Vol.2021-MPS-134 No.12
2021/7/27

深層学習による効率的な高精度量子化学計算結果予測手法の開発

WAN Mingda[1]　安尾信明[2]　関嶋政和[1]

概要：分子の化学的性質を予測することは、量子化学計算の主な目的の 1 つである。密度汎関数法（DFT）は、G4(MP2) のような複雑な計算手法に比べて高速な計算手法であるが、計算精度は十分ではない。本研究では、DFT と G4(MP2) で計算された原子化エネルギーの差を、深層学習を用いた予測によって補正可能であることを示した。QM9 データセットを用いた実験では、1 万個の学習データにより、テストデータの平均絶対誤差を 1kcal·mol^{-1} 以下にすることができ、G4(MP2) レベルの計算と同等の結果が得ることが出来た。この結果は、DFT と機械学習を用いた補正を組み合わせることで、計算精度では十分でない結果を補正可能である可能性を示唆している。

1. はじめに

分子の特性を迅速かつ正確に評価することは、材料科学、生物学、創薬などの多くの分野で非常に重要である。例えば、リガンドの薬理活性は、その原子の空間的配置や電子的性質、そしてこれらの原子が標的タンパク質とどのように相互作用するかに依存することが多い。量子化学計算では、分子の状態、エネルギー、原子結合、反応性などを原子レベルで直接扱うことができ、分子の電子状態を記述する方法である。量子化学の定式化により、分子の原子スケールでの性質を理解することができるようになった。

電子密度を知ることによって、基底状態の正確なエネルギーを得ることができる密度汎関数理論 (DFT)[1] は以下の式 (1) に示される定理に基づいている。

$$E[n] = T_s[n] + U_H[n] + V_{ext}[n] + E_{xc}[n] \qquad (1)$$

現在最もよく用いられている汎関数 B3LYP?は非常に成功した汎関数の例であり、欠点もあるものの、計算時間の割にそこそこ良い結果を与えることが多い。

実際の量子化学計算では、採用する近似計算の手法によって計算精度や計算時間が決定される。計算手法の精度（信頼性）はコスト（計算時間）とトレードオフの関係にあり、欲しい精度と許容できる計算時間との兼ね合いから実際の計算条件が決定される。DFT で満足な結果が得られない場合は、計算時間と引き換えに、さらに高精度の計算手法が必要となる。

[1] 東京工業大学 情報理工学院 情報工学系
[2] 東京工業大学 物質・情報卓越教育院

量子化学計算結果予測における手法は様々提案され、精度の改善が行われている。しかし、機械学習モデルを改善しても、DFT 計算を基に作成したデータセットで訓練されたモデルは DFT の誤差以上の精度で予測を行うことはできない。これは、DFT 計算を元に訓練されたモデルの誤差（ML 誤差）は、ラベルとの予測誤差と DFT の誤差の和で表されるためである（図 1）。

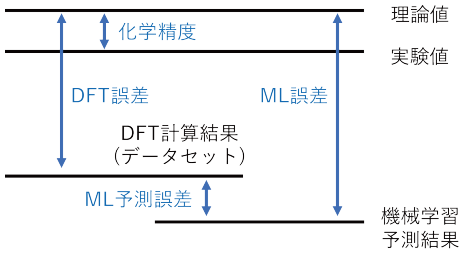

図 1　DFT 計算の結果を基にしたモデルの予測誤差

そして、DFT 計算の誤差は 5kcal·mol^{-1} 程度と、化学誤差よりも大きいことが知られている。[2] したがって、DFT 計算を基に訓練された機械学習モデルは、仮に訓練が適切であっても、化学精度と同程度の精度を達成できないという問題がある。本研究では、高速な量子化学計算手法と高精度量子化学計算の結果（原子化エネルギー）の関係を学習することにより、より高速かつ高精度に化合物のエネルギーを予測する手法を開発と評価を行った。

2. 評価実験

2.1 高精度量子化学計算

本研究では、化学精度と許容できる計算コストとの兼ね

合いから、高精度量子化学計算方法 G4MP2[3] を使用した。本研究で用いたデータベースは QM9[2] の全化合物である。G4MP2 の計算には B3LYP/6-31G(2df,p) レベルの分子三次元構造を利用しており、構造最適化の問題が最小限に抑えられる。

図 2 は QM9 データベースにおける化合物の類似度の分布である。データセットからランダムに 100,000,000 ペアをサンプリングし、類似度を計算した。Fingerprint として MACCS Key, 類似度指標として Tanimoto 係数を用いている。

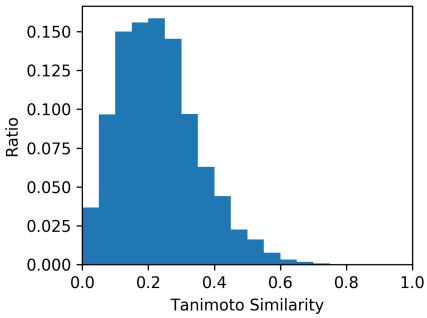

図 2 QM9 データセットにおける化合物の類似度分布

最初は QM9 の xyz ファイルから、Open Babel パッケージ [4] を用いて三次元分子構造を生成した。スピン多重度と電荷が正しくない場合は正しく書き直し、Gaussian 16 量子化学プログラム [5] のデフォルトの分子構造最適化設定を全 133,885 分子に使用した。うち 26 分子は構造最適化が収束しなかったため、キーワード SCF(NoVarAcc,NoIncFock) を使って収束させた。NoVarAcc は最初のステップからの完全な積分精度を使い、NoIncFock は incremental Fock matrix（収束加速）を不使用に設定する。これらのキーワードは正確な結果を得るための収束基準とは無関係である

本研究では 4 種類の原子化エネルギーを用いる。この内訳は内部エネルギー (0K)$\Delta U_{0\mathrm{K}}$、内部エネルギー (298.15K)$\Delta U_{298.15\mathrm{K}}$、エンタルピー (298.15K)$H_{298.15\mathrm{K}}$、自由エネルギー (298.15K)$G_{298.15\mathrm{K}}$ である。すべての分子に対して、高精度計算である G4MP2 と DFT（基底関数は B3LYP/6-31G(2df,p) を用いた）の両手法を用いて 4 種類の原子化エネルギーを計算した。

2.2 特徴量

2.2.1 ECFP 作成

ECFP の実装は、RDKit[6] によって生成された ECFP4 フィンガープリントを使用した。ECFP の表現は固定長のビット列であり、本研究はデフォルトの長さ、1024 を用いた。

2.2.2 Coulomb matrix 作成

原子電荷 $\{Z\}$ と原子座標 $\{\boldsymbol{R}\}$ から構成される Coulomb matrix(CM) を式 (2) に示す。原子電荷と物理距離計算は RDKit[6] で実装した。CM は三次元空間での分子の並進と回転に対して不変だが、機械学習モデルに使用するためには 2 つの問題がある。1 つは CM の次元は分子の原子数に依存すること、もう 1 つは CM 内の原子の順番が定義されていない、つまり 1 つの分子から多くの CM を定義できることである。

$$\boldsymbol{M}_{ij} = \begin{cases} 0.5Z_i^{2.4} \text{ for } i = j \\ \frac{Z_i Z_j}{|\boldsymbol{R}_i - \boldsymbol{R}_j|} \text{ for } i \neq j \end{cases}. \tag{2}$$

最初の問題は分子内に「見えない原子」を導入することによって改善した。このような原子は電荷がゼロ、他の原子と相互作用がない。$Z = 0$ のため、実際に、CM は 0 に埋め込むこと（ゼロパディング）に相当する。分子の原子と見えない原子の合計は定数 n にした。本研究のデータセットは最大原子数が 29 のため、定数 $n = 29$ に設定し、CM のサイズは 29×29 で固定した。CM 内の原子の順番については、本研究ではソートした Coulomb matrix を用いた。

2.3 学習モデル

本研究では、グラフベースのモデルである畳み込みニューラルネットワークを用いた。この手法では局所構造を捉えることができるとされており [7]、畳み込み層では、画像などに適用され、局所特徴を抽出することが可能である。同様に、原子の局所的な相関関係は利用されうる。しかしながら、画像と対照的に、分子の原子はグリッド上に配置されていない。また、畳み込みカーネルは連続性が必要である。分子エネルギーの特徴から、特定環境の原子エネルギーの和で表現される。

$$E^{\mathrm{est}} = \Sigma_i E_i^{\mathrm{est}} \tag{3}$$

原子 i のエネルギーは 2 つの完全接続層で予測できる。

$$E_i^{\mathrm{est}} = W_1 \mathbf{h}_i + \mathbf{b}_1 \tag{4}$$

$$\mathbf{h}_i = \tanh\left(W_2 \mathbf{c}_i^{(T)} + \mathbf{b}_2\right) \tag{5}$$

実験では隠れ層 $\mathbf{h}_i$ は 15 個のノードをもつネットワークを用いた。$\mathbf{c}_i^{(T)}$ は原子 i の係数ベクトル、T は更新の回数である。負数も予測したいため、活性化関数を tanh とした。

最初は原子電荷 Z_i に従い、分子の各原子 i に初期係数ベクトル $\mathbf{c}_i^{(0)}$ を割り当てた。本研究は 30 個の係数を持つ記述子を使用した。初期係数ベクトル $\mathbf{c}^{(0)} \sim \mathcal{N}(0, 1/\sqrt{30})$ に従い、ランダムで生成した。

原子 i と原子 j の距離行列をガウス展開した。本研究で

は $\Delta\mu = \sigma = 0.2\mathring{A}$、$\mu_{\min} = -1\mathring{A}$、$\mu_{\max} = 9.2\mathring{A}$ で設定した。

$$\hat{\mathbf{d}}_{ij} = \left[\exp\left(-\frac{(d_{ij} - (\mu_{\min} + k\Delta\mu))^2}{2\sigma^2}\right)\right]_{0\le k\le \mu_{\max}/\Delta\mu} \tag{6}$$

原子 i と他の原子の相互作用を考慮し、係数ベクトルの更新は下式のように行われる。

$$\mathbf{c}_i^{(t+1)} = \mathbf{c}_i^{(t)} + \sum_{j\neq i} \mathbf{v}_{ij} \tag{7}$$

相互作用 $\mathbf{v}_{ij}$ の定義は：

$$\mathbf{v}_{ij} = \tanh\left(W^{fc}\left(\left(W^{cf}\mathbf{c}_j + \mathbf{b}^{f_1}\right) \circ \left(W^{df}\hat{\mathbf{d}}_{\mathrm{ij}} + \mathbf{b}^{f_2}\right)\right)\right) \tag{8}$$

ここで、$\circ$ は要素ごとの行列積を表す。ここでは 60 個のノードを使用した。本研究において $T = 2$ を利用し、係数ベクトルを 2 回更新した。深層学習モデルは tensorflow[8] で構築した。

2.4 評価方法

本研究は回帰問題であり、2 つの広く使われる評価指標を用いた。

- 平均絶対誤差（MAE、Mean Absolute Error）

$$MAE = \frac{\sum_i |y_{\mathrm{obs},i} - y_{\mathrm{pred},i}|}{n} \tag{9}$$

- 二乗平均平方根誤差（RMSE、Root Mean Squared Error）

$$RMSE = \sqrt{\frac{\sum_i (y_{\mathrm{obs},i} - y_{\mathrm{pred},i})^2}{n}} \tag{10}$$

本研究では、元データを training / validation / testing セットに分割した。Training セットはモデルを訓練するために使用され、validation セットはハイパーパラメータを調整するために使用され、testing セットはモデルを評価するために使用された。本研究は学習に用いるデータ数を変化させながら行った。各データ数の場合に対して、training / validation データをランダムに 7/3 に分け、残りのすべてを testing データとして使用した。

3. 結果

図 3 は B3LYP と G4MP2 の原子化エンタルピー (298.15K) の対応関係を示している。偏差が 51.31 kcal·mol^{-1} の「外れ値」があるが、90% 以上の分子誤差は 10 kcal·mol^{-1} 以下である。

G4MP2 を基準とした際の B3LYP 計算結果の MAE と RMSE を表 1 に示す。B3LYP の原子化エンタルピー (298.15K) は G4MP2 から 4.66 kcal·mol^{-1} 程度外れており、この結果は Ramakrishnan の研究と一貫している。Ramakrishnan の研究によれば、100 個のランダム分子に対して、B3LYP と G4MP2 の MAE = 5.0 kcal·mol^{-1}、RMSE = 6.1 kcal·mol^{-1} である。[2]

表 1 G4MP2 と B3LYP の原子化エネルギーの MAE と RMSE(kcal·mol^{-1})

	MAE	RMSE
$\Delta U_{0\mathrm{K}}$	4.67	5.86
$\Delta U_{298.15\mathrm{K}}$	4.66	5.85
$\Delta H_{298.15\mathrm{K}}$	4.66	5.85
$\Delta G_{298.15\mathrm{K}}$	4.68	5.86

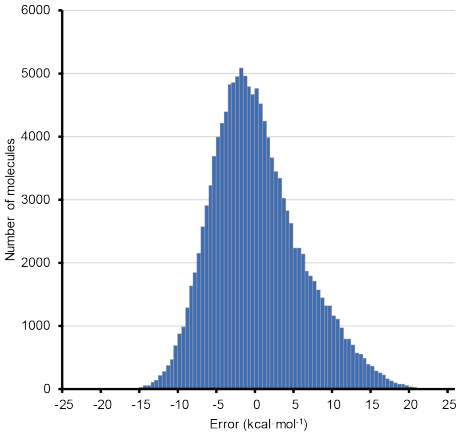

図 3 B3LYP と G4MP2 の原子化エンタルピー (298.15K) の誤差ヒストグラム

直接 G4MP2 の原子化エンタルピー (298.15K) をラベルにし、batch size=16、learning rate=0.001、epoch=200 で訓練したが、MAE は 4.35kcal·mol^{-1} となった。さらに訓練データを増え、10 万分子を訓練データにする場合でも、直接予測の MAE は 2.17kcal·mol^{-1} と、化学精度の 2 倍以上の結果となった。エネルギーの差を計算することで、データセットのノイズを減らし、予測が容易になる可能性がある。以下の実験ではすべてエネルギー差をラベルに使用した。

データ量と精度の評価を行うため、異なる training / validation セットを用い、それぞれ 10 回の実験（batch size=64、learning rate=0.001、epoch=50）を行いった。Training / validation セットサイズ 0 のポイントは B3LYP と G4MP2 の MAE と RMSE であり、灰色のエリアは目標である化学精度以下の領域を示している。Training / validation セットのサイズが大きくなると共に、予測誤差は減少する。Training / validation セットのサイズが 10,000 の場合、モデルは残り 123,885 化合物の testing セットで化学精度と同等の精度を示した。

4. 考察

本研究では、高精度量子化学計算結果（原子化エネルギー）を高速化かつ高精度に予測することを目的とし、 1) 13 万化合物を含む高精度量子化学データベースの構築、2) 化学精度と同等の精度で予測が可能な深層学習モデルの開発 を行った。

従来の量子化学計算結果予測モデルは DFT 計算の誤差によって制約され、適切に訓練したとしても化学精度を達成できない可能性があった。そこで、我々は 133,885 分子の熱化学計算を G4MP2 理論レベルで行い、Δ 機械学習を用いた補正モデルによって B3LYP の計算結果から G4MP2 の計算結果を予測した。

結果として、10,000 化合物以上を用いて訓練することで、残る 12 万化合物について化学精度と同程度の精度をもつ機械学習モデルを構築することが可能であることを示した。また、10 万分子を学習した場合、最終的な精度は約 0.30 kcal·mol^{-1} と、化学精度よりはるかに低くなることを示した。高価な G4MP2 の代わりに、機械学習は非常に短時間で同じレベル精度の結果を得ることができた。これらの結果から、低精度手法と補正モデルの併用によって高精度手法を代替できる可能性が示唆された。

参考文献

[1] Hohenberg, P. and Kohn, W.: Inhomogeneous Electron Gas, *Phys. Rev.*, Vol. 136, pp. B864–B871 (online), DOI: 10.1103/PhysRev.136.B864 (1964).
[2] Ramakrishnan, R., Dral, P. O., Rupp, M. and Von Lilienfeld, O. A.: Quantum chemistry structures and properties of 134 kilo molecules, *Scientific data*, Vol. 1, p. 140022 (2014).
[3] Curtiss, L. A., Redfern, P. C. and Raghavachari, K.: Gaussian-4 theory using reduced order perturbation theory, *The Journal of Chemical Physics*, Vol. 127, No. 12, p. 124105 (online), DOI: 10.1063/1.2770701 (2007).
[4] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. and Hutchison, G. R.: Open Babel: An open chemical toolbox, *Journal of cheminformatics*, Vol. 3, No. 1, p. 33 (2011).
[5] Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Petersson, G. A., Nakatsuji, H., Li, X., Caricato, M., Marenich, A. V., Bloino, J., Janesko, B. G., Gomperts, R., Mennucci, B., Hratchian, H. P., Ortiz, J. V., Izmaylov, A. F., Sonnenberg, J. L., Williams-Young, D., Ding, F., Lipparini, F., Egidi, F., Goings, J., Peng, B., Petrone, A., Henderson, T., Ranasinghe, D., Zakrzewski, V. G., Gao, J., Rega, N., Zheng, G., Liang, W., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Throssell, K., Montgomery, Jr., J. A., Peralta, J. E., Ogliaro, F., Bearpark, M. J., Heyd, J. J., Brothers, E. N., Kudin, K. N., Staroverov, V. N., Keith, T. A., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A. P., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Millam, J. M., Klene, M., Adamo, C., Cammi, R., Ochterski, J. W., Martin, R. L., Morokuma, K., Farkas, O., Foresman, J. B. and Fox, D. J.: Gaussian˜16 Revision B.01 (2016). Gaussian Inc. Wallingford CT.
[6] Landrum, G.: RDKit: Open-source cheminformatics.
[7] LeCun, Y., Bengio, Y. et al.: Convolutional networks for images, speech, and time series, *The handbook of brain theory and neural networks*, Vol. 3361, No. 10, p. 1995 (1995).
[8] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015). Software available from tensorflow.org.