DEIM2022 I44-4 (day4 p23)

結晶構造を対象としたグラフニューラルネットワークにおけるグラフ識別性能の限界

河野　圭祐†　　小出　智士†　　塩川　浩昭††　　天笠　俊之††

† (株) 豊田中央研究所　〒 480–1118 愛知県長久手市横道 41-1
†† 筑波大学計算科学研究センター　〒 305–0006 茨城県つくば市天王台 1-1-1
E-mail: †{kawano,koide}@mosk.tytlabs.co.jp, ††{shiokawa,amagasa}@cs.tsukuba.ac.jp

あらまし　本研究では結晶構造に対するグラフニューラルネットワークにおけるグラフ識別性能の限界を理論的に示す．結晶構造のもつ周期性や回転，並進不変性を取り扱うために，グラフを用いて結晶構造を表現し，グラフニューラルネットワークによって物性値を予測する研究が行われている．本研究では結晶構造をグラフによって表現する段階と，グラフを入力とするグラフニューラルネットワークによって物性値を予測する段階の 2 つの段階においてグラフ識別性能が低下することを理論的に示す．また，グラフ識別性能を向上させる特徴量 (node orbits) を用いることで，物性値の予測精度がどのように変化するのかを検証する．

キーワード　Message passing neural network，WL-test，グラフ識別性能，Node orbits

1　はじめに

結晶構造は特定の原子の並びが一定の周期で繰り返す構造であり，金属や半導体といった様々な材料において出現する．結晶構造を入力とし，その結晶の様々な物性値を予測することができれば，材料開発の効率化に寄与することが期待できる．深層学習をはじめとする機械学習技術は近年様々な分野に応用されており，これを結晶構造に対する物性値予測に対しても適用することで高精度な物性値予測が行える可能性がある．一方で，結晶構造は画像などの広く機械学習が利用されている対象と比較して，以下の特異な性質を持っている．

- 無限に繰り返す周期構造を持っている．
- 3 次元方向の回転，並進に対して物性値が不変である．

これらの性質は CNN などの広く利用されている機械学習モデルでは取り扱うことが難しい．結晶構造を扱う方法としてグラフを用いて結晶データを表現し，グラフニューラルネットワークを用いて物性値の予測を行う方法が研究されており，従来の密度汎関数理論を用いた手法よりも高精度な機械学習モデルが提案されている [16].

結晶構造は原子をノード，その間の結合をエッジとみなすと，無限のノードとエッジをもつグラフ (以下では無限グラフとよぶ) として表現することができる．グラフは一般に座標などの値を持たないデータ形式であるため，回転や並進などの操作に対して不変である．また，結晶構造の無限に繰り返す周期構造を取り扱うために，繰り返しの単位 (ユニットセル) 内に存在する原子のみをノードとする有限なグラフ (商グラフ) によって結晶構造を表現する方法が提案されている [13, 15, 16].

グラフニューラルネットワークはグラフを入力とするニューラルネットワークであり，近年様々な手法が提案されている [1,2,4,5,8,10–12,17]．グラフニューラルネットワークの代表的な手法として message passing neural network (以下 MPNN) [4] があり，結晶構造に対する機械学習モデルでも利用されている [16]．MPNN は様々なタスクにおいて高い予測性能を持つことが実験的に示されている一方で，MPNN は Weisfeiler-Lehman 同型テスト (WL-test) で識別することができない 2 つの同型でないグラフに対して異なる出力を割り当てることができないことが示されている [17]．MPNN のグラフ識別性能を向上させる目的で，様々な MPNN の拡張が提案されている [1, 2, 10, 11]. 一方で，無限に繰り返す結晶構造，およびその商グラフに対応するグラフニューラルネットワークの性質は明らかになっていない．

本研究では結晶構造の物性値予測を行う MPNN について，そのグラフ識別性能の限界が一般的な MPNN よりも低いことを示し，グラフ構造特徴量を用いて予測精度を向上させることを試みる．本研究の貢献は以下である．

（1） 商グラフとして表現した時点で識別不可になってしまう異なる結晶構造が存在することを示す．

（2） WL-test よりも真にグラフ識別性能が低い relative WL-test を定義し，一般に結晶構造に対して用いられる MPNN のグラフ識別性能が relative WL-test 以下であることを証明する．

（3） MPNN のグラフ識別性能を relative WL-test や WL-test よりも向上させるためには，商グラフへ変換する前の，結晶構造そのものや無限グラフから追加の情報を取得する必要があることを理論的に示し，実際にグラフ構造特徴量を無限グラフから求め，学習に利用することで予測精度がどのように変化するのかを確認する．

2章では結晶構造とグラフについて，また MPNN について説明する．3章では結晶構造に対するグラフおよび MPNN に対する理論的な検討を行い，グラフ識別性能の向上のためのグラフ構造特徴量を用いた MPNN を提案する．4章で関連研究について述べ，5章で実験結果を示し，6章にまとめを示す．

2 準　　備

2.1 結晶構造

結晶構造は $\mathbb{R}^3$ 空間上で周期的に繰り返している原子の並びとして表現される．この原子の並びは一般に繰り返しの単位 (ユニットセル) を用いて記述される．あるユニットセル U は $U := (L, \mathcal{X}), L \in \mathbb{R}^{3\times3}, \mathcal{X} := \{(\mathbf{x}_n, a_n) \mid \mathbf{x}_n \in \mathbb{R}^3, a_n \in \mathcal{A}, n \in [\![N]\!]\}$ と表すことができる．ここで，L はユニットセルの形状を表す 3 つの 3 次元ベクトルに対応し，$\mathcal{X}$ はユニットセル 内に存在する原子を表す．n 番目の原子は原子の種類 a_n と位置 $\mathbf{x}_n$ を用いて表される．$\mathcal{A}$ は原子の種類すべてを表す集合であり，$[\![N]\!]$ は 1 から N までの整数の集合を表す．ここで，ユニットセル の大きさを $|U| = |\mathcal{X}|$ と定義する．可算無限個の原子をもつ結晶構造 $\mathcal{M}$ はユニットセル を用いて $\mathcal{M} := \left\{(\tilde{\mathbf{x}}_{n,\mathbf{z}}, a_n) \mid \mathbf{z} \in \mathbb{Z}^3, (\mathbf{x}_n, a_n) \in \mathcal{X}, n \in [\![N]\!]\right\}$ と表すことができる．ここで，$\tilde{\mathbf{x}}_{n,\mathbf{z}} = \mathbf{x}_n + L\mathbf{z}$ である．また，$(n, \mathbf{z})$ の原子の近傍の原子を $\mathcal{N}_{\mathcal{M}}(n, \mathbf{z})$ とする．具体的には例えば原子間の距離しきい値 $\delta > 0$ を用いた $\mathcal{N}_{\mathcal{M}}(n, \mathbf{z}) = \{(n', \mathbf{z}') \mid \|\tilde{\mathbf{x}}_{n,\mathbf{z}} - \tilde{\mathbf{x}}_{n',\mathbf{z}'}\|_2 \le \delta\}$ などを考えることができる．また，$|\mathcal{N}_{\mathcal{M}}(n, \mathbf{z})|$ は任意の $n \in [\![N]\!], \mathbf{z} \in \mathbb{Z}^3$ に対してたかだか有限であるとする．

1 つの結晶構造に対応するユニットセル は 1 つではなく無数に存在している．例えば，あるユニットセルがあるとき，それを縦横に複数個並べたものも，同じ結晶構造を表すユニットセルとみなせる．そこで，ある結晶構造 $\mathcal{M}$ に対応するユニットセル のうち，最小のものを $U^* = (L^*, \mathcal{X}^*)$ とする．つまり，$|U^*| \le |U|, \forall U \in \mathcal{U}(\mathcal{M})$，ここで $\mathcal{U}(\mathcal{M})$ は $\mathcal{M}$ に対応するユニットセルの集合である．最小のユニットセルはある結晶構造に対して 1 つとは限らないことに注意されたい．

2.2 結晶構造に対応する無限グラフ

結晶構造 $\mathcal{M}$ に対応する無限グラフ を $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ と記す．ここで，ノード集合は結晶構造の中にある原子に対応し，$\mathcal{V} = \{v_i \mid i \in \mathbb{Z}_+\}$ である．エッジ集合は $\mathcal{E} = \{(v_i, v_{i'}) \mid i \in \mathbb{Z}_+, i' \in \mathcal{N}_{\mathcal{G}}(i)\}$ である．ここで，$\mathcal{N}_{\mathcal{G}}(i)$ は v_i の隣接ノードのインデックス集合であり，$\mathcal{N}_{\mathcal{G}}(i) = \{p^{-1}_{\mathcal{G}\to\mathcal{M}}(n', \mathbf{z}') | (n', \mathbf{z}') \in \mathcal{N}_{\mathcal{M}}(p_{\mathcal{G}\to\mathcal{M}}(i))\}$ である．ただし，$p_{\mathcal{G}\to\mathcal{M}} : \mathbb{Z}_+ \mapsto ([\![N]\!], \mathbb{Z}^3)$ は無限グラフのノード番号 i と結晶構造における原子の番号 $(n, \mathbf{z})$ の間の全単射であり，$p^{-1}_{\mathcal{G}\to\mathcal{M}}$ は逆写像である．一般性を失わずに $1 \le i \le N$ の範囲で $p_{\mathcal{G}\to\mathcal{M}}(i) = (i, \mathbf{0})$ が成り立つとする．

無限グラフ $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ と $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ があるとき，$\mathcal{G}$ の任意の 2 頂点 $(v_i, v_{i'})$ に対して，$(v_i, v_{i'}) \in \mathcal{E} \Leftrightarrow (f(v_i), f(v_{i'})) \in \tilde{\mathcal{E}}$ となるような全単射 $f : \mathcal{V} \mapsto \tilde{\mathcal{V}}$ が存在するとき，この 2 つの無限グラフ は同型であるとし $\mathcal{G} = \tilde{\mathcal{G}}$ と記す．

2.3 結晶構造に対応する商グラフ

無限グラフ は無限のノードと無限のエッジを含んだグラフであり，直接機械学習モデルで取り扱うことは難しい．そこで，商グラフ と呼ばれるグラフを用いた結晶構造の表現が研究されている [16]．ある結晶構造 $\mathcal{M}$ に対応するユニットセル U および無限グラフ $\mathcal{G}$ があるとき，これらに対応する商グラフは $G(\mathcal{G}, U) := (V, E), V := \{v_n \mid n \in [\![N]\!]\}$, $E := \{\!\{(v_n, v_{n'}) \mid n' \in \mathcal{N}_G(n), n \in [\![N]\!]\}\!\}$ と定義される．ここで，$\{\!\{\}\!\}$ はマルチセットを表す．以後，G は有限グラフを表し，$\mathcal{G}$ は無限グラフを表す．商グラフにおける隣接ノード集合は $\mathcal{N}_G(n) = \{\!\{n' \mid (n', \mathbf{z}) \in \mathcal{N}_{\mathcal{M}}(n, \mathbf{0})\}\!\}$ と表される．任意の i に対して $\mathcal{N}_{\mathcal{M}}(i)$ がたかだか有限個であることに注意すると，ノード V とエッジ E はともにたかだか有限個である．ノード数，エッジ数がともに有限であるグラフのことを以後有限グラフと呼ぶ．

定性的には商グラフはユニットセル 内部に存在する原子のみをノードとして持つグラフである．エッジとして，ユニットセル内に存在する原子間の結合に加えて，ユニットセル 外の原子への結合を対応するユニットセル内の原子への結合に置き換えることで表現する．このため，1 組のノード間に複数のエッジが存在することが特徴である．

2 つの商グラフ $G(\mathcal{G}, U) = (V, E)$ と $\tilde{G}(\tilde{\mathcal{G}}, \tilde{U}) = (\tilde{V}, \tilde{E})$ があるとき，G の任意の 2 頂点 $(v_n, v_{n'})$ に対して，$\{(v_n, v_{n'})\} \subset E$ と $\{(f(v_n), f(v_{n'}))\} \subset \tilde{E}$ が $|\{(v_n, v_{n'})\}| = |\{(f(v_n), f(v_{n'}))\}|$ を満たす 全単射 $f : V \mapsto \tilde{V}$ が存在するとき，2 つの商グラフ は同型であるとし $G = \tilde{G}$ と記す．

2.4 WL-test

WL-test [14] は各ノードについて隣接するノードのラベルを伝搬させることによって，グラフの同型性を判定する方法である．WL-test では同型なグラフは必ず同型と判定されるが，同型でないグラフが同型であると判定されることもある．WL-test [14] の k ステップ目において，グラフ G の n 番目のノードに割り当てられるラベル (WL ラベル) を $w^{(k)}_n(G)$ とする 2 つのグラフ $G, \tilde{G}$ に対して求まるユニークな WL ラベルの集合に適当な順序を導入した順序集合を $\mathcal{W} = [\mathcal{W}_j \mid j \in [\![|\mathcal{W}|]\!]]$ とする．このとき，WL ラベルのカウントベクトル $\mathbf{c}(G)$ の要素 $c_j(G)$ は $c_j(G) = |\{v_n \mid v_n \in V, w^{(K)}_n(G) = \mathcal{W}_j\}|$ と表すことができる．WL-test は $\mathbf{c}(G)$ と $\mathbf{c}(\tilde{G})$ を比較し，これらが一致しているとき同型，そうでないとき同型でないと判定する方法である．

2.5 MPNN

MPNN [4] はグラフに対するニューラルネットワークの 1 つであり，グラフ $G = (V, E)$ の各ノード $v_n, n \in [\![N]\!]$ に対して状態 $h^{(k)}_n$ をわりあて，近傍のノードの情報を集約することで状態を更新する．ここで k は MPNN の層に対応するインデックスである．グラフ G 上のノード特徴量を $\mathbf{h}^{(1)}_n \in \mathbb{R}^{d_m}, n \in [\![N]\!]$, エッジ特徴量を $\mathbf{e}_{n,n'} \in \mathbb{R}^{d_e}, n, n' \in [\![N]\!]$ とするとき，MPNN は message, update, readout という 3 つの関数によって構成される．

$$\mathbf{m}^{(k+1)}_n = \sum_{n' \in \mathcal{N}(n)} f^{(k)}_M(\mathbf{h}^{(k)}_n, \mathbf{h}^{(k)}_{n'}, \mathbf{e}_{n,n'}) \qquad \text{(message)} \qquad (1)$$

$$\mathbf{h}^{(k+1)}_n = f^{(k)}_U(\mathbf{h}^{(k)}_n, \mathbf{m}^{(k+1)}_n) \qquad \text{(update)} \qquad (2)$$

$$\mathbf{o} = f_R(\{\mathbf{h}^{(K)}_n | n \in [\![N]\!]\}) \qquad \text{(readout)} \qquad (3)$$

ここで，$f^{(k)}_M$ は近傍のノードから情報を伝搬する message 関数であり，主に多層パーセプトロンなどが用いられる．Message 関数の入力は入力グラフのノード数に依存しない固定長になっ

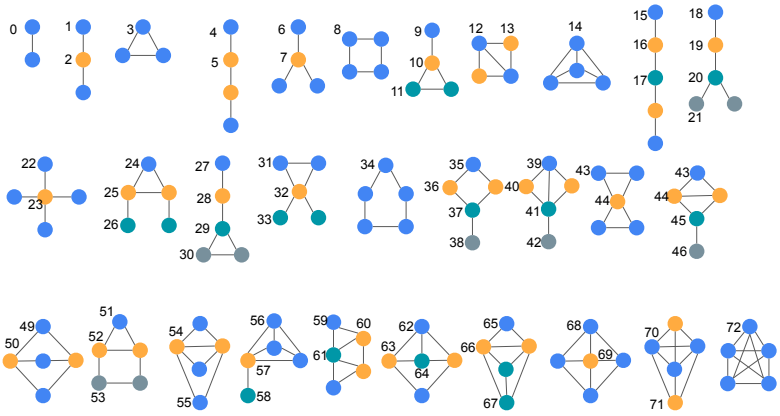

図 1　グラフ構造特徴量 (Orbits)．数字はノード orbits のインデックス ξ を表す．対称性より各グラフレットにおいて同じ色のノードは区別されないことに注意されたい．

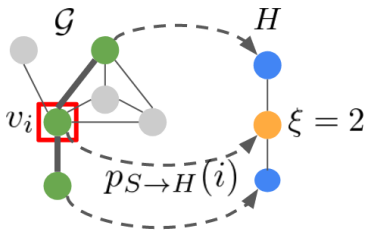

図 2　グラフ構造特徴量 (orbits) のカウント方法

ていることに注意されたい．$f_U^{(k)}$ は update 関数であり，message 関数の出力と状態 $\mathbf{h}_n^{(k)}$ から状態を更新する．ノードごとの出力ではなく，グラフに対して 1 つの出力を得たい場合，MPNN では message 関数と update 関数を K 回繰り返し適用したのち，最後に readout 関数 f_R によって状態をまとめることで出力 o を得る．ここで，f_R には mean や sum などの関数が用いられる．

2.6　CGCNN

CGCNN は結晶構造に対応する結晶構造を入力とし，形成エネルギーやバンドギャップを予測するグラフニューラルネットワークである．CGCNN では商グラフを用いて結晶グラフを表現する．このとき，結晶構造において求めた原子間距離をエッジ特徴量として入力する．つまり，エッジ特徴量は $\mathbf{e}_{n,n',\mathbf{z}'}$ は同じ $v_n, v_{n'}$ 間のエッジであったとしても異なる値になることがある．CGCNN は MPNN をベースとした手法であり，以下の 3 つの関数を用いる．

$$\mathbf{m}_n^{(k+1)} = \sum_{n'\in\mathcal{N}_G(n)} \sigma(\mathbf{y}_{n,n'}^{(k)} W_f^{(k)} + \mathbf{b}_f^{(k)}) \odot g(\mathbf{y}_{n,n'}^{(k)} W_s^{(k)} + \mathbf{b}_s^{(k)}) \quad (4)$$

$$\mathbf{h}_n^{(k+1)} = \mathbf{h}_n^{(k)} + \mathbf{m}_n^{(k+1)} \quad (5)$$

$$\mathbf{o} = \frac{1}{N} \sum_{n\in[\![N]\!]} \mathbf{h}_n^{(K)} \quad (6)$$

ここで，$\mathbf{y}_{n,n'}^{(k)} = \mathbf{h}_n^{(k)} \oplus \mathbf{h}_n'^{(k)} \oplus \mathbf{e}_{n,n',\mathbf{z}'}$ である．

2.7　グラフ構造特徴量

ノード orbits カウントはグラフ $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ 中の各ノード v_i が特定のグラフ構造の形成に何回関与したかというカウントによって，各ノードの役割を定量化するものである．$\mathcal{H}$ をサイズ k 以下のすべての連結グラフ H(グラフレット) の集合とする．ここで，ノード v_i のグラフレット H への関与の仕方はグラフレットによって複数存在することがある．例えば，長さ 3 の鎖を H とするとノード v_i が鎖の端のノードに対応する場合と中央のノードに対応する場合の 2 つの関与の仕方が存在する．本研究ではあるノードがグラフレットのどの部分と対応したかという関与の仕方ごとにカウントを行い，カウントを並べたものノード特徴量とする．以後，関与の仕方をノード orbits と呼ぶ．

図 1に最大 5 ノードで構成されるグラフレットおよび関与の仕方を示す．グラフ $\mathcal{G}$ における各ノード v_i に対応するノード orbits $\mathbf{u}_{i,\xi}$ は，グラフ $\mathcal{G}$ の v_i を含む連結な誘導部分グラフ $S \in \mathcal{S}(\mathcal{G}, i)$ とグラフレット $H \in \mathcal{H}$ の間に，ノード間の同型写像 $p_{S\to H}$ が存在するとき，以下のように定義される．

$$\mathbf{u}_{i,\xi} = |\{S \mid \xi = \mathrm{Orb}(\tilde{v}_j), \exists p_{S\to H}(i) = j, S \in \mathcal{S}(\mathcal{G}, i)\}| \quad (7)$$

ここで，$\tilde{v}_j$ は H の j 番目のノードであり，$\xi = \mathrm{Orb}(\tilde{v}_j)$ はノード orbits に対応するインデックスである．例を図 2に示す．グラフ $\mathcal{G}$ のあるノード v_i に対するノード orbits 特徴量を求める手順は以下である．(1) グラフ $\mathcal{G}$ からノード v_i を含む部分グラフ S(図 2中，緑丸) を抽出する．(2) 部分グラフ S と同型となるグラフレット H の間で同型写像 $p_{S\to H}$ を求める．(3) v_i と対応づけられたグラフレット中のノード $\tilde{v}_j$ のもつノード orbits のインデックス ξ を得て，対応するカウントを増やす．(5) v_i を含むすべての連結なサイズ k 以下の部分グラフ S について，(2), (3) を実行する．

3　結晶構造のための MPNN に対する理論的検討および無限グラフにおけるグラフ構造特徴量

CGCNN などでも用いられている mean 関数を Readout 関数として用いる MPNN のグラフ識別性能が，WL-test よりも真に悪いことを示す．次に，グラフ識別性能を向上させるためには，商グラフに変換する前の無限グラフや結晶構造そのものから，特定の条件を満たす情報を抽出し，MPNN に入力する必要があることを示す．最後に実際に条件を満たす特徴量であるグラフ構造特徴量 (orbits) を無限グラフに対して求め，MPNN に入力する方法について説明する．

3.1　Relative WL-test

証明のために，WL ラベルの相対頻度 (頻度ベクトル)

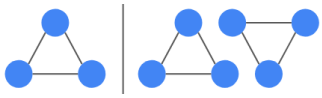

図 3　WL-test では識別可能な一方で，relative WL-test では識別不可能な 2 つのグラフの例

$\mathbf{r}(G) = \frac{1}{N}\mathbf{c}(G)$ を比較する relative WL-test を導入する．relative WL-test は 2 つのグラフ $G, \tilde{G}$ に対して，$\mathbf{r}(G)$ と $\mathbf{r}(\tilde{G})$ を比較し，これらが一致しているとき同型，そうでないとき同型でないと判定する方法である．以下の命題は有限グラフに対する relative WL-test のグラフ識別性能が WL-test よりも真に低いことを示している．ただし，WL-test は 2 つの同型なグラフを必ず同型であると判定することに注意されたい．

Proposition 1. 任意の有限グラフのペア G と $\tilde{G}$ について，*WL-test* で同型と判定されるならば，*relative WL-test* でも同型と判定される．

Proof. G と $\tilde{G}$ が WL-test で同型と判定されるとき，$\mathbf{c}(G) = \mathbf{c}(\tilde{G})$ が成り立つ．また，WL-test で同型と判定される 2 つの有限グラフのノード数は等しい．よって，頻度ベクトルの定義より $\mathbf{r}(G) = \mathbf{r}(\tilde{G})$ が成り立つ． □

Proposition 2. *WL-test* で同型でないと判定されるが，*Relative WL-test* では同型であると判定されるような同型でない有限グラフのペア G，$\tilde{G}$ が存在する．

このようなグラフの例を図3 に示す．ノード数が異なる 2 つのグラフは WL-test では同型でないと判定されるが，relative WL-test では WL ラベルの割合を比較するため，このような有限グラフを同型と判定してしまう．

3.2　商グラフ を入力とする MPNN のグラフ識別性能

本節では，無限グラフを商グラフへと変換する段階 (Proposition 3) と readout 関数を用いる MPNN を用いる段階 (Proposition 4) の 2 段階で，グラフ識別能力が低下していることを示し，それらのグラフ識別性能の低下に対処するための追加の特徴量に必要な条件を明らかにする．

Proposition 3. 商グラフ $G, \tilde{G}$ が同型であっても，対応する無限グラフ $\mathcal{G}$ と $\tilde{\mathcal{G}}$ が同型であるとは限らない．

図 4 にこのようなグラフの組の例を示す．

Proposition 4. *Readout* 関数に *mean* を用いる *MPNN* $\psi : \{G\} \mapsto \mathbb{R}$ のグラフ識別性能は *relative WL-test* 以下である．

Proof. Relative WL-test で同型と判定される 2 つのグラフ $G(V, E)$ と $\tilde{G}(\tilde{V}, \tilde{E})$ について考える．ここで，MPNN は mean を readout 関数として用いているので $\psi(G) = \frac{1}{N}\sum_{n=1}^{N} h_n^{(K)}$ である．[17] の Lemma 2 の証明より，MPNN の隠れ状態は WL ラベル $w_n^{(K)}(G)$ と写像 ϕ を用いて $h_n^{(K)} = \phi(w_n^{(K)}(G))$ と書ける．よって，$\psi(G) = \frac{1}{N}\sum_{n=1}^{N} \phi(w_n^{(K)}(G)) = \sum_{j=1}^{|\mathcal{W}|} r_j(G)\phi(\mathcal{W}_j)$ である．

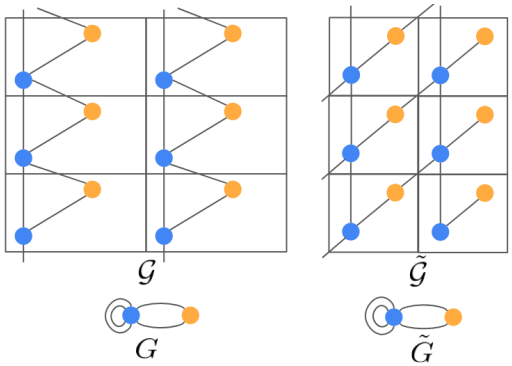

図 4　商グラフ $G, \tilde{G}$ が同型であるにも関わらず，対応する無限グラフ $\mathcal{G}$ と $\tilde{\mathcal{G}}$ が異なる例．ここでは簡単のため二次元の格子を考える．四角の枠はユニットセルに対応する．

仮定より $\mathbf{r}(G) = \mathbf{r}(\tilde{G})$ が成り立つので，$\mathcal{W}$ が $G, \tilde{G}$ で共通であることに注意すると $\psi(G) = \psi(\tilde{G})$ である．ゆえに relative WL-test で同型と判定される 2 つのグラフに対する ψ の出力は等しい． □

本研究ではこれらのグラフ識別性能の低下を抑えるために，最初のグラフ識別性能の低下が起こる前の無限グラフからグラフ識別に貢献する特徴量を抽出し，対応する商グラフに対する MPNN へ入力することを提案する．抽出する特徴量を選ぶために以下の Lemma, Proposition を示す．

Lemma 1. 無限グラフ $\mathcal{G}$ において，ある組 $p_{\mathcal{G}\to\mathcal{M}}, n \in [\![N]\!], i \in \mathbb{Z}_+, \mathbf{z} \in \mathbb{Z}^3$ が $i = p^{-1}_{\mathcal{G}\to\mathcal{M}}(n, \mathbf{z})$ を満たすならば，任意の k について $w_i^{(k)}(\mathcal{G}) = w_n^{(k)}(\mathcal{G})$ が成り立つ．

Proof. $k = 0$ のとき，対称性より v_n と v_i のノードラベルが同じであることに注意すると，$w_i^{(0)}(\mathcal{G}) = w_n^{(0)}(\mathcal{G})$ が成り立つ．ある k において，$i = p^{-1}_{\mathcal{G}\to\mathcal{M}}(n, \mathbf{z})$ の関係を満たす任意の (n, i) で $w_i^{(k)}(\mathcal{G}) = w_n^{(k)}(\mathcal{G})$ が成り立つとする．このとき，対称性より i の隣接ノードの WL ラベルのマルチセットは n の隣接ノードの WL ラベルのマルチセットと等しい．すなわち，$\{\!\{w_j^{(k)}(\mathcal{G}) \mid j \in \mathcal{N}_{\mathcal{G}}(i)\}\!\} = \{\!\{w_j^{(k)}(\mathcal{G}) \mid j \in \mathcal{N}_{\mathcal{G}}(n)\}\!\}$．ゆえに，$w_i^{(k+1)}(\mathcal{G}) = w_n^{(k+1)}(\mathcal{G})$ であり，帰納法から任意の k に対して $w_i^{(k)}(\mathcal{G}) = w_n^{(k)}(\mathcal{G})$ が成り立つ． □

Proposition 5. 無限グラフ $\mathcal{G}$ と対応する商グラフ G において $\mathbf{r}(\mathcal{G}) = \mathbf{r}(G)$ が成り立つ．

Proof. 無限グラフ $\mathcal{G}$ および商グラフ G において $n \in [\![N]\!]$ を満たす $k = 0$ ステップの WL ラベルについて，$p_{\mathcal{G}\to\mathcal{M}}(n) = (n, \mathbf{0})$ であることから $w_n^{(0)}(\mathcal{G}) = w_n^{(0)}(G)$ が成り立つ．また，n 番目のノードの近傍ノードはそれぞれ以下のように書ける．

$$\mathcal{N}_{\mathcal{G}}(n) = \{p^{-1}_{\mathcal{G}\to\mathcal{M}}(n', \mathbf{z}') | (n', \mathbf{z}') \in \mathcal{N}_{\mathcal{M}}(n, \mathbf{0})\} \tag{8}$$

$$\mathcal{N}_{G}(n) = \{p^{-1}_{\mathcal{G}\to\mathcal{M}}(n', \mathbf{0}) | (n', \mathbf{z}') \in \mathcal{N}_{\mathcal{M}}(n, \mathbf{0})\} \tag{9}$$

ある k ステップにおいて，$w_n^{(k)}(\mathcal{G}) = w_n^{(k)}(G)$ $(1 \le n \le N)$ が成り立つとき，Lemma 1より無限グラフ $\mathcal{G}$ における n 番目のノードの隣接ノードの WL ラベルのマルチセットは

$$
\begin{aligned}
&\{\!\{ w_j^{(k)}(\mathcal{G}) \mid j \in \mathcal{N}_{\mathcal{G}}(n) \}\!\} \\
=&\{\!\{ w_j^{(k)}(\mathcal{G}) \mid j = p^{-1}_{\mathcal{G}\to\mathcal{M}}(n', \mathbf{z}'), (n', \mathbf{z}') \in \mathcal{N}_{\mathcal{M}}(n, \mathbf{0}) \}\!\} \\
=&\{\!\{ w_{n'}^{(k)}(\mathcal{G}) \mid n' = p^{-1}_{\mathcal{G}\to\mathcal{M}}(n', \mathbf{0}), (n', \mathbf{z}') \in \mathcal{N}_{\mathcal{M}}(n, \mathbf{0}) \}\!\} \\
=&\{\!\{ w_{n'}^{(k)}(G) \mid n' \in \mathcal{N}_G(n) \}\!\}
\end{aligned}
\tag{10}
$$

である．$k+1$ ステップでの WL ラベルは k ステップでの WL ラベルと近傍の WL ラベルのマルチセットの関数であるので，$w_n^{(k+1)}(\mathcal{G}) = w_n^{(k+1)}(G)$ が成り立つ．よって，任意の k について $w_n^{(k)}(\mathcal{G}) = w_n^{(k)}(G)$ が成り立つ．

一般には無限個のノードを含む無限グラフに対して頻度ベクトルを求めることはできない．しかし，Lemma 1より，$\mathcal{G}$ では WL ラベルの相対頻度は $\{\!\{ w_n^{(k)}(G) \mid n \in [\![N]\!] \}\!\}$ の頻度をカウントしたものと等しいため，N 個のノードの WL ラベルの相対頻度を用いて頻度ベクトル $\mathbf{r}(\mathcal{G})$ を求めることができ，また，$\mathbf{r}(\mathcal{G}) = \mathbf{r}(G)$ が成り立つ． □

3.3 無限グラフ におけるグラフ構造特徴量

Relative WL-test では捉えることのできない特徴量として，本研究では無限グラフから node orbits [3] を抽出しノード特徴量として MPNN に入力する．Node orbits のうちサイクルを含むものは MPNN では捉えることのできない情報であることに注意されたい．無限グラフに含まれる無限個のノードのそれぞれに対して orbits を計算することは不可能である．一方で，結晶グラフはユニットセルを単位として繰り返す構造をもっているため，ユニットセルの中に存在する原子に対応するノードに対して node orbits を計算するだけでよい．そこで，我々はノード $v_n, n \in [\![N]\!]$ のそれぞれに対して，l−hop で到達可能なノード集合 $\mathcal{N}_l(v_n)$ を求め，v_n に対応する node orbits を求めた．

4 関 連 研 究

4.1 グラフニューラルネットワーク

グラフニューラルネットワークの先駆けとして，グラフから固定サイズのノードを抜き出して CNN に入力する方法 [12] や，グラフラプラシアンを用いる方法 [8] が提案されている．その後，個々のノードに対して近傍のノードから情報 (メッセージ) を集め，それを利用して個々のノードの状態を更新していく方法 (MPNN [4]，graph convolutional neural network [5]) が提案された．一方で，MPNN のグラフ識別性能は WL-test 以下であることが理論的に示されている [17].

MPNN のグラフ識別性能を向上させることを目的として，MPNN を拡張した手法が提案されている．k-node の部分グラフをノードとする高次グラフを入力とする higher-order graph neural network は k-WL test と同程度のグラフ識別性能を持つことが示されている [10]．また，ノード特徴量としてランダムに割り振られたインデックスを付与することで，グラフ識別性能を向上させる研究が行われている [2, 11]．これらの手法は MPNN のグラフ識別性能を向上させる一方で追加の計算量が大きい．より簡便な方法として，WL-test では捉えることができないグラフ構造特徴量を特徴量として MPNN に入力することで，グラフ識別性能を向上させる研究が行われている [1]．我々の提案する方法は，MPNN のグラフ識別性能を向上させることを試みるものであり，この点においてこれらの研究と関連している．一方で，我々の方法は結晶構造を表す無限グラフを対象にしている点がこれらの研究とは大きく異なっている．無限グラフに対して上述の方法を適用する方法は自明ではないことに注意されたい．また，商グラフや MEAN 関数を readout 関数とする MPNN と relative WL-test との関係を示したことも我々の貢献であることを強調しておく．

4.2 結晶構造，分子に対するグラフニューラルネットワーク

Crystal graph convolutional neural network (CGCNN) [16] は結晶構造に対応する商グラフを入力として，MEAN 関数を readout 関数とする MPNN によって結晶構造の物性値を予測するモデルである．CGCNN ではエッジ特徴量として結晶構造における原子間距離を用いる．この特徴量は商グラフへ変換される前の結晶構造そのものから得られる特徴量でありグラフ識別性能に寄与している．また，DimeNet [9] は原子間の結合をノードとするグラフニューラルネットワークによって，分子の原子化エネルギーなどを予測する回帰問題の予測するモデルである．原子間の結合をノードとすることは，2 ノードの部分グラフをノードとする高次グラフを入力とするグラフニューラルネットワークに対応する．我々は CGCNN のグラフ識別性能を向上させるために node orbits をノード特徴量に加える．Dimenet のように高次グラフを用いると学習，予測に必要な計算量が大きくなっていまうのに対して，我々の方法は node orbits を一度を計算してしまえば，CGCNN とほぼ同じ計算時間で学習，予測が可能である．

5 数 値 実 験

本節ではグラフ構造特徴量 (orbits) を加えて MPNN のグラフ識別性能を向上させることで，物性値予測問題の精度がどのように変化するのかを確認する．ここでは $l = 4$ とし，5 ノード以下のグラフレットに対応する orbits をノード特徴量として利用した．5 ノード以下のグラフレットに対応する orbits は 73 種類あるが，多くの特徴量を入力として追加することは，過学習の原因となり，予測モデルの汎化性能を低下させてしまうことがある．そこで，我々は relative WL-test では識別不可能なグラフの性質に対応する，木構造で表せないグラフレットに対応する orbits のみを用いる場合 (non-tree) と，サイクルに対応する orbits のみを用いる場合 (cycle)，およびベースラインであるグラフ構造特徴量を用いない場合 (none) の 3 つの設定において実験を行う．CGCNN にこれらの特徴量を追加し，モデルを学習した後，(1) formation energy (2) absolute energy (3) band gap の 3 つの物性値の予測精度を比較する．ミニバッチサイズを 256 とし，Adam [7] を用いて 200 エポック学習した．初期の学習率は 0.01 とし，100 エポック後に 0.1 倍した．モデルへの入力は CGCNN [16] に従い，具体的にはノード特徴量として原子量などを持ち，エッジ特徴量として原子間の距離をもつ商

	Formation energy ($\times 10^{-2}$)	Absolute energy ($\times 10^{-2}$)	Band gap ($\times 10^{-1}$)
None	$5.22 \pm 8.88 \times 10^{-2}$	$\mathbf{6.84} \pm 1.94 \times 10^{-1}$	$\mathbf{3.31} \pm 6.20 \times 10^{-2}$
Cycle	$\mathbf{5.20} \pm 6.64 \times 10^{-2}$	$6.95 \pm 1.54 \times 10^{-1}$	$3.34 \pm 5.24 \times 10^{-2}$
Non-tree	$5.24 \pm 9.57 \times 10^{-2}$	$6.93 \pm 1.80 \times 10^{-1}$	$3.43 \pm 6.94 \times 10^{-2}$

表 1　テストデータに対する平均絶対誤差

	Formation energy ($\times 10^{-2}$)	Absolute energy ($\times 10^{-2}$)	Band gap ($\times 10^{-1}$)
None	$4.19 \pm 1.48 \times 10^{-1}$	$5.43 \pm 1.90 \times 10^{-1}$	$1.70 \pm 1.89 \times 10^{-1}$
Cycle	$4.12 \pm 1.43 \times 10^{-1}$	$5.31 \pm 1.70 \times 10^{-1}$	$1.64 \pm 9.89 \times 10^{-2}$
Non-tree	$\mathbf{3.88} \pm 1.28 \times 10^{-1}$	$\mathbf{4.91} \pm 2.43 \times 10^{-1}$	$\mathbf{1.44} \pm 1.26 \times 10^{-1}$

表 2　学習データに対する平均絶対誤差

グラフである．出力は各物性値である．学習に用いたロスは物性値と予測の二乗誤差である．その他のモデルのパラメータは CGCNN [16] の公開されている実装[1]のデフォルト値に従った．

学習，検証，テストに用いたデータは Materials project [6] から取得した．(1)formation energy および (2)absolute energy の予測では合計 28046 のデータ，(3)band gap 予測では 16458 データを用いた．用いたデータの詳細については [16] を参考にされたい．学習データとして 80%のデータを用い，検証データとテストデータとしてそれぞれ 10%ずつのデータを用いた．200 エポックの学習期間のうち，検証データのロスが最も小さくなったときの重みパラメータを用いてテスト精度を求めた．

3 つの設定 (cycle，non-tree，none) のそれぞれについて，異なるランダムな初期値を用いてネットワークを 10 回学習した際のテストデータに対する平均絶対誤差を表 1に示す．表に示すように，今回用いたグラフ構造特徴量ではテストデータに対する予測誤差の有意な改善にはいたらなかった．テストデータに対する予測誤差が改善しなかった原因として過学習の問題が考えられる．表 2に訓練データに対する平均絶対誤差を示す．表より，訓練データに対しては，特に non-tree の設定はグラフ構造特徴量を一切用いなかった場合と比較して誤差が減少している．以上から，特徴量を追加することによって過学習が発生してしまっていることがわかる．

6 ま と め

本研究では，結晶構造に対するグラフニューラルネットワークにおけるグラフ識別性能の限界を示した．商グラフへの変換する段階および，mean 関数を readout 関数に用いた MPNN を用いる段階の 2 つの段階において，グラフ識別性能が低下していることを理論的に示した (Proposition 3および Proposition 4)．また，これらのグラフ識別性能の低下に対処するためには，relative WL-test では捉えられない情報を MPNN に伝えることが必要であることを示した．実際に relative WL-test では捉えられない情報であるグラフ構造特徴量をノード特徴量として追加して，物性値の予測精度を評価した．しかしながら，過学習が原因でテスト精度の向上にはいたらなかった．今後の課題として，正則化などのアプローチによる過学習への対処および，より適した特徴量の選択がある．

文　献

[1] Giorgos Bouritsas, Fabrizio Frasca, Stefanos Zafeiriou, and Michael M Bronstein. Improving graph neural network expressivity via subgraph isomorphism counting. *arXiv preprint arXiv:2006.09252*, 2020.

[2] George Dasoulas, Ludovic Dos Santos, Kevin Scaman, and Aladin Virmaux. Coloring graph neural networks for node disambiguation. *arXiv preprint arXiv:1912.06058*, 2019.

[3] Jerry Ray Dias. *Molecular orbital calculations using chemical graph theory*, Vol. 19. Springer, 1993.

[4] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

[5] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1025–1035, 2017.

[6] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, Vol. 1, No. 1, p. 011002, 2013.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[9] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations (ICLR)*, 2020.

[10] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 4602–4609, 2019.

[11] Ryan Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Relational pooling for graph representations. In *International Conference on Machine Learning*, pp. 4663–4673. PMLR, 2019.

[12] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pp. 2014–2023. PMLR, 2016.

[13] Stephen C Power, Igor A Baburin, and Davide M Proserpio. Isotopy classes for 3-periodic net embeddings. *Acta Crystallographica Section A: Foundations and Advances*, Vol. 76, No. 3, pp. 275–301, 2020.

[14] Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, Vol. 12, No. 9, 2011.

[15] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

[16] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, Vol. 120, No. 14, p. 145301, 2018.

[17] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

1：https://github.com/txie-93/cgcnn/tree/f42ab233c4ee0c416879d6bc2d22a264418413ad