機械学習を用いた分子構造のデータ解析に関する研究

情報科学研究科　情報基礎科学専攻　小林・佐藤研究室
C3IM1030 深澤祐輔

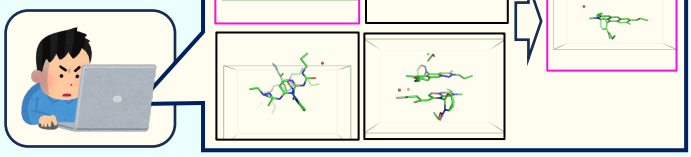
背景・課題

有機化合物のX線結晶構造解析

- 薬学や材料・生命科学などの様々な分野における新しい知見の獲得や技術発展への貢献
  - タンパク質や酵素の構造を理解することによる新薬・新規治療法の開発
  - 高分子材料の理論的な設計の実現
- 結晶構造の理解には，原子スケールの構造解析が必要
- XFEL(X線自由電子レーザ)より得られた回析像から候補となる分子構造を算出

X線結晶構造解析の課題

- 算出される分子構造データの数は膨大
- データ選別の大変さ
  - 研究者が一つ一つ目視で確認し，正解データを選別
  - 長時間の作業となり、研究者への負担大



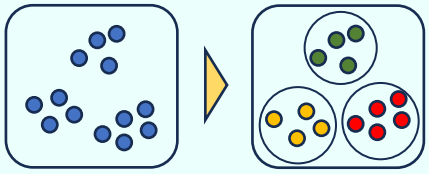
研究の目的

- 分子データの選別作業における作業量の削減のため，AIを駆使した高精度なデータ解析技術の確立

提案手法：分子のグラフ化による特徴量抽出とそれを用いたクラスタリング手法

クラスタリング

- 教師なし学習の一つ
- 分子データを類似構造データごとにグループ（クラスタ）分け
- クラスタごとに分子データ構造に意味付け



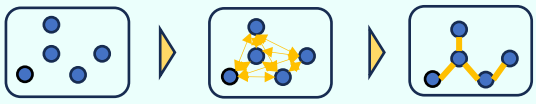
クラスタリングに用いる分子データの特徴量

- 分子を構成する原子の接続関係（トポロジ）やそのサイズなど，分子の特徴を適切に捉えることができる特徴量を用いることが必要
  例：分子を構成する原子の数，ベンゼン環の数
- X線構造解析から得られる分子データは，分子を構成する原子の三次元座標情報のみで表現されるため，そこから原子間の接続情報を推定
- 推定された接続情報をもとに，分子の特徴量を求める

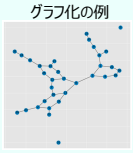
特徴量抽出：分子データのグラフ化

- グラフ：ノード(点)同士を，エッジ(辺)によってつないだデータ構造
- 原子同士の接続といった分子構造を適切に表現することが可能となり，サイズやトポロジといった特徴量を抽出することが可能

グラフ化の流れ



グラフ化の例



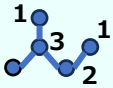
グラフ化によって得られる特徴量

ノード数	エッジ数
次数１の原子の数	ベンゼン環の数
次数２の原子の数	連結成分の数
次数３の原子の数	

分子データのサイズを表す特徴量

分子データのトポロジを表す特徴量

次数
ノードから出るエッジの数



サイズ・トポロジを表す特徴量を用いることで適切なクラスタリングが可能！

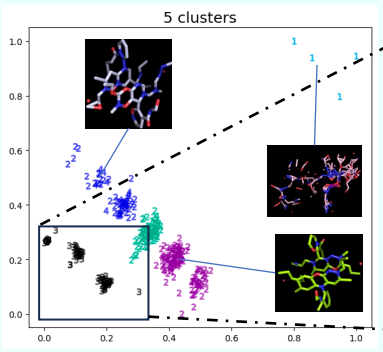
評価

実験条件

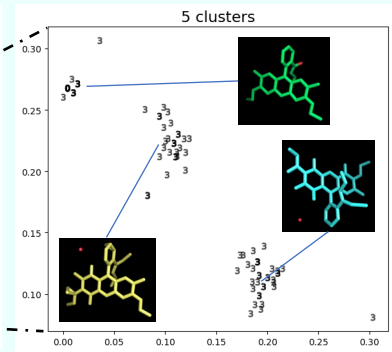
- 結晶分子データ
  - **ローダミン6G(339個)**

  - 接続距離：1.6Å未満
  - 正解データに0，その他は1~4のいずれかのラベルを付与
- クラスタリング
  - k-means(クラスタ数は5に設定)
- 可視化アルゴリズム
  - 主成分分析(PCA)

評価結果



クラスタリング結果.



クラスタリング結果の拡大図.

＊**同じ色に塗られている番号は同じクラスタにグループ分けされる**

- 同じクラスタ内のデータラベルがほぼ一致
  - グラフ化によって分子構造の特徴量を効果的に抽出でき，類似性，非類似性を正しく捉えたクラスタリングができたため

全339個の分子データから，正解データを含むクラスタ内の83個まで確認するべきデータ数を削減

⇨ 研究者の作業量の軽減が見込まれる

結論

- 分子データのグラフ化により，適切なクラスタリングが可能
  - クラスタリングによって，正解候補データが集まったクラスタから正解データを選択でき、分子構造解析を効率よく行うことができる
    → 作業時間の短縮，研究者の負担軽減に貢献

今後の研究計画

- 様々な構造を持つ分子データへの適用やその解析精度の向上
  - 分子データのクラスタリングにおいてより有効な特徴量の考案
  - グラフニューラルネットワーク(GNNs)を用いた半教師あり学習の有効性・適用可能性の評価

分野横断セミナー　セッション４　2024/3/6
テクノフェスティバル　Bグループ　2024/3/7

(URL) https://www.cal.is.tohoku.ac.jp
(E-mail) yusuke.fukasawa.q1@dc.tohoku.ac.jp