

環状ペプチドの量子化学計算データ生成と機械学習力場への適用性検証

Quantum Mechanical Data Generation for Cyclic Peptides and Evaluation of Applicability to Machine Learning Force Fields

松本 大夢¹ 渡邊 千鶴² 大田 雅照² 奥脇 弘次³ 田 雨時⁴
 Hiromu Matsumoto Chizuru Watanabe Masateru Ohta Koji Okuwaki Yu-Shi Tian
 高谷 大輔⁴ 池口 満徳^{2,5} 福澤 薫⁴ 本間 光貴² 加藤 幸一郎¹
 Daisuke Takaya Mitsumori Ikeguchi Kaori Fukuzawa Teruki Honma Koichiro Kato

1. はじめに

分子シミュレーションは創薬研究において、生体内分子間相互作用の原子レベル解析を通じた医薬品候補化合物の合理的設計を可能にする基盤技術である。これらのシミュレーションは、標的タンパク質の動的構造変化、リガンド結合親和性、および実験的観測が困難な過渡的相互作用の解析等において不可欠な役割を果たす[1]。

現在、生体分子系の動的挙動解析における標準的手法として分子動力学 (MD) シミュレーションが広く用いられているが、その計算精度は採用する分子力場の物理化学的妥当性に大きく依存する。これに対し、量子化学計算は電子の波動関数を第一原理的に取り扱うことで高精度な分子物性予測を実現するが、計算コストが系のサイズ N に対して $O(N^3) \sim O(N^7)$ で増大するスケーリング則により、生体高分子への直接的適用は著しく制限される。この計算効率の問題に対する解決策として提案されたフラグメント分子軌道 (FMO) 法は、巨大分子系を局所的なフラグメントに分割し、フラグメント間の静電的および交換相互作用を取り込む「分割統治」戦略により、計算コストを $O(N) \sim O(N^2)$ に削減することに成功している[2]。しかしながら、創薬研究で要求されるマイクロ秒からミリ秒オーダーの構造サンプリングを実現するには、FMO法を用いてなお計算負荷は依然として過大である。

この計算効率と精度のトレードオフを解決する新たなアプローチとして、機械学習力場 (Machine Learning Force Field: MLFF) が注目を集めている[3]。Behler-Parrinello型に代表される MLFF は、原子周囲の局所的化学環境を回転・並進・置換に対して不変な対称性関数により記述し、深層ニューラルネットワークを用いてポテンシャルエネルギーおよび原子に作用する力を高速に予測する手法である[4]。MLFF は水クラスター、有機小分子、金属表面吸着系等において量子化学計算と同等の精度を達成している。しかし、従来の研究では学習に用いる教師データとして密度汎関数理論 (DFT) 計算が主に用いられており、タンパク質等の巨大生体分子系への展開は教師データ生成の計算コストにより制限されていた[5]。

本研究では、FMO法により得られた高精度量子化学データを教師データとして活用し、生体高分子系に適用可能な機械学習力場を構築する新規手法を提案する。原子に作用する力を含む FMO 計算結果を学習データセットとして用いることで、従来の MLFF では取り扱いが困難であった大規模生体分子系においても、量子化学的精度を維持しつつ計算効率の飛躍的向上を実現する汎用的力場の開発が可能となる。

2. 計算手法及び結果

ここでは、環状ペプチドの構造生成及び分子系のポテンシャルエネルギー/原子上に働く力のデータセット構築に始まり、人工ニューラルネットワークを用いた機械学習力場訓練の概要、ポテンシャルエネルギーと力に関する FMO 計算結果の再現結果について述べる。

2.1 データセットの構築

2.1.1 環状ペプチド構造

本研究でデータセット構築に用いた環状ペプチドの構造を図1に示す。本ペプチドは、12個のアミノ酸残基から構成される環状構造を有する。その一次構造 (配列: IML-SAR-MEA-VAL-PRO-H7V-V99-THR-V85-MVA-ASP-DMN) は、4つの標準アミノ酸 (太字) に対し、サルコシン (SAR) をはじめとする8つの非標準アミノ酸が含まれる点を特徴とする。

本分子は、その構造内に芳香環、およびフッ素 (F) や塩素 (Cl) といったハロゲン原子を多数含む。これらの官能基に由来する π - π 相互作用やハロゲン結合といった電子の効果は、古典的な分子力場では正確な記述が困難である。したがって本分子系は、量子化学的アプローチの有効性を検証する上で理想的な対象といえる。

構造多様性を確保するため、系統的なコンフォメーション探索を実施し、200個の代表的なコンフォメーションを選定した。これらの構造は、主鎖の二面角分布および側鎖配向において十分な多様性を有することを確認している。

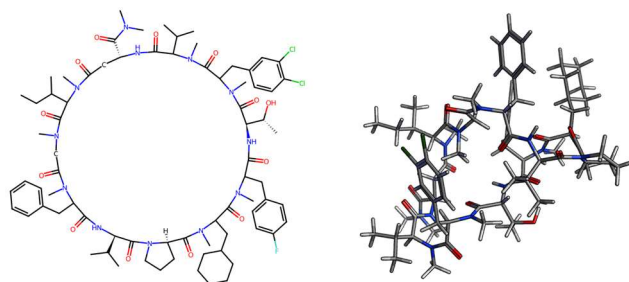


図1 環状ペプチドの2次元構造式/3次元立体構造

- | | |
|-------------|---|
| 1. 九州大学 | Kyushu University |
| 2. 理化学研究所 | RIKEN (The Institute of Physical and Chemical Research) |
| 3. (株) JSOL | JSOL Corporation |
| 4. 大阪大学 | Osaka University |
| 5. 横浜市立大学 | Yokohama City University |

2.1.2 FMO 計算による教師データ生成

2.1.1.で生成した200構造について、FMO2-MP2/6-31G'レベルでの真空中における量子化学的ポテンシャルエネルギーと原子上の力を計算した。使用ソフトウェアはABINIT-MPである。

2.2 機械学習力場の検証

2.2.1 モデルの概要

本研究では、FMO計算により得られた環状ペプチドの量子化学データを再現する機械学習力場として、High-dimensional neural network potentials (HDNNP)を採用した。HDNNPは原子種ごとに独立したニューラルネットワークを用いて原子エネルギーを予測し、その総和により系全体のポテンシャルエネルギーを算出する手法である：

$$E = \sum_i^N E_i^{atom}$$

各原子エネルギーの予測には、Behlerらにより提案されたAtom-Centered Symmetry Functions (ACSF)を記述子として用いた[6]。ACSFは着目原子周辺の化学環境を、分子の並進・回転・置換対称性を保持しながら記述する手法であり、今回の検証では動径対称関数($G^{(2)}$)と角度対称関数($G^{(4)}$)から構成される：

$$G_i^{(2)} = \sum_{j \neq i} e^{-\eta(R_{ij}-R_s)^2} \cdot f_c(R_{ij})$$

$$G_i^{(4)} = 2^{1-\zeta} \sum_{j,k \neq i} (1 + \lambda \cos \theta_{ijk})^\zeta \cdot e^{-\eta(R_{ij}^2 + R_{ik}^2 + R_{jk}^2)} \cdot f_c(R_{ij})f_c(R_{ik})f_c(R_{jk})$$

ここで、 f_c はカットオフ関数である。本研究では動径成分、角度成分各16の記述子を使用し、カットオフ半径は動径10.0 Å、角度5.0 Åに設定した。

2.2.2 実装と計算条件

本研究では、各原子種に対して3層の隠れ層(128-64-32ユニット)を持つ全結合型ニューラルネットワークを使用した。実装はchainerを用いた。データセットは訓練:検証:テスト = 8:1:1の比率で分割し、バッチサイズ16で学習を実施した。モデルの汎化性能は5分割交差検証により評価し、記述子の標準化処理により学習の安定性向上を図った。原子上の力はエネルギーの座標微分により解析的に計算される($\mathbf{F}_i = -\nabla_{\mathbf{r}_i} E$)。

学習では以下の損失関数 \mathcal{L} を最小化する：

$$\mathcal{L} = (1 - \lambda) \cdot \mathcal{L}_{energy} + \lambda \cdot \mathcal{L}_{force}$$

ここで、 λ はエネルギーと力の相対的な重みを制御するパラメータである。エネルギーと力を同時に学習することで、ポテンシャルエネルギー面の値だけでなくその勾配情報も適切に再現し、より物理的に妥当なモデルの構築が可能となる。分子系のサイズを考慮し、 $\lambda = 0.9$ と設定した。

2.2.3 結果

図2に本手法により構築されたHDNNPモデルの予測性能を示す。左図はテストデータセットにおけるエネルギー予測値と参照値(FMO計算値)の相関を、右図は原子に作用する力の各成分についての相関をそれぞれ表している。単一の環状ペプチド構造に対するモデル性能として、エネ

ルギー予測における相関係数 $R = 0.71$ 、力の予測における相関係数 $R = 0.85$ が得られた。この結果は、HDNNPが環状ペプチドのFMO計算結果に含まれるエネルギーと力の傾向を学習可能であることを示している。特に、力の予測において高い相関を示したことは、エネルギー面の勾配情報を適切に学習できていることを示唆しており、分子動力学シミュレーションへの応用において重要な意味を持つ。

しかしながら、これらの結果は単一分子での検証にとどまっており、より多様な環状ペプチドへの適用可能性については今後の系統的な検証が必要である。また、記述子パラメータや力の学習の割合の最適化、ネットワーク構造の改良等により、さらなる精度向上の余地が残されている。

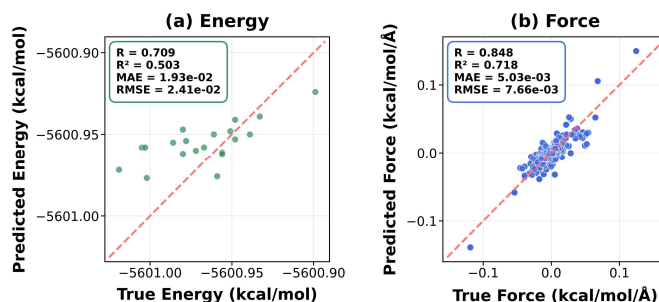


図2 (a)エネルギー、(b)力についての学習結果

3. おわりに

本研究では、フラグメント分子軌道(FMO)法による高精度量子化学計算データを教師データとして、機械学習力場の構築可能性を検証した。単一の環状ペプチドを対象とした実験において、人工ニューラルネットワークモデルがFMO-MP2レベルで算出されたポテンシャルエネルギーおよび原子に働く力の傾向を学習可能であることを実証した。特に、エネルギーと力に対して、それぞれモデルが学習能力を示したことは、本アプローチの基本的な妥当性を示唆している。

謝辞

本研究は、ライフインテリジェンスコンソーシアム(LINC)およびFMO創薬コンソーシアム(FMODD)の一環として実施された。また、本研究の一部は、日本医療研究開発機構(AMED)の創薬等ライフサイエンス研究支援基盤事業(BINDS)による支援(課題番号:JP23ama121030)を受けて行われた。本研究はJSPS科研費JP25KJ1950の助成を受けたものである。

参考文献

- [1] Karplus M., McCammon J.A., "Molecular dynamics simulations of biomolecules", Nature Structural Biology, Vol.9, No.9 (2002).
- [2] Kitaura K., Ikeo E., Asada T., Nakano T., Uebayasi M., "Fragment molecular orbital method: an approximate computational method for large molecules", Chemical Physics Letters, Vol.313, No.3-4 (1999).
- [3] Noé F., Tkatchenko A., Müller K.-R., Clementi C., "Machine learning for molecular simulation", Annual Review of Physical Chemistry, Vol.71 (2020).
- [4] Behler J., Parrinello M., "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces", Physical Review Letters, Vol.98, No.14 (2007).
- [5] Schütt K.T., Sauceda H.E., Kindermans P.J., Tkatchenko A., Müller K.R., "SchNet - A deep learning architecture for molecules and materials", The Journal of Chemical Physics, Vol.148, No.24 (2018).
- [6] Behler J., "Atom-centered symmetry functions for constructing high-dimensional neural network potentials", The Journal of Chemical Physics, Vol.134, No.7 (2011).