情報処理学会論文誌　データベース　Vol.13　No.3　22–31 (July 2020)

サーベイ論文

深層学習を用いた新物質探索に関するサーベイ

奥野 智也[1,a)]　佐々木 勇和[1]　鈴木 雄太[2]

受付日 2019年12月10日, 採録日 2020年3月14日

概要：所望の物理化学的な性質を持つ新たな物質の探索は化学，創薬，物質・材料科学などの分野において重要な課題である．従来のアプローチは研究者の勘や経験に大きく依存し，また時間的なコストが高いという問題がある．そのため，探索の効率化を目的として，機械学習やデータマイニングなどの情報科学の技術を取り入れた研究がさかんに行われている．近年では深層学習技術を用いた高精度化が進んでいる．そこで，本稿では新物質探索における深層学習技術を網羅的に調査し体系的にまとめることを目的とする．新物質の探索技術を (1) 物質構造からその性質を識別する分類と回帰技術，および (2) 性質から物質を導出する生成技術に大別し，それぞれの技術の適用分野，データの分類，および深層学習のモデルについて述べる．さらに，既存技術の制約や問題点を述べ，今後の課題を明確にする．

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

A Survey on Material Discovery by Deep Neural Networks

Tomoya Okuno[1,a)]　Yuya Sasaki[1]　Yuta Suzuki[2]

Received: December 10, 2019, Accepted: March 14, 2020

Abstract: The discovery of new materials with desired physicochemical properties is an important task in several fields such as chemistry, drug discovery, and materials science. A conventional approach depends on intuition and experience of researchers. The problem of conventional approach is very time-consuming. Therefore, for improving the efficiency of the exploration, it is actively addressed to apply informatics technology such as machine learning and data mining to material discovery. Recent developments of deep learning achieve high performance compared with conventional techniques. In this paper, we comprehensively survey deep learning techniques for material discovery and systematically summarize them. The techniques are categorized into two parts (1) classification and regression that predict properties from material structures and (2) generation that derives the materials from the property. We review application fields, data representation, and deep learning models. Finally, we discuss the constraints and problems of the existing techniques, and we clarify future challenges.

Keywords: materials informatics, chemoinformatics, deep learning

1. はじめに

所望の物理化学的な性質や機能を持つ新物質の探索は，化学，創薬，材料科学などの分野において基礎的な目標の1つである．探索の候補となる物質の空間は広大であり，たとえば，ドラッグライクな化合物の候補は，10^{23} から 10^{60} 個存在するといわれている [39]．従来の物質探索は，研究者の勘や経験による実験設計と，実験やシミュレーションによる物質の性能評価を交互に繰り返す方法をとる．しかしながら，このようなアプローチは時間的なコストが高いため探索範囲に限界がある．

新物質探索をより効率的に行うために，情報科学と物質を扱う科学分野との融合分野が注目を集めている．たとえばケモインフォマティクス [36] やマテリアルズインフォマティクス [42] がその例である．これらの分野では物質から性質を求める従来のアプローチに加えて，物質探索の効率

[1] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka university, Suita, Osaka 565–0871, Japan
[2] 総合研究大学院大学高エネルギー加速器科学研究科
School of High Energy Accelerator Science, SOKENDAI (The Graduate University for Advanced Studies), Tsukuba, Ibaraki 305–0801, Japan
[a)] okuno.tomoya@ist.osaka-u.ac.jp

化を図るために所望の性質から物質を求めるという逆方向のアプローチが研究されている．

情報科学の技術が用いられるようになった背景には 2 つの要因がある．1 つ目の要因はシミュレーションから主に得られるデータ量の増加である．第一原理計算や分子動力学法といったシミュレーションの課題は計算コストの高さであったが，計算機の処理能力向上によって，計算が容易な性質に関しては候補物質の大規模なシミュレーションが可能となり，その結果データ量が増加している．さらに実験データも含め，得られたデータは整備，公開されており，化学では PubChem [25]，材料分野では，Materials Project Database [19] や，Automatic Flow of Materials Discovery Library [8] などのデータベースが利用可能である．

2 つ目の要因は，機械学習やデータマイニング技術の性能向上である．機械学習やデータマイニング技術の応用は，ケモインフォマティクスでいち早く研究されている [7]．近年は深層学習が多く取り入れられ，活発に研究されている．特に，深層生成モデルを使った手法は，物質探索の新たな手法として重要である [48]．

深層学習技術の急速な発展にともない，化学や物質・材料科学と情報科学の融合分野においても深層学習技術が調査されている [5], [12], [34], [45], [46], [48]．一方で，新物質探索における課題の検討や深層学習モデルの比較は十分に行われていない．そこで，本稿では，機械学習を用いた新物質探索に興味がある化学や物質・材料科学の研究者，および機械学習技術の新たな応用として新物質探索に取り組む情報科学の研究者を対象とし，新物質探索における深層学習の応用研究の網羅的な調査と体系的な整理を目的とする．また，共通の課題と分野ごとの課題を明確にすることで，今後の研究の方向性を示す．

物質探索における深層学習の応用技術は大きく 2 つに分類できる．1 つ目は物質構造の分類や物質の性質予測を行う分類・回帰技術である．分類・回帰技術では，物質構造を特徴量として物質の性質を RNN や GNN を用いて学習し，未知の物質構造に対してその性質を予測する．これは物質から性質を決定する従来の実験やシミュレーションと同じプロセスである．2 つ目は物質構造を学習しそれを模倣したデータを生み出す生成技術である．生成技術では，GAN や VAE などの生成モデルを用いて物質を模倣することで，既存物質に近く所望の性質を持つ可能性が高い物質を生成する．さらに，最適化手法と併用することで物質生成の効率化が可能となる．分類・回帰技術と生成技術の比較を**図 1** に示す．物質空間は探索対象となる物質のなす空間であり，物質の性質・機能空間は物質空間と対応して性質の大小の分布を表す空間（物性値の大小を色の濃淡で表現）である．分類・回帰技術は，物質空間から性質・機能空間への射影であり，物質から物性値を導き出す．一方で生成技術は性質・機能空間から物質空間への射影であり，物性値から物質を導き出す．

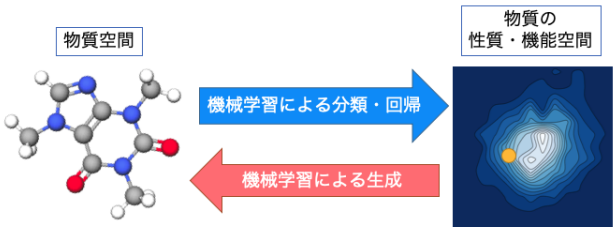

図 1　物質探索手法の比較
Fig. 1　Two material discovery methods.

それぞれの 2 つの技術の詳細な分類については，データとする物質，さらに深層学習モデルに入力するデータ構造にて分類する．まず，データとする物質においては分子と結晶に大別する．分子は 2 つ以上の原子が結合した電気的に中性な物質であり，有機化合物などがその典型である．結晶は原子や分子が集まって規則正しく配列した物質であり，金属がその典型である．次に，データ構造に関しては，原子と原子間の関係性をモデル化したグラフ，原子の位置をモデル化する 3 次元データ，また化学構造を 1 次元的に表現する文字列があり，それぞれのデータ構造に対して深層学習モデルが適応されている．

本稿の構成は以下のとおりである．2 章で，深層学習の既存技術の分類とデータの表現について，3 章で分類・回帰技術，4 章で生成技術について，5 章にて他の技術と深層学習との比較について述べる．最後に，6 章で今後の課題について述べ，7 章で本稿のまとめる．

2.　既存技術の分類

新物質探索における識別・予測技術と生成技術において，それぞれの技術が対象とする物質，データの表現方法，および深層学習モデルの観点から分類を行う．**表 1** にそれぞれの技術の概要をまとめ，技術の詳細においては，3 章と 4 章にて述べる．

以下では，データ表現，深層学習モデル，および他分野における深層学習技術との違いを述べる．

2.1　対象の物質

探索の対象となる物質の形態は分子と結晶に大別される．本稿では，分子は 2 つ以上の原子が何らかの力により結合した電気的に中性な物質，結晶は原子や分子が周期的に配列した物質と考える．分子は比較的少数の原子から構成され[*1]，分子の構造が興味の対象となる物質としては，医薬品などの有機化合物が代表的な例である．実際の物質における結晶は，アボガドロ数程度のオーダの個数（10^{23} 個）の原子により構成されるきわめて巨大な系であるが，これは比較的少数の原子をある対称操作（並進・回転・鏡映など）に従って配置した単位格子（繰り返し単位）が多

[*1]　本稿ではポリマーなどの高分子は対象としない．

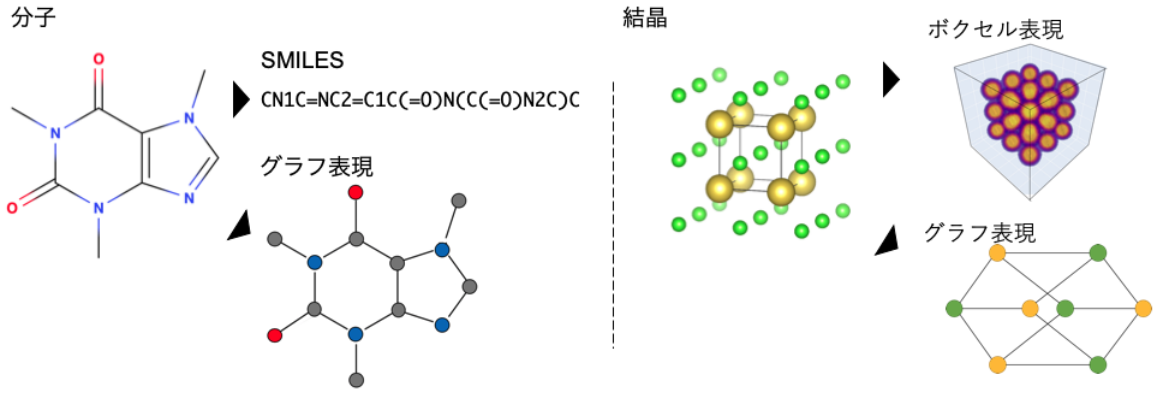

図 2　分子と結晶のデータ表現例
Fig. 2　Data representation of molecules and crystals.

表 1　既存の深層学習による物質探索の概要．データは論文内にて使用されている実験データの種類を表す
Table 1　Summary of deep learning technology for material discovery.

識別・予測技術

論文	データ	表現	モデル
文献 [14]	分子	文字列	RNN
文献 [6], [11], [22], [24], [44]	分子	グラフ	GNN
文献 [6], [55]	結晶	グラフ	GNN
文献 [43]	結晶	FP	FC

生成技術

論文	データ	表現	モデル
文献 [4], [49]	分子	文字列	RNN
文献 [40]	分子	文字列	RNN+RL
文献 [9], [13], [29]	分子	文字列	RNN+VAE
文献 [16]	分子	文字列	RNN+GAN+RL
文献 [31], [33]	分子	グラフ	RNN
文献 [21], [50]	分子	グラフ	GNN+VAE
文献 [10], [57]	分子	グラフ	GNN+GAN+RL
文献 [23]	結晶	ボクセル	CNN+VAE

数配列されたものとして表現することができる．結晶構造が興味の対象となる物質の例としては，金属やセラミックなどがあげられる．分子に対応する物性は，その分子構造によって発現する．一方，結晶性物質の物性は，単一の単位格子ではなくその無数の集合によってマクロに発現する特質であることに留意する必要がある．

2.2　データ表現

分子や結晶のデータ表現の方法について述べる．物質は 3 次元空間上に存在する原子により構成され，原子間に働く相互作用によりその構造が決まる．物質をどのように表現するかにより深層学習モデル設計および精度に大きな影響がある．物質のデータ表現は下記のものがある．

- 文字列：化学構造を文字列として 1 次元的に表現する方法．
- グラフ：原子を節点，距離や結合の種類を辺とした重み付き有向グラフとして表現．
- フィンガープリント（FP）：分子の構造の有無をビット列で表現．
- 座標データ：物質を構成する原子の座標で表現．
- ボクセル：画像で用いられるピクセルを 3 次元に拡張した表現．

図 2 に分子および結晶の物質構造の例を示す．グラフや文字列による表現は物質を扱う深層学習モデルでよく用いられる表現である．グラフ表現では，原子を節点，距離や結合の種類を辺とする．また，節点の情報には原子の種類に加えて，電荷など，各原子に付随する情報を選択的に用いる．辺の情報には，一重結合，二重結合などの原子間の結合の種類，あるいは原子間の距離が用いられる．文字列の表現は SMILES 記法 [53] が用いられる．SMILES 記法では節点を原子，辺を結合とした分子グラフを深さ優先探索により線形な文字列として表現する方法である．座標データは物質を構成する原子の 3 次元的な位置を表すデータである．フィンガープリントは特定の分子の構造の有無をビット列で表現し，特徴的な分子構造を含んでいるか，もしくはいくつ含んでいるかを表すことができる．

2.3　深層学習モデル

新物質探索において様々な深層学習モデルが適応されている．代表的に用いられているものを下記に示す．

- 全結合ニューラルネットワーク（FC）：順伝播型ニューラルネットワークであり，すべてのユニットが結合している．最も単純なニューラルネットワークである．
- 再帰ニューラルネットワーク（RNN）[32]：データのシーケンスを繰り返し入力し学習する．自然言語処理における文字列や時系列データにおけるストリーミングデータなどのシーケンスデータを扱う分野にてよく用いられる．
- 畳込みニューラルネットワーク（CNN）[28]：畳み込

み層とプーリング層から構成されているニューラルネットワークである．画像認識にて広く用いられる．
- グラフニューラルネットワーク（GNN）[58]：グラフニューラルネットワークはグラフ構造を扱う深層学習モデルの総称であり，隣接節点との関係性を用いて学習する．
- 敵対生成ネットワーク（GAN）[15]：生成モデルの一種であり，generator と discriminator から構成される．generator は訓練データに近いデータを生成し，discriminator は訓練データなのか generator により生成されたデータなのかを識別する．これにより訓練データには存在しないが，訓練データに近いデータを生成可能である．
- 変分オートエンコーダ（VAE）[27]：生成モデルの一種であり，入力データよりも少ない次元数の特徴を抽出する．

分類と回帰においては，RNN と GNN がそれぞれのデータ表現方法に応じて用いられている．一方，生成においては RNN のみ，もしくは GAN や VAE が用いられることが多く，GAN や VAE を用いる場合，そのアーキテクチャの一部に RNN や GNN をデータ表現方法に応じて用いる．

また，このような深層学習モデルと併用して，物質探索のガイドの役割で強化学習 [2] が用いられる．強化学習（RL）は，設計された報酬を最大化する行動や選択を学習することが可能である．

2.4　他分野と新物質探索の違い

最後に，深層学習技術の適用に関して，画像や自然言語処理といった代表的な領域と，新物質探索における技術的な違いを述べる．

- **データ構造**：データの大きさが任意長であり，また順序がない．
- **物理的な制約**：生成される物質には物理化学的な制約がある．

まず，データの入力順序の問題がある．分子や結晶内の原子はデータの順番が一意に決定しない．これはグリッド構造で順序を規則的に割り当てることのできる画像データとの大きな違いである．さらに，別の問題として，データの大きさの問題がある．深層学習モデルの入力層は固定長であることが多いが，分子や原子に含まれる原子の個数は様々であり扱うことが難しい．これらのデータ構造上の問題は深層学習モデルの選択，開発に大きく関わる．

次に，物質には物理的な制約の問題が存在する．たとえば，原子の結合には原子価の制約があり，また，結晶は空間的な対称性を満たす必要がある，このような制約があるため，生成モデルでは，可能な限り物理化学的に正しい物質を生成するようなアーキテクチャが必要である．

これらの問題により，単純に既存技術を適用するだけでは所望の物質を探索することは難しく，新物質探索のために新たな深層学習技術が開発されている．

3.　分類・回帰技術

分類・回帰技術は，ある物質が与えられたときに物質の性質を予測することが可能である．与えられた物質データから正確に物性値を高精度に予測するために様々なモデルが提案されている．以下では，物質の性質予測タスクにおける深層学習モデルについて述べる．

3.1　分子における分類・回帰技術

まず，分子における分類・回帰技術について述べる．分子に関する分類・回帰技術に関する機械学習を研究はさかんに研究されている．深層学習以外の機械学習を用いる手法では，特徴量を自動的に抽出できないため，多様な特徴量が提案されており，たとえば，Coulomb matrix, Bag of bonds, フィンガープリントなどがある [45]．一方，深層学習モデルではより文字列やグラフ表現などの表現が用いられることが多い．以下では入力するデータ構造に分けて技術を紹介する．

3.1.1　文字列表現を用いた分類・回帰技術

Goh [14] らは，SMILES を入力とした RNN を用いて，物性予測を行っている．また，文字列の一部をマスクすることで，性質予測の精度に対して，どの文字が重要かを調べている．評価実験では水和物のデータセットを用いた溶媒和自由エネルギーの予測では第一原理計算を用いた場合よりも高い性能を示している．

SMILES 表現 [53] はどの原子から文字列化するかによって複数の表現を取りうる．この問題に対処するために，分子グラフと一意に対応するように SMILES を拡張した表記方法として，canonical SMILES がある [38]．また，機械学習で用いる場合には，同じ分子グラフから SMILES が複数種類作成可能であることを利用し，データオーグメンテーションを行う方法がある [3]．Kimber ら [26] は，オーグメンテーションが行われた SMILES を用いて畳み込みニューラルネットワークを用いた提案モデルの評価を行っている．

3.1.2　グラフ表現を用いた分類・回帰技術

Duvenaud ら [11] は，フィンガープリントの一種である circular fingerprint のアルゴリズムを拡張して，グラフ表現を扱う畳み込みニューラルネットを構築した．このモデルは特徴量設計を必要としないフィンガープリント（neural fingerprints）の作成と分子の性質の予測を可能にした．検証実験では溶解度，薬効，有機太陽光発電率のラベル付きデータ，それぞれ 1,144 件，10,000 件，20,000 件を用いて，従来用いられてきた extended-connectivity circular fingerprint（ECFP）を用いた場合と比較し，溶解度，有機太陽光発電率については予測精度向上，薬効については同

等の予測性能を達成している．またKearnesら [24] もグラフ畳み込みニューラルネットワークを用いて特徴量を作成する手法を提案している．

文献 [11], [24] では，それぞれグラフ畳み込みニューラルネットワークを提案しているが，一般にグラフ畳み込みニューラルネットワーク以外にも様々なグラフを扱うモデルが存在する．たとえば，アテンション機構を用いたモデル [52] や，ゲート機構を取り入れたGG-NNs [30] がある．

Gilmerら [22] はGG-NNsにアテンション機構を持つモデルを組み合わせることで，QM9データセットを用いた13種類の物性値予測実験においてKearnesら [24] の手法やGG-NNsやその他機械学習による手法よりも高い性能を記録している．Ryuら [44] も同様に，アテンション機構とゲート機構を取り入れたグラフ畳み込みニューラルネットワークを用いた手法を提案している．この手法ではアテンション機構の一部に各分子内の原子のペアによって決まる相互作用情報を持った辞書を用いている．実験では，分配係数の予測においてGG-NNsと同等の精度でより短期間で学習できることを示している．また，学習した原子の特徴量および分子の潜在表現の距離を可視化することで，原子の機能や分子を構造特性相関について提案手法がより深く特徴をとらえられることを確認している．

3.1.3　その他の表現を用いた分類・回帰技術

3次元的な原子配置の座標データを用いた予測器も存在する．座標データを用いる場合，画像のようにデータがグリッドに固定されていないため処理が難しくなる．Schüttら [47] は，任意の位置を持つ対象をモデリング可能なcontinuous-filter convolutional（cfconv）layersを用いて，分子内の任意の位置に存在する原子間の相互作用をモデリングしたSchNetを提案している．SchNetでは活性化関数に解析的に微分可能な関数を用いることで出力されるエネルギーの入力の座標に対する微分係数，すなわち力の計算を可能にしている．検証実験では，QM9データセットを用いたエネルギー予測問題においてGilmerら [22] の手法と比較して高い性能を示している．

3.2　結晶における分類・回帰技術

分子における分類・回帰技術と比較して，結晶における分類・回帰技術の深層学習の適用事例は少ない．結晶は主に材料科学分野にて研究されているが，分子に比べて利用可能なデータが少ないことや，結晶構造データの取り扱いの難しさが一因である．

Ryanら [43] はフィンガープリントを入力とした全結合ニューラルネットワークによって，結晶構造のトポロジーに基づいた潜在表現を得られることを確認した．また，Yeら [56] は各原子の電気陰性度とイオン半径のみを記述子に用いて，DFTで計算されたガーネットとペロブスカイトのデータに絞り形成エネルギーを予測した．その結果ガーネットでは7–10 meV/atom，ペロブスカイトでは20–34 meV/atomの精度で予測可能であることを示した．

分子の場合と同様に結晶の場合にもグラフニューラルネットワークを用いたモデルが存在する．CGCNN [55] は結晶構造を分子の場合と同様にグラフで表現し，グラフ畳み込みニューラルネットワークで扱うモデルである．DFTシミュレーションで計算された形成エネルギー，バンドギャップなど6つの物性値の予測タスクを行っている．その結果，形成エネルギーの予測では28,046件の訓練データを用い，平均絶対誤差0.039 eVの精度を出している．

Chenら [6] は文献 [22], [55] のモデルを一般化したグラフニューラルネットワークを用いて，分子と結晶の両方を扱えるモデルを提案している．さらに，気温などのグローバルなデータを学習に取り入れて実験を行っている．結晶の物性値予測実験ではCGCNNやSchNetとの比較を行っており，たとえば形成エネルギーの予測実験では，60,000件の結晶データを訓練に用いて平均絶対誤差0.028 eV/atomの予測精度を達成している．これはSchNetの結果を上回る精度である．

Kajitaら [23] は異なる結晶構造に含まれるどのような場の量も3次元のボクセルに変換できるボクセル記述子を開発し3次元の畳み込みニューラルネットワークによってハートリーエネルギーの予測を行った．また，1次元表現を用いた手法も存在する．NOMADのKaggleコンペティション [51] で優勝したモデルは結晶をグラフ表現に直し，さらにn-gramを用いてシーケンスに直した手法を取っている．

別のアプローチとして，Ziettiら [59] はシミュレーションによって計算した回折画像を入力とした畳み込みニューラルネットワーク（CNN）を用いた結晶構造の分類を行っている．また，Jhaら [20] はDFT計算データだけではなく実験データを利用した転移学習モデルを提案し，材料の形成エネルギー予測を行っている．

4.　生成技術

新物質探索では所望の性質に合わせて物質を探索する必要があり，これまでに機械学習を用いて物質探索のガイドをする手法が開発されてきた．近年は強化学習を用いた手法も提案されている [10]．生成モデルはこれらの技術と組み合わせることで，物質探索を効率化することが期待できる．生成モデルとして用いられるVAEやGANのモデルを**図 3** に示す．VAEは物質をエンコードして潜在表現を得たのち，潜在表現上で性質に関して最適化を行った後，デコードして物質を生成する手法が取られる．GANでは強化学習などによって生成物質に付随する性質の制御を行う．

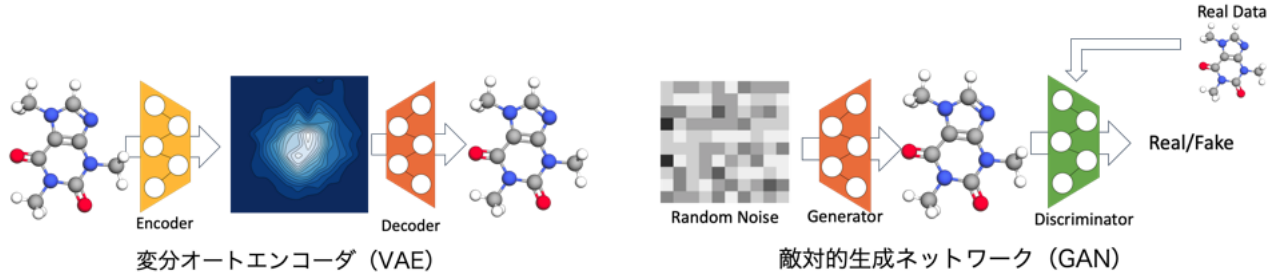

図 3　生成技術で用いられる深層学習モデルの模式図
Fig. 3　Popular frameworks for deep generative tasks.

4.1　分子における生成技術

分子における生成技術も分類・回帰技術と同様に各データ構造に対して様々な深層学習モデルが適用されている．

4.1.1　文字列表現を用いた生成技術

Bjerrum ら [4] は LSTM を含む RNN による生成モデルを提案している．学習データは ZINC12 データセット [18] を用い，モデルから生成される分子と訓練データの間で SA スコアなどの性質の分布の比較を行っている．Segler ら [49] も同様に，LSTM を含む RNN モデルによって新たな分子を生成している．さらに，この手法では訓練データに似た分子が生成されてしまうため，所望の性質を持つ少数の分子を用いて転移学習を行うことで，所望の性質に近い分子を生成可能であることを確かめている．さらに，Popova ら [40] は，RNN を用いた生成モデルに対して強化学習による探索手法を提案している．この手法では予測モデルと生成モデルから構成され，生成モデルにより生成された分子に対して予測モデルが数値的な報酬を割り当て，生成モデルは期待報酬を最大化するように訓練される．

Bombarelli ら [13] は，VAE を用いて，SMILES で表現された分子の集合を連続的な潜在表現にエンコードする手法を提案している．この潜在変数空間上で，所望の性質による勾配を用いた探索と，潜在表現から SMILES を再生成が可能であり，物質探索のプロセスを自動化している．VAE のエンコーダとデコーダには RNN や 1 次元の畳み込みニューラルネットが用いられている．

しかしながら，Bombarelli らの手法 [13] はデコード時に文法的に間違った SMILES を生成することがある．そこで GrammerVAE（GVAE）[29] では．エンコード，デコード時に SMILES を直接用いるのではなく，文脈自由文法の生成規則を表す構文木を利用し，生成される SMILES の有効性を高めた．また，探索にはベイズ最適化を用いて実験している．さらに，生成される SMILES の有効性を上げる手法として Dai ら [9] は文法と意味論的な制約を高めるために，コンパイラの理論を応用した手法を提案している．

Guimaraes ら [16] は，RNN を用いてシーケンスデータを扱う GAN を構築し，強化学習で最適化する手法を提案している．

4.1.2　グラフ表現を用いた生成技術

グラフ生成モデルは，全ノードを一度に生成する手法と，ノードを 1 つずつ生成する手法の大きく 2 つに分類できる．Simonovsky ら [50] は分子グラフを生成する VAE をベースとしたモデルを提案した．グラフは線形な表現を一意に作ることができないので，生成時は全ノードが結合した確率グラフを生成してからグラフを得る．MolGAN [10] は GAN を用いた分子グラフを生成するモデルである．最適化には（深層）強化学習を用いた勾配法である DDPG を用いている．You ら [57] はグラフ畳み込みニューラルネットワークと GAN を用いた生成モデルと強化学習を合わせた手法を提案している．

物理化学的に正しい分子グラフを生成することは難しい課題である．Ma ら [35] は，VAE で生成した分子グラフの有効性を上げる正則化フレームワークを提案している．Junction tree VAE [21] は木分解を用いて分子グラフ内の化学的に有効なサブグラフが壊れないようにしたモデルである．これにより化学的に有効な物質のみを生成することに成功している．

Liu ら [33] は，GGNNs と VAE を用いて，ノードを 1 つずつ生成する生成モデル（CGVAE）を提案している．Liu ら [31] は任意のグラフをシーケンスに生成できる生成モデルを提案し，分子グラフを生成する実験を行っている．実験では，SMILES をシーケンスに生成する手法と比較し，生成した分子の有効性や新規性が向上することを確認している．

4.2　結晶における生成技術

結晶における生成技術はまだまだ発展途上であり，提案技術の数は非常に少ない．CrystalGAN [37] では，GAN を用いた結晶構造生成モデルを提案している．しかしながらこの手法では高コストなシミュレーションでの評価を避けるため，限られた物質領域でしか実験が行われていない．結晶に対する新たな手法として，Hoffman ら [17] は結晶構造を 3 次元のデータとして扱う，VAE を用いた生成モデルを提案している．

5. 議論

本章では，深層学習と他の技術について比較する．まず，深層学習以外の機械学習技術と比較を行い，次にシミュレーションと比較を行う．

5.1 深層学習以外の機械学習との比較

深層学習以外の機械学習を用いた手法と深層学習を用いた手法の性能について議論する．深層学習は学習用のデータ量を十分に用意できる場合，高い精度で予測することが可能である．文献 [54] では，17 種類のデータセットを用いて幅広く機械学習の手法を評価しており，実験では 17 種類のうち 11 種類のデータセットにおいてグラフベースの深層学習を用いた手法が最高性能を収めている．その一方で，データサイズが小さく複雑なタスクに対してはグラフベースの深層学習手法が最適ではないことや，データに偏りが大きいデータセットの場合はカーネル SVM のような従来の手法が優れている場合もあることも示している．

また，深層学習の利点として特徴量設計が不要なことがあげられる．従来の機械学習は，特徴量設計がモデルの予測精度に大きく依存しているため，タスクとモデルに合わせた特徴量設計が必要である．これに対し，深層学習では特徴量設計の必要はなく，タスクに合わせてモデルの内部で学習される．一方で，深層学習における高精度な予測のためには，長時間の訓練やハイパーパラメータのチューニングが必要であることが一般的である．そのため，実験データの追加や修正があった場合，即座に対応することが難しい．そのため，大きな実験指針の決定には有効であるが，実験中におけるインタラクティブな意思決定においては従来の機械学習が向いている場合も多い．

次に生成技術においては，従来の機械学習技術で達成することは難しく，現状では深層学習による技術が多く提案されている．生成技術の実用上の応用可能性については，分子に対する生成技術は有効性の問題は解決しつつあり，実際の化学や創薬における物質探索への応用が期待できる．一方で，結晶の生成技術に関する研究は十分に行われておらず，実際の問題への応用にはまだ時間を要するだろう．

5.2 シミュレーションとの比較

物質・材料研究では，実験とあわせて，シミュレーションが広く用いられおり，シミュレーションにより計算された物性データが大規模に蓄積されている．本稿で述べた分類・回帰技術では物性の真値に対して予測を行ったわけではなく，シミュレーションによって計算された誤差を含む物性値データを用いて学習および精度検証を行っている．そのため，精度においてはシミュレーションと直接比較することは難しいが，たとえば，CGCNN の誤差は，シミュレーションによる物性値を真値とした場合において，シミュレーション誤差以内の精度で予測可能である．しかし，より精密なシミュレーション計算における誤差の範囲には収まっておらず，深層学習の精度は十分とはいえない．一方で，深層学習は学習済みのモデルがあれば，シミュレーションと比較して高速に物性値を推定することが可能である．そのため，高速かつ精度を大きく下げずに予測でき仮想的なスクリーニングに有用である．

6. 今後の課題

本章では，既存技術と現状をふまえ，今後の課題についてまとめる．

6.1 3 次元構造を扱う深層学習技術

3 章および 4 章では，グラフ表現やグラフを文字列に直す SMILES を用いた手法が高い精度を出すことを説明した．現状の概観として，分子のような化合物の構造式については，SMILES などのデータの表現方法が確立され，構造を機械学習の枠組みで取り扱うための様々な手法が発展している．その一方で，結晶構造についてはいまだそのデータ表現を模索している段階にある．

現在は結晶のデータ表現として，グラフが主流といえるが，グラフには 3 次元構造の情報を一部失うという情報の損失がある．たとえば，立体異性体は異なる原子の配置を持つ分子であるが，同じグラフとして表現される．また，3 次元的な配置は，結晶構造や受容体を考えるうえでも重要である．

3 次元構造を扱う深層学習技術として大きく分けて 2 つの手法がある．まず 1 つ目はグラフに座標や原子間の距離などの属性値を付与する方法である．たとえば，グラフのエッジを結合の種類ではなく，原子間の距離を用いたモデルがある [55]．次に，グラフではなく物質を 3 次元データと見なす方法である．点群や，ボクセルのような 3 次元データはコンピュータビジョンなどの分野でさかんに研究されており，PointNet などの 3 次元データを扱う識別モデルや生成モデルが提案されている [1]．そのため，物質を 3 次元データとして見なした深層学習による物質探索は今後の有望な技術の 1 つである．

材料開発において結晶構造を取り扱う機会は多く，結晶構造からの物性予測，あるいは望む物性に対応する構造の生成は実用上きわめて重要なタスクであるため，結晶に対してどのようなデータ表現を用いるべきかは今後の新物質探索における重要なタスクの 1 つである．

6.2 合成プロセスの予測

有機化合物については，化学式から物性予測，あるいは望む物性を持ちそうな化合物を予測するためのデータセット（QM9 など）がすでに普及している．しかし，提案された新規化合物が実際に現実的なコストで合成可能であるか

は別問題であり，現状ではその合成経路探索は熟練者による直感や，きわめて高コストな実験をともなうため，研究のボトルネックとなり得る．すなわち，機械学習による予測と，実世界における実験のタイムスケールのギャップを埋める工夫が必要となる．したがって今後は，合成可能性も含めた予測および，候補物質の合成プロセス（レシピ）の予測がトレンドとなると見込まれる．特に合成プロセスの予測についてはいまだ発展途上で，その記述方法を含めて模索がなされている段階である．そのため，深層学習による合成プロセスの予測が期待されている．一方で，現状で利用可能なオープンデータでは合成可能性やプロセスについての情報は網羅されておらず，学習データ量が足りないという問題があるため，材料科学と情報科学の連携が今後ますます重要となる．

6.3　失敗データの収集と公開

物質合成など何らかの実験が失敗する条件を知ることは，その成功条件を探索するうえで非常に重要である．しかし，実験の失敗条件を論文として発表することの価値は低いとみなされており，研究室内の実験ノートやコンピュータに保存されるのみで，失敗のデータが公開されることは少ない．

Raccuglia ら [41] は，この事実に着目し，研究室内の実験ノートから失敗した実験条件をクロールし，そのデータを用いて機械学習モデルを構築することで，成功する実験条件を提案することに成功した．そのため，成功した実験データのみ公開するのではなく，失敗した実験データの公開する取り組みが広がりつつある．最近の取り組みとして，EU では公的研究費の配分ポリシーとして実験結果のオープンデータ化を推進し，実験データの永続化・公開のためのデータリポジトリの整備が行われている．一例として，EU と CERN によって運営される zenodo[*2]などがある．学術出版の面からも，実験データの公開を主目的とした雑誌（Scientific Data など）が創刊されるなど，新しい取り組みがなされており，コミュニティ全体において，失敗も含めた実験データの価値が認識されようとしている．

一方で，材料開発は有益な材料を開発することが目的であり，成功が期待される実験条件のみを対象として実験を行うのが基本である．いい換えれば，あえて失敗しそうな実験をすることはない．したがって，公開されるデータセットには既知の（成功する）実験条件の周辺を重点的にサンプリングしたことによるバイアスが含まれることに留意する必要がある．深層学習において，研究者らが成功を期待する実験のみが訓練データに含まれている場合，学習結果においても研究者らが想像する範疇から抜け出せない可能性が高い．そのため，深層学習などの機械学習の利用を前提として，失敗する可能性が高い実験を含む様々な実験データを取得することが重要と考えられる．

[*2] https://zenodo.org/

7.　おわりに

新物質探索の研究は，自然言語やグラフ，3 次元データを扱う深層学習モデルとの関わりが深い．そのため，化学や材料科学だけではなく，情報科学においてもさかんに研究されている．

本調査では，分類・回帰技術と，生成技術に大きく分けて，深層学習を用いた新物質探索の研究動向について調査した．分類・回帰技術は，いくつかの物性値においてシミュレーションの誤差以下の精度での予測を可能にしている．また，生成技術は，分類・回帰技術や最適化手法と組み合わせることにより，新物質探索の効率化と自動化に貢献している．

本稿が新物質探索における深層学習技術の発展の助けとなれば幸いである．

謝辞　本研究の一部は，JST ACT-I（JPMJPR18UE）および科学研究費（20K19805）の支援を受けて実施された．

参考文献

[1] Ahmed, E., Saint, A., Shabayek, A.E.R., Cherenkova, K., Das, R., Gusev, G., Aouada, D. and Ottersten, B.: A survey on Deep Learning Advances on Different 3D Data Representations, arXiv preprint arXiv:1808.01462 (2018).

[2] Arulkumaran, K., Deisenroth, M.P., Brundage, M. and Bharath, A.A.: Deep Reinforcement Learning: A Brief Survey, *IEEE Signal Processing Magazine*, Vol.34, No.6, pp.26–38 (2017).

[3] Bjerrum, E.J.: Smiles enumeration as data augmentation for neural network modeling of molecules, arXiv preprint arXiv:1703.07076 (2017).

[4] Bjerrum, E.J. and Threlfall, R.: Molecular generation with recurrent neural networks (RNNs), arXiv preprint arXiv:1705.04612 (2017).

[5] Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O. and Walsh, A.: Machine learning for molecular and materials science, *Nature*, Vol.559, No.7715, pp.547–555 (2018).

[6] Chen, C., Ye, W., Zuo, Y., Zheng, C. and Ong, S.P.: Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals, *Chemistry of Materials*, Vol.31, No.9, pp.3564–3572 (2019).

[7] Chen, W.L.: Chemoinformatics: Past, Present, and Future, *Journal of Chemical Information and Modeling*, Vol.46, No.6, pp.2230–2255 (2006).

[8] Curtarolo, S., Setyawan, W., Wang, S., Xue, J., Yang, K., Taylor, R.J.K., Hart, G.L.W., Sanvito, S., Nardelli, M.B., Mingo, N. and Levy, O.: AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations (2012).

[9] Dai, H., Tian, Y., Dai, B., Skiena, S. and Song, L.: Syntax-directed variational autoencoder for structured data, arXiv preprint arXiv:1802.08786 (2018).

[10] De Cao, N. and Kipf, T.: MolGAN: An implicit generative model for small molecular graphs, arXiv preprint

arXiv:1805.11973 (2018).
[11] Duvenaud, D.K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A. and Adams, R.P.: Convolutional networks on graphs for learning molecular fingerprints, *Advances in Neural Information Processing Systems*, pp.2224–2232 (2015).
[12] Elton, D.C., Boukouvalas, Z., Fuge, M.D. and Chung, P.W.: Deep learning for molecular design—A review of the state of the art, *Molecular Systems Design & Engineering*, Vol.4, No.4, pp.828–849 (2019).
[13] Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P. and Aspuru-Guzik, A.: Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, *ACS Central Science*, Vol.4, No.2, pp.268–276 (2018).
[14] Goh, G.B., Hodas, N.O., Siegel, C. and Vishnu, A.: Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties, arXiv preprint arXiv:1712.02034 (2017).
[15] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets, *Advances in Neural Information Processing Systems*, pp.2672–2680 (2014).
[16] Guimaraes, G.L., Sanchez-Lengeling, B., Outeiral, C., Farias, P.L.C. and Aspuru-Guzik, A.: Objective-reinforced generative adversarial networks (organ) for sequence generation models, arXiv preprint arXiv: 1705.10843 (2017).
[17] Hoffmann, J., Maestrati, L., Sawada, Y., Tang, J., Sellier, J.M. and Bengio, Y.: Data-Driven Approach to Encoding and Decoding 3-D Crystal Structures, arXiv preprint arXiv:1909.00949 (2019).
[18] Irwin, J.J., Sterling, T., Mysinger, M.M., Bolstad, E.S. and Coleman, R.G.: ZINC: A Free Tool to Discover Chemistry for Biology, *Journal of Chemical Information and Modeling*, Vol.52, No.7, pp.1757–1768 (2012).
[19] Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G. and Persson, K.A.: Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Materials*, Vol.1, No.1, p.011002 (2013).
[20] Jha, D., Choudhary, K., Tavazza, F., Liao, W.-K., Choudhary, A., Campbell, C. and Agrawal, A.: Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning, *Nature Communications*, Vol.10, No.1, pp.1–12 (2019).
[21] Jin, W., Barzilay, R. and Jaakkola, T.: Junction tree variational autoencoder for molecular graph generation, arXiv preprint arXiv:1802.04364 (2018).
[22] Jørgensen, P.B., Jacobsen, K.W. and Schmidt, M.N.: Neural message passing with edge updates for predicting properties of molecules and materials, arXiv preprint arXiv:1806.03146 (2018).
[23] Kajita, S., Ohba, N., Jinnouchi, R. and Asahi, R.: A universal 3D voxel descriptor for solid-state material informatics with deep convolutional neural networks, *Scientific Reports*, Vol.7, No.1, pp.1–9 (2017).
[24] Kearnes, S., McCloskey, K., Berndl, M., Pande, V. and Riley, P.: Molecular graph convolutions: Moving beyond fingerprints, *Journal of Computer-Aided Molecular Design*, Vol.30, No.8, pp.595–608 (2016).
[25] Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., Wang, J., Yu, B., Zhang, J. and Bryant, S.H.: PubChem Substance and Compound databases, *Nucleic Acids Research*, Vol.44, No.D1, pp.D1202–D1213 (2015).
[26] Kimber, T.B., Engelke, S., Tetko, I.V., Bruno, E. and Godin, G.: Synergy effect between convolutional neural networks and the multiplicity of smiles for improvement of molecular prediction, arXiv preprint arXiv:1812.04439 (2018).
[27] Kingma, D.P. and Welling, M.: Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).
[28] Krizhevsky, A., Sutskever, I. and Hinton, G.E.: Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, pp.1097–1105 (2012).
[29] Kusner, M.J., Paige, B. and Hernández-Lobato, J.M.: Grammar variational autoencoder, *International Conference on Machine Learning*, pp.1945–1954 (2017).
[30] Li, Y., Tarlow, D., Brockschmidt, M. and Zemel, R.: Gated graph sequence neural networks, arXiv preprint arXiv:1511.05493 (2015).
[31] Li, Y., Vinyals, O., Dyer, C., Pascanu, R. and Battaglia, P.: Learning deep generative models of graphs, arXiv preprint arXiv:1803.03324 (2018).
[32] Lipton, Z.C., Berkowitz, J. and Elkan, C.: A critical review of recurrent neural networks for sequence learning, arXiv preprint arXiv:1506.00019 (2015).
[33] Liu, Q., Allamanis, M., Brockschmidt, M. and Gaunt, A.: Constrained graph variational autoencoders for molecule design, *Advances in Neural Information Processing Systems*, pp.7795–7804 (2018).
[34] Liu, Y., Zhao, T., Ju, W. and Shi, S.: Materials discovery and design using machine learning, *Journal of Materiomics*, Vol.3, No.3, pp.159–177 (2017). High-throughput Experimental and Modeling Research toward Advanced Batteries.
[35] Ma, T., Chen, J. and Xiao, C.: Constrained generation of semantically valid graphs via regularizing variational autoencoders, *Advances in Neural Information Processing Systems*, pp.7113–7124 (2018).
[36] Mitchell, J.B.O.: Machine learning methods in chemoinformatics, *WIREs Computational Molecular Science*, Vol.4, No.5, pp.468–481 (2014).
[37] Nouira, A., Sokolovska, N. and Crivello, J.-C.: Crystalgan: learning to discover crystallographic structures with generative adversarial networks, arXiv preprint arXiv:1810.11203 (2018).
[38] O'Boyle, N.M.: Towards a Universal SMILES representation: A standard method to generate canonical SMILES based on the InChI, *Journal of Cheminformatics*, Vol.4, No.1, p.22 (2012).
[39] Polishchuk, P.G., Madzhidov, T.I. and Varnek, A.: Estimation of the size of drug-like chemical space based on GDB-17 data, *Journal of Computer-Aided Molecular Design*, Vol.27, No.8, pp.675–679 (2013).
[40] Popova, M., Isayev, O. and Tropsha, A.: Deep reinforcement learning for de novo drug design, *Science Advances*, Vol.4, No.7 (2018).
[41] Raccuglia, P., Elbert, K.C., Adler, P.D.F., Falk, C., Wenny, M.B., Mollo, A., Zeller, M., Friedler, S.A., Schrier, J. and Norquist, A.J.: Machine-learning-assisted materials discovery using failed experiments, *Nature*, Vol.533, No.7601, pp.73–76 (2016).
[42] Rajan, K.: Materials Informatics: The Materials "Gene"

and Big Data, *Annual Review of Materials Research*, Vol.45, No.1, pp.153–169 (2015).
[43] Ryan, K., Lengyel, J. and Shatruk, M.: Crystal Structure Prediction via Deep Learning, *Journal of the American Chemical Society*, Vol.140, No.32, pp.10158–10168 (2018).
[44] Ryu, S., Lim, J., Hong, S.H. and Kim, W.Y.: Deeply learning molecular structure-property relationships using attention-and gate-augmented graph convolutional network, arXiv preprint arXiv:1805.10988 (2018).
[45] Sanchez-Lengeling, B. and Aspuru-Guzik, A.: Inverse molecular design using machine learning: Generative models for matter engineering, *Science*, Vol.361, No.6400, pp.360–365 (2018).
[46] Schmidt, J., Marques, M.R., Botti, S. and Marques, M.A.: Recent advances and applications of machine learning in solid-state materials science, *NPJ Computational Materials*, Vol.5, No.1, pp.1–36 (2019).
[47] Schütt, K., Kindermans, P.-J., Felix, H.E.S., Chmiela, S., Tkatchenko, A. and Müller, K.-R.: Schnet: A continuous-filter convolutional neural network for modeling quantum interactions, *Advances in Neural Information Processing Systems*, pp.991–1001 (2017).
[48] Schwalbe-Koda, D. and Gómez-Bombarelli, R.: Generative Models for Automatic Chemical Design, arXiv preprint arXiv:1907.01632 (2019).
[49] Segler, M.H.S., Kogej, T., Tyrchan, C. and Waller, M.P.: Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks, *ACS Central Science*, Vol.4, No.1, pp.120–131 (2018).
[50] Simonovsky, M. and Komodakis, N.: GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders, *Lecture Notes in Computer Science*, pp.412–422 (2018).
[51] Sutton, C., Ghiringhelli, L.M., Yamamoto, T., Lysogorskiy, Y., Blumenthal, L., Hammerschmidt, T., Golebiowski, J., Liu, X., Ziletti, A. and Scheffler, M.: Nomad 2018 kaggle competition: Solving materials science challenges through crowd sourcing, arXiv preprint arXiv:1812.00085 (2018).
[52] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. and Bengio, Y.: Graph attention networks, arXiv preprint arXiv:1710.10903 (2017).
[53] Weininger, D.: SMILES, a Chemical Language and Information System, 1. Introduction to Methodology and Encoding Rules, *Journal of Chemical Information and Computer Sciences*, Vol.28, No.1, pp.31–36 (1988).
[54] Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K. and Pande, V.: MoleculeNet: A benchmark for molecular machine learning, *Chemical science*, Vol.9, No.2, pp.513–530 (2017).
[55] Xie, T. and Grossman, J.C.: Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties, *Physical Review Letters*, Vol.120, p.145301 (2018).
[56] Ye, W., Chen, C., Wang, Z., Chu, I.-H. and Ong, S.P.: Deep neural networks for accurate predictions of crystal stability, *Nature Communications*, Vol.9, No.1, pp.1–6 (2018).
[57] You, J., Liu, B., Ying, Z., Pande, V. and Leskovec, J.: Graph convolutional policy network for goal-directed molecular graph generation, *Advances in Neural Information Processing Systems*, pp.6410–6421 (2018).
[58] Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C. and Sun, M.: Graph neural networks: A review of methods and applications, arXiv preprint arXiv:1812.08434 (2018).
[59] Ziletti, A., Kumar, D., Scheffler, M. and Ghiringhelli, L.M.: Insightful classification of crystal structures using deep learning, *Nature Communications*, Vol.9, No.1, pp.1–10 (2018).

奥野 智也

大阪大学大学院情報科学研究科博士前期課程．2019 年関西大学システム理工学部卒業．深層学習を応用した材料研究に興味を持つ．

佐々木 勇和　（正会員）

大阪大学大学院情報科学研究科助教．2014 年大阪大学情報科学研究科博士後期課程修了．博士（情報科学）．データベースシステム，グラフデータ処理，都市コンピューティングに関する研究に従事．ACM, IEEE, 日本データベース学会の各会員．

鈴木 雄太

総合研究大学院大学高エネルギー加速器科学研究科博士後期課程．2019 年東京理科大学基礎工学研究科修士課程修了．修士（工学）．機械学習を応用した物質計測技術開発，材料データのデータマイニング等，マテリアルズインフォマティクスに関する研究に従事．

（担当編集委員　小山 聡）