結晶グラフ畳み込みニューラルネットワークによる熱膨張係数予測

石田純一[i]

Prediction of Thermal Expansion Coefficient using Crystal Graph Convolutional Neural Network

Junichi ISHIDA

近年，材料の物性を高い精度で予測できる機械学習技術としてグラフ畳み込みニューラルネットワーク(GCNN)が注目を集めている．GCNN は有機・無機を問わず幅広い化合物に適用可能なことから，数多くの実施例が報告されている．本稿では，燃料電池材料開発への適用を想定した GCNN を用いた無機化合物の体積熱膨張係数の予測と解析事例を報告する．燃料電池は主に電極，電解質材料の積層によって構成されており，これらの材料の熱膨張係数のミスマッチは燃料電池の早期劣化につながることから緻密な制御が必要となる．そこで，オープンソースの材料データベース AFLOW に格納されている 5534 個の無機化合物の熱膨張係数を取得し，結晶グラフ畳み込みニューラルネットワークを用いた学習を行うことで熱膨張係数の予測モデルの作成を試みた．本稿では計算の概要，予測結果，並びに GCNN 等の機械学習技術を用いたマテリアルズインフォマティクスの今後の展望についても概説する．

(**キーワード**): 燃料電池材料，物性予測，機械学習，マテリアルズインフォマティクス，グラフ畳み込みニューラルネットワーク

1　はじめに

材料科学分野は，実験・理論・数値計算の多角的な視点から研究が行われることで，現代に至るまで飛躍的な進歩を遂げてきた．過去 10 年ではペロブスカイト型太陽電池，鉄系超伝導体，トポロジカル絶縁体といった材料群で多くの新規化合物が報告され，それらの機能・物性の究明が著しい速度で進展した．更に，化合物情報をデータベースに蓄積し，機械学習技術を活用して材料研究を行うデータ駆動型の手法が一分野を築きつつあり，材料科学は今もなお変動期にあると言える．

材料探索の手法に焦点を当てると，計算技術の進歩もあり現在では実験とコンピュータシミュレーションを併用した材料探索が一般的に行われている．所望の機能を備えた化合物を発見するため，材料科学者は物理・化学的知見に基づいて機能性材料の候補物質を予測するが，ここに第一原理計算や分子動力学計算といった手法を援用し，電子状態や動的性質を解析することで材料探索に役立てる例が多い．また，網羅的に材料物性をシミュレーションするバーチャルハイスループットスクリーニングと呼ばれる手法を活用することで，これまで見落とされてきた新規の機能性材料が予測され，実際に合成される例も増えている．

シミュレーションを用いた材料探索を行う上でのボトルネックとなり得るのが計算コストである．第一原理計算では物性の高精度予測が可能だが，大規模系や材料の時間発展解析では計算コストが増大し材料探索を円滑に行う上での大きな制約となる．一方，現在多分野で活用が進められているニューラルネットワークを始めとした機械学習技術は，学習用

[i] サイエンスソリューション部デジタルエンジニアリングチーム　コンサルタント　博士（工学）

データの構築に係るコストは無視できないものの，学習済みモデルによる推論時の計算コストが第一原理計算に比べて低い上，物性値に関する高い予測性能を発揮し得ることから材料科学分野でも重要な研究対象と見做されている．特に材料の構造に関する情報を効率的に学習可能な手法として，グラフ畳み込みニューラルネットワーク(GCNN)と呼ばれる新しい技術が近年登場し，有機・無機問わず幅広い材料に適用されている．GCNN を始めとした機械学習技術の実際の材料探索現場への活用は現在模索段階にあると考えられるが，実験・理論・演繹的な数値計算に続く第 4 の手法として今熱い注目を集めている．

本稿では GCNN の概要を説明するとともに，具体的な燃料電池材料開発現場への適用を想定した無機化合物の熱膨張係数の予測結果を報告する．燃料電池材料開発では電極や電解質材料の積層構造の設計が極めて重要だが，それらの材料の熱膨張による格子のミスマッチは燃料電池の早期劣化につながるため緻密な制御が必要とされている．そのため高速・高精度で熱膨張係数を予測できる機械学習モデルの開発は材料探索の上で重要である．以下では予測にあたっての計算の手順や予測精度，今後の材料科学分野における機械学習の適用の方向性についても概説する．

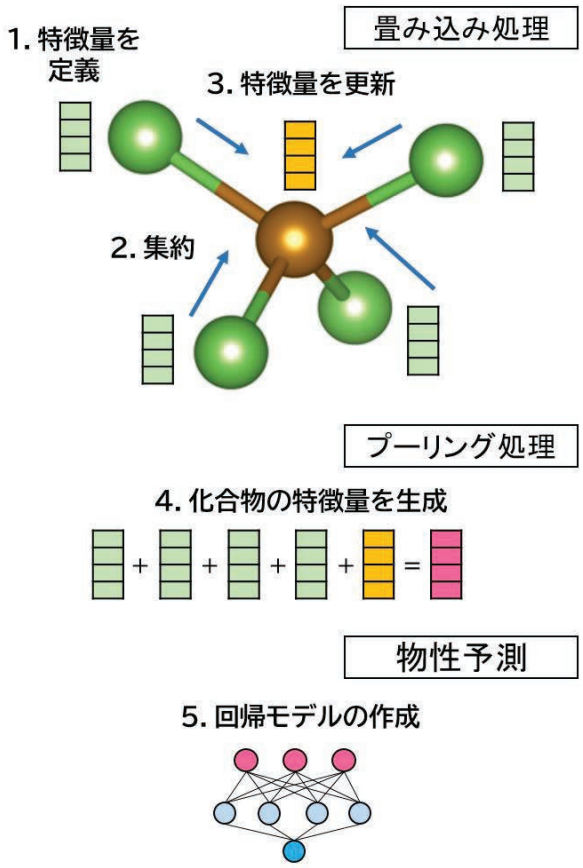

図 1　グラフ畳み込みニューラルネットワークにおける処理内容の概念図

2　グラフ畳み込みニューラルネットワーク

2.1　概要

GCNN を用いた物性予測のワークフローの概要を**図 1** に示す．GCNN を化合物の物性予測に適用する場合，構造情報をグラフ構造として表現する必要がある．グラフの頂点（ノード）は化合物の構成原子に対応しており，原子番号・イオン半径といった原子の複数の性質を表現するための特徴量が与えられる．原子同士の化学結合はグラフの辺（エッジ）に対応しており，原子間距離などの情報を含んだ特徴量が与えられる．このように定義されたノード・エッジ特徴量は特徴行列と呼ばれる行列形式で表現される．なお，一般にグラフネットワークでは解析者がノード間の結合の有無を定義するが，材料系へ適用する場合は適当なカットオフ距離を設定することでノード間の結合を定義する事例が多い．

適切な特徴行列が定義されたのち(**図 1.1**)，ある原子を取り巻く隣接した原子や化学結合の特徴量は 1 つのベクトルに集約され(**図 1.2**)，中心原子の特徴量に更新処理が施される(**図 1.3**)．集約・更新方法には様々な手法が提案されているが，特徴量の結合や多層パーセプトロンモデルによる変換操作などが行われる．GCNN における畳み込み処理ではこうした集約・更新処理を各原子に施すことで周囲の環境を組み込んだ原子の特徴量が生成される．

畳み込み処理を施された各原子の特徴量は，ベクトルの加算等によるプーリング処理によってグラフ（化合物）全体を表すベクトルに変換される(**図 1.4**)．その特徴量ベクトルに対し多層パーセプトロンによる非線形変換を施すことで，熱膨張係数等の化合物の様々な物性予測が可能となる(**図 1.5**)．

2.2　結晶グラフ畳み込みニューラルネットワーク

本稿では GCNN を結晶系の無機材料へ適用した例として知られる結晶グラフ畳み込みニューラルネットワーク（CGCNN）モデルを用いて解析を行った[1]．結晶系材料をグラフとして表現する場合，周期性を考慮する必要があるため無機材料を表すグラフは無向多重グラフとなり，結晶学的に異なる原子との結

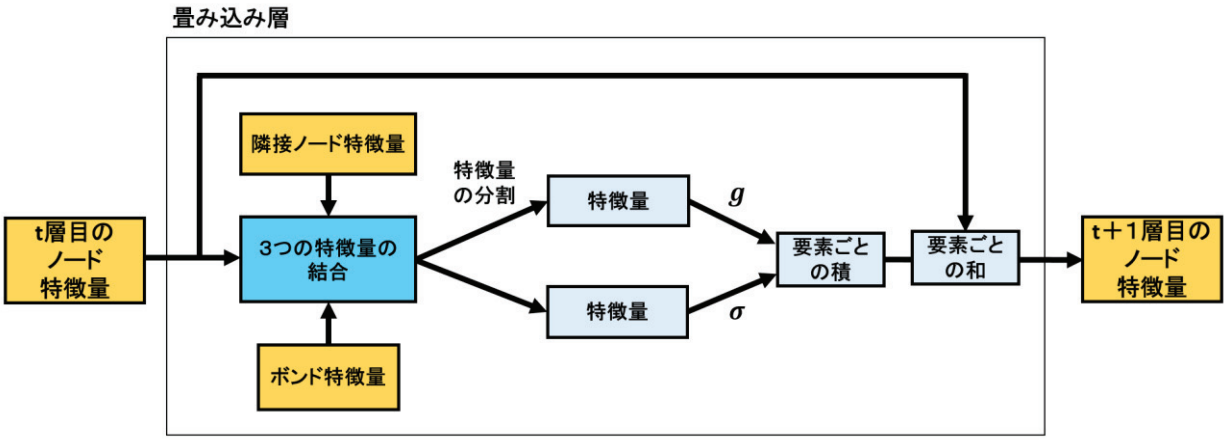

図 2　結晶グラフ畳み込みニューラルネットワークにおける畳み込み層の概念図（図 1 記載畳み込み処理に対応）

合によって分子を表すグラフが構成される．グラフ構造は結晶系材料を表す共通データフォーマットである Crystallographic Information File（CIF）の化合物構造情報を元に行列形式として変換可能であり，プログラム上で CIF の読み込みや加工を行う場合には Python 言語で記述された材料分析用オープンソースライブラリ pymatgen が有用である．

CGCNN ではある原子の特徴量をそれ自身の特徴量，配位している隣接原子の特徴量，化学結合の特徴量を集約して更新するが，これらの更新操作は以下の式で表される．

$$\boldsymbol{v}_i^{(t+1)} = \boldsymbol{v}_i^{(t)} + \sum_{j,k} \sigma(\boldsymbol{z}_{(i,j)_k}^{(t)} \boldsymbol{W}_f^{(t)} + \boldsymbol{b}_f^{(t)}) \odot g(\boldsymbol{z}_{(i,j)_k}^{(t)} \boldsymbol{W}_s^{(t)} + \boldsymbol{b}_s^{(t)}) \quad (1)$$

$$\boldsymbol{z}^{(t)}{}_{(i,j)_k} = \boldsymbol{v}_i^{(t)} \oplus \boldsymbol{v}_j^{(t)} \oplus \boldsymbol{u}^{(t)}{}_{(i,j)_k} \quad (2)$$

ここに，$\boldsymbol{v}_i^{(t)}$は畳み込み第 t 層の i 番目の原子のノード特徴量，$\boldsymbol{u}^{(t)}{}_{(i,j)k}$は i, j 番目の原子の k 番目の結合の特徴量を表す．g はソフトプラス関数などの活性化関数，σはシグモイド関数，$\boldsymbol{W}$は重み行列，$\boldsymbol{b}$ はバイアスである．また，⊙は行列の要素ごとの積，⊕は行列の結合を表す．これを可視化した畳み込み層におけるデータ処理の概念図が**図 2** であり，**図 1** に記載した畳み込み処理に対応している．**図 2** に示した畳み込み処理を複数回繰り返すことで，原子はより遠くの周辺情報を取り込みながら自己の特徴量を獲得する．なお，本モデルでは化学結合を表すエッジの特徴量は更新されず，固定したものとして扱われている．また結合したベクトルへの全結合層による処理やバッチノーマライゼーションといった処理は**図 2** では省略している．

原子の特徴量が更新された後，化合物に属する各原子の特徴量の平均を取ることで化合物全体を表すベクトル形式の特徴量を生成される．ここから更に多層パーセプトロンによる全結合層を経由することで物性値が出力される．

3　データセット

学習に必要なデータセットは，無機材料データベースとして知られる Automatic-FLOW for Materials Discovery(AFLOW)から取得した [2]．AFLOW データベースには 300 万種を超える材料の構造，電子物性，熱的性質，機械的性質などの物性値や化合物の CIF が格納されており，ブラウザ上での操作に加え一般に公開されている REST-API を用いることで登録データに機械的にアクセスできるため，大量の材料データの統計解析や機械学習を行うことが可能である [3]．本解析では REST-API を用いて化合物の CIF およびそれに対応した熱膨張係数を取得した．ここで熱膨張係数はハイスループット計算用ライブラリである Automatic Gibbs Library(AGL)によって計算された準調和振動近似を用いたシミュレーション値であり，各化合物の 300K の体積熱膨張係数(1/K)に相当している．

取得した化合物データサイズは計 5534 個となった．データセットに含まれる化合物は 1 元系 96 個，2 元系 2490 個，3 元系 2948 個となり，4 元系以上の化合物は含まれていなかった．ここから，空間群と化学式が一致している類似化合物を除外することで計 5409 個の化合物を学習対象として選定した．学習データに含まれる熱膨張係数の最大値，最小値，平均値はそれぞれ 5.67×10^{-4}, -3.17×10^{-6}, 5.70×10^{-5} となった．

4　学習の実施

学習には GitHub で公開されている CGCNN のソースコードを活用し，学習結果の解析のため一部修正を行った[4]. ハイパーパラメータはグリッドサーチ法によって決定し，畳み込み層数 M は $2 \leq M \leq 4$，バッチサイズは 2^n $(1 \leq n \leq 5)$，隠れ層の特徴量ベクトルの次元数は 2^n $(1 \leq n \leq 5)$，更新方法として ADAM による勾配降下法を採用し，学習率は 10^{-n} $(1 \leq n \leq 3)$の範囲で探索した．また訓練・検証・テストデータは 6:2:2 の割合で分割した．エポック数は全学習で 200 と統一した．結果，畳み込み層数 3，バッチサイズ 64，隠れ層特徴量ベクトル次元数 64，学習率 0.01 の条件で相関係数 R が最大化した．以下ではこの条件を用いた学習結果を示す．

5　学習結果

学習済みモデルに対するテストデータを用いた真値-予測値の散布図を**図 3 上図**に示す．予測値は 2×10^{-4} 未満の領域に集中しており，真値との高い正の相関示した(R＝0.9572)．真値が 4×10^{-4} を超える熱膨張係数の大きい材料に対しても比較的大きな値を予測していることから，良い対応関係を示していることが見て取れる．

また，学習曲線の振る舞いを**図 3 下図**に示す．学習の進展に伴い損失関数(平均二乗誤差)は振動を伴いながら訓練用・検証用データに対して共に減少傾向を示した．200 エポック終了時には訓練用・検証用データともに損失関数の減少が概ね停止している様子が見て取れた．学習曲線の振動が目立つが，学習率が大きく一回のパラメータ更新に伴う変化が大きいこと，検証用データ数が少ないためバッチ毎の損失の平均を取得した際に分散が大きくなるといった要因が推測される．

CGCNN ではプーリング処理を施すことで結晶系全体を表す特徴量ベクトル $\boldsymbol{v}_c$ が生成される．学習済みネットワークからテストデータに対する $\boldsymbol{v}_c$ を取り出し，それを t 分布型確率的近傍埋め込み法(t-sne)によって 2 次元上に次元削減してプロットした結果が**図 4** である．ここで，カラーバーはそれぞれのデータに対する熱膨張係数の予測値に対応している．図の左上には熱膨張係数が小さい物質群が位置しており，右下に進むほど熱膨張係数の大きい物質群がプロットされている様子が見て取れる．このことから

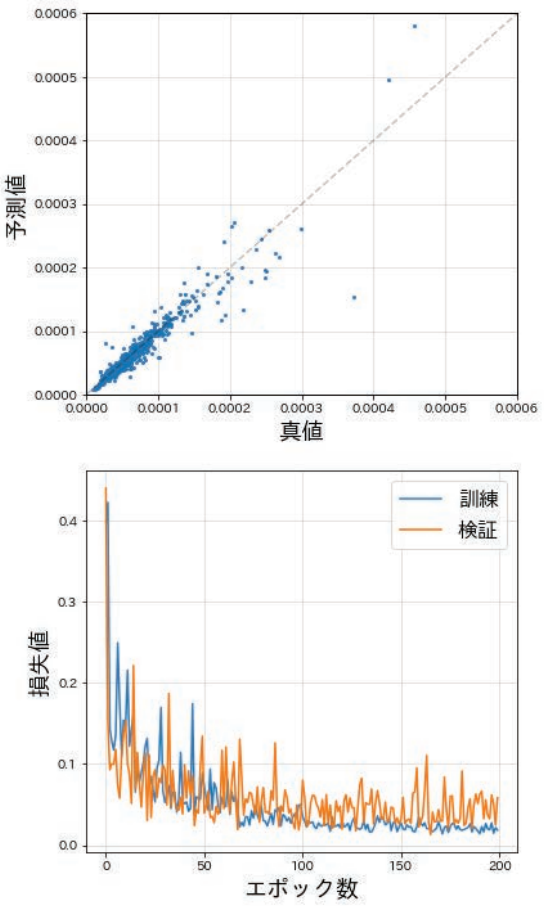

図 3　（上図）テストデータに対する真値-予測値の相関図（下図）最適化されたハイパーパラメータを用いた訓練用データと検証用データに対する学習曲線

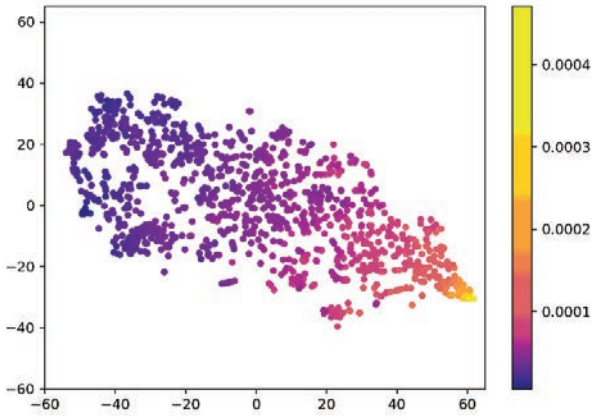

図 4　プーリング後のベクトルに対する t-sne による次元削減の結果(カラーバーは各データ点の熱膨張係数の予測値).

結晶系のベクトル $\boldsymbol{v}_c$ には結晶系と物性値を結びつける特徴量が埋め込まれており，熱膨張係数が近い物質群は類似したベクトル表現，値が大きく異なる化合物は遠く離れたベクトル表現を獲得できたことが分かった．

6　マテリアルズインフォマティクスへの GCNN 活用の展望

本稿では予測対象を熱膨張係数に絞ったが，CGCNN の報告文献では生成エンタルピーや弾性率，バンドギャップといった結晶系の他の物性に関しても高い精度で予測できることが報告されている[1]．また Facebook 社とカーネギーメロン大学が共同で開発した電極触媒データベース Open Catalyst 2020 において，原子のエネルギーや力の予測精度の検証に CGCNN が活用された[5]．ここでは原子座標を含む距離情報を連続関数であるガウス基底関数で表現することにより，エネルギーを原子位置に関して微分することで力の計算が可能となっている．力の予測精度が向上すれば計算コストの大きい第一原理分子動力学計算を機械学習で高速化できるため，触媒の緩和構造の計算が容易になるものと期待される．このように CGCNN を始めとしたグラフニューラルネットワークは材料の物性予測において非常に強力なモデルであり，材料の適応範囲も広がりを見せている．材料探索の過程においてコストの大きい計算を容易に行えないような状況下では活用の検討を行う価値は十分にあるものと考えられる．

更に，近年は単なる物性予測の手法に留まらず，分子生成器などの複雑なモデルの中で化合物情報を読み込むための 1 つのモジュールとして GCNN を活用する動きが広がっている．文献[6]では強化学習の手法の 1 つである方策勾配法によって所望の物性を備えた化合物を自動生成するグラフ畳み込み方策ネットワークモデルを提案している．GCNN を用いることで化合物情報を読み込み，原子の選択・化学結合の追加といった出力を得ることで最適な戦略決定を行っている．また文献[7]では敵対的生成ネットワーク(Generative Adversarial Network, GAN)を用いることで物性を最適化するような化合物の特徴行列・隣接行列を生成し，それを GCNN の入力データとすることで物性の予測を行っている．このような，GCNN，強化学習，GAN といった様々な手法を組み合わせた分子生成器の開発は材料分野だけでなく IT 分野からも現在盛んに研究が行われている．現時点では基礎研究レベルにあり分子生成器を用いた機能性材料の実際の開発事例は報告されていないようだが，分子生成器由来の化合物が誕生する日もそう遠くないのではないかと期待される．

7　おわりに

本稿では，燃料電池開発への適用を想定した無機材料の熱膨張係数予測をグラフ畳み込みニューラルネットワークを用いて行った．相関係数 0.95 を超える高い予測精度を示し，熱膨張係数の近い材料は近い位置に，値が大きく異なる材料は遠い位置に，各種材料の特徴量ベクトルに従って 2 次元図上にマッピングすることができた．

今回作成したモデルでは不純物等を含まない化学量論比に従う無機材料を学習対象としたが，固体酸化物燃料電池では化学ドーピングを施すことでイオンの移動経路となる欠損サイトを導入し，物性を最適化するといった操作が一般的に行われている．そのため，不純物がドーピングされた系における熱膨張係数の高精度予測は今後の重要な課題である．これを実現するためには，シミュレーションベースの学習済みモデルを構築したのち，少量の実験データに対して転移学習を実施や，シミュレーションによりドーピングされた系の熱膨張係数の教師データを作成し学習を実施するといった戦略が想定される．

引 用 文 献

1) Xie, Tian, and Jeffrey C. Grossman. "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties." *Physical review letters* 120.14 (2018): 145301.
2) Automatic-FLOW for Materials Discovery, aflow.org
3) Curtarolo, Stefano, *et al*. "AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations." *Computational Materials Science* 58 (2012): 227-235.
4) GitHub Repository, https://github.com/txie-93/cgcnn
5) Zitnick, C. Lawrence, et al. "An Introduction to Electrocatalyst Design using Machine Learning for Renewable Energy Storage." *arXiv preprint arXiv:2010.09435* (2020).
6) You, Jiaxuan, *et al.* "Graph convolutional policy network for goal-directed molecular graph generation." *Advances in neural information processing systems*. 2018.
7) De Cao, Nicola, and Thomas Kipf. "MolGAN: An implicit generative model for small molecular graphs." *arXiv preprint arXiv:1805.11973* (2018).