# マテリアルズインフォマティクスにおける分子記述子の基礎

（統計数理研究所　ものづくりデータ科学研究センター）○林　慶浩
Fundamentals of molecular descriptors in materials informatics (*Data Science Center for Creative Design and Manufacturing, Institute of Statistical Mathematics*) ○ Yoshihiro Hayashi

This talk presents an outline of molecular descriptors used in the field of materials informatics. Molecular descriptors are numerical vectors that represent the characteristics of a molecule based on its chemical structure and physicochemical properties. Representative molecular descriptors are explained in an overview of their algorithms, features, and points to note when using them.

Predefined fingerprints quantify the pattern of chemical structures based on the presence/absence or frequency of predefined fragments. Some of the most famous examples are MACCS Keys [1]. While intuitive for chemists, if the defined fragment set is redundant for the target group of compounds, it results in a sparse vector representation where most elements are zero.

Enumerative fingerprints, such as Extended Connectivity Fingerprint (ECFP) [2], are a descriptor that solves this problem. ECFP counts all substructures up to the N-th proximity atoms. In ECFP, all substructures up to the N-th proximity atoms are counted, so that the fragment set is defined according to the input compound group. In addition, the atomic features are propagated to neighboring atoms during the counting process, resulting in a condensed representation of the local environment of the atoms.

In machine learning using these descriptors, features are first created, and then machine learning is applied using the features as input. On the other hand, in recent years, an approach to predict physical properties using graph-based neural networks such as Graph Convolutional neural network (GCNN), which considers chemical structures as graphs and uses the graph data as input for machine learning, has been developed [3,4]. In GCNN, a vector representation of the local environment of an atom is created by repeatedly performing convolution operations on nearby nodes (atoms) in the input graph, as shown in Figure 1. The weights for this convolution are estimated from the data. The output layer is then constructed through a neural network. In other words, the process of creating a vector of descriptors in conventional approaches is replaced by a neural network with an architecture that reflects the graph structure of the molecule, enabling learning of physical properties from chemical structures.

Detailed description will be presented in the talk on the day.

*Keywords : Materials Informatics; Molecular Descriptor; Machine Learning*

本講演では，マテリアルズインフォマティクス（MI）の分野で用いられる分子記述子について概説する．分子記述子は，その分子の特徴を化学構造や物理化学的性質に基づく数値ベクトルとして表現したものである．いくつかの代表的な分子記述子について，アルゴリズムの概要やそれぞれの特徴，使用時の注意点などを解説する．

事前定義型のフィンガープリントは，あらかじめ定義されたフラグメントの有無や頻度に基づき化学構造のパターンを数値化する．有名なものとして，MACCS Keys [1]

などがある．化学者にとって直感的である一方で，定義されたフラグメント集合が解析対象の化合物群に対して冗長な場合，ほとんどの要素が 0 である疎なベクトル表現となる．

この問題を解決する記述子として，ECFP（Extended Connectivity Fingerprint）[2] を代表とする，列挙型のフィンガープリントが挙げられる．ECFP では第 N 近接までの全ての部分構造を数え上げる．このため，入力された化合物群に応じてフラグメント集合が定義される．また，数え上げの際に，原子特徴量を隣接原子へ伝播させることで，原子の局所的な環境を縮約した表現となる．

これらの記述子を用いた機械学習では，まず特徴量を作り，その後それを入力とした機械学習を適用する．一方で，近年では化学構造をグラフと捉えて，このグラフデータをそのまま機械学習の入力とする，GCNN（Graph Convolutional neural network）といったグラフ系ニューラルネットワークを用いて物性を予測するアプローチが開発されている[3,4]．GCNN では，Figure 1 に示したように入力されたグラフ上で近いノード（原子）に対する畳み込み演算を繰り返し行うことで，原子の局所的な環境のベクトル表現とする．この畳み込みの重みはデータから推定される．その後ニューラルネットワークを通して出力層を構築する．つまり，既存アプローチにおける記述子というベクトルを作る過程を，分子のグラフ構造を反映したアーキテクチャを持つニューラルネットワークに置き換えることで，化学構造から物性の学習を可能とする．

詳細な解説は当日の講演で発表する．

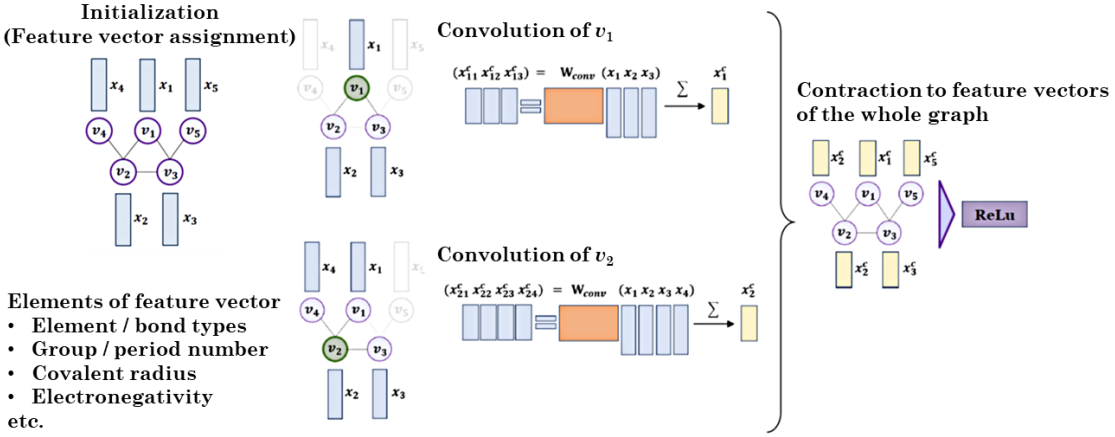


Figure 1. Computation of convolutional layer of graph neural network.

[1] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1273.
[2] D. Rogers, and M. Hahn, *J. Chem. Inf. Comput. Sci.*, **2010**, *50*, 742.
[3] D. Duvenaud, D. Maclaurin, J. A.-Iparraguirre, R. G.-Bombarelli, T. Hirzel, A. A.-Guzik, R. P. Adams, *arXiv*:1509.09292, **2015**.
[4] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, K.-R. Müller, *arXiv*:1706.08566, **2017**.