構造的・電子的記述子を用いた機械学習による NMR 核磁気遮蔽定数の予測と解析

(早大先進理工 [1]・早大理工総研 [2]・(株)三菱ケミカルホールディングス [3]) ○小川 陽太郎 [1]・速水 雅生 [2]・中野 匡彦 [2,3]・清野 淳司 [1,2]

Prediction and Analysis of NMR Nuclear Magnetic Shielding Constants by Machine Learning with Structural and Electronic Descriptors ([1]*School of Advanced Science and Engineering, Waseda Univ.*, [2]*WISE, Waseda Univ.*, [3]*Mitsubishi Chemical Corp.*) ○Yohtaro, Ogawa[1]; Masao, Hayami[2]; Masahiko, Nakano[2,3]; Junji, Seino[1,2]

Toward the practical application of automated experiments, it is important to develop automatic analysis techniques for spectra to identify compounds and determine their structures. In this study, we focused on NMR spectra. We generated a large dataset of nuclear magnetic shielding constants by quantum chemical calculations for compounds in the QM9 database. Machine learning models to predict the properties were constructed.

Keywords : NMR Spectrum; Machine Learning; Quantum Chemical Calculation; Materials Informatics; Structural and Electronic Descriptors

【緒言】NMR スペクトルは有機分子やタンパク質など化合物の同定・構造決定のために重要な技術である。近年開発が進む自動実験の実用化に向けて、機械学習を用いた高速なスペクトル予測手法の開発が進んでいる。より一般的な化合物に適用できる手法とするには、実験だけでなく量子化学計算などにより、膨大なデータを生成することは重要である。本研究では、量子化学計算により算出した核磁気遮蔽定数を、構造的/電子的記述子を用いて高速に予測する機械学習モデルを構築した。

【方法】C, O, N, F 原子が 9 個以下になるよう自動生成された QM9 データベース内の約 13 万分子・約 235 万個の H 原子に対して B3LYP/6-31G**レベルで核磁気遮蔽定数の計算を行い、目的変数値を収集した。記述子として、(i)最近接の 28 原子までの距離と各元素種、(ii)RDKit と alvaDesc により生成した 2 次元構造が主となる分子記述子 1124 種、(iii)B3LYP/6-31G**レベルの軌道エネルギーや占有数などの電子的記述子 66 種を用いた。機械学習手法は XGBoost を用いた。

【結果】各記述子を用いた学習モデルによる MAE と R^2 を Table 1 に示す。(i)と(iii)はともに高精度であり、距離情報と電子状態情報が類似していると推察される。また、(iii)を 10 アミノ酸から成るシニョリンに適用した (Figure 1)。学習データ/記述子が不足している一部を除いて、概ね核磁気遮蔽定数を予測でき、小さい分子のデータのみの学習からタンパク質などの大きな分子の結果を予測できる可能性が示された。

Table 1. Performance of the models for test molecules in the QM9 database.

Descriptors	MAE / ppm	R^2
(i)	0.20	0.98
(ii)	1.19	0.31
(iii)	0.30	0.95

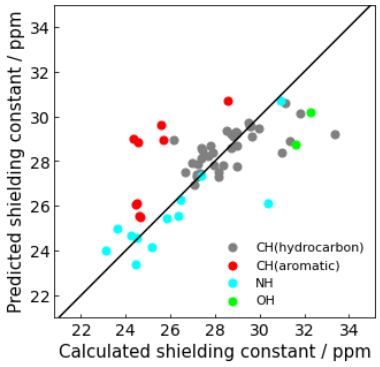

Figure 1. True vs prediction plots for ^{1}H shielding constants of chignolin.