

結晶の物性値予測における点群深層学習の応用

奥野 智也[†] 佐々木勇和[†] 原田 圭[†] 吉村 政志^{††}

[†] 大阪大学大学院情報科学研究科 〒565-0871 大阪府吹田市山田丘 1-5

^{††} 大阪大学レーザー科学研究所 〒565-0871 Japan 大阪府吹田市山田丘 2-6

E-mail: [†]{okuno.tomoya,sasaki,harada.kei}@ist.osaka-u.ac.jp, ^{††}yoshimura-m@ile.osaka-u.ac.jp

あらまし 材料の特性を理解する上でミクロな結晶構造に由来する物理的な性質（物性）を知ることは重要であり、近年は深層学習を用いて結晶データから物性値を予測する手法が研究されている。代表的な手法として、結晶データをグラフで表現しグラフ畳み込みニューラルネットワークで予測する手法が提案されているが、結晶のグラフ表現は3次元情報を一部失うという問題がある。そこで本稿では結晶格子に含まれる原子を点群として表現し、物性値予測を行うニューラルネットワークを構築する。提案手法は、代表的な点群深層学習モデルである PointNet を回帰問題に応用しており、また結晶データの任意のデータサイズに対応している。評価実験では、シミュレーション計算で作成されたデータセットを用いて物性値の予測を行い、既存手法およびシミュレーションデータ自体に含まれるの誤差との比較を行う。

キーワード マテリアルズインフォマティクス, PointNet, 深層学習, 結晶

1 はじめに

新たな機能や性質を持つ材料を開発することは、材料科学分野における基礎的な目標である。従来の研究開発プロセスは、研究者の勘や経験に依る実験設計及び性能評価により進められる。しかしながら、このようなプロセスは偶発性に依存するため時間的なコストが高い。例えば、リチウムイオンバッテリーはアイデアが考案された1970年代から実際に実用化された1990年代までの約20年間の試行錯誤の上で完成している[1]。そのため、研究開発の効率化を目的として材料科学と情報科学の融合分野であるマテリアルズインフォマティクス(MI)が注目を集めており、これまでに実験やシミュレーション計算のデータを集めたデータベースの整備や機械学習などの情報科学の技術の活用が進められている。

材料開発において材料の結晶構造や物理化学的な性質(以下、物性)、およびそれらの関係性を知ることは重要である。マテリアルズインフォマティクスの分野では、機械学習を用いて結晶構造と物性の関係をモデル化し、材料の物性を予測手法が研究されている[2]。予測対象となる性質は、マクロな物性では熱伝導率や超電導臨界温度、ミクロな物性では結晶の形成エネルギーやバンドギャップなど多岐にわたる。また、深層学習の急速な発展に伴ってニューラルネットワークを用いた手法も研究されている。ニューラルネットワークを用いる手法では特徴量はモデル内部で学習されるため、機械学習を用いた手法よりもエンドツーエンドな予測器を構築することができるという利点がある。

深層学習を用いて物性値を取り扱う代表的な手法として、Crystal Graph Convolutional Neural Network(以下、CGCNN)[3]が挙げられる。CGCNNでは結晶構造を、原子を節点、原子間距離を枝としたグラフとして表現し、深層ニュー

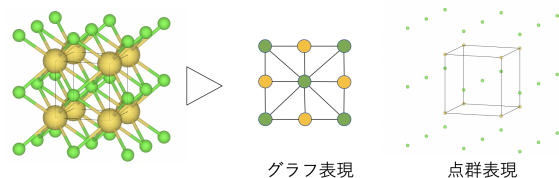


図1 結晶構造のグラフ表現と点群表現のイメージ¹

ラルネットワークの一種であるグラフ畳み込みニューラルネットワークを用いて処理を行うことで、物性値予測や結晶構造の分類を行なっている。しかしながら、CGCNNでは3次元的な結晶構造をグラフに変換したデータ表現を用いるため、原子や分子間の距離情報のみでは表現できない3次元的な配置の関係情報を一部を失っていると考えられる。グラフおよび点群によるデータ表現のイメージを図1に示す。3次元的な点群による結晶データの表現は情報の欠損が起きないため、より高い物性値予測性能が期待できる。

そこで本稿では点群を取り扱う深層学習モデルの代表的な手法である PointNet を回帰問題に応用し、結晶構造を3次元データとして直接取り扱う物性値予測手法を提案する。本手法で用いる結晶の点群表現では、結晶の最小繰り返し単位である単位格子内の各原子集合を点群とみなし、各点には原子の情報を与える。また、結晶データの点群表現を取り扱う深層学習モデルには、データの性質上の問題から次の3つの制約が生じる：(1) 各点の座標の任意の回転に関してモデルの出力が不変であること、(2) モデルへの点の入力順序に関してモデルの出力が不変であること、および(3) 取り扱う点群データのサイズが可変であることである。(1)と(2)は一般の点群データを扱う場合と同様の制約であり、本稿では点群を取り扱う深層学習モデ

²: 本画像の一部は VESTA [4] を用いて作成された

ルである PointNet を用いることによりこれらを満たす。(3) は一般の点群データとは異なる制約である。一般の点群データは 3 次元モデルの表面から任意の数の点をサンプリングすることで作成されるため基本的に固定長である。一方で各結晶の単位格子に含まれる原子数は多様であるためデータサイズが変化する。提案手法では PointNet 内の MaxPooling 層を利用し、自然な拡張を行うことで (3) の可変なデータサイズに対応する。

評価実験では、Materials Project データベース [5] からデータセット作成し、既存手法である CGCNN と物性値予測性能の比較評価を行う。その結果提案手法は既存手法である CGCNN にはおよばないが、いくつかの物性値についてはシミュレーションデータ自体に含まれる誤差以下の精度で予測可能であることを示す。

本稿の構成は以下の通りである。2 章で関連研究、3 章で事前知識、4 章で提案手法、5 章で評価実験について述べ、6 章で結論を述べる。

2 関連研究

本章では創薬および化学における深層学習を用いた物性値予測手法、材料科学における深層学習を用いた結晶の物性値予測手法、および点群を取り扱う深層学習技術について説明する。

2.1 創薬および化学における深層学習を用いた物性値予測

創薬や化学の分野において分子の性質や機能を予測することは重要であり、機械学習技術及び深層学習技術が盛んに研究されている。分子の表現には原子を節点、原子間の結合を枝とした分子グラフや、それを SMILES 記法を用いて変換した文字列表現が用いられることが多い [6]。Goh [7] らは、分子を SMILES 記法によって文字列化された分子を、再帰型ニューラルネットワークに入力し、物性値予測を行っている。また、文献 [8] [9] では、分子グラフを処理するグラフニューラルネットワークを提案し、物性値の予測を行っている。また、分子において文字列やグラフによる表現は 3 次元情報を一部損失しており、立体配座による異性体を区別できないという問題がある。そこで 3 次元情報に着目した研究も行われている。Schütt [10] らは分子に含まれる原子の位置を考慮して相互作用をモデリングした深層学習モデルを提案している。また、Samanta ら [11] は分子グラフを取り扱う深層生成モデルの入力データに原子の座標を情報に用いている。これらの創薬や化学で用いられる手法では結晶のような原子数の大きな分子ではなく比較的原子数の少ない分子が取り扱われる。

2.2 材料科学における深層学習を用いた結晶の物性値予測

材料科学の分野において結晶を取り扱うことは多く、結晶構造と物性値の関係を知ることは重要である。そのため、機械学習や深層学習を用いて結晶構造から物性予測を行う研究が盛んに行われている。Ryan ら [12] は結晶構造や物性値から作成した特徴量を入力としたニューラルネットワークによって、トポロジーに基づく潜在表現が抽出可能であることを確認した。さらに、分子をグラフで扱う深層学習モデルと同様に、

CGCNN [3] は結晶データに対してグラフ畳み込みニューラルネットワークを用いて、物性値予測及び結晶構造の分類を行っており、シミュレーションの予測誤差以内の精度でいくつかの物性値が予測可能であることを示している。また、Chen ら [13] は分子と結晶を扱うことが可能なグラフニューラルネットワークを用いた MEGNET を提案している。

一方、結晶を 3 次元データとして扱う手法も提案されている。Hoffman ら [14] はボクセルによって表現された結晶構造を、3 次元畳み込みニューラルネットワークと変分オートエンコーダを用いて処理する生成モデルを提案している。しかしながらボクセルによる表現は非常にスパースであり、計算コストが高いという問題がある。本稿で用いる点群を用いた手法では、スパースな表現にならないため、ボクセルと比較して計算コストが低い。また、特徴量設計やグラフ表現、ボクセル表現に変換する必要がないことから、提案手法はよりエンドツーエンドなモデルであるといえる。

2.3 点群を取り扱う深層学習技術

点群を扱う深層学習の研究では、3 次元モデルの表面からサンプリングされた点群の分類や、セマンティックセグメンテーションなどのタスクが盛んに研究されている。PointNet [15] は代表的なモデルであり、点群のデータ構造の特徴である回転、順序に対する不変性、及び点の間の相互作用を考慮している。さらに点群に含まれる各点の周辺情報により着目した、PointNet の改良手法が盛んに研究された。PointNet++ [16] や PointCNN [17] はその例である。Yifan ら [18] はグラフ畳み込みニューラルネットワークを用いて点群を処理する手法を提案している。また、Zaheer ら [18] は、点群に限定せず一般の集合を扱うことが可能なモデルを提案している。

3 事前知識と問題定義

本稿では結晶データから物性値を予測する回帰問題に取り組む。本章では、問題定義、および比較手法の説明を行う。

3.1 問題定義

まず本論文で用いる用語の定義を行う。

定義 1 (結晶/結晶構造) 原子や分子が空間的に規則正しく周期的に配置された物質を結晶と呼ぶ、また、その配置構造を結晶構造という。物質の多くは固体状態で結晶構造を示すため、材料開発において結晶構造は重要な指標となる。結晶構造は一般的に平行移動の繰り返し単位 (単位格子) で記述され、単位格子に含まれる原子数は結晶ごとに異なる。本稿で用いるデータは単純単位格子 (*primitive cell*) で記述されている。

定義 2 (物性値) 物質の持つ物理的な性質を物性と呼ぶ。物性には大きく分けて力学的性質、熱的性質、電気・磁氣的性質、光学的性質がある。

定義 3 (点群) コンピュータ上で扱われる点の集合、すなわち座標の集合を点群 (*point cloud*) と呼ぶ。これは物体の表面を

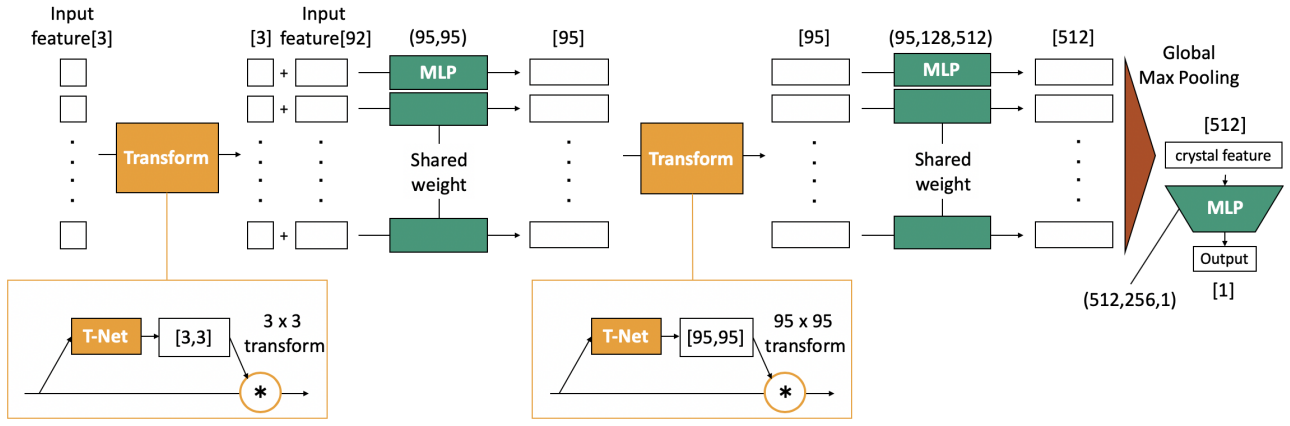


図 2 本実験で用いる Pointnet のアーキテクチャ

計測することなどで得られるデータ形式である。一般の点群はそれに含まれる点の配置関係が重要な情報であり、これはデータ生成時のサンプリング対象の物体の形状によって規定されている。一方、本研究で用いる点群は結晶データは結晶の最小繰り返し単位（単位格子）に含まれる原子の集合である。この点群について、一般の点群とは異なる点は次の 2 点である。(1) 各点、すなわち各原子の間には物理的な相互作用が働き、それらの配置関係は結晶全体のエネルギーの局所的な最小点であることである、すなわち、配置関係の背後にある各点間の関係性が一般の点群と本質的に異なる。(2) 各点の情報は原子の座標情報に加えて原子の情報を用いることである。これは結晶構造を十分に表現するためには原子の座標だけではなく、原子の種類などの情報が必要となるためである。

問題定義（結晶の物性値予測タスク） 結晶データ X が与えられたとき、それに由来する物性値 y を予測するタスクを結晶の物性値予測タスクと定義する。予測した物性値を物性値 y' とすると予測器は単純に

$$y' = f(X) \quad (1)$$

と表現できる。この予測値 y' と真値 y の差を表す誤差関数 $Error(y, y')$ を可能な限り小さくなるような予測器を構築することが本タスクの目的である。

3.2 Crystal Graph Convolutional Neural Network

比較手法となる CGCNN の説明を行う。CGCNN では結晶内の原子を節点とし、距離が近い原子間に枝を張ることでグラフを構築しモデルに入力する。また、中心となる節点情報を v_i 、 v_i の周辺にある原子の節点情報を v_j 、枝情報を $u_{(i,j)_k}$ とするとグラフ畳み込み処理は以下の式で表される。

$$\mathbf{v}_i^{(t+1)} = \mathbf{v}_i^{(t)} + \sum_{j,k} \sigma \left(z_{(i,j)_k}^{(t)} \mathbf{W}_f^{(t)} + \mathbf{b}_f^{(t)} \right) \odot g \left(z_{(i,j)_k}^{(t)} \mathbf{W}_s^{(t)} + \mathbf{b}_s^{(t)} \right) \quad (2)$$

ただし、 $z_{(i,j)_k}^{(t)} = \mathbf{v}_i^{(t)} \oplus \mathbf{v}_j^{(t)} \oplus \mathbf{u}_{(i,j)_k}$ 、 $\sigma(\cdot)$ はシグモイド関

数、 $g(\cdot)$ は Softplus などの活性化関数、 \mathbf{W} 、 \mathbf{b} はそれぞれ学習する重みとバイアス項である。畳み込み処理を行った後は、全てのノードの特徴ベクトルの総和をとったのちに、全結合層によって物性値もしくは分類ラベルを出力する。

4 提案手法

本章では、まず結晶データの性質とモデルの設計について説明する。次に、PointNet を用いた提案モデルのアーキテクチャについて説明する。

4.1 データの性質とモデルの設計

深層学習モデルを考える際、データ構造の性質によってモデルのアーキテクチャ制約が加わる。結晶内の原子の座標データ集合は 3 つのデータ構造上の性質を持つ。

- 回転不変性：結晶構造は原子の座標への任意の回転に対して不変であるため、入力される座標データの任意の回転した場合、モデルから同じ値が出力される必要がある。
- 順序普遍性：点の入力順序に対して出力される値が不変である。画像データの場合にはグリッドに固定されたデータ構造を持つため、順序が一意に決定される。一方グリッドに固定されていないデータ構造のグラフや点群に対しては入力の順序が任意に決定できるため、モデルの出力がデータのどの入力順序に対しても不変になる必要がある。
- 大きさの可変性：結晶格子に含まれる原子の数は結晶ごとに異なる。そのため、どのような数の原子が含まれる結晶に対しても処理が可能なモデルが必要になる。

回転不変性及び順序普遍性は点群深層学習で取り扱われる点群データと同様の性質であり、既存の点群深層学習モデルを用いて扱うことができる。そのため本稿では代表的な点群深層学習モデルである PointNet を用いることで回転不変性及び順序普遍性の問題を解決する。一方、データの大きさ可変性は結晶特有の課題であり、これに対処する実装方法は 2 通り考えられる。1 つ目はダミーデータを用いて結晶内の原子数をデータセット内最大値に合わせて固定長に揃える方法である。ダミーデータには同一の結晶内に含まれる原子を用いることで、Global

MaxPooling 時にダミーデータの情報を消すことができる。2 つ目は PointNet 内の Global MaxPooling 処理をサイズが異なる場合にも処理を行えるよう拡張する方法である。本手法では実装が容易であり拡張性が高いため前者の方法を用いてデータの大きさの可変性の問題に対応する。

4.2 提案モデル

提案モデルでは点群データを扱う深層学習モデルである PointNet のアーキテクチャを座標データの大きさの可変性に対応させることで、結晶に含まれる原子の座標データの処理を可能にする。提案モデルの概要を図 2 に示す。提案モデルは点群で表された結晶データを入力として、重みが共有された MLP 層、Global Max Pooling 層、および T-Net の 3 部分からなる。ここで、点群の各点には各原子の座標に加えて原子の情報を結合した特徴量を与える。まず、入力される全ての点群に対して重みを共有した MLP 処理を行う。MLP 層の重みを共有することで座標データの順序不変性に対応する。MLP 層により変換された各点の特徴量に対して点群内の各成分の最大値をとる Global Max Pooling 処理を行う。なお、Global Max Pooling 層では座標データの大きさ可変性に対応した処理を行う。Pooling 処理の結果から得られる結晶データの特徴量に対して MLP 処理を適用し、最後に出力次元数を 1 とすることで物性値の推定値とする。また、T-Net では座標データの回転不変性に対応するために座標変換行列の学習を行う。以下では、各層の詳細について説明する。

重みが共有された MLP 層：点群を取り扱うためには入力される点群の順序によらない処理を行う必要がある。PointNet では入力される全ての点に対して重みを共有した多層パーセプトロン (MLP) による処理を行うことで、入力順序に依存しない処理を行っている。 \mathbb{R} を原子の特徴量の次元を持つ空間、 x を一つの原子、 n を結晶が含む原子の個数とする。

$$f(x_1, x_2, \dots, x_n) \approx g(h(x_1), h(x_2), \dots, h(x_n)) \quad (3)$$

PointNet では点群の入力を扱う一般の関数 $f: 2^{\mathbb{R}^C} \rightarrow \mathbb{R}$ を、全ての点で共有された全結合層 $h: \mathbb{R}^C \rightarrow \mathbb{R}^K$ と、点の順序によらない処理で情報圧縮を行う関数 $g: \underbrace{\mathbb{R}^K \times \mathbb{R}^K \times \dots \times \mathbb{R}^K}_n \rightarrow \mathbb{R}$ によって近似を行う。PointNet では関数 g に Global Max Pooling が用いられる。

Global Max Pooling 層：Global Max Pooling 層では、重みが共有された MLP 層により変換された各点の特徴量を、特徴量の各成分に対して点群内での最大値をとることで情報を集約する。これにより点群の情報を点の順序に関係なく集約することができる。さらに、最大値をを求める計算は任意のデータの大きさに対して行えるため、提案手法では結晶の任意のデータのサイズに対して処理が行えるよう Global Max Pooling 層を自然に拡張している。

T-Net：T-Net では点群の入力から PointNet の重みが共有された MLP 層と Global Max Pooling 層により座標変換行列を学習の行う。T-Net から出力される変換行列は元の点群と積算される。この処理は特徴量空間にも適用が可能であり、入力す

る点群の座標とモデルの途中で得られる潜在表現に対して一度ずつこの処理を行う。

5 評価実験

PointNet の性能評価を行うために、結晶データに対して物性値の予測を行う。以下では、用いたデータセット、原子（点群の各点）の特徴量、評価方法および実験結果について説明する。

5.1 データセット

結晶構造と物性値のデータは、Materials Project [5] のデータベースから取得した。このデータはシミュレーション計算により求められたものである。データセットには結晶構造データと物性値が含まれる。予測対象となる形成エネルギーなどの物性値、および結晶構造データは第一原理計算により計算されており、それ自体に誤差が存在する [19]。

本実験においては、Materials Project のデータベースには不安定な結晶も多数含まれているため、より安定な物質へと変化の際に放出するエネルギーが 0eV の安定な結晶を収集したデータセットを用いる。また、本実験で予測する物性値は形成エネルギー (formation energy)、バンドギャップ、ポアソン比の 3 つである。物性値によって Materials Project に存在するデータ数がことなるためデータは各物性値ごとに収集し、その内テストデータを 20 % として、全てのデータが一度ずつテストデータとなるよう入れ替えながら 5 回実験を行う。各実験ではテストデータ以外の 80 % のうち 60 % を訓練データ、20 % を検証データとしてランダムに選択する。

5.2 原子の特徴量

結晶内の原子は座標に加えてそれ自体の特徴量が重要である。そのため、提案手法で用いる点の特徴量には、連続値の原子の 3 次元的な座標に加え、CGCNN で用いられている one-hot 表現の特徴量を結合した特徴量を用いる (表 1)。

表 1 原子の特徴量

特徴・性質	単位	範囲	カテゴリ数
座標 (x,y,z)	Å	-	-
族	-	1-18	18
周期	-	1-9	9
電気陰性度	-	0.5-4.0	10
共有結合半径	pm	25-250	10
価電子	-	1-12	12
電離エネルギー	eV	1.3-3.3	10
電子親和力	eV	-3-3.7	10
ブロック	-	s,p,d,f	4
原子体積	cm ³ /mol	1.5-4.3	10

5.3 評価

評価には平均絶対誤差 (MAE) を用い、PointNet と既存手法である CGCNN の物性値予測精度との比較を行う。最終的な評価指標はテストデータを入れ替えて行った 5 回の実験の平均絶対誤差の平均をとることで算出する。

表 2 物性値予測の実験結果)

物性値	単位	訓練データ数	提案手法	CGCNN	シミュレーション計算誤差 (参考値)
Formation Energy	eV/atom	20603-20604	0.064	0.046	0.081-0.136 [19]
Band gap	eV	20603-20604	0.306	0.240	0.6 [20]
Poisson ratio	-	4087-4089	0.046	0.043	-

5.4 実験結果

評価実験の結果を表 2 に示す。表 2 内の CGCNN の結果は提案手法と同じ訓練データ、検証データ、テストデータを用いた場合の実験結果であり文献 [3] の結果ではないことに注意されたい。また、シミュレーションの計算誤差は収集したデータセットに基づくものではないため、参考値としている。

実験結果は、予測した全ての物性値において CGCNN に及ばない結果となった。例えば、形成エネルギーの予測における平均絶対誤差は、提案手法では 0.064 であるのに対し、CGCNN は 0.046 となっており既存手法である CGCNN の方が高い予測性能を示している。この結果の考察については 5.5 にて議論する。次に、データそのものに含まれ得るシミュレーション計算誤差と比較すると、形成エネルギーとバンドギャップについてはシミュレーション計算誤差以下の精度で予測可能であることがわかる。ここで、シミュレーション計算誤差は実験値に対する誤差であり、提案手法の誤差はシミュレーション計算データに対する誤差であることから、単純比較はできないためシミュレーション誤差はあくまで参考値であることに注意されたい。

5.5 考察

本節では、5.4 節の評価実験結果から提案手法の問題点について考察する。

PointNet と CGCNN のアーキテクチャの大きく異なる点は局所的なコンテキストを取り込む構造を持つかどうかである。文献 [16] において、PointNet の問題点として異なるスケールでの局所的なコンテキストをうまくつかめないことが指摘されている。これは PointNet では MaxPooling 層によって一度に点群全体の情報を集約することに起因している。一方、CGCNN では各節点に対して相互作用の強さを考慮した畳み込み処理を複数回行った後に全体の情報を集約するため、各節点間で情報が伝搬し、局所的なコンテキストを取り込むことができると考えられる。したがって、評価実験における CGCNN と PointNet の予測精度の差は局所的なコンテキストを学習に取り込んでいるかどうかであると考えられる。言い換えると、物性予測では局所的なコンテキストがなくともいくつかの物性値においてシミュレーション計算誤差程度の予測が可能であるが、結晶からリッチな特徴量を抽出し高精度に物性値を予測するためには局所的なコンテキスト、すなわち結晶格子の一部に着目した局所的な情報や各原子の周辺情報を陽に取り込むニューラルネットワークのアーキテクチャが必要であるといえる。

6 結論

本稿では、点群を取り扱う深層学習モデルである PointNet を応用し結晶の物性値予測を行った。評価実験においては、既

存手法の CGCNN の性能には及ばないが、いくつかの物性値においてはシミュレーションの計算誤差以下の精度で予測可能であることを示した。

本研究で提案した深層学習モデルには改良の余地がある。まず、5.5 節で議論したように PointNet では結晶格子の一部に着目した局所的な情報や各原子の周辺情報をうまく学習に取り込めていないと考えられる。このような局所的なコンテキストを取り込むために、PointNet++ [16] で用いられているような階層的に異なるスケールで情報集約を行うアーキテクチャを用いる、もしくはグラフ畳み込みニューラルネットワークと PointNet を併用することで物性値予測性能が向上する可能性がある。また、別のアプローチとして、結晶 1 つに対し予測対象の物性値は複数存在しているため複数の性質を予測するマルチタスク学習に変えることで 1 つの物性値に寄らない汎用的な特徴量を抽出することができ、予測精度が向上する可能性がある。これらは今後の課題としたい。

また、本研究で用いた手法は結晶以外の一般の分子に対しても容易に応用が可能である。分子のデータ表現はグラフや文字列が主に用いられるが、提案手法のように 3 次元のデータ表現とそれを直接取り扱う深層学習モデルを用いることで、立体配座による異性体などの 3 次元情報が重要な分子を区別できるメリットがある。また、今後の方針として点群を取り扱う深層生成モデルを用いて所望の性質を持つ結晶を生成する問題に取り組むことは面白い試みであると思われる。既に創薬の分野では、分子のグラフ表現や文字列表現を直接生成することを目的とした深層生成モデルの応用研究が進められており、深層生成モデルは材料科学においても今後の有望な技術のひとつである。

文献

- [1] Charles Ward. Materials genome initiative for global competitiveness. 2012.
- [2] Jonathan Schmidt, Mário R. G. Marques, Silvana Botti, and Miguel A. L. Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, Vol. 5, No. 1, p. 83, 2019.
- [3] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, Vol. 120, p. 145301, 2018.
- [4] Koichi Momma and Fujio Izumi. VESTA3 for three-dimensional visualization of crystal, volumetric and morphology data. *Journal of Applied Crystallography*, Vol. 44, No. 6, pp. 1272-1276, 2011.
- [5] 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.

- [6] Daniel C. Elton, Zois Boukouvalas, Mark D. Fuge, and Peter W. Chung. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, Vol. 4, No. 4, p. 828–849, 2019.
- [7] Garrett B. Goh, Nathan O. Hodas, Charles Siegel, and Abhinav Vishnu. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties, 2017.
- [8] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Proceedings of the NIPS*, pp. 2224–2232. Curran Associates, Inc., 2015.
- [9] Peter Bjørn Jørgensen, Karsten Wedel Jacobsen, and Mikkel N. Schmidt. Neural message passing with edge updates for predicting properties of molecules and materials, 2018.
- [10] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Saucedo Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Proceedings of the NIPS*, pp. 991–1001, 2017.
- [11] Bidisha Samanta, Abir DE, Gourhari Jana, Pratim Kumar Chattaraj, Niloy Ganguly, and Manuel Gomez Rodriguez. Nevae: A deep generative model for molecular graphs. *Proceedings of the AAAI*, Vol. 33, p. 1110–1117, 2019.
- [12] Kevin Ryan, Jeff Lengyel, and Michael Shatruk. Crystal structure prediction via deep learning. *Journal of the American Chemical Society*, Vol. 140, No. 32, pp. 10158–10168, 2018. PMID: 29874459.
- [13] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, Vol. 31, No. 9, p. 3564–3572, 2019.
- [14] Jordan Hoffmann, Louis Maestrati, Yoshihide Sawada, Jian Tang, Jean Michel Sellier, and Yoshua Bengio. Data-driven approach to encoding and decoding 3-d crystal structures. *arXiv preprint arXiv:1909.00949*, 2019.
- [15] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE CVPR*. IEEE, 2017.
- [16] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Proceedings of the NIPS*, pp. 5099–5108. Curran Associates, Inc., 2017.
- [17] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Proceedings of the NIPS*, pp. 820–830, 2018.
- [18] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Proceedings of the NIPS*, editor = I. Guyon and U. V. Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett, pages = 3391–3401, year = 2017,.
- [19] Scott Kirklin, James E. Saal, Bryce Meredig, Alex Thompson, Jeff W. Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, Vol. 1, No. 1, p. 15010, 2015.
- [20] Anubhav Jain, Geoffroy Hautier, Charles J. Moore, Shyue Ping Ong, Christopher C. Fischer, Tim Mueller, Kristin A. Persson, and Gerbrand Ceder. A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science*, Vol. 50, No. 8, pp. 2295 – 2310, 2011.