

人工知能による化学データのデジタル化と大規模分子設計

○隅田 真人(1), Xu Xin (1), Zhou Dazhi (1)

(1)新知能統合研究センター 分子情報科学チーム

1. 本課題の研究の背景、目的、関係する課題との関係
人類が長年蓄積してきた化学情報は膨大な規模に達しているが、その多くは論文内のテキストや図表(画像)として埋没しており、高度な活用には至っていない。近年のマルチモーダル大規模言語モデル(LLM)の台頭は、これら多様な形式のデータから情報を抽出・統合し、研究開発を加速させる可能性を秘めている。

特に化学分野においては、化学構造式画像から正確な構造情報(SMILES等)や物性予測を行う「Chemical Image Recognition」の精度向上が急務である。しかし、化学構造式は微細な線画や立体配置を示す記号で構成される離散的な情報であるため、一般的な画像認識モデルでは特徴抽出が困難という課題がある。

本研究の目的は、実験値に代わる高精度な「教師データ」として、大量の量子化学計算データ(物性値、電子状態、スペクトル等)を生成し、それをLLM(ChatGPTやGemini等の基盤モデルを想定)に学習(ファインチューニング)させることにある。これにより、画像から複雑な化学的文脈を理解できる、次世代の化学支援AIの構築を目指した。

2. 具体的な利用内容、計算方法

本研究では、約 50 万分子の多様な有機化合物を対象とし、以下のステップで計算を実施した。

1. **データ生成:** 自動量子化学計算ワークフローツール「QCforever」を用い、Gaussian 16 による構造最適化および物性算出(HOMO-LUMO ギャップ、励起エネルギー、赤外・紫外線吸収スペクトル等)を自動化した。
2. **学習データセットの構築:** 算出された数値データと分子のグラフィックスデータを紐付け、LLM 学習用のトークン形式に変換した。
3. **モデルの学習:** HOKUSAI の GPU ノード(112 コア/GPU 搭載環境)を用い、上記データセットを用いた LLM のファインチューニングを試行した。計算手法としては、限られたリソースで効率的に学習を行うため、Parameter-Efficient Fine-Tuning (PEFT/LoRA) 等の手法の適用を検討した。

3. 結果

HOKUSAI の GPU ノードにおける実行結果として、以下の知見が得られた。

- **GPU 学習の制約:** 現行の HOKUSAI の利用ルール(1 ノード、最大 72 時間)の条件下では、50 万件規模のマルチモーダルデータを用いた LLM のフルパラメータ、あるいは大規模なファインチューニングを完遂するには計算リソースが著しく不足していることが判明した。具体的には、初期の学習ループの段階でタイムアウトとなり、モデルの収束を確認するまでには至らなかった。
- **ワークフローの確立:** 一方で、CPU サーバー群を用いた「QCforever」による自動計算パイプラインは極めて安定して動作した。これにより、人手を介さずに数万件規模の量子化学計算データベースを構築するプロセスを確認できたことは大きな収穫である。GPU での学習は困難であったが、学習に供するための「高品質なデータの量産体制」を HOKUSAI 上で確立することができた。

4. まとめ

本年度の試行により、GPU コア数 112、利用期間 3 日間というリソース割り当ては、数億~数十億パラメータ規模の LLM を大規模データセットで再学習させるには不十分であることが分かった。しかしながら、CPU サーバーを活用した自動計算によるデータベース構築手法については十分な実用性を確認できた。

5. 今後の計画・展望

今後は、HOKUSAI のリソース特性を最大限に活かすため、以下の方針で研究を継続する。

1. **データ生成拠点としての活用:** HOKUSAI の強みである多数の CPU ノードを活用し、さらに大規模な物性データベースの構築に注力する。
2. **ハイブリッド・コンピューティング:** HOKUSAI で生成した大規模データを外部のより大規模な GPU クラスター、あるいはモデル蒸留等の軽量化技術と組み合わせ、効率的な学習プロセスを再設計する。
3. **分子設計への応用:** 構築したデータベースに基づき、特定の物性(光電子材料や医薬品候補等)を持つ分子を AI が自動提案する、生成 AI と量子化学計算の統合システムの開発を進める予定である。