Vol.2019-BIO-57 No.11
2019/3/9

分子グラフ上の距離を考慮したグラフ畳込みニューラルネットワークによる化合物活性予測

伊井 良太[1]　柳澤 渓甫[1]　大上 雅史[1]　秋山 泰[1,a)]

概要：標的となるタンパク質に対して薬理活性を有する化合物を計算機上で発見するバーチャルスクリーニングでは，機械学習がよく用いられている．グラフ畳込みニューラルネットワークの一種である Weave module が 2016 年に Kearnes らによって提案された．Weave module は原子単体に注目した特徴 (アトム特徴) だけでなく原子ペアに着目した特徴 (ペア特徴) も用いて離れた原子の情報を取り入れられる．しかし，離れた距離にある原子ペアはグラフ上の距離が現実における立体的距離と相関するどうかは不確実である．本研究では，既存の Weave module に対して 3 つの改良手法を提案した．1 つ目は環構造内の原子に関するグラフ上の距離の修正，2 つ目はペア特徴の畳込みでグラフ上の距離によって異なる重み行列を用いること，3 つ目はペア特徴からアトム特徴に変換する際に取り込むペア特徴に対して距離による重み付けを行ったことである．実験結果より，提案手法は Weave module に対するわずかな性能向上が見られ，距離表現の工夫が化合物活性予測に有用である可能性を示した．

キーワード：グラフ畳込みニューラルネットワーク，リガンドベース・バーチャルスクリーニング，機械学習，深層学習

Graph convolutional neural networks considering distance on molecular graph for compound activity prediction

Ryota Ii[1]　Keisuke Yanagisawa[1]　Masahito Ohue[1]　Yutaka Akiyama[1,a)]

Abstract: Machine learning is often used in virtual screening that finds compounds having pharmacological activity on a target protein. Weave module is a type of graph convolutional neural networks, proposed by Kearnes *et al.* in 2016. It uses not only features focusing on atoms alone (atom features) but also features focusing on atom pairs (pair features), and can take information of non-adjacent atoms. However, the correlation between the distance on the graph and the 3-dimensional coordinate distance is uncertain. In this study, we proposed three improvements for modifying the weave module. First, the distances between ring atoms on the graph were modified to bring the distances on the graph closer to the coordinate distance. Second is to use different weight matrices depending on the distance on the graph in the convolution layers of pair features. The third is to use a weighted sum by distance when converting from pair features to atom features. Experimental results show the performance of the proposed method is slightly improved compared to weave module, and the improvement of distance representation might be useful for compound activity prediction.

Keywords: graph convolutional neural network, ligand-based virtual screening, machine learning, deep learning

1. 導入

医薬品の研究開発では 1 つの薬を作るのに最低でも 10

[1] 東京工業大学 情報理工学院 情報工学系，
Department of Computer Science, School of Computing, Tokyo Institute of Technology
[a)] akiyama@c.titech.ac.jp

年以上もの年月がかかり，開発費用は数百億円から数千億円以上かかるとされている [1]．大規模な化合物ライブラリの中から創薬ターゲットとなるタンパク質に対して活性を持つ化合物をハイスループットスクリーニングによって選別する手法が普及しているが [2]，膨大な数の化合物をスクリーニングするには大きなコストがかかる．そこで，計算機を使って効率よく活性を持つ化合物（ヒット化合物）を予測することのできる，バーチャルスクリーニングが期待されている [3]．

バーチャルスクリーニングの枠組みの 1 つとして，既知の活性情報を教師ラベルとして機械学習による予測を行うリガンドベース法がある．特に近年では，化合物の各原子をノード，結合をエッジとしたグラフとみなし，ニューラルネットワークを介して特徴抽出ができるようになった [4–6]．ここでは，グラフ構造上の畳込み演算によって畳込みニューラルネットワークを実現するグラフ畳込みニューラルネットワーク (graph convolutional neural network, GCN) が用いられる．

GCN を用いた化合物の特徴抽出では，David らの neural graph fingerprints (NGF) [4]，Han らによる GCN [5]，Kearnes らの Weave module [6] などがよく用いられる．これらは通常の Fingerprint のように一定規則に基づく化合物記述子（特徴ベクトル）を生成せず，分子構造の学習によって特徴ベクトルを柔軟に表現できるという長所がある．

GCN において，David らや Han らの手法では分子グラフ内のエッジの特徴を考慮しておらず，ノード 1 近傍の構造を学習することに焦点を当てている．対して，Kearnes らの Weave module では原子単体に注目した特徴 (アトム特徴) だけでなく，離れた原子との特徴 (ペア特徴) も用いて相互の特徴ベクトルを変換していくことで離れた原子間の特徴を取り入れることができている．しかし，Kearnes らの Weave module は化合物内の原子の組み合わせを考えたとき，ある原子から離れたペアとなる原子の数は距離ごとに異なっており，Weave module の入力となるペア特徴ではその点を考慮していない．

本研究では Kearnes らの Weave module において分子グラフ上の距離特徴を効果的に利用するために，環構造内の原子に関するグラフ上の距離の修正，ペア特徴の畳込みおよびその集約を改良することで効率的に離れた原子間の特徴を反映した GCN 手法を提案することを目的とする．

2. 先行研究：Weave module

2016 年に Kearnes らが提案した Weave module [6] の構造を**図 1** に示す．Weave module は図 1 中で①～⑦で示された 7 つの変換操作によって構成される．本研究では，初期特徴の生成方法，および変換操作③（ペア特徴から中間アトム特徴に変換する操作）を改良対象とした．これらの操作に関して以下に説明し，詳細は文献 [6] に委ねる．

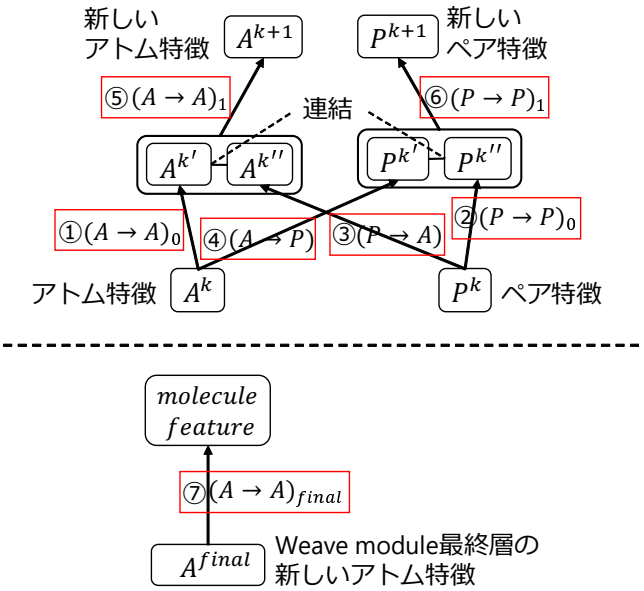

図 1 Weave module [6]
Fig. 1 Weave module [6]

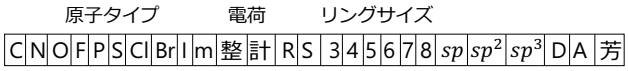

図 2 アトム特徴の入力ベクトル（d_a^0 次元）
Fig. 2 The input vector of atom features

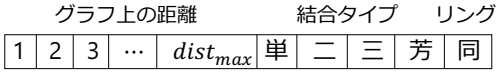

図 3 ペア特徴の入力ベクトル（d_p^0 次元）
Fig. 3 The input vector of pair features

2.1 初期特徴ベクトルの生成方法

図 1 において，ニューラルネットワークの入力に相当する初期のアトム特徴 $\boldsymbol{A}^0$ およびペア特徴 $\boldsymbol{P}^0$ は，原子タイプや結合タイプなどのグラフ構造の簡単な記述子が用いられる．これらは行列の形をしており，分子内最大原子数が n_{max} の場合，$\boldsymbol{A}^0$ のサイズは 1 つの原子に対応する d_a^0 次元特徴ベクトル（横ベクトル）を縦に n_{max} 個並べた $\boldsymbol{A}^0 \in \mathbb{R}^{n_{max} \times d_a^0}$ となり，$\boldsymbol{P}^0$ のサイズは 1 つの原子ペアに対応する d_p^0 次元特徴ベクトル（横ベクトル）を縦に n_{max}^2 個並べた $\boldsymbol{P}^0 \in \mathbb{R}^{n_{max}^2 \times d_p^0}$ である．初期のアトム特徴 $\boldsymbol{A}^0$ およびペア特徴 $\boldsymbol{P}^0$（の各行）の構成を，それぞれ**図 2**，**図 3** に示す．$dist_{max}$ は原子ペアで表現する最大の距離である．

2.2 変換操作③：ペア特徴から中間アトム特徴への変換

Weave module 第 k 層目において，以下の変換操作により原子 i との原子ペア全てに対して畳込み操作を行い，それらを足し合わせることで原子 i に対する中間アトム特徴を計算する．

$$\boldsymbol{a}_i^{k''} = \sum_j f(\boldsymbol{W}_{PA}^k \boldsymbol{p}_{(i,j)}^k + \boldsymbol{b}_{PA}^k) \quad (1)$$

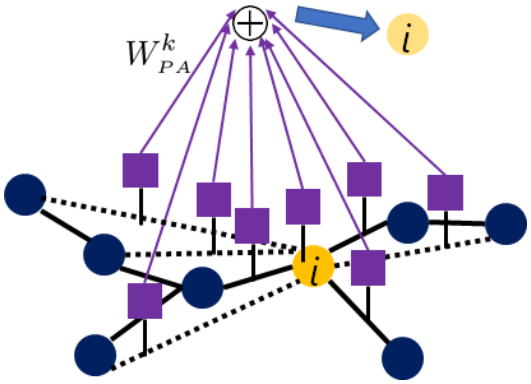

図 4　ペア特徴からアトム特徴への変換
Fig. 4　Converting from pair features to atom features

$\boldsymbol{p}^k_{(i,j)} \in \mathbb{R}^{d^k_p}$ は k 層目における原子ペア (i,j) のペア特徴入力ベクトル，$\boldsymbol{a}^{k''}_i \in \mathbb{R}^{d_{PA}}$ は原子 i のアトム特徴出力ベクトル，$\boldsymbol{W}^k_{PA} \in \mathbb{R}^{d_{PA} \times d^k_p}$ は重み行列，$\boldsymbol{b}^k_{PA} \in \mathbb{R}^{d_{PA}}$ はバイアスベクトルである．$f(\cdot)$ はベクトルの要素全てに ReLu 関数を適用する活性化関数である．この変換操作の模式図を図 4 に示す．全ての原子 $i = 1, ..., n_{max}$ に対して $\boldsymbol{a}^{k''}_i$ を求め，縦に並べたものがアトム特徴 $\boldsymbol{A}^{k''} \in \mathbb{R}^{n_{max} \times d_{PA}}$ となる．

2.3　問題点

Weave module に存在する問題点を以下に挙げる．

(1) **環構造内の原子に関するグラフ上の距離**
環構造内の原子ペア間におけるグラフ上の距離と現実の立体的距離が相関しているかの不確実性がある

(2) **ペア特徴の畳込み**
グラフ上の距離の長さに関わらずすべてのペア特徴に対して一様な重みを使用する

(3) **ペア特徴の集約**
畳込みんだペア特徴に対してペアとなる原子をすべて一様に足し込んでおり，ペア間の距離による違いが反映されない

3.　提案手法

本研究では，2.3 で挙げた Weave module の問題点 (1)〜(3) を解決する 3 つの改良について提案する（提案 1〜提案 3）．

3.1　提案 1: 環構造内の原子に関するグラフ上の距離の修正

Weave module のペア特徴では，原子ペア間における距離をグラフ上の最短経路の長さで定義する．分子内においてペアとなる原子を探索した場合，鎖状構造に比べて環構造は実際の分子配座において形状が大きく変わらないため，グラフ上の距離と物理上の距離で差があると考えた．例えば，ベンゼンの分子構造（図 5 中央）では 6 つの C–C 間の結合距離は 1.39 Å ですべて等しく，結合角はすべて 120° である．そこで，結合を考慮せずに原子同士をつないだ場合，注目原子に対してオルト位，メタ位にある原子ペアは等距離（距離 1）とし，パラ位にある原子ペアはグラフ上の距離と比べてより近くにあると考えられるため距離 3 ではなく距離 2 とした（図 6 中央）．

これを実現するため，化合物内の環構造に含まれる原子に新たにエッジを付与することで分子グラフを再定義した．本提案における分子グラフの定義を以下の **Algorithm 1** に示す．

Algorithm 1 において，GetSymmSSSR() は多環系構造の化合物において，環を構成するすべての結合を含む単環構造集合のうちで各環が最小環員数で構成される環の最小集合を取得する関数，flatten() は 1 次元配列に変換する関数，Shortest_path_length() はある頂点 r だけを固定してその頂点と HOP までの長さの最短経路長を辞書型 $dict$ で返す関数である．HOP は注目原子から探索を停止する原子までのホップ数を表す．また，items() により，$dict$ 中の各要素の探索した原子 (キー) と最短経路長 (値) を取得する．GetSymmSSSR() は RDKit ライブラリ (version 2018.03.4) [7] で，Shortest_path_length() は NetworkX ライブラリ (version 2.2) [8] でそれぞれ実装されている．

本実験では，$HOP = 2$ と定義することで環構造を有する分子グラフの環構造内のすべての頂点間で距離が $\lceil d/2 \rceil$ であるとした．ここで d は原子ペア間における最短経路の長さを表す．なお，HOP が 3 以上の場合は，Algorithm1 の 9 行目において $v = 2, 3, ..., HOP$ を全て条件に（OR で）加えることで実現できる．これは，環構造内のある頂点から距離 $2, 3, ..., HOP$ にある頂点にエッジを追加したことになる．

再定義した分子グラフに対して，すべての 2 頂点間の最短経路を同時に計算する手法であるワーシャルフロイド法を用いることでグラフの全ての頂点の間の最短経路を求めて原子ペア間の距離特徴とした．図 5 のような環構造を有するフラン (五員環)，ベンゼン (六員環)，ナフタレン (多環) を例に，2D 構造に対応して再定義したグラフを図 6 に示す．五員環の場合は星形で完全グラフ，六員環の場合は六芒星となる．ナフタレンは星型が 2 つでき，2 つの環同士を結ぶエッジも存在する．

3.2　提案 2: 異なる重みを用いたペア特徴の畳込み

ペア特徴に対する重み付けをニューラルネットワークによる学習で決定するように改良した．Weave module では，注目原子から各距離に存在する原子ペア特徴に対して，距離の長さに関わらずすべて同じ重み行列を用いてペア特徴の畳込みを行っている．そこで，各ペア特徴を区別するために

Algorithm 1 環構造上の距離の定義

Input: 分子グラフ $\mathcal{G}$
Output: 再定義した分子グラフ $\mathcal{H}$

$\mathcal{H} \leftarrow \mathcal{G}$
$sssr \leftarrow$ GetSymmSSSR($\mathcal{G}$)
$r_n \leftarrow$ flatten($sssr$)
for each vertex v in $sssr$ **do**
  for r in $ring$ **do**
    $dict \leftarrow$ Shortest_path_length($\mathcal{G}$, r, HOP)
    $list \leftarrow$ [ ]
    for k, v in $dict$.items() **do**
      if $v = HOP$ **then**
        $list$.append(k)
      end if
    end for
    $v_d \leftarrow$ set($keys$) and set($ring\_flat$)
    for a in v_d **do**
      $\mathcal{H}$.add_edge(r, a)
    end for
  end for
end for

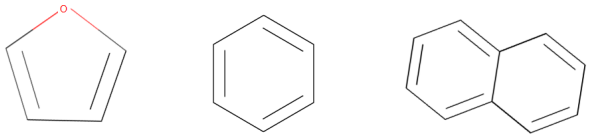

図 5　環構造の一例
Fig. 5　Examples of ring structure

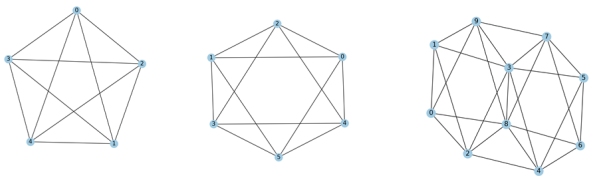

図 6　再定義した環構造の一例
Fig. 6　Examples of redefined ring structure

注目原子からの距離 $dist_0, dist_1, ..., dist_n, ..., dist_{max}, dist_\infty$ とラベル付けを行った．ここで，$dist_\infty$ は最大原子ペア距離 $dist_{max}$ を超えた距離を表す．これに対応して，ペア特徴の畳込みの際に，距離に基づいて異なる重み行列 $\boldsymbol{W}_{PA_{dist_0}}, \boldsymbol{W}_{PA_{dist_1}}, ..., \boldsymbol{W}_{PA_{dist_n}}, ..., \boldsymbol{W}_{PA_{dist_{max}}}, \boldsymbol{W}_{PA_{dist_\infty}}$ を使用するようにした．

提案 2 では，原子 i との原子ペアについて距離ごとに重み行列を使い分けて畳込みを行い，それらを足し合わせることで原子 i に対する中間アトム特徴を計算する．この変換操作を図 **7** に示す．Weave module（図 4）では原子 i のペアとなる距離 0–3 にある各原子ペアに対して，同じ重み行列 $\boldsymbol{W}_{PA}^k$ を用いていたところが，提案 2（図 7）では距離 1 の原子ペアには $\boldsymbol{W}_1$ の，距離 2 の原子ペアには $\boldsymbol{W}_2$ の重み行列を用いて畳込み演算を行い，得られたペア特徴を足し合わせることで原子 i の特徴ベクトルを更新する．

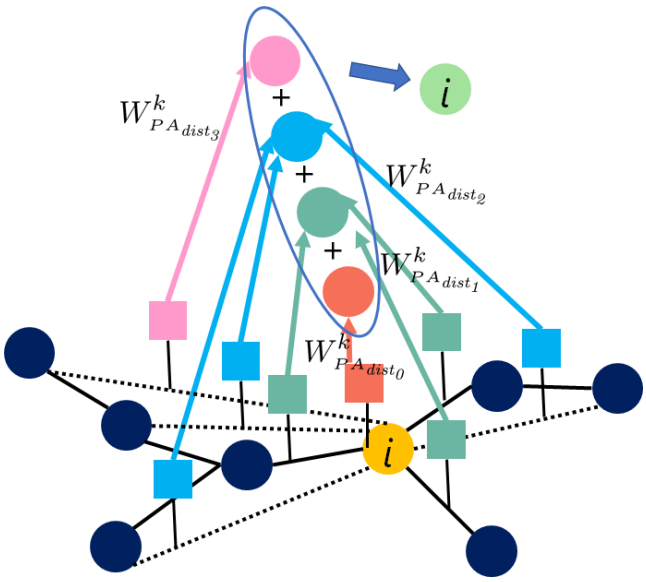

図 7　異なる重みを用いたペア特徴の畳込み
Fig. 7　Convolution of pair features using different weights

3.3　提案 3: 距離に基づくペア特徴の集約

分子グラフにおいて，ある原子からグラフ上の距離が遠くなるほど距離に対する不確実性が増すことから，原子同士の距離が遠い原子ペアは近い原子ペアに比べて重要性が低いと考えられる．そこで，提案 3 では，原子 i に対する中間アトム特徴 $\boldsymbol{a}_i^{k''}$ を求める際に，距離 d_{ij} が近いものほど大きい重み付けを行うような係数 $g(d_{ij})$ を 3 種類提案し，式 (1) を以下のように修正した．

$$\boldsymbol{a}_i^{k''} = \sum_j g(d_{ij}) f(\boldsymbol{W}_{PA}^k \boldsymbol{p}_{(i,j)}^k + \boldsymbol{b}_{PA}^k) \tag{2}$$

- ステップ関数
  注目原子から距離が $dist_{max}$ を超えたペア特徴は取り込まないような関数 $g(d) = 0 \ if \ d > dist_{max} \ else \ 1$.
- 一次関数
  注目原子から距離が 1 離れるごとにペア特徴に対する重み付けを定量的に小さくしていくような関数 $g(d) = -0.1d + 1$.
- 二次関数
  注目原子および注目原子の 1 近傍のペア特徴はそのまま取り込み，距離が遠くなるほどペア特徴に対する重み付けが距離の 2 乗で減衰する関数 $g(d) = 1/d^2$.

なお，式 (2) では提案 2 の重みの使い分けが反映されていないが，提案 2 と提案 3 は同時に用いることが可能である．

4.　評価実験

4.1　データセット

MoleculeNet [9] より，Biophysics のデータセット HIV，MUV，PCBA をそれぞれ選択した．

表 1　データセットの詳細
Table 1　Details of datasets

dataset	タスク	#Pos[*1]	#Neg[*1]	化合物数	除外数
HIV	1	1,319	39,065	40,384	743
MUV	17	489	249,397	93,087	0
PCBA	128	471,273	33,509,569	437,035	894

[*1]各タスク間で同じ化合物が異なるラベルで登録されているため，重複して数えた数を記載した．

- HIV
  4 万以上の化合物について HIV 複製を阻害する能力をテストした National Cancer Institute の抗 AIDS 剤スクリーニング [10] の結果データ．スクリーニング結果に基づき，確認された不活性 (CI)，確認された活性 (CA)，確認された中程度の活性 (CM) の 3 つのカテゴリーに分けられており，CA と CM のラベルを結合して非活性 (CI) と活性 (CA および CM) の分類ラベルとなっている．
- MUV (Maximum Unbiased Validation)
  化合物の活性評価実験の結果が収録された PubChem [11] BioAssay から収集されたデータセット．最近傍分析を適用して選別されている [12]．約 9 万の化合物に対する 17 のターゲットが含まれる．
- PCBA (PubChem BioAssay)
  PubChem BioAssay から収集されたデータセット．学習タスクは 128 種，化合物数は 3000 万以上含まれる．

分子データは SMILES 形式で提供される．本研究では RDKit [7] を使用して SMILES 形式から 2D 分子グラフに変換した．水素原子は省き，最大原子数 n_{max} を超える重原子数の化合物はデータセットから除外した．**表 1** に，各データセット内のタスク数，本研究で使用した活性化合物数，非活性化合物数，使用した化合物数，重原子数が $n_{max} = 60$ を超えた（除外された）化合物数を示す．

4.2　モデルのトレーニングと評価指標

グラフ畳込みニューラルネットワークモデルは深層学習ライブラリ Chainer Chemistry (version 0.4.0) [13] を用いて実装した．モデルのハイパーパラメータは**表 2** の通りであり，Kearnes らが使用していた値 [6] を設定した．これらのハイパーパラメータを設定したモデルについて，最大原子ペア距離 $dist_{max}$ を 1 から 5 まで検討した．

本研究の実験ではモデルの予測精度を，活性ありという予測確率の高い順に並べた化合物順序から式 (3) の ROC 曲線 [14] の曲線下面積 (AUC) および式 (4) の Enrichment Factor (EF) [15] によって評価した．

$$\mathrm{AUC} = 1 - \frac{1}{N_{\mathrm{Pos}}} \sum_{i=1}^{N_{\mathrm{Pos}}} \frac{N^i_{\mathrm{Neg}}}{N_{\mathrm{Neg}}} \tag{3}$$

$$\mathrm{EF}_{x\%} = \frac{N_{\mathrm{Pos},x\%}/N_{x\%}}{N_{\mathrm{Pos}}/N} \tag{4}$$

表 2　モデルのハイパーパラメータ
Table 2　Model hyperparameters

項目		設定値
分子内最大原子数 n_{max}		60
最大原子ペア距離 $dist_{max}$		1–5
Weave module 数 k		2
$d_{AA}, d_{PP}, d_{PA}, d_{AP}, d_A, d_P$		50
$d_{A_{final}}$		128
Fully-connected layers (層ごとのユニット数)		2000, 100
トレーニング	バッチサイズ	96
	Optimizer	Adam
	学習率	0.001
	エポック	100
train:valid:test	HIV	8 : 1 : 1
	PCBA, MUV	6 : 2 : 2
trial (試行回数) m	HIV	10
	PCBA, MUV	5

N_{Pos} は活性化合物の数，N_{Neg} は非活性化合物の数，N^i_{Neg} は i 番目の活性化合物よりも順位の高い非活性化合物の数，N は全化合物数，$N_{\mathrm{Pos},x\%}$ は上位 $x\%$ 内の活性化合物の数，$N_{x\%}$ はデータセット内の $x\%$ の化合物数（即ち $N_{x\%} = \frac{x}{100}N$）である．AUC は 0.5 でランダム，1.0 で完全正答の予測を示す．$\mathrm{EF}_{x\%}$ は化合物の順位付けによって活性化合物が上位 $x\%$ に何倍濃縮できたかを示す値となる．本研究では $\mathrm{EF}_{1\%}$ および $\mathrm{EF}_{5\%}$ を用いた．

各データセットは，表 2 に示す比率にて訓練データ (train)/検証データ (valid)/テストデータ (test) に分割した．データセット中の各タスクごとに，検証データに対して最も良い AUC が得られる epoch（学習チェックポイント）を選択し，テストデータに適用してタスクごとの平均 AUC を算出した．各タスクを $\mathcal{T}$，epoch を n，trial を $i(= 1, ..., m)$ と表すと，AUC の算出方法は以下の通りである．

$$n_{best,\mathcal{T}} = \operatorname*{argmax}_n \operatorname*{mean}_i \left(\mathrm{AUC}^{\mathrm{valid}}_{\mathcal{T},n,i}\right) \tag{5}$$

$$\mathrm{AUC} = \operatorname*{median}_{\mathcal{T}} \operatorname*{mean}_i \left(\mathrm{AUC}^{\mathrm{test}}_{\mathcal{T},n_{best,\mathcal{T}},i}\right) \tag{6}$$

ここで $\mathrm{AUC}^{\mathrm{valid}}_{\mathcal{T},n,i}$ は trial i における epoch n での訓練データによるネットワークを用いて検証データのタスク $\mathcal{T}$ を予測したときの AUC 値であり，$\mathrm{AUC}^{\mathrm{test}}_{\mathcal{T},n_{best,\mathcal{T}},i}$ は trial i における epoch $n_{best,\mathcal{T}}$ での訓練データによるネットワークを用いてテストデータのタスク $\mathcal{T}$ を予測したときの AUC 値である．各 trial i でのデータセットの分割は都度ランダムに行われる．評価指標の算出の流れを**図 8** に示す．

EF についても式 (6) と同様に算出した．

5.　実験結果

5.1　提案 1 および提案 2 の結果

まず，Weave module，提案 1，提案 2，および提案 1 と

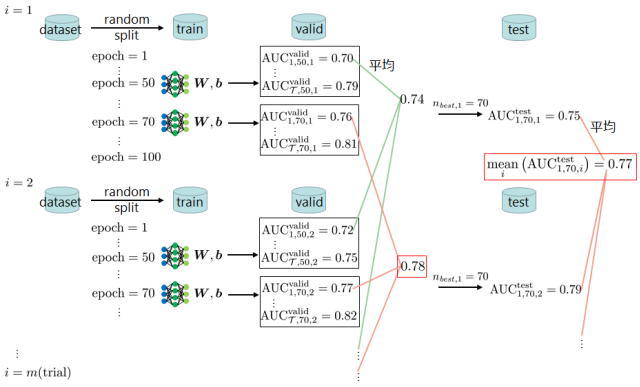

図 8　評価メトリクスの算出方法
Fig. 8　The method of calculating evaluation metrics

表 3　提案 1 および提案 2 による各データセットの AUC
Table 3　The AUC of each dataset using proposal 1 and proposal 2

dataset	model	distance				
		1	2	3	4	5
HIV	Weave	0.796	0.798	0.795	0.793	**0.801**
	提案 1	0.796	**0.803**	0.799	0.794	0.798
	提案 2	0.794	0.797	0.797	0.799	**0.806**
	提案 1+2	**0.806**	0.798	0.801	0.800	0.800
MUV	Weave	0.680	0.720	0.739	0.689	**0.743**
	提案 1	0.706	**0.783**	0.735	0.741	0.754
	提案 2	0.723	**0.738**	0.714	0.671	0.736
	提案 1+2	0.757	**0.760**	0.704	0.737	0.693
PCBA	Weave	0.822	**0.824**	0.821	0.821	0.823
	提案 1	0.821	**0.825**	0.823	0.823	0.824
	提案 2	0.822	0.821	0.820	0.822	**0.823**
	提案 1+2	0.819	0.821	**0.823**	0.822	0.821

提案 2 を合わせたモデルで，各データセットの AUC を比較した結果を**表 3** に示す．提案 1 では MUV データセットで Weave module よりも高い予測性能が得られたが，HIV および PCBA では Weave module と同等精度に留まった．提案 2 単独では Weave module とほとんど精度は変わらず，提案 1 と提案 2 の組み合わせでもわずかな精度向上に留まった．各データセットにおいて，最良の distance パラメータ（表 3 中の太字箇所）で Weave module に対して AUC に差があるかをウィルコクソンの符号順位検定により調べたところ，有意な差は得られなかった．

提案 1 について**図 9**，**図 10** より HIV データセットにおける EF の分布を確認した．$EF_{1\%}$ で distance = 2 のとき 19.2 と最も高くなった．$EF_{5\%}$ でも distance = 4, 5 がわずかに高くなり，離れた原子間の特徴が反映できたことを示唆している．しかし，MUV データセットについては，$EF_{1\%}$，$EF_{5\%}$ ともにそのような結果は見られず，Weave module とあまり変わらない性能だった．

図 11 および**図 12** は AUC 値の分布である．ただし図 12 では式 (6) における median 操作をする前のタスクごとの AUC による分布を示した．提案 1 と 2 を合わせたモデルを

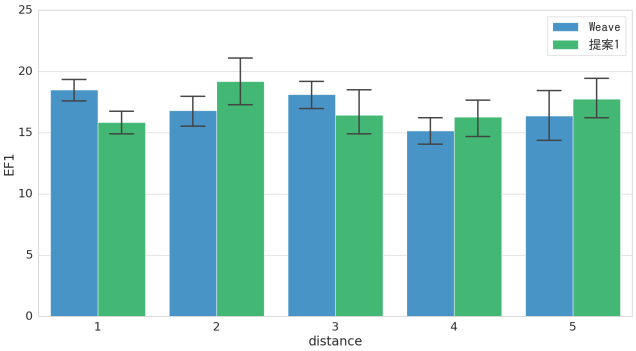

図 9　提案 1 による HIV データセットの $EF_{1\%}$
Fig. 9　The $EF_{1\%}$ of HIV dataset using proposal 1

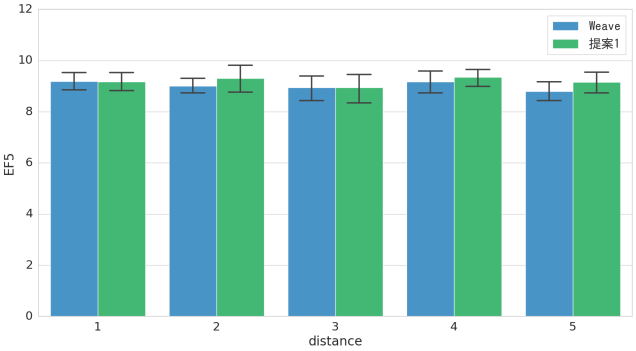

図 10　提案 1 による HIV データセットの $EF_{5\%}$
Fig. 10　The $EF_{5\%}$ of HIV dataset using proposal 1

Weave module や提案 1 と比較すると，MUV データセットの AUC では，distance = 1, 2 のような近傍の原子のみ重みを分けたことで AUC は高くなり，各タスクの AUC のばらつきが小さくなった．また，提案 1 と 2 を合わせたモデルは，HIV データセットの平均 $EF_{1\%}$ では distance = 5 のとき 18.8 と高くなり，distance = 4, 5 のときは提案 1 と比較してわずかな改善が見られた．

5.2　提案 3 の結果

ペア特徴の集約手法について Weave module と提案 3 の 3 種類の関数を組み込んだ場合において，HIV および MUV データセットに対して実験した結果を**表 4** に示す．一次関数と二次関数のモデルが，通常の加算に対し，両データセットで AUC ベースではわずかに高くなったが，Weave module に対する統計的有意差は得られなかった．また，ステップ関数は他モデルと比較して精度が低くなったことから，Weave module 層を積み重ねることで畳込みのサイズが拡大し，最大原子ペア距離を超えたペア特徴も重視して取り込まれていることが分かった．

6.　考察

6.1　グラフ上の距離を修正したことによる影響

環構造数がデータセット内の化合物数に対して 3 倍程度であったため，提案 1 によって環構造内の原子に関してグラフ上の距離を修正したことは意義があったと言える．

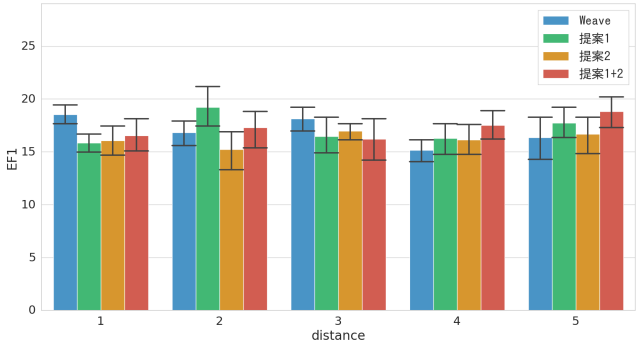

図 11　提案 2 による HIV データセットの $EF_{1\%}$
Fig. 11　The $EF_{1\%}$ of HIV dataset using proposal 2

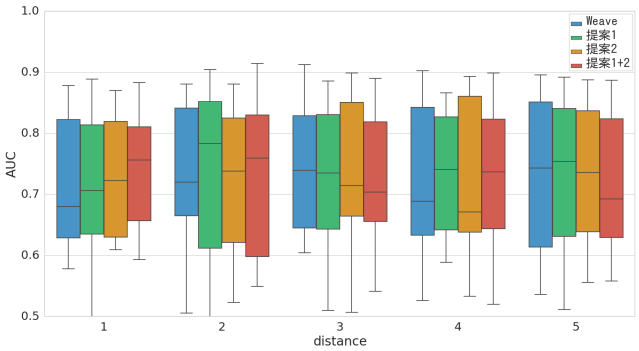

図 12　提案 2 による MUV データセットの AUC
Fig. 12　The AUC of MUV dataset using proposal 2

表 4　提案 3 による各データセットの AUC
Table 4　The AUC of each dataset using proposal 3

dataset	model	distance				
		1	2	3	4	5
HIV	Weave	0.796	0.798	0.795	0.793	**0.801**
	ステップ	0.766	0.767	0.765	0.769	**0.772**
	一次	0.799	0.798	0.803	0.799	**0.807**
	二次	0.796	0.791	**0.803**	0.798	0.803
MUV	Weave	0.680	0.720	0.739	0.689	**0.743**
	ステップ	0.629	**0.721**	0.692	0.677	0.690
	一次	0.731	**0.749**	0.687	0.713	0.729
	二次	**0.752**	0.742	0.713	0.722	0.702

本研究では環構造内の原子に着目して，共有結合に基づくのではなく，環構造内の原子間距離 d を $\lceil d/2 \rceil$ と定義してエッジを付加したことで，通常とは異なる分子構造に変換した．これにより，グラフ上の距離ほど離れていなかった原子ペアを物理上の距離に近づけることができた．更なる検討事項として，環構造内のグラフ上の距離を修正したことによって環構造は自動的に学習できていると考えられるため，ペア特徴の中に含まれる原子ペアが同じリングに所属するという特徴は省く方が良い可能性がある，さらに，本研究では環に 3 つの二重結合を持つようなベンゼン環と単結合のみを持つようなシクロヘキサンを区別していないため，環に含まれている結合の種類も考慮することで，モデルの予測性能が向上することが期待できる．また，多環の場合は，1 つ 1 つの環を区別することで直接結合されていない原子同士にはエッジを張らないようにすることなどが考えられる．

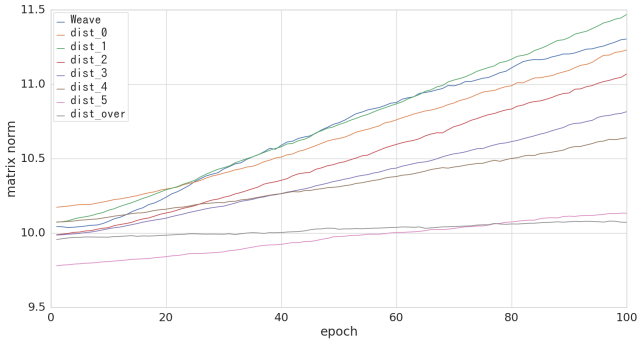

図 13　HIV における重み行列 ($\boldsymbol{W}^0_{PA}$ と $\boldsymbol{W}^0_{PA_{dist_n}}$) ノルムの比較
Fig. 13　Comparison of weight matrix norms ($\boldsymbol{W}^0_{PA}$ and $\boldsymbol{W}^0_{PA_{dist_n}}$) in HIV dataset

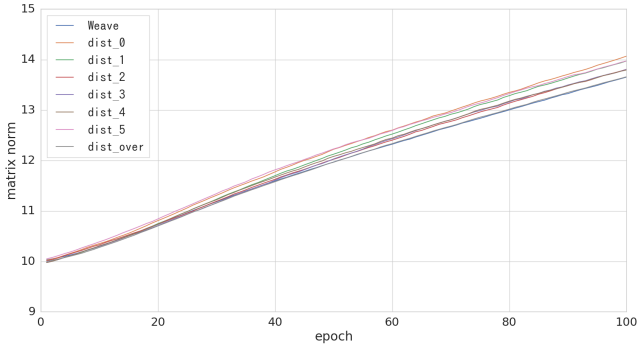

図 14　HIV における重み行列 ($\boldsymbol{W}^1_{PA}$ と $\boldsymbol{W}^1_{PA_{dist_n}}$) ノルムの比較
Fig. 14　Comparison of weight matrix norms ($\boldsymbol{W}^1_{PA}$ and $\boldsymbol{W}^1_{PA_{dist_n}}$) in HIV dataset

6.2　重み行列ノルムの推移

Weave module と提案 2 について，学習が進むにつれて，重み行列がどのように変化しているかを調査した．各重み行列 $\boldsymbol{W}$ のフロベニウスノルム $\|\boldsymbol{W}\|_F$（要素二乗和の平方根）を，HIV データセットについて，最大原子ペア距離が 5 のときのもので求めた．Weave module 第 0 層目を**図 13**，第 1 層目を**図 14** に示す．各図中の dist_n は距離 n のときの重み行列のノルム，dist_over は最大原子ペア距離 5 を超えたときの重み行列のノルムである．

図 13 より，Weave module と比較すると，dist_0，dist_1，dist_2 は傾きがおおよそ等しいが，dist_5 および dist_over は傾きが緩やかになった．Weave module 第 0 層目では離れた原子ペアは重み行列の値があまり変動していないことからあまり重要ではなく，注目原子の近傍のペア特徴をより重視して取り込んでいることがわかった．よって，近傍の原子ペア距離 0–2 とそれら以外の離れた原子ペアで重みを分けることでモデルの性能を改善できる可能性があると考えられる．図 14 より，Weave module と dist_n の傾きがおおよそ等しくなっていたことがわかる．Weave module 第 1 層目では重み行列の値が大きく変動していることから，

注目原子の近傍のペア特徴だけなく離れた原子ペア特徴も重視していることがわかった．ゆえに，Weave module 第 1 層目では距離ごとに重みを分ける必要がない可能性があり，Weave module 層が進むにつれて重み行列の構成を変える方法が有効である可能性がある．

7. まとめ

7.1 本研究の結論

本研究では，先行研究である Kearnes らの Weave module [6] におけるペア特徴からアトム特徴に変換する操作に対し，以下の改良を行った．

化合物内の環構造におけるグラフ上の距離の修正　環構造に含まれる原子ペアの距離 d を $\lceil d/2 \rceil$ とすることでグラフ上の距離を物理上の距離に近づけた．そして，最大原子ペア距離を 2 としたとき，概ね精度が高くなった．また，最大原子ペア距離を 4, 5 としたとき，離れた原子間の特徴が反映でき，分子グラフ上の距離特徴を効果的に利用できた．

グラフ上の距離ごとに異なる重みを用いたペア特徴の畳込み　畳込みの際に，距離ごとに重みを使い分けることによってモデルの一般化を試みた．化合物内の環構造におけるグラフ上の距離を修正した上で距離ごとに異なる重みで畳込み演算を行うことで，最大原子ペア距離を 1 または 2 にしたとき，概ね精度が高くなった．Weave module の第 0 層では注目原子の近傍原子群と離れた原子群で異なる重みを使用することに有用性がある可能性を見出した．

グラフ上の距離に基づくペア特徴の集約　畳込んだペア特徴を，距離に応じて勾配を付けることで注目原子近傍の原子を重要視したペア特徴の取り込みを行った．一次関数や二次関数の重みを使用してペア特徴を集約することでわずかに予測精度を改善できた．

7.2 今後の課題

今後の課題として以下の 4 点を挙げる．

- 化合物内の環構造における距離表現の更なる工夫，およびその際の環構造の区別を行う．
- 近傍の原子ペア群とそれら以外の離れた原子ペア群での重みの使い分け，およびそれを実行するにあたり，近傍と遠方の正確な境界を決定する．
- Weave module はノード自身の特徴だけでなく，ノード間の特徴も使用することによって，他のグラフ畳込みモデルとは違ったアプローチで記述子の生成段階からグラフ構造を学習する．しかし，特徴ベクトルの変換操作が複雑であることから，ペア特徴からアトム特徴への変換操作を改良するだけでは大幅な精度の改善が達成できない可能性がある．その他の変換操作においても距離特徴を活かした改良が必要な可能性がある．
- グラフ上の距離が重要となるタスクでのモデルの一般化に向けて，Weave module の他のグラフ畳込みモデルについて，ノードの特徴だけでなくノード間の特徴を新たに加えて重み分けのような畳込み演算を定義することでグラフ畳込みモデルの拡張を目指す．

謝辞　本研究の一部は，JST CREST「EBD: 次世代の年 ヨッタバイト処理に向けたエクストリームビッグデータの基盤技術」(JPMJCR1303)，JST リサーチコンプレックス推進プログラム，文部科学省 地域イノベーション・エコシステム形成プログラム，AMED 創薬等先端技術支援基盤プラットフォーム（BINDS）(JP18am0101112) の支援を受けて行われた．

参考文献

[1] Mullard A. New drug costs US $2.6 billion to develop. *Nat. Rev. Drug. Discov.*, 13(12), 877, 2014.
[2] Macarron R., Banks M. N., Bojanic D., *et al.* Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug. Discov.*, 10, 188–195, 2011.
[3] Leelananda S. P., Lindert S. Computational methods in drug discovery. *Beilstein J. Org. Chem.*, 12, 2694–2718, 2016.
[4] Duvenaud D., Maclaurin D., Aguilera-Iparraguirre J., *et al.* Convolutional networks on graphs for learning molecular fingerprints. In *Proc. NIPS*, 2215–2223, 2015.
[5] Altae-Tran H., Ramsundar B., Pappu A. S., *et al.* Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.*, 3, 283–293, 2017.
[6] Kearnes S., McCloskey K., Berndl M., *et al.* Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.*, 30(8), 595–608, 2016.
[7] Landrum G. RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org.
[8] Hagberg A. A., Schult D. A., Swart P. J. Exploring network structure, dynamics, and function using networkX. In *7th Python in Sci. Conf. (SciPy)*, 11–15, 2008.
[9] Wu Z., Ramsundar B., Feinberg E. N., *et al.* MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.*, 9, 513–530, 2018.
[10] AIDS Antiviral Screen Data. http://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data
[11] Wang Y., Xiao J., Suzek T. O., *et al.* PubChem's BioAssay database. *Nucleic Acids Res.*, 40(D1), D400–D412, 2012.
[12] Rohrer S. G., Baumann K., Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.*, 49(2), 169–184, 2009.
[13] Chainer Chemistry: A Library for Deep Learning in Biology and Chemistry. http://github.com/pfnet-research/chainer-chemistry
[14] Jain A. N., Nicholls A. Recommendations for evaluation of computational methods. *J. Comput.-Aided. Mol. Des.*, 22(3–4), 133–139, 2008
[15] Hamza A., Wei N. N., Zhan C. G. Ligand-based virtual screening approach using a new scoring function. *J. Chem. Inf. Model.*, 52, 963–974, 2012.