

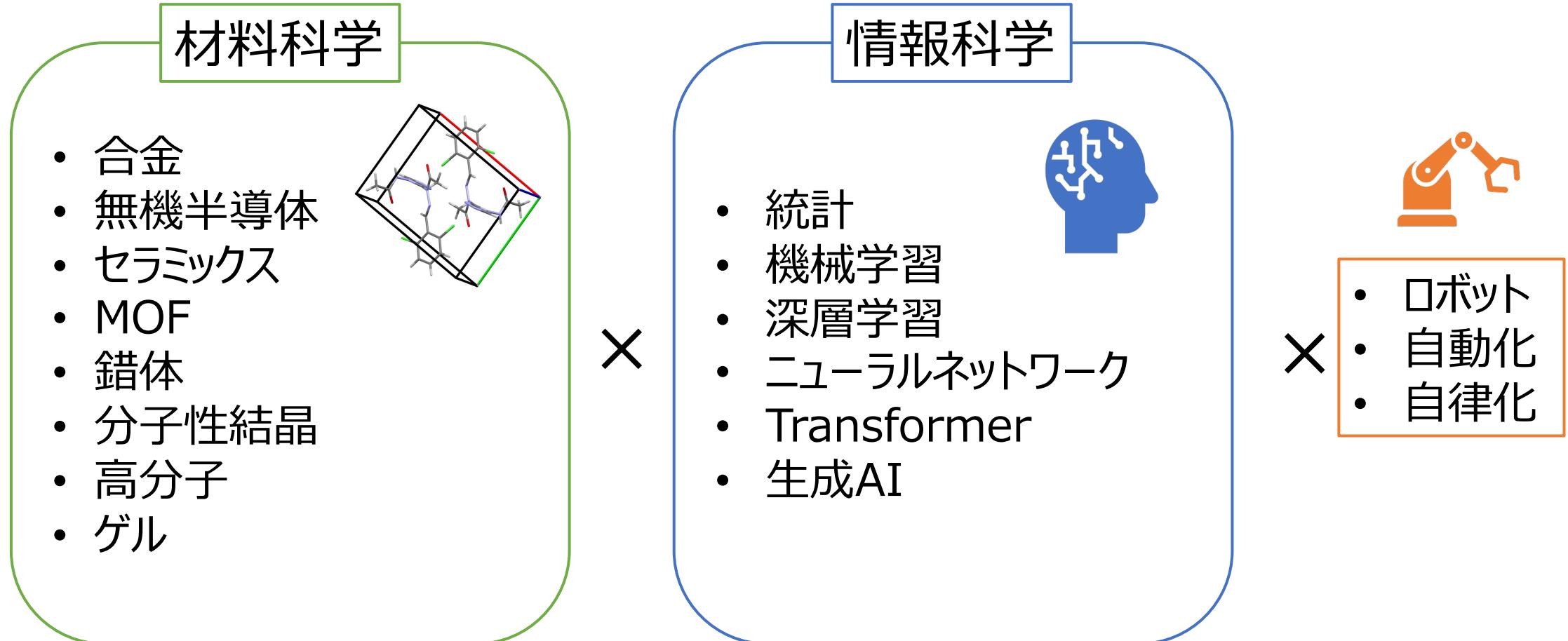
最近の研究内容

～有機固体のマテリアルズインフォマティクス～

谷口卓也

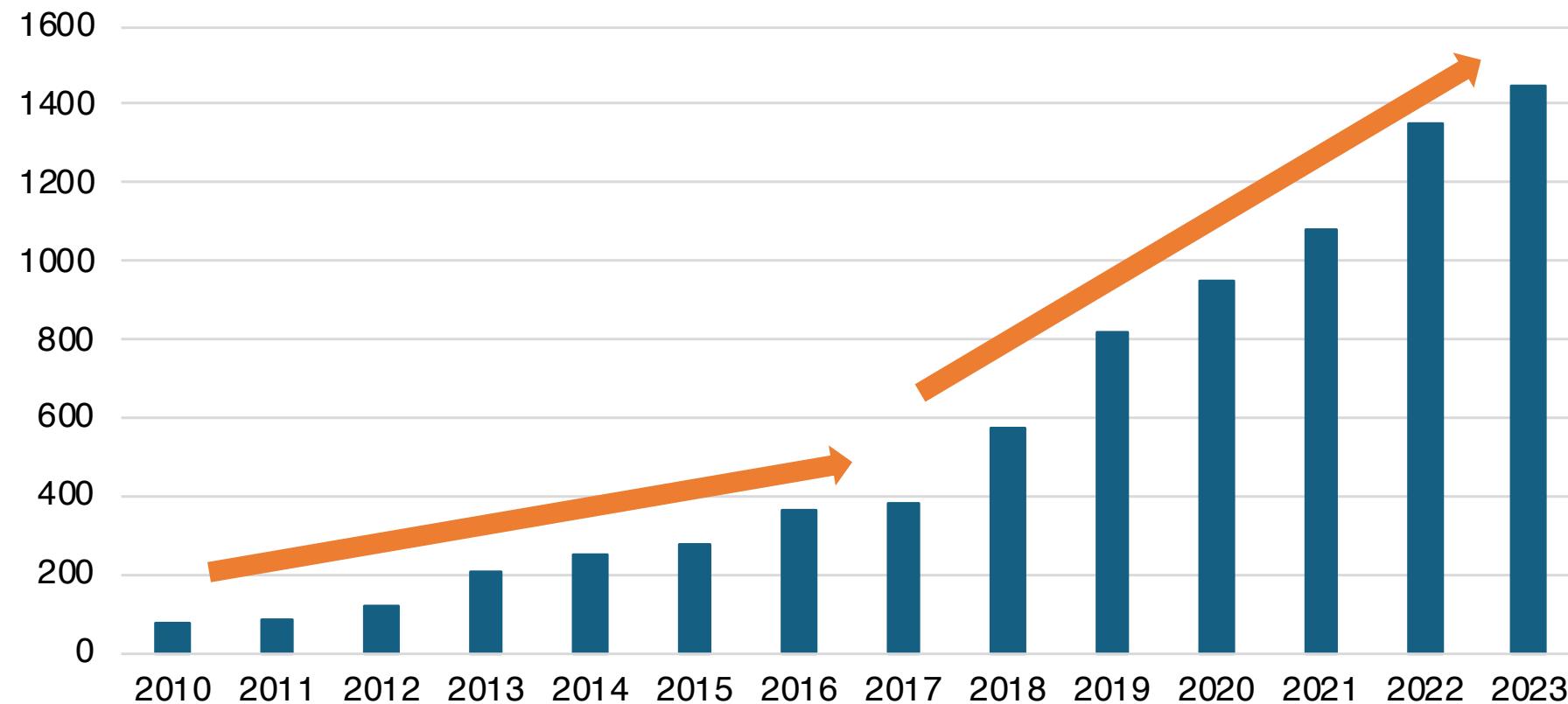
早稲田大学データ科学センター

マテリアルズ・インフォマティクス



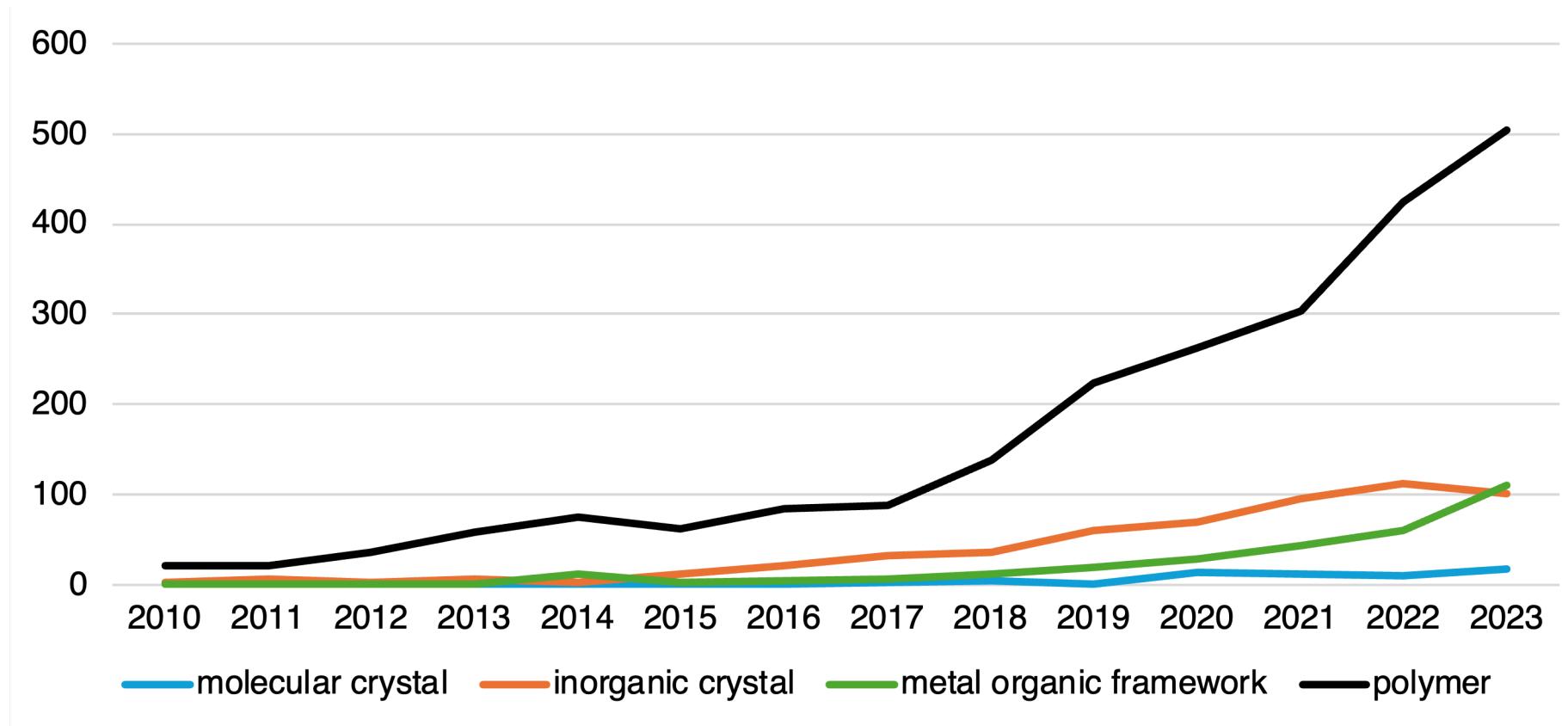
マテリアルズ・インフォマティクス

“Materials Informatics”の件数 (Google Scholar)



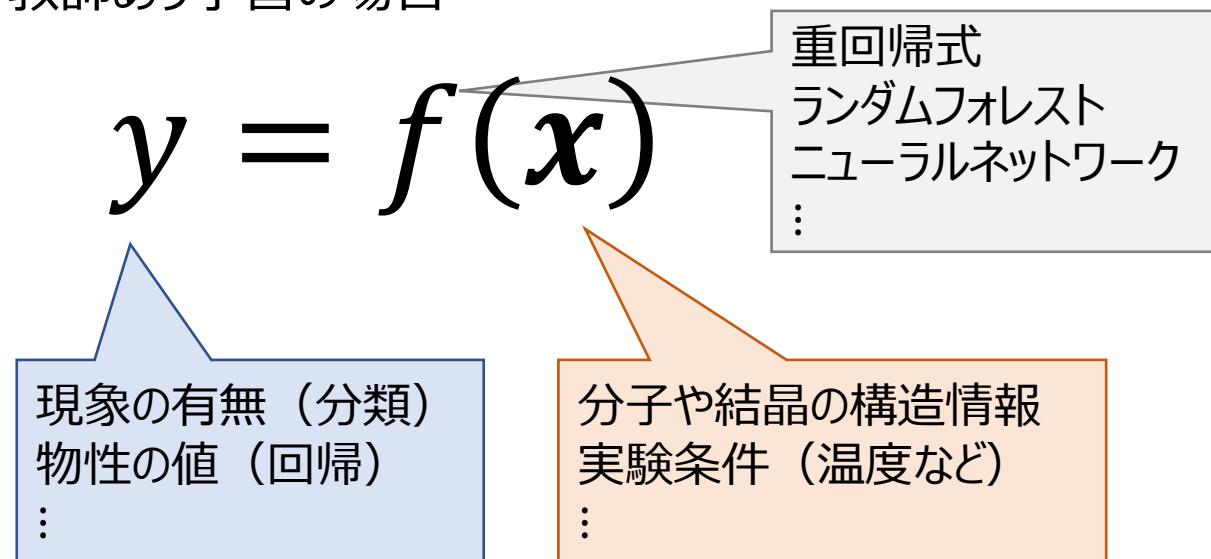
マテリアルズ・インフォマティクス

“Materials Informatics”&“○○”の件数 (Google Scholar)



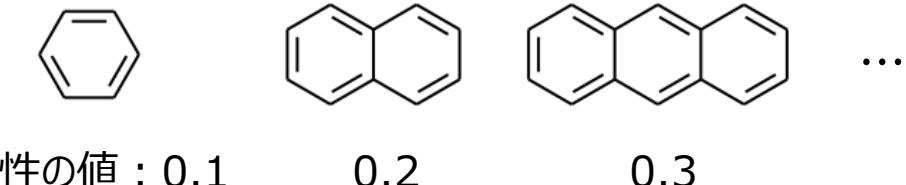
マテリアルズ・インフォマティクス

教師あり学習の場合

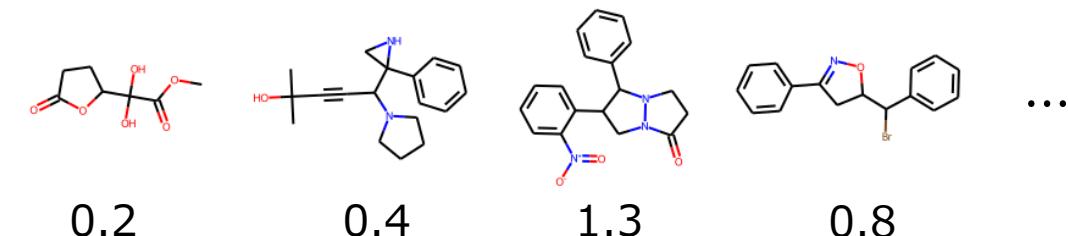


- データがどれくらい似ているか (データの近さ) に基づき関係式をつくり、実験に活用する
- データをどう表現するか(x)、関係式の作り方(f)、データの数や質も重要になる

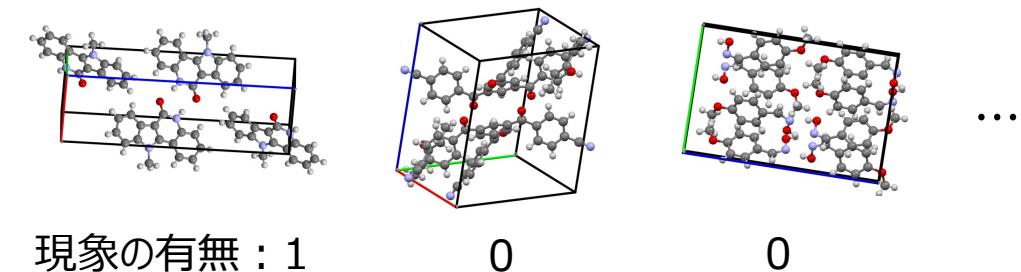
(例1) 似た分子間の比較



(例2) 似ていない分子間の比較



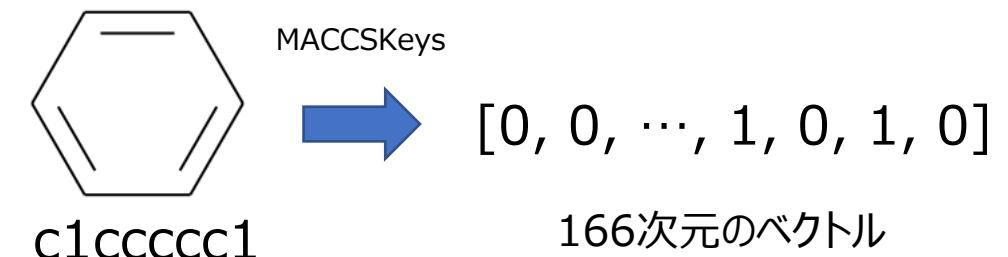
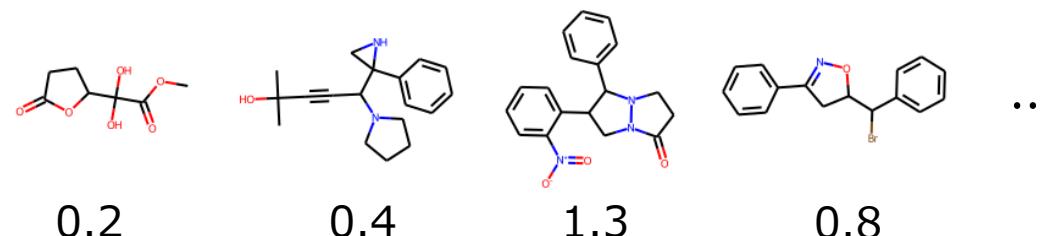
(例3) 似ていない結晶間の比較



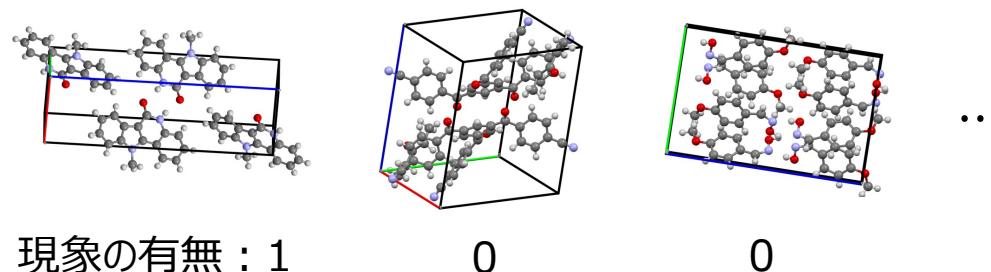
データの表現

ベクトルに変換してデータ間の似ている・似ていない（＝データの近さ）を評価したい

- 似ていない分子間の比較



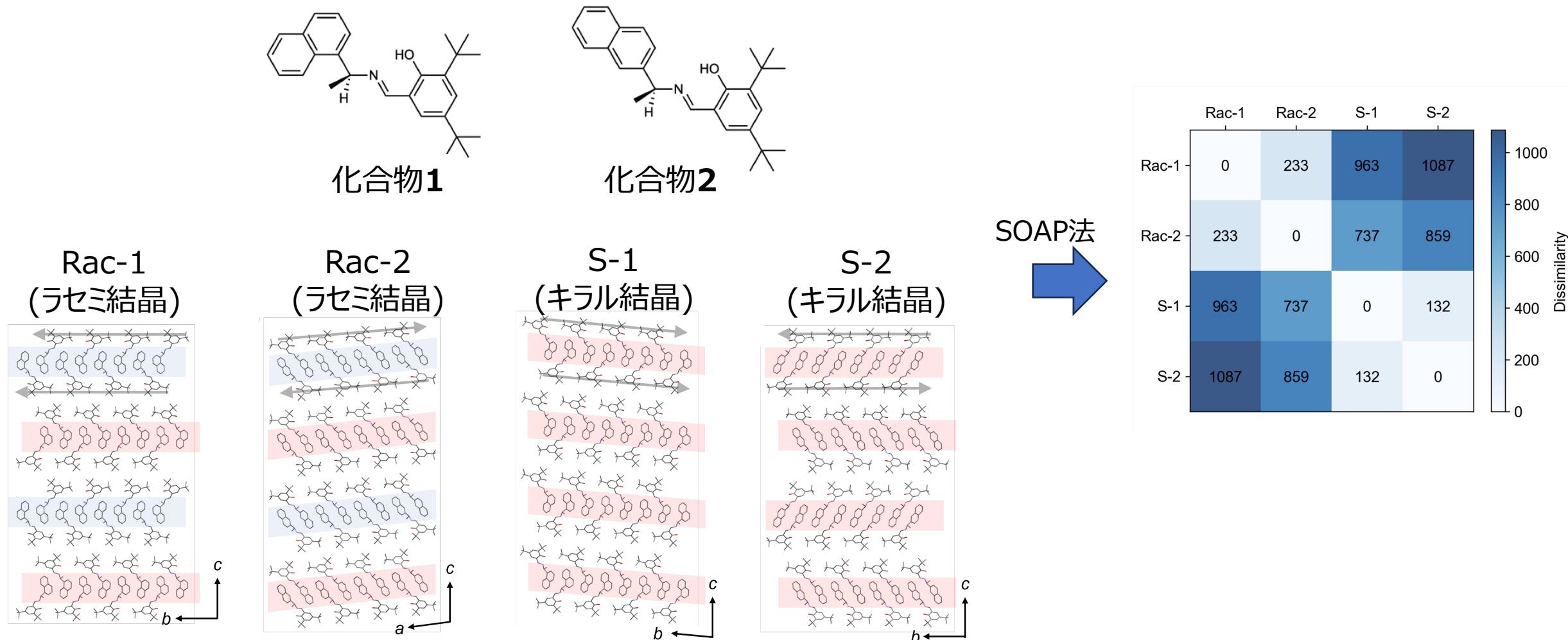
- 似ていない結晶間の比較



※このようなベクトルをデータから作るのがニューラルネットワーク

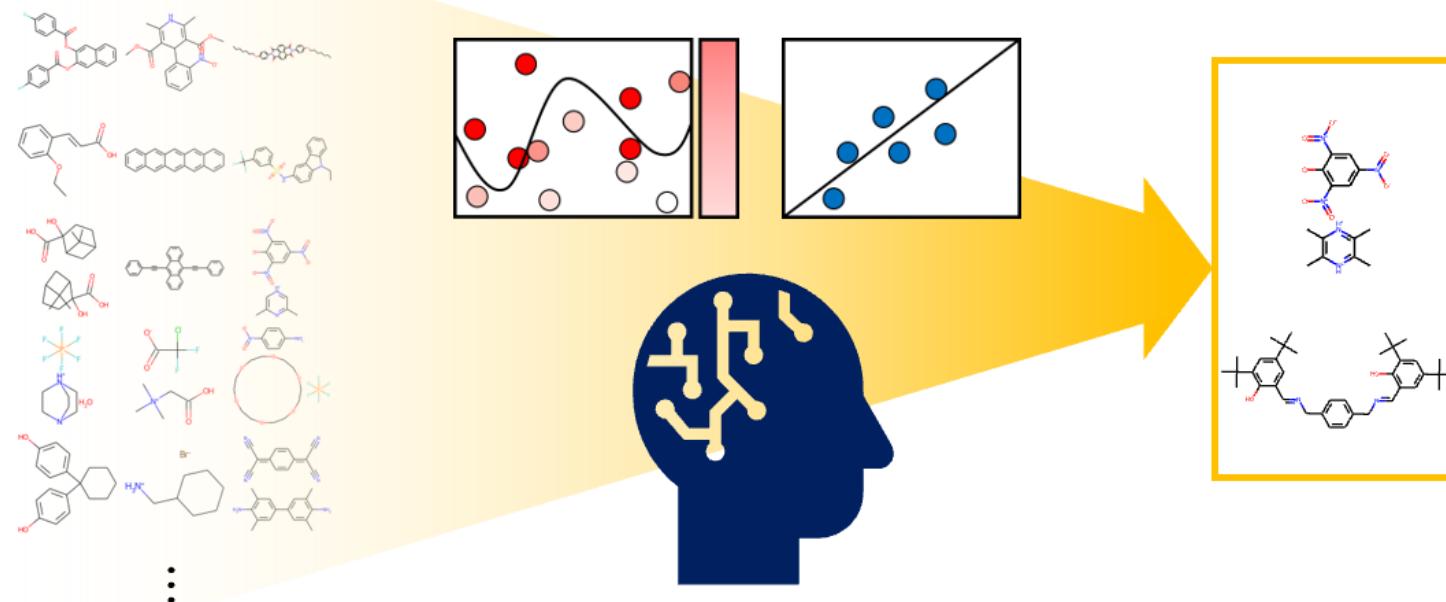
データの表現

ベクトルに変換してデータ間の似ている・似ていない（＝データの近さ）を評価したい



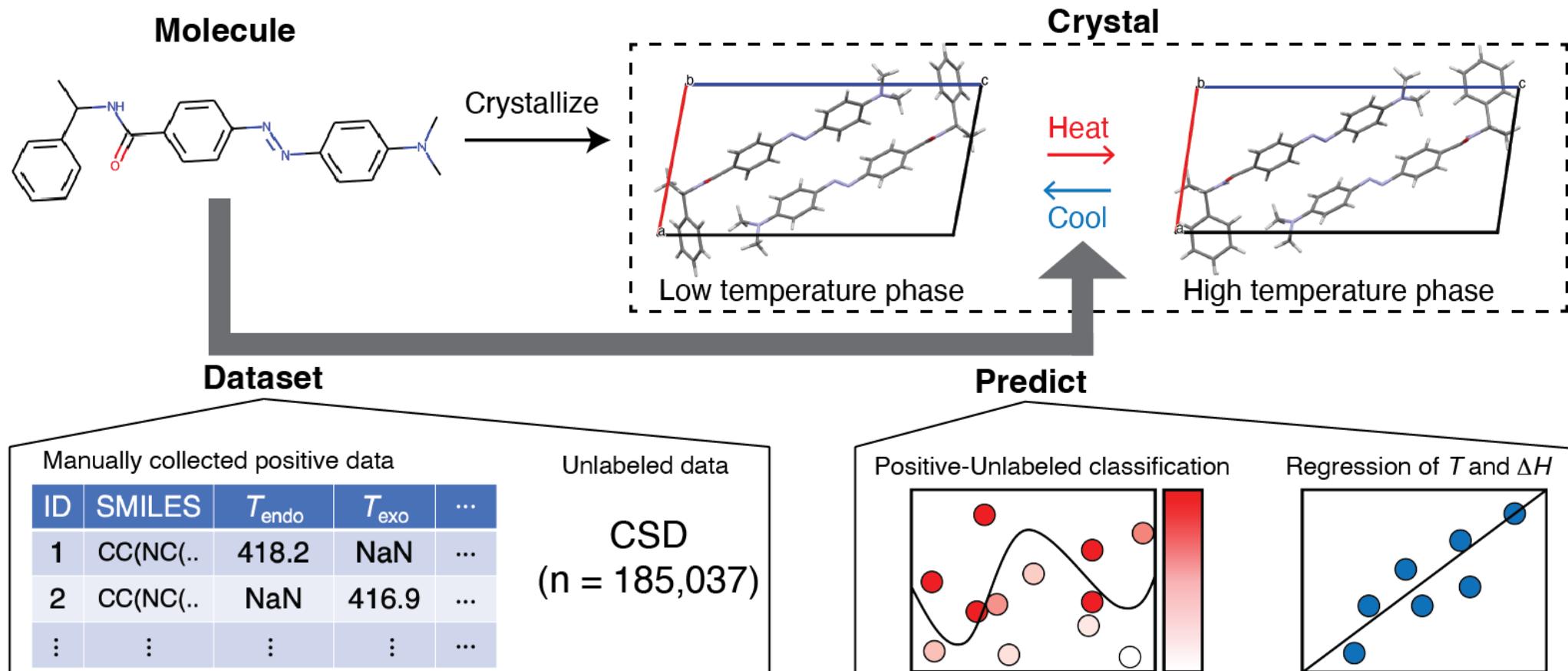
分子記述子による構造相転移スクリーニング

Molecular Screening by Machine Learning



Digital Discovery, 2023, 2, 1126-1133.

構造相転移



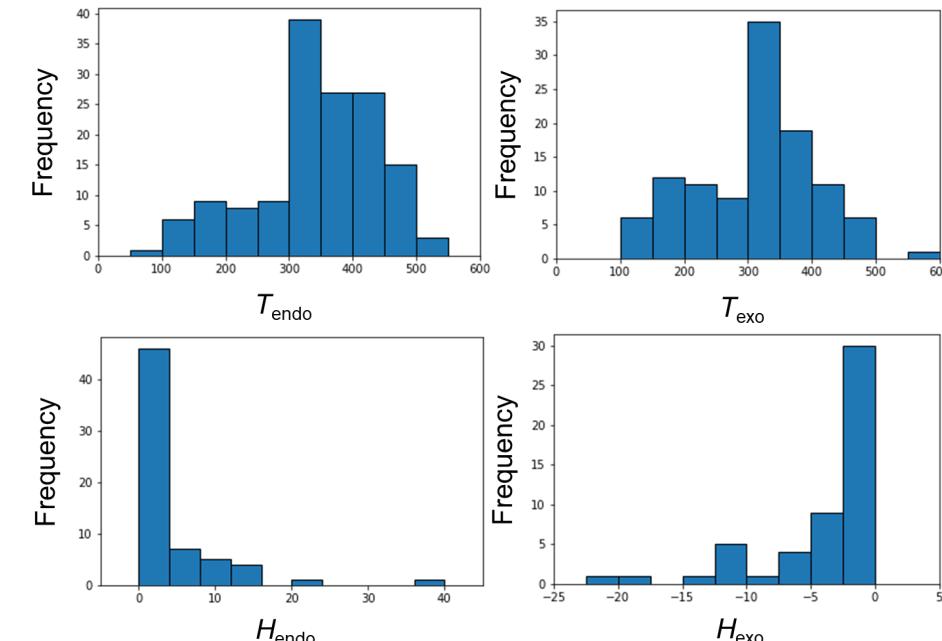
メリット：注目すべき分子の情報や分子デザインの知見が得られる

デメリット：結晶構造情報は含めていない

固相転移データセットの構築

Positiveデータ (相転移が起きることを確認済みのデータ)

- 学術論文から分子構造、転移温度、エンタルピーを収集
- 温度変化で相転移するデータのみ収集
- 全行程の中で最も時間のかかった工程
- ユニークな分子構造の数 : $n = 88$



Unlabelデータ (相転移が起きるか不明なデータ)

- ケンブリッジ結晶構造データベースから相転移の報告のない分子構造を抽出
- ユニークな分子構造の数 : $n \approx 180,000$

やりたいこと

適切な分類器を構築し、UnlabelデータからPositiveになりそうなデータを見つけたい

評価指標 (TPR×SE) の値

データの表現 x

関係式 f

	RF	NN	SVM	GBDT
Mordred	9.3	0.0	0.3	1.0
ECFP	18.9	299.7	4856.9	1.3
Avalon	25.7	415.5	11492.5	32.0
ErG	19.0	79.0	3408.2	33.8
RDKitDesc	49.7	107.2	NaN	0.7
MACCSKeys	9.8	71.1	2667.4	5.2
Estate	11.5	5.9	0.0	15.8

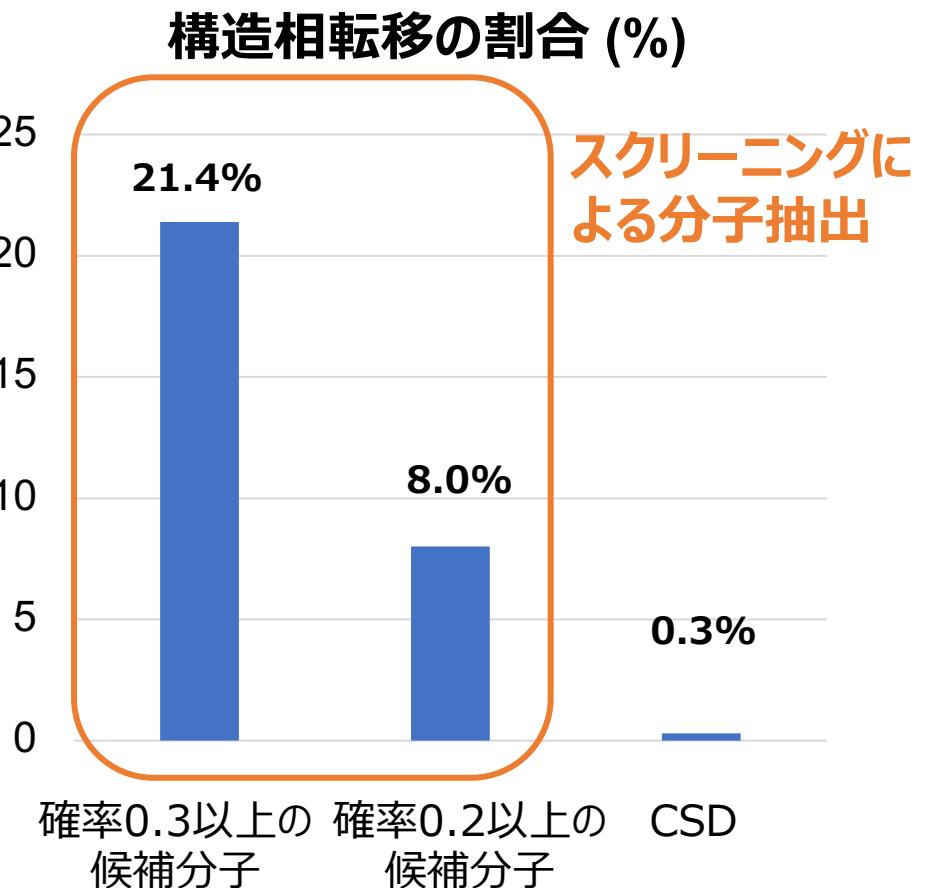
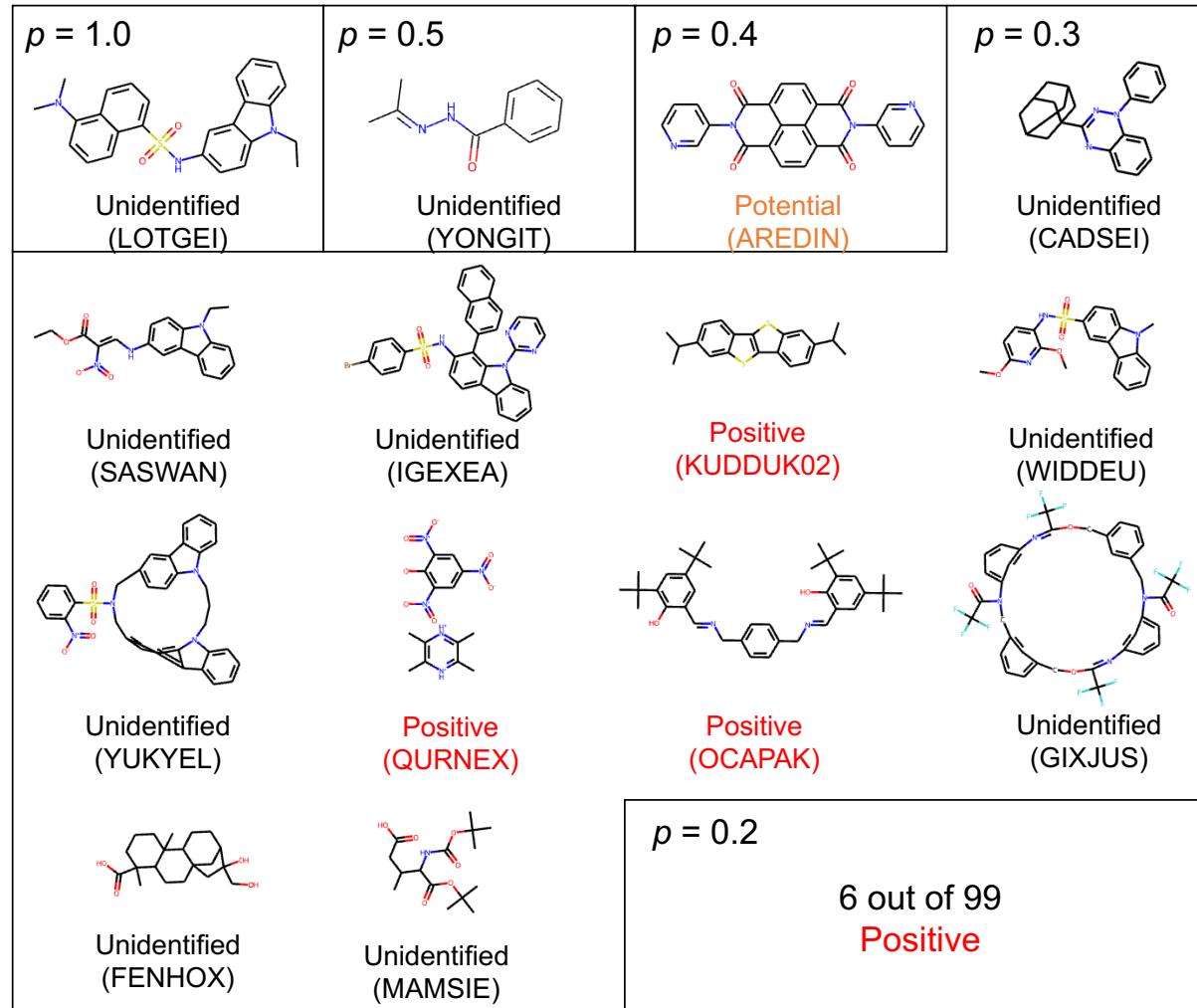
TPR: True Positive Rate
SE: Selection Effect

分類器の妥当性
提案器としての能力

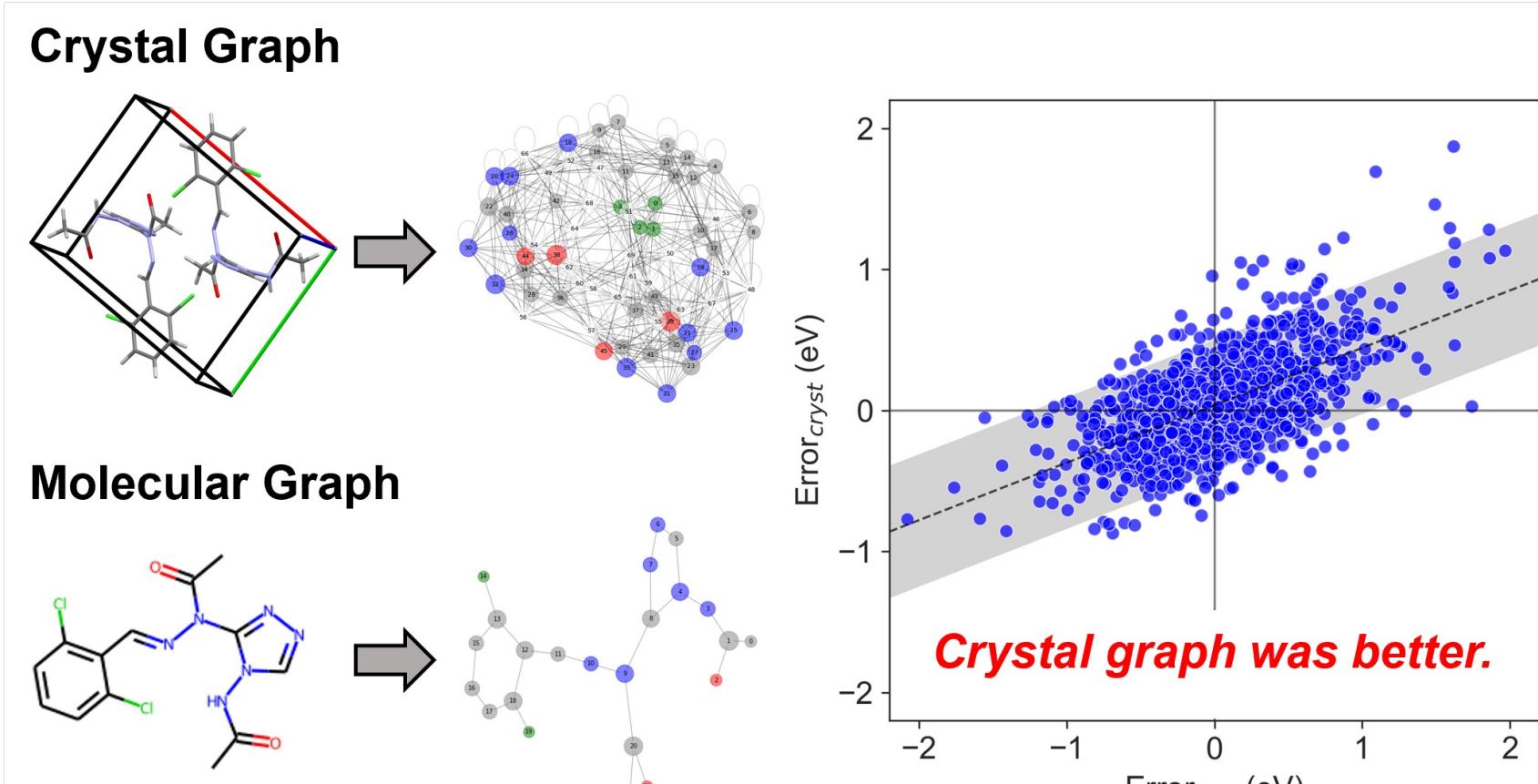
$$\text{TPR} \times \text{SE} = \frac{n_{pp}}{n_p} \times \frac{n_u}{n_{up}} = \frac{n_u}{n_p} \times \frac{n_{pp}}{n_{up}} = k \times \frac{n_{pp}}{n_{up}}$$

n_p : positiveデータの数
 n_u : unlabeledデータの数
 n_{pp} : positiveデータのうち、正しくpositiveと予測されたデータ数
 n_{up} : unlabeledデータのうち、positiveと予測されたデータ数

スクリーニング結果

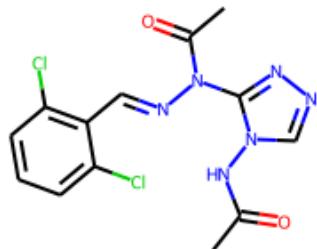


グラフニューラルネットワークによるバンドギャップ予測

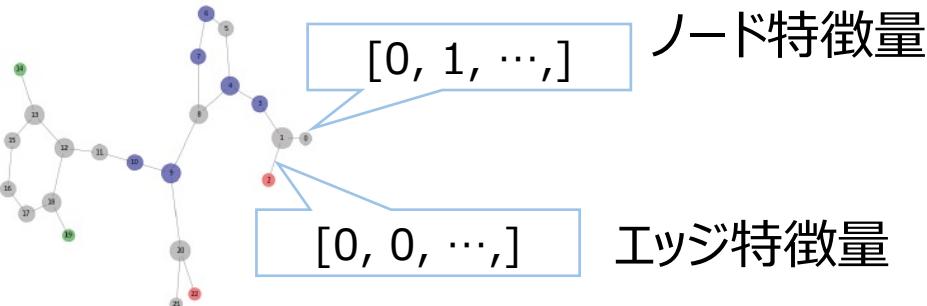


グラフデータ

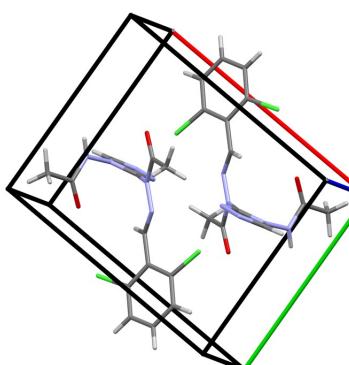
分子 → グラフデータ



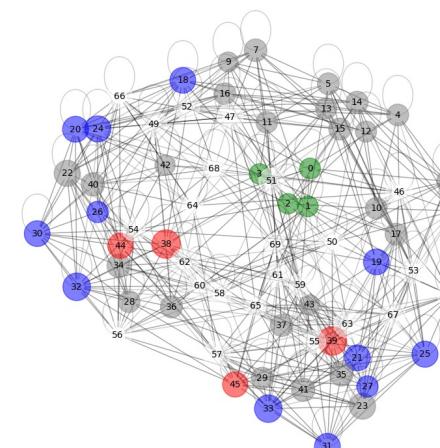
原子をノード
結合をエッジ



結晶 → グラフデータ



原子をノード
距離でエッジ



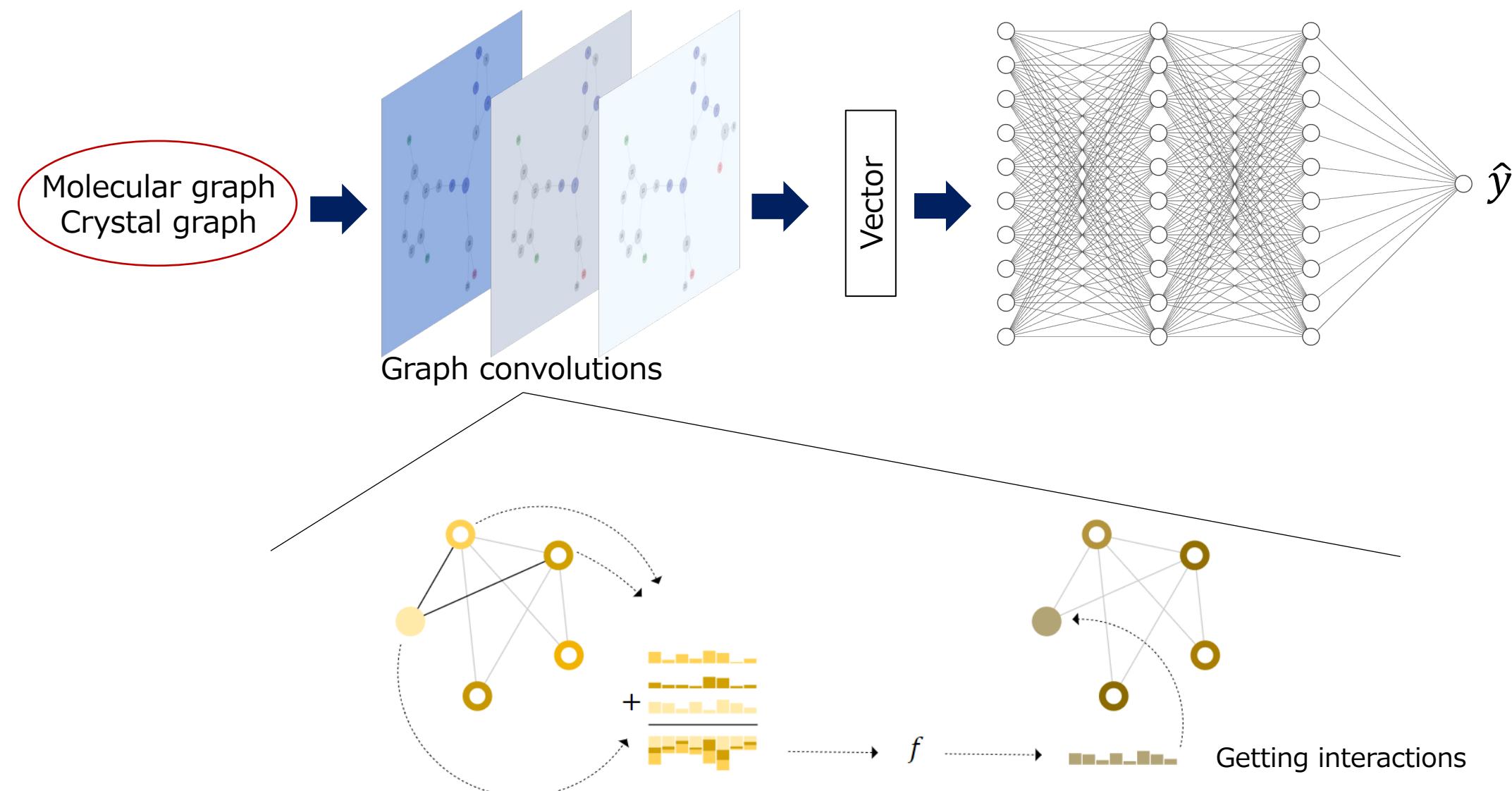
※様々なグラフデータ

- 文献の引用関係
- 実験フローチャート
- 人間関係
- 画像
- レコメンデーション

⋮

グラフニューラルネットワーク

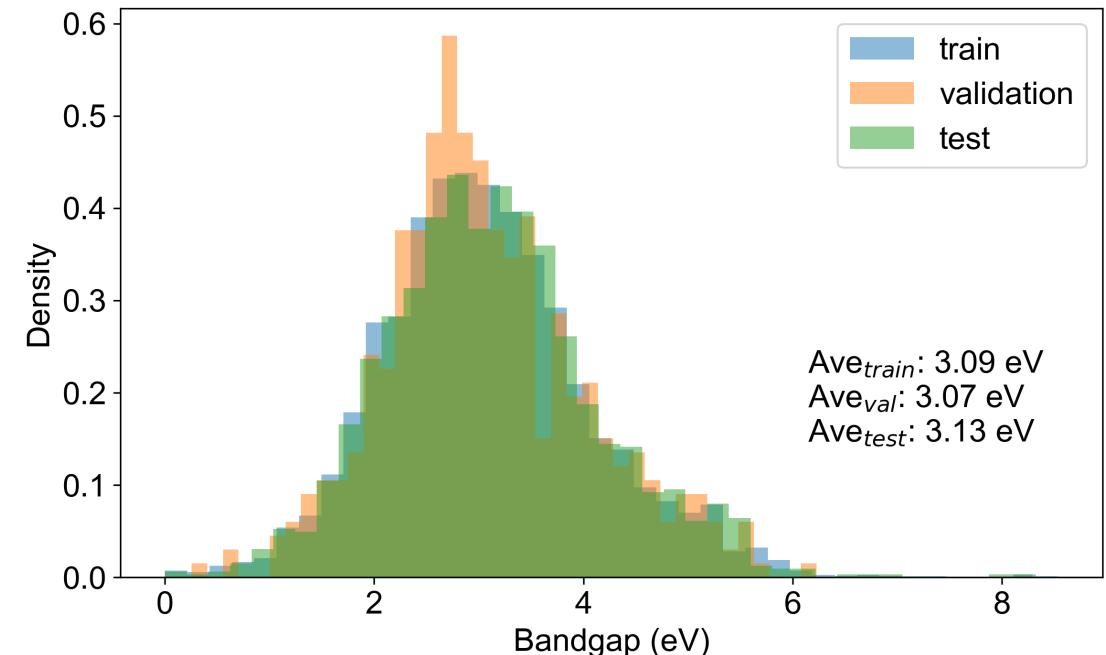
GNN



バンドギャップのデータセット

バンドギャップ (y)

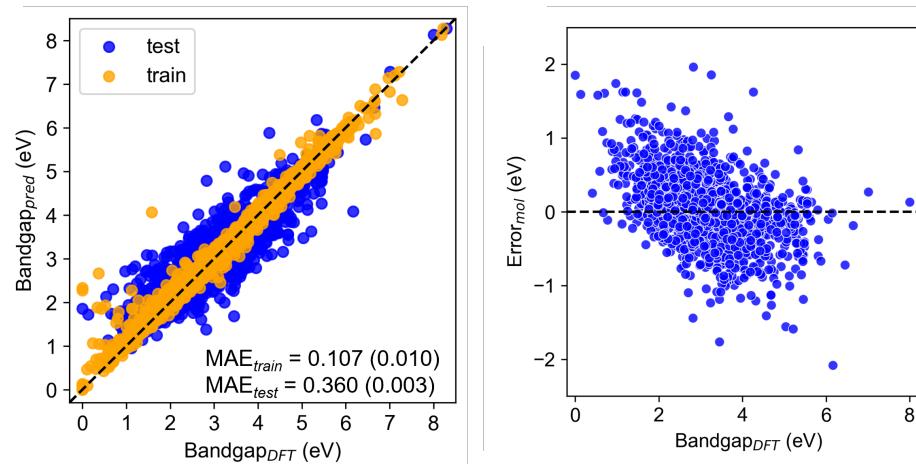
- 半導体特性、光吸収などに関わる固体の物性
- 有機太陽電池材料、発光材料などで重要な
なる
- Organic Materials Database (OMDB)
からデータセットを取得



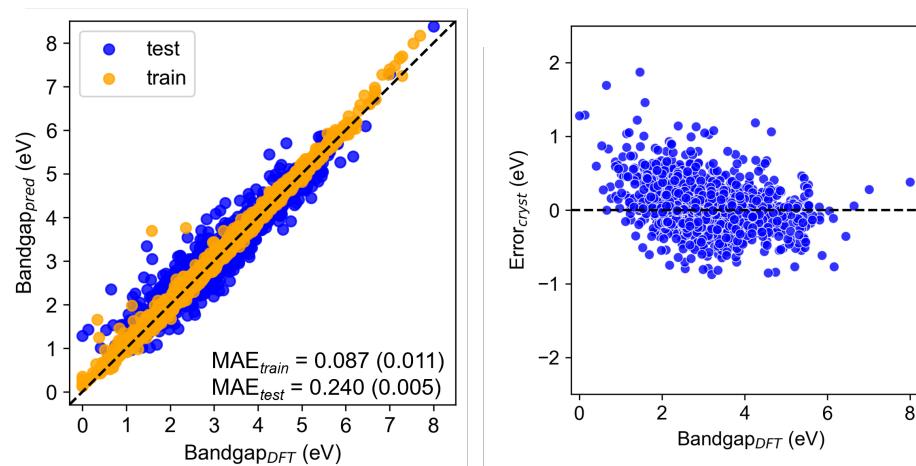
$N_{\text{total}} = 10472$
Train: Val: Test = 0.8: 0.05: 0.15

予測誤差の評価

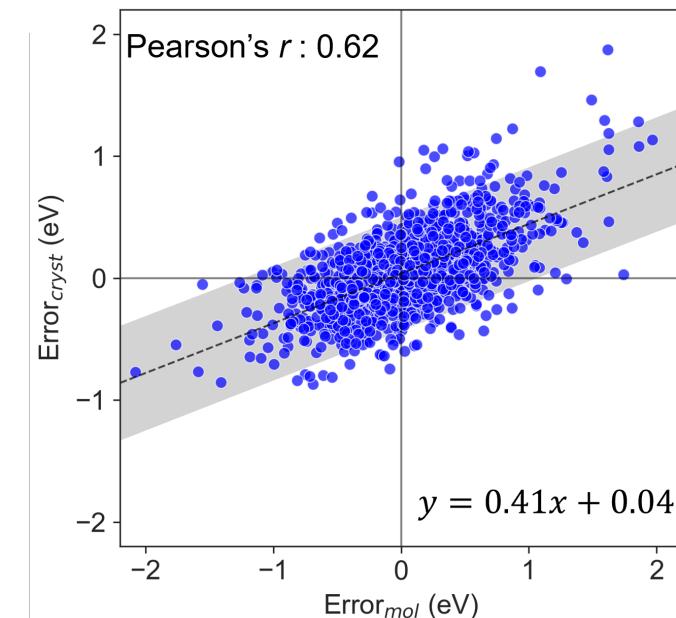
分子グラフ



結晶グラフ

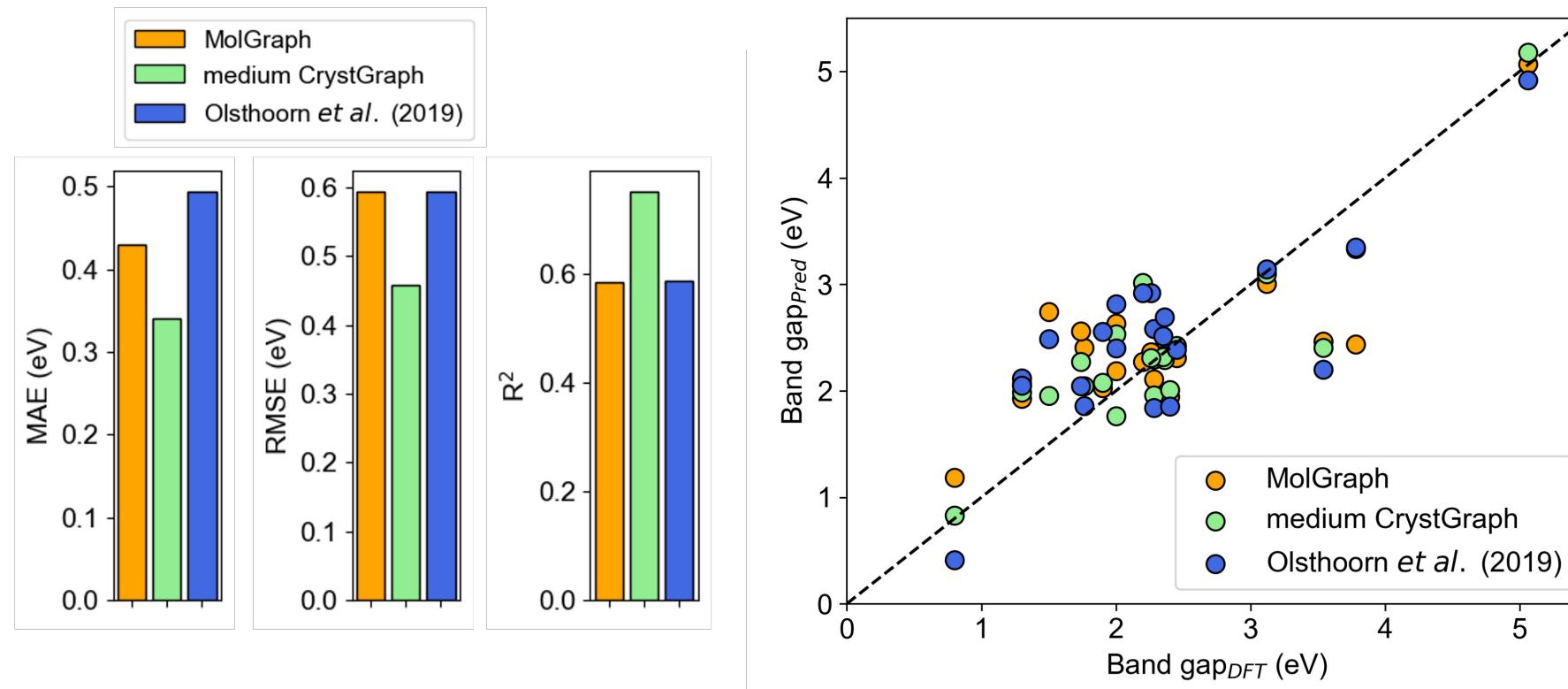


結晶グラフの誤差は分子グラフの約0.4倍

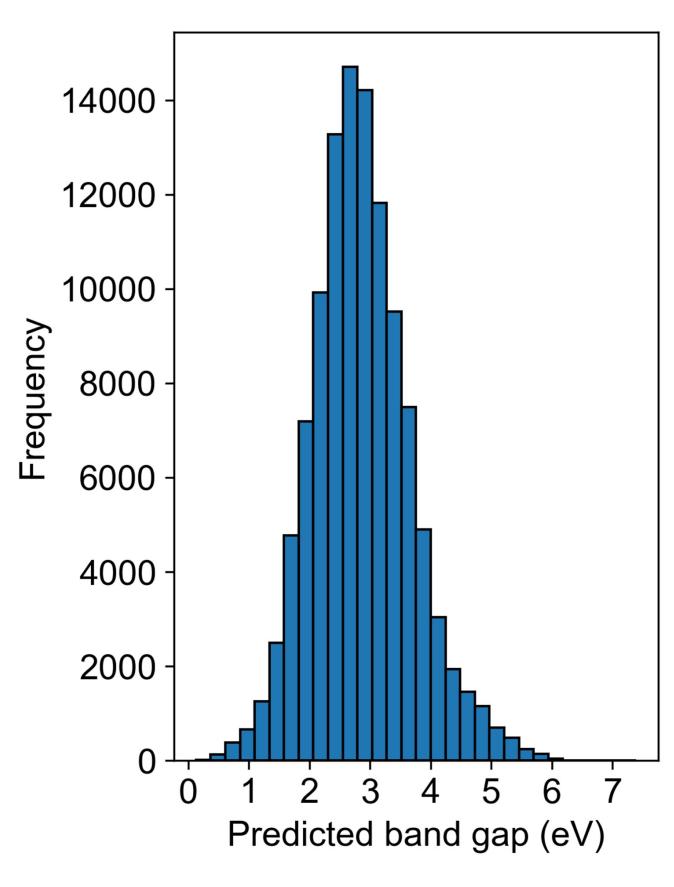


予測誤差の評価

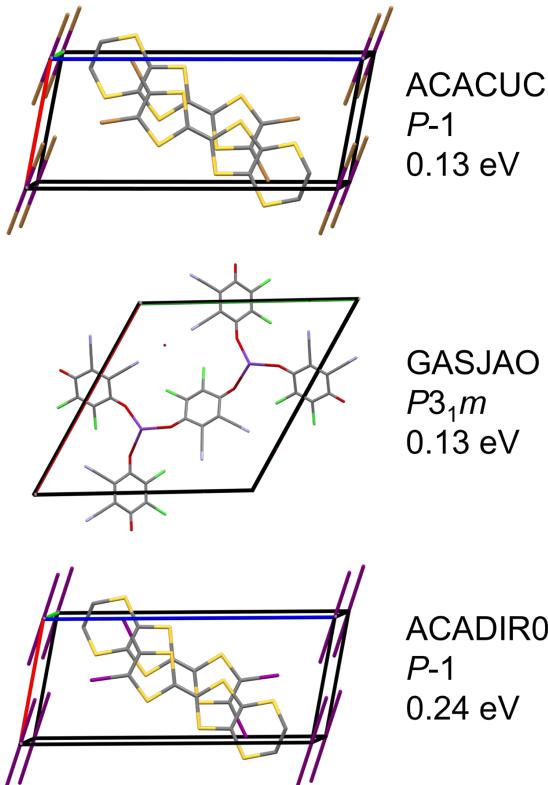
今回の結晶グラフモデルが最も良い予測誤差だった



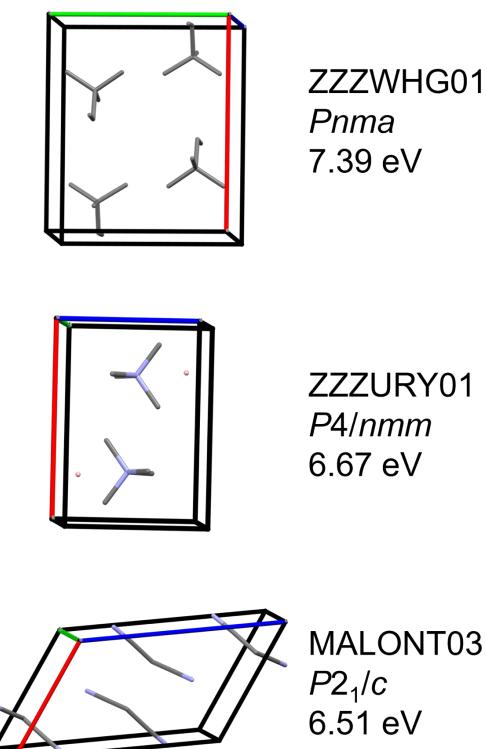
スクリーニング



Smallest predictions

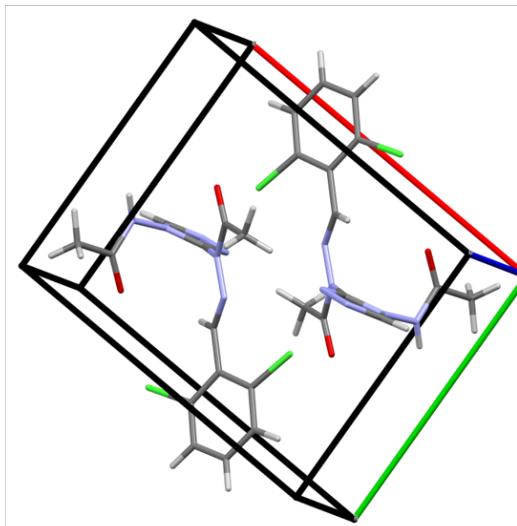


Largest predictions

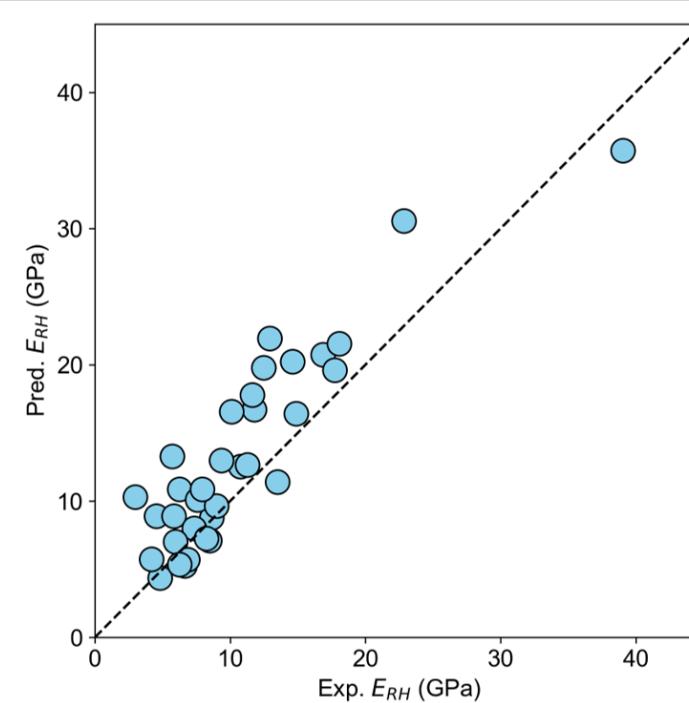


機械学習ポテンシャルによる弾性率スクリーニング

Molecular Crystals

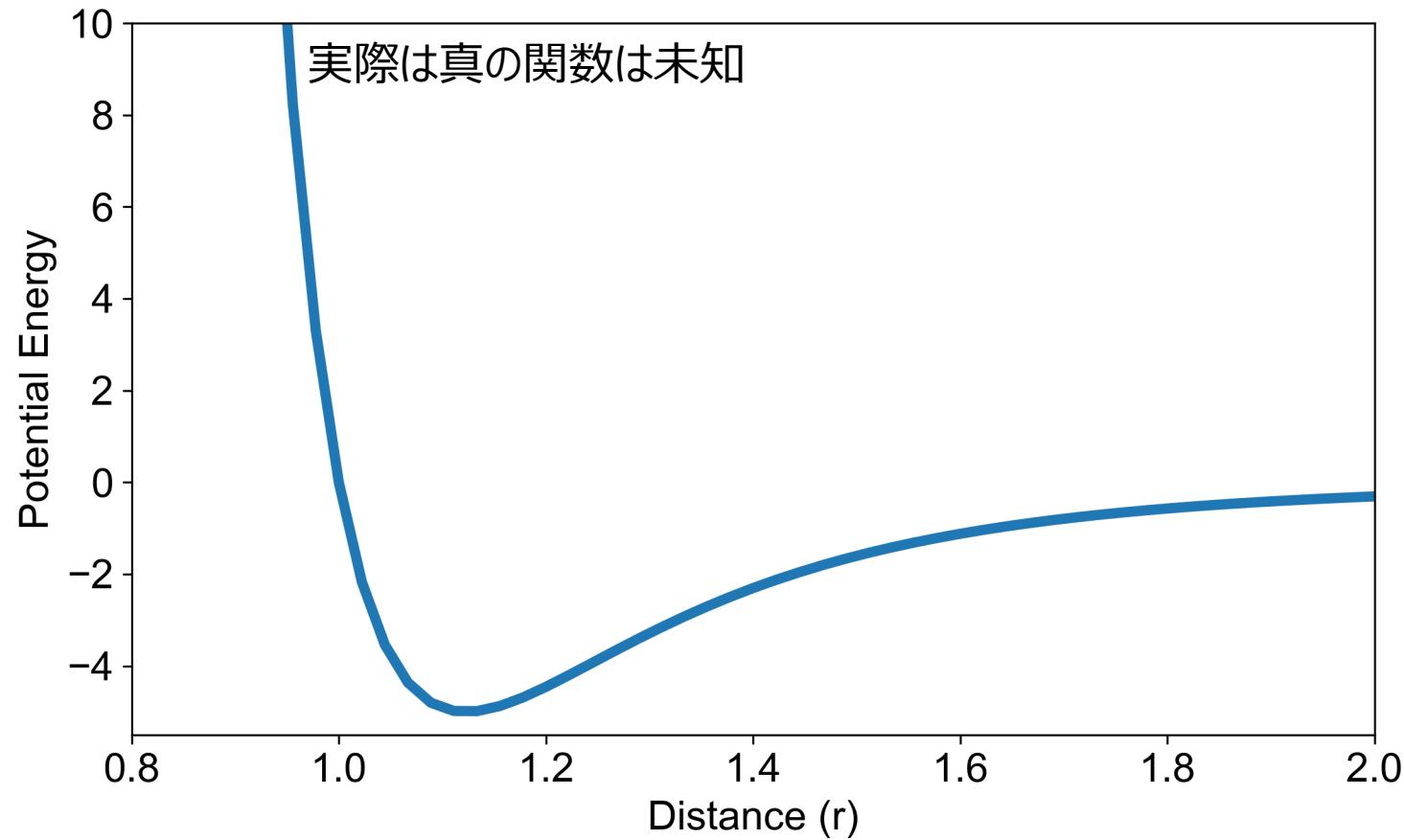


Elastic Modulus

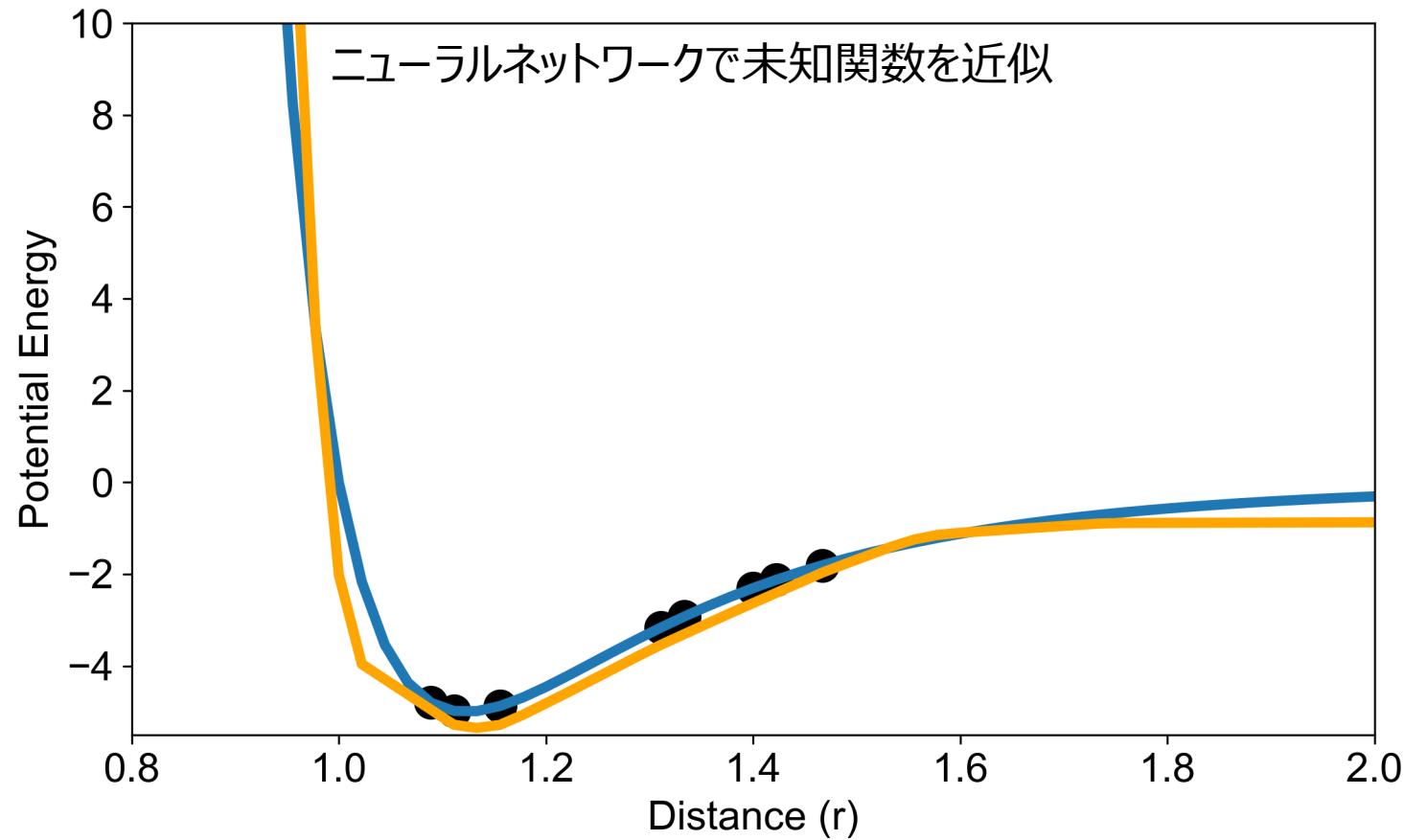


CrystEngComm, 2024, 26, 631-638.

機械学習ポテンシャル



機械学習ポテンシャル



弾性定数テンソル

$$C_{ij} = \frac{1}{V} \left(\frac{\partial^2 U}{\partial \epsilon_i \partial \epsilon_j} \right)$$

Voigtの表記法 (最大21個の独立成分)

$$\begin{pmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} \\ C_{21} & C_{22} & C_{23} & C_{24} & C_{25} & C_{26} \\ C_{31} & C_{32} & C_{33} & C_{34} & C_{35} & C_{36} \\ & & & C_{44} & C_{45} & C_{46} \\ & & & & C_{55} & C_{56} \\ & & & & & C_{66} \end{pmatrix}$$



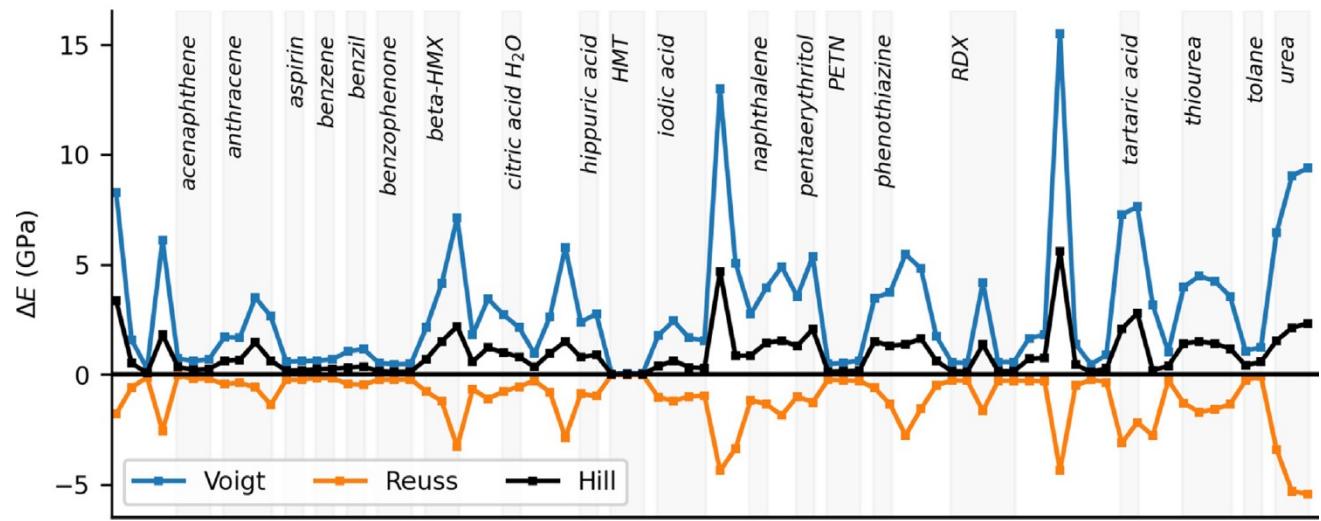
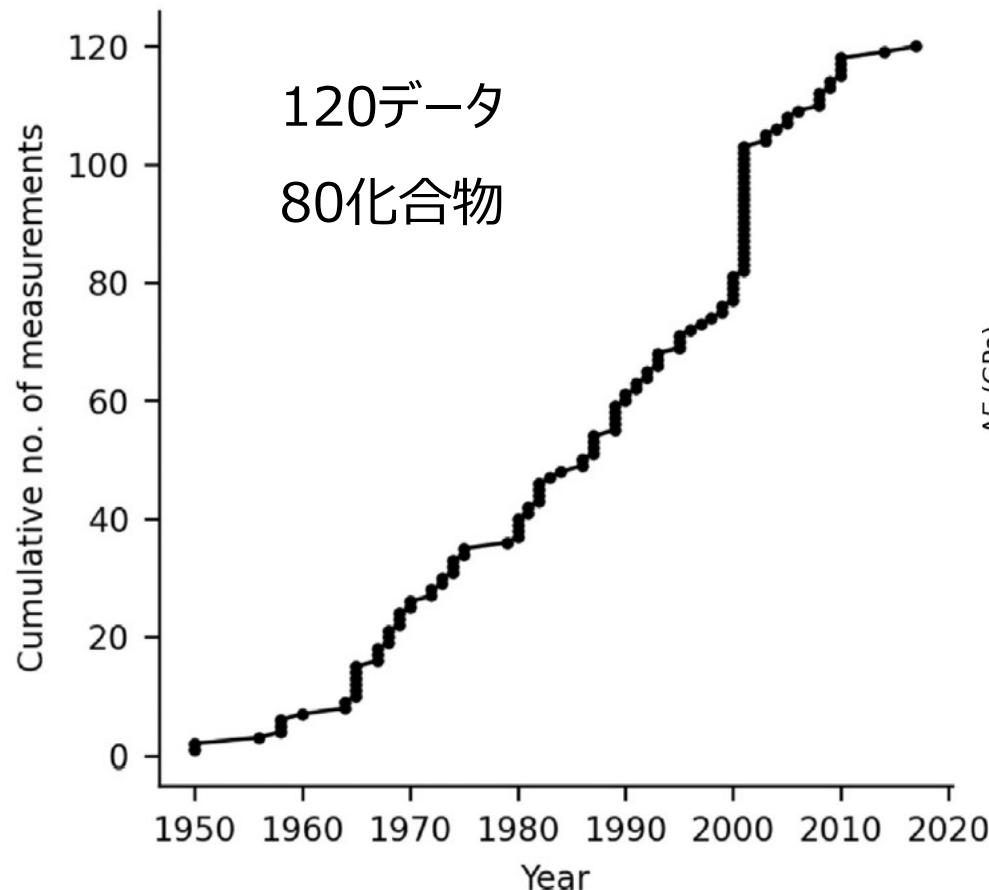
- 体積弾性率 (bulk modulus)
- 剛性率 (shear modulus)
- ヤング率 (Young's modulus)
- ポアソン比 (poisson ratio)
- 異方性 (anisotropy)

例) 体積弾性率

$$K = \frac{1}{9} \sum_{i=1}^3 \sum_{j=1}^3 C_{ij}$$

弾性定数テンソルのデータセット

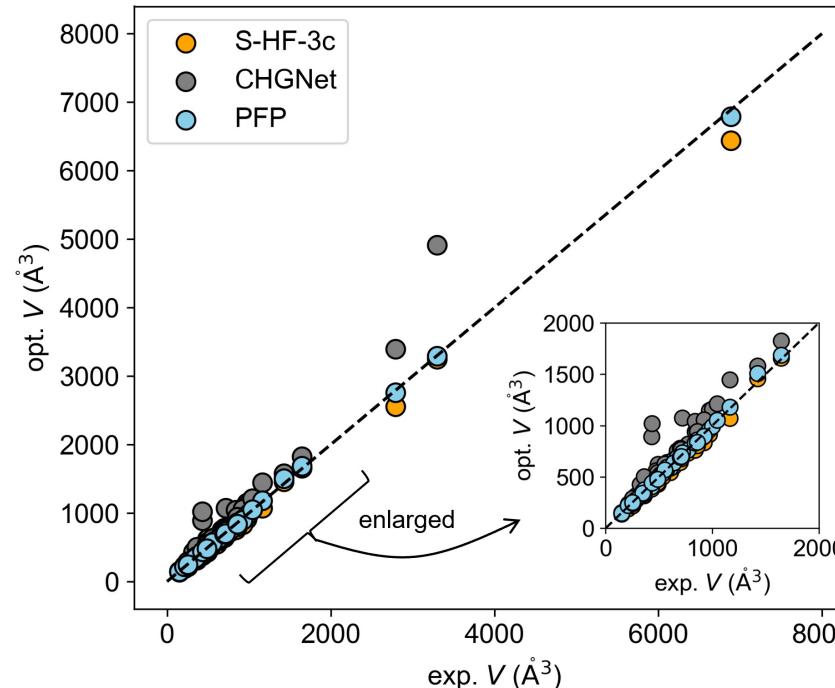
弾性テンソルが測定された分子結晶のデータ数



先行研究によると、分散力補正したHartree-Fock計算
(S-HF-3c) が概ね実験値の弾性率を再現

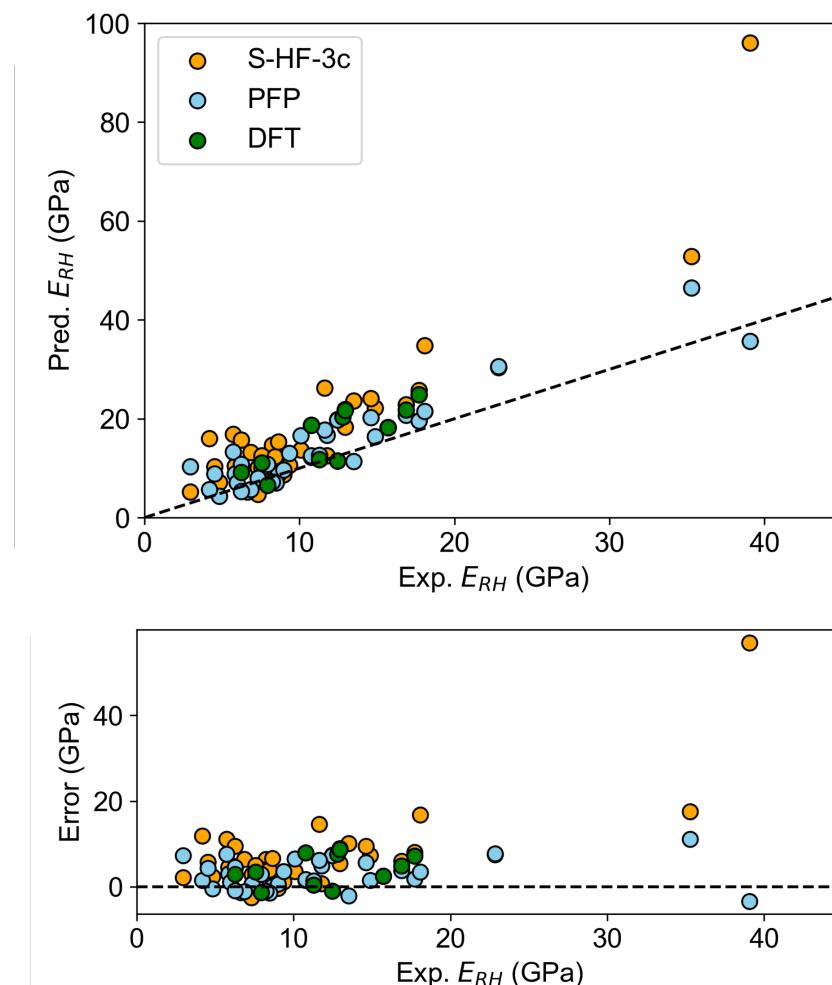
機械学習ポテンシャルの検証

結晶体積の再現性

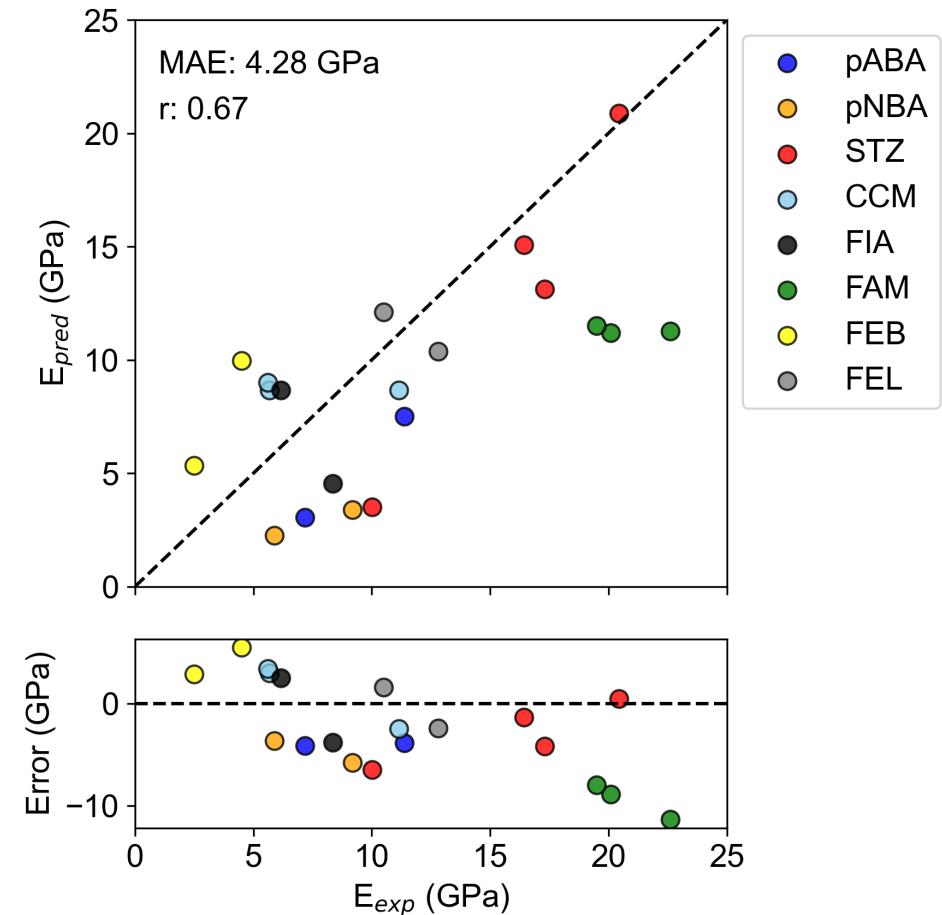
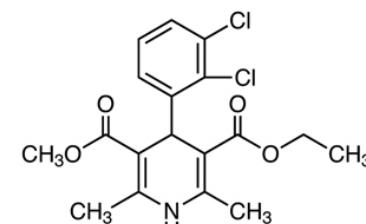
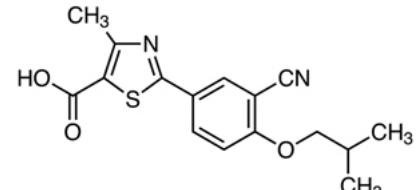
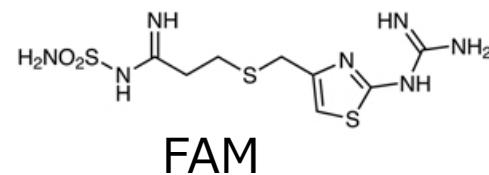
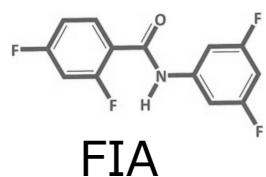
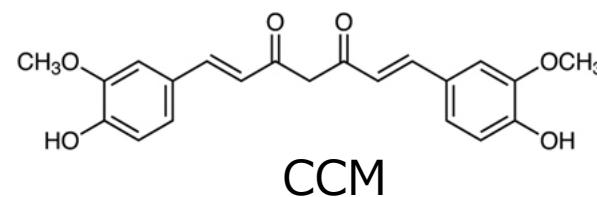
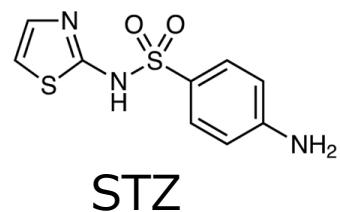
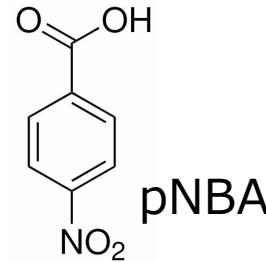
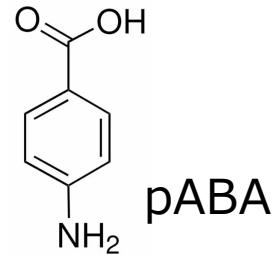


Dataset	MAE of $\Delta V_{\text{PFP}} (\%)$	MAE of $\Delta V_{\text{CHGNet}} (\%)$
Elasticity ($n = 44$)	2.15	19.88
X23 ($n = 23$)	4.27	25.19
Y2023 ($n = 20$)	1.88	33.00

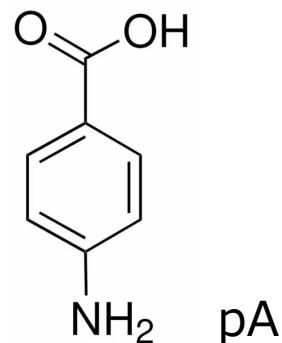
平均ヤング率 E_{RH} の比較



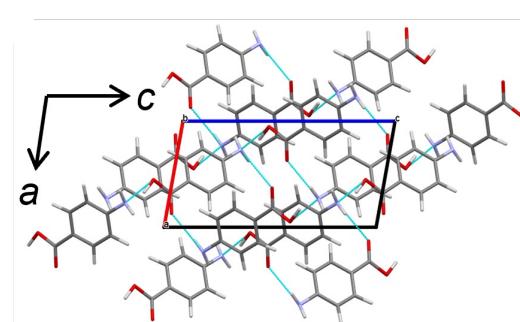
ヤング率の比較



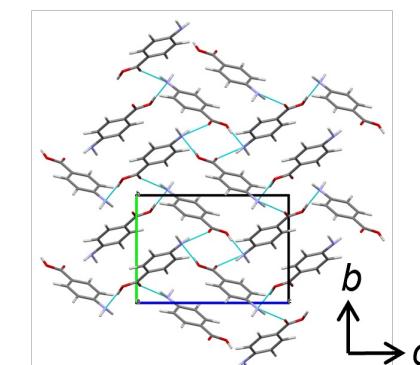
各軸方向のヤング率の大小



I型

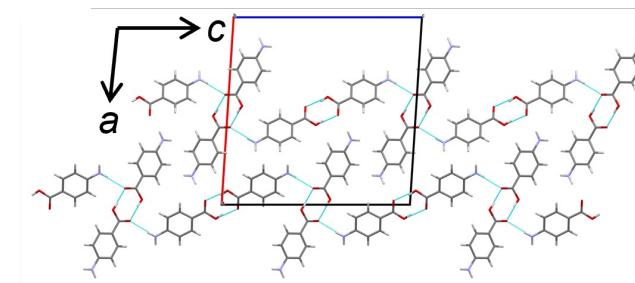


3D H-bonded network

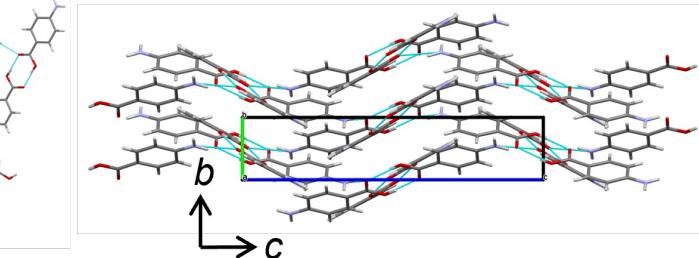


$$\begin{aligned}E_{[100]} &= 6.94 \text{ GPa} \\E_{[010]} &= 5.25 \text{ GPa} \\E_{[001]} &= 7.19 \text{ GPa}\end{aligned}$$

II型



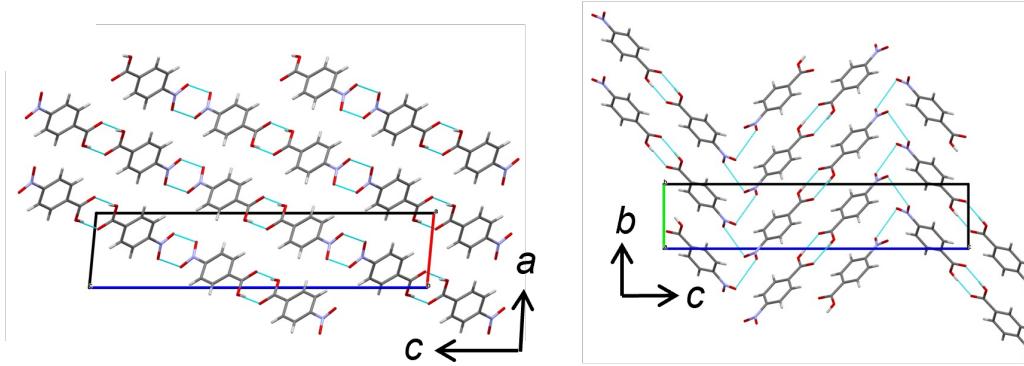
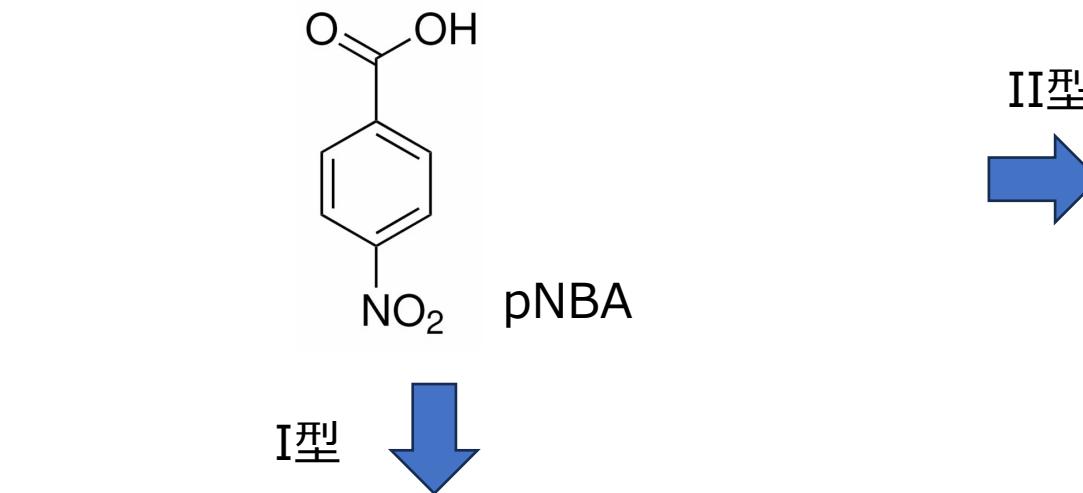
1D zig-zag H-bonded network



$$\begin{aligned}E_{[100]} &= 3.05 \text{ GPa} \\E_{[010]} &= 0.78 \text{ GPa} \\E_{[001]} &= 1.92 \text{ GPa}\end{aligned}$$

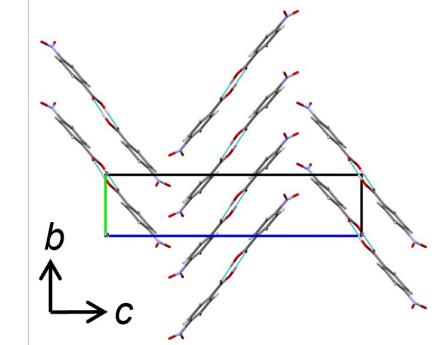
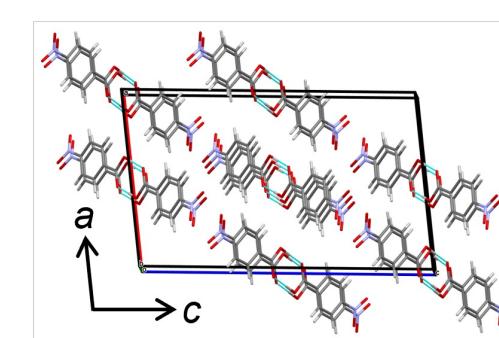
- 水素結合ネットワークの方向にはヤング率が大きい
 - 水素結合ネットワークの積層方向もヤング率が大きい

各軸方向のヤング率の大小



$$\begin{aligned}E_{[100]} &= 4.24 \text{ GPa} \\E_{[010]} &= 2.03 \text{ GPa} \\E_{[001]} &= 2.25 \text{ GPa}\end{aligned}$$

II型 

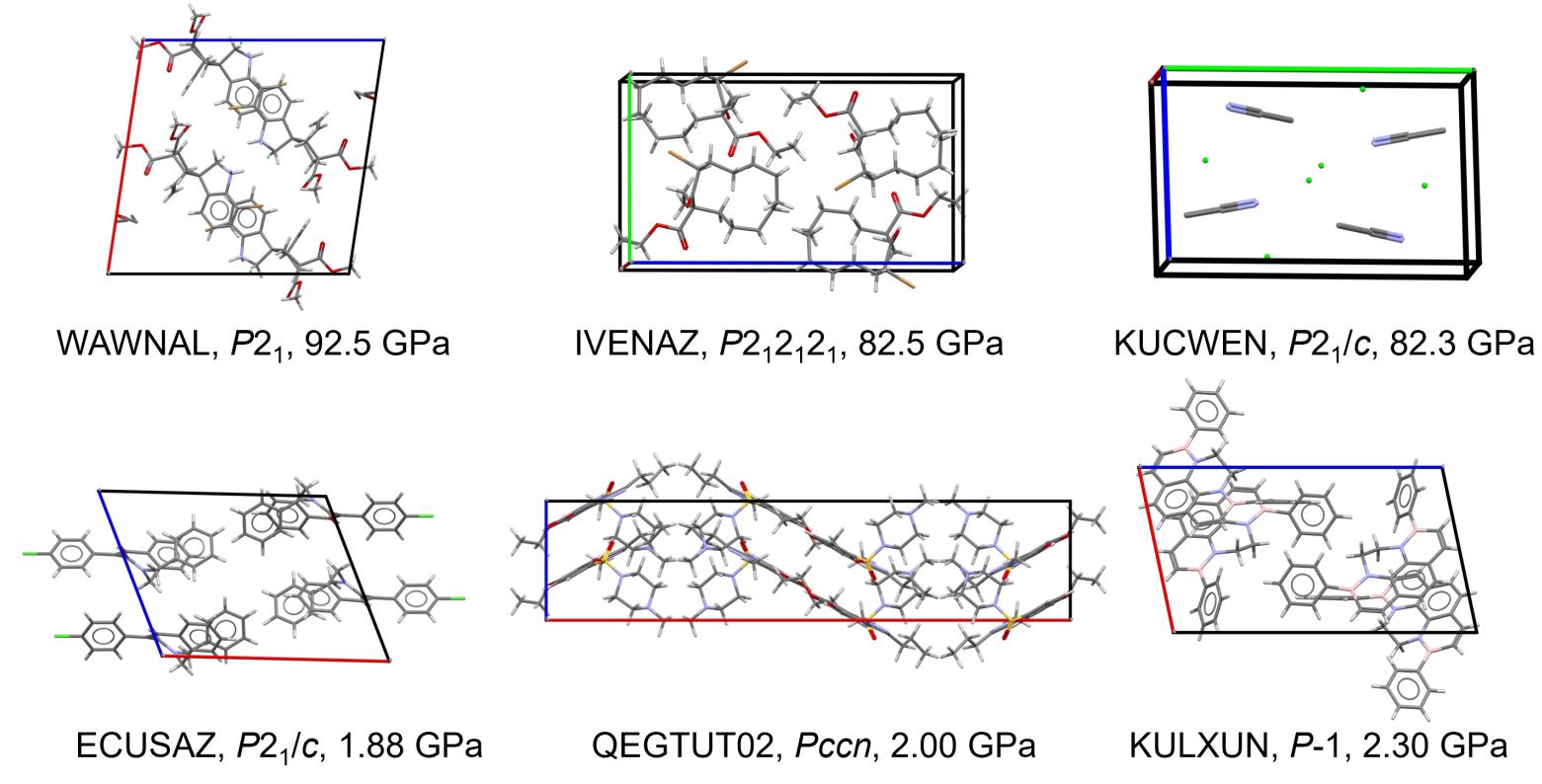
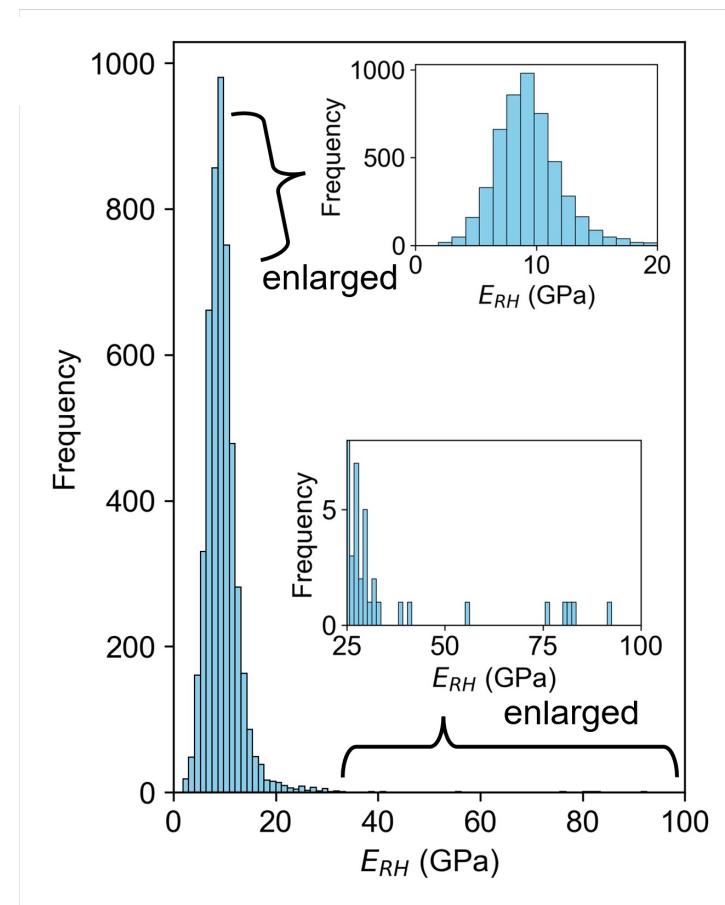


$$\begin{aligned}E_{[100]} &= 2.53 \text{ GPa} \\E_{[010]} &= 1.72 \text{ GPa} \\E_{[001]} &= 3.39 \text{ GPa}\end{aligned}$$

- 水素結合ネットワークの方向にはヤング率が大きい
- 水素結合ネットワークの積層方向もヤング率が大きい
- π スタッキングの方向はヤング率が小さい

弾性率スクリーニング

CSDの結晶のヤング率をランダムに5000個抽出



将来展望 未知の有機固体探索に向けて

