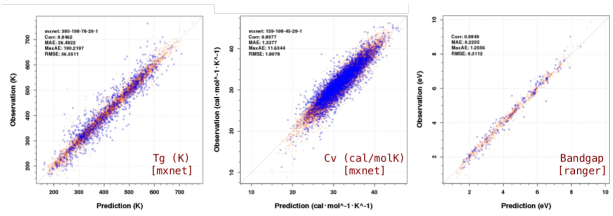
機械学習を用いた有機分子の物性値予測モデルライブラリの作成とその予測

統計数理研究所　山田　寛尚
統計数理研究所　Wu　Stephen
国立研究開発法人物質・材料研究機構　Chang　Liu
統計数理研究所　吉田　亮

[背景] 世の中には少なくとも，10^7 から 10^8 オーダーの有機分子があらゆるデータベースに登録されている．有機分子は製造，医療・医薬など様々な分野で利用されている．これら有機分子の物性もデータベースに登録されており，それら物性値は実験的な方法，計算科学的な方法を用いて得られたものである．しかしながら，物性値はすべてのデータベースに満遍なく存在するわけではなく，実験の困難さや，計算コストなどの理由により，データ数が少ないものも多くある．そこで，既に学習済みのモデルを用いて，他の領域に適応させることを可能とする方法である転移学習に注目した．本研究では，転移学習に用いる学習済みのモデルライブラリの作成とモデル性能の検証を行った．

[方法] 今回用いたデータベースは PoLyInfo（高分子，主に実験値）[1]，Polymer Genome（高分子，主に計算値）[2, 3]，QM9（小分子、計算値）[4]である．モデルはニューラルネット作成した．PoLyInfo の一部と Polymer Genome は random forest, gradient boosting, Elastic net を用いたモデルライブラリも作成した（Polymer Genome は random forest のみ）．作成したモデルの物性値は PoLyInfo より 4 種類，QM9 より 12 種類，Polymer Genome より 15 種類である．記述子は fingerprint を用いた．

[結果] PoLyInfo と Polymer Genome の各物性値について 1000 モデルずつ，QM9 の各物性値については約 500 モデル（熱容量については 1000 モデル作成）ずつ作成した．図はベストモデルのパフォーマンスを示す（3 種類抜粋）．左から PoLyInfo のガラス転移温度，小分子の熱容量，Polymer Genome のバンドギャップである．



ベストモデルのパフォーマンス

[参考文献]
[1] PoLyInfo: http://polymer.nims.go.jp
[2] Polymer Genome: https://www.polymergenome.org/
[3] T.D. Huan et al. (2016). A polymer dataset for accelerated property prediction and design, Sci. Data, 3, 160012.
[4] R. Ramakrishnan et al. (2014). Quantum chemistry structures and properties of 134 kilo molecules, Sci. Data 1, 140022.