境をローカルサーバー + mdx 上に構築
    - 予備実験として単一物性でのハイパーパラメータサーチを実施
- **今後の予定**
  - ニューラルネットワークアーキテクチャの自動構築手法の実装とその評価 (Q2)
  - 生成モデルによる、小規模データのマルチタスク手法の設計・実装・評価 (Q3)
  - 最新モデルによる精度追求 (Q4)