

分子軌道エネルギーを説明変数とした機械学習

寺前 裕之^{a*}, 玄 美燕^b, 高山 淳^b, 岡崎 真理^b, 坂本 武史^b

^a城西大学理学部化学科, 〒350-0295 坂戸市けやき台 1-1

^b城西大学薬学部薬学科, 〒350-0295 坂戸市けやき台 1-1

teramae@gmail.com

The values of the DPPH free radical scavenging ability values (IC_{50}) of ferulic acid and its derivatives have been estimated by machine learning with only molecular orbital (MO) energies as the explanatory variables. We use four machine learning regression methods, Partial Least Square (pls), Random Forest (rf), Multi-Layer Perceptron Neural Network with Optional Monotonicity Constraints (monmlp) and eXtreme Gradient Boosting with Linear model (xgbLinear), using R-package caret. We use 22 molecules for the training set and 6 molecules for the test set. The root mean square (RMS) errors of predicted values for the test set are used for estimating the precision of the training. The best result is obtained by xgbLinear just using two MO energies (HOMO and LUMO). It has been proved that the IC_{50} values can be predicted by the molecular orbital energies only.

(Received: January 2, 2023; Accepted for publication: February 6, 2023; Online publication: April 22, 2023)

Keywords : Antioxidant effect, Radical scavenging ability, Equilibrium geometries, Eigenvalues of molecular orbital, Machine learning

1 はじめに

機械学習の急速な発展により近年化学分野への応用の可能性が広がってきている。機械学習による予測には、説明変数のセットを適切に選ぶ必要があるが、応用の可能性を広げるためには、なるべく簡単に測定できるか、あるいは理論計算により求められることが望ましい。我々はどのような分子にも対応できる説明変数のセットとして、分子軌道法により求められた分子軌道エネルギーが使えるのではないかと考え、一連の研究を行ってきた。

我々は、2008年にベンゾジアゼピン(BZP)およびチエノジアゼピン系の抗不安薬について、抗不安性と抗痙攣性の強さと next-HOMO の軌道エネルギー値と相関があることを見いだしたことから、このような発想に至った [1]。以前の研究で同様にフェルラ酸(FA)の抗酸化作用に対して、DPPH フリーラジカル消去濃度(IC_{50})の置換基効果を説明できることを示した [2]。ただし、一つの軌道エネルギー値のみに対して強い相関関係を持つことは一般的には稀なため、複数の軌道エネルギー値と IC_{50} との関係を Random Forest (rf) 回帰法により機械学習を試みたところ、強い相関関係が見いだされ、分子軌道計算

のみで IC_{50} 値の予測が可能となった。

一方、BZP の薬効やその他の性質に対して、抗不安性と抗痙攣性以外に以前は関係性を検討できなかった筋弛緩性、さらには薬物動態学的パラメータである T_{max} と $T_{1/2}$ についても機械学習による予測が可能であることを報告した [3]。以上の機械学習を用いた研究では訓練データとテストデータの分割についてはサンプル数が少ないことから、複数の分割を試行して最良の結果を採用していたが、これは本来得られるデータよりも良すぎるデータが得られる事が多く、正確性に疑問が生じた。

そのため本研究ではフェルラ酸の抗酸化作用について再計算を行った結果について報告する。また、以前の研究ではフェルラ酸についてラジカルモデルを考えて UHF 法による分子軌道エネルギーも使用していたが、今回はラジカルではない元々のフェルラ酸誘導体の分子軌道エネルギーのみによる予測を試みた。

2 計算方法

分子軌道計算プログラム Gaussian16 RevB.01 [4] を使用し RHF/6-31G (d, p) レベルで構造最適化を行った。最適化された構造に対して振動数計算を行い、安定構造であ

Table 1. RMS errors of test set with number of orbitals

# of orbitals	2	4	6	8	10
xgbLinear	0.276	0.434	0.333	0.672	0.692
monmlp	0.542	0.667	0.948	0.860	0.944
rf	0.387	0.354	0.374	0.409	0.423
pls	0.568	0.511	0.468	0.489	0.489

Table 2. R^2 values of training set with number of orbitals

# of orbitals	2	4	6	8	10
xgbLinear	0.998	1.000	1.000	1.000	1.000
monmlp	0.530	0.888	0.955	1.000	1.000
rf	0.832	0.851	0.851	0.876	0.905
pls	0.333	0.324	0.325	0.344	0.334

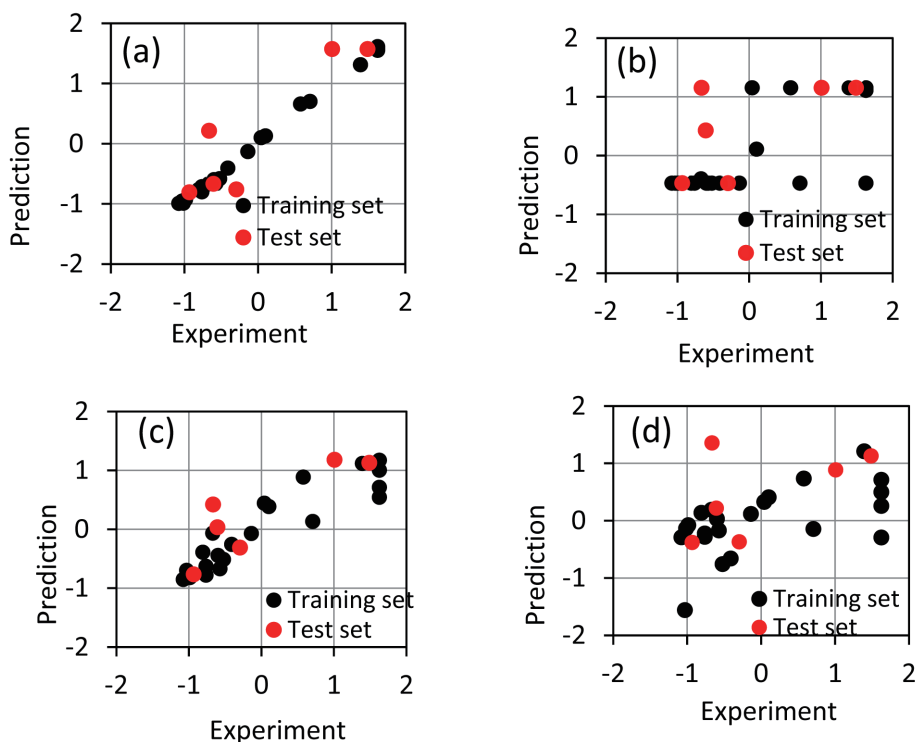


Figure 1 Plots of training and test sets of IC_{50} values of ferulic acid derivatives. Regression methods are (a) xgbLinear, (b) monmlp, (c) rf, and (d) pls, respectively. Numbers of orbitals are 2 for (a) and (b), and 4 for (c) and (d). The black circles denote the training set, and the red circles denote the test set.

ることを確認した。機械学習計算にはRのcaretパッケージ [5]を使用し分子軌道計算により得られた軌道エネルギー値を説明変数として、最大10軌道までを使用して解析を試みた。トレーニングセットは22分子、テストセットは6分子とした。交差検証には3-fold-validationを使用した。回帰法は、いくつかの方法を試しているが、本論文ではxgbLinear (線形勾配ブースティング), monmlp (単調多層パーセプトロンニューラルネットワーク), rf (ランダムフォレスト), pls (部分的最小自乗回帰)の4種類について報告する。

トレーニングセットとテストセットへの分割は、乱数を使用して行った。テストセットの予測値と実験値の差のroot mean square (RMS)エラー値を計算して、RMSエ

ラー値が最小値となる時の分子軌道エネルギーの数を原則として説明変数の最適な数として採用した。

3 結果と考察

Table 1に分子軌道数2-10個を説明変数とした場合のテストセットの予測値と実験値との誤差のRMS値, Table 2にはトレーニングセットの決定係数の値 R^2 値を示した。またFigure 1には $\log(IC_{50})$ 値の実験値と予測値の分散を示した。数値は標準化されている。

PLsについてはRMSエラー値が大きくまたトレーニングセットの決定係数も小さくなっており、適切でないと考えられる。なおRMSエラー値は $dB[\mu M]$ 単位で表している。XgbLinearとmonmlpはいずれもトレーニングセッ

トの決定係数が説明変数の数によっては1になっているが、その時の説明変数の数とRMSエラー値はかなり異なる。Monmlpは分子軌道数が8および10の場合に相関係数が1になっているが、RMSエラー値は最小にはなっておらず、最小となるのは2の場合である。つまり分子軌道数が2以上では説明変数である軌道数の増加はほぼ過学習に使われていると推定される。XgbLinearもRMSエラー値が最小となるのは軌道数が2の場合で決定係数は2-10のすべてでほぼ1になっている。このフェルラ酸のラジカル消去能の予測に関してはxgbLinearが最も優れた回帰法であると言える。RfはxgbLinearに次いで小さいRMSエラー値を示しているが、トレーニングセットの決定係数はやや小さくなるという結果になっている。なお以上のことはFigure 1においても明らかに示されている。

4 結論

機械学習を用いてフェルラ酸のラジカル消去能を分子軌道のエネルギー値を説明変数として予測することを試みた。以前までの研究とは違って、ラジカル状態の分子軌道エネルギーではなく、元々のフェルラ酸誘導体の分子軌道エネルギーのみを用いているため、ラジカル体の分子構造を求める必要が無くなり、より多くの分子に対しての応用が可能になったと考えられる。

4種類の回帰法を使用した、いずれの方法でもある程度の相関が得られ、分子軌道計算と機械学習の組み合わせは強力なツールであり、様々な分子物性の予測が期待できるであろうことがわかった。

参考文献

- [1] H. Teramae, K. Ohtawara, T. Ishimoto, U. Nagashima, *Bull. Chem. Soc. Jpn.*, **81**, 1094 (2008). DOI:10.1246/bcsj.81.1094
- [2] H. Teramae, M. Xuan, T. Yamashita, J. Takayama, M. Okazaki, T. Sakamoto, *J. Comput. Chem. Jpn.*, **18**, 211 (2019). DOI:10.2477/jccj.2019-0034
- [3] H. Teramae, M. Xuan, J. Takayama, M. Okazaki, T. Sakamoto, *AIP Conf. Proc.*, **2611**, 020007 (2022). DOI:10.1063/5.0119589
- [4] Gaussian 16, Revision B.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.
- [5] <https://cran.r-project.org/web/packages/caret/caret.pdf>