Journal Club (2023 年 6 月 26 日) まとめ

担当: 大谷 悠喜

発表論文:

Rice, G., Wagner, T., Stabrin, M. *et al.*

TomoTwin: generalized 3D localization of macromolecules in cryo-electron tomograms with structural data mining.

Nat Methods **20**, 871–880 (2023).

DOI: 10.1038/s41592-023-01878-z

研究目的および概要:

Cryogenic electronic tomography (Cryo-ET) によって得られるデータのアノテーションは研究者の手によって行われており、相当量の労力と時間を要する作業になっている。また、深層学習によるアノテーションを行う研究研究は行われているが、いずれも学習データには人の手によってアノテーションがつけられたデータセットが必要になっている。

そこで本研究では、人の手によるアノテーションが一切必要のない手法 TomoTwin を開発した。TomoTwin は学習データに TEM Simulator によるシミュレーションデータを使用することで人の手によるアノテーション作業を必要としないモデルの学習を行なった。同一の粒子 (タンパク質) を近い距離に、異なる粒子を遠くに射影するデータ空間の学習を行うことによって、任意のクラスターまたは特定のタンパク質のトモグラフィー画像上でのアノテーションを可能にした (図1)。

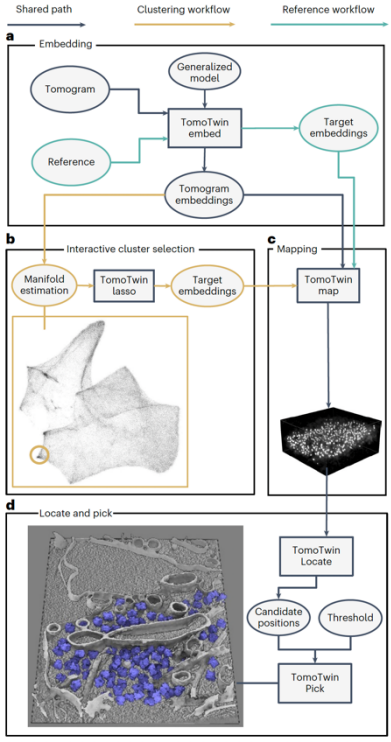

図1

TomoTwin によるアノテーションをもとに再構成した立体構造の再構成ができたこと、大きさの異なるタンパク質においても高い F1 スコアを示すこと等を示した。特に、単粒子再構成において異なる粒子が含まれることで精度が著しく低くなるのを防ぐためにアノテーションの偽陽性がないことを主張した。また、GUI との統合によって専門家以外の研究者も直感的にクラスターの選択や同一タンパク質の抽出などができるようにした (図2)。

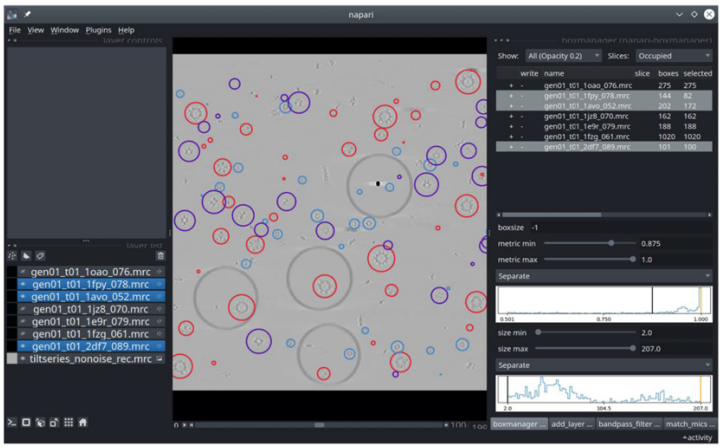

図2

先行研究と比べて何がすごい？ 技術やアプローチのキモはどこ？:

- 人の手による一切のアノテーションを必要としない学習データを使い深層学習モデルを作成した。学習データにシミュレーションデータを使ったのがアプローチのキモである。また、学習に使ったタンパク質の分子量のレンジを広く設定することによって、タンパク質の大きさ・形状に予測精度が左右されないモデルの作成に成功した。

どうやってこの手法/仮説の有効性を検証したのか:

- 学習に使っていないシミュレーションデータセットで検証することによって TomoTwin の性能の評価を行なった。また、立体構造が既知であるタンパク質と TomoTwin によるピッキングによって再構成した構造の比較を行なった。

その他、議論した内容（ネガティブコメントや limitation もあれば):

- Cryo-ET によって得られるデータにアノテーションをつける作業は専門家による職人技である上に非常に時間のかかる作業であるので、それを全自動で行える本手法は従来のボトルネックの一つを解消している。
- Cryo-EM によるタンパク質の立体構造の解像度は 3-5Åにあるのに対して TomoTwin の 13.7Åは低すぎる。Nature Methods に（なんとか）掲載されたのは他にも色々やったからだろう。
- Method に同じタイトルで全く異なる内容の項が存在しているなど研究以外の部分において初歩的なミスが散見された（共著が原因？）。

この研究をさらに発展させるとしたら:

- Snapshot は AlphaFold 等でも得られるようになってきたので、次はタンパク質の結合や立体構造の変化などの動的な情報を取得するような技術の開発が望まれる。