統計数理 第69巻 第1号

（通巻133号）

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

目　次

特集「マテリアルズインフォマティクスの最前線」

2021年6月

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

〒190-8562 東京都立川市緑町10-3　　電話 050-5533-8500（代）

本号の内容はすべて https://www.ism.ac.jp/editsec/toukei/ からダウンロードできます

ISSN 0912-6112

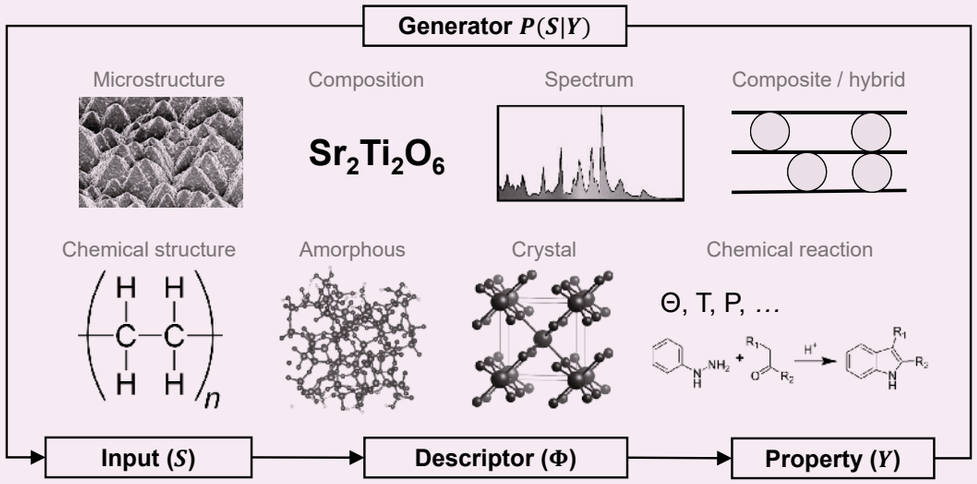
統計数理

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

第69巻 第1号 2021

統計数理 Vol. 69, No. 1

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS



統計数理研究所

統計数理

（年 2 回発行）

「統計数理」は，統計数理研究所における研究成果を掲載する統計数理研究所「彙報」として 1953 年に歴史を始め，1985 年に誌名を変更し今の形となりました．現在は，統計数理研究所の研究活動に限らず，広く統計科学に関する投稿論文を掲載し，統計科学の深化と発展，そして統計科学を通じた社会への貢献を目指しています．

投稿を受け付けるのは，次の 6 種です．

a.　原著論文　　b.　総合報告　　c.　研究ノート
d.　研究詳解　　e.　統計ソフトウェア　　f.　研究資料

投稿された原稿は，編集委員会が選定・依頼した査読者の審査を経て，掲載の可否を決定します．投稿規程，執筆要項は，本誌最終頁をご参照ください．

また，上記以外にも統計科学に関して編集委員会が重要と認める内容について，編集委員会が原稿作成を依頼することがあります．

その他，「統計数理」に関するお問い合わせは，各編集委員にお願いします．

All communications relating to this publication should be addressed to associate editors of the Proceedings.

大学共同利用機関法人　情報・システム研究機構
統計数理研究所
〒 190-8562 東京都立川市緑町 10-3　　電話 050-5533-8500（代）
https://www.ism.ac.jp/

印刷：笹氣出版印刷株式会社

表紙の図は本誌 6 ページ図 1 を改変したものです

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

Vol. 69,　No. 1

Contents

Special Issue : Frontiers of Materials Informatics

Letter

June, 2021
Research Organization of Information and Systems
The Institute of Statistical Mathematics
10–3 Midori-cho, Tachikawa, Tokyo 190–8562, JAPAN

統計数理(2021)
第 69 巻 第 1 号 1–3

巻頭言「特集 マテリアルズインフォマティクスの最前線」について

吉田 亮[†](オーガナイザー)

マテリアルズインフォマティクスは，材料科学とデータ科学の融合領域である．2011 年に米国にてマテリアルズ・ゲノム・イニシアチブ(Materials Genome Initiative: MGI)という国家プロジェクトが始動した．材料開発には，新素材の発見から製品化までにおよそ 10-20 年という年月を要すると言われている．同プロジェクトでは，材料開発に要する期間を半分に短縮するという目標が掲げられ，そのホワイトペーパーにおいて材料データ基盤の整備とデータ科学の技術導入が目標実現への鍵になると宣言された．これを機にマテリアルズインフォマティクスという学際領域が一躍脚光を浴びることになった．我が国では，2015 年に JST イノベーションハブ構築支援事業「情報統合型物質・材料開発イニシアティブ」(拠点：国立研究開発法人 物質・材料研究機構)が始動し，マテリアルズインフォマティクスの学術基盤の整備と人材創出，社会実装に向けた動きが急速に活発化した．統計数理研究所の研究者は，データ科学の学術基盤をもとにマテリアルズインフォマティクスの学術創生を促進してきた．さらに，2017 年には統計数理研究所において「ものづくりデータ科学研究センター」が設立された [1]．同センターでは，素材・化学企業を中心に多数の企業との共同研究を推進し，材料開発の最前線で実践・実証研究を展開している．データ科学の独自の視点から材料研究の諸問題に対するユニークな切り口を発見し，新しい科学的手法を創出・実践する．これが同センターに課されたミッションである．本特集を構成する 5 報の論文は，この一連の取り組みから生み出された成果の一部をまとめたものである．

材料研究のパラメータ空間は極めて広大である．例えば，有機低分子化合物のケミカルスペースには，約 10^{60} 個の候補分子が存在すると言われている．さらに，実用材料の研究では，プロセスや添加剤・溶媒選択などが設計変数に加わり，パラメータ空間の大きさは爆発的に増大する．マテリアルズインフォマティクスの問題の多くは，このような広大な探索空間から所望の特性を有するパラメータを同定することに帰着する．これまでは実験と物理法則に基づく計算機実験が材料研究の進歩を牽引してきた．研究者の経験や勘に基づき材料を設計し，計算と実験による物性評価に基づき設計指針を見直す．このようなアプローチでこれまでに数多くの革新的材料が発見されてきた．しかしながら，経験や勘に基づく試行錯誤的な設計，計算，実験というループだけでは，決して超えられない壁が存在する．ここに「データ」と「データの科学」を導入することで，材料開発のコストを大幅に削減し，革新的特性を持つ新材料を創製する．これがマテリアルズインフォマティクスに寄せられた社会からの期待である．

2015 年の「情報統合型物質・材料開発イニシアティブ」発足当時，マテリアルズインフォマティクスの世界は，ほぼ何もない広大なブルーオーシャンであった．当時この分野に集まった研究者には，データ駆動型科学が革新的な成果をもたらしうる研究シーズを探索・発見し，自らのアイデアを実証し，社会に対してデモンストレーションすることが求められた．同セン

[†] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3

ターの研究者らは，データ科学の学術基盤を活かし，この未踏領域を切り拓いてきた．こうした中で生み出されたいくつかの研究成果が本特集に取り上げられている．

吉田論文は，マテリアルズインフォマティクスの基本的なワークフローである順方向と逆方向の予測をデータ科学の視点から整理した．順問題の目的は, 材料の構造から特性を予測するモデルを導くことである．これに対し，逆問題でモデルの逆写像を求めて所望の特性を有する材料を予測する．このワークフローを材料の “表現・学習・生成” という観点から整理し，データ科学の諸問題を論じている．また，材料研究における様々な逆問題を解説している．

吉田・ウ・森川論文は，高分子材料の設計における機械学習の適用事例を示している．同グループは，ものづくりデータ科学研究センターの研究者が中心となり開発したベイズ推論に基づく分子設計のアルゴリズム [2] を適用し，従来の高分子に比べて約 80% の高熱伝導率を有する新しい高分子の合成に成功した [3, 4]．本研究は，機械学習が自律的に設計した高分子が実際に合成された数少ない事例の一つである．本論文の趣旨の一つは，データ科学の実践の過程から浮かび上がってきたマテリアルズインフォマティクスが抱える問題点を切り取ることである．

ウ・山田・林・ザメンゴ論文は，高分子材料のマテリアルズインフォマティクスのレビューを行っている．マテリアルズインフォマティクスの研究対象の中でも，高分子という材料には独特の難しさがあり，研究の進展・実践展開に大幅な遅れが生じている．データベース，高分子の複雑性と表現の問題，材料特性の予測，設計という四つの切り口から，高分子インフォマティクスの現状と諸問題を論じている．

劉・山田・ウ論文は，材料研究のスモールデータの問題を論じている．材料研究のデータの量は，データ科学の他の応用分野に比べると圧倒的に少ない．また，コミュニティ全体でコモンデータを創出しようという動向も極めて低調である．スモールデータの壁を乗り越えるデータ科学の技術である転移学習に着目し，複数の実例を紹介しながらその潜在的な学習能力を実証している [5, 6]．特に，転移学習を適用することで，データの範囲外に存在する材料の特性予測に成功した複数の事例を紹介していることが本論文の特色である．一般に革新的な材料の周辺にデータは存在しない．一方，従来のデータ科学では，基本的に入力が近ければ出力も近いという内挿的な予測を行う．転移学習から得られたモデルには，本論文で示された事例のように，外挿性が備わっていることがしばしば観測されている．

郭論文は，有機分子の合成経路の設計に関する機械学習の進展を解説している．コンピュータで合成経路を自動設計するという研究は，有機化学の分野で 50 年以上前から研究が進展してきた．その先駆者である有機化学者の Elias James Corey 博士は，2011 年にノーベル化学賞を受賞している．その後，脈々と継承されてきた研究の潮流が，近年の機械学習の進歩に合流することで，従来の発想とは全く異なる新しい技術を生み出そうとしている [7]．同論文を通じて，この転換期にデータ科学が果たした役割や技術変化の様子を垣間見ることができる．

本特集のタイトルは「マテリアルズインフォマティクスの最前線」である．最初に本特集を企画したのは 2019 年 3 月頃であった．この巻頭言を執筆している現在が 2021 年 6 月なので，本特集は 2 年以上前に企画されたことになる．マテリアルズインフォマティクスの学術体系はまだ発展途上の段階にある．このような技術の勃興が激しい創成期の学術領域において，2 年という年月はとても長く感じられる．本特集のトピックスは，企画当時は確かに「最前線」であった．しかしながら，2021 年 6 月現在，これらの研究はすでに最前線にはない．フロンティアの位置は時々刻々と変化している．本特集の著者らもすでに次のステージの研究に向かっている．5 年後に本特集を読み返したときに，我々はどう感じるのか．そんなことを本特集の著者らと話している．マテリアルズインフォマティクスはデータ科学，計算科学，実験科学の合流点に位置する．これらの全ての進歩が新しい科学的手法を生み出し，材料研究の在り方を刷新

していくに違いない．特に近年のデータ科学は，最先端の研究が応用分野に合流する時間差が急速に短くなってきている．5 年後の「マテリアルズインフォマティクスの最前線」が果たしてどうなっているのか．退屈とは無縁の日々を過ごせることは間違いなさそうである．

1. 統計数理研究所プレスリリース (2017).「ものづくりデータ科学研究センターの設立について」，https://www.ism.ac.jp/noesuisin/news/monodukuri-opening.html.
2. Ikebata, H., Hongo, K., Isomura, T., Maezono, R. and Yoshida, R. (2017). Bayesian molecular design with a chemical language model, *Journal Computer-Aided Molecular Design*, **31**, 379–391.
3. Wu, S., Kondo, Y., Kakimoto, M.-A., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., Schick, C., Morikawa, J. and Yoshida, R. (2019). Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm, *npj Computational Materials*, **5**, 66.
4. 統計数理研究所プレスリリース (2019).「機械学習の「記憶」を活用し、高分子の熱伝導性の大幅な向上に成功」，https://www.ism.ac.jp/ura/press/ISM2019-07.html.
5. Yamada, H., Liu, C., Wu, S., Koyama, Y., Ju, S., Shiomi, J., Morikawa, J. and Yoshida, R. (2019). Predicting materials properties with little data using shotgun transfer learning, *ACS Central Science.*, **5**, 1717–1730.
6. 統計数理研究所プレスリリース (2019).「物性予測タスク訓練済みモデルの包括的ライブラリ XenonPy.MDL を公開～転移学習で材料インフォマティクスのスモールデータの壁を乗り越える～」，https://www.ism.ac.jp/ura/press/ISM2019-09.html.
7. Guo, Z., Wu, S., Ohno, M. and Yoshida, R. (2020). Bayesian algorithm for retrosynthesis, *Journal Chemical Information and Modeling*, **60**, 4474–4486.

統計数理(2021)
第 69 巻 第 1 号 5–33

特集「マテリアルズインフォマティクスの最前線」
[総合報告]

マテリアルズインフォマティクス概説

吉田 亮†

(受付 2020 年 11 月 2 日；改訂 2021 年 4 月 2 日；採択 4 月 2 日)

要　　旨

マテリアルズインフォマティクスの問題の多くは，順問題と逆問題の形式に帰着する．順問題の目的は, 系の入力に対する出力の予測である．例えば，入力変数は材料の構造，出力変数は物性に相当する．これに対し，逆問題では文字通り逆方向の予測を行う．すなわち，出力の目標値を所与とし，それを達成する入力変数を予測する．データ科学の文脈では，このワークフローは材料の “表現・学習・生成” を行うことに相当する．記述子と呼ばれる特徴ベクトルを用いて材料の構造を “表現” し，データのパターンに基づいて構造から物性の数学的写像を “学習” する．さらに，モデルの逆写像を求めて所望の物性を有する材料を “生成” し，有望な候補を同定する．解析対象の変数は，分子，組成，結晶，混合物，プロセス，合成経路など，問題に応じて多様な形式をとる．本稿は，材料の表現・学習・生成という概念に基づき，マテリアルズインフォマティクスの諸問題と解析手法を概説する．

キーワード：物性，材料設計，合成，逆問題，記述子，生成モデル．

1. はじめに

一般に材料研究のパラメータ空間は極めて広大である．例えば，有機低分子化合物のケミカルスペースには，およそ 10^{60} 個の候補分子が存在すると言われている(Kirkpatrick and Ellis, 2004)．一方，公共の化合物データベースに登録されている有機化合物の個数は高々 10^8 のオーダーに過ぎない(Bolton et al., 2008; Wang et al., 2009; Irwin and Shoichet, 2005; Gaulton et al., 2012)．したがって，有機化合物のケミカルスペースには依然として広大な未踏領域が残されている．さらに，実用材料の研究開発では，プロセスや添加剤，溶媒選択などがパラメータに加わり，パラメータ空間の大きさは爆発的に増大する．マテリアルズインフォマティクス(MI: materials informatics)の問題の多くは，このような広大な探索空間から所望の特性を有する未知パラメータを同定することに帰着する．これは多目的最適化の問題である．一般の工業品設計との本質的な違いは，パラメータ空間の特殊性と多様性にある．パラメータは，組成，分子，結晶構造，混合物，材料の微細構造，プロセス条件など，問題に応じて多様な形式をとる．

MI の最も基本的なワークフローは，順方向と逆方向の予測からなる(図 1)．順問題の目的は, 系の入力 S に対する出力 Y の予測である．例えば，入力は材料(分子，組成，結晶など)，出力は物性や材料の構造的特徴に相当する．これまでの材料研究では，第一原理計算や分子動力学計算など，物理法則に基づくシミュレーションが順方向の予測を担ってきた．このような膨大なコストを伴う計算を統計モデルに代替させることが，MI の主要課題のひとつである．これに対し，逆問題では文字通り逆方向の予測を行う．すなわち，出力 Y の目標値を定め，順

† 統計数理研究所：〒190–8562 東京都立川市緑町 10–3

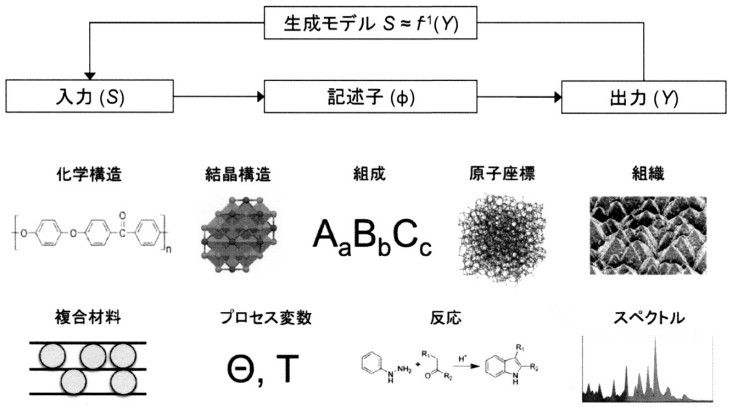


図 1. MI の基本的なワークフロー．入力 S(例えば，化学構造)から出力 Y(物性)の順方向の予測モデルを導き，出力 Y の目標値を近似的に達成する入力 S を逆向きに予測する．

方向のモデルの逆写像を求めることで，所望の出力を(近似的に)達成する入力 S を予測する．これらの計算は，材料の “表現・学習・生成” を行うことに相当する．記述子で材料構造の “表現” を行い，データのパターンから構造から物性の数学的写像を “学習” する．さらに，その逆写像を求めて所望の Y を有する材料 S を “生成” し，有望な候補を炙り出す．本稿では，材料の表現・学習・生成というコンセプトに基づいて，MI の様々な解析手法を概説していく．

MI のデータ解析の特殊性の一つは，変数の特殊性と高次元性にある．組成，分子，結晶構造など，一般に固定長ベクトルに基づく特徴表現が非自明な変数が解析対象になることが多い．したがって，我々が対峙する課題をデータ科学の枠組みに帰着させるには，変数の形式に応じて適切な記述子を用意しなければならない．また，逆問題を解くには，広大な探索空間を自由自在に走査できる S の生成モデルが必要になる．変数の形式の多様さゆえ，多くの場合，問題ごとに解析手法とソフトウェアを用意する必要がある．

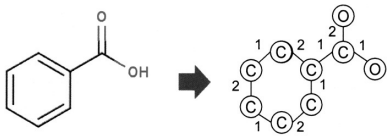
2. 物質・材料の表現

記述子は MI の最も基本的な要素技術である．入力 S の形式が多様であるがゆえ，解析対象ごとに様々な研究が進行している．また，後の節で紹介するように，出力変数が特殊な場合もある．ここでは，MI の最も基本的な入力変数として，化学構造，組成情報，結晶構造の記述子を解説する．

2.1 分子記述子

化学構造(2 次元構造)の最も自然な表現の形式は，ラベル付きの無向グラフである(図 2)．グラフ $G = (V, E, L_V, L_E)$ は，頂点集合 V とエッジ集合 E から構成される．頂点は原子，エッジは結合を表し，頂点には元素種($L_V = \{\mathrm{C}, \mathrm{O}, \mathrm{N}, , \ldots, \mathrm{F}\}$)，エッジには結合結合次数 $L_E = \{1, 2, 3\}$ を表す属性が与えられる．

SMILES(Simplified Molecular Input Line Entry System: Weininger, 1988)は，化学構造を文字列で記述する表記法である．原子を元素記号で表し，特別な文法に従うことで，環構造，分岐，結合次数，同位体，不斉中心などを厳密に記述できる．全ての化学構造は，SMILES の文字列に変換できる．例えば，図 3 に示したバニリン(vanillin $C_8H_8O_3$)の SMILES 表記は O=Cc1ccc(O)c(OC)c1 となる．環構造は，始点と終点の原子の後に同じ数字(ここでは 1)でラ



グラフ	$G = (V, E, L_v, L_e)$
頂点集合(原子)	$v \in V$
エッジ集合(結合)	$e \in E$
頂点ラベル(元素種)	$L_V = \{\mathrm{C}, \mathrm{O}, \mathrm{N}, \ldots, \mathrm{F}\}$
エッジラベル(結合次数)	$L_E = \{1, 2, 3\}$

図 2. 化学構造のグラフ表現.

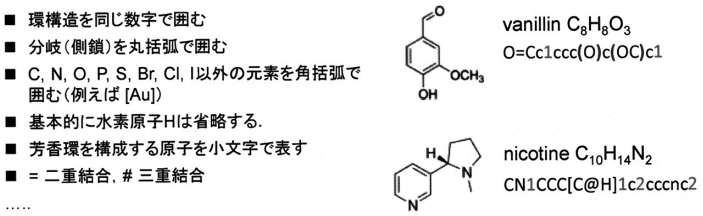


図 3. SMILES による化学構造の文字列表現.

ベリングされる．丸括弧は分岐(側鎖)を表す．等号 “=” は二重結合を表す．芳香環を構成する原子を小文字で表すというルールにより，環を構成する炭素は小文字の “c”，それ以外の炭素は大文字の “C” と表す．また，水素原子は原子価に基づいて暗黙に付加するという方針のもと，特別な場合を除いて，水素原子は省略される．SMILES の文法規則は，直感的に理解しやすく，少ないバイト長で一次元的(線形)に構造を表現できる．また，SMILES の文字列が与えられると，分子の構造式は一意に決まる．このような利点により，SMILES は化学の分野で最も広く利用されるデータ形式となった．

分子フィンガープリント(molecular fingerprint)は，化学構造の最も基本的な記述子である．部分構造(フラグメント)の集合 $\mathcal{F}$ に対し，各フラグメント $f_i \in \mathcal{F}$ の有無(バイナリ型)や頻度(カウント型)に基づき化学構造のパターンを数値化する(図 4)．バイナリ型記述子ベクトル $\phi(S)$ の要素 i は，S がフラグメント f_i を持てば 1，そうでなければ 0 をとる．カウント型記述子は f_i の個数を要素に持つ．通常，記述子ベクトルの長さは $O(10^2)$-$O(10^3)$ 程度となる．

これまでに数多くのフィンガープリントアルゴリズムが開発されてきた．これらの違いはフラグメント集合の構成方法による．表 1 に，Python のケモインフォマティクスライブラリ RDKit (Landrum, 2016)と R 言語のライブラリ rcdk (Guha et al., 2007)に実装されているフィンガープリント記述子の一覧を示す．フィンガープリントには，何らかの目的(例えば，物性予測)に基づいて選定された所与のフラグメント集合を用いるタイプ(事前定義型)と，入力された化合物の集合からある制約を満たす全てのフラグメントを列挙するタイプ(列挙型)がある．

事前定義型は，あるタスクを目的に事前に定義されたフラグメント集合を数え上げる．例えば，Klekota-Roth フィンガープリントの 4,860 個のフラグメントは，薬剤分子の薬理活性のデータに基づき，活性レベルが高い化合物に頻出する部分構造を選定したものである(Klekota and Roth, 2008)．また，PubChem フィンガープリントは，881 次元のバイナリ型記述子である

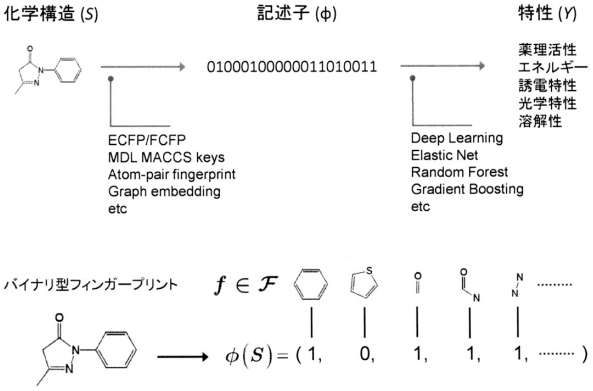


図 4. フィンガープリント記述子に基づく構造物性相関分析.

表 1. RDKit (Python)と CDK(Java)/rcdk(R 言語)ライブラリに実装されているフィンガープリント記述子の一覧.

パッケージ	名前	アルゴリズム
RDKit	RDK	Daylight-like fingerprinting
	Layered	Daylight-like fingerprinting
	Atom-pairs	Carhart et al. (1985)
	Morgan	Similar to ECFP/FCFP (Rogers and Hahn, 2010)
	MACCSkeys	166 bits MDL MACCS keys
	TopologicalTorsion	Topological torsion fingerprint (Nilakantan et al., 1987)
	Pattern	Pre-defined structural pattern
	E-state	Hall and Kier (1995)
CDK (Java) / rcdk (R)	standard	Paths of a given length (1024 bits)
	extended	standard + ring + atomic property
	maccs	166 bits MDL MACCS keys (Durant et al., 2002)
	circular	ECFP6 (Rogers and Hahn, 2010)
	pubchem	881 bits PubChem fingerprint (Bolton et al., 2008)
	graph	standard + connectivity
	kr	4,860 bits Klekota-Roth (Klekota and Roth, 2008)
	hybridization	standard + Info. on hybridization
	shortestpath	shortest paths between atoms
	signature	count type of fingerprint

(Bolton et al., 2008). 各要素は，特定の元素，環構造，結合，部分構造の個数に対する条件を満たす場合に 1 となる．PubChem はアメリカ国立衛生研究所(NIH)が保有する化合物のデータベースである．PubChem フィンガープリントは，類似構造の検索に用いられている．他にも構造検索由来のフィンガープリントとして，MDL 社が開発した MACCS Keys(166 フラグメント)も有名である(Durant et al., 2002).

列挙型のフィンガープリント記述子は，解析対象の化合物集合からある条件を満たすフラグメントを全列挙し，フラグメントの有無や個数で構造を表現する．代表例として，Morgan フィンガープリント(別名：Circular フィンガープリント)や ECFP(extended connectivity fingerprint)，FCFP(functional-class fingerprint)が列挙型に分類される(Rogers and Hahn, 2010).

事前定義型では，定義されたフラグメント集合が解析対象の化合物集合に対して冗長な場合，過剰に疎なベクトル表現になってしまう．一方，列挙型はデータに応じてフラグメント集合を規定するため，一般に柔軟性が高い．

ここで，列挙アルゴリズムの一例として，Morgan アルゴリズムに基づき設計された ECFP フィンガープリントの計算手法を解説する．計算手順を Algorithm 1 に示す．

Algorithm 1 ECFP フィンガープリント．

Input 化学構造 S，半径 R，フィンガープリントの長さ B
Output バイナリ型フィンガープリントベクトル $\phi(S)$

Initialize: 各原子に属性を表す整数ベクトル r_n $(n = 1, \ldots, N)$ を割り振る． #N は原子数
Initialize: $\phi(S) \to 0$ #フィンガープリントベクトルの初期化
for $r \in \{1, \ldots, R\}$ **do**
  for $n \in \{1, \ldots, N\}$ **do**
    $(r_n, \boldsymbol{r}_{\mathcal{A}_n}) = \mathrm{get}(n, \mathcal{A}_n)$ #対象原子 n と隣接原子 $\mathcal{A}_n$ の属性値ベクトルを取り出す.
    $v = \mathrm{concat}(r_n, \boldsymbol{r}_{\mathcal{A}_n})$ #属性値ベクトルの全ての要素をつなげる．
    $r_n = \mathrm{hash}(v)$ #ハッシュ関数を用いて，v を整数値 r_n に変換
    $i = \mathrm{mod}(r_n, B) + 1$ #剰余演算を行い，1 から B の範囲に
    $\phi_i(S) \leftarrow 1$
  end for
end for

各原子に属性を表すベクトルを割り振り，各原子に隣接原子の属性ベクトルを伝播させる．対象原子と隣接原子の属性ベクトルの全要素をつなげた整数値に対し，ハッシュ関数を適用し，ユニークな整数値を取得する．このハッシュ値は，対象原子の周辺構造を数値化したものと見なされる．最後にフィンガープリントのベクトルの長さ B でハッシュ値を除算し，その剰余のアドレスにビットを立てる．この操作を R 回繰り返せば，第 R 近接の全ての周辺構造を数え上げることができる．

ECFP の属性ベクトルは，(1)原子番号，(2)隣接する重元素原子の個数，(3)結合する水素原子の個数，(4)形式電荷，(5)環の構成原子かどうかを表す二値変数(0/1)からなる．一方，ECFP と同じ論文で発表された FCFP(functional-class fingerprint)では，リガンドの結合に関する特性(水素ドナー，アクセプター，極性，芳香族性の有無を表す二値変数)を属性ベクトルとする．

ECFP のアルゴリズムは，1965 年に発表された Morgan アルゴリズム(Morgan, 1965)に由来する．目的は，原子の周辺環境を縮約したユニークな属性値を算出することである．原子番号などの初期値を各原子に割り当て，隣接原子の属性値を縮約し，自己の属性値を更新する．このような計算を R 回繰り返し，第 R 近接までの隣接原子の結合パターンを反映した属性値を計算する．このようなグラフ上の演算は，ECFP などの列挙型記述子だけでなく，後述のグラフ畳み込みニューラルネットワークの計算にも継承されている．

ECFP に代表される多くのフィンガープリント記述子は，ある原子を中心とした部分構造のパターンを表現する．このような記述子は，分子の形などの大域的特徴の表現には適さない．したがって，実際のデータ解析では，局所構造の記述子に加えて，大域的な特徴や物理化学的な量的特徴を表す記述子を組み合わせてモデルを作る(Burden, 1989, 1997; Moreau and Broto, 1980; Moriwaki et al., 2018)．アトムペアフィンガープリント(atom-pair fingerprint)は，分子中の全ての重原子の組合せを数え上げる(Carhart et al., 1985)．ECFP などは局所的な部分構

造のみを数え上げの対象としているが，アトムペアフィンガープリントは遠く離れた原子間の情報を取り込むことができる．元素種，隣接する重原子の数，π 結合の数に基づき原子のタイプを類別化し，任意の 2 タイプの原子ペアとその距離として最短結合数を考える．トポロジカル二面角フィンガープリント(topological torsion fingerprint)は，二面角を形成する全ての 4 原子を数え上げる(Nilakantan et al., 1987)．アトムペアフィンガープリントの自然な拡張となっているが，トポロジカル二面角フィンガープリントは局所的なパターンのみを数え上げの対象としている．

ECFP や FCFP では，複数の部分構造がベクトルの一つの要素に割り当てられるケースが生じる．これをビット衝突問題(bit collision)という．ビットの重複は，剰余演算により強制的に長さ B のベクトルに縮約する操作から生じる．この問題を回避するために，主に 2000 年代にグラフを対象とした正定値カーネル(グラフカーネル)の研究が活発に行われた(Vishwanathan et al., 2010)．グラフカーネルは，ある大きさ以下の全ての部分構造を数え上げ，頻度情報を加算無限個の要素を持つベクトルに縮約する．パス(Gärtner et al., 2003; Kashima et al., 2003)や木構造(Mahé and Vert, 2009; Mahé et al., 2004; Yamashita et al., 2014)など，部分構造の型に制限が設けられる．多くの場合，動的計画法を適用することで，加算無限個のベクトルの内積(カーネル)を高速に計算できる．

2.2 組成・構造記述子

MI のデータ解析における最も基本的な入力変数は化学組成(あるいは原料組成)である．元素を大文字，組成を小文字で表し，組成式を $S = S_{c^1}^1 \dots S_{c^K}^K$ と表す．ここで，以下のような組成記述子のクラスを考える．

$$\phi_{f,\eta}(S) = f(c^1, \dots, c^K, \eta(S^1), \dots, \eta(S^K)). \tag{2.1}$$

右辺の $\eta(S^k)$ は，原子番号，電気陰性度，分極率など，元素 S^k の特徴量を表す(Seko et al., 2017; Ward et al., 2016)．元素特徴量 $\eta(S^1), \dots, \eta(S^K)$ と組成 $c^1, \dots, c^K$ に関数 f を適用して，記述子ベクトルの一つの要素を計算する．f は，加重平均，幾何平均，加重分散，最大プーリング，最小プーリング，加重和などに相当する．我々が開発している Python のライブラリ XenonPy には，58 種類の元素特徴量が実装されている．原子番号，結合半径，ファンデルワールス半径，電気陰性度，熱伝導率，バンドギャップ，分極率，沸点，融点などから構成される(Liu et al., 2021)．

式(2.1)において，結晶中の各原子の局所的な配位環境を表す特徴量を $\eta(S^k)$ に設定すれば，結晶構造の記述子となる．Seko et al. (2017) は局所構造の特徴量として，pRDF(partial radial distribution function)，GRDF(generalized radial distribution function)，AFS(angular Fourier series)を用いている．結晶構造の記述子は，その他にも数多く存在する．例えば，Behler's radial symmetry function (Behler and Parrinello, 2007)，Oganov's fingerprints (Oganov and Valle, 2009)，SOAP(smooth overlap of atomic positions)(Bartók et al., 2013)などが挙げられる．さらに，格子定数，バンドギャップ，状態密度など，物性の計算値や実験値を記述子 に含める方法も考えられる(Isayev et al., 2017)．しかしながら，結晶構造や物性値を含む記述子は，当然ながら計算コストが極めて高くなるため，そのモデルは材料スクリーニングの用途には適さない．

アモルファスやガラス，乱れのある系に対し，パーシステントホモロジーという数学理論を適用し, 原子配置の分布の位相情報を記述するという取り組みがある(平岡, 2015)．さらに，パーシステントホモロジーと機械学習を組み合わせることで，位相データ解析と呼ばれる方法論を体系化しようという試みがある．Kusano et al. (2016) は，分子動力学シミュレーション

から生成された原子配置データに位相データ解析を適用し，SiO_2 の液相-ガラス相の転移温度を検出している．

また，材料構造を画像で表現し，物性予測を画像認識の問題に帰着させるというアプローチがある．複合材料の微細組織を撮影した電子顕微鏡の画像を入力とし，畳み込みニューラルネットワークで材料特性を予測するという研究がある（Cang et al., 2017; Li et al., 2018b）．Hirn et al.（2017）は，近似電子密度に基づく記述子を提案している．計算が容易な元素単体の電子密度を計算しておき，これらの重ね合わせで物質全体の電子密度を近似する．この電子密度に散乱変換と呼ばれるウェーブレット変換を施し，原子の順番，並進，回転，対称性に対して不変な記述子を導いている．Carleo and Troyer（2017）は，入力をポテンシャル画像，出力を波動関数とし，深層学習による画像識別の問題に帰着されて多体電子系のシュレーディンガー方程式を解いている．

2.3 材料構造のグラフ表現とニューラルネットワーク

近年，材料の構造をグラフで表現し，グラフ系ニューラルネットワークを用いて物性を予測する研究がトレンドを形成している（Duvenaud et al., 2015; Schütt et al., 2017）．化学構造の自然な表現形式はラベル付き無向グラフである．また，結晶構造の周期的な原子配置は，結晶グラフ（crystal graph）と呼ばれる単位胞の原子の近接関係を核とする巡回型グラフで表現できる（Xie and Grossman, 2018）．一般に固定長ベクトルで表現できないグラフ形式の変数をニューラルネットワークの演算にどのように帰着させるか．これがグラフ系ニューラルネットワークの設計概念の中心をなす．

ここで，グラフを構成する N 個のノード $v_1, \ldots, v_N$ を属性値ベクトル $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ で特徴付ける．化学構造の場合，各ノードは原子，属性値ベクトルは原子の特徴量に相当する．例えば，元素種を表す one-hot ベクトル（全原子種の数と同じ長さを持つベクトルを用意し，該当する元素の要素のみを 1，その他を 0 とする）や原子量，電気陰性度などが属性値ベクトルを構成する．

グラフ畳み込みニューラルネットワーク（GCNN: graph convolutional neural networks）（Wu et al., 2020b）の核となる計算は，畳み込み層の演算である（図 5）．いま，N 個のノードに対し，第 l 層において隠れ変数 $\boldsymbol{h}_i^l \in \mathbb{R}^{k_l}$ $(i = 1, \ldots, N)$ が与えられているとする．第 l 層が入力層の

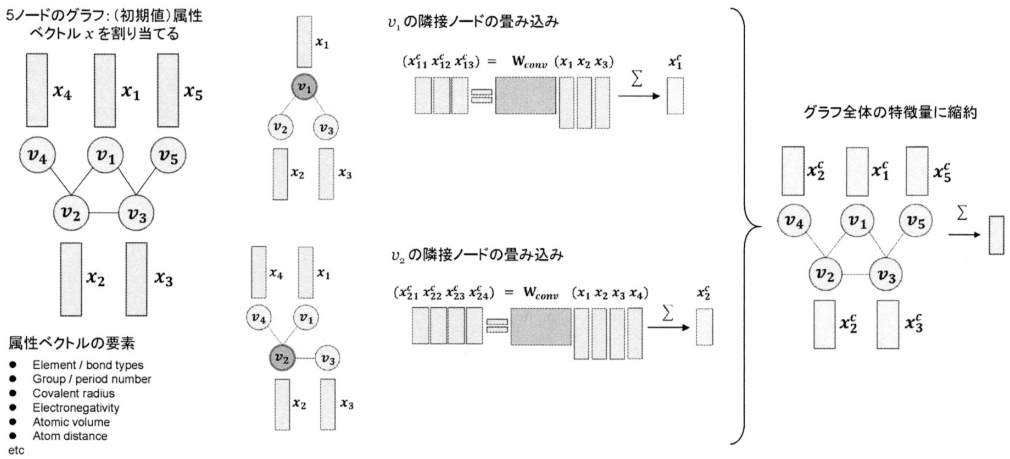


図 5．グラフニューラルネットワークの畳み込み層の計算．

場合，$\boldsymbol{h}_i^{(l)} = \boldsymbol{x}_i$ とする．これらのベクトルをグラフ上の畳み込み演算に基づき，第 $l+1$ 層の隠れ変数 $\boldsymbol{h}_i^{(l+1)} \in \mathbb{R}^{k_{l+1}}$ $(i = 1, \ldots, N)$ に変換する．ここで，ノード間の隣接関係を表す二値変数 a_{ij} を導入する．ノード i と j の間にエッジが存在すれば，a_{ij} は 1，そうでなければ 0 となる．このとき，第 $l+1$ 層の畳み込み演算は，以下のように表される．

$$\boldsymbol{h}_i^{(l+1)} = \sigma\left(W^{(l)} \sum_{j=1}^{N} a_{ij} \boldsymbol{h}_j^{(l)}\right). \tag{2.2}$$

$W^{(l)}$ は $k_{l+1} \times k_l$ の重み行列，σ は活性化関数である．この演算を各ノードに適用し，更新された長さ k_{l+1} のノード特徴量を得る．未知パラメータは $W^{(l)}$ である．

ここで示した例では，エッジの同一性を仮定していたが，分子や結晶構造のグラフは，エッジにも属性情報が与えられる．例えば，化学構造の場合，one-hot 表現などにより，エッジに結合次数(単結合，二重結合，三重結合など)の情報が付与される．また，結晶グラフには結合長(実数)が付与される．エッジ特徴量の取り扱いについては様々な方法が考えられる．例えば，式(2.2)に隣接するエッジの特徴量の畳み込み演算の項を加えればよい．また，エッジの特徴量は，隣接ノードの特徴量の畳み込み演算で更新される．

畳み込みの操作を L 回繰り返し，第 L 近接までのノードの属性値を合成してノード特徴量を計算する．最後に N 個の特徴量の総和 $\sum_{i=1}^{N} \boldsymbol{h}_i^{(L)}$ を取り，グラフ全体の特徴量に変換する．この時点で全てのグラフは固定長のベクトルに変換されるため，あとは従来型のニューラルネットワークを積層し，最後に出力層を構築する．

隣接ノードの属性値を段階的に集約していきながら，原子の局所環境を数値化していくという計算は，前述の Morgan アルゴリズムと類似性を持つ．GCNN では，畳み込み演算で集約の計算を実行する点と畳み込みの重み行列をデータから推定する点に特徴がある．

3. 物性の学習

3.1 仮想スクリーニング

実験や理論計算から得られたデータを用いて，構造 S から特性 Y を予測するモデル $Y = f(S)$ を導く．Y が連続変数の場合を回帰分析，離散変数の場合を判別分析という．ディープラーニング，ランダムフォレスト，エラスティックネット，ロジスティック回帰，サポートベクターマシン，ガウス過程回帰など，数多くの手法がある(例えば，機械学習の入門書(Friedman et al., 2001; Bishop, 2006)を参照)．膨大な数の候補材料のライブラリを作製した上で，訓練されたモデルを用いてスクリーニングを実施する．

機械学習を用いた材料スクリーニングは，創薬ではかなり古くから行われてきたが，材料研究に適用されるようになったのはごく最近のことである．Gómez-Bombarelli et al. (2016) は，第一原理計算のデータで学習したニューラルネットワークを用いて，400,000 個以上の候補物質のスクリーニングを実施し，高い外部量子収率を有する有機 LED の新規分子を発見した．Seko et al. (2015) は，第一原理計算で 101 個の無機化合物の格子熱伝導率を計算し，ベイズ最適化とガウス過程回帰を組み合わせて物性予測モデルを導いた．このモデルを用いて Materials Project (Jain et al., 2013)に登録されている 54,779 化合物のスクリーニングを行い，221 個の低熱伝導性物質を同定した．Carrete et al. (2014) は，32 個のハーフホイスラー化合物(half-Heusler compound)の熱伝導率の理論値を用いて，ランダムフォレストで回帰モデルを導き，AFLOW データベース(Curtarolo et al., 2012)に登録されている 450 化合物をスクリーニングした．Pilania et al. (2013) は，繰り返し単位が 4 ブロックの基本要素から構成される 175 個の高分子材料(ポリエチレン)に対し，第一原理計算で 8 種類の物性値(バンドギャップ，

生成エネルギー，誘電率など)を算出し，カーネルリッジ回帰を適用して各物性の予測モデルを構築した．このモデルを用いて，8 ブロックのポリマーユニットを持つ 29,365 個の高分子材料のスクリーニングを実施している．同研究グループは，その後，データセットを拡大し，Huan et al. (2016) において，データベース Polymer Genome (Chandrasekaran et al., 2020)を公開している．Wu et al. (2019)は，PoLyInfo という高分子物性データベースを用いて熱物性を予測するモデルを導き，高い熱伝導率を有する新規ポリマーを合成した．ベイズ推論に基づく分子生成アルゴリズムで仮想ライブラリを作製し，高い熱伝導率を持つと予想された 3 個のポリマーを絞り込み，実験検証を行った．少数のデータから熱伝導率の予測モデルを導くために，転移学習という解析手法を適用している．高分子のガラス転移温度や比熱など，熱伝導率と相関を持つ他の物性データから予測モデルを導き，小数のデータを用いて訓練済みモデルのファインチューニングを行い，高精度な熱伝導率の予測モデルを導いた．

3.2 スモールデータと転移学習

材料研究のデータの量は，データ科学の他の応用分野に比べると圧倒的に少ない．原因として，次の三点が考えられる：(1)データ取得の高コスト性; (2)研究者のニーズや設計パラメータ(作製方法，実験条件への依存性など)の多様性によるコモンデータベース創出の難しさ; (3)競合相手に対する情報秘匿の意識が高く，データ公開に対するインセンティブが研究者に働きにくい．したがって，オープンデータベースの開発が中々進まない．さらに，先端領域に近づくにつれて，スモールデータの傾向はより顕著になる．また，コミュニティ全体でコモンデータを創出しようという動向も極めて低調である．少なくとも短中期的には大学のラボや一企業で生産可能なデータが MI の標準的な解析対象になることが予想される．

転移学習は，あるタスクで事前に訓練されたモデルを他のタスクに転用するための解析手法である．少量のデータで機械学習のモデルを構築する際に広く使われるテクニックである．例えば，大量の画像データを用いて動物の種類を判定するニューラルネットワークを訓練し，少数の花の画像データを用いて訓練済みモデルを改変することで，花の種類の分類器を構築する．動物の分類器は，訓練の過程で汎用的な画像特徴量を獲得していることが期待され，その一部は花の分類器にも転用できる可能性がある．その場合，花の分類器を一から学習するのではなく，少数のデータで動物の分類器を修正すれば十分かもしれない．ヒトの脳には，少ない経験でも合理的に予測を行うメカニズムが備わっている，例えば，小さい頃からピアノを学んでいた人は，音楽に関する一般的な知識を獲得しているため，他の楽器の演奏技術を比較的容易に習得できる．転移学習はこのような学習過程を模倣する．

Yamada et al. (2019)では，4 つの具体例を示しながら，MI における転移学習の有効性を実証している．また同論文では，低分子，高分子，無機化合物の 45 種類の特性を対象に約 140,000 個の機械学習の予測モデルを開発し，訓練済みモデルライブラリ XenonPy.MDL を発表した．XenonPy は，同グループが開発している MI のオープンソースプラットフォームである．XenonPy には MI の様々なタスクを実行する機械学習のアルゴリズムが実装されており，ユーザーは API 経由で XenonPy.MDL の訓練済みモデルを再利用し，材料設計の様々なワークフローを構築できる．

Wu et al. (2019)では，転移学習を用いて高分子材料の熱伝導率の予測モデルを構築した．高分子物性データベース PoLyInfo に収録されている 28 種類のアモルファスポリマーの熱伝導率のデータを使用した．高分子のガラス転移温度，融点，比熱，粘度に加え，低分子化合物の比熱容量を元タスクとした．各々の元タスクに対して 100 個の異なるモデルを構築し，28 個の熱伝導率のデータを用いて訓練済みモデルを熱伝導率の予測モデルに転移した．交差検証により平均絶対誤差が最小の転移モデルを選定し，分子設計を実施した．最終的に 3 種類の新規高

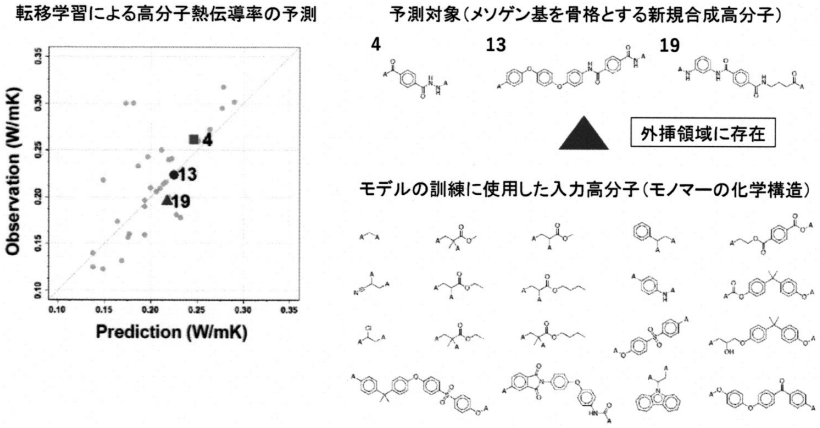


図 6. 転移学習による高分子熱伝導率の予測．左：3 種類の新規高分子に対する転移モデルの予測値と実測値．右：新規高分子のモノマーと転移学習で用いた訓練データの化学構造．

分子を合成し，熱伝導率の測定を実施した．図 6 に示すように，熱伝導率の実験値は転移モデルの予測値と概ね一致している．ここで注目すべき点は，合成した高分子との類似構造が訓練データにほとんど含まれていない点である．一般に機械学習は「入力が近ければ，出力も近い」という原理に基づき予測を行うため，訓練データの分布の近傍でのみ予測性能を有する．広範囲のケミカルスペースに適用できる汎用的な特徴量の獲得に有効な何らかの情報が元タスクのデータに含まれており，この特徴抽出器を再利用することで，訓練データの範囲外の入力に対しても予測性を持つモデルを構築できた．転移学習のモデルには，本事例のような外挿性が備わっていることがしばしば観測される（Yamada et al., 2019; Ju et al., 2019）．

4. 物質・材料の生成

構造から特性の順方向の予測モデル $Y = f(S)$ が得られたもとで，その逆写像 $S = f^{-1}(Y^*)$ を求めて，所望の特性 $Y = Y^*$ を有する構造 S を予測する．逆写像の計算では，厳密解だけでなく，$S \approx f^{-1}(Y^*)$ を満たす S も網羅的に抽出することが求められる．統計学者 John Turky は “An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem” という言葉を残している（Tukey, 1962）．物理の理論モデルとは違い，全ての統計モデルは正しくない．したがって，真に重要な仮説は厳密解の周辺に存在している可能性がある．厳密解を含む近傍の分布を網羅的に調べ上げ，専門家の知見に基づくスクリーニングなど，何らかの方法で分布の中から有望な候補を絞り込む．

4.1 化学構造の生成モデル

4.1.1 遺伝的アルゴリズム

データ科学的な手法に基づく化学構造の逆問題，すなわち分子設計は，ケモインフォマティクスという分野で古くから研究が進められてきた（Venkatasubramanian et al., 1994, 1995）．順方向のモデルを導くことを構造物性相関解析（structure-property relationship analysis）という．これに対し，その逆写像を求めることを逆構造物性相関解析（inverse structure-property relationship analysis）という．

ケモインフォマティクスの初期の頃には，逆構造物性相関解析において遺伝的アルゴリズムが広く用いられてきた．遺伝的操作を施して現在の化学構造を改変し，順方向のモデルを用いて目標値 Y^* への適合度を計算する．さらに適合度に応じて，次世代の候補構造を選抜する．適合度の算出には，構造物性相関解析で導いた物性予測モデルを用いることが多いが，任意のスコア関数を用いても構わない．例えば，薬剤分子の仮想ライブラリの作製では，候補構造の薬らしさを数値化するスコア関数（QED: quantitative estimation of drug-likeness（Bickerton et al., 2012; Wildman and Crippen, 1999））や合成可能性スコア（synthetic accessibility score（Ertl and Schuffenhauer, 2009））が適合度を構成する．前者は，RDKit の rdkit.Chem.QED module で計算できる．後者のコードは，GitHub で配布されている．

アルゴリズムの最も重要な構成要素は構造生成モデルである．構造生成モデルには，以下の要件を満たすことが求められる．

(1) 化学的に不適切な構造を生成しない．
(2) ケミカルスペースに存在する多様な化学構造を生成できる．
(3) 合成可能性が高く，化学的に安定な構造を生成できる．
(4) 任意の特徴を持つ構造群を生成できるようにモデルを柔軟にカスタマイズできる．ただし，要求される “特徴” のルールは，必ずしも明示的に書き下すことができない．

項目 (1) の要件は自明である．例えば，化学的に不適切な結合次数や配位数を持つ構造を排除する必要がある．合成研究者の独創力やセレンディピティが及ばない斬新な構造を発掘するために，生成モデルは項目 (2) を満たすことが望ましい．項目 (3) も自明である．項目 (4) は，生成モデルの汎用性についての要求性能である．例えば，有機薄膜太陽電池（OPV: organic photovoltaics）のドナー分子の探索では，OPV に特徴的な平面性の高い分子を生成することが求められる．耐熱性の高い高分子材料を探索したいときは，例えば，ポリイミド樹脂らしい構造を生成したいかもしれない．さらに，配向のしやすさなどの条件が加わることもある．このようなスペックは，そのルールを陽に書き下せない．個々の探索対象に対してフルスクラッチでモデルを構築するのではなく，パラメータなどを調整することで，様々なモデルを柔軟に構築できる汎用的な道具があれば便利である．

構造変換のための遺伝的操作では，元素やフラグメント（部分構造）をランダムに組み替える．以下に典型的な遺伝的操作の一覧を示す．

変異：選択された原子やフラグメントを他の要素に置き換える．
挿入：選択された位置に原子やフラグメントを追加する．
欠失：選択された原子やフラグメントを削除する．
伸長：選択された原子やフラグメントの複製を隣接位置に追加する．

化学構造の遺伝的操作の大きな特徴は，フラグメント単位での構造改変にある．例えば，実在の化合物からフラグメント（置換基，環構造など）の集合を抽出しておき，雛形 A-B-C の三つの構成単位 A, B, C にフラグメントを割り当てて仮想ライブラリを作製する．ここで，任意の化合物を断片化するアルゴリズムが必要になる．また，上述の変異，挿入，欠失，伸長などの遺伝的操作をフラグメント単位で実行する際にも断片化のアルゴリズムを適用する．

4.1.2 言語モデルによる分子生成

化学構造の遺伝的操作では，構造改変用の部品に既存化合物のフラグメントを使用することで，生成される構造の自由度を制限して探索空間を絞り込む．こうすることで，仮想ライブラリの合成可能性の向上を図る．しかしながら，探索空間の過度な絞り込みは，構造の新規性を

低下させるかもしれない．この点を克服するために，主に機械学習の研究者らが有機化学の世界に進出し，従来の発想とは全く異なるアプローチで分子生成の問題に取り組んでいる．2018 年頃を境にこの流れを汲んだ研究成果が相次いで発表された．

ここでは，Ikebata et al.（2017）で提案された確率的言語モデル（拡張 n グラム）による構造生成手法を紹介する．訓練データ集合に用いる既存化合物の化学構造を SMILES 形式で記述する．S は長さ p の文字列 $S = s_1 s_2 \dots s_p$ で表現される．この文字列集合を用いて n グラムのモデルを訓練し，既存分子に現れるパターン（頻出フラグメントや適切な化学結合のルールなど）を模倣した構造生成モデルを構築する．ここで，文字列 S の確率分布 $p(S)$ を条件付き確率の積で表現する．

$$p(S) = p(s_1) \prod_{i=2}^{p} p(s_i | s_{1:i-1}) \tag{4.1}$$

i 番目の文字 s_i の出現確率は先行する $s_{1:i-1} = s_1 \dots s_{i-1}$ に依存する．一般に同一の化学構造に対する SMILES の表現は一意ではない．このような構造的に等価な文字列を異なる S として扱う．言語モデルに基づく構造生成の基本的なコンセプトは，以下の通りである．既知の化合物の部分文字列の頻度から条件付き確率 $p(s_i|s_{1:i-1})$ を推定し，訓練されたモデルに化学言語のコンテキストを学習させる．所与の部分構造 $s_{1:i-1}$ に対し，モデルを用いて残りの文字列を生成する．条件付き確率に従い，終了コードが出現するまで文字を一個ずつ追加していく．言語モデルは SMILES の文法規則に合致する文字列を生成しなくてはいけない．ここで，環構造と側鎖などに関する分岐表現の文法規則が技術的な難しさになる．Ikebata et al.（2017）では，条件付き分布 $p(s_i|s_{1:i-1})$ のモデリングを工夫したり，SMILES 文字列の単語定義を改変することで，この問題の解決を図っている．

Ikebata et al.（2017）は，確率的言語モデルとベイズ推論を組み合わせて，所望の特性を有する分子を設計する手法を開発した．この手法は XenonPy の iQSPR-X というモジュールに実装されている（Wu et al., 2020a）．ベイズ推論による分子設計では，条件付き確率のベイズ則に基づいて順方向のモデルを逆方向の予測モデルに変換する．

$$p(S|Y \in U) \propto p(Y \in U|S) p(S) \tag{4.2}$$

訓練データを用いて S から Y の順方向の予測モデルを構築する．このモデルを用いて条件付き分布 $p(Y|S)$ を定める．このモデルから任意の S が所望の特性の範囲 U に入る確率 $p(Y \in U|S) = \int_U p(y|S)\mathrm{d}y$ を計算する．さらに，事前分布 $p(S)$ を用いて有望な探索空間を絞り込む．左辺の条件付き確率分布 $p(S|Y \in U)$ は事後確率分布である．この条件付き確率分布から SMILES の文字列 S をサンプリングすることで，所望の特性 $Y \in U$ を満たす新規分子を同定する．

4.1.3 深層生成モデルによる分子生成

近年，深層生成モデルと呼ばれるニューラルネットワークに注目が集まっている．特に，音楽（Jaques et al., 2017），画像（Choi et al., 2018; Zhu et al., 2017; Yi et al., 2017; Isola et al., 2017），アート作品（Elgammal et al., 2017）などの自動生成・変換・編集において，深層生成モデルは驚くべき性能を発揮することが分かってきた．このような時流の中，分子生成のタスクに深層生成モデルを適用する研究が 2018 年頃を境に急速に活発化した（サーベイ論文：Elton et al., 2019; Sanchez-Lengeling and Aspuru-Guzik, 2018）．これらの手法には，技術面で発展途上な点も多く残されている．しかしながら，MI の創成期と空前の機械学習ブームとの接点から生み出された象徴的な技術として，本稿ではこの話題を取り上げる．

深層学習に基づく SMILES 生成器を最初に提案したのは Gómez-Bombarelli et al. (2018)である(ArXiv でのプレプリント公開は 2016 年 10 月)．この論文では，変分自己符号器(VAE: variational autoencoder)という生成モデルが用いられている．モデルは，エンコーダとデコーダという二つのニューラルネットワークから構成される．エンコーダは，SMILES 文字列 S を固定長・実数型の潜在変数 Z に変換する．デコーダは，任意の潜在変数 Z から SMILES 文字列 S への変換を定める．エンコーダには，再帰型ニューラルネットワーク(RNN: recurrent neural network)あるいは言語用に設計された畳み込みニューラルネットワークが用いられる．デコーダには RNN を用いる．当該論文では，ZINC データベースから抽出した約 250,000 の市販分子や約 100,000 個の有機 EL 用の仮想ライブラリを用いて VAE を訓練している．訓練済みモデルから計算される潜在変数は化学構造の特徴量である．この固定長ベクトルを入力とし，特性 Y を予測するモデル $Y = f(Z)$ を構築できる．回帰モデルの入力 Z は連続変数となる．したがって，例えば，特性の最適化 $Z^* = \arg\max_Z f(Z)$ には，勾配計算を用いた連続最適化のアルゴリズムを適用できる．さらに，デコーダに Z^* を入力すれば，化学構造 S を得ることができる．このように VAE を導入することで，化学構造の(連続)表現，特性予測，最適化，構造生成というタスクを統一的なフレームワークの中でシームレスに実行できる．

Segler et al. (2018)は，LSTM(long short term memory)による化学構造の生成を初めに提唱した論文である(ArXiv でのプレプリント公開は 2017 年 1 月)．モデルは特筆すべき点はない普通の LSTM である．ChEMBL という化合物データベースに登録されている 140 万個の化合物の SMILES を訓練データに使用した．SMILES のトークンの数は 51 個である．初期構造(部分文字列)から開始して，モデルの出力確率に基づいて 1 文字追加し，さらに追加された文字を入力とする．この再帰計算を終了コードが出力するまで繰り返す．Yang et al. (2017)は，LSTM とモンテカルロ木探索(Monte Carlo tree search)を組み合わせて，分子設計アルゴリズム ChemTS を開発した．LSTM を用いて SMILES の文字をノードとする探索木を伸長・分岐させながら，報酬(目標特性との近さ)を最大にする文字列を探索するという手法である．

Segler et al. (2018)，Gómez-Bombarelli et al. (2018)のプレプリント公開を境に，深層学習に基づく分子生成の研究は大きなブームになった．Elton et al. (2019)の表 1 に，膨大な数の論文のリストと概要がまとめられている．SMILES 系のモデルだけでなく，グラフ系の深層ニューラルネットワークを用いたモデルも数多く提案されている．例えば，Kipf and Welling (2016)では，GCNN をエンコーダとし，潜在変数から隣接行列および元素・結合ラベルへのデコードをニューラルネットワークでモデル化している．2018 年に機械学習の国際会議 ICML(International Conference on Machine Learning)で発表された Jin et al. (2018)の JT-VAE (junction tree variational autoencoder)はグラフ系の分子生成手法のベンチマークモデルとなっている．化学構造を Junction Tree と呼ばれる木構造に変換するアルゴリズムを用いている．モデルは，木構造のデコーダ・エンコーダと，さらに木構造から元の分子グラフに変換するデコーダから構成される．

これらの手法は，大量の化学構造のデータから実在分子の骨格や結合パターンを学習し，広大な化学空間を走査できる生成モデルを構築する．このような生成モデルを用いて逆問題を解くことで，斬新な候補分子を同定できるかもしれない．ただし，このようなアプローチは，斬新な構造を得る代償として，化学的に不適切な構造や合成可能性が低い構造を大量に生成してしまう．例えば，VAE のデコーダをそのまま適用すると，SMILES の構文規則(環構造の開閉サイクル，分岐，許容原子価など)を満たさない無効な構造が大量に出力されることがある．Jin et al. (2018)は，VAE で生成した分子の内，約 99% が不適切な化学構造であったと報告している．実際には，生成された構造の化学的な妥当性を事後的に検査し，有効な構造だけを用いる．本来は離散変数として取り扱われるべき化学構造を強制的に連続変数に変換すること

で，潜在空間の中に存在しえない構造を含む大きなデッドゾーンが生じる．この問題は，RNN や VAE だけでなく，微分可能な非線形関数である全てのニューラルネットワークに共通する．ベンチマーク指標(適用可能な化学空間の広さや生成された構造の正しさなど)を開発し，分子生成モデルの性能を系統的に評価しようという試みも始まっている(Brown et al., 2019)．また，フラグメント組み換えや点変異に基づく旧世代の手法との系統的な比較も必要である．XenonPy に実装されている n-gram による分子生成(Ikebata et al., 2017; Wu et al., 2020a)も選択肢の一つである．n-gram による分子生成では，訓練データに含まれる局所構造しか出現しないため，化学的ルールに違反する構造はほとんど生成されない．現段階では実践展開の経験が不足しており，乱立するこれらの手法の優劣を論じるには時期尚早かもしれない．

4.1.4 適用例：高熱伝導性高分子の探索

Wu et al. (2019)は，Ikebata et al. (2017)のベイズ推論に基づく分子設計アルゴリズムを適用して，高熱伝導率を有する新規高分子を発見した．データ解析のワークフローを図 7 に示す．一般に高い熱伝導率を持つ高分子材料は，軟化温度(ガラス転移温度 Tg)や溶融温度(融点 Tm)が十分に高く，高温まで軟化あるいは溶融しない．具体的には，融解しても取り得る配座構造の変化の少ない剛直な高分子ほど，融解のエントロピーが小さくなり，融点が高くなる．高分子のガラス転移点は，高分子の分子間力や屈曲性，対称性によって支配される．環構造の割合の多い主鎖構造を持つ高分子材料は，融解熱に関わる分子間相互作用ないしは凝集力が大きく，Tg が高くなる．

仮想ライブラリの作製では，高い Tg と Tm を持つ芳香族ポリアミドをターゲットとした．PoLyInfo データベースからホモポリマーの Tg と Tm のデータを抽出し，5,917 および 3,234 件のデータからランダムに 80% のサンプルを選択し，ベイズ線形回帰で順方向のモデルを構築した．入力変数の記述子には，モノマーの分子骨格のパターンを数値化した分子フィンガープリントを用いた．ここでは，ECFP などの複数のフィンガープリントを合わせて使用した．さらに，PoLyInfo の 14,423 個のホモポリマーを用いて言語モデルを訓練し，Tg と Tm の範囲

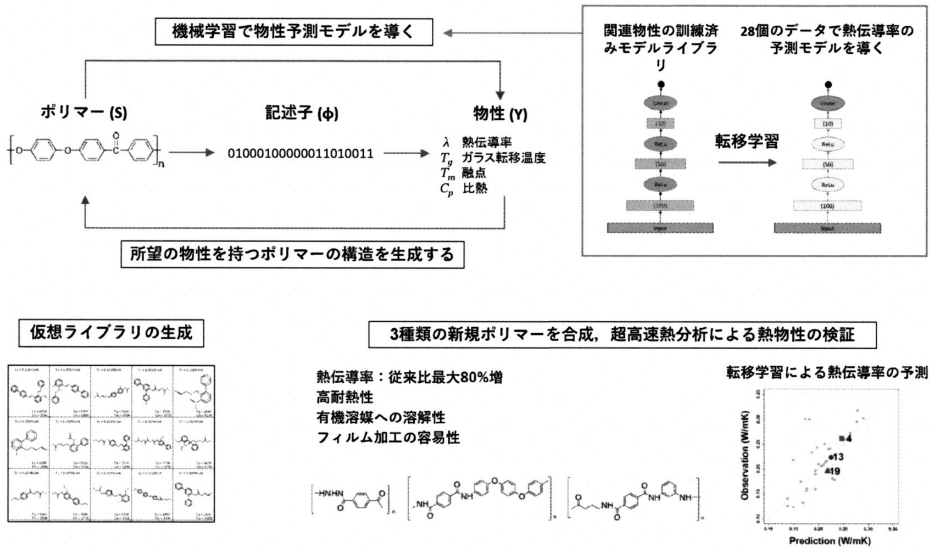


図 7．ベイズ推論に基づく高熱伝導性高分子の設計．

200-500°C，300600°C をターゲットに 1,000 種類の仮想ライブラリを作製した．ただし，溶融成形が可能な熱可塑性樹脂の設計では，耐熱性を若干犠牲する必要がある．このことから，事後選択の段階で Tg の温度の上限を 300 とした．

次に，機械学習モデルを用いて 1,000 個の候補分子の熱伝導率を推算した．前節で述べたように，PoLyInfo には熱伝導率のデータがたったの 28 件しかなかったため，通常の教師あり学習では物性予測モデルを構築できなかった．そこで，転移学習を導入して問題の解決を図った．事前学習には，高分子の Tg，低分子化合物の比熱容量などのデータを使用した．さらに，モノマー構造の液晶らしさや合成可能性のスコアリングを行い，最終的に 3 個の芳香族ポリアミドを絞り込み，合成と物性測定を実施した．合成された高分子の一つは，熱伝導率が 0.41 W/mK に達することが確認された．これは典型的な無配向のポリアミド系高分子と比較して約 80% の性能向上に相当する．さらに，高耐熱性や有機溶媒への溶解性，フィルム加工の容易性など，実用化に有利な様々な要求特性を併せ持つことが確認された．

4.2 材料微細組織の生成

材料の組織はプロセスと組成から決まる．さらに，材料組織が材料特性の主な支配要因となる．ここで，プロセス・組成，組織，特性の間をつなぐデータ科学のアイデアを述べる．材料組織をモデルの入出力として取り扱うために，電子顕微鏡の画像を用いるとしよう(図 8)．鉄鋼製品や高分子複合材の研究において，素材の表面や内部組織の解析に走査電子顕微鏡(SEM)や透過電子顕微鏡が用いられる．例えば，組成・温度依存的に制御される結晶粒の微細化を行い，材料の機械的性質を向上させたい．この問題を解くために，材料組織を画像として取り扱う．こうすることで，画像認識やコンピュータビジョンの解析手法を適用できる．例えば，材料組織から特性の予測は，入力が画像，出力が実数値となる．このタスクは通常の画像認識の問題設定(回帰)と同じである．Li et al. (2018b)は，畳み込みニューラルネットワークを用いてこの問題にアプローチしている．モデルの訓練では，ImageNet (Deng et al., 2009)という一般物体認識用に用意された大量の画像データで学習した VGG16 (Simonyan and Zisserman, 2014)という訓練済みモデルのファインチューニングを行っている．また，プロセス・組成から材料組織を予測する場合，入力は実数のベクトル，出力は行列(グレースケール画像)やテンソル形式(カラー画像)の画像データとなる．これは，多次元出力変数の回帰の問題である．あるいは，コンピュータビジョンにおける画像生成というタスクに帰着する(Li et al., 2018a; Cang et al., 2017)．

ここで，深層生成モデルを用いた材料の微細組織の予測の例を紹介する(Banko et al., 2020)．材料は，Cr が主成分の金属板に Al でコーティングした薄膜である．材料は，Cr が主成分の金属板に Al でコーティングした薄膜である．Cr 金属板と Al 金属板を向かい合わせに設置し，

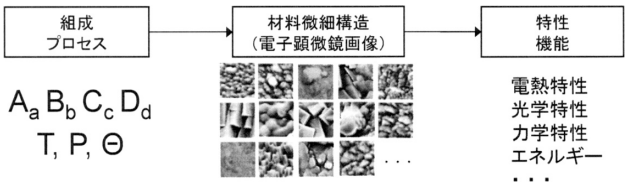


図 8. プロセスと組成からの材料微細組織の予測と組織から特性の予測．材料組織を電子顕微鏡画像で表現することで，前者のタスクは画像生成，後者のタスクは画像認識に帰着する．

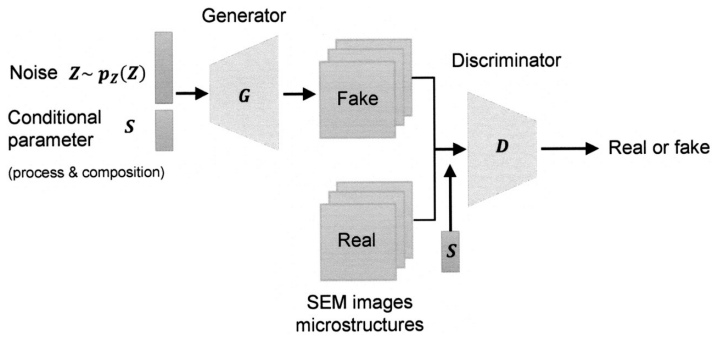


図 9. Conditional GAN のネットワークダイアグラム.

マグネトロンスパッタリング法で Ar ガスを Al 金属板に高速で吹きかけ，飛び出した Al 原子を Cr 金属板に吸着させる．モデルの入力は，組成 $\mathrm{Cr}_{1-c}\mathrm{Al}_c\mathrm{O}_d\mathrm{N}$ とプロセスを表す 6 次元の実数ベクトルである：

(1) Cr, Al の組成比 c
(2) O の組成 d
(3) Al を吸着させる際の温度
(4) Al を吸着させる際の圧力
(5) Al 金属板に入射させる時の Ar イオンの平均エネルギー
(6) Ar ガスの電離度

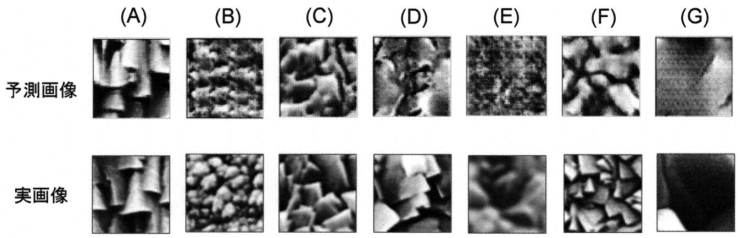
出力は材料組織の SEM 画像である．学習用の元データの数は 123 個である．ここから，128 ピクセル ×128 ピクセルの部分画像をランダムに 128 個抽出し，計 123×128 枚の画像をモデルの学習に使用する．

問題の形式は，入力が 6 次元の実数ベクトル S，出力が 128×128 の行列 Y という多次元出力変数の回帰分析である．ここで Banko et al. (2020) に従い，Conditional Generative Adversarial Networks (cGANs) (Mirza and Osindero, 2014) というニューラルネットワークを用いて，この問題を解く．図 9 に示すように，cGAN は generator (G) という画像生成モデルと discriminator (D) という判別モデルから構成される．$G = G(S, Z)$ の入力は，プロセスと組成を表すパラメータ S とランダムノイズ Z である．畳み込みニューラルネットワークを中心に構成されたモデルを介し，入力変数は SEM 画像に変換される．$D = D(Y, S)$ の入力は SEM 画像 Y とプロセスと組成を表すパラメータ S である．実際の SEM 画像あるいは G が生成した偽画像と入力変数が与えられたもとで，畳み込みニューラルネットワークでその真偽を判定する．G と D の学習は，次の minmax 戦略に基づいて実施される．

$$\min_G \max_D \mathbb{E}_{(Y,S)\sim p_{\mathrm{data}}(Y,S)}[\log D(Y,S)] + \mathbb{E}_{Z\sim p(Z),S\sim p_{\mathrm{data}}(S)}[\log(1 - D(G(S,Z),S))]. \tag{4.3}$$

第 1 項が大きくなるのは，$D(Y, S)$ が大きくなるとき，すなわち，本物の画像を正しく本物であると識別できた場合である．第 2 項が大きくなるのは，$1 - D(G(S, Z), S)$ が大きくなる，すなわち，G が生成する偽画像を偽物と識別できたときである．D は識別性能が最適になるように学習される．G は第 2 項を小さくするように学習される．すなわち，D を誤判定させるように学習が進行する．G と D を交互に訓練する中で，高品質の偽画像を生成する G を導く．

ノイズ $Z \sim p(Z)$ を抽出し，訓練された生成モデル $Y = G(S, Z)$ を用いることで，任意の組



入力変数

組成: $Cr_{1-x}Al_xO_yN$

プロセスパラメータ

- 蒸着温度 (T_d)
- 蒸着圧力 (P_d)
- 平均イオンエネルギー (E_i)
- 電離度 (I_d)

	(A)	(B)	(C)	(D)	(E)	(F)	(G)
T_d [℃]	20	200	300	500	500	600	800
P_d [Pa]	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Al [at.%]	0	17	0	20	0	27	0
O [at.%]	0	10	3	0	0	10	3
E_i [eV]	2	8.5	3	42	210	7	6
I_d [a.u.]	0.2	0.5	0.1	1	1	0.5	0.2

図 10. cGAN による材料組織の予測．7 種類の組成・プロセスに対する実際の SEM 画像と予測画像を示す．

成・プロセス S に対する材料組織の SEM 画像を予測する．図 10 は，7 種類の組成・プロセスを入力したときの SEM 画像の予測結果である．各組成・プロセスに相当するデータは，訓練データから除去している．結晶粒形やサイズの大まかな傾向を予測することに成功している．

cGAN は回帰のテクニックの一種である．特に，出力変数が多次元で入出力の関係が非常に複雑な系において有用なアプローチとなる．出力の空間が高次元の場合，有限個のサンプルでは情報が不足する．そこで，真のデータを模倣した人工的なデータを生成する．このように拡大したデータセットにモデルを適合させることで過学習を抑制する．出力変数の形式は，SEM 画像の場合は行列であったが，原理的にはベクトルでもテンソルでも構わない．材料研究では，スペクトルや物性の時空間イメージングなど，関数型やテンソル型の出力を取り扱う問題設定がある．このような形式の材料データの研究は現時点ではあまり進んでいないが，cGAN やその関連手法が有望なアプローチになりうる．

4.3 有機化合物の合成経路探索

化学構造の設計の次に解くべきタスクは，合成経路の設計である．ここでは例として，以下の 2 ステップの合成反応を考える．

$$S_1 + S_2 \to X + S_3 \to Y \tag{4.4}$$

第 1 ステップでは，二つの反応物 S_1 と S_2 が中間生成物 X を合成する．これに反応物 S_3 をあたえ，最終生成物 Y を合成する．合成経路設計の目的は，標的分子 Y に到達可能な反応物 $S = (S_1, S_2, S_3)$ の組を同定することである．反応物は商用化合物のリストから選択される．通常，$O(10^6)$ 個ほどの商用化合物を取り扱う．したがって，問題は $O(10^{6\times 3})$ の候補経路から構成される探索空間 $\mathcal{T}$ 上の組み合わせ最適化に帰着する．

ここで，合成経路の設計における順問題と逆問題を定式化する．順問題の目的は，反応物の組 S から生成物 Y の予測モデル $Y = f(S)$ を導くことである．一方，逆問題の目的は，生成物のターゲット $Y = Y^*$ が与えられたもとで，その逆写像 $S = f^{-1}(Y^*)$ を求めることである．

深層学習の技術的進歩は，合成反応の順方向の予測精度の向上に大きく貢献した．ここで，

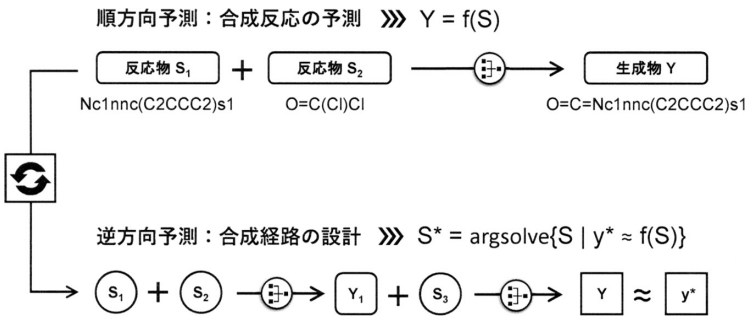


図 11．SMILES 表現に基づく合成反応予測と逆方向予測による合成経路の設計．

表 2．様々なモデルの合成反応予測の性能(top-1, top-3, top-5, top-10 accuracies [%])．

Model	Top-1	Top-3	Top-5	Top-10
Template-based (Coley et al., 2017)	71.8	86.7	90.8	94.6
WLDN (Jin et al., 2017)	79.6	87.7	89.2	-
Modified WLDN (Coley et al., 2019)	85.6	90.5	92.8	93.4
Molecular Transformer (Schwaller et al., 2019b)	90.4	94.6	95.3	-

反応物と生成物の SMILES 表現に基づくアプローチを取り上げる．図 11 に示すように，1 ステップの合成反応 $S_1 + S_2 \to Y$ において，反応物の組 $S = (S_1, S_2)$ を SMILES 文字列に変換し，両者をピリオドで連結する．また，生成物の化学構造 Y も SMILES 文字列に変換する．ここで，1 ステップの合成反応の予測は文字列から文字列への写像を求める問題に帰着する．入出力変数をこのように定義することで，機械翻訳のニューラルネットワークを用いて予測モデルを構築できる．USPTO という米国特許化合物(1978-)の合成反応データベース(Lowe, 2012)に約 100 万件のデータが収録されている．このデータを用いて Transformer (Vaswani et al., 2017)という機械翻訳のモデルを訓練する．表 2 に示すように，あるベンチマークセットにおいて，その予測精度は 90% 以上に達することが分かっている(Schwaller et al., 2019a)．その他にも，化学反応のルールを陽に取り込んだテンプレートベースと呼ばれる手法やグラフ変換の深層学習のアイデアが合成反応予測の研究に導入され，予測性能の改善が図られている(表 2)．

Guo et al. (2020)で，合成反応の順方向モデル $Y = f(S)$ の逆写像を求め，任意の生成物 $Y = Y^*$ を合成する反応物の組を探索するアルゴリズムを提案している．ここで，事後分布 $p(S|Y = Y^*)$ を以下のようにモデリングする．

$$p(S|Y = Y^*) \propto p(S, Y = Y^*) = \frac{1}{Z} \exp\left(- \frac{E(Y^*, f(S))}{T} \right). \tag{4.5}$$

ギブズ分布のエネルギー E は，標的生成物 Y^* のフィンガープリント記述子と順方向モデルの予測生成物の間の非類似度(ユークリッド距離など)を表す．温度パラメータ T は，候補反応物の多様性を制御するハイパーパラメータである．事後分布は，商用化合物の組み合わせの上に定義される．例えば，式(4.4)の 2 ステップの反応経路の設計では，定義域 $\mathcal{T}$ は $O(10^{6\times3})$ 個の離散点から構成される．したがって，事後分布は次のような形で表される．

$$p(S|Y=Y^*) \propto \sum_{s_i \in \mathcal{T}} p(S=s_i, Y=Y^*) I(S=s_i). \tag{4.6}$$

指示関数 $I(\cdot)$ は，引数が真であれば 1，そうでなければ 0 をとる．つまり，事後分布は，膨大な数の候補点 $s_i \in \mathcal{T}$ の上に確率 $p(S=s_i|Y=Y^*) \propto p(S=s_i, Y=Y^*)$ を持つ離散分布となる．この確率分布は厳密に計算できないので，n 個の代表的な候補点 $\hat{\mathcal{S}} = \{\hat{s}_i | i = 1, \ldots, n\}$ を選び，以下のように近似する．

$$\hat{p}(S|Y=Y^*) \propto \sum_{i=1}^{n} p(S=\hat{s}_i, Y=Y^*) I(S=s_i). \tag{4.7}$$

近似に用いられる候補点の集合 $\hat{\mathcal{S}}$ は，事後確率 $p(S=\hat{s}_i, Y=Y^*)$ ができるだけ大きく，多様な反応経路を含むことが望ましい．Guo et al. (2020) は，この近似分布を導くために逐次型のモンテカルロ計算アルゴリズムを開発した．

Guo et al. (2020) では，USPTO のデータを用いて包括的な数値実験を実施し，既知の合成経路に対する予測性能や提案された経路の合成可能性を検証している．図 12 は，一つの標的分子に対する 2 ステップの反応経路の解析例を抜粋したものである．この例では，6,000 以上の反応経路が予測された．また，反応経路のパターンを分類してグループを代表する合成経路を選定し，有機合成の知見に基づき候補経路の合成可能性の評価を実施した (図 12)．複数の標的分子に対してこのような評価を行い，約 35% の候補経路が化学的に妥当と結論付けた．デー

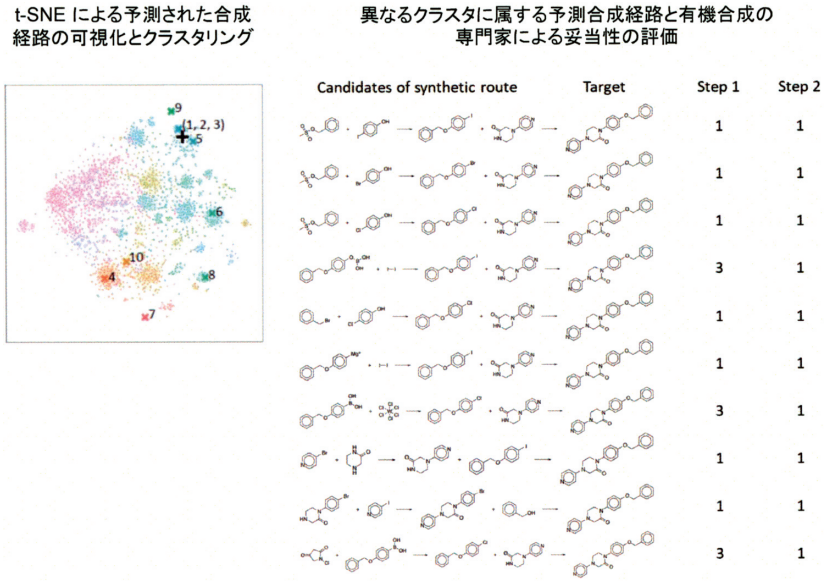


図 12. 機械学習による合成経路の設計．左：標的生成物に対する 6,613 個の 2 段階合成経路を予測し，t-SNE を用いて二次元空間にマッピングした結果．+ はデータベースにある既知の反応経路を示す．X-means クラスタリング (Dau Pelleg, 2000) で予測された経路を 98 個のグループに分類し，同定されたクラスタを色分けして表示している．シンボル × は、右図に示す 10 個の候補経路の位置を表す．右：標的生成物に対する 10 個の候補経路．反応の各ステップに対し，有機合成の専門家が三段階評価 (1: 可，2: 不明，3: 不可) を実施した．

タ解析の結果を仮説的知見として提示し，ユーザーである専門家の創造性を掻き立てることで意思決定を支援する．データ科学をこのような目的に活用する場合，提示される仮説やシナリオには多様性が求められる．最終的な意思決定を専門家に委ねるのであれば，仮説には多少の誤りが含まれていても構わない．仮説の精度を多少犠牲にしても，様々なシナリオを提示し，専門家の経験や知識だけでは到達できない斬新な発想を誘導すべきである．

4.4 結晶構造探索

任意の組成に対する最安定あるいは準安定の結晶構造を予測する問題を考える．結晶構造予測のソフトウェアとして広く普及している USPEX（Oganov and Glass, 2006; Oganov et al., 2011; Lyakhov et al., 2013）や CALYPSO（Wang et al., 2010; Zhang et al., 2017）は，第一原理計算と進化計算の融合アルゴリズムを実装している．複数の初期構造を用意し，第一原理計算や分子動力学計算を用いて構造最適化を行う．構造最適化では初期構造を起点にエネルギー計算を行い，エネルギーが徐々に低くなるように原子位置を変位させて局所的に安定な構造を求める．これらの候補構造のエネルギーを適合度とし，有望な候補を選抜する．さらに，選抜された候補構造に遺伝的操作（変異や交叉）を施し，次世代の候補構造を生成する．このループを繰り返しながら，エネルギーが最も低い構造を同定する．USPEX は結晶構造の遺伝的操作に独自の変異および交叉のアルゴリズムを採用している．一方，CALYPSO は粒子群最適化法（particle swarm optimization）というアルゴリズムを実装している．USPEX との違いは，swarm shift という遺伝的操作を採用している点にある．

USPEX と CALYPSO は第一原理計算によるエネルギー評価を何度も反復する必要があるため，膨大な計算時間を伴う．一方，CrySPY（Yamashita et al., 2018）は，機械学習に基づくサロゲートモデルを導入することで，探索の効率化を実現している（図 13）．USPEX や CALYPSO と違い，CrySPY は結晶構造の遺伝的操作を行わない．探索を開始する前に生成器を用いて候補構造を準備しておく．空間群と組成が与えられたもとで，可能な Wyckoff 位置の組と原子座標をランダムに生成する．このとき原子間距離に下限を設ける．探索範囲は候補構造の中に限られる．ベイズ最適化で構造とエネルギーのデータを段階的に蓄積しながら，ガウス過程回帰モデルの予測性能を徐々に改善していく．このサロゲートモデルを用いて低エネルギーに達する可能性が高い有望な候補構造を絞り込む．あるいは，期待値の低い候補を探索対象から除外する．

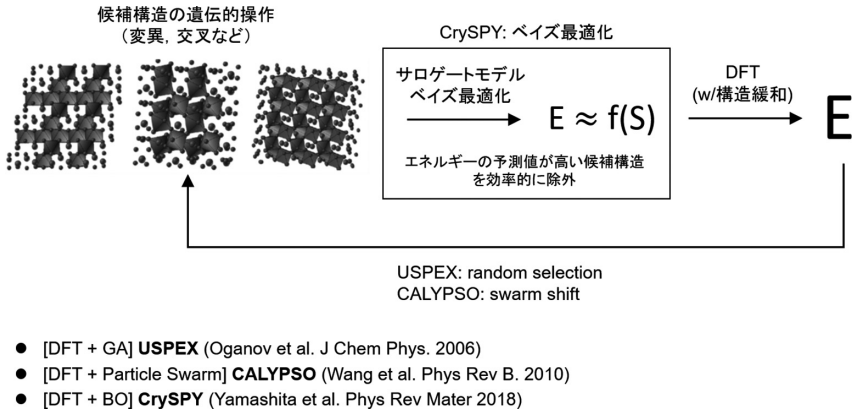


図 13．結晶構造探索アルゴリズム（USPEX, CALYPSO, CrySPY）のワークフロー．

4.5 ボルツマン生成器

熱力学的平衡状態における分子や分子集合系の存在確率は，原子間ポテンシャル $U(S)$ のボルツマン分布に従う．

$$\pi(S) = \frac{1}{D(T)} \exp\left(- \frac{U(S)}{T} \right). \tag{4.8}$$

S は系を構成する原子の座標を要素とするベクトル，T は温度を表す．正規化定数 $D(T)$ は温度の関数であり，解析的には求まらない．原子間ポテンシャルは原子間相互作用を表す．古典分子動力学法の典型的なポテンシャル関数は，原子間のファンデルワールス力を表すレナード-ジョーンズ型ポテンシャル(Lennard-Jones potential)や電荷間のクーロン力を表す静電ポテンシャルの和によって記述される．図 14 にポテンシャル関数の具体例を示す．ポテンシャル関数には複数のパラメータが含まれる．古典分子動力学法では，経験的に決められたパラメータを使用する．

分子動力学シミュレーションやモンテカルロ法を用いて，確率分布 $\pi(S)$ から原子座標をサンプリングすることで系の様々な平衡状態を推測できる．しかしながら，ある平衡状態から別の平衡状態への遷移が稀にしか起こらない複雑な系では，これらの方法では現実的な時間内に相転移を再現できない．そこで，Noé et al. (2019)は，フローモデルと呼ばれる深層生成モデルを用いて，$\pi(S)$ の近似分布 $p(S)$ を構築する手法を提案している(図 15)．この近似分布から N 個の独立なサンプル $\{S_i | i = 1, \ldots, N\}$ を生成し，重点サンプリング(importance sampling, Liu (2008))でボルツマン分布の経験分布を得る．

$$\hat{\pi}(S) = \frac{\sum_{i=1}^{N} w_i I(S = S_i)}{\sum_{i=1}^{N} w_i}, \quad w_i = \frac{\pi(S_i)}{p(S_i)}. \tag{4.9}$$

正規化された重み $w_i / \sum_i w_i$ の計算には，正規化定数 $D(T)$ は必要ないことに注意せよ．

フローモデルは確率変数の変数変換を行うニューラルネットワークである．S と同じ次元を持つ潜在変数 $Z \in \mathbb{R}^p$ は確率密度関数 $p_z(Z)$ に従うと仮定する．多くの場合，$p_z(Z)$ には多変量正規分布や一様分布などの単純な確率分布を仮定する．そして，ニューラルネットワークを用いて潜在変数に変数変換 $S = f(Z)$ を施し，S が従う複雑な確率分布 $p_s(S)$ を構築する．p 層

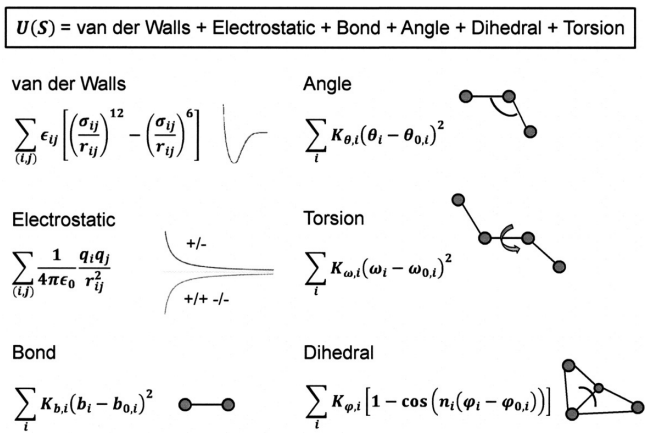


図 14. 古典分子動力学法のポテンシャル関数の例．

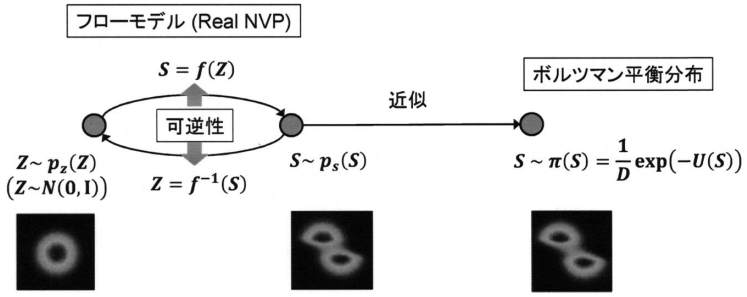


図 15. フローモデルによるボルツマン分布の近似.

の合成関数 $f = f_1 \circ f_2 \ldots \circ f_p$ は，可逆なニューラルネットワークでモデル化される．すなわち，$Z = f^{-1}(S)$，$f^{-1} = f_p^{-1} \circ f_{p-1}^{-1} \ldots \circ f_1^{-1}$ となる．このとき，ヤコビアンに基づく確率変数の変数変換で S の確率密度関数を次のように計算できる．

$$
\begin{aligned}
p_s(S) &= p_z(f^{-1}(S)) \left| \det \frac{\partial f^{-1}(S)}{\partial S} \right| \\
&= p_z(Z) \left| \det \frac{\partial f(Z)}{\partial Z} \right|^{-1} \\
&= p_z(Z) \prod_{k=1}^{p} \left| \det \frac{\partial f_k(Z_{k-1})}{\partial Z_{k-1}} \right|^{-1}
\end{aligned}
\tag{4.10}
$$

記号 det は行列式を表す．第 1 行から第 2 行の式変形では，逆関数のヤコビアンの公式を用いている．第 3 行の変形は，合成関数の微分の連鎖律より導かれる．$Z_0 = Z$, $Z_k = f_k^{-1} \circ f_{k-1}^{-1} \ldots \circ f_1^{-1}(Z)$ である．ヤコビアン $\frac{\partial f(Z)}{\partial Z}$ は，$p \times p$ の行列である．p は原子数 $\times$ 3 となり，通常の系では数千から数万オーダーとなる．そこで，既存のフローモデルでは，ヤコビアンの行列式が効率的に計算できるようなモデリングが行われている．例えば，Noé et al. (2019)で使用している Real NVP (Dinh et al., 2016)では，カップリング・レイヤーというモデルが各層に設定される．これにより，ヤコビアンは上三角行列となり，対角要素の積で行列式を計算できる．重点サンプリングにおいて，$p_s(S)$ は重みの計算で必要になることに注意せよ．これを簡単に計算できることが重点サンプリングにおいて必須事項となる．また，分子動力学シミュレーションや実験により S の観測データが与えれている場合，$p_s(S)$ を簡単に計算できれば，最尤推定を容易に実行できる．Noé et al. (2019)では，これを training by sample と呼んでいる．また，目標分布であるボルツマン分布の変数変換は，以下のようになる．

$$
\begin{aligned}
\pi_z(Z) &= \pi_s(f(Z)) \left| \det \frac{\partial f(Z)}{\partial Z} \right| \\
&= \frac{1}{D(T)} \exp\left(-\frac{U(f(Z))}{T} \right) \prod_{k=1}^{p} \left| \det \frac{\partial f_k(Z_{k-1})}{\partial Z_{k-1}} \right|
\end{aligned}
\tag{4.11}
$$

この確率密度関数を用いて，$\pi_z(Z)$ と $p_z(Z)$ のカルバック・ライブラー情報量が最小になるようにパラメータを推定する．具体的には，潜在変数の N 個のサンプル $\{Z_i | i = 1, \ldots, N\}$ を生成し，尤度 $\sum_i \log \pi_z(Z_i)$ を最大にするようにパラメータを推定する．Noé et al. (2019)では，これを training by energy と呼んでいる．

5. まとめ

本稿では，物質・材料の表現・学習・生成という観点から MI の概説を試みた．材料研究のデータ解析における入出力の変数は多様な形式をとる．多様であるがゆえ，問題毎に方法論とツールを開発していくことが求められる．一方で，記述子と生成モデルが整備されれば，順問題と逆問題の計算を実行するだけである．あとは，様々な問題に対して道具を用意し，実践を展開していけばよい．さらに，このありきたりなワークフローにデータ科学，計算科学，実験科学の学術的進歩が合流することで，新しい科学的手法が生み出され，科学的発見をもたらす．特に近年，データ科学の最先端の研究が応用分野に合流する時間差が急速に短くなってきている．本稿で取り上げた話題においても，その一端を垣間見ることができるだろう．

現在の MI は依然として黎明期にある．材料研究には，データ科学が革新的な発見を実現できる多くの課題が未発見のまま残されているに違いない．材料組織の予測と制御，触媒反応，合成実験のプロセス制御など，本稿でもそのごく一部を取り上げたが，実質的にこれらの研究はまだ始まっていないに等しい．このような未踏領域に足を踏み入れ，データ科学のユニークな視点から問題を発掘し，新しい科学的手法を創出する．すなわち，材料研究の諸問題をデータ科学の順問題・逆問題の形式に定式化する．これこそが MI に求められる最も重要なミッションである．

言うまでもなく，データ駆動型研究において最も重要なものはデータである．データ科学の他の応用分野に比べると材料研究のデータ量は圧倒的に少ない．データ科学が本格的に導入して間もないこともあり，データベースの整備も発展途上の段階にある．また，科学的成果と産業応用が密接に結びついているため，研究者は情報秘匿の意識が高く，一部の領域では，今後もデータの共有化が進まない可能性がある．そのような領域では，スモールデータの壁をいかに乗り越えていくかという課題が永遠に残される．一方で，データは無限に湧き出る石油でもある．データの量と多様性は決して減少することなく単調に増加し続ける．それと同時に，データを持つものと持たないものの間に格差が生じることになる．データ駆動型研究の本質はパワーゲームである．ビッグデータとスモールデータが混在する領域で，MI の在り方を俯瞰することも重要かもしれない．

謝　辞

本研究は国立研究開発法人・科学技術振興機構戦略的創造研究推進事業（CREST）JPMJCR19I3，科研費 19H01132，19H05820，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）JPNP16010 の助成を受けた．本論文をまとめるにあたり，統計数理研究所ものづくりデータ科学研究センターの皆様には，多くの議論にお付き合いいただきました．心よりお礼申し上げます．特に，総合研究大学院大学複合科学研究科 統計科学専攻岩山めぐみ氏と統計数理研究所ものづくりデータ科学研究センター Guo Zhongliang 氏には，本論文で示した図表の一部を提供していただいた．心よりお礼申し上げます．

参 考 文 献

Banko, L., Lysogorskiy, Y., Grochla, D., Naujoks, D., Drautz, R. and Ludwig, A. (2020). Predicting structure zone diagrams for thin film synthesis by generative machine learning, *Communications Materials*, **1**(1), 1–10.

Bartók, A. P., Kondor, R. and Csányi, G. (2013). On representing chemical environments, *Physical Review B*, **87**(18), p.184115.

Behler, J. and Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces, *Physical Review Letters*, **98**(14), p.146401.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. and Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs, *Nature Chemistry*, **4**(2), 90–98.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer-Verlag, New York.

Bolton, E. E., Wang, Y., Thiessen, P. A. and Bryant, S. H. (2008). PubChem: Integrated platform of small molecules and biological activities, *Annual Reports in Computational Chemistry*, **4**, 217–241.

Brown, N., Fiscato, M., Segler, M. H. and Vaucher, A. C. (2019). GuacaMol: Benchmarking models for de novo molecular design, *Journal of Chemical Information and Modeling*, **59**(3), 1096–1108.

Burden, F. R. (1989). Molecular identification number for substructure searches, *Journal of Chemical Information and Computer Sciences*, **29**(3), 225–227.

Burden, F. R. (1997). A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix, *Quantitative Structure-Activity Relationships*, **16**(4), 309–314.

Cang, R., Xu, Y., Chen, S., Liu, Y., Jiao, Y. and Yi Ren, M. (2017). Microstructure representation and reconstruction of heterogeneous materials via deep belief network for computational material design, *Journal of Mechanical Design*, **139**(7), p.071404.

Carhart, R. E., Smith, D. H. and Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: Definition and applications, *Journal of Chemical Information and Computer Sciences*, **25**(2), 64–73.

Carleo, G. and Troyer, M. (2017). Solving the quantum many-body problem with artificial neural networks, *Science*, **355**(6325), 602–606.

Carrete, J., Li, W., Mingo, N., Wang, S. and Curtarolo, S. (2014). Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling, *Physical Review X*, **4**(1), p.011019.

Chandrasekaran, A., Kim, C. and Ramprasad, R. (2020). Polymer genome: A polymer informatics platform to accelerate polymer discovery, *Machine Learning Meets Quantum Physics*, 397–412, Springer, Cham.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S. and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8789–8797.

Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. and Jensen, K. F. (2017). Prediction of organic reaction outcomes using machine learning, *ACS Central Science*, **3**(5), 434–443, DOI: http://dx.doi.org/10.1021/acscentsci.7b00064.

Coley, C., Jin, W., Rogers, L., Jamison, T. F., Jaakkola, T. S., Green, W. H., Barzilay, R. and Jensen, K. F. (2019). A graph-convolutional neural network model for the prediction of chemical reactivity, *Chemical Science*, **10**, 370–377, DOI: http://dx.doi.org/10.1039/C8SC04228D.

Curtarolo, S., Setyawan, W., Hart, G. L., Jahnatek, M., Chepulskii, R. V., Taylor, R. H., Wang, S., Xue, J., Yang, K., Levy, O. et al. (2012). AFLOW: An automatic framework for high-throughput materials discovery, *Computational Materials Science*, **58**, 218–226.

Dau Pelleg, A. M. (2000). X-means: Extending k-means with efficient estimation of the number of clusters, *Proceedings of the 17th International Conference on Machine Learning*, 727–734, https://www.cs.cmu.edu/ dpelleg/download/xmeans.pdf.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database, *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 248–255.

Dinh, L., Sohl-Dickstein, J. and Bengio, S. (2016). Density estimation using real nvp, arXiv preprint arXiv:1605.08803.

Durant, J. L., Leland, B. A., Henry, D. R. and Nourse, J. G. (2002). Reoptimization of MDL keys for use

in drug discovery, *Journal of Chemical Information and Computer Sciences*, **42**(6), 1273–1280.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A. and Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints, *Advances in Neural Information Processing Systems*, **28**, 2224–2232.

Elgammal, A., Liu, B., Elhoseiny, M. and Mazzone, M. (2017). CAN: Creative adversarial networks, generating "ar" by learning about styles and deviating from style norms, arXiv preprint arXiv:1706.07068.

Elton, D. C., Boukouvalas, Z., Fuge, M. D. and Chung, P. W. (2019). Deep learning for molecular design — A review of the state of the art, *Molecular Systems Design & Engineering*, **4**(4), 828–849.

Ertl, P. and Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, *Journal of Cheminformatics*, **1**(1), p.8.

Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The Elements of Statistical Learning*, 1, Springer Series in Statistics, Springer, New York.

Gärtner, T., Flach, P. and Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives, *Learning Theory and Kernel Machines*, 129–143, Springer, Berlin, Heidelberg.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. et al. (2012). ChEMBL: A large-scale bioactivity database for drug discovery, *Nucleic Acids Research*, **40**(D1), D1100–D1107.

Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T. D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M. A., Chae, H. S., Einzinger, M., Ha, D.-G., Wu, T. et al. (2016). Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach, *Nature Materials*, **15**(10), 1120–1127.

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. (2018). Automatic chemical design using a data-driven continuous representation of molecules, *ACS Central Science*, **4**(2), 268–276.

Guha, R. et al. (2007). Chemical informatics functionality in R, *Journal Statistical Software*, **18**(5), 1–16.

Guo, Z., Wu, S., Ohno, M. and Yoshida, R. (2020). Bayesian algorithm for retrosynthesis, *Journal of Chemical Information and Modeling*, **60**(10), 4474–4486.

Hall, L. H. and Kier, L. B. (1995). Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information, *Journal of Chemical Information and Computer Sciences*, **35**(6), 1039–1045.

平岡裕章 (2015). データに潜む幾何構造：パーシステントホモロジー(特集 自然の中の幾何構造), 数理科学, **53**(6), 48–53.

Hirn, M., Mallat, S. and Poilvert, N. (2017). Wavelet scattering regression of quantum chemical energies, *Multiscale Modeling & Simulation*, **15**(2), 827–863.

Huan, T. D., Mannodi-Kanakkithodi, A., Kim, C., Sharma, V., Pilania, G. and Ramprasad, R. (2016). A polymer dataset for accelerated property prediction and design, *Scientific Data*, **3**(1), 1–10.

Ikebata, H., Hongo, K., Isomura, T., Maezono, R. and Yoshida, R. (2017). Bayesian molecular design with a chemical language model, *Journal of Computer-aided Molecular Design*, **31**(4), 379–391.

Irwin, J. J. and Shoichet, B. K. (2005). ZINC — A free database of commercially available compounds for virtual screening, *Journal of Chemical Information and Modeling*, **45**(1), 177–182.

Isayev, O., Oses, C., Toher, C., Gossett, E., Curtarolo, S. and Tropsha, A. (2017). Universal fragment descriptors for predicting properties of inorganic crystals, *Nature Communications*, **8**(1), 1–12.

Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134.

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. (2013). The Materials Project: A materials genome approach to accelerating materials innovation, *APL Materials*, **1**(1), p.011002, http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi, DOI: http://dx.doi.org/10.1063/1.4812323.

Jaques, N., Gu, S., Bahdanau, D., Hernández-Lobato, J. M., Turner, R. E. and Eck, D. (2017). Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control, *International Conference on Machine Learning*, 1645–1654.

Jin, W., Coley, C. W., Barzilay, R. and Jaakkola, T. (2017). Predicting organic reaction outcomes with Weisfeiler-Lehman network, *Advances in Neural Information Processing Systems*, **30**, 2608–2617, https://papers.nips.cc/paper/6854-predicting-organic-reaction-outcomes-with-weisfeiler-lehman-network.

Jin, W., Barzilay, R. and Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation, arXiv preprint arXiv:1802.04364.

Ju, S., Yoshida, R., Liu, C., Hongo, K., Tadano, T. and Shiomi, J. (2019). Exploring diamond-like lattice thermal conductivity crystals via feature-based transfer learning, arXiv preprint arXiv:1909.11234.

Kashima, H., Tsuda, K. and Inokuchi, A. (2003). Marginalized kernels between labeled graphs, *Proceedings of the 20th International Conference on Machine Learning*, 321–328.

Kipf, T. N. and Welling, M. (2016). Variational graph auto-encoders, arXiv preprint arXiv:1611.07308.

Kirkpatrick, P. and Ellis, C. (2004). Chemical space, *Nature*, **432**(823).

Klekota, J. and Roth, F. P. (2008). Chemical substructures that enrich for biological activity, *Bioinformatics*, **24**(21), 2518–2525.

Kusano, G., Hiraoka, Y. and Fukumizu, K. (2016). Persistence weighted Gaussian kernel for topological data analysis, *International Conference on Machine Learning*, 2004–2013.

Landrum, G. (2016). RDKit: Open-source cheminformatics software, https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4.

Li, X., Yang, Z., Brinson, L. C., Choudhary, A., Agrawal, A. and Chen, W. (2018a). A deep adversarial learning methodology for designing microstructural material systems, *Proceedings of ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, p.V02BT03A008.

Li, X., Zhang, Y., Zhao, H., Burkhart, C., Brinson, L. C. and Chen, W. (2018b). A transfer learning approach for microstructure reconstruction and structure-property predictions, *Scientific Reports*, **8**(1), 1–13.

Liu, C., Fujita, E., Katsura, Y., Inada, Y., Ishikawa, A., Tamura, R., Kimura, K. and Yoshida, R. (2021). Machine learning to predict quasicrystals from chemical compositions, *Advanced Materials* (in press), https://doi.org/10.1002/adma.202102507.

Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, New York.

Lowe, D. M. (2012). Extraction of chemical structures and reactions from the literature, Ph.D. Thesis, Department of Chemistry, University of Cambridge. DOI: http://dx.doi.org/10.17863/CAM.16293.

Lyakhov, A. O., Oganov, A. R., Stokes, H. T. and Zhu, Q. (2013). New developments in evolutionary structure prediction algorithm USPEX, *Computer Physics Communications*, **184**(4), 1172–1182.

Mahé, P. and Vert, J.-P. (2009). Graph kernels based on tree patterns for molecules, *Machine Learning*, **75**(1), 3–35.

Mahé, P., Ueda, N., Akutsu, T., Perret, J.-L. and Vert, J.-P. (2004). Extensions of marginalized graph kernels, *Proceedings of the Twenty-first International Conference on Machine Learning*, p.70.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784.

Moreau, G. and Broto, P. (1980). The autocorrelation of a topological structure: A new molecular

descriptor, *New Journal of Chemistry*, **4**(6), 359–360.

Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service, *Journal of Chemical Documentation*, **5**(2), 107–113.

Moriwaki, H., Tian, Y.-S., Kawashita, N. and Takagi, T. (2018). Mordred: A molecular descriptor calculator, *Journal of Cheminformatics*, **10**(1), 1–14.

Nilakantan, R., Bauman, N., Dixon, J. S. and Venkataraghavan, R. (1987). Topological torsion: A new molecular descriptor for SAR applications. Comparison with other descriptors, *Journal of Chemical Information and Computer Sciences*, **27**(2), 82–85.

Noé, F., Olsson, S., Köhler, J. and Wu, H. (2019). Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning, *Science*, **365**(6457), p.eaaw1147.

Oganov, A. R. and Glass, C. W. (2006). Crystal structure prediction using ab initio evolutionary techniques: Principles and applications, *The Journal of Chemical Physics*, **124**(24), p.244704.

Oganov, A. R. and Valle, M. (2009). How to quantify energy landscapes of solids, *The Journal of Chemical Physics*, **130**(10), p.104504.

Oganov, A. R., Lyakhov, A. O. and Valle, M. (2011). How evolutionary crystal structure prediction works and why, *Accounts of Chemical Research*, **44**(3), 227–237.

Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. and Ramprasad, R. (2013). Accelerating materials property predictions using machine learning, *Scientific Reports*, **3**(1), 1–6.

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints, *Journal of Chemical Information and Modeling*, **50**(5), 742–754.

Sanchez-Lengeling, B. and Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering, *Science*, **361**(6400), 360–365.

Schütt, K., Kindermans, P.-J., Felix, H. E. S., Chmiela, S., Tkatchenko, A. and Müller, K.-R. (2017). Schnet: A continuous-filter convolutional neural network for modeling quantum interactions, *Advances in Neural Information Processing Systems*, **30**, 992–1002.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C. and Lee, A. A. (2019a). Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction, *ACS Central Science*, **5**(9), 1572–1583.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C. and Lee, A. A. (2019b). Molecular Transformer: A model for uncertainty-calibrated chemical reaction prediction, *ACS Central Science*, **5**(9), 1572–1583, DOI: http://dx.doi.org/10.1021/acscentsci.9b00576.

Segler, M. H., Kogej, T., Tyrchan, C. and Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks, *ACS Central Science*, **4**(1), 120–131.

Seko, A., Togo, A., Hayashi, H., Tsuda, K., Chaput, L. and Tanaka, I. (2015). Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization, *Physical Review Letters*, **115**(20), p.205901.

Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. and Tanaka, I. (2017). Representation of compounds for machine-learning prediction of physical properties, *Physical Review B*, **95**(14), p.144110.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

Tukey, J. W. (1962). The future of data analysis, *The Annals of Mathematical Statistics*, **33**(1), 1–67.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need, *Advances in Neural Information Processing Systems*, **30**, 5998–6008.

Venkatasubramanian, V., Chan, K. and Caruthers, J. M. (1994). Computer-aided molecular design using genetic algorithms, *Computers & Chemical Engineering*, **18**(9), 833–844.

Venkatasubramanian, V., Chan, K. and Caruthers, J. M. (1995). Evolutionary design of molecules with desired properties using the genetic algorithm, *Journal of Chemical Information and Computer*

Sciences, **35**(2), 188–195.

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R. and Borgwardt, K. M. (2010). Graph kernels, *The Journal of Machine Learning Research*, **11**, 1201–1242.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J. and Bryant, S. H. (2009). PubChem: A public information system for analyzing bioactivities of small molecules, *Nucleic Acids Research*, **37**(suppl_2), W623–W633.

Wang, Y., Lv, J., Zhu, L. and Ma, Y. (2010). Crystal structure prediction via particle-swarm optimization, *Physical Review B*, **82**(9), p.094116.

Ward, L., Agrawal, A., Choudhary, A. and Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials, *npj Computational Materials*, **2**(1), 1–7.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *Journal of Chemical Information and Computer Sciences*, **28**(1), 31–36.

Wildman, S. A. and Crippen, G. M. (1999). Prediction of physicochemical parameters by atomic contributions, *Journal of Chemical Information and Computer Sciences*, **39**(5), 868–873.

Wu, S., Kondo, Y., Kakimoto, M.-A., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., Morikawa, J. and Yoshida, R. (2019). Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm, *npj Computational Materials*, **5**(1), 1–11.

Wu, S., Lambard, G., Liu, C., Yamada, H. and Yoshida, R. (2020a). iQSPR in XenonPy: A Bayesian molecular design algorithm, *Molecular Informatics*, **39**(1-2), p.1900107.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Philip, S. Y. (2020b). A comprehensive survey on graph neural networks, *IEEE Transactions on Neural Networks and Learning Systems*, **32**(1), 4–24.

Xie, T. and Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, *Physical Review Letters*, **120**(14), p.145301.

Yamada, H., Liu, C., Wu, S., Koyama, Y., Ju, S., Shiomi, J., Morikawa, J. and Yoshida, R. (2019). Predicting materials properties with little data using shotgun transfer learning, *ACS Central Science*, **5**(10), 1717–1730.

Yamashita, H., Higuchi, T. and Yoshida, R. (2014). Atom environment kernels on molecules, *Journal of Chemical Information and Modeling*, **54**(5), 1289–1300.

Yamashita, T., Sato, N., Kino, H., Miyake, T., Tsuda, K. and Oguchi, T. (2018). Crystal structure prediction accelerated by Bayesian optimization, *Physical Review Materials*, **2**(1), p.013803.

Yang, X., Zhang, J., Yoshizoe, K., Terayama, K. and Tsuda, K. (2017). ChemTS: An efficient python library for de novo molecular generation, *Science and Technology of Advanced Materials*, **18**(1), 972–976.

Yi, Z., Zhang, H., Tan, P. and Gong, M. (2017). DualGAN: Unsupervised dual learning for image-to-image translation, *Proceedings of the IEEE International Conference on Computer Vision*, 2849–2857.

Zhang, Y., Wang, H., Wang, Y., Zhang, L. and Ma, Y. (2017). Computer-assisted inverse design of inorganic electrides, *Physical Review X*, **7**(1), p.011017.

Zhu, J.-Y., Park, T., Isola, P. and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks, *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232.

Materials Informatics: A Review and Perspectives

Ryo Yoshida

The Institute of Statistical Mathematics

In this paper, we present an overview of materials informatics, focusing on machine learning technologies to several inverse problems in materials research. The objective of the forward problem is to predict the output of a system with respect to its input. For example, the input variable corresponds to the structure of a given material and the output variable corresponds to its properties. In the inverse problem, we identify promising candidate materials that exhibit any given desired properties by solving the inverse mapping of the forward model. This is a conventional workflow of data science, but one distinct feature of data analysis in materials research lies in the high dimensionality and specificity of the variables. In general, the search space for candidate materials is extremely vast. In addition, in many cases, we deal with variables that are non-trivial to be represented into fixed-length vectors, such as composition, molecules, and crystal structures. In this paper, we describe the essence of machine learning for solving inverse problems by introducing various examples.

Key words: Physical property, materials design, synthesis, inverse problem, descriptor, generative models.

統計数理(2021)
第 69 巻 第 1 号 35–47

特集「マテリアルズインフォマティクスの最前線」
[総合報告]

機械学習による機能性分子の自動設計：高熱伝導高分子材料の探索

吉田 亮[1]・ウ ステファン[1]・森川 淳子[2]

(受付 2021 年 2 月 2 日；改訂 4 月 15 日；採択 4 月 15 日)

要　　旨

ベイズ推論や機械学習の解析技術を適用し，所望の特性を持つ化学構造を設計する．実験やシミュレーションから得られたデータを用いて，構造から特性の順方向の予測モデルを構築する．これに条件付き確率のベイズ則を適用し，特性から構造の逆方向の予測モデルを導く．さらに，このモデルから仮説分子を発生させることで，所望の特性を有する有望な候補を同定する．我々は，このようなアプローチを実践し，熱伝導率が 0.41 W/mK に達する可塑性プラスチックポリマーを発見した．これは典型的な無配向のポリアミド系高分子と比較して約 80% の性能向上に相当する．本稿では，データ科学の非専門家を対象にベイズ推論に基づく分子設計アルゴリズムの技術解説を行い，高分子熱物性の研究への適用例を解説する．

キーワード：分子設計，ベイズ推論，転移学習，ポリマー，熱伝導率．

1. はじめに

有機低分子のケミカルスペースには，およそ 10^{60} 個の候補分子が存在すると言われている(Kirkpatrick and Ellis, 2004)．分子設計の目的は，この広大な探索空間から目標の特性を示す有望な仮説分子を同定することである．実験やシミュレーションから得られたデータを用いて，分子の化学構造 S から特性 Y を予測する統計モデル $Y = f(S)$ を構築する．これを構造物性相関解析という．さらに，このモデルの逆写像 $S = f^{-1}(Y)$ を求めて，所望の特性 $Y \in U$ を満たす構造 S を求める．これを逆構造相関物性解析という．この逆問題を解く最もシンプルなアプローチは，仮想スクリーニングである．膨大な数の候補分子(仮想ライブラリ)を作製し，統計モデルを用いて大規模なスクリーニングを実施する．一般に，物理実験や物理モデルに基づく計算機実験に比べると，統計モデルは計算速度が圧倒的に速く，膨大な数の候補分子を対象にスクリーニングを実施できる．しかしながら，探索空間が極めて大きい場合，大量のライブラリを用いたとしても，目標値に中々近づけないことが多い．そこで逆向きの計算が必要になる．逆向きの計算は，目標値周辺に分布する構造を重点的に探索することを目的とする．

逆写像 $S = f^{-1}(Y)$ を求める方法の一つは，分子フラグメントの集合と遺伝的アルゴリズムを組み合わせた手法である(Venkatasubramanian et al., 1994, 1995; Douguet et al., 2000; Lameijer et al., 2006)．複数の初期分子を用意し，分子の部分構造をランダムに改変し(変異)，分子間で部分構造の組み換え(交叉)，予測された特性が目標値に近い分子を優先的に複製(選択)しなが

[1] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3
[2] 東京工業大学 物質理工学院：〒 152–8550 東京都目黒区大岡山 2–12–1

ら世代交代を進めていき，徐々に目標値に近づけていく．例えば，Mannodi-Kanakkithodi et al. (2016)は，7 種類の要素(CH2, NH, CO, C6H4, C4H2, CS, O)から構成される n ブロックのポリマーユニットを対象に，所望の誘電特性を有するポリマーを設計している．Venkatraman and Alsberg (2018)は，分子フラグメントと遺伝的アルゴリズムを用いて，屈折率を対象としたポリマー設計の事例を報告している．また，グラフの数え上げアルゴリズムに基づく探索手法も研究されている（Miyao et al., 2016）．

上述のアプローチでは，構造改変の計算に既存分子のフラグメントを用いることで，分子構造の自由度に制限を加え，探索空間を絞り込み，設計された分子の合成可能性の向上を図っている．しかしながら，探索空間の過度な絞り込みは，構造の新規性の低下を引き起こす．そこで近年，特に 2018 年頃を境に機械学習の分野から端を発する形で従来とは全く異なる手法が出現した．SMILES という形式で分子の化学構造を文字列で表し，確率的言語モデルを用いて文字列のパターンを学習する．こうすることで，既存分子に現れるパターン(頻出フラグメントや不適切な化学結合のルールなど)を模倣した構造生成器を構築できることが分かってきた．我々は 2017 年に，自然言語処理の n-gram という言語モデルを SMILES 用に拡張し，訓練された SMILES 生成モデルを用いた分子設計のアルゴリズムとソフトウェアを発表した（Ikebata et al., 2017）．さらにほぼ同時期に，ディープラーニングに基づく言語生成モデルを活用した分子設計手法が相次いで発表されることとなった(変分自己符号器（Gómez-Bombarelli et al., 2018），再帰的ニューラルネットワーク（Segler et al., 2018))．

本稿は，著者らが開発した確率的言語モデルとベイズ推論に基づく分子生成の研究を解説する（Ikebata et al., 2017）．このアルゴリズムは XenonPy という Python ライブラリの iQSPR-X というモジュールに実装されている（Wu et al., 2020）．構造物性相関を捉えた順方向のモデルに条件付き確率のベイズ則を適用し，逆方向のモデルを導く．次に，既存化合物の化学構造を SMILES 形式で文字列で表し．この文字列集合を用いて言語モデルを訓練し，既存分子に内在する構造パターンを模倣した生成モデルを構築する．最後に，言語モデルを用いて逐次モンテカルロ法（Liu, 2008)の提案分布を設計し，逆方向のモデルから所望の特性を有する仮説分子を発生する．

実問題への適用例として，高い熱伝導を有する高分子材料を発見した研究を紹介する（Wu et al., 2019）．我々はベイズ推論に基づく分子設計アルゴリズムを用いて，高熱伝導率をターゲットに仮想ライブラリを作製した．その中から 3 種類の芳香族ポリアミドを合成し，最大で熱伝導率 0.41 W/mK に達する新規ポリマーを発見した．観測された熱伝導率は，典型的な無配向状態のポリアミド系高分子と比較して約 80% の性能向上に相当する．さらに，高耐熱性や有機溶媒への溶解性，フィルム加工の容易性など，実用化に求められる諸特性を併せ持つことが実験的に確認された．

2. ベイズ推論に基づく分子設計

ベイズ推論に基づく分子設計は，以下に示す条件付き確率のベイズ則に基づいて順方向と逆方向の予測を行う．

$$p(S|Y \in U) \propto p(Y \in U|S)p(S). \tag{2.1}$$

訓練データ集合 $\mathcal{D} = \{(Y_i, S_i)|i = 1, \ldots, n\}$ を用いて S から Y の順方向の予測モデルを構築する．このモデルを用いて条件付き確率分布 $p(Y|S)$ を定める．このモデルから任意の S が所望の特性の範囲 U に入る確率を計算したものが右辺の $p(Y \in U|S) = \int_U p(y|S)\mathrm{d}y$ に相当する．ベイズ推論の文脈では，$p(Y \in U|S)$ はパラメータ S の尤度関数と呼ばれる．さらに，事前確

率分布 $p(S)$ を介して有望な探索空間を絞り込む．左辺の条件付き確率分布 $p(S|Y \in U)$ は事後確率分布と呼ばれる．事後確率分布は尤度関数と事前確率分布の積に比例する．この条件付き確率分布から S をサンプリングすることで，所望の特性 $Y \in U$ を満たす新規分子を同定する．

2.1 尤度関数：順方向の予測

尤度関数の構築について，Ikebata et al. (2017) ではベイズ型の線形回帰モデルを使用している．

$$Y = \boldsymbol{\phi}(S)^{\mathsf{T}}\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(\epsilon|0,\tau), \tag{2.2}$$

$$\boldsymbol{\beta}|\tau \sim N(\boldsymbol{\beta}|\mathbf{0},\tau\boldsymbol{V}_0), \tag{2.3}$$

$$\tau \sim \mathrm{IG}(\tau|a_0,b_0). \tag{2.4}$$

式 (2.2) の $\boldsymbol{\phi}(S)$ は分子の構造的特徴を表す記述子ベクトル，$\boldsymbol{\beta}$ は回帰係数ベクトルを表す．観測ノイズ ϵ は平均 0，分散 τ の正規分布 $N(\epsilon|0,\tau)$ に従う．式 (2.2) より，S と未知パラメータ $\boldsymbol{\beta}$，τ が所与のもとで，Y の条件付き確率分布 $p(Y|S,\boldsymbol{\beta},\tau)$ は平均 $\boldsymbol{\phi}(S)^{\mathsf{T}}\boldsymbol{\beta}$，分散 τ の正規分布に従う．式 (2.3) は，回帰係数 $\boldsymbol{\beta}$ が平均ベクトル $\mathbf{0}$，共分散行列 $\tau\boldsymbol{V}_0$ の多変量正規分布に従うことを表す．ここで，$\boldsymbol{\beta}$ の事前分布の共分散行列は，もう一方の未知パラメータ τ に依存することに注意せよ．これは，この後に示す事後分布や予測分布を簡潔な形に導くための便宜的な措置である．式 (2.4) の τ の事前分布は，形状パラメータ a_0，尺度パラメータ b_0 の逆ガンマ分布 $\mathrm{IG}(\tau|a_0,b_0)$ に従うと仮定する．ここで，線形モデル Y の切片項は 0 と仮定していることに注意せよ．この仮定にデータを適合させるために，Y の観測データの平均がゼロとなるように中心化 $(Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i \to Y_i)$ を施す．

以上の仮定のもとで，τ が所与のもとでの $\boldsymbol{\beta}$ の事後分布は，以下の多変量正規分布になる．

$$p(\boldsymbol{\beta}|\tau,\mathcal{D}) \propto \prod_{i=1}^{n} p(Y_i|S_i,\boldsymbol{\beta},\tau)p(\boldsymbol{\beta}|\tau) \propto N(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta,\tau\boldsymbol{\Sigma}_\beta). \tag{2.5}$$

ここで，事後分布の平均と共分散行列は，

$$\boldsymbol{\mu}_\beta = \boldsymbol{\Sigma}_\beta\boldsymbol{\Phi}\boldsymbol{y}, \tag{2.6}$$

$$\boldsymbol{\Sigma}_\beta = (\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}} + \boldsymbol{V}_0^{-1})^{-1}. \tag{2.7}$$

$\boldsymbol{y} \in \mathbb{R}^n$ は，出力変数の n 個の観測値を要素に持つベクトル，$\boldsymbol{\Phi} = (\boldsymbol{\phi}(S_1)\cdots\boldsymbol{\phi}(S_n)) \in \mathbb{R}^{p\times n}$ は，n 個の記述子ベクトルを列ベクトルに持つ行列である．また，τ の事後分布は，以下の逆ガンマ分布となる．

$$p(\tau|\mathcal{D}) = \mathrm{IG}(\tau|a_\tau,b_\tau), \tag{2.8}$$

$$a_\tau = a_0 + \frac{n}{2}, \tag{2.9}$$

$$b_\tau = b_0 + \frac{1}{2}(\boldsymbol{y}^{\mathsf{T}}\boldsymbol{y} - \boldsymbol{\mu}_\beta^{\mathsf{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta). \tag{2.10}$$

次に Y の予測分布を示す．任意の S に対する Y の予測分布は，事後確率分布を用いて次のように計算できる．

$$\begin{aligned} p(Y|S) &= \int p(Y|S,\boldsymbol{\beta},\tau)p(\boldsymbol{\beta}|\tau,\mathcal{D})p(\tau|\mathcal{D}))\mathrm{d}\boldsymbol{\beta}\mathrm{d}\tau \\ &= \mathrm{T}(Y|\mu_Y(S),\sigma_Y(S),\nu_Y). \end{aligned} \tag{2.11}$$

ここで，$\mathrm{T}(Y|\mu_Y(S),\sigma_Y(S),\nu_Y)$ は，平均 $\mu_Y(S)$，尺度パラメータ $\tau_Y(S)$，自由度 ν_Y の t 分布

の確率密度関数を表す．

$$\mu_Y(S) = \boldsymbol{\phi}(S)^\mathsf{T}\boldsymbol{\mu}_\beta, \tag{2.12}$$

$$\sigma_Y(S) = \frac{b_\tau}{a_\tau}(\boldsymbol{I} + \boldsymbol{\phi}(S)^\mathsf{T}\boldsymbol{\Sigma}_\beta\boldsymbol{\phi}(S)), \tag{2.13}$$

$$\nu_Y = 2a_\tau. \tag{2.14}$$

ベイズ線形回帰の事後分布と予測分布の導出については，ベイズ統計学の一般的な教科書を参照せよ（例えば，Gelman et al., 2013）．

XenonPy では，ユーザーが作成した任意の尤度関数を逆解析に組み込むことができる（参考文献に示した Liu and Wu, 2021 を参照）．ベイズ線形回帰モデル以外の選択肢として，ガウス過程（Gaussian process）に基づくノンパラメトリックベイズ回帰が挙げられる（Rasmussen, 2003; 持橋・大羽, 2019）．その他のモデリングの方法としては，ニューラルネットワーク，ランダムフォレスト，エラスティックネット回帰（$\ell 1$, $\ell 2$ 正則化回帰），勾配ブースティング，ロジスティック回帰，サポートベクターマシンなども選択肢となる（Hastie et al., 2009）．しかしながら，これらの非ベイズ的なモデルは確率モデルではないため，条件付き確率分布 $p(Y|S)$ を定義できない．そこで，アドホックな解決方策として，ブートストラップ法を適用してモデルの不確かさを定量化することが考えられる．例えば，決定論的な回帰モデル $f(S)$ が与えれたもとで，同時確率分布を次のようにモデリングする．

$$p(Y, S) \propto \exp\left(-\frac{(Y - f(S))^2}{\sigma^2(S)}\right)p(S). \tag{2.15}$$

右辺の $p(S)$ 以外の項は，平均 $f(S)$，分散 $\sigma(S)^2$ の正規分布の確率密度関数に相当する．分散 $\sigma^2(S)$ を決めるために，モデルの訓練時にブートストラップ分散を計算する．具体的な手順は，以下の通りである．

(a) 訓練データ集合から m 個（$m < n$）のサンプルの復元抽出を B 回行い（ブートストラップ），サンプル集合 $\mathcal{D}_1, \ldots, \mathcal{D}_B$ を作成する．
(b) 各 $\mathcal{D}_b$ を用いてモデル $f_b(S)$ を訓練する（$b = 1, \ldots, B$）．
(c) B 個のモデルの平均と分散を式 (2.15) の $f(S)$，$\sigma^2(S)$ とする．

$$f(S) = \frac{1}{B}\sum_{b=1}^{B} f_b(S), \quad \sigma^2(S) = \frac{1}{B}\sum_{b=1}^{B}(f_b(S) - f(S))^2. \tag{2.16}$$

入力空間において訓練データが疎な領域のサンプルは，その周辺に類似サンプルがない．したがって，ブートストラップサンプリングにおいてそのサンプルが選択されなければ，代替するものがないため，近傍の S の $f_1(S), \ldots, f_B(S)$ のばらつきは大きくなる．逆に，訓練データが密な領域では，類似サンプルが多数あるため，$f_1(S), \ldots, f_B(S)$ のばらつきは小さくなる．このようなモデルの不確かさの定量の仕方は，データ科学の様々な局面で現れる（例えば，決定論的なモデルを用いたベイズ最適化（Hutter et al., 2011））．しかしながら，このような手法はあくまで経験的なアプローチであり，理論的な根拠は乏しい．

2.2 確率的言語モデルによる構造生成

Ikebata et al. (2017) で提案された確率的言語モデル（拡張 n グラム）による構造生成について述べる．訓練集合に用いる既存化合物の化学構造を SMILES 形式で記述する．この文字列集合を用いて n グラムのモデルを訓練し，既存化合物に現れるパターン（頻出フラグメントや不適

表 1. SMILES の正式なルールと iQSPR-X の修正ルールの対応表.

	正規ルール	修正方法
環構造の始点	$n \in \{1,2,\ldots\}$	&
環構造の終点	$n \in \{1,2,\ldots\}$	&_n (n は始点のインデックス)
原子 A の後の結合記号	=A (2 重結合), #A (3 重結合)	=A や #A を 1 文字とする
終了コード	N/A	$
角括弧内の文字	[abcde]	[abcde] を 1 文字とする

切な化学結合のルールなど)を模倣した構造生成モデルを構築する.

SMILES 記法に基づき，分子を長さ g の文字列 $S = s_1 s_2 \cdots s_g$ に変換する．SMILES の正規のルールに加えて，全ての文字列には終了コード '$' が付与される．終了コードを入れてモデルを訓練することで，再帰的な文字列伸長処理を自動的に終了させる．例えば，文字列 $\cdots$CCC = O の右側への伸長は化学結合のルールに抵触する．既存化合物に内在する化学のルールをモデルに学習させることで，右側伸長を実行する際に自動的に '$' を付加する．さらに，環の始点と終点を示す数字を '&' と '&_n' で表現する．改訂された表現規則を表 1 に示す.

ここで，文字列 S の事前分布 $p(S)$ を次のように条件付き確率の積で表現する.

$$p(S) = p(s_1) \prod_{i=2}^{p} p(s_i | s_{1:i-1}). \tag{2.17}$$

i 番目の文字 s_i の出現確率は，先行する $s_{1:i-1} = s_1 \cdots s_{i-1}$ に依存する．一般に，同一の化学構造に対する SMILES の表現は一意ではない．このような構造的に等価な文字列を異なる S として扱う.

言語モデルに基づく構造生成器の基本コンセプトは，以下の通りである．既知の化合物の部分文字列の頻度から条件付き確率 $p(s_i|s_{1:i-1})$ を推定し，訓練されたモデルに化学言語のコンテキストを学習させる．所与の部分構造 $s_{1:i-1}$ に対し，モデルを用いて残りの文字列を生成する．条件付き確率に従い，終了コードが出現するまで文字を一個ずつ追加していく.

言語モデルは SMILES の文法規則に抵触しない文字列を生成する必要がある．ここで，環構造と側鎖などに関する分岐表現の文法規則が，モデリングの技術的な難しさとなる．以下では，具体例に基づいて問題点を説明する.

(a) 閉じていない環と分岐のシンボルは文法エラーとなる．例えば， $s_{1:6}$ = CC(C(C を伸長する場合，右側のどこかに 2 つの閉記号 ')' を含む必要がある.

(b) SMILES 文字列の隣接文字は，化学構造上で必ずしも隣接するとは限らない．例えば，CCCCC(CCCC)C を考えてみる．括弧内の部分文字列は主鎖からの分岐を表す．主鎖を構成する 6 つの炭素は，文字列上では分岐要素の前後に分かれて配置している．この場合，最終文字 s_{12} = C の出現確率は，分岐の文字よりも主鎖の文字の方に影響を受けるべきである．つまり，s_i の条件付き確率は，文脈依存的に $s_{1:i-1}$ との関係性が決まるべきである．条件に一つ以上の環が現れる場合も同様である(例えば，c1ccc2ccccc2c1C).

これらの技術的課題を解決するために，拡張 n グラムは条件付き確率を以下のようにモデリングする.

$$p(s_i|s_{1:i-1}) = \prod_{k=1}^{20} p(s_i|\phi_{n-1}(s_{1:i-1}), \mathcal{A}_k)^{I(s_{1:i-1} \in \mathcal{A}_k)}. \tag{2.18}$$

ここで, $I(\cdot)$ は，引数が真であれば 1，そうでなければ 0 をとる指示関数を表す．モデルは，20

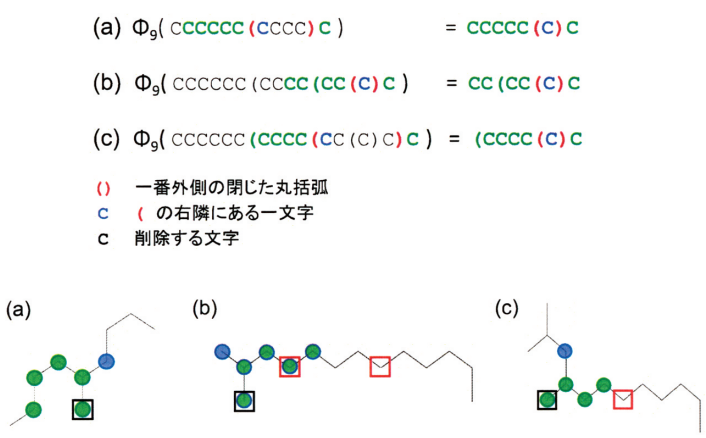


図 1. $\phi_{n-1}(\cdot)$ による部分文字列選択.

種類のサブモデル $p(s_i|\phi_{n-1}(s_{1:i-1}), \mathcal{A}_k)$ $(k = 1, \ldots, 20)$ からなる．条件の部分文字列 $s_{1:i-1}$ の状態が，相互に排他的な条件 $\mathcal{A}_k$ $(k = 1, \ldots, 20)$ のいずれに属するかを調べる．該当する一つのモデルが有効になる．20 個の条件 $(= 2 \times 10)$ は, 部分文字列 $s_{1:i-1}$ 内の閉じていない分岐の有無と閉じていない環の数 $\{0, 1, \ldots, 9\}$ の組からなる．モデルの訓練では，全訓練データを 20 パターンに分類した部分文字列を用いて，前後の文字列の頻度を計算して条件付き確率を推定する．こうすることで，例えば $s_{1:i-1} =$ CCCC(CC(の場合，訓練されたモデルの条件付き確率は，右側に二つの閉記号 ‘)’ を生成するような偏りを持つ．

$\phi_{n-1}(s_{1:i-1})$ は，$s_{1:i-1}$ から長さ $n-1$ の部分文字列を選択する演算子であり，以下の二つの操作から構成される．

(a) **縮小**：$s_{1:i-1}$ が，閉じた括弧で囲まれた部分文字列 $t = t_1 \cdots t_q$ を含んでおり，t 自体が他の閉じた括弧で囲まれることがないとする．言い換えれば，t は一番外側の閉じた括弧の内側にある部分文字列である．部分文字列 t は，その最初の文字 t_1 を除くすべての文字を削除され，$t \to t = t_1$ に縮小される．つまり，t_1 は, 一番外側の閉じた括弧の始点 ‘(’ の右隣の文字である．

(b) **抽出**：$\phi_{n-1}(s_{1:i-1})$ は $s_{1:i-1}$ の縮小文字列の最後の $n-1$ 個の文字を出力する．

図 1 に ϕ の適用例を示している．この操作により，任意の閉じた最外殻の括弧内の部分文字列は，分岐点に隣接する原子を表す 1 文字に縮小される．こうすることで，s_i の出現確率は化学構造上の隣接した $n-1$ 個の要素に依存するようになる．

2.3 所望の特性を有する化学構造の生成

訓練された拡張 n グラムと順方向のモデルを用いて逐次モンテカルロ法(SMC: sequential Monte Carlo)を実行し，事後分布から化学構造をサンプリングする．SMC の一般的な解説については，Liu (2008) などを参照せよ．Algorithm 1 に，iQSPR-X に実装されている SMC のアルゴリズムを示す．これは，Del Moral et al. (2006) の提案手法に基づいて設計されている．一般に複数の系統の化学構造が高い事後確率を有する．SMC でこのような多様な化学構造を検出するために，逆温度の非減少列 $0 \le \beta_1 \le \beta_2 \le \cdots \le \beta_{T-1} \le \beta_T = 1$ を用いて，尤度関数のアニーリング $p(Y \in U|S)^{\beta_t}$ を行う．逆温度が低下するにつれ，尤度関数は平坦になる．小さな $\beta_1 \approx 0$ から開始して，逆温度をゆっくりと 1 に近づけていき，最終的に $\beta_t = 1$ $(\forall t \ge s)$ に

到達したタイミングで事後分布にブリッジする．

提案分布 $g(s_i^*|s_i^{t-1})$ を用いて，ステップ $t-1$ の粒子 s_i^{t-1} を新しい s_i^* に置き換える．各粒子の目標特性への適合度 w_i は，順方向のモデルを用いて評価される．この重みに比例するように選択確率を定め，$\{s_i^*\}_{i=1}^p$ のリサンプリングを行い，新しい粒子集合 $\{s_i^t\}_{i=1}^p$ をえる．粒子の更新とリサンプリングを T 回繰り返し，$n = p \times T$ のサンプルを生成する．これらを用いて事後分布を近似する．

Algorithm 1 逐次モンテカルロ法による化学構造の生成．

Input 反復数 T，粒子数 p，冷却計画 $\{\beta_t\}_{t=1}^T$，有効サンプルサイズの閾値 E
Output $\mathcal{P}_t$ $(t = 1, \ldots, T)$

p 個の粒子(候補構造)の初期値 $\mathcal{P}_0 = \{s_i^0\}_{i=1}^p$ を生成する．
for $t = 1, \ldots, T$ **do**
  for $i = 1, \ldots, p$ **do**
    構造生成モデル g を用いて，候補粒子 $s_i^* \sim g(s_i^*|s_i^{t-1})$ を生成する．
    尤度関数に基づいて各粒子の重みを更新する：

$$w_i^t = w_i^{t-1} \frac{p(Y = y^*|S = s_i^*)^{\beta_t}}{p(Y = y^*|S = s_i^{t-1})^{\beta_{t-1}}}.$$

  end for
  重みを正規化する：$W_i \propto w_i^t$, $\sum_i W_i = 1$
  有効サンプルサイズ ESS $= p(\sum_i W_i^2)^{-1}$ を計算
  if ESS $\geq E$ **then**
    確率 W_i で $\{s_i^*\}_{i=1}^p$ のリサンプリングを行い，粒子を更新 $\mathcal{P}_t = \{s_i^t\}_{i=1}^p$
    重みを初期化 $w_i^t = 1/p$ $(i = 1, \ldots, p)$
  else
    $\mathcal{P}_t = \{s_i^*\}_{i=1}^p$
  end if
end for

SMC の構成要素の内，提案分布 $g(S^*|S)$ による構造改変が重要な機能を担う．iQSPR-X のデフォルト設定では，拡張 n グラムを用いて以下の計算を実行する．

(a) 一様乱数 $z \sim U(0, 1)$ を抽出する．S が文法的に正しく，z が並べ替え実行確率 κ(デフォルト 0.2)以下であれば，SMILES の文字列を並べ替える $S \to S^*$．そうでなければ，並べ替えを行わず，現在の S を S^* とする．最初の文字をランダムに選択し，Open Babel や RDKit の関数を適用することで，SMILES 文字列の並べ替えを実行する．
(b) S^* の右端の m 文字を削除して $S^{**} = s_{1:g-m}^*$ とする．二項確率 η(デフォルト 0.5)，最大長 L(デフォルト 5)の二項分布 $m \sim B(m|L, \eta)$ に従うように削除する長さ m を決める．
(c) 短縮された文字列の右側に $L - m$ 個の文字を追加する．個々の文字は，言語モデル $s_i \sim p(s_i|s_{1:i-1})$ に従う．終端符号が出現したら伸長を停止して S' をえる．

詳しくは，Ikebata et al. (2017)を参照せよ．

3. 高熱伝導率を持つ高分子の探索

Wu et al. (2019)では，iQSPR-X を適用して新規の高熱伝導性高分子材料を発見した．解析のワークフローを図 2 に示す．一般に，高い熱伝導率を持つ高分子材料は，軟化温度(ガラス転移温度 Tg)や溶融温度(融点 Tm)が十分に高く，高温まで軟化あるいは溶融しない．具体的

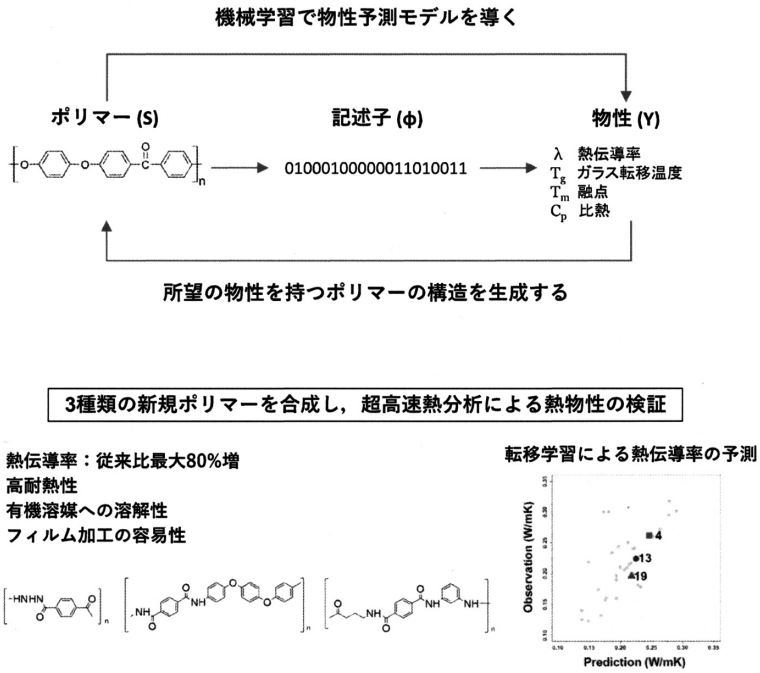


図 2．iQSPR-X を用いた高熱伝導率を持つ熱可塑性樹脂の発見．

には，融解しても取り得る配座構造の変化の少ない剛直な高分子ほど，融解のエントロピーが小さくなり，融点が高くなる．高分子のガラス転移点は，高分子の分子間力や屈曲性，対称性によって支配される．環構造の割合の多い主鎖構造を持つ高分子材料は，融解熱に関わる分子間相互作用ないしは凝集力が大きく，Tg が高くなる．

仮想ライブラリの作製では，高い Tg と Tm を持つ芳香族ポリアミドをターゲットとした．PoLyInfo データベース（Otsuka et al., 2011）からホモポリマーの Tg と Tm のデータを抽出し，5,917 および 3,234 件のデータからランダム 80% を選択して，ベイズ線形回帰で順方向のモデルを構築した．モノマー構造の記述子には ECFP（Rogers and Hahn, 2010）などの複数のフィンガープリントを合わせて使用した．さらに，PoLyInfo の 14,423 個のホモポリマーを用いて言語モデルを訓練し，Tg と Tm の範囲 200–500°C，300–600°C をターゲットに 1,000 種類の仮想ライブラリを作製した．ただし，溶融成形が可能な熱可塑性樹脂の設計では，耐熱性を若干犠牲する必要がある．このことから，事後選択のステップでは，Tg の温度の上限を 300 とした．

次に，機械学習モデルを用いて 1,000 個の候補分子の熱伝導率を推算した．熱伝導率の予測モデルの学習には，高分子物性データベース PoLyInfo に収録されている 28 種類のアモルファスポリマーの熱伝導率のデータを使用した(データの選定と前処理の方法については，Wu et al., 2019 を参照)．データ数が極端に少ないため，通常の教師あり学習では物性予測のモデルを構築できなかった．そこで，転移学習を導入して問題の解決を図った．高分子のガラス転移温度，融点，密度，粘度に加え，低分子化合物の定容比熱容量を元ドメインとした．高分子の物性データは PoLyInfo，低分子化合物の比熱は QM9 という第一原理計算の物性データベースから抽出した（Ramakrishnan et al., 2014）．サンプル数は表 2 に示した通りである．後述の

表 2. 高分子熱伝導率の転移学習に使用したデータセット.

物性	材料	データベース	サンプル数	備考
ガラス転移温度 (Tg)	ポリマー	PoLyInfo	5,917	
融点 (Tm)	ポリマー	PoLyInfo	3,234	
定圧比熱 (Cp)	ポリマー	PoLyInfo	58	アモルファス
熱伝導率 (λ)	ポリマー	PoLyInfo	19	アモルファス，10–35°C
定容比熱 (Cv)	低分子	QM9	133,805	第一原理計算，25°C

手順で各々の元ドメインに対して 100 個の異なるモデルを構築した．28 個のデータを用いて，これらのモデルを熱伝導率の予測モデルに転移した．10 分割交差検証で転移モデルの汎化性能を評価し，平均絶対誤差(mean absolute error, MAE)が最小のモデルを抽出した．

モデルの入力には化学構造のみを用いた．様々なフィンガープリント記述子を連結した後，その中からランダムに抽出した最大で 500 個の要素を機械学習モデルの入力変数とした．ニューラルネットワークの構造もランダムに決めた．ピラミッド型の構造に制限し，ニューロン数と層数をランダムに選択した．このような訓練済みモデルをランダムに 100 個作り，ショットガンアプローチで目標ドメインの MAE を最小にする転移モデルを選定した．

図 3 は，各々の元ドメインにおいて目標ドメインの MAE が最も小さかった転移モデルの交差検証の結果を示している．さらに図 3 には，28 個のデータで直接訓練されたモデルの交差検証の結果も示されているが，汎化性能は極めて低い．一方，全ての元ドメインにおいて，転移モデルは転移学習を介さないモデルの汎化精度を大きく上回っている．

さらに，モノマー構造の液晶らしさや合成可能性などのスコアリングを行い，最終的に 3 個の芳香族ポリアミドに候補を絞り込み，合成・実験検証を行った．選択したのは，全芳香族ポリアミド(分子構造 4)，芳香族ポリヒドラジド(13)，および脂肪属―芳香族ポリアミド(19)の 3 種類である．分子構造 4, 13 はジカルボン酸とジアミンの反応により，分子構造 19 は自己縮合 AB 型モノマーから高分子を合成した．分子構造 4, 13 から合成した高分子には物性の報告事例はなく，分子構造 19 は新規な化学構造であり，全く新しい高分子の合成に成功したことになる．予測された熱伝導率の値は，実験による測定結果と良好な一致を示し(図 4)，さらに熱処理による結晶化の促進により，熱伝導率は 0.41 W/mK に達することが確認された．これは典型的な無配向状態のポリアミド系高分子と比較して約 80% の性能向上に相当する．さらに，高耐熱性や有機溶媒への溶解性，フィルム加工の容易性など，実用に求められる諸特性を併せ持つことも実験的に確認された．

4. まとめ

本稿は，ベイズ推論に基づく分子設計の手法を解説し，機械学習で設計した高分子材料が実際に合成・検証された実例を紹介した．近年，材料研究とデータ科学の融合が急速に進行し，実証的観点からその有効性や可能性について様々な検討が行われている．しかしながら，材料研究の他の領域に比べると，高分子材料の研究とデータ科学の学融合は大きな遅れをとっている．その背景には，高分子物性の世界ではデータ駆動型研究に資するデータベースがほとんど存在しないという事実がある．現在，材料科学の様々な分野では，機械学習への活用を目的としたデータベースの整備が急速に進んでいる(Materials Project（Jain et al., 2013), QM9 (Ramakrishnan et al., 2014)，など)．しかしながら，高分子材料のデータベースの整備はほとんど進んでおらず，本研究で利用した PoLyInfo 以外には，高分子物性を系統的に収集したデータベースは存在しない．さらに，分子動力学シミュレーションなどの高分子物性の理論計算で

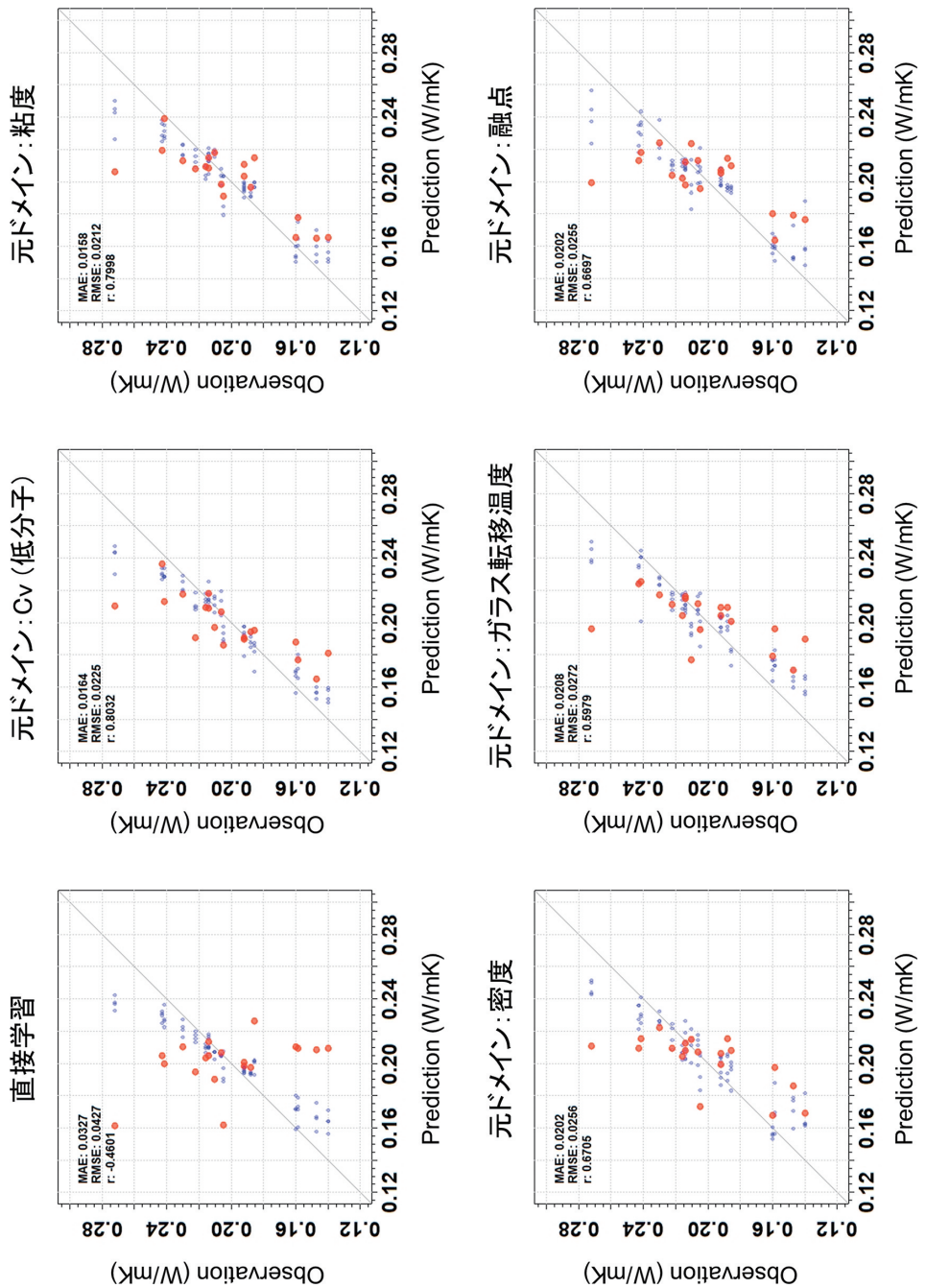


図 3．様々な元ドメイン（高分子のガラス転移温度，融点，密度，粘度，低分子化合物の定容比熱（Cv））から高分子熱伝導率への転移と転移学習を介さない通常の機械学習の 10 分割交差検証の結果．横軸は交差検証の予測値，縦軸は実測値．青点は訓練，赤点は交差検証の結果を表す．

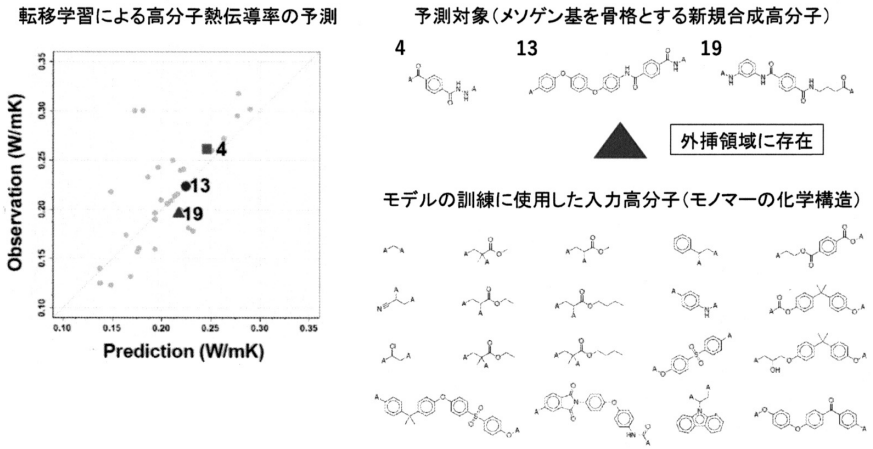


図 4. 左：3 種類の新規ポリマーに対する転移モデルの予測値と実測値．右：新規高分子のモノマーと転移学習で用いた訓練データの化学構造．

は，現時点において，全自動化によるハイスループットデータ生成は技術的に難しいと言われている．少なくとも中長期的観点においては，高分子材料のデータ駆動型研究では，スモールデータの壁をいかに突破するかが鍵を握ることになる．

本稿で紹介した研究では，合成の容易性という観点から 3 種類の高分子のみを選定・合成したが，我々が作製した仮想ライブラリには，他にも有望な候補物質が残されている可能性がある．また，この研究で適用した機械学習の解析技術は汎用的であり，任意の特性をターゲットに同様の解析を行うことができる．これから数年以内に，同様のアプローチから多くの埋蔵物質が発見され，その中から従来の常識を覆すような新しい高分子材料が発掘されることが期待される．

謝　辞

本研究は JST CREST JPMJCR19I3，科研費 19H01132 の助成を受けた．本論文をまとめるにあたり，統計数理研究所ものづくりデータ科学研究センターの皆様には，多くの議論にお付き合いいただきました．心よりお礼申し上げます．

参　考　文　献

Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers, *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), **68**(3), 411–436.

Douguet, D., Thoreau, E. and Grassy, G. (2000). A genetic algorithm for the automated generation of small organic molecules: Drug design using an evolutionary algorithm, *Journal of Computer-Aided Molecular Design*, **14**(5), 449–466.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, 2nd ed., Chapman and Hall/CRC, New York.

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. (2018). Automatic chemical design using a data-driven continuous representation of molecules, *ACS Central Science*, **4**(2), 268–276.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*: *Data Mining, Inference, and Prediction*, Springer-Verlag, New York.

Hutter, F., Hoos, H. H. and Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration, *International Conference on Learning and Intelligent Optimization*, 507–523, Springer, Berlin, Heidelberg.

Ikebata, H., Hongo, K., Isomura, T., Maezono, R. and Yoshida, R. (2017). Bayesian molecular design with a chemical language model, *Journal of Computer-Aided Molecular Design*, **31**(4), 379–391.

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. (2013). The Materials Project: A materials genome approach to accelerating materials innovation, *APL Materials*, **1**(1), p.011002, http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi, DOI: http://dx.doi.org/10.1063/1.4812323.

Kirkpatrick, P. and Ellis, C. (2004). Chemical space, *Nature*, **432**, p.823.

Lameijer, E.-W., Kok, J. N., Bäck, T. and IJzerman, A. P. (2006). The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules, *Journal of Chemical Information and Modeling*, **46**(2), 545–552.

Liu, C. and Wu, S. (2021). XenonPy-iQSPR tutorial, https://xenonpy.readthedocs.io/en/stable/tutorials/7-inverse-design.html.

Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*, Springer Verlag, New York, Berlin, Heidelberg.

Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. and Ramprasad, R. (2016). Machine learning strategy for accelerated design of polymer dielectrics, *Scientific Reports*, **6**, p.20952.

Miyao, T., Kaneko, H. and Funatsu, K. (2016). Inverse QSPR/QSAR analysis for chemical structure generation (from y to x), *Journal of Chemical Information and Modeling*, **56**(2), 286–299.

持橋大地, 大羽成征 (2019).『ガウス過程と機械学習』, MLP 機械学習プロフェッショナルシリーズ, 講談社.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. and Yamazaki, M. (2011). PoLyInfo: Polymer database for polymeric materials design, *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 22–29.

Ramakrishnan, R., Dral, P. O., Rupp, M. and Von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules, *Scientific Data*, **1**(1), 1–7.

Rasmussen, C. E. (2003). Gaussian processes in machine learning, *Summer School on Machine Learning*, 63–71, Springer, Berlin, Heidelberg.

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints, *Journal of Chemical Information and Modeling*, **50**(5), 742–754.

Segler, M. H., Kogej, T., Tyrchan, C. and Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks, *ACS Central Science*, **4**(1), 120–131.

Venkatasubramanian, V., Chan, K. and Caruthers, J. M. (1994). Computer-aided molecular design using genetic algorithms, *Computers & Chemical Engineering*, **18**(9), 833–844.

Venkatasubramanian, V., Chan, K. and Caruthers, J. M. (1995). Evolutionary design of molecules with desired properties using the genetic algorithm, *Journal of Chemical Information and Computer Sciences*, **35**(2), 188–195.

Venkatraman, V. and Alsberg, B. K. (2018). Designing high-refractive index polymers using materials informatics, *Polymers*, **10**(1), p.103.

Wu, S., Kondo, Y., Kakimoto, M.-A., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., Morikawa, J. and Yoshida, R. (2019). Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm, *npj Computational Materials*, **5**(1), 1–11.

Wu, S., Lambard, G., Liu, C., Yamada, H. and Yoshida, R. (2020). iQSPR in XenonPy: A Bayesian molecular design algorithm, *Molecular Informatics*, **39**(1-2), p.1900107.

Machine Learning for Automated Molecular Design with Application to the Discovery of New Polymers with High Thermal Conductivity

Ryo Yoshida[1], Stephen Wu[1] and Junko Morikawa[2]

[1]The Institute of Statistical Mathematics
[2]School of Materials and Chemical Technology, Tokyo Institute of Technology

We aim to design chemical structures with desired properties by applying analytical techniques of Bayesian inference and machine learning. Based on data obtained from experiments or simulations, we derive a model that forwardly predict physical, chemical, electronic, thermodynamic, mechanical properties of any give chemical structure. The Bayes rule of conditional probability is applied to this forward model to derive the backward prediction model from property to structure. By generating hypothetical molecules from this model, we identify promising candidates that exhibit the desired properties. We have successfully applied this approach to discover new plastic polymers with thermal conductivity reaching 0.41 W/mK. This corresponds to a performance improvement of about 80% compared to a conventional unoriented polyamide polymer. In this paper, we describe the technology of the Bayesian molecular design algorithm, and then illustrate its application to the study of polymer thermophysical properties.

Key words: Molecular design, Bayesian inference, transfer learning, polymer, thermal conductivity.

統計数理(2021)
第 69 巻 第 1 号 49–63

特集「マテリアルズインフォマティクスの最前線」
[総合報告]

材料研究における転移学習の応用

劉 暢[1]・山田 寛尚[1,2]・ウ ステファン[1,3]

(受付 2020 年 11 月 6 日;改訂 2021 年 5 月 10 日;採択 5 月 12 日)

要　　旨

材料研究のデータ量は,機械学習の他の応用分野に比べると圧倒的に少ない.このスモールデータの壁を乗り越えるために,転移学習を活用する.化学的特性,電気的特性,熱力学的特性,機械特性など,材料特性の間には物理化学的な依存関係が存在している.限られたデータからある特性を予測するために,十分なデータが利用可能な代理特性のモデルを事前に学習し,このモデルを目標物性の予測に利用する.このように他のドメインから獲得したモデルや特徴表現を目標ドメインの予測に利用することで,非常に少ないデータでも高い予測能力を持つモデルを構築できることがある.本稿では,高分子や無機材料を含む様々な問題設定で転移学習を適用し,その潜在的能力をデモンストレーションする.特に,転移学習を適用することで,訓練データ集合の分布の範囲を大きく逸脱した領域において予測能力を獲得した事例を報告する.

キーワード:転移学習,物性予測,スモールデータ,有機材料,ポリマー,無機材料.

1. はじめに

従来の材料研究では,候補材料の特性を評価するために分子動力学計算や第一原理計算のような物理モデルに基づく計算機実験が活用されてきた.しかしながら,一般にシミュレーションは膨大な計算コストを要するため,網羅的な材料スクリーニングへの適用は難しい.そこで,計算コストが小さい統計モデルに特性評価の計算を代替させて,膨大な数の候補材料のスクリーニングを実現しようという研究が様々な材料系を対象に進行している(Carrete et al., 2014; Seko et al., 2015; Gómez-Bombarelli et al., 2016; Hansen et al., 2016; Oliynyk et al., 2016; Sumita et al., 2018; Matsumoto et al., 2018; Wu et al., 2019; Liu et al., 2021).多いときで,仮想ライブラリを含む数億オーダーの候補材料を対象にスクリーニングが実施される.機械学習の目的は,候補材料 S の特性 Y を計測した実験や物理シミュレーションのデータ集合を用いて,予測モデル $Y = f(S)$ を導くことである.問題形式は単純な教師あり学習である.

データ駆動型材料研究に立ちはだかる最も大きな壁は,限られたデータ量とデータの多様性の不足の問題である.画像認識や自然言語処理などのデータ科学の他の応用分野と比べると,材料研究に利用可能なデータの量は圧倒的に少ない.例えば,本稿で示す無機材料の熱伝導率の研究では,データ数がたったの 45 個しかない.データが少ない主な原因として,次の三点が考えられる.

[1] 統計数理研究所:〒 190–8562 東京都立川市緑町 10–3
[2] 東京薬科大学 薬学部:〒 192–0392 東京都八王子市堀之内 1432–1
[3] 総合研究大学院大学 複合科学研究科統計科学専攻:〒 190–8562 東京都立川市緑町 10–3

- 実験や計算機実験のコストが高い．
- 材料組成の選択や材料作製のプロセス設計(温度依存性，添加物・溶媒選択)など，材料特性の決定には非常に多くの因子が関与する．したがって，一般に設計空間が極めて広大になる．さらに，個々の研究者の研究対象が大きく異なるため，社会全体でコモンデータを創出しようという動きが起こりにくい．
- 科学的成果と産業応用の垣根が低いため，競合相手に対する情報秘匿の意識が高く，データ公開に対するインセンティブが研究者に働きにくい．

以上の理由により，コミュニティが協力してコモンデータを創出しようという動向は極めて低調である．さらに，研究対象が先端領域に近づくにつれ，スモールデータの傾向はより顕著になる．少なくとも中長期的には，大学の研究室や一企業で生産可能なデータがマテリアルズインフォマティクスの標準的な解析対象になると予想される．

本稿では，転移学習という方法論を用いて材料研究のスモールデータの問題にアプローチする（Agrawala and Choudhary, 2016; Hutchinson et al., 2017; Oda et al., 2017; Jalem et al., 2018; Yonezu et al., 2018; Kailkhura et al., 2019; Segler et al., 2018; Cubuk et al., 2019; Li et al., 2018; Kaya and Hajimirza, 2019)．転移学習はあるドメイン(元ドメイン)の学習モデルを別のドメイン(目標ドメイン)に利活用するための方法論である．目標ドメインのデータ量が不足している場合，一旦，十分な量のデータを利用できる元ドメインのモデルを構築する．この訓練済みモデルの特徴量や推定されたパラメータを目標ドメインのスモールデータで改変して最終的なモデルを導く．データ量が少なくてフルスクラッチでの学習は難しいが，関連する元ドメインの訓練済みモデルを適切に利用することでデータ量の不足を補う．このようなアプローチがスモールデータの壁を乗り越える有効な手段になりうることが，様々な分野で実証されつつある．本稿では，有機高分子や無機化合物の物性予測など，様々な材料研究における転移学習の成功事例を紹介する．特に転移学習が有する外挿的な予測能力をデモンストレーションする．なお，本稿で示す解析結果の詳細な説明については，著者らの原著論文 Yamada et al. (2019) を参照せよ．

2. 転移学習

2.1 ニューラルネットワークを用いた教師あり転移学習

本稿はニューラルネットワークを用いた教師あり転移学習に焦点を絞る．特別な工夫は何も施さず，教科書的なテクニックのみを用いる．一般にニューラルネットワークの学習では，入力層に近い下層のニューロンが一般的な特徴量を表し，出力層に近づくにつれてドメイン固有の特徴量に変換されていく．ニューラルネットワークの転移学習はこの性質を利用する．

ここで，転移元のドメインを元ドメイン(source task)，転移先の目標ドメイン(target task)と呼ぶことにする．元ドメインの教師データを用いて，モデル $Y_s = f_s^L(X)$ を構築する．Y_s と X は系の元ドメインの出力変数と入力変数，$f_s(X) = f_L \circ f_{L-1} \ldots \circ f_1(X)$ は L 層のニューラルネットワークを表す．この訓練済みモデルの第 K 層($K < L$)までの部分モデル $\phi(X) = f_K \circ f_{K-1} \ldots \circ f_1(X)$ を目標ドメインのモデルの記述子として利用する方法を特徴抽出による転移学習と呼ぶ．すなわち，目標ドメインの教師データを用いて $Y_t = f_t \circ \phi(X)$ という形式のモデルを学習する．ここで，Y_t は目標ドメインの出力変数，f_t は任意のモデルである．図 1 は，転移学習のワークフローを模式的に表したものである．元ドメインの学習過程で，ニューラルネットワークは Y_s の予測に有用な特徴量 $\phi(X)$ を獲得する．元ドメインと目標ドメインの間に共通のメカニズムが存在すれば，この特徴量は Y_t の予測にも活用できることが期待される．$\phi(X)$ の次元を十分に小さくとることができ，かつ線形モデルのような簡素な

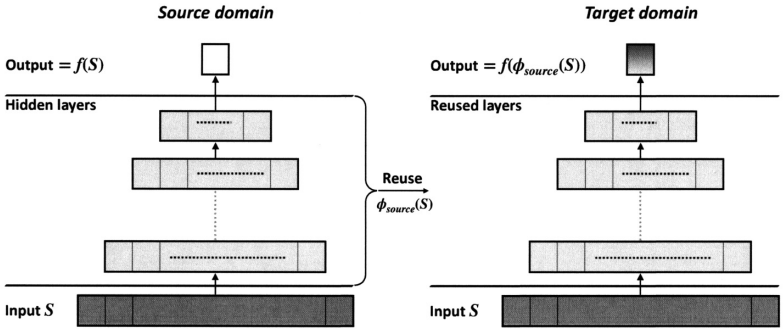


図 1. 元ドメインの訓練済みモデルの一部(この例では出力層を除いた部分モデル)を目標ドメインの特徴抽出器として活用するタイプの転移学習.

モデルで f_t を記述できる系であれば，目標ドメインのモデルをフルスクラッチで学習するよりも，少ないデータ量で高い予測精度のモデルを構築できるかもしれない.

次にファインチューニング(fine-tuning)と呼ばれる転移学習の方法を解説する．ファインチューニングは訓練済みモデルの重みを初期値とし，目標ドメインのデータセットを用いて再学習する．目標ドメインの学習の際は，パラメータ更新を早期に停止し，重みを大きく変化させずに，訓練済みモデルを微修正する．適切な早期停止を実施するために，目標ドメインのデータを訓練用と検証用に分割した上で，訓練用データを用いて低い学習率でパラメータを更新しながら，検証用データの予測精度の変化をモニタリングし，精度が最も高くなるタイミングでパラメータ更新を停止する．理論的な根拠は明確ではないが，ファインチューニングは経験的に有効なアプローチであることが知られており，特に画像認識や機械翻訳等の訓練済みモデルを対象に広く用いられている.

2.2 訓練済みモデルライブラリ XenonPy.MDL

本グループは，XenonPy (https://xenonpy.readthedocs.io/)という Python パッケージを開発している．XenonPy は，様々な材料を対象に材料設計のワークフローを構築するために必要なモジュール群を実装している．本稿では XenonPy の解説は行わないが，ここでは，サブモジュールの一つである XenonPy.MDL という物性予測タスクを対象とする訓練済みモデルライブラリを紹介する．このライブラリには 2020 年時点で，低分子，高分子，無機材料の 45 種類の特性を予測する約 140,000 個の訓練済みモデルが実装されている(表 1)．ユーザーは API (Application Programming Interface)を用いて訓練済みモデルを取得し，XenonPy を経由して材料設計の様々なワークフローを構築できる．脳の学習メカニズムとの対比で言えば，多様且つ包括的な訓練済みモデル群を実装することは，多くの経験から記憶の集合体を獲得することに相当する．モデルの多様性が増すほど強力な転移学習を実現できる可能性が高まる.

3. 転移学習の適用例

ここからは，重要物性である熱伝導率と屈折率を対にした転移学習の三つの適用例を取り上げ，転移学習の有効性を示す.

3.1 無機化合物の熱伝導率

熱伝導率は，伝熱，対流，放熱等の熱の移動を考える上で，材料の断熱性能を表す指標であ

表 1. XenonPy.MDL に収録されている訓練済みモデルの抜粋．低分子，高分子，無機材料の 45 種類の特性を予測するモデルを実装している（2020 年 10 月時点）．

Material type	Database	Property	Model type	Model parameters	No. of models	Best model correlation	No. of descriptors	Descriptor type
Organic	PoLyInfo (polymer)	Glass transition temperature	RF-R	RF setup 1	1,000	0.950	max 500*	rcdk-all
			GB-R	GB setup	1,000	0.950	max 500*	rcdk-all
			EN-R	EN setup	1,000	0.920	max 500*	rcdk-all
			NN-R	NN setup 1	1,000	0.950	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.955	2,048	RDKit-5
		Density	NN-R	NN setup 1	1,000	0.910	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.859	2,048	RDKit-5
		Viscosity	NN-R	NN setup 1	1,000	0.890	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.613	2,048	RDKit-5
		Melting temperature	NN-R	NN setup 1	1,000	0.880	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.885	2,048	RDKit-5
		Heat capacity (const. pressure)	NN-R	TL setup 1	25,000	0.992	max 400-600#	rcdk-all
		Thermal conductivity	NN-R	TL setup 1	25,000	1.000	max 400-600#	rcdk-all
	QM9 (small molecule)	Heat capacity at constant volume	NN-R	NN setup 1	∼500	0.900	max 400-600#	rcdk-all
		LUMO	NN-R	NN setup 1	∼500	0.950	max 400-600#	rcdk-all
		HOMO-LUMO gap	NN-R	NN setup 1	∼500	0.940	max 400-600#	rcdk-all
		Zero point vibrational energy	NN-R	NN setup 1	∼500	0.940	max 400-600#	rcdk-all
		Internal energy at 0 K	NN-R	NN setup 1	∼500	0.920	max 400-600#	rcdk-all
		Enthalpy at 298.15 K	NN-R	NN setup 1	∼500	0.910	max 400-600#	rcdk-all
		Free energy at 298.15 K	NN-R	NN setup 1	∼500	0.910	max 400-600#	rcdk-all
		HOMO	NN-R	NN setup 1	∼500	0.880	max 400-600#	rcdk-all
		Internal energy at 298.15 K	NN-R	NN setup 1	∼500	0.880	max 400-600#	rcdk-all
		Isotropic polarizability	NN-R	NN setup 1	∼500	0.870	max 400-600#	rcdk-all
		Electronic spatial extent	NN-R	NN setup 1	∼500	0.800	max 400-600#	rcdk-all
		Dipole moment	NN-R	NN setup 1	∼500	0.740	max 400-600#	rcdk-all
	Polymer Genome (polymer)	Bandgap	RF-R	RF setup 2	1,000	0.964	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.985	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.983	2,048	RDKit-5
		Dielectric constant	RF-R	RF setup 2	1,000	0.965	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.982	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.958	2,048	RDKit-5
		Ionic dielectric constant	RF-R	RF setup 2	1,000	0.898	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.934	max 400-600#	rcdk-all
		Electronic dielectric constant	RF-R	RF setup 2	1,000	0.930	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.947	max 400-600#	rcdk-all
		Refractive index	RF-R	RF setup 2	1,000	0.953	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.985	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.981	2,048	RDKit-5
		Atomization energy	RF-R	RF setup 2	1,000	0.974	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.986	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.992	2,048	RDKit-5
		Density	RF-R	RF setup 2	1,000	0.961	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.982	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.989	2,048	RDKit-5
		Ionization energy	RF-R	RF setup 2	1,000	0.922	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.962	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.940	2,048	RDKit-5
		Electron affinity	RF-R	RF setup 2	1,000	0.955	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.978	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.987	2,048	RDKit-5
		Cohesive energy	RF-R	RF setup 2	1,000	0.839	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.943	max 400-600#	rcdk-all
		Melting temperature	RF-R	RF setup 2	1,000	0.920	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.94	max 400-600#	rcdk-all

表 1.（つづき）

Material type	Database	Property	Model type	Model parameters	Num. of models	Best model correlation	Num. of descriptors	Descriptor type
Organic	Polymer Genome (polymer)	Glass transition temperature	RF-R	RF setup 2	1,000	0.937	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.962	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.931	2,048	RDKit-5
		Hildebrand solubility parameter	RF-R	RF setup 2	1,000	0.951	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.962	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.879	2,048	RDKit-5
		Molar heat capacity	RF-R	RF setup 2	1,000	0.989	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.991	max 400-600#	rcdk-all
		Molar volume	RF-R	RF setup 2	1,000	0.965	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.984	max 400-600#	rcdk-all
	PHYSPROP	Boiling point	NN-R	NN setup 1	1,000	0.782	max 400-600#	rcdk-all
	MD database	Solvation free energy	NN-R	NN setup 1	1,000	0.94	max 400-600#	rcdk-all
	Jean-Claude Bradley	Melting temperature	NN-R	NN setup 1	1,000	0.84	max 400-600#	rcdk-all
Inorganic	Materials Project	Volume	NN-Py	NN setup 3	3,600%	0.997	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.606	N/A	N/A
		Formation energy per atom	NN-Py	NN setup 3	3,600%	0.997	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.977	N/A	N/A
		Total energy per atom	NN-Py	NN setup 3	3,600%	0.996	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.963	N/A	N/A
		Density	NN-Py	NN setup 3	3,600%	0.994	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.996	N/A	N/A
		Fermi energy	NN-Py	NN setup 3	3,600%	0.968	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.933	N/A	N/A
		Magnetization	NN-Py	NN setup 3	3,600%	0.923	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.723	N/A	N/A
		Bandgap	NN-Py	NN setup 3	3,600%	0.910	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.936	N/A	N/A
	Citrination datasets id:152062	Total dielectric constant	NN-Py	NN setup 3	3,600%	0.565	290/150	XenonPy
		Electronic dielectric constant	NN-Py	NN setup 3	3,600%	0.504	290/150	XenonPy
		Refractive index	NN-Py	NN setup 3	3,600%	0.762	290/150	XenonPy
	Shiomi data	Scattering phase space	NN-Py	NN setup 3	~1,200	0.912	290/150	XenonPy
		Lattice thermal conductivity	NN-Py	NN setup 3	~1,200	0.998	290/150	XenonPy
			NN-Py	TL setup 2	~200	0.999	290	XenonPy

訓練済みモデルの概要．RF-R，GB-R，EN-R，NN-R はそれぞれ，ランダムフォレスト(ranger)，勾配ブースティング(xgboost)，Elastic Net 回帰(glmnet)，ディープニューラルネットワーク(MXnet)を表す．ここで括弧内のシンボルは R のパッケージ名を表す．NN-Py と RF-Py はそれぞれ，ディープニューラルネットワーク(PyTorch)，ランダムフォレスト(scikit-learn)を表す．括弧内のシンボルは Python のパッケージ名である．CGCNN-Py は crystal graph convolution neural network(PyTorch)を表す．RF setup 1 は回帰木の数(nTree)を 100–800，特徴量の数(mTry)を 20–100 の範囲でランダムに選んでいる．RF setup 2 の場合は，nTree が 50–500，mTry が 50–500 である．GB setup は学習率(*eta*)を 0.1–1，決定木の深さの最大を 3–10，学習回数(nround)を 50–200 の範囲に設定している．EN setup は Elastic Net の正規化パラメータ(λ)をランダムに選択し，*alpha* を 0–1 の範囲に設定している．NN setup 1 は，ニューラルネットワークの学習のエポック数を 3,000–4,000 の範囲に設定．隠れ層を 3 もしくは 4 に設定し，最初の隠れ層のニューロン数の最大値を 400，隠れ層の最終層のニューロン数を 10–30 の範囲に設定している．NN setup 2 は最初の隠れ層のニューロン数の最大値を 1,640 に設定し，他のパラメータは NN setup 1 と同様に設定している．NN setup 3 はエポック数を 1,000–3,000，隠れ層の数を 3–6 の範囲に設定し，最初の隠れ層のニューロン数は 348 に固定している．隠れ層の最終層のニューロン数の最小値を 5 に設定している．TL setup 1 はランダムフォレストの入力を元ドメインの隠れ層の最終層のニューロンを利用する．nTree と mTry は訓練データ数と訓練データ数の半分の範囲に設定している．TL setup 2 は RF-Py の入力に SPS のベスト訓練済みモデルの全ての隠れ層を連結し，全隠れ層からランダムに選択したニューロンを利用している．nTree は 200 に設定し，選択されたニューロン数の最大値は隠れ層の総ニューロン数の平方根を取った値に設定している．rcdk-all は rcdk で利用可能なフィンガープリント(standard, extended, graph, hybridization, maccs, estate, pubchem, kr, circular)を連結したものを利用したものを表す．RDKit-5 は Atom-Pair，Topological-Torsion フィンガープリント，Morgan フィンガープリント(特徴量ベースの不変量の有・無)，RDKit に含まれる基本的なフィンガープリントを利用したものを表している．XenonPy は XenonPy パッケージに搭載されている化学組成と RDF 記述子を利用したものを表す．(*) 11,106 bits の全記述子の中で 90% 以上が 0 であるフィンガープリントを取り除き，残った記述子の中からランダムに選択(最大 500 個)．(#) *と同様の処理を行い，400–600 もしくは 1,500–3,000 の範囲でランダムに選択．(%) compositional 記述子モデル；1,200，RDF for stable 記述子モデル；1,200，compositional for unsable 記述子モデル；1,200．

る．我々のグループの過去の研究から，転移学習により高い熱伝導率を持つ無機化合物の同定に成功した例を紹介する（Ju et al., 2021）．45 化合物の格子熱伝導率（LTC: lattice thermal conductivity）を予め第一原理計算で算出し，このデータを用いて化学組成から LTC を予測するモデルを構築する．データは 45 件しかないため，何の工夫もない機械学習では予測精度を引き出すことは難しい．そこで，散乱位相空間（SPS: scattering phase space）という中間物性を元ドメインに定め，転移学習で問題解決を図る．LTC に比べると SPS の第一原理計算のコストは圧倒的に低く，320 化合物に対する物性データを用意した．図 2(a)に示すように，SPS と LTC の間には弱い負の相関が存在する．

転移学習によるモデル構築とスクリーニングの手順および結果は，以下の通りである．

(1) XenonPy の *xenonpy.descriptor.Compositions* モジュールで算出した 290 次元の組成記述子を入力とする（記述子の詳細については，Liu et al., 2021 を参照）．
(2) ニューラルネットワークの最大層数を 5 とし，ランダムに 100 個のネットワーク構造を生成し，SPS のモデルを訓練する．ネットワーク構造は，入力層（290 ニューロン）から出力層（1 ニューロン）にかけてニューロンの数が単調減少するように制約する（ピラミッド型）．ハイパーパラメータの詳細については，XenonPy に組み込まれているサンプルコードを参照してほしい．
(3) 100 個の訓練済みモデルを，45 件のデータを用いて LTC の予測モデルに転移する．このとき 10 分割交差検証（クロスバリデーション）を実施し，検証用データセットに対して平均絶対誤差（MAE: mean absolute error）が最も小さい転移モデルを抽出する（図 2(b)）．
(4) 転移モデルで Materials Project の約 140,000 化合物の LTC を予測し，14 個の化合物を同定（選択基準の詳細については，Ju et al., 2021 を参照）．
(5) 第一原理計算で 14 化合物の LTC を検証．

同定された 14 個の化合物の LTC 値に対する検証結果を図 3 に示す．14 個の化合物の LTC は最高で 3,000 W/mK を超える水準に到達している．一方，訓練に使用した 45 個の化合物の LTC は 400 W/mK に満たない領域に分布していることがわかる（図 3 のヒストグラム）．機械学習は「入力が近ければ，出力も近い」という原理に従い予測を行うため，一般的にモデルは訓

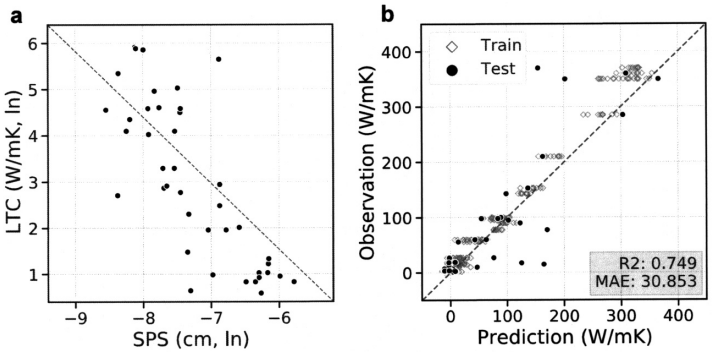


図 2. SPS の訓練済みモデルから LTC への転移学習．(a) SPS（元ドメイン）と LTC（目標ドメイン）の同時分布．SPS と LTC は自然対数の値をプロット．(b) 検証用データに対する SPS（元ドメイン）と転移モデルによる LTC（目標ドメイン）の予測（10 分割交差検証）．横軸と縦軸は予測値と実測値を表す．ダイアモンド（白）とサークル（黒）はそれぞれ訓練データとテストデータを表す．

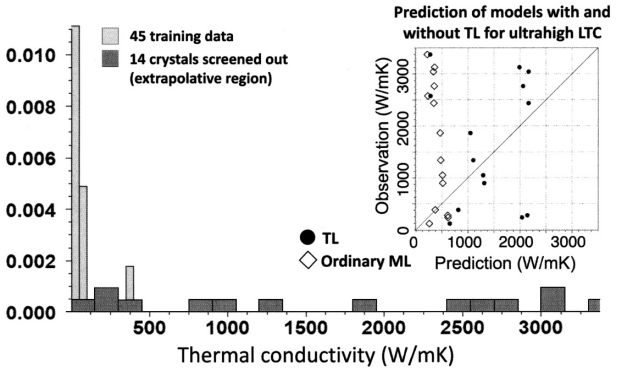


図 3. 転移モデルによる高熱伝導化合物のスクリーニング結果を第一原理計算で検証した結果．ヒストグラム：45 点の訓練データ（灰色）とスクリーニングで同定された 14 化合物の LTC の分布（黒）．LTC は第一原理計算で算出した．散布図：直接訓練したモデル（Ordinary ML）と転移モデル（Transfer Learning: TL）を用いた 14 化合物の予測値の比較．

練データの分布の近傍でのみ予測能力を有する．実際，45 個のデータのみを用いて構築したニューラルネットワーク（直接訓練）は 14 個の化合物の LTC を全く予測できないことがわかる．一方，SPS を経由した転移学習のモデルは，14 個の化合物の LTC をある程度予測できることがわかる．転移学習の予測モデルには，本事例のように外挿性が備わっているケースがしばしば観測されている（Yamada et al., 2019）．元ドメインの 320 個のデータに汎用的な特徴量の獲得に寄与する何らかの情報が含まれており，この特徴抽出器を再利用することで訓練データの水準を大きく超えた領域においても予測性能を有するモデルを構築できた．確かなことは言えないが，これがこの結果に対する自然な解釈である．どのような状況で外挿性が発現するのかは分からない．

3.2 結晶性ポリマーと無機結晶化合物の屈折率

次に結晶性ポリマーと無機結晶化合物における転移学習の事例を紹介する．ここでの予測対象は，有機分子と無機化合物の屈折率である．

Polymer Genome（Mannodi-Kanakkithodi et al., 2018）という高分子物性のデータベースには，第一原理計算で算出した 853 個の高分子の屈折率が収録されている．無機化合物の屈折率については，Citrination（https://citrination.com/）というデータベースから抽出した 1,056 件のデータを使用した．高分子，無機化合物ともに構造情報を一切利用せず，モデルの入力変数は組成のみとし，XenonPy の 290 次元組成記述子を用いて屈折率を予測した．

図 4 は，全サンプルの記述子行列と物性値の関係を可視化したヒートマップである．この図から高分子と無機化合物の各々の記述子と物性の相関パターンが読み取れるが，高分子と無機化合物の間に共通性はほとんどないことが分かる．図 5（b）は，t-SNE（van der Maaten and Hinton, 2008）という次元圧縮の手法を用いて 290 次元の記述子ベクトルを二次元平面に配置した結果である．この図からも分かるように，高分子と無機化合物の記述子ベクトルは，特徴空間上でかなり離れた位置に分布しており，無機化合物の組成パターンのわずか一部にのみ，高分子とのオーバーラップがみられる．

無機化合物から高分子への転移学習と高分子から無機化合物への転移学習を行った．4 層の隠れ層からなるピラミッド型のニューラルネットワークで元ドメインの訓練済みモデルを構築

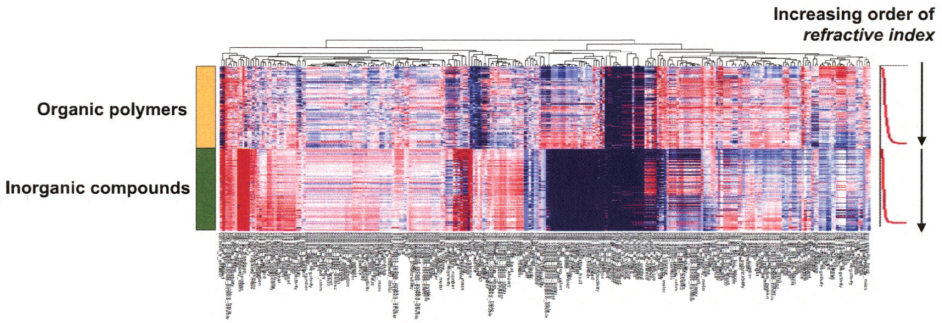


図 4. 1,056 個の無機化合物と 853 個の高分子の組成記述子(横軸)のヒートマップ表示．屈折率の昇順にサンプルを上から下に並び替えている．屈折率と相関の高い記述子には，何らかの特徴的なパターンがみられる．高分子と無機化合物の間には，ほとんど共通性がないことが分かる．

し，最上位の隠れ層を記述子に用いて．ランダムフォレストで目標ドメインのモデルに転移した．ハイパーパラメータ等の詳細は，論文 Yamada et al. (2019)を参照せよ．元ドメインの訓練には全データを使用し，転移の際は全データの 80% を訓練，残りをテストデータに使用した．

まずは，無機化合物から高分子への転移学習の結果をみてみる．図 5(a)に示すように，無機のデータで訓練したモデルに高分子の組成を入力しても屈折率をほとんど予測できない(図 5(a)上図)．一方，無機化合物のモデルを高分子に転移したモデルは，高分子の屈折率を高い精度で予測できる(図 5(a)下図)．図 4 や図 5(a)で示したように，一見すると無機化合物と高分子の屈折率の間には共通性はほとんどなさそうである．それにもかかわらず，転移学習の結果は両者の間に何らかの共通性が存在することを示唆しているが，この結果は極めて非直感的である．

一方，高分子から無機化合物に転移したモデルは，無機化合物の屈折率を全く予測できないことがわかる(図 6(a)(b))．この転移の不可逆性は，転移学習の本質の一つを捉えている．高分子の組成データには計 17 種類の元素しか含まれておらず，C, H, O, Cl の元素に偏っている．一方，無機化合物の組成データは遷移金属を含む計 63 種類の元素から構成されている．したがって，構成元素の包含関係としては，高分子の組成は無機化合物の部分集合ということになる．したがって，高分子の訓練済みモデルには 17 種類以外の元素に対する表現能力が備わっておらず，無機化合物の入力組成は外挿領域に存在すると考えられる．

3.3 高分子の熱伝導率

高分子物性データベース PoLyInfo (Otsuka et al., 2011)に収録されている 19 個のアモルファスポリマーの熱伝導率のデータを使用する．データの選定と前処理の方法については，論文 Wu et al. (2019)を参照せよ．

高分子のガラス転移温度，融点，定圧比熱容量，粘度に加えて，低分子化合物の定積比熱容量を元ドメインとした．高分子の物性データは PoLyInfo，低分子化合物の比熱のデータは QM9 (Ramakrishnan et al., 2014; Ruddigkeit et al., 2012)という第一原理計算の物性データベースから抽出した．後述の手順で各々の元ドメインに対して 1,000 個の異なるモデルを構築し，これらのモデルを 19 個のデータを用いて熱伝導率の予測モデルに転移した．ここで 5 分割交差検証を適用して転移学習モデルの汎化性能を評価し，MAE が最小のモデルを抽出する．

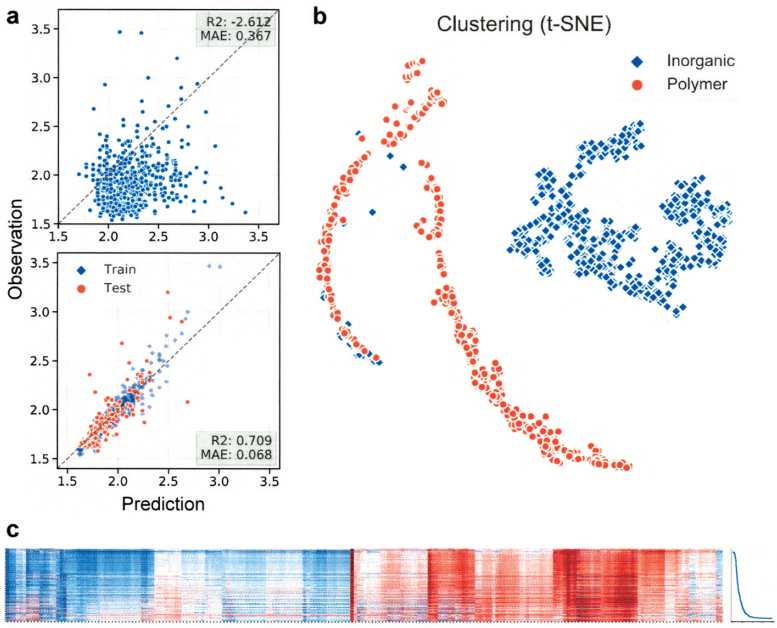


図 5. 無機化合物から高分子屈折率への転移学習．(a)上図は無機のデータで訓練したモデルに高分子の組成を入力した際の屈折率の予測値を表す．下図は，無機化合物の訓練済みモデルを高分子のモデルに転移した結果．訓練データとテストデータはダイアモンド(青色)とサークル(橙色)で表される．両図とも，横軸は予測値，縦軸は観測値を表す．(b)t-SNE による無機化合物と高分子の組成記述子の二次元平面への可視化．無機結晶化合物と結晶性ポリマーはダイアモンド(青色)とサークル(橙色)で表される．(c)無機化合物の訓練済みモデルに 853 個の高分子の組成を入力した際の隠れ層(特徴量)をヒートマップで可視化した結果．853 個のサンプルは屈折率の大きさ順に並び替えている．転移学習では，これらの特徴量を記述子とした．

モデルの入力にはモノマーの化学構造のみを用いた．RDKit に実装されている 9 種類のフィンガープリント記述子(ECFP, FCFP, MACCS など)を連結し，11,106 次元の記述子ベクトルを構築した．その中からランダムに抽出した 400~600 個の要素を機械学習モデルの入力変数とした．ニューラルネットワークの構造もランダムに決めた．ピラミッド型の構造に制限してニューロン数と層の数をランダムに選択した．各々の元ドメインにおいて，このような訓練済みモデルをランダムに 100 個作り，最終隠れ層を記述子としてランダムフォレストでモデルを構築した．100 個の転移モデルの内，目標ドメインの 5 分割交差検証の平均 MAE を最小にする転移モデルを選定した．

図 7 に，各々の元ドメインからの転移の結果を示している．それぞれ，目標ドメインの MAE が最も小さかった転移モデルの 5 分割交差検証の結果を示している．さらに，図 7 には，転移学習を適用せずに直接訓練したモデルの 5 分割交差検証の結果も示している．転移学習を経由しない通常のモデルの汎化性能は極めて低い．一方，全ての元ドメインにおいて，転移モデルは通常のモデルの汎化精度を大きく上回った．

Wu et al. (2019)では，低分子化合物の定圧比熱容量からの転移モデルを選択し，モデルの逆問題を解き，高い熱伝導率に達すると予想されるモノマー分子を設計した．最終的に 3 種類

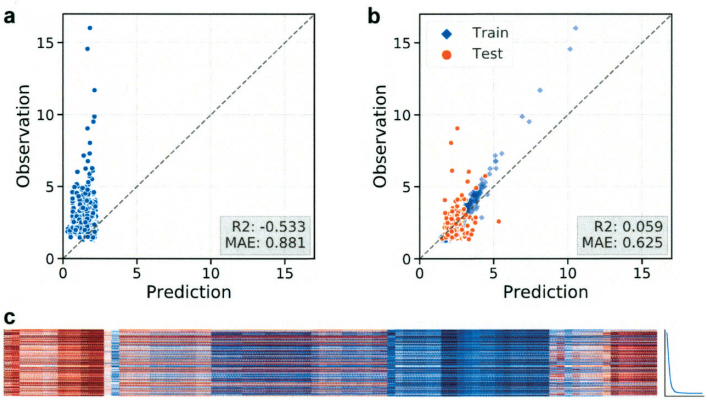


図 6. 高分子から無機化合物への転移学習．(a)高分子のデータで訓練したモデルに無機化合物の組成を入力した屈折率の予測値を表す．横軸は予測値，縦軸は観測値を表す．(b)高分子の訓練済みモデルを無機化合物のモデルに転移した結果．訓練データとテストデータはダイアモンド(青色)とサークル(橙色)で区別される．(c)高分子屈折率の訓練済みモデルに無機化合物の組成を入力した際の隠れ層(特徴量)をヒートマップで可視化した結果．サンプルは屈折率の大きさ順に並び替えている．転移学習では，これらの特徴量を記述子とした．

の芳香族ポリアミドに候補を絞り込み，ポリマー合成と物性測定を実施した．合成された高分子の一つは，熱伝導率が 0.41 W/mK に達することが確認された(詳細は Wu et al., 2019 を参照)．これは典型的な無配向のポリアミド系高分子と比較して約 80% の性能向上に相当する．図 8 に示すように，熱伝導率の実験値は転移モデルの予測値と概ね一致している．ここで注目すべきは，合成した高分子と類似する化学構造は 19 個の訓練データにほとんど含まれていない点である．無機化合物の熱伝導率のケースと同様に，転移学習の外挿性が観察された．

4. まとめ

本稿では，材料研究の複数のタスク(スモールデータに基づく低分子・高分子・無機化合物の物性予測)を例に取り上げ，転移学習が有する潜在的な予測性能を実験的に示した．材料研究のスモールデータの限界を乗り越える上で，転移学習の活用は有望な解決方策を与える．本研究では，低分子，高分子，無機化合物の 45 種類の特性を対象に約 140,000 個の機械学習の予測モデルを開発し，訓練済みモデルライブラリ XenonPy.MDL に実装した．XenonPy のユーザーは，API を用いてライブラリから転移元モデルの候補を取得し，転移されたモデルを用いて材料設計のワークフローを構築できる．優れた研究者が過去の経験から大量且つ多様な記憶を獲得しているのと同様に，転移学習の成功の鍵は，包括的な訓練済みモデルライブラリを実装することにあると考えている．

本稿では特に転移学習が有する外挿性獲得のメカニズムに注目した．一般に革新的な材料の周辺にはデータが存在しない．しかしながら，機械学習のモデルは訓練データとテストデータの類似性に基づいて予測を行うため，周辺にデータが存在しない外挿領域では予測能力を失う．一方，本稿で示したように，転移学習を巧みに適用することで，極めて少ない訓練データでも時に外挿的といえる予測モデルを構築できる．物理化学的な関連性を持つ元ドメインにおいて，広い物質空間に分布する大量のデータを用いて事前学習を行う．この過程で広い物質

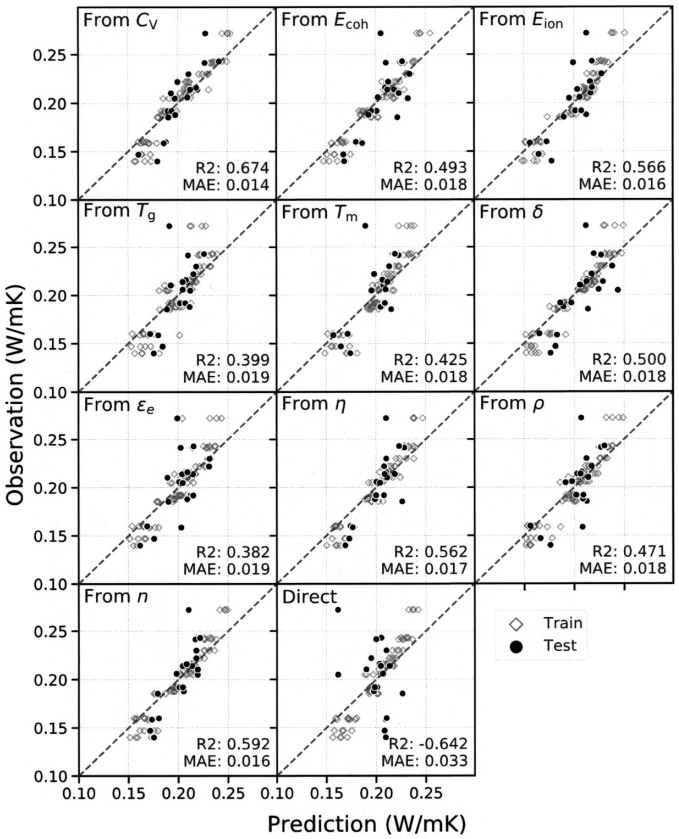


図 7. 様々な元ドメイン(高分子ポリマーの凝集エネルギー E_{coh}，イオン化エネルギー E_{ion}，ガラス転移温度 T_{g}，融点 T_{m}，溶解度パラメーター δ，誘電率 ϵ_e，粘度 η，密度 ρ，屈折率 n ，低分子化合物の定圧比熱 C_{V})から高分子熱伝導率への転移と通常の教師あり学習(Direct)の 5 分割交差検証の結果．横軸は予測値，縦軸は観測値を表す．ダイアモンド(白)とサークル(黒)は，それぞれ訓練データとテストデータを表す．

空間で適用可能な汎用的な特徴表現を獲得できれば，これを目標ドメインに転移することで，データの範囲外での予測能力を有するモデルを構築できる．このことが実験的に観測された．しかしながら，転移学習の外挿性獲得のメカニズムはほぼ未解明である．外挿性の発現には元ドメインの選択や訓練データの分布のパターンが関与していることが示唆されているが，この現象を説明できる理論はまだ確立されていない．

謝　辞

本研究は科研費 18K18017 の助成を受けて遂行されたものです．また，論文をまとめるにあたり，統計数理研究所ものづくりデータ科学研究センターの吉田亮教授と野口瑶特任研究員から有益な助言をいただきました．データの共有については，東京大学の塩見淳一郎教授，東京工業大学の森川淳子教授，物質・材料研究機構の小山幸典主幹研究員から支援を受けました．この場を借りて深く御礼申し上げます．

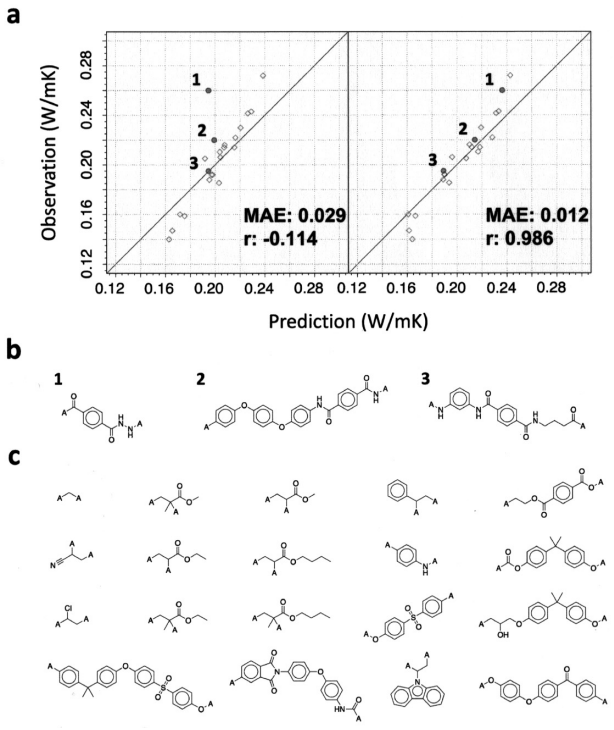


図 8. 転移モデルによる熱伝導率の予測精度．(a)新たに合成した 3 種類の高分子(サークルで表示，番号を付与)に対する通常の教師あり学習のモデル(左)と転移モデル(右)の予測値と実測値．(b)合成した 3 種類の芳香族ポリアミドのモノマー．(c)転移学習に用いた 19 個の訓練データの化学構造．

参 考 文 献

Agrawala, A. and Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science, *APL Materials*, **4**(5), p.053208, DOI: http://dx.doi.org/10/gd7d53.

Carrete, J., Li, W., Mingo, N., Wang, S. and Curtarolo, S. (2014). Finding unprecedentedly low-thermal-conductivity half-heusler semiconductors via high-throughput materials modeling, *Physical Review X*, **4**(1), 011019–011019, DOI: http://dx.doi.org/10/gbfqzz.

Cubuk, E. D., Sendek, A. D. and Reed, E. J. (2019). Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data, *The Journal of Chemical Physics*, **150**, p.214701, DOI: http://dx.doi.org/10/gf3k4k.

Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T. D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M. A., Chae, H. S., Einzinger, M., Ha, D.-G., Wu, T., Markopoulos, G., Jeon, S., Kang, H., Miyazaki, H., Numata, M., Kim, S., Huang, W., Hong, S. I., Baldo, M., Adams, R. P. and Aspuru-Guzik, A. (2016). Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach, *Nature Materials*, **15**(10), 1120–1127, DOI: http://dx.doi.org/10/f859z9.

Hansen, E. C., Pedro, D. J., Wotal, A. C., Gower, N. J., Nelson, J. D., Caron, S. and Weix, D. J. (2016). New ligands for nickel catalysis from diverse pharmaceutical heterocycle libraries, *Nature Chemistry*, **8**(12), 1126–1130, DOI: http://dx.doi.org/10/f9dvx3.

Hutchinson, M. L., Antono, E., Gibbons, B. M., Paradiso, S., Ling, J. and Bryce Meredig (2017). Overcoming data scarcity with transfer learning, arXiv:1711.05099.

Jalem, R., Kanamori, K., Takeuchi, I., Nakayama, M., Yamasaki, H. and Saito, T. (2018). Bayesian-driven first-principles calculations for accelerating exploration of fast ion conductors for rechargeable battery application, *Scientific Reports*, **8**(1), p.5845, DOI: http://dx.doi.org/10/gdfwfv.

Ju, S., Yoshida, R., Liu, C., Wu, S., Hongo, K., Tadano, T. and Shiomi, J. (2021). Exploring diamond-like lattice thermal conductivity crystals via feature-based transfer learning, *Physical Review Materials* (in press).

Kailkhura, B., Gallagher, B., Kim, S., Hiszpanski, A. and Han, T. Y.-J. (2019). Reliable and explainable machine learning methods for accelerated material discovery, arXiv:1901.02717.

Kaya, M. and Hajimirza, S. (2019). Using a novel transfer learning method for designing thin film solar cells with enhanced quantum efficiencies, *Scientific Reports*, **9**, p.5034, DOI: http://dx.doi.org/10/gjw4n7.

Li, X., Zhang, Y., Zhao, H., Burkhart, C., Brinson, L. C. and Chen, W. (2018). A transfer learning approach for microstructure reconstruction and structure-property predictions, *Scientific Reports*, **8**, p.13461.

Liu, C., Fujita, E., Katsura, Y., Inada, Y., Ishikawa, A., Tamura, R., Kimura, K. and Yoshida, R. (2021). Machine learning to predict quasicrystals from chemical compositions (preprint, in review), DOI: http://dx.doi.org/10.21203/rs.3.rs-240290/v1.

Mannodi-Kanakkithodi, A., Chandrasekaran, A., Kim, C., Huan, T. D., Pilania, G., Botu, V. and Ramprasad, R. (2018). Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond, *Materials Today*, **21**(7), 785–796, DOI: http://dx.doi.org/10/gd7q4v.

Matsumoto, R., Hou, Z., Hara, H., Adachi, S., Takeya, H., Irifune, T., Terakura, K. and Takano, Y. (2018). Two pressure-induced superconducting transitions in SnBi2Se4 explored by data-driven materials search: New approach to developing novel functional materials including thermoelectric and superconducting materials, *Applied Physics Express*, **11**(9), 093101–093101, DOI: http://dx.doi.org/10/gjw4nw.

Oda, H., Kiyohara, S., Tsuda, K. and Mizoguchi, T. (2017). Transfer learning to accelerate interface structure searches, *Journal of the Physical Society of Japan*, **86**(12), p.123601, DOI: http://dx.doi.org/10/gjv2v9.

Oliynyk, A. O., Antono, E., Sparks, T. D., Ghadbeigi, L., Gaultois, M. W., Meredig, B. and Mar, A. (2016). High-throughput machine-learning-driven synthesis of Full-Heusler Compounds, *Chemistry of Materials*, **28**(20), 7324–7331, DOI: http://dx.doi.org/10/f88n5s.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. and Yamazaki, M. (2011). PoLyInfo: Polymer database for polymeric materials design, *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 22–29, DOI: http://dx.doi.org/10/fhvjd8.

Ramakrishnan, R., Dral, P. O., Rupp, M. and von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules, *Scientific Data*, **1**(1), p.140022, DOI: http://dx.doi.org/10/gdq9k4.

Ruddigkeit, L., van Deursen, R., Blum, L. C. and Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, *Journal of Chemical Information and Modeling*, **52**(11), 2864–2875, DOI: http://dx.doi.org/10/f4d9mt.

Segler, M. H. S., Kogej, T., Tyrchan, C. and Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks, *ACS Central Science*, **4**(1), 120–131, DOI: http://dx.doi.org/10/gcwpxd.

Seko, A., Togo, A., Hayashi, H., Tsuda, K., Chaput, L. and Tanaka, I. (2015). Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization, *Physical Review Letters*, **115**(20), 205901–205901, DOI: http://dx.doi.org/10/f8d2ww.

Sumita, M., Yang, X., Ishihara, S., Tamura, R. and Tsuda, K. (2018). Hunting for organic molecules with artificial intelligence: Molecules optimized for desired excitation energies, *ACS Central Science*, **4**(9), 1126–1133, DOI: http://dx.doi.org/10/gfcpxs.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using T-SNE, *Journal of Machine Learning Research*, **9**, 2579–2605.

Wu, S., Kondo, Y., aki Kakimoto, M., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., Schick, C., Morikawa, J. and Yoshida, R. (2019). Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm, *npj Computational Materials*, **5**(1), DOI: http://dx.doi.org/10/gf6mkg.

Yamada, H., Liu, C., Wu, S., Koyama, Y., Ju, S., Shiomi, J., Morikawa, J. and Yoshida, R. (2019). Predicting materials properties with little data using shotgun transfer learning, *ACS Central Science*, **5**, 1717–1730, DOI: http://dx.doi.org/10/ggrbd7.

Yonezu, T., Tamura, T., Takeuchi, I. and Karasuyama, M. (2018). Knowledge-transfer-based cost-effective search for interface structures: A case study on Fcc-Al [110] tilt grain boundary, *Physical Review Materials*, **2**(11), p.113802, DOI: http://dx.doi.org/10/gjw4nx.

Application of Transfer Learning in Materials Research

Chang Liu[1], Hironao Yamada[1,2] and Stephen Wu[1,3]

[1]The Institute of Statistical Mathematics
[2]School of Pharmacy, Tokyo University of Pharmacy and Life Sciences
[3]Department of Statistical Science, School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies, SOKENDAI

The digital transformation of materials research has resulted in a broad array of materials property databases; however, the available databases do not include advances realized in machine learning. Transfer learning is a machine learning framework with potential to break the barrier and identify various properties that are physically interrelated. For a given target property to be predicted from a limited supply of training data, models on related proxy properties are pre-trained using enough data to capture the common features relevant to the target task. Repurposing such machine-acquired features for a target task results in an outstanding predictive power even with exceedingly small data. We demonstrate transfer learning in various real-world applications, including property prediction of polymers and inorganic materials. In particular, we show several examples in which transfer learning is applied to obtain a predictive capability in a domain that greatly deviates from the training data distribution.

Key words: Transfer learning, novel material design, crystalline, molecular, polymer.

統計数理(2021)
第 69 巻 第 1 号 65–82

特集「マテリアルズインフォマティクスの最前線」
[総合報告]

高分子インフォマティクスの諸問題

ウ ステファン[1,2]・山田 寛尚[1,3]・林 慶浩[1]・ザメンゴ マッシミリアーノ[4]

(受付 2020 年 10 月 30 日；改訂 2021 年 4 月 4 日；採択 4 月 6 日)

要　　旨

高分子はモノマー分子の設計やプロセス制御により様々な物理化学的特性を発現する．多様な機能を有する高分子の用途は，プラスチックやゴムのような日用品から電子材料や光学材料の部材など，極めて多岐に渡る．このような多機能性により，高分子は現代社会に欠かすことができない材料となっている．高分子インフォマティクスは，高分子科学，コンピュータサイエンス，機械学習の接点から生まれた学際領域である．高分子物性データと機械学習を組み合わせることで，機能性高分子の設計や材料創生のプロセスを加速させることが高分子インフォマティクスに課されたミッションである．近年，高分子材料の研究にデータ駆動型アプローチを導入する事例が増えてきているが，利用可能なデータが少ないことや統一的な構造表現の方法が欠如していること，高分子の構造物性相関が複雑な階層性を有することなど，様々な技術的・社会的課題が顕在化しつつある．本研究では，高分子物性データベース，高分子構造の数値表現，材料特性の予測，設計という四つの観点から，高分子インフォマティクスの現状と展望を論じる．

キーワード：高分子インフォマティクス，機械学習，ハイスループットスクリーニング，逆設計，実験計画法．

1. はじめに

高分子材料は日常生活の様々な場面に利用されている．材料の用途は，レジ袋やペットボトル，電子機器，光学材料，航空宇宙産業の構造部品に至るまで多岐に渡る．高分子(ポリマー)は，繰り返し単位であるモノマー(低分子化合物)が繋がった鎖状あるいは網目状の巨大分子である．ポリマー鎖は多種多様な構造を形成する．構造を制御することで，柔軟な材料から硬く変形しにくい材料を作製できる．このような構造多様性が高分子の物理的・化学的特性に寄与している．高分子には，単一モノマーが連なるもの以外に，2 種類以上のモノマーから構成される共重合体や環状高分子のような特異なトポロジーを形成するものも存在する．現代社会で利用されている高分子材料は．常に高機能化が求められてきた．高分子工学や高分子科学は，新しい高分子の発見を機に，その科学的理解，制御，設計を目的に発展してきた．これまでに数多くの高分子が発見されてきたが，一般に，高分子は天然高分子と合成高分子に分類される．本稿は後者に焦点を当てる．

[1] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3
[2] 総合研究大学院大学 複合科学研究科統計科学専攻：〒 190–8562 東京都立川市緑町 10–3
[3] 東京薬科大学 薬学部：〒 192–0392 東京都八王子市 1432–1
[4] 東京工業大学 物質理工学院：〒 152–8550 東京都目黒区大岡山 2–12–1

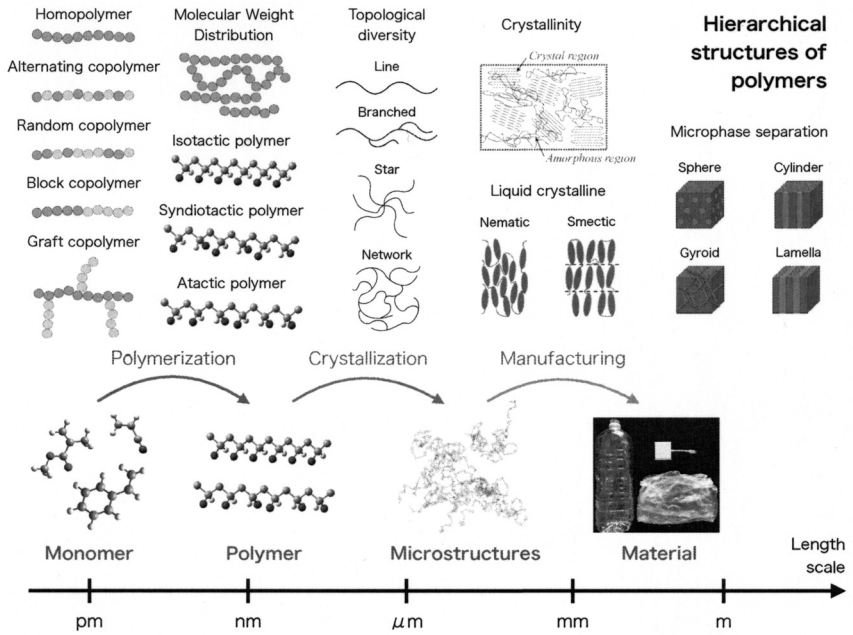


図 1．様々なスケールにおけるポリマーの構造．

高分子の研究は，20 世紀前半に盛んになり，これまでにいくつかの大きなパラダイムシフトを経験してきた．研究の初期は，1953 年にノーベル賞を受賞したシュタウディンガー（Hermann Staudinger）がその発展を大きく牽引してきた（Feldman, 2008）．当時は試行錯誤から導いた経験則を手掛かりに，新しい高分子が発見されてきた．その中で実験データの蓄積も進み，ポリマーの設計指針を与える統計モデルが開発されるようになった．代表的なモデルには，1974 年のノーベル賞受賞者フローリー（Paul John Flory）の研究や原子団寄与法がある（Flory, 1969; van Krevelen and te Nijenhuis, 2009; Bicerano, 2002）．また，ここ数十年，計算機の能力が大きく進歩したことで，物理モデルを用いた計算機実験による特性評価も広く実施されるようになった（Saha and Bhowmick, 2019）．例えば，異なる長さのポリマーの特性を計算機実験で評価できるようになった（Steinhauser and Hiermaier, 2009; Gartner and Jayaraman, 2019）．他の材料系に比べると高分子の実験データの取得は膨大なコストを伴うため，計算機実験による大規模データベースの創生が待ち望まれる．さらに近年は，高分子科学，コンピュータサイエンス，機械学習の融合領域である高分子インフォマティクスの学術創生が活発化している．ビッグデータに基づくデータ駆動型アプローチで材料創生のペースを大幅に加速させるというビジョンが掲げられている．しかしながら，高分子インフォマティクスの実践には，高分子構造の複雑な階層性に起因する様々な課題が立ちはだかっている（Audus and de Pablo, 2017; Kumar et al., 2019a）．

ポリマー設計のプロセスは次の三つのステップから構成される：モノマーの設計（重合），微細組織の設計（結晶化），材料加工（製造）（図 1）．分子のサイズは有機材料の特性に大きな影響を与えるが，ポリマーの「サイズ効果」はモノマーの分子サイズとは直接相関しない．例えば，エチレンは炭素 2 個から構成される炭化水素であり，非常に小さな分子である．これを重合したものが，ポリ袋などに使われるポリエチレンである．ポリエチレン自体は非常に大きな分子になり，モノマーの分子サイズとは関係ない．代わりに，ポリマーの分子量分布（MWD,

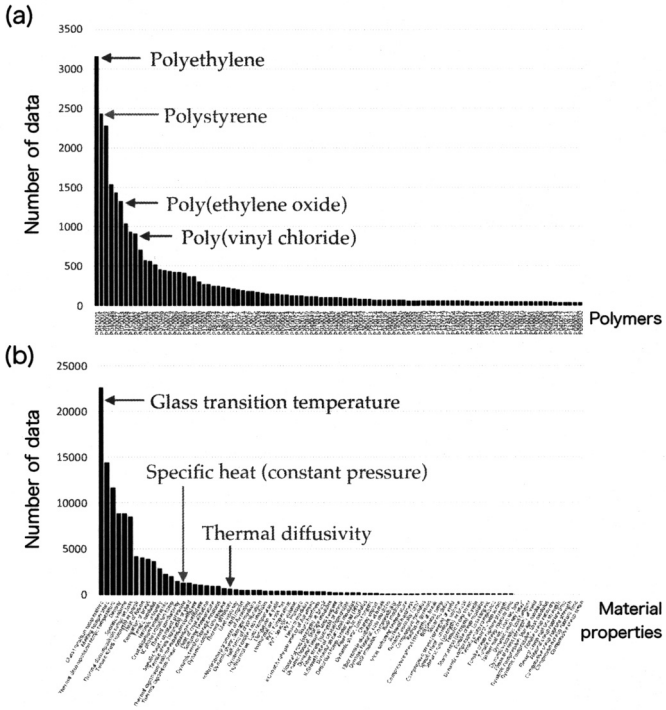


図 2. PoLyInfo の 54,151 件のデータの内訳(2016 年 4 月時点)．(a)データ数が多い上位 100 ポリマーの頻度を降順にプロットしている．(b)83 種類の熱物性のデータ数を降順にプロットしている．

molecular weight distribution)が，分子の大きさと特性を関連付ける．例えば，重合プロセスを制御することで，同一モノマーから異なる分子量分布を持つポリマーを合成できる．分子量分布はポリマー特性の制御パラメータとなる(Imrie et al., 1994; Nunes et al., 1982; Fetters et al., 1994)．また，重合されたポリマー鎖は集合体となり，結晶化のプロセスを経て多様な構造を形成する．最終的に，結晶構造がポリマーの特性に影響を与える．例えば，Yi et al. (2018)は，太陽電池の性能を向上させるために，ポリ(3-ヘキシルチオフェン)分子の結晶性と配向性を制御し，さらに延伸や添加剤の混合などの製造プロセスを経ることで，その特性を向上させた(Pascu and Vasile, 2005)．このように実用材料のポリマー設計のパラメータ空間は複合的である．例えば，ポリマーが単一または複数のモノマーから構成されているか(ホモポリマー，コポリマー)，重合プロセスの温度，添加剤や充填剤の種類，成形方法など，設計空間は異なる階層の異種パラメータから構成される．ただし，実際の研究では，探索空間を絞り込むために一部のパラメータだけに着目し，他のパラメータを無視することが多い．

このような広大な設計空間を対象にデータ駆動型研究を実践するには，量的にも質的にも包括的なデータセットが必要となる．しかしながら，オープンな高分子物性データベースは極めて少ない(Audus and de Pablo, 2017)．また，ポリマーの種類や特性に大きな偏りがあることも多い．例えば，世界最大の高分子物性データベースである PoLyInfo (National Institute for Materials Science, 2011; Otsuka et al., 2011)では，熱特性のデータの約 30% が 10 種類のポリマーから構成されており，その内，40% 以上がガラス転移温度の測定値である(図 2)．高分子材料のデータ駆動型研究には，限られたデータに基づく機械学習の解析技術が必要不可欠にな

る．本稿では，高分子物性データベース，高分子構造の数値表現(記述子)，材料特性の予測，設計という四つの観点から，高分子インフォマティクスの現状と諸問題を論じる．また，我々が開発している Python オープンソースライブラリ XenonPy (Liu et al., 2016)の高分子材料研究への適用事例を解説する．

2. 高分子インフォマティクスにおける機械学習

統計的推測の基本問題は，データ集合 D が与えられたもとで，入力 x から出力 y への写像 f を推定することである．例えば，x はポリマーの構造を数値化したベクトル(記述子)，y は特性である．(f, y, x, D) の設定によって，機械学習の問題設定は次のように分類される．

- **教師あり学習**　D として x と y のサンプル(ラベル付きデータ集合)が与えられ，これを用いて x から y の写像 f を学習する．このような問題設定を教師あり学習という．出力 y が実数の場合を「回帰」，クラスラベルの場合を「判別」という．例えば，モノマーの化学構造を記述子 x で表し，ポリマーのガラス転移温度 y を予測する(Wu et al., 2016; Kim et al., 2018)．
- **教師なし学習**　データ集合 D が x のみのサンプルからなるとき，教師なし学習という問題に帰着する．教師なし学習の代表的な手法は，クラスタリングと次元削減である．例えば，Xu et al. (2019)は，教師なし学習でポリマーの相転移を研究している．クラスタリングは，ラベルなしデータ集合から，出力のクラスラベル y を予測する問題である．ラベルの情報がないので，一般に f の構造やデータ生成過程に強い仮定をおく必要がある．写像 f の推定を目的とせず，単に x の分布特性を調べたり，教師あり学習の補助解析の道具(例えば，特徴抽出)として活用することもある．
- **強化学習**　強化学習では，ある目標を達成するために，対話的な環境から戦略を学習することを目的とする．我々は，現在の状態 s を観測し，行動 a を選択することで，目標達成度に応じた報酬を得る．行動を選択すると環境の状態は確率的に遷移する．したがって，報酬も確率的に決定する．確率的な状態遷移と報酬決定のメカニズムは未知であり，対話的にデータを蓄積しながら学習を進めていく．強化学習では，データ集合から状態価値関数と行動価値関数を推定する．前者の入力は $x = s$，後者は $x = (s, a)$ となる．Li et al. (2018)は，強化学習を適用して適応的に実験を計画することで，ポリマーの MWD を制御することに成功している．

科学的問いを明文化し，問題の背景にある物理化学的知識を整理し，利用可能なデータや計算資源を把握し，機械学習の問題設定として定式化し，適切な記述子と学習アルゴリズムを選択することが，高分子インフォマティクスの研究の本質である．

3. データベース

データ駆動型研究における最も重要な構成要素はデータである．データの質と量によって，データ科学の最高到達点が決まる．一般に機械学習の予測は外挿領域の信頼性が低い．言い換えれば，予測対象と訓練データの類似度が低くなるにつれ，予測精度は低くなる．材料研究の目標は新しい材料の発見であるが，革新的な材料は常に外挿領域に存在する．外挿の実現可能性を高めるには，広大な探索空間を包含する高品質なデータが必要になる．表 1 に高分子物性を含むデータベースの一覧を示す．一覧に示したもの以外にも，高分子に関する大量の出版物や合成技術を集めたデータベース(例えば，NIST Synthetic Polymer MALDI Recipes Database (NIST, 2014))が存在するが，高分子インフォマティクスに適用可能なデジタルデータとして

表 1．高分子データベースの一覧(2020 年 9 月 15 日現在)．

データベース（URL）	概要
PoLyInfo (polymer.nims.go.jp)	国立研究開発法人物質・材料研究機構（NIMS）が提供している学術文献から抽出したデータをまとめた高分子物性データベース（18,044 件の文献データ）．18,015 種類のモノマーから重合されたポリマー群の物性データ 367,711 点を収録している(National Institute for Materials Science, 2011; Otsuka et al., 2011)．
Polymer Genome - Khazana (khazana.gatech.edu)	24 の出版物から抽出した実験データと第一原理計算で算出した物性値を提供しているプラットフォーム．データベースには，1,412 種類のポリマー/有機材料と 2,657 種類の無機材料の特性データが収録されている(Huan et al., 2016; Kim et al., 2018)．
Polymer Property Predictor and Database (pppdb.uchicago.edu)	CHiMaD が提供しているデータベース．文献から抽出した 263 件の Flory-Huggins χ パラメータと 212 件のガラス転移温度のデータを含む．
NanoMine (materialsmine.org)	ポリマーコンポジットの微細組織構造の組成，プロセス，電子顕微鏡データ，物性を含むデータベースならびにデータ共有のためのプラットフォーム(Zhao et al., 2016, 2018)．
Cambridge Structural Database (www.ccdc.cam.ac.uk/structures)	有機・無機材料の結晶構造データベース．100 万以上の構造を収録しており，その内の約 11% が高分子である．
CROW (polymerdatabase.com)	ポリマーの熱物性データを含むデータベース．文献から抽出した実験データや定量的構造活性相関解析から算出した計算物性のデータを含む．
Polymers: A Property Database (poly.chemnetbase.com)	Wiley 出版社の書籍 “*Polymers: A Property Database*” の付録として提供されている高分子物性データ(Ellis and Smith, 2020)．
Citrination (citrination.com)	ポリマーの機械的特性や固体表面エネルギーなど，様々なデータを公開しているマテリアルズインフォマティクスのプラットフォーム．
CAMPUS (campusplastics.com)	ポリマー 9,236 種を含む市販の材料特性データベース．
Identify software (netzsch-thermal-analysis.com)	600 以上の市販ポリマーの示差走査熱量測定線による熱分析のデータを収録した市販ソフトウェアとデータベース．

利用できるものは，ほぼ全て表中に記載されている．

一般的な機械学習の応用分野のデータベースと比較すると，高分子インフォマティクスで利用できるデータベースは量も多様性も著しく小さい．高分子科学の歴史は長く，その中で大量のデータが蓄積されてきたはずだが，ハンドブックや出版物に記録されている歴史的なデータのほとんどはデジタル化されておらず，データの多くは公開もされていない．また，学術コミュニティでデータを共有化しようという取り組みも極めて低調である．これこそが高分子インフォマティクスの発展を阻害している最も大きな要因である(Audus and de Pablo, 2017)．

今後，埋蔵データの整理および学術コミュニティにおけるデータ共有が進み，シミュレーション技術やスーパーコンピュータの演算性能の進歩により計算機実験のデータの蓄積が大幅に進むことで，高品質で大規模なオープンデータが創出されることを期待したい．さらには，ハイスループット実験の技術(Oliver et al., 2019)に人工知能やロボットを組み合わせることで(Burger et al., 2020)，高分子の実験の効率化が進み，そのようなデータのオープン化が進むことを期待したい．

4. 記述子

高分子インフォマティクスにおけるもう一つの重要な構成要素は記述子の選択である．記述子の目的はモデルの入力変数の特徴をコンパクトな形で符号化することである．しかしながら，ポリマーの場合，階層構造が複雑であるため，その符号化は容易ではない(Baer et al., 1987)．ポリマーのユニークな表現の例として，ポリマーマークアップ言語がある．これを利用することで，ポリマーの組成情報から加工パラメータまでの情報を厳密に記述できる(Adams et al., 2008)．この表現は，データベースの構築には適しているが，モデルの入力変数として使用するには扱いづらい．記述子の良さは，材料を一意に表現する能力とタスクに対する学習性能，計算コストのトレードオフから決まる(Zhou et al., 2019)．例えば，インクジェット沈着の制御に有益なポリイミドの探索において，ポリマー鎖の組成や化学構造に加えて，材料組織の微細構造の特徴を表す記述子が重要になる(Hart et al., 2015)．また，特定の相転移特性を持つポリマーを設計するには，原子レベルの構造を識別する記述子が必要になる(Ramprasad et al., 2017)．記述子は現象の背後にある物理的・化学的特性に応じて選択されるべきである．このような階層的な記述子を構成するために，Materials Knowledge System という Python パッケージが開発されている(Brough et al., 2017)．

温度，添加剤や溶媒の選択，膜厚など，ポリマーの製造プロセスに関するパラメータは数値で与えられることが多く，そのようなパラメータを記述子に含めることは容易である．一方，ポリマー鎖の分子骨格や微細組織の構造は固定長の数値ベクトルによる表現方法が自明ではなく，記述子を定めるには工夫が必要である．畳み込みニューラルネットワークを用いれば，電子顕微鏡による材料微細組織の撮像データを入力とし，特性予測モデルを構築できる(Wang et al., 2020)．また，原子配置のような点群形式のデータをパーシステントホモロジーという位相情報記述子で表現する研究も進んでいる(Buchet et al., 2018)．高分子の 1 次構造や高次構造を何らかの方法でグラフの形で表現できれば，グラフデータの正定値カーネルを活用することもできる(Vishwanathan et al., 2010)．ただし，高分子構造を表現する上で高分子鎖の長さの表現が重要になることがある．そのような場合，適切なグラフ表現の定め方は自明ではない．

高分子インフォマティクスで最も広く用いられている記述子は，モノマーの化学構造の表現を対象としたものである．元々は低分子化合物のために開発された物性記述子やフィンガープリントをモノマーの表現に直接適用する．これらの記述子は Python のケモインフォマティクスライブラリ RDKit などを使えば計算できる．通常は，化学構造のグラフ表現(隣接行列など)や線形表記法(SMILES, simplified molecular input line entry system)(Weininger, 1988)による文字列が入力のインタフェースとなっている(Miccio and Schwartz, 2020)．しかしながら，モノマーの化学構造を直接入力すると，モノマー間の連結部の構造情報が無視されてしまう．この問題を解決するために，n 個のモノマーを連結したオリゴマーを入力することが考えられるが，n の選び方に明確な基準がない．n を大きくとれば，一般にポリマーの適切な表現に近づいていくと考えられるが，物理量を算出するような記述子は，n の選択によりその意味が異なってくる．また，分子が大きくなると計算量が大きくなってしまう．分子フィンガープリン

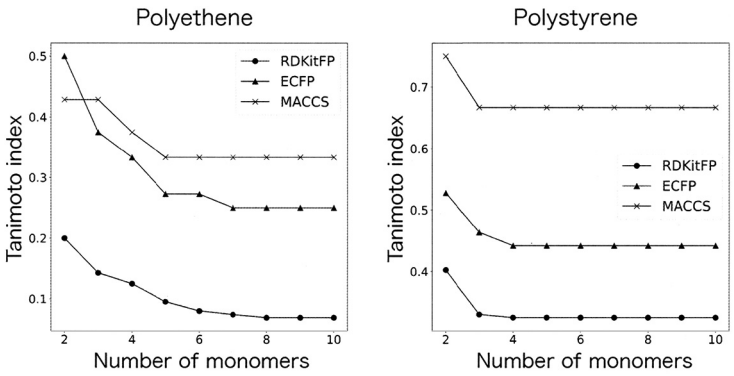


図 3. オリゴマー化におけるモノマー数を変えたときのフィンガープリントの変化．Tanimoto 類似度を用いてオリゴマーとモノマーのフィンガープリントの差を評価した．XenonPy に実装されている三つの RDKit フィンガープリントを使用した．“RDKitFP” は標準フィンガープリント，“ECFP” は Morgan フィンガープリント，“MACCS” は MACCS キーを表す．

トは，化学構造の最も基本的な記述子である．部分構造（フラグメント）の集合に対し，各フラグメントの有無（バイナリ型）や頻度（カウント型）に基づき化学構造のパターンを数値化する．モノマーの連結部の情報をフィンガープリントに反映するためにオリゴマー化を行うと，n の選択によってフラグメントの数が変わる．n を大きくしていくとフラグメントの数は収束していくが，ポリマーやフィンガープリントの種類によって収束の速度が変わるので，適切な n の選択は難しい．図 3 は，異なるフィンガープリントの n に対する収束の振る舞いを示している．Wu et al. (2016) はモノマーが無限に繰り返すポリマー鎖を仮定した記述子の計算アルゴリズムを提案しているが，いくつかのフィンガープリントについては偏りの問題を解決できていない．

ポリマーの記述子には，いくつかの未解決の課題がある．ポリマー鎖は主鎖と側鎖に分けることができる．この二つの成分を区別することは，ポリマーの特性を記述する上で非常に重要である．しかしながら，ポリマーによっては主鎖と側鎖の定義が曖昧であり，自動識別のアルゴリズムを作成することは簡単ではない．もう一つの課題は，コポリマーの記述子である．交互共重合体の場合，複数のモノマーが繰り返し単位となる．これを単に「メタモノマー」と考えればよいが，分子が大きくなるため記述子の計算負荷が大きくなる．また，ブロック共重合体やグラフト共重合体の記述子は確立されていない．

記述子の選択は解きたいタスクやデータの取得コストによって決まる．材料研究における機械学習の主な用途は候補材料のスクリーニングである．通常，候補材料の個数は，多いときで数億のオーダーになることもある．量子化学計算に基づく物性記述子などは大規模スクリーニングの用途には適さない．当然ながら，実験値を含む記述子も使うべきではない．材料探索を目的とする場合，そのような変数は入力ではなく予測対象の出力変数として問題を定式化すべきである．

5. 特性予測

高分子インフォマティクスの中心的な課題は，ポリマーの特性予測である．予測対象の特性は，ガラス転移温度，融点，粘度，熱物性，電気特性，光学的特性など多岐に渡る（Willbourn,

表 2. 様々なポリマー特性に対する機械学習の予測モデル．ΔE は原子化エネルギー，ϵ_{gap} はバンドギャップ，κ は誘電率，ρ は密度，HOMO は最高占有分子軌道，LUMO は最低未占有分子軌道，ϵ_{opt} は光学的ギャップ，η は屈折率，δ は溶解度パラメータ，T_g はガラス転移温度，E_g はガラス弾性率，E_r はゴム弾性率，$\tan\delta_{max}$ は力学的損失正接のピーク．記述子については，Mix は Kim et al. (2018)で用いられた複数の記述子の混合，ICD は無限連鎖記述子(Wu et al., 2016)，Str は Jørgensen et al. (2018)でカスタマイズされた文字列型記述子，D&P は Dragon 記述子(Mauri et al., 2006)と PaDEL 記述子(Yap, 2011)の組み合わせ，Img は 2 次元微細構造画像の直接使用を表す．モデルについては，GP はガウス過程，SVM はサポートベクターマシン，PLS は部分最小二乗回帰，VAE は Jørgensen et al. (2018)で提案された変分オートエンコーダの隠れ層を記述子とする回帰モデル，CNN は畳み込みニューラルネットワークを表す．論文で報告されているモデルの平均二乗誤差(RMSE)，平均絶対誤差(MAE)，決定係数(R^2)を示す．CV-5 は 5 回の交差検証，Split-X は全データセットからテストデータを X% ランダムに分割，Select-27 は 27 個のデータポイントをテストデータとして手動で選択していることを表す．* これらのモデルは，平均絶対パーセント誤差が報告されている．

特性	データ数	記述子	モデル	テスト方法	RMSE	MAE	R^2	Unit
ΔE (Kim et al., 2018)	392	Mix	GP	CV-5	0.01	0.01	0.999	eV/atom
ϵ_{gap} (Wu et al., 2016)	155	ICD	SVM	Split-20	—	—	0.88	eV
ϵ_{gap} (Kim et al., 2018)	382	Mix	GP	CV-5	0.3	0.23	0.971	eV
ϵ_{gap} (Jørgensen et al., 2018)	3,989	Str	VAE	CV-5	—	74	—	meV
κ (Wu et al., 2016)	155	ICD	SVM	Split-20	—	—	0.96	—
κ (Kim et al., 2018)	384	Mix	GP	CV-5	0.48	0.32	0.815	—
ρ (Kim et al., 2018)	173	Mix	GP	CV-5	0.05	0.03	0.938	g/cm^3
HOMO (Jørgensen et al., 2018)	3,989	Str	VAE	CV-5	—	66	—	meV
LUMO (Jørgensen et al., 2018)	3,989	Str	VAE	CV-5	—	43	—	meV
ϵ_{opt} (Jørgensen et al., 2018)	3,989	Str	VAE	CV-5	—	70	—	meV
η (Kim et al., 2018)	384	Mix	GP	CV-5	0.08	0.05	0.892	—
η (Khan et al., 2018)	221	D&P	PLS	Split-30	—	0.004	0.899	—
η (Lightstone et al., 2020)	527	Mix	GP	Select-27	0.05	—	0.88	—
δ (Kim et al., 2018)	113	Mix	GP	CV-5	0.56	0.4	0.955	$MPa^{1/2}$
T_g (Wu et al., 2016)	270	ICD	SVM	Split-20	—	—	0.95	K
T_g (Kim et al., 2018)	451	Mix	GP	CV-5	17.74	12.79	0.944	K
E_g (Wang et al., 2020)	11,000	Img	CNN	Split-15	—	0.68	—	%*
E_r (Wang et al., 2020)	11,000	Img	CNN	Split-15	—	3.12	—	%*
$\tan\delta_{max}$ (Wang et al., 2020)	11,000	Img	CNN	Split-15	—	3.58	—	%*

1976)．

いくつかのポリマー特性については，高分子科学の理論や実験に基づく観察から導かれた経験式が存在する．Python パッケージ thermo には，このような複数のモデルが実装されている(Caleb Bell and Contributors, 2016–2020)．原子団寄与法は，統計モデルに基づくポリマー特性の予測手法であり，かなり古くから研究が進められてきた(van Krevelen and te Nijenhuis, 2009)．分子内の特定の結合原子群(原子団)のポリマー特性への寄与を線形モデルで記述する．原子団の間の相互作用をモデルに組み込むこともある．入力変数に化学構造のフィンガープリント記述子を用いた機械学習モデルは，原子団寄与法の発展形と考えられる．機械学習では，Elastic Net，サポートベクターマシン，ランダムフォレスト，ニューラルネットワークなどのモデルを訓練データから推定する．表 2 に，様々なポリマー特性に対する機械学習の予測モデ

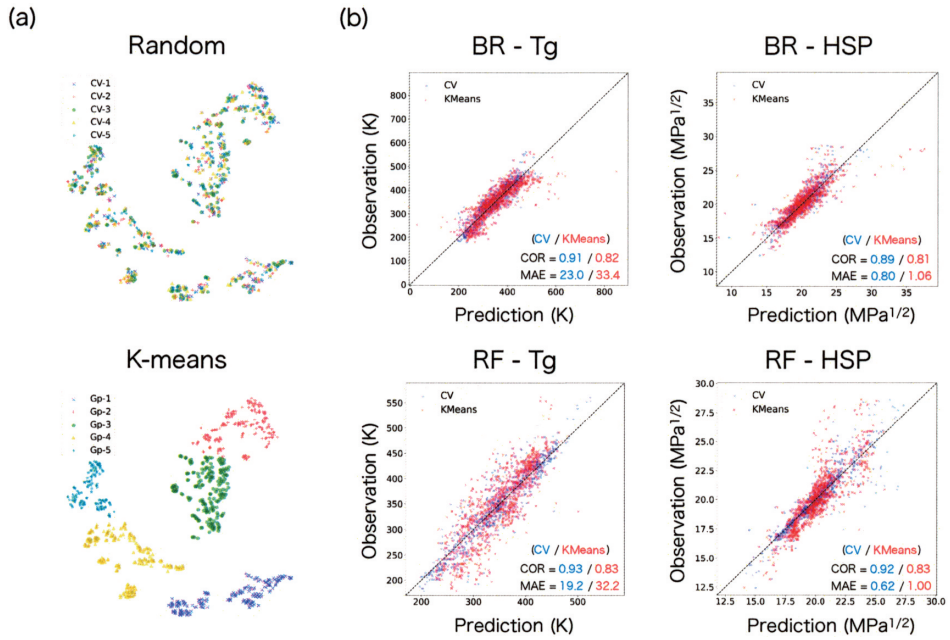


図 4. Polymer Genome (Kim et al., 2018)のデータを用いた機械学習モデルの外挿性能の検証．(a)t-SNE(perplexity=30)を適用して，RDKit の 200 次元記述子ベクトルを 2 次元空間に投影した．上図はこのデータセットをランダムに 5 分割した結果，下図は K-means による 5 分割の結果を表す．ガラス転移温度(Tg)と Hildebrand 溶解度パラメータ(HSP)の交差検証の際，分割したデータを訓練とテストデータセットに分配した．(b)ベイズリッジ回帰(BR)とランダムフォレスト(RF)に基づく Tg と HSP の予測値と観測値のプロット．赤い十字点は，K-means クラスタリング(Lloyd, 1982)に基づく交差検証の結果．青い丸はランダム分割による交差検証の結果．

ルをまとめている．

機械学習の予測は基本的に内挿である．したがって，一般的に訓練データに近い領域でのみ，その予測は有効である．ケモインフォマティクスの分野で研究されてきたモデルの適用領域(applicability domain, AD)という概念は，統計モデルの信頼性の高い領域を同定するために使用される(Sheridan et al., 2004)．統計学における不確かさ(uncertainty)の概念は，予測の妥当性と強く相関する一般的な指標を与える(Chatfield, 1995)．図 4 は，機械学習のモデルの外挿領域における予測性能を示した実験結果である．PoLyInfo からガラス転移温度と Hildebrand 溶解度パラメータのデータを抽出し，ベイズリッジ回帰とランダムフォレストを使って予測モデルを作った．外挿性能を評価するために，二種類の交差検証でモデルの性能を評価した．一つ目は，ランダムにデータを 5 個に分割し，その内の 1 個をテストデータセット，残りの 4 個を訓練データとした．この操作を 5 回繰り返し，モデルの予測性能を 5 個のテストセットの平均予測精度で評価した．二つ目は，訓練データとテストデータの分布が異なるようにデータ分割を行った．t-SNE(t-distributed Stochastic Neighbor Embedding) (van der Maaten and Hinton, 2008)を用いて記述子ベクトルを 2 次元空間に投影し，K-means クラスタリングでデータセットを 5 分割した．これらを訓練とテストセットに分配した．ランダムな分割に比べて，外挿領域の予測精度は大きく劣化することが確認された．この問題の解決法として，転移学習という

手法を導入することが考えられる．転移学習では，あるタスクで学習されたモデルを目標タスクの予測に利活用することで，予測性能の向上を図る．Yamada et al. (2019)は，ポリマーを含む様々な材料科学の問題に転移学習を適用している．転移学習は，グローバルな材料空間からローカルな領域への転移，豊富なデータで学習した特性予測モデルから利用可能なデータが限られた特性予測の問題への転移，計算データから実験データへの転移などに適用できる．転移学習はマルチフィデリティ学習と呼ばれることもあり，ポリマーの結晶化傾向(Venkatram et al., 2020)やバンドギャップ(Patra et al., 2020)の予測にも活用されている．

6. ポリマー設計

近年，機械学習によるポリマー設計の適用事例は増加傾向にあるが，モノマーの設計から製造工程までをエンドツーエンドで設計した事例はまだ報告されていない．そのような中，ポリマー設計の各工程で設計効率を向上させるために，高分子インフォマティクスの解析技術を利用した事例がいくつか存在する．例えば，Wu et al. (2019)では，高い熱伝導率をターゲットにモノマー構造を機械学習で設計し，ポリマー合成ならびに熱伝導率の実験検証を行っている．Li et al. (2018)は，ポリマーの MWD を実験的に制御するために，強化学習で最適戦略を導く手法を開発している．機械学習を用いた設計戦略には，ハイスループットスクリーニング，逆設計，実験計画法の 3 種類がある．本節では，これらの手法を高分子設計に適用した事例を紹介する．

6.1 ハイスループットスクリーニング

ハイスループットスクリーニングは，大量の候補材料から有望な特性を持つ候補を絞り込むことを目的とする．高分子インフォマティクスでは，候補ポリマーのライブラリを構築した上で，特性予測モデルを用いて目標値に達する可能性が高い候補を同定する．探索空間が小さい場合は，全ての候補を検証すればよい．候補ポリマーのライブラリは，データベースから選定するか，あるいは構造生成モデルを用いて仮想ポリマーのライブラリを構築する．例えば，分子のフラグメント集合を定義し，それらの網羅的な組み合わせを考えて仮想ライブラリを作製する．GDB-17 (Ruddigkeit et al., 2012)や PubChem (Kim et al., 2016)のような化合物データベースから，断片化のアルゴリズムを適用してフラグメント集合を得ることができる．また，探索空間をさらに拡張するために，近年は機械学習の生成モデルを用いてライブラリを作製するアプローチもよく見られる．フラグメント法では，構造改変用の部品に既存化合物のフラグメントを使用することで，生成される構造の自由度を制限して探索空間を絞り込む．こうすることで，仮想ライブラリの合成可能性の向上を図る．しかしながら，探索空間の過度な絞り込みは，構造の新規性を低下させるかもしれない．この点を克服するために，主に機械学習の研究者らが有機化学の世界に進出し，従来の発想とは全く異なるアプローチで分子生成の問題に取り組んでいる．Ikebata et al. (2017)は，確率的言語モデル(拡張 n グラム)による構造生成手法を提案している．訓練データ集合に用いる既存化合物の化学構造を SMILES 形式で記述する．この文字列集合を用いて言語モデルを訓練し，既存分子の頻出部分構造や分子骨格のパターンを模倣した構造生成モデルを構築する．Wu et al. (2020)は Ikebata et al. (2017) の言語モデルをポリマーライブラリの生成に活用している．また，有機化合物をグラフや SMILES で表現した上で，グラフ生成や言語生成用のディープニューラルネットワークを用いてライブラリを作製するという研究も実践されている(Cao and Kipf, 2018; You et al., 2018; Popova et al., 2018)．

候補ライブラリを作成した後，特性予測モデルを用いて目標特性を持つ候補をスクリーニン

グする．ハイスループットスクリーニングをポリマーに適用した先行研究は数多くある．最近の例では，高屈折率ポリマーの探索(Khan et al., 2018; Jabeen et al., 2017; Afzal et al., 2019)や共役ポリマーの光電子特性のスクリーニング(Wilbraham et al., 2018)などがある．このようなポリマー設計の戦略では，対象となるケミカルスペースが広大な場合，目標に達する候補を同定するためにかなり大きなライブラリが必要になる．問題に応じて，合成可能性が高く，十分な多様性を有する高品質のライブラリを構築することが重要なポイントになる．

6.2 逆設計

特性予測モデルは入力 x(ポリマー)から出力 y(特性)へのマッピングを定める．これに対し逆設計では，y の目標範囲を x のサブドメインにマッピングする．逆写像を求めるために，遺伝的アルゴリズムのような探索的手法やベイズ推定に基づき y の目標範囲に達する確率が高い x をサンプリングする．いずれのアプローチも基本的に以下のような反復法で問題を解く．

(1)初期候補を選択する．
(2)生成モデルを用いて現在の候補を改変し，新しい候補を提案する．
(3)特性予測モデルを用いて，候補ポリマーの予測特性と目標特性との近さ(尤度)を評価する．
(4)尤度に応じて候補を選抜・更新する．
(5)ステップ(2)に戻る．

ステップ(2)では，ハイスループットスクリーニングと同様に新しい候補を生成するモデルが必要となる．探索空間が非常に大きい場合，ハイスループットスクリーニングで絨毯爆撃的に膨大な数の候補をテストしたとしても，目標特性に近い候補にヒットしないことがある．あるいは，多くの見過ごしが生じる．逆設計では，目標特性に近い入力変数の領域を重点的に探索することで，計算効率の向上を図る．表 3 は，機械学習によるポリマーの逆設計に関するいくつかの成功事例をまとめている．

一般に逆設計は不良設定問題である．したがって，何らかの正則化を施した上で逆問題を解く必要がある．合成可能性が高いポリマーや探索対象の分子骨格のパターンを絞り込むことで，不良設定の問題を緩和する．ただし，そのような正則化を施したとしても，一意性の欠如などの問題は完全には解決しない．逆設計の計算の目的は仮説生成である．目標値に最も近い候補を同定することではなく，目標値の周辺に分布する候補のアンサンブルを得ることが目的である．すなわち，最適化ではなく，数え上げの問題である．モデルには誤りがある．また，逆設計は不良設定問題となっている．したがって，モデルの上で目標値に最も近い候補は，現実において目標値に最も近いとは言えない．不適切な解が含まれていたとしても，多様なアンサンブル(仮説)をマイニングし，多様なシナリオを専門家に提案することを重視する．そこで重要になるのは，広大な探索空間から多様な解を同定できる探索手法である．特に探索空間が高次元の場合，通常の手法は局所的なモードにとらわれてしまうことが多い．この問題に対しては，探索空間の大きさを適切に制限する，探索空間を低次元空間に射影する，アニーリングなどの緩和手法を採用するなどの対策法が考えられる．

6.3 実験計画法

実験計画の目的は，設計目標に達するまでの実験の量を最小にすることである．機械学習のモデルが広い設計空間をカバーするために，不必要な実験をできるだけ減らし，データ生成の効率化を図りたい．実験を行う候補をランダムに選ぶのではなく，再帰的に実験計画を策定していく．すなわち，新しい試験候補を適切に選び，実験結果を既存のデータセットに追加し

表3. 機械学習を用いたポリマーの逆設計の例.

論文	目標物性	手法
Mannodi-Kanakkithodi et al. (2016)	バンドギャップと誘電率	遺伝的アルゴリズムを用いた最適化
Jørgensen et al. (2018)	光学的バンドギャップとLUMO	ディープニューラルネットワークにおける埋め込み空間の勾配に基づく最適化
Pilania et al. (2019)	ガラス転移温度	遺伝的アルゴリズムを用いた最適化
Kumar et al. (2019b)	曇点	粒子群最適化を用いた最適化
Wu et al. (2019)	熱伝導率	逐次モンテカルロ法によるサンプリング
Schadler et al. (2020)	三つの異なる誘電特性	遺伝的アルゴリズムを用いた最適化
Wu et al. (2020)	バンドギャップと誘電率	逐次モンテカルロ法によるサンプリング

て，次のラウンドの実験に移行する．これは化学者が日常的に行っている設計過程そのものである．これをデータ科学の枠組みで定式化し，よりシステマティックに実行する．

実験計画法は，かなり古くから研究されてきたデータ科学の基本問題である．文脈により，クリギング(kriging)，ベイズ最適化(Bayesian optimization)，能動学習(active learning)という呼称が用いられてきた(Brochu et al., 2010)．ここで直観的な説明を示す．特性 y と入力 x に対し，回帰関数 $f(x)$ を有限個のデータ点から推定する．多くの場合，$f(x)$ にはガウス過程モデルが仮定される．まず適当にデータ点を生成し，推定値 $f(x)$ とその分散 $\sigma(x)$ を計算する．基本的には，データ $\{y_i, x_i\}_i$ を順に追加しながら，最小のステップ数で推定値の分散を最小にする問題を考える．直観的には，現時点の分散 $\sigma(x)$ が大きな領域から重点的にデータ点を選択すれば，次のステップの推定精度をより大きく改善できることが期待される(厳密には分散ではなく，分散を元に計算されるある効用関数)．このように推定値の更新と分散が大きい領域からの重点的なサンプリングを繰り返しながら，段階的に推定値を改善していく．厳密な説明ではないが，これが実験計画法に共通するアイデアである．

実験設計に一般的に利用されている手法はベイズ最適化と強化学習である．前者は，予測の不確実性が高い候補を重点的に探索し，それに応じて効用関数を最適化する．後者は，問題を設計目標達成時の報酬を伴うゲームとして扱い，エージェントは最大の報酬(設計目標達成)を得るために最良の戦略(実験の最小化)を学習しようとする．表4に，ポリマー設計における異なる階層の実験計画法の適用例を示す．

実験計画法アルゴリズムは，十分に大きなデータベースが存在しない場合に有望な解決策を与える．しかしながら，ポリマーの実験は時間的コストが大きいため，実験計画のサイクルを何度も繰り返すことは容易ではない．特に，新規ポリマーの合成には膨大な時間を要するため，実験計画のサイクルを回すには，既に合成されたポリマーに候補を限定しなければならない．また，分子動力学シミュレーションのような物理モデルによる計算機実験も他の材料系に比べると計算コストが極めて大きい．特に，系ごとにパラメータ調整を行う必要があり，計算機実験の自動化が難しい．実験計画による高分子研究を実践するには，合成，物性測定，計算機実験の自動化およびハイスループット化の壁を乗り越える必要がある．

表 4. ポリマー設計の実験計画法の適用例.

論文	目標物性	探索空間	手法
Li et al. (2017)	繊維の質，長さ，直径（中央値）	五つの合成プロセスパラメータ	ベイズ最適化
Li et al. (2018)	MWD	5 種類の化学試薬の量	強化学習
Wang et al. (2018)	界面境界の誘電特性と粘弾性	特性モデルのハイパーパラメータ	ベイズ最適化
Minami et al. (2019)	ガラス転移温度	選択された 3 種類のポリマーの混合比	ベイズ最適化
Kim et al. (2019)	ガラス転移温度	データベース内の 736 個の事前定義済み候補	ベイズ最適化

7. おわりに

本稿では，高分子インフォマティクスの関連研究のレビューを行った．ただし，高分子インフォマティクスの現状は依然として黎明期にあり，本稿で取り上げた研究は萌芽的段階にあるものが多い．高分子インフォマティクスが従来の高分子研究を変革する可能性を有することは間違いない．しかしながら，短中期的にはそれは実現しそうにない．十分に大きなデータセットが存在するなら，データ科学の技術は高分子材料の研究開発プロセスを大幅に加速できるかもしれないが，現状は必要条件を満たしていない．

データ駆動型研究における最も重要なリソースはデータである．しかしながら，高分子物性データベースの構築には，他の材料系にはない技術的な難しさがある．その一つは，高分子の材料としての多様性である．高分子はプロセスを制御することで，様々な高次構造を形成する．分子量分布も制御の対象である．また，温度や圧力などを変えると異なる構造を形成する．さらに，実際の材料のほとんどは，単体のポリマーではなく，他の材料との複合材として機能を発現している．このようなプロセス依存性が高分子材料の多様な機能を生み出す源泉となっている．一方，この多様性がデータ駆動型研究にとっては阻害要因となる．設計空間は分子骨格とプロセスの組み合わせから構成される．設計空間があまりにも広大過ぎるため，現在のリソースでは，包括的なデータベースを作ることは難しい．また，個々の研究者が興味を持っている設計空間も多種多様であるため，学術コミュニティ全体でコモンデータを作り出そうという動きも起こりにくい．このことが，オープンデータベースの開発が進まない一因になっていると考えられる．

また，データ科学による予測可能性も高分子の多様性による制限を受ける．化学構造だけでは特性が決まらないため，高次構造を無視したモデルの予測能力には限界がある．一方，高次構造を実験で観測することは容易でないため，モデルの入力変数に含めることは現実的ではない．そもそも実験をしないと計算できないモデルは，材料探索の用途には使えない．本来，高次構造はモデルの入力変数ではなく，予測の対象として定式化されるべきである．しかしながら，系統的に収集されたデータが存在しないため，高次構造の予測可能性に関する研究は全くと言っていいほど進んでいない．このような量的にも質的にも限られたデータセットからデータ科学の解析技術で何をどこまでできるのかを明らかにしていくことも高分子インフォマティクスの学術的課題の一つである．

謝 辞

本研究は科研費 19H01132，18K18017，JST CREST JPMJCR19I3 の助成を受けたものである。論文で示した計算結果には，東京工業大学 TSUBAME3.0，東京大学物性研究所及び自然科学研究機構計算科学研究センターのスーパーコンピュータを使用させていただいたことに深謝する．

参 考 文 献

Adams, N., Winter, J., Murray-Rust, P. and Rzepa, H. S. (2008). Chemical markup, XML and the World-Wide Web. 8. Polymer markup language, *Journal of Chemical Information and Modeling*, **48**(11), 2118–2128.

Afzal, M. A. F., Haghighatlari, M., Ganesh, S. P., Cheng, C. and Hachmann, J. (2019). Accelerated discovery of high-refractive-index polyimides via first-principles molecular modeling, virtual high-throughput screening, and data mining, *The Journal of Physical Chemistry C*, **123**(23), 14610–14618.

Audus, D. J. and de Pablo, J. J. (2017). Polymer informatics: Opportunities and challenges, *ACS Macro Letters*, **6**(10), 1078–1082.

Baer, E., Hiiltner, A. and Keith, H. D. (1987). Hierarchical structure in polymeric materials, *Science*, **235**(4792), 1015–1022.

Bicerano, J. (2002). *Prediction of Polymer Properties*, Marcel Dekker, New York.

Brochu, E., Cora, V. M. and de Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning, *arXiv:1012.2599*.

Brough, D. B., Wheeler, D. and Kalidindi, S. R. (2017). Materials Knowledge Systems in Python — A data science framework for accelerated development of hierarchical materials, *Integrating Materials and Manufacturing Innovation*, **6**(1), 36–53.

Buchet, M., Hiraoka, Y. and Obayashi, I. (2018). *Persistent Homology and Materials Informatics*, 75–95, Springer Singapore, Singapore.

Burger, B., Maffettone, P. M., Gusev, V. V., Aitchison, C. M., Bai, Y., Wang, X., Li, X., Alston, B. M., Li, B., Clowes, R., Rankin, N., Harris, B., Sprick, R. S. and Cooper, A. I. (2020). A mobile robotic chemist, *Nature*, **583**(7815), 237–241.

Caleb Bell and Contributors (2016–2020). thermo: Chemical properties component of Chemical Engineering Design Library (ChEDL), https://github.com/CalebBell/thermo, Last checked: September 25, 2020.

Cao, N. D. and Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs, *ArXiv*, abs/1805.11973.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference, *Journal of the Royal Statistical Society. Series A* (*Statistics in Society*), **158**(3), 419–466.

Ellis, B. and Smith, R. (2020). *Polymers*: *A Property Database*, 2nd ed., CRC Press, Boca Raton.

Feldman, D. (2008). Polymer history, *Designed Monomers and Polymers*, **11**(1), 1–15.

Fetters, L. J., Lohse, D. J., Richter, D., Witten, T. A. and Zirkel, A. (1994). Connection between polymer molecular weight, density, chain dimensions, and melt viscoelastic properties, *Macromolecules*, **27**(17), 4639–4647.

Flory, P. J. (1969). *Statistical Mechanics of Chain Molecules*, John Wiley and Sons, New York.

Gartner, T. E. and Jayaraman, A. (2019). Modeling and simulations of polymers: A roadmap, *Macromolecules*, **52**(3), 755–786.

Hart, L. R., Harries, J. L., Greenland, B. W., Colquhoun, H. M. and Hayes, W. (2015). Molecular

design of a discrete chain-folding polyimide for controlled inkjet deposition of supramolecular polymers, *Polymer Chemistry*, **6**, 7342–7352.

Huan, T. D., Mannodi-Kanakkithodi, A., Kim, C., Sharma, V., Pilania, G. and Ramprasad, R. (2016). A polymer dataset for accelerated property prediction and design, *Scientific Data*, **3**(1), p.160012.

Ikebata, H., Hongo, K., Isomura, T., Maezono, R. and Yoshida, R. (2017). Bayesian molecular design with a chemical language model, *Journal of Computer-Aided Molecular Design*, **31**, 379–391.

Imrie, C. T., Karasz, F. E. and Attard, G. S. (1994). The effect of molecular weight on the thermal properties of polystyrene-based sidechain liquid-crystalline polymers, *Journal of Macromolecular Science — Pure and Applied Chemistry*, **31**(9), 1221–1232.

Jabeen, F., Chen, M., Rasulev, B., Ossowski, M. and Boudjouk, P. (2017). Refractive indices of diverse data set of polymers: A computational QSPR based study, *Computational Materials Science*, **137**, 215–224.

Jørgensen, P. B., Mesta, M., Shil, S., García Lastra, J. M., Jacobsen, K. W., Thygesen, K. S. and Schmidt, M. N. (2018). Machine learning-based screening of complex molecules for polymer solar cells, *The Journal of Chemical Physics*, **148**(24), p.241735.

Khan, P. M., Rasulev, B. and Roy, K. (2018). QSPR modeling of the refractive index for diverse polymers using 2D descriptors, *ACS Omega*, **3**(10), 13374–13386.

Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. and Ramprasad, R. (2018). Polymer genome: A data-powered polymer informatics platform for property predictions, *The Journal of Physical Chemistry C*, **122**(31), 17575–17585.

Kim, C., Chandrasekaran, A., Jha, A. and Ramprasad, R. (2019). Active-learning and materials design: The example of high glass transition temperature polymers, *MRS Communications*, **9**(3), 860–866.

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. (2016). PubChem substance and compound databases, *Nucleic Acids Research*, **44**(D1), D1202–1213.

Kumar, J. N., Li, Q. and Jun, Y. (2019a). Challenges and opportunities of polymer design with machine learning and high throughput experimentation, *MRS Communications*, **9**(2), 537–544.

Kumar, J. N., Li, Q., Tang, K. Y. T., Buonassisi, T., Gonzalez-Oyarce, A. L. and Ye, J. (2019b). Machine learning enables polymer cloud-point engineering via inverse design, *npj Computational Materials*, **5**(1), p.73.

Li, C., Rubín de Celis Leal, D., Rana, S., Gupta, S., Sutti, A., Greenhill, S., Slezak, T., Height, M. and Venkatesh, S. (2017). Rapid Bayesian optimisation for synthesis of short polymer fiber materials, *Scientific Reports*, **7**(1), p.5683.

Li, H., Collins, C. R., Ribelli, T. G., Matyjaszewski, K., Gordon, G. J., Kowalewski, T. and Yaron, D. J. (2018). Tuning the molecular weight distribution from atom transfer radical polymerization using deep reinforcement learning, *Molecular Systems Design & Engineering*, **3**, 496–508.

Lightstone, J. P., Chen, L., Kim, C., Batra, R. and Ramprasad, R. (2020). Refractive index prediction models for polymers using machine learning, *Journal of Applied Physics*, **127**(21), p.215105.

Liu, C., Wu, S. and Yoshida, R. (2016). XenonPy, https://xenonpy.readthedocs.io/en/latest/, Last checked: September 25, 2020.

Lloyd, S. P. (1982). Least squares quantization in PCM, *Information Theory, IEEE Transactions*, **28**(2), 129–137.

Mauri, A., Consonni, V., Pavan, M. and Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations, *MATCH Communications in Mathematical and in Computer Chemistry*, **56**, 237–248.

Miccio, L. A. and Schwartz, G. A. (2020). From chemical structure to quantitative polymer properties prediction through convolutional neural networks, *Polymer*, **193**, p.122341.

Minami, T., Kawata, M., Fujita, T., Murofushi, K., Uchida, H., Omori, K. and Okuno, Y. (2019).

Prediction of repeat unit of optimal polymer by Bayesian optimization, *MRS Advances*, **4**(19), 1125–1130.

National Institute for Materials Science (2011). PoLyInfo, http://polymer.nims.go.jp/index_en.html, Last checked: September 25, 2020.

NIST (2014). Synthetic Polymer MALDI Recipes Database, https://maldi.nist.gov/, Last checked: September 25, 2020.

Nunes, R. W., Martin, J. R. and Johnson, J. F. (1982). Influence of molecular weight and molecular weight distribution on mechanical properties of polymers, *Polymer Engineering & Science*, **22**(4), 205–228.

Oliver, S., Zhao, L., Gormley, A. J., Chapman, R. and Boyer, C. (2019). Living in the fast lane — High throughput controlled/living radical polymerization, *Macromolecules*, **52**(1), 3–23.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. and Yamazaki, M. (2011). PoLyInfo: Polymer database for polymeric materials design, *Proceedings of the 2011 International Conference on Emerging Intelligent Data and Web Technology*, 22–29, IEEE, Tirana, Albania.

Pascu, M. and Vasile, C. (2005). *Practical Guide to Polyethylene*, Smithers Rapra Publishing, Shrewsbury.

Patra, A., Batra, R., Chandrasekaran, A., Kim, C., Huan, T. D. and Ramprasad, R. (2020). A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap, *Computational Materials Science*, **172**, p.109286.

Pilania, G., Iverson, C. N., Lookman, T. and Marrone, B. L. (2019). Machine-learning-based predictive modeling of glass transition temperatures: A Case of polyhydroxyalkanoate homopolymers and copolymers, *Journal of Chemical Information and Modeling*, **59**(12), 5013–5025.

Popova, M., Isayev, O. and Tropsha, A. (2018). Deep reinforcement learning for de novo drug design, *Science Advances*, **4**(7), p.eaap7885.

Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. and Kim, C. (2017). Machine learning in materials informatics: Recent applications and prospects, *npj Computational Materials*, **3**(1), p.54.

Ruddigkeit, L., van Deursen, R., Blum, L. C. and Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, *Journal of Chemical Information and Modeling*, **52**(11), 2864–2875.

Saha, S. and Bhowmick, A. K. (2019). An Insight into molecular structure and properties of flexible amorphous polymers: A molecular dynamics simulation approach, *Journal of Applied Polymer Science*, **136**(18), p.47457.

Schadler, L. S., Chen, W., Brinson, L. C., Sundararaman, R., Gupta, P., Prabhune, P., Iyer, A., Wang, Y. and Shandilya, A. (2020). A perspective on the data-driven design of polymer nanodielectrics, *Journal of Physics D*: *Applied Physics*, **53**(33), p.333001.

Sheridan, R. P., Feuston, B. P., Maiorov, V. N. and Kearsley, S. K. (2004). Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR, *Journal of Chemical Information and Computer Sciences*, **44**(6), 1912–1928.

Steinhauser, M. and Hiermaier, S. (2009). A review of computational methods in materials science: Examples from shock-wave and polymer physics, *International Journal of Molecular Sciences*, **10**(12), 5135–5216.

van der Maaten, L. J. P. and Hinton, G. E. (2008). Visualizing high-dimensional data using t-SNE, *Journal of Machine Learning Research*, **9**(86), 2579–2605.

van Krevelen, D. W. and te Nijenhuis, K. (2009). *Properties of Polymers*: *Their Correlation with Chemical Structure*; *Their Correlation with Chemical Structure*; *Their Numerical Estimation and Prediction from Additive Group Contributions*, 4th ed., Elsevier, Amsterdam.

Venkatram, S., Batra, R., Chen, L., Kim, C., Shelton, M. and Ramprasad, R. (2020). Predicting crystallization tendency of polymers using multifidelity information fusion and machine learning, *The*

Journal of Physical Chemistry B, **124**(28), 6046–6054.

Vishwanathan, S., Schraudolph, N. N., Kondor, R. and Borgwardt, K. M. (2010). Graph kernels, *Journal of Machine Learning Research*, **11**(40), 1201–1242.

Wang, Y., Zhang, Y., Zhao, H., Li, X., Huang, Y., Schadler, L. S., Chen, W. and Brinson, L. C. (2018). Identifying interphase properties in polymer nanocomposites using adaptive optimization, *Composites Science and Technology*, **162**, 146–155.

Wang, Y., Zhang, M., Lin, A., Iyer, A., Prasad, A. S., Li, X., Zhang, Y., Schadler, L. S., Chen, W. and Brinson, L. C. (2020). Mining structure–property relationships in polymer nanocomposites using data driven finite element analysis and multi-task convolutional neural networks, *Molecular Systems Design & Engineering*, **5**, 962–975.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *Journal of Chemical Information and Computer Sciences*, **28**(1), 31–36.

Wilbraham, L., Berardo, E., Turcani, L., Jelfs, K. E. and Zwijnenburg, M. A. (2018). High-throughput screening approach for the optoelectronic properties of conjugated polymers, *Journal of Chemical Information and Modeling*, **58**(12), 2450–2459.

Willbourn, A. H. (1976). Molecular design of polymers, *Polymer*, **17**(11), 965–976.

Wu, K., Sukumar, N., Lanzillo, N. A., Wang, C., "Rampi" Ramprasad, R., Ma, R., Baldwin, A. F., Sotzing, G. and Breneman, C. (2016). Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: Toward optimized dielectric polymeric materials, *Journal of Polymer Science Part B*: *Polymer Physics*, **54**(20), 2082–2091.

Wu, S., Kondo, Y., Kakimoto, M.-a., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., Schick, C., Morikawa, J. and Yoshida, R. (2019). Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm, *npj Computational Materials*, **5**(1), p.66.

Wu, S., Lambard, G., Liu, C., Yamada, H. and Yoshida, R. (2020). iQSPR in XenonPy: A Bayesian molecular design algorithm, *Molecular Informatics*, **39**(1-2), p.1900107.

Xu, X., Wei, Q., Li, H., Wang, Y., Chen, Y. and Jiang, Y. (2019). Recognition of polymer configurations by unsupervised learning, *Physical Review E*, **99**(4), p.043307.

Yamada, H., Liu, C., Wu, S., Koyama, Y., Ju, S., Shiomi, J., Morikawa, J. and Yoshida, R. (2019). Predicting materials properties with little data using shotgun transfer learning, *ACS Central Science*, **5**(10), 1717–1730.

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, *Journal of Computational Chemistry*, **32**(7), 1466–1474.

Yi, A., Chae, S., Hong, S., Lee, H. H. and Kim, H. J. (2018). Manipulating the crystal structure of a conjugated polymer for efficient sequentially processed organic solar cells, *Nanoscale*, **10**, 21052–21061.

You, J., Liu, B., Ying, R., Pande, V. and Leskovec, J. (2018). Graph convolutional policy network for goal-directed molecular graph generation, *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 6412–6422, Curran Associates Red Hook, NY, USA.

Zhao, H., Li, X., Zhang, Y., Schadler, L. S., Chen, W. and Brinson, L. C. (2016). Perspective: NanoMine: A material genome approach for polymer nanocomposites analysis and design, *APL Materials*, **4**(5), p.053204.

Zhao, H., Wang, Y., Lin, A., Hu, B., Yan, R., McCusker, J., Chen, W., McGuinness, D. L., Schadler, L. and Brinson, L. C. (2018). NanoMine schema: An extensible data representation for polymer nanocomposites, *APL Materials*, **6**(11), p.111108.

Zhou, T., Song, Z. and Sundmacher, K. (2019). Big data creates new opportunities for materials research: A review on methods and applications of machine learning for materials design, *Engineering*, **5**(6), 1017–1026.

Challenges in Polymer Informatics

Stephen Wu[1,2], Hironao Yamada[1,3], Yoshihiro Hayashi[1] and Massimiliano Zamengo[4]

[1]The Institute of Statistical Mathematics
[2]Department of Statistical Science, School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies, SOKENDAI
[3]School of Pharmacy, Tokyo University of Pharmacy and Life Sciences
[4]School of Materials and Chemical Technology, Tokyo Institute of Technology

Polymers can exhibit a wide range of functional properties based on different design of monomer and controlling of their manufacturing processes. Their broad applications range from the plastic bags and bottles used in daily life to a variety of electronics, and even structural components in the aerospace industry. Polymer informatics is an interdisciplinary research field of polymer science, computer science, information science and machine learning that serves as a platform to exploit existing polymer data for efficient design of functional polymers. Despite the increasing examples of data-driven approach to polymer design, there has been notable challenges of the development of polymer informatics attributed to the complex hierarchical structures of polymers, such as the lack of open databases and unified structural representation. In this paper, we review and discuss the applications of machine learning on different aspects of the polymer design process through four perspectives: polymer databases, representation (descriptor) of polymers, predictive models for polymer properties, and polymer design strategy.

Key words: Polymer informatics, machine learning, high-throughput screening, inverse design, experimental design.

統計数理(2021)
第 69 巻 第 1 号 83–97

特集「マテリアルズインフォマティクスの最前線」
[総合報告]

反応予測と合成経路設計の機械学習

郭 中梁†

(受付 2020 年 11 月 4 日；改訂 2021 年 5 月 6 日；採択 5 月 7 日)

要　　旨

有機合成において，反応物から生成物を予測することを反応予測という．一方，合成目標の最終生成物から逆方向に反応経路を設計することを合成経路設計という．反応予測と合成経路の自動設計は 50 年以上前から研究されてきたが，近年，有機合成の研究分野に機械学習の先進技術が導入されたことで，モデルの予測精度が著しく向上した．本稿では，2017 年以降に発表された化学反応を対象とする機械学習の解析技術を解説する．特に化学反応の順方向と逆方向予測の問題設定の違いに注目しながら，当該分野の研究動向を概説していく．また，我々のグループが発表したベイズ推論に基づく合成経路設計手法を簡単に紹介する．

キーワード：反応予測，合成経路設計，機械学習，ベイズ推論．

1. はじめに

化合物の合成手法の研究は有機化学の最も重要なテーマの一つである．材料や医薬品の研究で新規の化合物が設計され，望ましい性質や有効性が見込まれる場合，その化合物を合成し，実証する必要がある．そのため，有効な合成経路を設計することが重要な問題となる．近年，機械学習の技術進歩に伴い，2017 年頃から様々な新規分子設計手法が開発されてきた(Ikebata et al., 2017; Jin et al., 2018; Cao and Kipf, 2018; You et al., 2018)．これらの手法は，既存のデータを利用して深層ニューラルネットワークなどのモデルを訓練し，所望の特性を満たす分子を自動設計する．従来の専門家による分子設計に比べると，機械学習の分子設計手法は短時間で大量の候補分子を提案できる．一方でその予測精度は訓練データとモデルに依存する．そこで，大量の候補分子の特性を網羅的に検証する必要が生じることとなり，合成経路の自動設計や合成・計測実験のハイスループット化の実現が喫緊の課題として浮かび上がってきた．本稿では，機械学習を活用した合成経路の自動設計に関する近年の研究動向を解説する．

これまでの合成経路の計画立案は専門家の知識と経験に基づく試行錯誤によって行われてきた．一方で合成計画の自動策定の研究の起源は 50 年以上前に遡る．1969 年に Corey と Wipke は Organic Chemical Simulation of Synthesis という最初のコンピュータ支援合成計画プログラムを発表した(Corey and Wipke, 1969)．その後，LHASA (Pensak and Corey, 1977)，SYNCHEM (Gelernter et al., 1977)，WODCA (Gasteiger et al., 2000)などのプログラムが開発されてきた．初期の合成経路設計システムには，有機化学の知識を手動でコーディングした反応ルールや，データベースから自動抽出された大量の反応テンプレートを用いたルールベース型アルゴリズムが実装されている．また，入力された化合物にどのルールを適用するかを決め

† 統計数理研究所：〒190–8562 東京都立川市緑町 10–3（現 愛知県がんセンター研究所：〒 464–8681 名古屋市千種区鹿子殿 1 番 1 号）

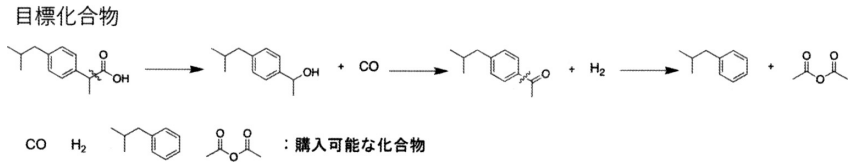


図 1．合成経路設計の例．目標化合物は反応ルールに基づいて前駆体分子に変換される．前駆体分子は，購入可能な化合物に到達するまで変換が繰り返される．

るために，知識ベースのモデル，さらに最近では機械学習モデルが使用されている．選択された化学結合の変換ルールを合成対象の化合物に適用し，合成目標の化合物を前駆体分子に分解する．この変換を購入可能な化合物に到達するまで繰り返し，合成経路を設計する(図 1)．

ルールベース型アルゴリズムの問題点の一つは内挿的な予測しかできないことにある．当然ながら，ルールセットに書き込まれた既存の知識を超えた新規の反応を予測できない．そこで近年，機械学習の解析技術を導入し，より広い反応空間をカバーする合成経路設計アルゴリズムの研究が活発化している．本稿はこの一連の研究動向を概説する．有機合成では，反応物から生成物を予測することを反応予測という．モデルは，反応物を入力として，その生成物を予測する．第 2 節では機械学習を利用した反応予測モデルを紹介する．一方，合成目標の生成物から逆方向に反応経路を探索することを合成経路設計という．合成経路設計では，目標化合物から出発し，合成しやすい前駆体分子への変換を繰り返す．最終的に購入可能な化合物に到達するまでこの操作を反復して合成経路を策定する．これを逆合成解析ということもある．機械学習のモデルは，目標化合物を入力とし，その反応物を予測する．第 3 節では，合成経路設計の機械学習の解析技術をサーベイする．特に，合成経路設計に内在するデータ科学の問題点を整理し，合成経路設計の問題定義を再考する．それを踏まえ，我々が近年発表したベイズ推論に基づく合成経路設計の手法を解説する．第 4 節では，多段階合成経路設計における機械学習を紹介する．第 5 節では，分子設計における反応予測モデルの活用方法を紹介する．第 6 節では，合成反応の反応条件を予測する機械学習の研究を紹介する．さらに，機械学習と自動合成・自動計測の技術を組み合わせた当該分野の最新の取り組みを紹介する．第 7 節はまとめの節である．

2. 化学反応と反応予測モデル

化学反応は一般に反応式 $\mathrm{A}+\mathrm{B} \xrightarrow{\mathrm{C}} \mathrm{D}+\mathrm{E}$ によって表される．これは反応物 A と反応物 B が反応条件 C のもとで，生成物 D と E を生成することを意味する．化学反応は本質的に分子軌道と分子軌道が相互作用し，原子同士の結合が切断されたり，新しい結合が作られる現象である．反応物に存在する官能基のペアに注目することで，有機合成の専門家はある程度の精度で化学反応を予測できるといわれている．ルールベースの反応予測システムは，このような専門家の推論過程を模倣したものである．グラフマッチングなどの手法を用いて，反応物に存在する官能基のペアを見つけ，ルールセットと照らし合わせ，分子内また分子間がどのように相互作用するかを予測する．ルールセット中の複数のルールが適用できる場合，スコアリング関数を利用して各ルールのスコアを計算し，適用すべきルールを決定する．従来のルールベースシステムのスコアリング関数は専門家の経験則から設計されたものであるが，ルールセットのサイズが大きくなるにつれ，適切なルールを選択することが難しくなる．そこで，機械学習を利用して既存の反応データからスコアリング関数を求める手法が提案されている(Segler and Waller, 2017; Coley et al., 2017b)．さらに，ルールセットや既存知識を全く用いずに，膨大な

数のデータから合成反応のパターンを end-to-end で学習する手法が提案されている．後述するように，このようなアプローチでは機械翻訳の深層ニューラルネットワークやグラフ畳み込みニューラルネットワークが用いられる．

2.1 ルールベースの反応予測モデル

ルールベースの反応予測モデルはルールセットとスコアリング関数から構成される．初期の頃は専門家の知識を手作業でまとめたルールセットを利用してきたが，近年では，大規模な反応データからルールセットを自動抽出する手法が確立されている（Coley et al., 2017b）．化学反応の大規模データベースとして，Reaxys や Scifinder が広く利用されている．しかしながら，これらのデータベースは商用のため，アクセスが容易でなく，機械学習系の研究では USPTO（United States Patent and Trademark Office dataset）（Lowe, 2012）というオープンデータセットが最も広く用いられている．このデータセットはアメリカの特許情報から抽出された数百万件の合成反応を収録している．各反応のデータは反応物と生成物の化学構造を含む．また一部の反応には，触媒や溶媒，温度など，反応条件の情報を含む．反応物と生成物，触媒などの化学構造は，SMILES（Simplified Molecular Input Line Entry System）記法（Weininger, 1988）に基づいた文字列で記述されている（図 2）．SMILES は，化学構造を文字列で記述する表記法である．原子を元素記号で表し，特別な文法に従うことで，環構造，分岐，結合次数，同位体，不斉中心などを厳密に記述できる．全ての化学構造は SMILES 形式の文字列に変換できる．

ルールセットの抽出では，まずはじめに反応物と生成物に含まれる原子同士の対応付け（atom-mapping）を行う．化学反応を介して原子と原子の結合が変化するが，反応前後で原子の種類と数は一致する．したがって，反応物の原子と生成物の原子の間には，1 対 1 の対応関係が存在する．Indigo（Pavlov et al., 2011）などのツールは，反応物と生成物を比較し，反応の種類を判断した上で，反応物と生成物の原子間の対応関係を atom-mapping という形式で表記する．ただし，一部の反応では，二つ以上の生成物が存在し，主生成物のみがデータとして記録されており，副産物の情報が欠けているケースがある．すなわち，生成物の全ての原子は反応物に含まれているが，反応物の原子の一部は生成物に含まれない．その場合，既存のツールを使うことで，生成物と反応物の間で対応がとれる原子のみをマッピングできる．

Atom-mapping の結果を用いることで，反応の前後で結合が変化した原子を同定できる．この原子を “反応中心” と定義する．さらに，反応中心の周辺原子が構成する部分構造の変化を “反応ルール” と定義する（図 3）．RDKit（Landrum et al., 2006）などのツールを適用することで，反応ルールは SMARTS という文字列の形式で表現される．SMARTS はデータベースでの部分構造検索を目的に SMILES を拡張した表記法である．反応ルールを構成する部分構造を表すために SMARTS 表記を利用できる．

ルールセットを利用した反応予測の最初のステップは，与えられた反応物に対し，適用可能な反応ルールを選択することである．各反応ルールの部分構造と反応物の化学構造をマッチン

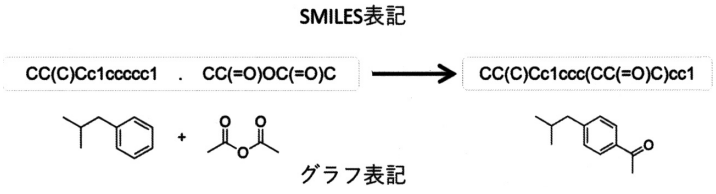


図 2. 化学反応の SMILES 表記とグラフ表記．二つの反応物の SMILES 表記はピリオドで連結される．

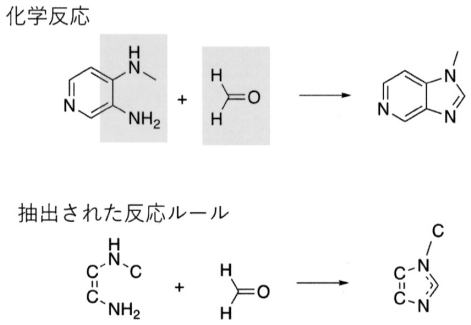


図 3. 反応中心と反応ルールの抽出.

グし，反応中心が反応物に存在するかどうかを判別する．部分構造の検索は SMARTS 表記に基づいて行われる．適用可能なルールが複数存在する場合は，それぞれのルールを適用して複数の生成物を出力する．さらに，入力された反応物と出力された生成物から反応の確からしさをスコアリング関数で評価し，最もスコアが高い反応の生成物を最終出力とする．

近年，機械学習でスコアリング関数をデータから導くアプローチが広く適用されている．Segler and Waller (2017)は教師あり学習で任意の反応物から適用すべき反応ルールのクラスラベルを予測する問題を考えている．Extended-Connectivity Fingerprint (Rogers and Hahn, 2010)という記述子を用いて，反応物の化学構造を表現する．与えられた反応データから教師データを作成し，化学構造の記述子から反応ルールのクラスラベルを出力するモデルを訓練する．論文では，エキスパートシステムとロジスティック回帰，二種類のニューラルネットワークを用いて，最も可能性の高い反応ルールを予測している．反応ルールを適用して導いた候補生成物と真の生成物を比較し，予測精度を比較している(Segler and Waller, 2017)．ルールセットのサイズが 103 種類の場合(103 種類の多クラス分類問題)，エキスパートシステムの予測精度はたったの 0.07 であるのに対し，ロジスティック回帰の精度は 0.86，二種類のニューラルネットワークの精度は共に 0.92 に達することが報告されている．さらに，ルールセットのサイズが 8,720 種類に増えた場合，エキスパートシステムの精度は 0.02 であるのに対し，ロジスティック回帰は 0.41，二種類のニューラルネットワークの精度は 0.78，0.77 に達することが報告されている．専門家が手動で設計したスコアリング関数を利用するエキスパートシステムに比べて，機械学習のスコアリング関数は高い予測性能を示すことが分かっている．Coley et al. (2017b)は，反応中心における結合の形成または切断を数値化した記述子を用いてニューラルネットワークを訓練し，スコアリング関数を構築した．ルールセットのサイズが 1,689 種類の場合，ニューラルネットワークの精度は 0.685 に達すると報告されている．ルールベースの反応予測に機械学習を導入することで，スコアリング関数の精度を向上できる．また，ルールセットのサイズが大きい場合，専門家の知識に基づく選択は困難になる．機械学習を導入することでこの問題を克服し，幅広い反応をカバーできるようになった．

2.2 エンドツーエンドな反応予測

ルールベースの反応予測モデルはルールセットのサイズによって予測できる反応の数が決まる．反応データの数が増加するにつれ，ルールセットのサイズが大きくなり，スコアリング関数の設計が難しくなる．また，ルールセットのサイズが大きくなるにつれ，部分構造検索と生成物予測に要する計算時間が大きくなる．しかしながら，ルールセットのサイズを制限すれば，予測可能な反応の種類が少なくなる．このように，ルールベースの反応予測は拡張性の問

題を内包している．この問題を克服するために，深層ニューラルネットワークを用いた反応予測モデルが提案されることとなった．

Jin et al. (2017)は，分子をラベル付きグラフとみなし，反応予測のタスクをグラフ変換の問題に定式化した．ルールセットを利用する代わりに，グラフニューラルネットワークを用いて，反応物中の反応中心を同定し，生成物を予測する．具体的には，まず反応データを用いて反応中心を予測するグラフニューラルネットワークを訓練する．さらに，予測された反応中心の原子の結合を切断したり，異なる反応中心の原子間を結合することで，生成物を予測する．反応中心の原子が複数存在する場合，確率が高い K 個の原子に対して，結合の形成または切断を行う．これは分子グラフのノード間のエッジを増減する操作に相当する．K が大きくなるにつれて，生成されるグラフの数が組み合わせ数に応じて増加するが，化学的な制約を利用することで存在しえない分子を除外できる．従来のルールベースの反応予測モデルは部分構造検索で反応中心を判別するが，グラフニューラルネットワークに基づく方法では，データから訓練されたモデルで反応中心を予測する．また，グラフニューラルネットワークを利用する方法では，異なるグラフ変換によって異なる生成物が予測されるが，ルールベースの手法と同様にスコアリング関数を利用して，予測された生成物をランキングし，確からしい生成物を絞り込む．グラフニューラルネットワークを利用する手法は，USPTO データセットのベンチマークにおいて，85.6% の精度で生成物を予測できると報告されている．グラフニューラルネットワークを利用することで，ルールセットアプローチの拡張性の問題を解消できる．しかしながら，可能なグラフ変換を列挙する操作に時間が掛かるという問題がある．典型的な計算環境では，1 反応の予測に約 0.5 秒の実行時間を要する．例えば，反応予測モデルを利用して大量の候補分子のバーチャルスクリーニングを実施するには，計算効率の大幅な改善が求められる．

分子の SMILES 表現に基づき，機械翻訳で使われる深層ニューラルネットワークを活用し，反応を予測する研究がグラフニューラルネットワークの研究とほぼ同時期に発表された(Schwaller et al., 2018)．機械翻訳のニューラルネットワークのアーキテクチャはエンコーダとデコーダと呼ばれるモジュールから構成される(Sutskever et al., 2014)．日本語から英語への翻訳の場合，エンコーダは入力として日本語の文章を受け取り，その意味を単語埋め込みベクトルにエンコードする．デコーダはそのベクトルを受け取り，英語の文章にデコードする(図 4)．ニューラルネットワークには日本語の文字列が入力され，翻訳後の英語の文字列が出力される．機械翻訳のニューラルネットワークを利用して，反応物の SMILES 文字列から生成物の SMILES 文字列を予測する．日本語のように単語分割されていない文章の機械翻訳では，単語分割を含む tokenization という前処理が必要になる．SMILES も区切りがない文字列なので，単語分割を行う必要がある．Schwaller et al. (2019)は SMILES 表現の各原子が一つの

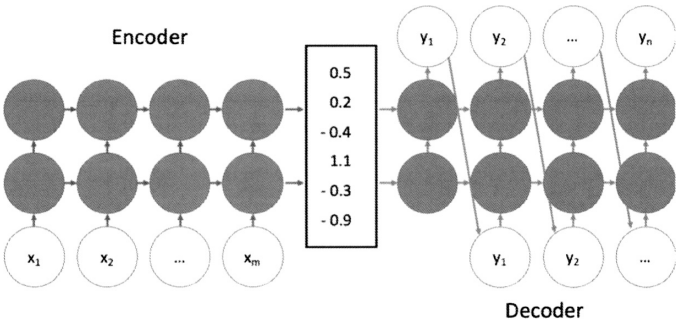


図 4．機械翻訳モデルのアーキテクチャ．

単語となるような tokenization を施した上で，機械翻訳のニューラルネットワークを適用している．USPTO データセットから訓練されたモデルは，90.4% の精度で生成物を予測できることが報告されている．機械翻訳のニューラルネットワークは生成物を直接出力するため，グラフニューラルネットワークに基づく手法のようなグラフの変換や列挙の操作を必要としないため，計算が速い．機械翻訳のニューラルネットワークの一種である Transformer (Vaswani et al., 2017)を利用した場合，1 反応の予測に要する平均実行時間は約 30ms と報告されている(Guo et al., 2020)．機械翻訳のニューラルネットワークの問題点の一つは，SMILES の文法規則に適合しない不適切な文字列を出力することである．Transformer を利用した反応予測モデルの検証では，出力の 0.5% が無効な SMILES という報告がある(Schwaller et al., 2019)．また，反応中心や反応進行のメカニズムがブラックボックス化されているため，ルールベースのモデルやグラフニューラルネットワークを利用した手法に比べると，予測結果の解釈性が低いという問題もある．

3. 合成経路の自動設計

合成経路の設計では，合成対象を単純な構造の前駆体分子に順に変換していきながら，前駆体分子が全て購入可能な分子になった時点で工程を完了する．途中で購入できない分子が現れた場合，さらに単純な前駆体分子に変換していく．このタスクのことを逆合成解析と呼ばれる．ルールベースの合成経路設計ソフトウェアは，ルールセットから合成対象の化合物に適用可能なルールを選び出し，前駆体分子となる反応物を推定する．適用可能なルールが複数存在する場合，反応予測モデルと同様にスコアリング関数によって確からしいルールを選択する．順方向の反応予測と同様に，ここでも既存の知識に基づいて設計されたスコアリング関数から機械学習ベースのアプローチへの展開が進行している．さらに，設計のタスクを回帰問題に帰着させた上で，入力変数である目的化合物から出力変数の前駆体分子を直接予測するアプローチも検討されている．ここで注意すべき点は，一つの化合物を合成する合成経路が複数存在するという事実である．したがって，一つの入力信号に対して複数の出力パターンが存在する系を取り扱うことになる．利用可能な試薬や商用化合物のリスト，実験環境などに応じて，合成科学者が合成経路の最終的な選択を行う．その場合，機械学習に求められるのは，合成目標に到達しうる複数の多様な候補経路を提示し，合成科学者の創造力を刺激することである．

3.1 ルールベースの合成経路設計

ルールベースの合成経路設計は，ルールベースの反応予測と同様のルールセットを利用する．データから抽出されたルールは，反応前後で結合などが変化する原子とその周辺構造の情報を与える．順方向の反応予測では入力反応物とルール中の反応中心の構造を照らし合わせ，適用可能なルールを見つけ出す．一方，合成経路設計では，生成物と反応後の反応中心の構造を照合し，適用可能なルールを生成物に適用して反応物を推測する．適用可能なルールが複数存在する場合には，スコアリング関数を利用してランキングを行う．Coley et al. (2017a)は，データベースの反応との構造類似度に基づくスコアリング関数を提案している．目的化合物と反応データ中の生成物の Tanimoto 類似度(Tanimoto, 1958)を計算する．さらに，ルールを適用して導いた予測反応物と反応データ内の反応物の Tanimoto 類似度を計算する．これらの類似度を用いてスコアを算出する．非常にシンプルなアプローチである．論文では，1 ステップの反応における生成物から反応物の予測精度は 37.3% と報告されている．Dai et al. (2019)はグラフニューラルネットワークに基づくスコアリング関数を提案している．適用可能なルールに対して，目的化合物，反応ルール，予測される反応物のセットをニューラルネットワークの入力と

してスコアを計算し，スコアの高いルールを選択する．このスコアリング関数に基づく逆合成予測の精度は 52.5% に達すると報告されている．この二つの論文の結果は，いずれも USPTO データセットから 5 万個の反応を取り出し（本項では USPTO-50k と略），その内の 80% を訓練セット，10% が検証用セット，10% がテストセットとしている．しかしながら，USPTO には 100 万以上の反応データが記録されており，データセットのサイズを大きくとると，ルールセットのサイズも増加する．その場合は順方向の反応予測と同様に，スコアリング関数の設計が難しくなり，さらに，グラフ変換や数え上げの操作に要する計算コストが問題になる．

3.2 エンドツーエンドな逆合成解析

逆合成解析は目的化合物から出発し，単純な分子に変換していくことで，最終的に入手可能な分子によって目的化合物を合成する合成経路を設計する手法である．購入できない分子から単純な前駆体分子への変換は生成物から反応物を予測するタスクとみなすことができる．以降，反応物から生成物を予測することは反応の進行方向と同じ方向で順方向の予測とし，生成物から反応物を予測することは逆方向の予測とする．順方向の反応予測で利用される深層ニューラルネットワークモデルは逆方向の逆合成解析でも利用されている．特に機械翻訳のモデルと SMILES を組み合わせたエンドツーエンドな逆合成解析手法が数多く提案されている．Liu et al.（2017）は，seq2seq という機械翻訳の深層ニューラルネットワークを利用して，反応物の SMILES から生成物の SMILES への変換を直接予測することを試みた．USPTO-50k データセットでは，seq2seq は 37.4% の予測精度を達成した．ただし，入力変数には反応物に加えて適用する反応のタイプ（ルール）も含まれているため，実際の適用ケースに比べると問題設定が簡単化されている．また，seq2seq が出力する反応物の SMILES 文字列の内，10% 以上が SMILES の文法規則に適合しないことが報告されている．これは機械翻訳モデルを適用する際の特有の問題である．そこで Zheng et al.（2020）では，seq2seq の代わりに Transformer を利用することで，USPTO-50k データセットでの予測精度は 43.3% に向上することを示した．また，文法規則に適合しない SMILES 文字列を修正する機能を導入し，さらなる予測精度の向上を図ったが，Transformer はほとんど無効な SMILES 文字列を生成しないため，最終的に予測精度は 43.7% に留まった．Lin et al.（2020）は異なる tokenization の仕方を利用した時の予測精度を調べた．例えば，ある分子の SMILES 表現「Nc1nc2[nH]c(CCCc3csc(C(=O)O)c3)cc2c(=O)[nH]1」は Schwaller et al.（2018）が提案した tokenization 処理によって「N c 1 n c 2 [n H] c (C C C c 3 c s c (C (= O) O) c 3) c c 2 c (= O) [n H] 1」になるが，Lin et al.（2020）は「[*]」を一つの単語として扱うべきと考え，「N c 1 n c 2 [nH] c (C C C c 3 c s c (C (= O) O) c 3) c c 2 c (= O) [nH] 1」になる tokenization 処理を提案した．しかしながら，Schwaller et al.（2018）が提案した tokenization 処理に基づいて訓練された Transformer は予測精度が高く，生成された SMILES の中に無効なものが少ない．

3.3 ベイズ推論を利用した合成経路設計

逆方向の予測モデルから生成された候補反応物は，多くの場合，購入可能な化合物ではない．したがって，購入可能な化合物に達するまで，前駆体分子への変換を繰り返す必要がある．例えば，逆方向のモデルによって目的化合物が A と B に分解された場合，両方の反応物は一般的に購入不可能な化合物となる．その場合，さらに A と B 両方の合成経路を特定しなければならない．したがって，合成経路のパス数が大きくなる傾向にあるため，その結果，実験コストが高くなり，失敗の確率も高くなる．そこで我々は，反応物の一方または両方を購入可能なものに制限した上で合成経路を設計するためのベイズ推論の枠組みを構築した（Guo et al., 2020）．

合成経路設計のもう一つの難しさは，問題の不良設定性から生じる．不良設定問題になる理

表 1. 代表的な順方向の反応予測モデルと逆方向の逆合成解析モデルの性能比較．template-based モデルは 2.1 節で紹介したルールベースの反応予測モデルの一例である．WLDN (Weisfeiler-Lehman Difference Network) と modified WLDN は 2.2 節で紹介したグラフニューラルネットワークに基づく反応予測モデルである．Molecular Transformer は 2.2 節で紹介した機械翻訳モデルを利用した予測モデルである．また similarity モデルは 3.1 節で紹介したルールベースの合成経路設計手法の一例である．SCROP (self-corrected retrosynthesis predictor) と Lin et al. 2020 は 3.2 節で紹介した機械翻訳ベースの合成経路予測モデルである．GLN (conditional graph logic network) は 3.1 節で紹介したグラフニューラルネットワークをスコアリング関数に利用したルールベースの合成経路設計手法である．

Task	Model	Top-1	Top-3	Top-5	Top-10
順方向の予測モデル	template-based (Coley et al., 2017b)	71.8	86.7	90.8	94.6
	WLDN (Jin et al., 2017)	79.6	87.7	89.2	-
	modified WLDN (Coley et al., 2019a)	85.6	90.5	92.8	93.4
	Molecular Transformer (Schwaller et al., 2019)	90.4	94.6	95.3	-
逆方向の予測モデル	similarity (Coley et al., 2017a)	37.3	54.7	63.3	74.1
	SCROP (Zheng et al., 2020)	43.7	60.0	65.2	68.7
	Lin et al. 2020 (Lin et al., 2020)	43.1	64.6	71.8	78.7
	GLN (Dai et al., 2019)	52.5	69.0	75.6	83.7

由の一つは，反応の副生成物がしばしばデータセットに記録されていないことによる．例えば，反応 A + B → C + D において，主生成物 C のみがデータに記載されており，副生成物 D を省略されていることが多い，すなわち，A + B → C だけがデータとして観測される．当然ながら，A と B には欠落構造である D が含まれているので，C のみから A と B を予測することは一般には不可能である．表 1 にこれまでに紹介した順方向の反応予測モデルと逆方向のモデルの予測精度をまとめた．逆方向のモデルの予測精度 (top 1) は 37% から 52% までの範囲にあり，順方向の予測に比べると精度は圧倒的に低い．前述のデータの欠損による不良設定問題が一因となっている．

一方，順方向の反応予測は，逆方向予測に比べるとかなり高い精度を達成している．Transformer の予測精度は 90.4% に達しており，これは逆方向予測の約 2 倍の精度となっている．逆方向の生成物から反応物へのマッピングは本質的に一対多である．すなわち，同じ生成物には複数の合成経路が存在するため，逆方向の予測精度が低下することは避けられない．したがって，逆方向の予測精度は順方向の予測精度より低くなることは当然である．一方，反応物が決まれば，一つの生成物が決まる．ここで，反応における順方向と逆方向の予測は目的が異なるため，精度を単純に比較することには意味がないことに留意しなくてはならない．逆合成解析の目的は，目的化合物の多様な合成経路を提案することである．しかしながら，逆方向の予測精度が高いほど，提示される予測反応物はデータに記述されているものに偏り，結果として，候補経路の多様性が制限されてしまう．

ここでは例として，以下の 2 ステップの合成反応を考える．

$$S_1 + S_2 \to X + S_3 \to Y \tag{3.1}$$

第 1 ステップでは，二つの反応物 S_1 と S_2 が中間生成物 X を合成する．これに反応物 S_3 を

あたえ，最終生成物 Y を合成する．合成経路設計の目的は，標的分子 Y に到達可能な反応物 $S = (S_1, S_2, S_3)$ の組を同定することである．

ベイズ推論に基づく合成経路設計は順方向の反応予測モデルの構築と逆方向の予測の二つのステップから構成される(Guo et al., 2020)．順方向の反応予測モデルは，反応物の組み合わせ S からその生成物 Y への写像 $Y = f(S)$ を定める．そのの逆写像 $S = f^{-1}(Y^*)$ を求めることで，目的化合物 Y^* に到達する反応物の組み合わせ S を同定する．反応物は商用化合物のリストから選択される．通常，$O(10^6)$ 個ほどの商用化合物を取り扱う．したがって，経路上の反応物の個数を p とすれば，問題は $O(10^{6\times p})$ の候補経路から構成される探索空間 $\mathcal{T}$ 上の組み合わせ最適化に帰着する．解空間の複雑度は，反応のステップ数に応じて指数関数的に増加する．

Guo et al. (2020)では，合成経路設計の問題をベイズ推論に帰着させるために，事後分布 $p(S|Y = Y^*)$ を以下のようにモデリングした．

$$p(S|Y = Y^*) \propto p(S, Y = Y^*) = \frac{1}{Z} \exp\left(-\frac{E(Y^*, f(S))}{T}\right). \tag{3.2}$$

ギブズ分布のエネルギー E は，標的生成物 Y^* のフィンガープリント記述子と順方向モデルの予測生成物の間の非類似度(ユークリッド距離など)を表す．温度パラメータ T は，候補反応物の多様性を制御するハイパーパラメータである．事後分布は，商用化合物の組み合わせの上に定義される．この確率分布は厳密に計算できないので，Guo et al. (2020)では，近似分布を導くために逐次型モンテカルロ計算のアルゴリズムを開発した．

論文では，USPTO のデータを用いて包括的な数値実験を実施し，既知の合成経路に対する予測性能や提案された経路の合成可能性を検証している．順方向のモデルの予測精度が約 87% のとき，既知の反応物を 47.5% の精度で同定できることを報告している．また複数の反応経路が存在することを想定して設計された逐次モンテカルロ法を適用することで，目標化合物に対する多様な反応経路を同定できる可能性を示している．1 ステップの反応経路の設計では，提案手法が一つの目標化合物に対して平均で約 500 個の候補経路を検出したことを報告している．さらに，有機合成の知見に基づき候補経路の合成可能性の評価を実施し，35-50% の候補経路が化学的に妥当であると結論付けた．

4. 多段階の合成経路設計

合成経路設計では，一つの化合物に対して，複数の変換が適応可能で，変換が繰り返して適用されると，合成経路の探索木が構築される．多段階の合成経路設計では，合成経路の探索木を構築しながら，効率的に購入可能な化合物からなる合成経路を見つけなければならない．Segler et al. (2018)は深層ニューラルネットワークとモンテカルロ木探索(MCTS: Monte Carlo tree search)を組み合わせた手法を提案した．最良優先探索などの木探索アルゴリズムでは，探索木のノードの優先度を決めるためにあらかじめスコア関数を設計する必要がある．しかしながら，合成経路設計の問題においては，スコア関数の設計方法は自明ではない．例えば，一般により単純な前駆体分子がより良い選択とされるが，反応性の高い官能基を保護するために複雑な前駆体分子を経由する合成経路が望ましいこともある．この問題を回避するために，Segler et al. (2018)は MCTS と三種類のニューラルネットワークを組み合わせた合成経路設計手法を提案した．探索木のノードは反応を実行する化合物の集合，エッジは反応に相当する．ルートノードは目標化合物である．現在選択されているノードに含まれる化合物は，expansion policy network というニューラルネットワークによって前駆体化合物に変換される．これを新たな子ノードとする．既存の反応データから学習された expansion policy network は，生成物を入力とし，その反応物を出力する逆向きの反応予測モデルである．また，expansion policy

network が提案した反応の妥当性を判定するために，in-scope filter network というモデルが用いられ，反応が進行しないと予想されるノードは除外される．新たに付け加えられた子ノードに rollout policy network というモデルを適用し，ノードに含まれる購入不可能な化合物が前駆体分子に変換される．この変換を何度も繰り返し，生成される前駆体分子が購入可能な化合物のリストにヒットするか，あるいはあらかじめ設定された深度に達した段階で伸長を停止する．その結果に基づいて，追加された子ノードのスコアが計算される．MCTS の各ステップでは，スコアが最も高い子ノードが選択され，上記 3 つのニューラルネットワークを繰り返して適用し，探索木を構築していく．ニューラルネットワークの訓練には，Reaxys というデータベースに記録されている 1,240 万件の 1 ステップの反応データを用いている．論文で示された二重盲検法では，45 名の有機化学者に既存の反応経路と機械学習が予測した経路を選択させた結果，両者の間に有意な差がみられなかった．

Mikulak-Klucznik et al. (2020)は，複雑な天然物の合成経路を設計するために，Chematica というシステムを拡張し，エキスパートシステムと機械学習のハイブリッド型システムを開発した．純粋な機械学習のアプローチとは異なり，Chematica は 100,000 以上のハンドコーティングされた反応ルールを用いる(Szymkuć et al., 2016)．天然物の合成経路を計画するには，立体化学的な制御が必要であるが，これは機械学習的なアプローチでは困難である．そこで，Mikulak-Klucznik et al. (2020)は，機械学習と量子化学計算を利用してハンドコーディングされたテンプレートの適用可能性を評価し，反応の位置選択性を検討した．反応空間の探索はビームサーチとスコア関数の組み合わせによって行われる．各探索深度においてスコアの最も良いノードが所定の数だけ保持し，探索が実行される．スコア関数によって，保持されるノードが異なるため，探索アルゴリズムの特性はスコア関数に大きく左右される．複数のスコア関数を利用したい場合，Mikulak-Klucznik et al. (2020)は複数のキューを用意し，各キューにスコア関数を決め，ビームサーチを行った．例えば，一つのキューは幅優先で探索し，もう一つのキューは深さ優先で探索することで，一つ目のキューが発見した有望な合成経路の開始点を二つ目のキューで素早く完了させることができる．このアルゴリズムでは，計算時間のほとんどが探索ではなく，立体選択性の評価やスコア関数によるノード評価に費やされる．有機化学者を対象に実施されたチューリングテストでは，このハイブリッドシステムが設計した合成経路は，人間が設計したものとほとんど見分けがつかなかった．また提案された三つの天然物の合成法が実験室で有効であることが実証された．

5. 反応予測モデルを利用した分子設計

反応予測の機械学習モデルは新規分子の設計にも利用されている．分子設計の目的は，所望の特性を有する化学構造を同定することである．Gottipati et al. (2020)は，強化学習を利用して合成経路と化学構造を同時に設計する手法を提案した．初期分子の候補集合を用意し，反応予測の機械学習モデルを適用して，一連の合成反応の生成物を計算する．さらに，別途用意したモデルを用いて最終生成物の物理化学的特性を評価する．問題は所望の特性を有する初期分子の同定である．このタスクを解くために，強化学習の actor-critic 法が用いられた(Sutton and Barto, 2018)．actor は各ステップの反応を決定するモデルである．モデルは二つのサブアクションを規定するニューラルネットワーク f と π から構成される．f はステップ t の生成物 R_t から最適な反応テンプレート T_t を予測する．π は R_t と T_t からアクション a_t を求める．critic であるニューラルネットワーク Q は，状態とアクションがもたらす将来の報酬(Q 値)を予測する．環境は現在の分子 R_t と予測されたテンプレート T_t，アクション a_t を入力とし，報酬 r_t と次の状態の生成物 R_{t+1} を計算し，さらにエピソードの終了判定を行う．ここで

のアクションは，現在の分子と購入可能な分子の反応を予測することになる．事前に定義された購入可能な分子のリストから予測されたアクションに最も近い k 個の反応物を選択し，反応予測モデルを用いて k 個の生成物を予測する．各生成物の報酬は所望の特性との近さを表すスコア関数で計算される．最大報酬に対応する生成物が R_{t+1} として出力する．反応ステップ数が最大数に達したとき，あるいは有効なテンプレートがない時点で，エピソードは終了する．Gottipati et al.（2020），この強化学習のアプローチを利用し，分子設計のいくつかのベンチマーク問題において最高性能に達することを示した．

Bradshaw et al.（2019）は深層生成モデルと反応予測モデルを組み合わせた分子設計の手法を提案した．ここでの生成モデルは反応物のペアを生成する．反応予測モデルは反応物から生成物を予測することで新規分子を提案する．生成モデルには Wasserstein Auto-Encoder（Tolstikhin et al., 2018）が用いられる．Wasserstein Auto-Encoder は Variational Auto-Encoder（Kingma and Welling, 2014）と同様にエンコーダとデコーダから構成される．エンコーダは購入可能な反応物のペアを潜在空間にエンコードし，デコーダは潜在空間の特徴ベクトルから元の反応物のペアを復元する．モデルの訓練に使用される損失関数は次のように定義される．

$$L = E_{\mathbf{x}\sim\mathcal{D}} E_{q(\mathbf{z}|\mathbf{x})}[c(\mathbf{x}, p(\mathbf{x}|\mathbf{z}))] + \lambda D(E_{\mathbf{x}\sim\mathcal{D}}[q(\mathbf{z}|\mathbf{x})], p(\mathbf{z})) \tag{5.1}$$

$\mathbf{x}, \mathbf{z}$ はそれぞれ反応物の特徴ベクトルと潜在空間にエンコードされた特徴ベクトルを表す．$q(\mathbf{z}|\mathbf{x})$ はエンコーダで，$p(\mathbf{x}|\mathbf{z})$ はデコーダである．コスト関数 c は復号された反応物のペアがエンコードされた反応物のペアと類似するように強制するためのもので，ダイバージェンス D はデータのエンコーディングにおける潜在変数 z の周辺分布が潜在空間における事前分布 $p(z)$ と一致するように強制するための項である．$\mathcal{D}$ はデータの経験分布で，λ は損失関数の第一項と第二項の相対的な重要度を決めるチューニングパラメータである．反応予測モデルには Molecular Transformer が利用されている．さらに目標特性に達する分子を設計するために，潜在空間の表現を記述子とし，最終生成物の特性を予測するニューラルネットワークを構築する．潜在空間の勾配情報を利用して目標特性に到達する分子の潜在変数を求め，デコーダを用いて潜在変数から反応物に変換する．最後に，反応予測モデルを用いて反応物から生成物を予測する．

本節で述べた手法は，提案された合成経路の化学的な妥当性の検証が不足している．一般に反応予測モデルの訓練に用いられる反応データセットには反応性の高いデータのみが記載されており，進行しない反応のデータは存在しない．したがって，反応予測のモデルは，入力化合物の反応が進行するという条件のもとでの生成物を予測する．そのため，任意の入力化合物の組を与えたときに反応の有無を判定するもう一つのモデルが必要になる．しかしながら，反応に失敗した事例を包括的に収集したデータセットが存在しないため，現時点においてはモデルの構築は難しい．化学反応の広大な空間の中，ほとんどの反応物の組は反応しない．実験検証なしに，不完全なモデルから予測された合成経路の妥当性を評価することは難しい．機械学習の研究者と有機化学の研究者の連携は，この分野において真に実用的な手法を開発する上で必要不可欠である．

6. 反応条件の予測と自動合成実験

ここまでの議論では主に反応物と生成物のみを考えてきたが，合成反応の設計変数には触媒や溶媒の選択並びに温度・圧力などの反応条件が含まれる．Gao et al.（2018）は，生成物と反応物が与えられたもとで，触媒や溶媒などの反応条件の予測を行った．モデルは一般的な多クラスの分類問題にニューラルネットワークを適用したものである．モデルの入力は反応物と生

成物のフィンガープリント記述子，出力は利用可能な触媒や溶媒の選択確率となっている．このような単純なモデルでランキングした Top 10 の反応条件の内，69.6% の精度で既知の反応条件が含まれる．また，Coley et al. (2019b)は，合成経路探索と反応条件予測の機械学習モデルを実装した自動合成実験のロボットを開発し，実験検証を行った．

7. まとめ

合成反応の予測や合成経路の自動設計の問題は，有機化学の世界において 50 年以上も研究されてきた非常に歴史の長い研究テーマである．学術研究の創成期から脈々と受け継がれてきたルールベースの予測手法は，近年の機械学習の進歩と合流することで，従来の技術とは全く異なるレベルに進化しようとしている．特に機械学習に基づく反応予測と合成経路設計の研究は 2017 年頃を機に急速に活発化した．これらの研究は有機化学や計算化学の研究者が全く思いもよらない着想で有機合成の問題にアプローチしている．本稿ではこの転換期に注目し，当該分野における機械学習の先進応用とその問題点を論じた．今後，モデルの予測結果の実験検証が行われ，化学の新しい発見がうまれると予想される．また，機械学習とロボットによるハイスループット実験の技術が合流していくことで，全自動合成システムが実用化されるに違いない．このような学術の潮流を見据えて，データセットの整備，特にモデルの訓練に必要な反応条件，反応収量，また失敗データなどの包括的なデータセットを整備していくことが求められる．全自動合成システムが実現すれば，実験計画法などの統計手法と組み合わせることで，革新的な化学反応や新材料の発見が大いに期待される．

参 考 文 献

Bradshaw, J., Paige, B., Kusner, M. J., Segler, M. and Hernández-Lobato, J. M. (2019). A model to search for synthesizable molecules, *Advances in Neural Information Processing Systems*, **32**, 7937–7949, https://proceedings.neurips.cc/paper/2019/file/46d0671dd4117ea366031f87f3aa0093-Paper.pdf.

Cao, N. D. and Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs, *arXiv*, http://arxiv.org/abs/1805.11973.

Coley, C. W., Rogers, L., Green, W. H. and Jensen, K. F. (2017a). Computer-assisted retrosynthesis based on molecular similarity, *ACS Central Science*, **3**(12), 1237–1245, DOI: http://dx.doi.org/10.1021/acscentsci.7b00355.

Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. and Jensen, K. F. (2017b). Prediction of organic reaction outcomes using machine learning, *ACS Central Science*, **3**(5), 434–443, DOI: http://dx.doi.org/10.1021/acscentsci.7b00064.

Coley, C. W., Jin, W., Rogers, L., Jamison, T. F., Jaakkola, T. S., Green, W. H., Barzilay, R. and Jensen, K. F. (2019a). A graph-convolutional neural network model for the prediction of chemical reactivity, *Chemical Science*, **10**(2), 370–377, DOI: http://dx.doi.org/10.1039/C8SC04228D.

Coley, C. W., Thomas, D. A., Lummiss, J. A. M., Jaworski, J. N., Breen, C. P., Schultz, V., Hart, T., Fishman, J. S., Rogers, L., Gao, H., Hicklin, R. W., Plehiers, P. P., Byington, J., Piotti, J. S., Green, W. H., Hart, A. J., Jamison, T. F. and Jensen, K. F. (2019b). A robotic platform for flow synthesis of organic compounds informed by AI planning, *Science*, **365**(6453), p.eaax1566, http://science.sciencemag.org/content/365/6453/eaax1566.abstract, DOI: http://dx.doi.org/10.1126/science.aax1566.

Corey, E. J. and Wipke, W. T. (1969). Computer-assisted design of complex organic syntheses, *Science*, **166**(3902), p.178, http://science.sciencemag.org/content/166/3902/178.abstract, DOI: http://dx.doi.org/10.1126/science.166.3902.178.

Dai, H., Li, C., Coley, C., Dai, B. and Song, L. (2019). Retrosynthesis prediction with conditional graph logic network, *Advances in Neural Information Processing Systems*, **32**, 8872–8882, https://proceedings.neurips.cc/paper/2019/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf.

Gao, H., Struble, T. J., Coley, C. W., Wang, Y., Green, W. H. and Jensen, K. F. (2018). Using machine learning to predict suitable conditions for organic reactions, *ACS Central Science*, **4**(11), 1465–1476, DOI: http://dx.doi.org/10.1021/acscentsci.8b00357.

Gasteiger, J., Pförtner, M., Sitzmann, M., Höllering, R., Sacher, O., Kostka, T. and Karg, N. (2000). Computer-assisted synthesis and reaction planning in combinatorial chemistry, *Perspectives in Drug Discovery and Design*, **20**(1), 245–264, DOI: http://dx.doi.org/10.1023/A:1008745509593.

Gelernter, H. L., Sanders, A. F., Larsen, D. L., Agarwal, K. K., Boivie, R. H., Spritzer, G. A. and Searleman, J. E. (1977). Empirical explorations of SYNCHEM, *Science*, **197**(4308), 1041–1049, DOI: http://dx.doi.org/10.1126/science.197.4308.1041.

Gottipati, S. K., Sattarov, B., Niu, S., Pathak, Y., Wei, H., Liu, S., Liu, S., Blackburn, S., Thomas, K., Coley, C., Tang, J., Chandar, S. and Bengio, Y. (2020). Learning to navigate the synthetically accessible chemical space using reinforcement learning, *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, **119**, 3668–3679, http://proceedings.mlr.press/v119/gottipati20a.html.

Guo, Z., Wu, S., Ohno, M. and Yoshida, R. (2020). Bayesian algorithm for retrosynthesis, *Journal of Chemical Information and Modeling*, **60**(10), 4474–4486, DOI: http://dx.doi.org/10.1021/acs.jcim.0c00320.

Ikebata, H., Hongo, K., Isomura, T., Maezono, R. and Yoshida, R. (2017). Bayesian molecular design with a chemical language model, *Journal of Computer-Aided Molecular Design*, **31**(4), 379–391, DOI: http://dx.doi.org/10.1007/s10822-016-0008-z.

Jin, W., Coley, C., Barzilay, R. and Jaakkola, T. (2017). Predicting organic reaction outcomes with Weisfeiler-Lehman network, *Advances in Neural Information Processing Systems*, **30**, 2607–2616, https://proceedings.neurips.cc/paper/2017/file/ced556cd9f9c0c8315cfbe0744a3baf0-Paper.pdf.

Jin, W., Barzilay, R. and Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation, *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, **80**, 2323–2332, http://proceedings.mlr.press/v80/jin18a.html.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes, *2nd International Conference on Learning Representations, Conference Track Proceedings*, Banff, Alberta, Canada, http://arxiv.org/abs/1312.6114.

Landrum, G. et al. (2006). RDKit: Open-source cheminformatics, https://www.rdkit.org.

Lin, K., Youjun, X., Pei, J. and Lai, L. (2020). Automatic retrosynthetic route planning using template-free models, *Chemical Science*, **11**, 3355–3364, DOI: http://dx.doi.org/10.1039/C9SC03666K.

Liu, B., Ramsundar, B., Kawthekar, P., Shi, J., Gomes, J., Luu Nguyen, Q., Ho, S., Sloane, J., Wender, P. and Pande, V. (2017). Retrosynthetic reaction prediction using neural sequence-to-sequence models, *ACS Central Science*, **3**(10), 1103–1113, DOI: http://dx.doi.org/10.1021/acscentsci.7b00303.

Lowe, D. M. (2012). Extraction of chemical structures and reactions from the literature, PhD Thesis, University of Cambridge, DOI: http://dx.doi.org/10.17863/CAM.16293.

Mikulak-Klucznik, B., Gołębiowska, P., Bayly, A. A., Popik, O., Klucznik, T., Szymkuć, S., Gajewska, E. P., Dittwald, P., Staszewska-Krajewska, O., Beker, W., Badowski, T., Scheidt, K. A., Molga, K., Mlynarski, J., Mrksich, M. and Grzybowski, B. A. (2020). Computational planning of the synthesis of complex natural products, *Nature*, **588**(7836), 83–88, DOI: http://dx.doi.org/10.1038/s41586-020-2855-y.

Pavlov, D., Rybalkin, M., Karulin, B., Kozhevnikov, M., Savelyev, A. and Churinov, A. (2011). Indigo: Universal cheminformatics API, *Journal of Cheminformatics*, **3**(1), p.P4, DOI: http://dx.doi.org/10.1186/1758-2946-3-S1-P4.

Pensak, D. A. and Corey, E. J. (1977). LHASA — Logic and heuristics applied to synthetic analysis, *Computer-Assisted Organic Synthesis*, 1–32, American Chemical Society, Washington, D. C., DOI: http://dx.doi.org/10.1021/bk-1977-0061.ch001.

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints, *Journal of Chemical Information and Modeling*, **50**(5), 742–754, DOI: http://dx.doi.org/10.1021/ci100050t.

Schwaller, P., Gaudin, T., Lányi, D., Bekas, C. and Laino, T. (2018). "Found in Translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models, *Chemical Science*, **9**(28), 6091–6098, DOI: http://dx.doi.org/10.1039/C8SC02339E.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C. and Lee, A. A. (2019). Molecular Transformer: A model for uncertainty-calibrated chemical reaction prediction, *ACS Central Science*, **5**(9), 1572–1583, DOI: http://dx.doi.org/10.1021/acscentsci.9b00576.

Segler, M. H. S. and Waller, M. P. (2017). Neural-symbolic machine learning for retrosynthesis and reaction prediction, *Chemistry — A European Journal*, **23**(25), 5966–5971, DOI: http://dx.doi.org/10.1002/chem.201605499.

Segler, M. H. S., Preuss, M. and Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI, *Nature*, **555**(7698), 604–610, DOI: http://dx.doi.org/10.1038/nature25978.

Sutskever, I., Vinyals, O. and Le, Q. V. (2014). Sequence to sequence learning with neural networks, *Advances in Neural Information Processing Systems*, **27**, 3104–3112, https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning*: *An Introduction*, MIT Press, Cambridge, Massachusetts.

Szymkuć, S., Gajewska, E. P., Klucznik, T., Molga, K., Dittwald, P., Startek, M., Bajczyk, M. and Grzybowski, B. A. (2016). Computer-assisted synthetic planning: The end of the beginning, *Angewandte Chemie International Edition*, **55**(20), 5904–5937, DOI: http://dx.doi.org/10.1002/anie.201506101.

Tanimoto, T. T. (1958). *An Elementary Mathematical Theory of Classification and Prediction*, International Business Machines Corporation, New York, New York.

Tolstikhin, I. O., Bousquet, O., Gelly, S. and Schölkopf, B. (2018). Wasserstein auto-encoders, *6th International Conference on Learning Representations, Conference Track Proceedings*, Vancouver, British Columbia, Canada, https://openreview.net/forum?id=HkL7n1-0b.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need, *Advances in Neural Information Processing Systems*, **30**, 5998–6008, https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *Journal of Chemical Information and Modeling*, **28**(1), p.31, DOI: http://dx.doi.org/10.1021/ci00057a005.

You, J., Liu, B., Ying, Z., Pande, V. and Leskovec, J. (2018). Graph convolutional policy network for goal-directed molecular graph generation, *Advances in Neural Information Processing Systems*, **31**, 6410–6421, https://proceedings.neurips.cc/paper/2018/file/d60678e8f2ba9c540798ebbde31177e8-Paper.pdf.

Zheng, S., Rao, J., Zhang, Z., Xu, J. and Yang, Y. (2020). Predicting retrosynthetic reactions using self-corrected Transformer neural networks, *Journal of Chemical Information and Modeling*, **60**(1), 47–55, DOI: http://dx.doi.org/10.1021/acs.jcim.9b00949.

Machine Learning in Reaction Prediction and Synthetic Route Design

Zhongliang Guo

The Institute of Statistical Mathematics

In organic chemistry, predicting the products from the reactants is called reaction prediction, while the design of synthetic routes in the opposite direction from the final products, which are the target molecule, is called synthetic route design. Reaction prediction and synthetic route design have been studied for more than 50 years. In recent years, advances in machine learning have significantly improved the accuracy of predicting chemical reactions. In this paper, we review the applications of machine learning in chemical reactions published since 2017. In particular, because the prediction of chemical reactions in the forward and reverse directions is different in mathematical formulation, we consider the differences in the application of machine learning methods. We also introduce a method for designing synthetic routes based on Bayesian inference presented by our group.

Key words: Reaction prediction, synthetic route design, machine learning, Bayesian inference.

統計数理(2021)
第 69 巻 第 1 号 99–118

［研究ノート］

ソーシャルメディア上のテキスト情報を考慮した社会ネットワーク分析モデル

―次数異質性モデルへの拡張―

五十嵐 未来[†]・照井 伸彦[†]

(受付 2020 年 5 月 25 日；改訂 10 月 6 日；採択 11 月 2 日)

要　　旨

近年，社会ネットワークをモデル化して分析する際に，ネットワーク情報だけでなく，人々がソーシャルメディア上などで生成するテキスト情報を考慮してコミュニティ構造を捉えることの重要性が増している．テキスト情報を考慮することにより，ネットワーク上で密にエッジが形成されている構造の中に，人々が持つ興味や関心に応じた複数のまとまりが存在するというような複雑なコミュニティ構造を持つ社会ネットワークの分析が可能となる．本研究では，これをモデル化した先行研究によるネットワークデータとテキストデータの同時利用モデルを拡張し，社会ネットワークにおいて一般的な性質であるエッジの生成されやすさがノードごとに異なる次数異質性を考慮したモデルを提案する．Twitter を用いた実証分析では，テキスト情報の活用及び次数異質性の考慮が予測性能に与える影響を検証するため，複数の比較モデルと共に比較実験を行い，提案モデルが優れた予測性能を持つことを示した．

キーワード：社会ネットワーク分析，コミュニティ検出，テキスト解析，トピックモデリング，ベイズ推定，ノード次数異質性．

1. はじめに

Social Networking Sites(SNS)の流行や電子商取引サイトの台頭などにより，消費者を取り巻く社会ネットワークを分析し，その構造を把握することは，企業のマーケティング活動における重要な位置を占めるようになっている．社会ネットワーク分析の手法は，統計学や社会学の分野を中心に長年研究されており，コミュニティ構造の抽出に代表されるように，ネットワークデータを要約し理解するための統計モデルが多く提案されている(e.g., Snijders and Nowicki, 1997; Airoldi et al., 2008)．これらのモデルでは，ネットワーク上のノードとエッジを観測データとして扱い，他と比べてエッジ密度が高くなるノード集合として定義されるコミュニティ構造を抽出する．また，社会ネットワークにおけるノードは，人々のことを表しており，人々の属性や行動といった付随的なデータを考慮することで，ネットワークモデルの精緻化を目指す研究も熱心に取り組まれている(e.g., Handcock et al., 2007)．中でも，近年は，ソーシャルメディアの流行や口コミ機能を搭載した電子商取引サイトの台頭などにより，ノードごとに固有のデータとしてユーザー生成コンテンツ(User-Generated-Contents，UGC)，特にテキスト情報

[†] 東北大学 経済学研究科：〒 980–8576 宮城県仙台市青葉区川内 27–1

をネットワーク情報と組み合わせた社会ネットワーク分析モデルが多く提案されている(e.g., Liu et al., 2009; Bouveyron et al., 2018).

ネットワーク情報だけでなくテキスト情報も考慮したモデルを構築することの利点は，一方の情報だけでは捉えることが難しいコミュニティ構造を識別できることである．ネットワーク情報のみを考慮する従来の研究では，他と比べてエッジ密度が高くなるノード集合をコミュニティと定義するが，テキスト情報も考慮してネットワークモデルを構築する近年の研究では，エッジの密度だけでなく，トピック比率などを用いたテキスト情報の類似性も考慮してコミュニティを定義している．例えば，Igarashi and Terui (2020)では，そのようなコミュニティをトピックベース・コミュニティと名付け，ネットワークのみ，もしくはテキストのみを利用するモデルと比較して，両者を考慮するモデルの方が精度よくコミュニティの分割が可能となることを示している．より具体的にトピックベース・コミュニティの概念を説明するために，例えば，ある学校の同級生で構成されるコミュニティを想定する．そこでは，学生らは互いに何らかの関係性を持った密度の高いネットワークが形成されているはずである．したがって，ネットワーク情報のみを考慮したモデルを用いると，そのようなネットワーク上には，一つのコミュニティが存在していると認識される．しかし，それと同時に，学生らは音楽や読書，スポーツといった様々な趣味を持っていることが考えられるため，共通の趣味を持った学生らをまとめて複数のコミュニティが存在するとみなす方が，より意味のあるセグメンテーションとなる可能性がある．ソーシャルメディアに代表されるオンライン上の社会ネットワークでは，上で説明したような，現実世界における社会的なつながりの上に存在する，興味や関心などに基づいたつながり，つまりトピックベース・コミュニティが点在していることが考えられる．したがって，ソーシャルメディア上に生成されたテキストコンテンツからそのユーザーの興味や関心を推定してネットワーク情報と結びつけることは，社会ネットワーク分析モデルを精緻化させることに繋がるだけでなく，複雑多様な現代のオンライン社会ネットワークの構造を理解するうえで有意義な分析となりうる．

また，社会ネットワークが持つ性質の一つとして，次数がノードごとに異質的であるという性質(次数異質性)がある．これは，多くの人々が少数の人とだけネットワーク上で関係性を持つ一方で，限られた人々が多くの人々と関係性を持つ傾向にあるという性質である．このように，現実の社会ネットワークにおいて，エッジが結ばれる確率は一定ではなく，ノードごとに異質であると仮定する方が，より現実に即したモデリングと言えるが，確率的ブロックモデル(Snijders and Nowicki, 1997; Igarashi and Terui, 2020)など代表的なネットワークモデルの多くでは，次数異質性を考慮していない．本研究では，ブロックモデルに即したエッジ生成確率をノードごとに異なるパラメータとして推定することで Igarashi and Terui (2020)のモデルを拡張し，次数異質性を考慮したモデルを提案する．実証分析では，現実のソーシャルメディアとして Twitter から得られたデータセットを用い，提案モデルである次数異質性を考慮した上でネットワーク情報とテキスト情報を結びつける社会ネットワークモデルを推定する．また，提案モデルに含まれる次数異質性及びテキスト情報の利用が外挿予測に与える効果を検証するために，提案モデルからそれぞれの特徴を除いた比較モデルとの比較実験を行う．

以下，2 節では，社会ネットワーク分析に関係する先行研究をまとめ，本研究の目的と位置づけを明確にする．3 節では，提案モデルを説明し，4 節ではその推定法を導出する．続いて，5 節では，Twitter データを利用した実証研究を報告し，最後に，6 節で結論と今後の課題を述べる．

2. 先行研究

2.1 社会ネットワーク分析モデルの進展

統計学や社会学などを中心として，古くから社会ネットワークをモデル化し，その構造を把握するための研究が続いている．中でも代表的なものが，確率的ブロックモデル（Stochastic Block Models, SBM, Wang and Wong, 1987; Snijders and Nowicki, 1997）である．SBM は，ノードが K 個のコミュニティのうち一つだけに属することを仮定しており，ノード i が属するコミュニティを $z_i \in \{1, \ldots, K\}$ とすると，ノード i と j の間にエッジが生成される確率は，$\psi_{z_i z_j}$ で表される．これは，$K \times K$ 行列 Ψ の (z_i, z_j) 成分であり，エッジ確率を表すパラメータである．

SBM は，様々な文脈でモデルの拡張が行われている．SBM がノードに単一のメンバーシップを仮定していたのに対し，Airoldi et al.（2008）は，各ノードが他ノードとの関係性ごとに異なるコミュニティに属することを許容する混合メンバーシップ確率的ブロックモデル（Mixed Membership Stochastic Blockmoldels, MMSB）を提案している．ノード i から j の関係性において，ノード i が属するコミュニティを s_{ij}（sender），ノード j が属するコミュニティを r_{ji}（receiver）とすると，両者の間にエッジが生成される確率は，$\psi_{s_{ij} r_{ji}}$ で表される．この拡張により，MMSB はコミュニティの重なりを考慮することができ（SBM ではコミュニティが重なることはない），より現実に即したモデリングが可能となっている．

また，社会学の文脈では，ノード間の関係性が性別や年齢といったノード固有の特徴量の影響を受けて決まることも知られている（Hoff et al., 2002; Handcock et al., 2007; Krivitsky et al., 2009）．しかし，本研究では，ソーシャルメディアに代表されるようなオンライン上の社会ネットワークに着目しているため，そのような特徴量は考慮しない．Twitter のような匿名型ソーシャルメディアでは，ユーザーは年齢や性別といった個人情報を隠した状態でアカウントを登録することができ，そのような状況において他者と関係を結ぶ際に考慮できる情報は，相手が形成しているネットワークとメディア上に投稿したコンテンツのみである．ただし，そのようなノードの属性を示すデータが利用可能であれば，提案モデルに取り込むことは容易であり，社会学的視点からの分析も可能である．

2.2 ネットワークとテキスト情報の同時モデリングに関する研究

前節で挙げた社会ネットワークモデルに関する研究では，ネットワーク情報のみに着目してモデルを提案しているが，近年，Twitter や Facebook といったオンライン上の社会ネットワーク構造をより深く理解するために，ネットワークとテキスト情報をどちらも考慮するモデルが盛んに研究されている．例えば，Chang and Blei（2010）は，ノードに固有のテキスト情報に対してトピックモデルを適用し，ノードのテキストに割り当てられたトピック割合の類似度に応じてノード間のエッジ生成確率が定義される関係トピックモデル（Relational Topic Model, RTM）を提案している．ただし，RTM の目的が，ネットワーク情報を加味してテキスト情報におけるトピックを推定するのに対して，Igarashi and Terui（2020）及び本研究のモデルは，テキスト情報を考慮してネットワーク上のコミュニティ構造を把握する点で対照的である．

Chang and Blei（2010）のようにテキスト情報を潜在的ディリクレ配分法（latent Dirichlet allocation,LDA, Blei et al., 2003）やその拡張モデルを用いてネットワークモデルに取り込むという方法は他にもいくつかの研究で見られる．例えば，Liu et al.（2009）は，Topic-Link LDA を提案しており，ノード固有のテキスト情報を考慮してコミュニティ構造を検出するという点で本研究と同じ目的を持っている．ただし，SBM と同様に，ノードが単一のコミュニティに属することを仮定した限定的な研究である．また，Liu et al.（2009）では，エッジ生成確率

表 1. 提案モデルと既存モデルの比較.

	観測データ	メンバーシップ	グラフの方向性	次数異質性
Blei et al. (2003)	テキストのみ	混合	-	-
Snijders and Nowicki (1997)	ネットワークのみ	単一	両方可能	考慮せず
Airoldi et al. (2008)	ネットワークのみ	混合	両方可能	考慮せず
Chang and Blei (2010)	ネットワーク／テキスト	混合	無向グラフのみ	考慮せず
Liu et al. (2009)	ネットワーク／テキスト	単一	無向グラフのみ	考慮せず
Bouveyron et al. (2018)	ネットワーク／テキスト	単一	両方可能	考慮せず
Zhu et al. (2013)	ネットワーク／テキスト	混合	両方可能	考慮せず
Igarashi and Terui (2020)	ネットワーク／テキスト	混合	両方可能	考慮せず
Karrer and Newman (2011)	ネットワークのみ	単一	両方可能	ノードごとの期待次数パラメータを導入
本研究	ネットワーク／テキスト	混合	両方可能	エッジ確率を異質パラメータとして定義

が，ノード固有のトピック及びコミュニティ割合の類似度によって定義されているため，エッジの向きが逆になってもその生成確率が変わらない，つまり無向グラフを想定しているのに対し，本研究を含めたブロックモデルにおいては，$K \times K$ 行列のエッジ確率パラメータを用いたネットワークモデリングにより，グラフの方向性にかかわらずモデルを適用可能である．他にも，Bouveyron et al. (2018)は，SBM にテキスト情報のモデルを加える形で拡張した Stochastic Topic Block Model(STBM)を提案している．

これらは単一のメンバーシップを仮定した SBM の拡張モデルであるが，Zhu et al. (2013)は，ノードの混合メンバーシップを仮定し，テキストとネットワーク情報の両者を考慮するネットワーク分析モデルを提案している．本研究における提案モデルとの相違点は，Zhu et al. (2013)は，エッジに割り当てられるコミュニティと単語に割り当てられるトピックが同一の分布に従っているという点であり，言い換えれば，コミュニティとトピックの次元を同一のものとして扱っている．しかし，現実の社会ネットワークでは，コミュニティとトピックが必ずしも互いに対応しているとは限らない．例えば，音楽とスポーツに興味のある人々が同じコミュニティ内に存在するネットワークを考える．このようなコミュニティを Zhu et al. (2013)のモデルで検出したとすると，一つのコミュニティに対して，音楽とスポーツという複数の意味的まとまりをもつトピックが対応してしまい，トピックの解釈性に欠ける．一方で，Igarashi and Terui (2020)及び本研究では，コミュニティとトピックがそれぞれ異なる分布に従うことを仮定しており，上記のようなネットワークに対しても，一つのコミュニティと，音楽トピック及びスポーツトピックのように別々に複数トピックを対応させることができる．3 節では，その詳細な定式化を説明する．

これらの既存モデルを踏まえて，本研究では，Igarashi and Terui (2020)によるノードの混合メンバーシップを仮定したネットワークとテキストの同時モデリングを拡張し，エッジ確率をノードごとに異質なパラメータとするモデルを検討する．これにより，社会ネットワークが一般的に有する次数異質性を考慮したモデリングが可能となる．エッジ生成確率の異質性については，Karrer and Newman (2011)が，ノードごとの期待次数をパラメータとして導入し，関係するノードに応じてエッジ生成確率が異質となるような補正を行うモデルを提案している．一方，本研究は，エッジ生成確率自体をノードごとに異質なパラメータとして直接推定する．

表 1 では，ここまでに議論した本研究と先行研究との比較をまとめている．まず，ネットワークやテキストどちらかのみを観測データとして扱うモデルと比較すると，本研究で提案するモデルは，その両者を考慮して社会ネットワーク分析を行うものであり，前述したようにどちらか一方の情報だけでは捕捉することが難しいネットワーク構造を明らかに出来る可能性がある．また，その両情報を扱う既存モデルと比較すると，ノードに混合メンバーシップを許容

している点，グラフの有向無向にかかわらず適用可能な点，そして社会ネットワークにおける次数異質性を考慮したモデリングを行っている点が本研究の特徴である．

3. モデル

本節では，まず提案モデルの基礎となる Igarashi and Terui (2020)のモデルを説明し，次にその差異を明らかにしながら本研究で使用するモデルの説明を行う．また，両モデルで共通して，観測されるデータは，ネットワーク情報を表す隣接行列 A，及びノードに固有のテキスト情報を表す単語の Bag-of-Words 集合 W の二つである．

まず，D 個のノードを持つ有向グラフを考えると，その隣接行列 A は，$D \times D$ 行列であり，行列の各要素はノード間の関係性を示す二値変数である．つまり，$a_{ij} = 0$ はエッジが存在しないことを表し，$a_{ij} = 1$ は存在することを表す．また，自己ループは考えないこととし，全ての i について $a_{ii} = 0$ である．Igarashi and Terui (2020)では，ノード i から j への関係性において，その送り手 i が潜在的なコミュニティ $s_{ij} \in \{1, \ldots, K\}$($K$ はコミュニティ数)に属し，受け手 j が潜在コミュニティ $r_{ji} \in \{1, \ldots, K\}$ に属することを仮定する．また，これら潜在コミュニティの行列表現を $S = (s_{ij}), R = (r_{ji})$ とする．モデルの生成過程において，送り手及び受け手のコミュニティはカテゴリカル分布，$s_{ij} \mid \eta_i \sim Categorical\,(\eta_i)$，$r_{ji} \mid \eta_j \sim Categorical\,(\eta_j)$ に従う．ただし，$\eta_i = (\eta_{i1}, \ldots, \eta_{iK})^\top$ はノード i のコミュニティ所属割合を表すパラメータであり，$\sum_k \eta_{ik} = 1$ を満たす．このコミュニティ分布の行列表現は $H = (\eta_1, \ldots, \eta_D)$ で表される．H は事前分布としてディリクレ分布 $\eta_i \mid \gamma \sim Dirichlet(\gamma)$ に従うことを仮定しており，$\gamma = (\gamma_1, \ldots, \gamma_K)^\top$ は推定にあたって調整が必要なハイパーパラメータである．

ノード i と j 間の関係性 a_{ij} は，s_{ij} と r_{ji} が所与の時，ベルヌーイ分布，$a_{ij} \mid s_{ij} = k, r_{ji} = k', \psi_{kk'} \sim Bernoulli\,(\psi_{kk'})$ に従うことを仮定する．ただし，$\psi_{kk'}$ は，送り手のコミュニティが k，受け手のコミュニティが k' の時にエッジが生成される確率を示す．また，エッジ確率の $K \times K$ 行列表現は $\Psi = (\psi_{kk'})$ で表され，行列の各要素は，事前分布としてベータ分布，$\psi_{kk'} \mid \delta_{kk'}, \epsilon_{kk'} \sim Beta(\delta_{kk'}, \epsilon_{kk'})$ に従う．このとき，δ, ϵ は Ψ と同じ次元を持つハイパーパラメータである．

従って，コミュニティ分布 H を所与としたときのネットワークデータに対する条件付尤度は以下で定義される．

$$
\begin{aligned}
(3.1)\ & p(A, S, R, \Psi \mid H) \\
&= p(A \mid S, R, \Psi) p(S \mid H) p(R \mid H) p(\Psi \mid \delta, \epsilon) \\
&= \prod_{i=1}^{D} \left\{ \prod_{j=1, j \neq i}^{D} \{ p(a_{ij} \mid s_{ij}, r_{ji}, \Psi) p(s_{ij} \mid \eta_i) p(r_{ji} \mid \eta_j) \} \right\} \prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{kk'} \mid \delta_{kk'}, \epsilon_{kk'}).
\end{aligned}
$$

続いて，ノード固有のテキスト情報について考える．ここでは，ノード i が生成したテキストについて，文章内の単語の順番を無視して，つまり Bag-of-Words の形式で保存した M_i 個の単語を観測データとする．ノード i に関する m 番目の単語 w_{im} は潜在的なコミュニティ $x_{im} \in \{1, \ldots, K\}$ 及びトピック $z_{im} \in \{1, \ldots, L\}$($L$ はトピック数)を持つことを仮定する．単語コミュニティと単語トピックの配列表現はそれぞれ X と Z で表され，各配列の要素は M_i 次元のベクトルである．モデルの生成過程において，単語コミュニティ x_{im} はカテゴリカル分布 $x_{im} \mid \eta_i \sim Categorical(\eta_i)$ に従う．ここで，η_i が単語コミュニティ x_{im} だけでなく，ノードコミュニティ s_{ij}, r_{ji} を生成するパラメータであったことを思い出すと，η_i はネットワークデータとテキストデータのモデルに共通するパラメータであり，両者の情報をつなげる役割を果たしている．一方，単語トピックは単語コミュニティが所与の状態でカテゴリカル分布

$z_{im} \mid x_{im} = k, \Theta \sim Categorical(\theta_k)$ に従う．このとき，$\theta_k = (\theta_{k1}, \ldots, \theta_{kL})^\top$ は，コミュニティ k に関するトピック割合を示すパラメータであり，$\sum_l \theta_{kl} = 1$ を満たす．このトピック分布の行列表現は $\Theta = (\theta_1, \ldots, \theta_K)$ であり，事前分布はディリクレ分布 $\theta_k \mid \alpha \sim Dirichlet(\alpha)$ に従う．

単語トピック z_{im} を所与として，それに対応する単語 $w_{im} \in \{1, \ldots, V\}$（$V$ は総単語数）は，単語トピックに対応するカテゴリカル分布 $w_{im} \mid z_{im} = l, \Phi \sim Categorical(\phi_l)$ に従う．ただし，$\phi_l = (\phi_{l1}, \ldots, \phi_{lV})^\top$ は，そのトピックにおいて単語が生成される確率を表す単語分布であり，$\sum_v \phi_{lv} = 1$ を満たす．単語分布の行列表現は $\Phi = (\phi_1, \ldots, \phi_L)$ であり，その事前分布はディリクレ分布 $\phi_l \sim Dirichlet(\beta)$ に従う．

従って，テキストデータに対する条件付尤度は，同じくコミュニティ分布 H を所与として，以下で定義される．

(3.2) $p(W, X, Z, \Theta, \Phi \mid H)$

$$
\begin{aligned}
&= p(W \mid Z, \Phi) p(Z \mid X, \Theta) p(X \mid H) p(\Theta \mid \alpha) p(\Phi \mid \beta) \\
&= \prod_{i=1}^{D} \left\{ \prod_{m=1}^{M_i} \{ p(w_{im} \mid z_{im}, \Phi) p(z_{im} \mid x_{im}, \Theta) p(x_{im} \mid \eta_i) \} \right\} \prod_{k=1}^{K} p(\theta_k \mid \alpha) \prod_{l=1}^{L} p(\phi_l \mid \beta).
\end{aligned}
$$

コミュニティ分布 H を所与とすることで，式(3.1)及び(3.2)の条件付尤度が独立となる仮定を置いているため，Igarashi and Terui (2020)の結合分布は，式(3.1)と(3.2)及び H の密度を掛け合わせることで以下のように得られる．

(3.3) $p(A, W, S, R, X, Z, H, \Psi, \Theta, \Phi)$

$$
\begin{aligned}
&= \prod_{i=1}^{D} \left\{ \prod_{j=1, j \neq i}^{D} \{ p(a_{ij} \mid s_{ij}, r_{ji}, \Psi) p(s_{ij} \mid \eta_i) p(r_{ji} \mid \eta_j) \} \right. \\
&\qquad \left. \prod_{m=1}^{M_i} \{ p(w_{im} \mid z_{im}, \Phi) P(z_{im} \mid x_{im}, \Theta) p(x_{im} \mid \eta_i) \} \right\} \times \\
&\prod_{i=1}^{D} p(\eta_i \mid \gamma) \prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{kk'} \mid \delta_{kk'}, \epsilon_{kk'}) \prod_{k=1}^{K} p(\theta_k \mid \alpha) \prod_{l=1}^{L} p(\phi_l \mid \beta).
\end{aligned}
$$

Igarashi and Terui (2020)のモデルでは，ユーザーが生成したテキストコンテンツを考慮しながらネットワーク上のコミュニティ構造を把握する，つまりトピックベース・コミュニティを見つけることを目的としている．このとき，ノード間にエッジが生成される確率を，$a_{ij} = 1 \mid s_{ij} = k, r_{ji} = k' \sim Bernoulli(\psi_{kk'})$ として全てのノードに対して同質的であることを仮定している．しかし，前節でも説明したように，現実の社会ネットワークにおいては，次数がノードによって大きく異なることが一般的であり，Igarashi and Terui (2020)では，この性質を考慮できていないため，現実のネットワークデータに対して十分に適合できない可能性がある．

本研究では，この問題を解決するために，エッジ生成確率の部分を $a_{ij} \mid s_{ij} = k, r_{ji} = k' \sim Bernoulli(\psi_{jkk'})$ としてモデルを拡張する．このとき，$\psi_{jkk'}$ は，送り手のコミュニティが k で，受け手のコミュニティが k' の時にエッジが生成される確率を示し，受け手のノード j に依存する異質なパラメータである．この定式化により，例えば，受け手 j がコミュニティ k の中で多くのエッジを集める，いわゆるハブノードである場合に，$\psi_{jkk'}$ が大きな値を取ることでそれを表現する．これにより，提案モデルは，社会ネットワークにおける次数分布の異質性を反映し，ノードごとの次数の多寡に応じてエッジ確率パラメータを異質的に推定することで，

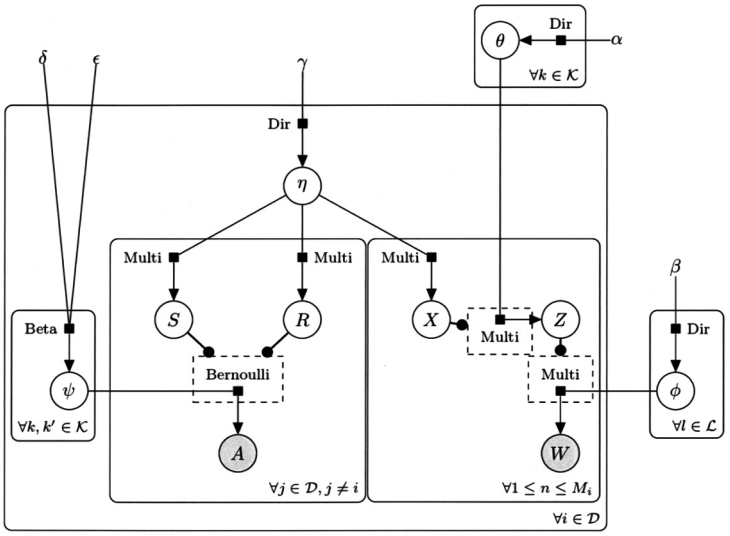


図 1. 提案モデルのグラフィカル表現.

より現実の社会ネットワークに即したモデリングが可能となる．したがって，次数異質性を提案モデルのパラメータを用いて言い換えれば，たとえ両ノードが同じようなコミュニティ分布 (η_i, η_j) を持っていたとしても，ノードによってエッジの繋がりやすさが異なるということであり，まさしくエッジ確率パラメータである $\psi_{i\cdot\cdot}, \psi_{j\cdot\cdot}$ が次数異質性を表していると解釈できる．また，エッジ確率の $K \times K$ 行列表現は $\Psi_i = (\psi_{ikk'})$ で表され，行列の各要素は，事前分布としてベータ分布 $\psi_{ikk'} \mid \delta_{kk'}, \epsilon_{kk'} \sim Beta(\delta_{kk'}, \epsilon_{kk'})$ に従うことを仮定する．

本研究で用いるモデルは，上述の点以外は Igarashi and Terui (2020) の定式化を採用する．このとき，コミュニティ分布 H を所与としたときのネットワークデータに対する尤度は，式 (3.1) から次式に変更される．

$$
\begin{aligned}
(3.4)\ \ & p(A, S, R, \Psi \mid H) \\
& = p(A \mid S, R, \Psi) p(S \mid H) p(R \mid H) p(\Psi \mid \delta, \epsilon) \\
& = \prod_{i=1}^{D} \left\{ \prod_{j=1, j\neq i}^{D} \{ p(a_{ij} \mid s_{ij}, r_{ji}, \Psi_j) p(s_{ij} \mid \eta_i) p(r_{ji} \mid \eta_j) \} \prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{ikk'} \mid \delta_{kk'}, \epsilon_{kk'}) \right\}.
\end{aligned}
$$

図 1 は提案モデルのグラフィカル表現である．

4. 条件付き事後分布とパラメータ推定

先行研究において，トピックモデルを推定するための手法は，変分ベイズ法や逐次学習法など多く提案されている．その中でも最も広く使われているものの一つが，崩壊型ギブスサンプリング (collapsed Gibbs sampling, CGS, Griffiths and Steyvers, 2004) である．これは，潜在変数の事後分布を導出する過程でモデルパラメータを積分消去し，サンプリングを効率的に行う手法である．以下では，本研究の提案モデルに対する CGS のための条件付き事後分布を導出する．

提案モデルにおける，コミュニティ分布 H，エッジ確率 Ψ，トピック分布 Θ，単語分布 Φ の

4 つのパラメータについては，事前分布との共役性に基づき，条件付き事後分布を既知の分布として導出することができる．ただし，その詳細な導出過程は付録 A に譲る．また，それ以外の潜在変数として，送り手及び受け手の潜在コミュニティ S, R，単語の潜在コミュニティ X 及び潜在トピック Z の 4 つがあるが，これらの条件付き事後分布は，付録 A で導出した事後分布を用いて以下のように導出される．

$$
\begin{aligned}
(4.1)\quad & p(s_{ij}=k, r_{ji}=k' \mid a_{ij}, A_{\setminus ij}, S_{\setminus ij}, R_{\setminus ji}, X, \gamma, \delta, \epsilon) \\
&\propto \int\int p(s_{ij}=k, r_{ji}=k' \mid \eta_i, \eta_j) p(x_i, x_j \mid \eta_i, \eta_j) p(\eta_i, \eta_j \mid S_{\setminus ij}, R_{\setminus ji}, X, \gamma) d\eta_i d\eta_j \\
&\quad \times \int p(a_{ij} \mid \psi_{jkk'}) p(\psi_{jkk'} \mid A_{\setminus ij}, S_{\setminus ij}, R_{\setminus ji}, \delta, \epsilon) d\psi_{jkk'} \\
&= \frac{N_{ik\setminus ij} + M_{ik} + \gamma_k}{\sum_t \left(N_{it\setminus ij} + M_{it} + \gamma_t\right)} \times \frac{N_{jk'\setminus ji} + M_{jk'} + \gamma_{k'}}{\sum_t \left(N_{jt\setminus ji} + M_{jt} + \gamma_t\right)} \times \\
&\quad \frac{\left(n^{(+)}_{jkk'\setminus ij} + \delta_{kk'}\right)^{\mathbb{I}(a_{ij}=1)} \left(n^{(-)}_{jkk'\setminus ij} + \epsilon_{kk'}\right)^{\mathbb{I}(a_{ij}=0)}}{n^{(+)}_{jkk'\setminus ij} + n^{(-)}_{jkk'\setminus ij} + \delta_{kk'} + \epsilon_{kk'}},
\end{aligned}
$$

$$
\begin{aligned}
(4.2)\quad & p(x_{im}=k, z_{im}=l \mid W, S, R, X_{\setminus im}, Z_{\setminus im}, \alpha, \beta, \gamma) \\
&\propto \int p(s_i, r_i \mid \eta_i) p(x_{im}=k \mid \eta_i) p(\eta_i \mid S, R, X_{\setminus im}, \gamma) d\eta_i \times \int p(z_{im}=l \mid \theta_k) \\
& p(\theta_k \mid X_{\setminus im}, Z_{\setminus,im}, \alpha) d\theta_k \times \int p(w_{im}=v \mid \phi_l) p(\phi_l \mid W_{\setminus im}, Z_{\setminus im}, \beta) d\phi_l \\
&= \frac{N_{ik} + M_{ik\setminus im} + \gamma_k}{\sum_t \left(N_{it} + M_{it\setminus im} + \gamma_t\right)} \times \frac{M_{kl\setminus im} + \alpha_l}{\sum_q \left(M_{kq\setminus im} + \alpha_q\right)} \times \frac{M_{lv\setminus im} + \beta_v}{\sum_u \left(M_{lu\setminus im} + \beta_u\right)}.
\end{aligned}
$$

ただし，式(4.1)における N_{ik} は，ノード i が持つ $D-1$ 個の関係性において，送り手及び受け手の潜在コミュニティとして k が割り当てられた回数を表し，M_{ik} は，ノード i の単語コミュニティに k が割り当てられた回数を表す．$n^{(+)}_{ikk'}$ は，ノード i に関する $D-1$ 個の関係性のうち，コミュニティ k, k' が割り当てられたエッジの数，$n^{(-)}_{ikk'}$ は，コミュニティ k, k' が割り当てられ，かつエッジのない関係性の数を表す．式(4.2)における M_{kl} は，コミュニティ k が割り当てられた単語のうちトピック l が割り当てられた回数，M_{lv} は，語彙 v にトピック l が割り当てられた回数を表す．また，添え字の $\setminus$ はこれらのカウントから，当該データを除くことを意味する．

CGS では，式(4.1)及び(4.2)に従って，各関係性及び単語に対して潜在コミュニティとトピックを繰り返しサンプリングする．最終的に，初期値に依存する稼働期間を除いたサンプルを用いて，積分消去していた 4 つのパラメータの期待値を計算することで推定値を得る．

5. 実証分析

5.1 使用データ

ここでは，現実のオンライン社会ネットワークに対して，提案モデルを用いた分析が有益であることを示すために，Twitter データを使った実証分析を行う．本節では，まず分析に用いたデータセットの概要と前処理について説明する．本研究では，任天堂株式会社が Twitter 上で保持している英語版公式アカウントを中心とするネットワークを対象として，以下の手順でデータを収集及び加工した．

表 2. WAIC によるモデル比較.

	$L=5$	$L=6$	$L=7$	$L=8$	$L=9$	$L=10$
$K=5$	4422206.32	4340879.93	4321068.95	4333535.35	4354814.11	4553144.83
$K=6$	4333313.32	4333488.66	4351008.38	4309479.01	4302773.27	4280703.13
$K=7$	4313265.58	4285253.01	**4272682.48**	4346780.91	4301005.75	4414800.13
$K=8$	4320416.87	4282485.37	4326300.05	4324393.23	4321806.29	4426226.19
$K=9$	4429170.84	4329997.66	4439594.82	4407656.85	4296128.61	4301655.85
$K=10$	4361219.83	4342899.53	4282056.30	4306509.44	4306244.12	4406655.34

まず，2018 年 5 月 1 日時点でのフォロー関係に従って，任天堂のアカウントをフォローしているユーザーからランダムにサンプリングを行った．続いて，サンプルされたユーザーをフォローしている別のユーザーからもランダムにサンプリングを行った．そして，それらのユーザーで形成されるネットワークにおいて，入次数と出次数の平均が 3 以下のユーザーを外れ値とみなしてデータセットから除外した．結果として，3,500 人のユーザーが残り，ネットワーク内におけるエッジの総数は 68,949 本であった．これらのユーザーで形成される有向グラフをネットワーク情報として使用する．

次に，テキストデータの作成方法を説明する．まず，上でサンプルされた 3,500 人分のアカウントに対して，2017 年 9 月 1 日から 2018 年の 2 月 28 日[1)] までに投稿した投稿内容からテキスト部分を全て抜き出した．これらのテキストデータに対して，文章から単語集合への分解，小文字への統一，数字，記号，及び主要なストップワード(a，the，I など)の削除，活用形から語幹への統一(stemming)の順に前処理を行った．さらに，処理済みのテキストデータのうち，コーパス内での頻度が 20 以下，あるいは 20 人以下のユーザーにしか使われていない低頻度の単語と，50 人以上のユーザーに使われている高頻度の単語を，トピック推定への悪影響を避けるためにデータセットから除いた．結果として，コーパス内には 9,001 種類の単語が残り，ノードごとの平均単語数は 98.2 であった．次節では，提案モデルにおけるコミュニティ数，トピック数の決定方法を説明したのち，作成したデータセットに対する提案モデルの推定結果について議論する．

5.2 分析結果

提案モデルを含めて，一般にブロックモデルを用いて分析する際には，事前にコミュニティ数(及び本研究ではそれに加えてトピック数)を決める必要がある．先行研究では，コミュニティ数の決定を情報量基準を用いたモデル比較として捉え，BIC による方法(Handcock et al., 2007; Saldaña et al., 2017)，integrated completed likelihood による方法(Daudin et al., 2008; Bouveyron et al., 2018)，変分ベイズによる方法(Latouche et al., 2012)など様々な手法が提案されている．しかし，本研究では，近年新たな情報量基準として提案され，現在では数多くの領域で使われている広く使える情報量基準(widely applicable information criterion, WAIC, Watanabe, 2010)をモデル比較の基準として採用した．提案モデルに対する WAIC の詳細は付録 B に譲る．表 2 は，コミュニティ数及びトピック数を 5 から 10 の範囲で設定し，5.1 節で作成したデータセットに対して WAIC を計算した結果である(K がコミュニティ数を，L がトピック数を表し，太字は表内で最少の値を意味する)．ただし，この時の繰り返し数は 5,000 回であり，そのうち 2,000 回を初期値に依存する稼働期間として除いた．また，ハイパーパラメータの設定は，それぞれ，$\alpha_l = 0.1, \forall l$，$\beta_v = 0.1, \forall v$，$\gamma_k = 1.0, \forall k$，$\delta_{kk'} = \epsilon_{kk'} = 0.1, \forall k, k'$ である．その結果，コミュニティ数 7，トピック数 7 のモデルが選ばれたため，以降ではこのモデルを用いた Twitter データの分析結果を議論する．

表 3. 提案モデルを用いて推定された単語分布において最も高い値を持つ上位 10 個の単語.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
nonfollow	teamemmmmsi	trapadr	criticalrol	iartg	growthhack	savvi
blackclov	dokkan	vevo	zeldathon	amread	digitalmarket	lube
hunterxhunt	twitchkitten	ddrive	orton	erotica	gdpr	foodporn
jojosbizarreadventur	vgc	leed	fursuitfriday	asmsg	smm	oiler
mkleosaga	roku	spinrilla	dramaalert	momlif	contentmarket	austria
wnf	wizebot	ifb	sdlive	hemp	gamedesign	tfc
hori	ryzen	gainwithpyewaw	htgawm	writerslif	podernfamili	crowdfir
mdva	freebiefriday	gainwithxtiandela	sml	bookreview	socialmedialmarket	tranc
hyrulesaga	streamersconnect	horford	robloxdev	kindleunlimit	bigdata	tock
nyxl	nbaliv	suav	yoongi	bookboost	emailmarket	texfil

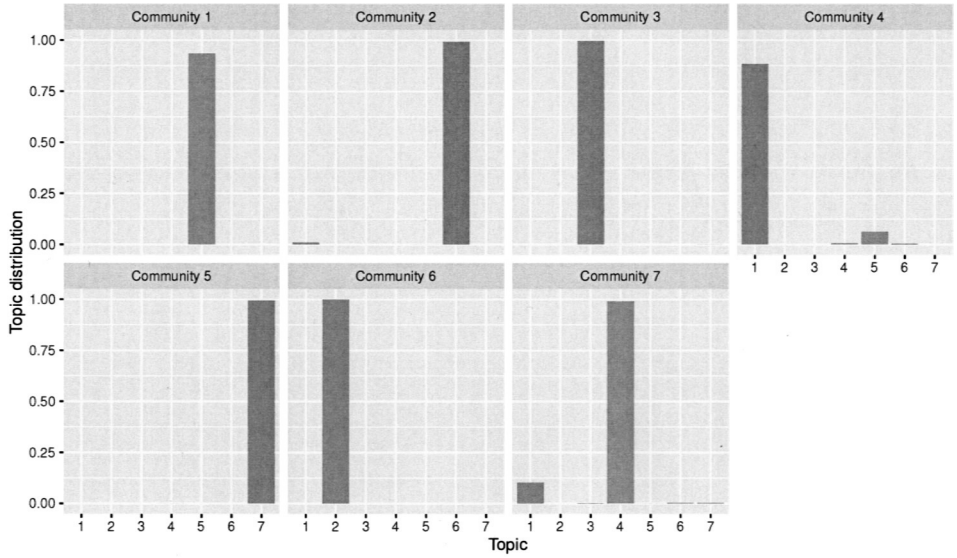


図 2. 提案モデルにおける各コミュニティのトピック分布に関する推定結果.

まず，ノードに依存しないグローバルパラメータ(単語分布 ϕ 及びトピック分布 θ)を見ることで，人々が検出されたコミュニティ内でどのようなことに関心を持っているのかが分かる．表 3 は，推定された単語分布の値が最も高い上位 10 個の単語をトピック毎に並べたものであり，これによってトピックの意味を解釈することができる．各トピックを代表する関連単語には下線が引かれており，トピックの意味は以下の通りに解釈できる[2)]．トピック 1 はアニメーションに関するトピック(代表的な単語は blackclov，hunterxhunt，jojosbizarreadventur など)，トピック 2 はストリーミング配信全般に関するトピック(代表的な単語は teamemmmmsi，twitchkitten，roku など)，トピック 3 は音楽に関するトピック(代表的な単語は vevo，spinrilla など)，トピック 4 はゲームストリーミング配信に関するトピック(代表的な単語は criticalrol，zeldathon など)，トピック 5 は読書に関するトピック(代表的な単語は amread，bookreview，kindleunlimit など)，トピック 6 はビジネスに関するトピック(代表的な単語は digitalmarket，smm，contentmarket など)，そしてトピック 7 はスポーツに関するトピック(代表的な単語は oiler，tfc など)と言える．

また，図 2 は，推定された各コミュニティのトピック分布であり，各コミュニティ内におけ

表 4. LDA を用いて推定された単語分布において最も高い値を持つ上位 10 個の単語.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
trapadr	teamemmmmsi	vevo	nonfollow	podernfamili	growthhack	gamedesign
ddrive	tfc	spinrilla	dokkan	iartg	digitalmarket	leed
ifb	twitchkitten	htgawm	zeldathon	amread	gdpr	savvi
gainwithpyewaw	hori	beck	criticalrol	asmsg	smm	lube
gainwithxtiandela	roku	orton	vgc	erotica	contentmarket	momlif
blackclov	mkleosaga	sdlive	fursuitfriday	foodporn	socialmediamarket	quoteoftheday
hunterxhunt	wnf	horford	dramaalert	dogsoftwitt	bigdata	austria
jojosbizarreadventur	wizebot	suav	sml	oiler	cto	hemp
yoongi	ryzen	herewego	robloxdev	writerslif	emailmarket	tranc
hoseok	streamersconnect	drippin	spforstreami	amiga	fintech	tock

るトピックの割合を確認することができる．これを見ると，コミュニティ毎のトピック分布が一つのトピックに集中しており，かつコミュニティ間で被りのない推定結果となっている．これは，提案モデルが，エッジが高密度かつテキストのトピックが類似するようなノード集合，つまりトピックベース・コミュニティを抽出するような構造になっているためと推察されるが，図 2 からのみでは判別できない．そこで，ネットワーク情報とテキスト情報を両方考慮する提案モデルの同時アプローチに対して，ネットワーク情報のみを考慮してコミュニティ構造を抽出する MMSB モデルと，テキスト情報のみを考慮してトピック構造を抽出する LDA モデルの結果を統合する独立アプローチとの比較を行うことで，図 2 の推定結果をさらに掘り下げていく．以下では，同時アプローチと独立アプローチに対して，単語分布によるトピックの解釈，コミュニティごとのトピック分布，推定されるコミュニティ構造をそれぞれ比較する．

まず，表 4 に，LDA によって抽出されたトピックに関する代表的な単語が示されている．これを見ると，提案モデルの推定結果である表 3 と同じ単語が同一のトピックに多く並んでおり，ネットワークとテキストを両方考慮したモデリングと，テキストのみのモデリングで同一のトピックを抽出していることが確認できる．

次に，MMSB モデルと LDA の結果を統合してコミュニティ毎のトピック分布を評価する．LDA モデルは，文書を単語集合とみなして文書ごとのトピック分布を推定するモデルであるが，ここではノードに単語集合が付随するとみなすため，ノードごとにトピック分布が推定される．また，MMSB モデルによってもノードごとにコミュニティへの所属割合が推定されている．したがって，コミュニティ所属割合で重みづけてトピック分布を足し合わせることで，提案モデルで推定されるようなコミュニティごとのトピック分布を事後的に導出することができる．LDA モデルによって推定されたノードごとのトピック分布を $\hat{\lambda}_i^{(ind)}$，MMSB モデルによって推定されたノードごとのコミュニティ分布を $\hat{\eta}_i^{(ind)}$ とすると，独立アプローチにおけるコミュニティごとのトピック分布 $\theta_k^{(ind)}$ は以下で導出される．

$$\theta_k^{(ind)} = \sum_{i=1}^{D} \hat{\lambda}_i^{(ind)} \times \hat{\eta}_{ik}^{(ind)}, \qquad k = 1, \ldots, K. \tag{5.1}$$

その結果は図 3 に示され，提案モデルの推定結果とは大きく異なり，一つのコミュニティに複数のトピックが対応していることが分かる．

そして，次数異質性を考慮したネットワークのみのモデルと提案モデルのコミュニティ内エッジ密度を比較する．両モデルとも相手ノードによって所属コミュニティが異なる混合メンバーシップを仮定しているため，推定されたコミュニティ分布に従って，ノードごとに最も高い値を持つコミュニティに所属するとしてエッジ密度を計算する．両モデルのエッジ密度及び

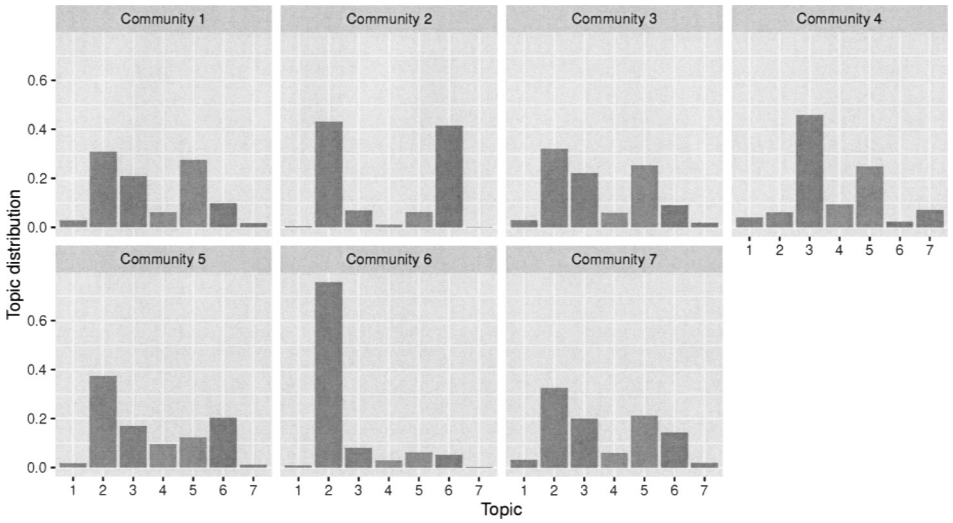


図 3. ネットワーク情報のみを考慮した比較モデルにおける各コミュニティのトピック分布.

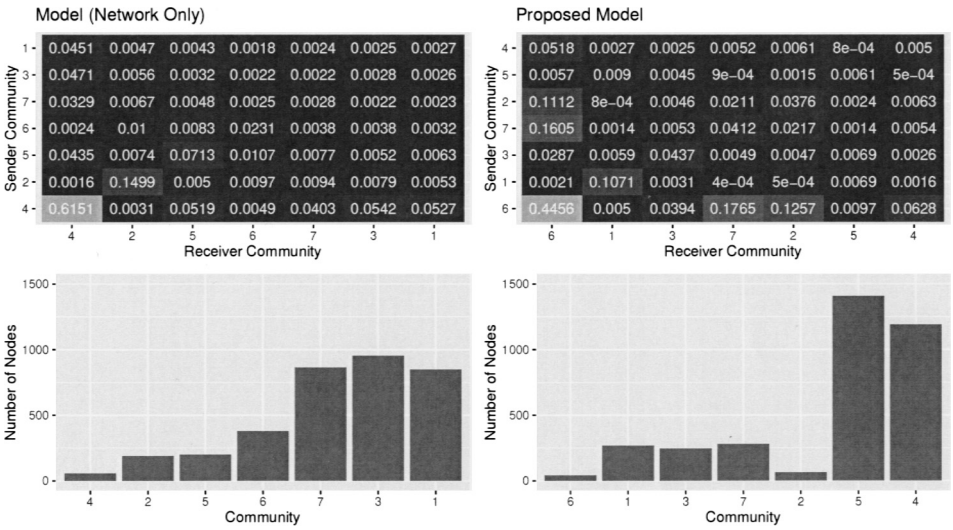


図 4. 提案モデル及び比較モデルにおける各コミュニティ内のエッジ密度と各コミュニティに所属するノード数.

コミュニティを構成するノードの数が図 4(左上と左下の図がネットワークのみを扱う MMSB モデルによる結果であり，右上と右下の図が提案モデルによる結果である)に示されている．なお，比較のため，コミュニティ内エッジ密度(対角成分)の大きい順にコミュニティ番号を並び変えている．これを見ると，多くのノードが所属するエッジ密度の低いコミュニティについて，ネットワークのみのモデルでは三つ，提案モデルでは二つで推定しているため，両モデルで値が全体的に少し異なっている．しかし，それ以外のコミュニティ構造については両者とも同様の構造を捉えており，例えば，少数のノードで構成されているコミュニティ(ネットワークのみのモデルでは 4 番，提案モデルでは 6 番のコミュニティ)や，他コミュニティとの関係

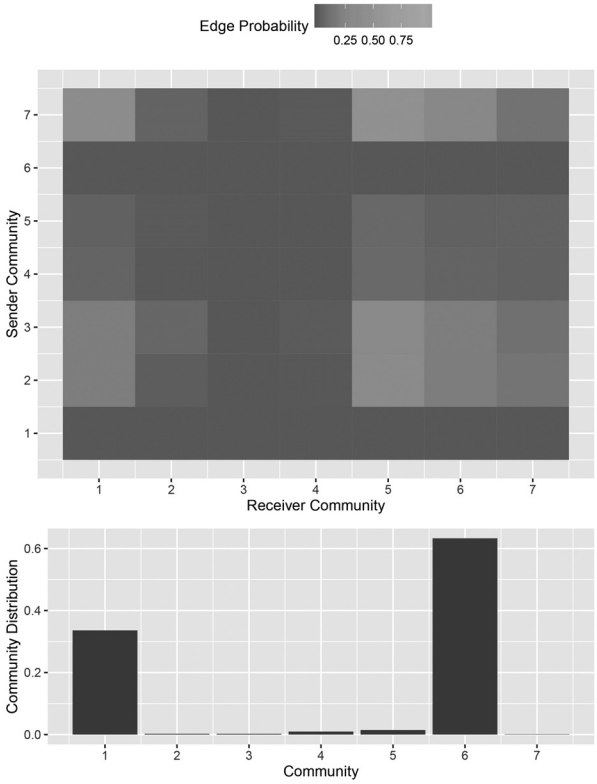


図 5. ノード 1 のエッジ確率とコミュニティ分布に関する推定結果.

は希薄であるものの，内部では比較的高い密度でつながっている中規模のコミュニティ(ネットワークのみモデルにおける 2，5 番，提案モデルにおける 1，3 番のコミュニティ)などが共通して推定されている．

以上により，提案モデルは，ネットワークあるいはテキストのみを考慮するモデルと全体的には同様のコミュニティ構造及びトピック構造を捕捉しながら，コミュニティ内のトピック構造をより明確に表現するモデルであると言える．ただし，これは今回用いたデータセットにおける結果であり，一般のネットワークにおける性質を検証するには理論解析など，さらなる議論が必要である．

最後に，各ノードについて異質なローカルパラメータ(エッジ確率の ψ)の推定結果を確認する．図 5 及び図 6 は，ノード番号 1 番と 237 番に関するエッジ確率とコミュニティ分布の推定結果である．また，ノード 1 の入次数は 6，出次数は 0 であり，ノード 237 の入次数は 657，出次数は 37 である．推定結果は，この両ノードの次数異質性を如実に表しており，ノード 1 が主に属するコミュニティ(コミュニティ 1 と 6)に関するエッジ確率は低い値で推定されているのに対して，ノード 237 が主に属するコミュニティ(コミュニティ 1 と 5)に関するエッジ確率は高い値で推定されている．このように，エッジ確率のパラメータに次数異質性を考慮した仮定を導入することで，より柔軟にネットワークモデルを表現できるようになり，テストデータに対する予測性能も向上することが期待される．次節では，これを検証するために，提案モデルからテキストの活用及び次数異質性の考慮という特徴を除いた比較モデルと共に比較実験を

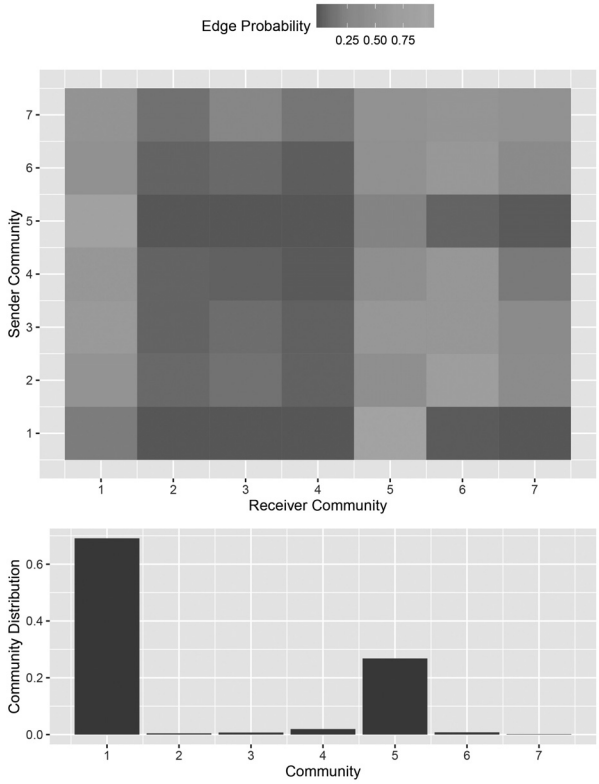


図 6. ノード 237 のエッジ確率とコミュニティ分布に関する推定結果.

行う.

5.3 予測性能の検証

本節では，提案モデルのテストデータに対する予測性能及びテキスト情報の活用と次数異質性の考慮が予測性能に与える効果を比較モデルとの比較を通して検証する．提案モデル（以降モデル IV）がテキスト情報の活用と次数異質性の考慮をどちらも含んだモデルであるのに対して，比較モデルとして，テキスト情報を活用せずかつ次数異質性を考慮しないモデル（以降モデル I），テキスト情報を活用せず次数異質性を考慮するモデル（以降モデル II），そしてテキスト情報を活用するが次数異質性は考慮しないモデル（以降モデル III）の 3 種類を検討する．

5.2 節では，全てのネットワーク，テキストデータを学習データとしてモデルの推定を行ったが，ここでは，各ノードが持つ $D-1$ 個の関係性のうち，90% を学習データとしてモデルの推定に使い，残りの 10% をテストデータとした．テキストデータについては，前節同様全てのデータを学習データとして用いた．また，繰り返し数やハイパーパラメータの設定も前節と同じ条件で推定している．これらの条件の下で学習データに対する推定を行い，各パラメータの推定値を得た．推定されたコミュニティ分布とエッジ確率を $\hat{H}, \hat{\Psi}$ と表すと，例えば提案モデルについては，テストデータ $a_{ij} \in A^{test}$ に対する予測確率は以下で計算できる．

$$p(a_{ij}=1) = \sum_{k=1}^{K}\sum_{k'=1}^{K} \hat{\eta}_{ik}\hat{\eta}_{jk'}\hat{\psi}_{jkk'} \tag{5.2}$$

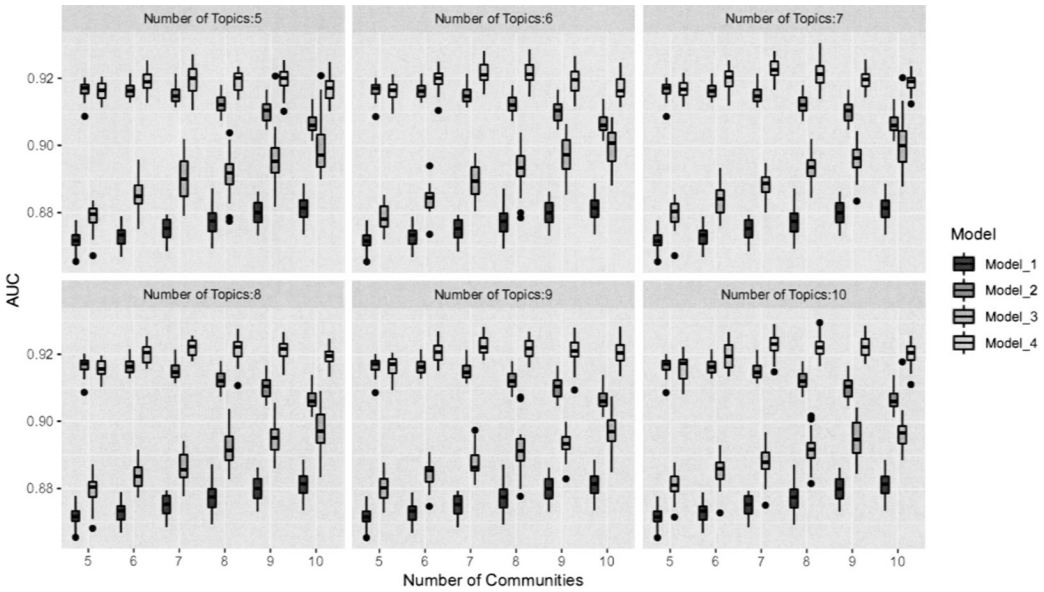


図 7. AUC による予測性能の比較.

3 種類の比較モデルについても同様に，コミュニティ分布とエッジ確率の積によって予測確率を計算できる．

ここでは，コミュニティ数とトピック数をそれぞれ 5 から 10 の値で設定し，各モデルの Area Under the Curve(AUC)をグリッド状に計算した．また，各モデル，各コミュニティ数及びトピック数について，テストデータを入れ替えながらモデルの推定と AUC の計算を 30 回繰り返した．図 7 はその時の AUC をまとめた箱ひげ図である．

まず，次数異質性を考慮しないモデル(モデル I，モデル III)と考慮するモデル(モデル II，IV)の比較を行う．次数異質性を考慮しない場合，コミュニティ数の増加と共に AUC が上昇している．一般的に，最適なクラスタ数に達するまでは，クラスタ数を増やすほどクラスタリングの精度は上昇するため，これは期待通りの結果と言える．一方で，次数異質性を考慮するモデルは，全体的に AUC が高く，異質性を導入することで，ネットワークモデルとしての予測性能が上がることが示されている．しかし，モデル II ではコミュニティ数の増加と共に AUC が下降しており，次数異質性が予測性能を悪化させる側面も持ち合わせている可能性がある．これは，異質性を導入することでノード一つ一つのエッジ確率を推定するためのデータが少なくなるため，コミュニティ数の増加と共にその影響が顕在化していると推察され，本研究で提案する形で異質性を導入するデメリットとも言える．その場合，階層ベイズを用いるなどして，エッジ確率の異質性を作り出す要因としてノード間に共通する構造を取り込むことで改善の可能性がある．

次に，テキスト情報を考慮しないモデル(モデル I，モデル II)と考慮するモデル(モデル III，モデル IV)を比較すると，テキスト情報を考慮することで予測精度が向上していることが分かる．さらに，その効果はコミュニティ数に対して一定ではなく，コミュニティ数が増えるにしたがって，テキスト効果による予測精度の上昇幅は大きくなっている．つまり，ネットワーク上のコミュニティを細かく分割していく際に，ネットワーク情報だけでは分割が難しい場面であっても，テキスト情報を使うことで，外挿予測に対しても頑健な，より精緻なクラスタリン

グが可能となる．提案モデルであるモデル IV に関しては，前述した次数異質性による予測性能への悪影響が予想されるが，テキスト効果による予測性能の改善が相殺しあうことで，コミュニティ数が増えても予測性能が落ちることなく，高い AUC を維持していると考えられる．また，トピック数については，その数が変化しても各モデル（モデル III，IV）の予測性能には大きな影響を与えていない．

結果として，テキスト情報と次数異質性を考慮する提案モデルがそれぞれの効果を考慮しない比較モデルとの比較の中で，最も予測性能に優れたモデルであることが示された．ただし，解決されていない課題もあり，次数異質性のモデルへの反映させ方によっては予測性能を悪化させる可能性があることについては，モデルの改善も含めて今後の課題としたい．

6. おわりに

本研究では，社会ネットワーク分析をより現実に即した有意義な分析とするために，ネットワーク情報だけでなく，人々の興味や関心を表すソーシャルメディア上のテキスト情報を考慮し，さらに，社会ネットワークに特有の次数異質性を加味したモデルを提案した．先行研究における既存モデルと比較したとき，本研究で提案するモデルの特色としては，ネットワーク上の各ノードが持つテキスト情報を利用している点，ノードがそれぞれの関係性によって異なるコミュニティに属することを許容している点，無向グラフか有向グラフにかかわらず適用できる点，そして次数異質性を考慮し，エッジ確率のパラメータがノードごとに異質であることを仮定している点が挙げられる．これによって，次数がノードによって大きく異なる一般的な社会ネットワークに対しても十分な適合性能を有しながら，エッジが密に集まっており，かつノードごとに固有のテキストデータが同一の分布から生成される，トピックベース・コミュニティの検出が可能となる．

実証分析の結果，崩壊型ギブスサンプリングによって推定される提案モデルは，現実の Twitter データに対して，解釈可能なコミュニティ及びトピック構造を捉えるだけでなく，次数異質性やテキスト情報を考慮しない比較モデルよりも優れた予測性能を持っていることが示された．さらに，この結果から，オンラインの社会ネットワーク分析において，ネットワーク上の大まかなコミュニティ構造を超えて，さらに細かくクラスターを分析していく場合は，各ノードが持つテキスト情報を加味することが外挿予測を向上させるという点で有益であることが分かった．

本研究では，オンライン上の社会ネットワークに着目したため，人々がネットワークを形成する際には，相手のネットワーク情報とテキスト情報のみを考慮するという仮定を置き，相手の年齢や性別といった属性情報，あるいは行動や態度といった情報は，これらのデータが利用できないことから提案モデルの考慮から外していた．一方で，社会ネットワーク分析に関する先行研究の文脈では，そのようなノード固有の（あるいはノード間の）特徴量がネットワーク形成に影響していることが多くの研究で示されている（Hoff et al., 2002; Handcock et al., 2007）．本研究では，エッジ形成の関数が，関係性を結ぶ両者のコミュニティ分布，及び関係性を受け取る側のエッジ確率で構成されていたが，先行研究を参照すると，ここに属性や行動情報といったノード固有の特徴量を組み込む拡張は有意義であり，これらの情報をモデルに取り込むことは直接的に可能である．データの利用可能性と合わせて今後の課題としたい．

注．

[1] テキストデータの前処理の段階で，大半のユーザーが，2018 年 3 月に開かれた Nintendo Direct という新商品発表イベントに関する投稿を行っていることが判明した．したがっ

て，本研究では，このような多くのユーザーで共通する同一の事象に対する投稿がトピックの推定に与える影響を避けるため，テキストデータの観測期間を 2018 年 2 月 28 日までとした．

[2)] 本文中で挙げた各トピックの代表的な単語の解説は次の通りである．トピック 1 の単語は，それぞれ，ブラッククローバー，ハンターハンター，ジョジョの奇妙な冒険という著名なアニメ・漫画作品のタイトルである．トピック 2 の単語は，配信者のグループ名や配信プラットフォーム名である．トピック 3 の単語は，ミュージックビデオなどのコンテンツを配信しているサイト名である．トピック 4 の単語は，ダンジョンズ&ドラゴンズやゼルダの伝説など特定のゲームをプレイし配信するプロジェクトである．トピック 5 の単語は，読書に関する Twitter 上のハッシュタグである．トピック 6 の単語は，オンラインマーケティングに関する情報を発信する際のハッシュタグである．トピック 7 の単語は，データセットの観測期間中に話題になったスポーツ大会の名前やアイスホッケーのチーム名である．

謝　辞

本研究は JSPS 科研費 18J20698 及び(A)17H01001 の助成を受けている．改訂に際して，査読者より大変有意義なコメントを頂戴した．

付　録

A. 条件付き事後分布の導出

3 節では，式(4.1)及び(4.2)で潜在コミュニティ及び潜在トピックの条件付き事後分布を導出した．これらの事後分布を得るためには，まず，コミュニティ分布，エッジ確率，トピック分布，単語分布の 4 つのパラメータについて，条件付き事後分布を導出する必要がある．事前分布との共役性に基づいて，これらの事後分布は以下のように導出される．

$$p(\eta_i \mid S, R, X, \gamma) = \frac{\Gamma\left(\sum_k N_{ik} + M_{ik} + \gamma_k\right)}{\prod_k \Gamma(N_{ik} + M_{ik} + \gamma_k)} \prod_{k=1}^{K} \eta_{ik}^{N_{ik}+M_{ik}+\gamma_k} \tag{A.1}$$

$$p(\psi_{ikk'} \mid A, S, R, \delta, \epsilon) = \frac{\Gamma(n_{ikk'}^{(+)} + n_{ikk'}^{(-)} + \delta_{kk'} + \epsilon_{kk'})}{\Gamma(n_{ikk'}^{(+)} + \delta_{kk'})\Gamma(n_{ikk'}^{(-)} + \epsilon_{kk'})} \times \psi_{ikk'}^{\mathbb{I}(a_{ij}=1)}(1 - \psi_{ikk'})^{\mathbb{I}(a_{ij}=0)} \tag{A.2}$$

$$p(\theta_k \mid X, Z, \alpha) = \frac{\Gamma\left(\sum_l M_{kl} + \alpha_l\right)}{\Pi_l \Gamma(M_{kl} + \alpha_l)} \prod_{l=1}^{L} \theta_{kl}^{M_{kl}+\alpha_l} \tag{A.3}$$

$$p(\phi_l \mid W, Z, \beta) = \frac{\Gamma\left(\sum_v M_{lv} + \beta_v\right)}{\Pi_v \Gamma(M_{lv} + \beta_v)} \prod_{v=1}^{V} \phi_{lv}^{M_{lv}+\beta_v}. \tag{A.4}$$

B. 提案モデルに対する WAIC の定義式

提案モデルに対する WAIC の定義式は以下の通りである．

$$WAIC = -2\sum_{i=1}^{D}(lpd^{(i)} - p_{waic}^{(i)}), \tag{B.1}$$

$$lpd^{(i)} = \log\left(\frac{1}{G}\sum_{g=b+1}^{G}\prod_{j=1}^{D} p(a_{ij} \mid H^{(g)}, \Psi_j^{(g)}) \prod_{m=1}^{M_i} p(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)})\right)$$

$$p_{waic}^{(i)} = \frac{G}{G-1}\left(\frac{1}{G}\sum_{g=b+1}^{G}\left(\sum_{j=1}^{D}\log p(a_{ij} \mid H^{(g)}, \Psi_j^{(g)})^2 + \sum_{m=1}^{M_i}\log p(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)})^2\right)\right.$$
$$\left. - \left(\frac{1}{G}\sum_{g=b+1}^{G}\left(\sum_{j=1}^{D}\log p(a_{ij} \mid H^{(g)}, \Psi_j^{(g)}) + \sum_{m=1}^{M_i}\log p(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)})\right)\right)^2\right).$$

ただし，$p(a_{ij} \mid H^{(g)}, \Psi_j^{(g)})$ と $p(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)})$ は，CGS によるサンプルのうち g 回目の繰り返しにおけるサンプルで推定したパラメータを用いて計算される尤度であり，以下で定義されている．

$$p(a_{ij} \mid H^{(g)}, \Psi_j^{(g)}) = \sum_{k=1}^{K}\sum_{k'=1}^{K} \eta_{ik} \cdot \eta_{jk'}^{(g)} \cdot \psi_{jkk'}^{(g)\mathbb{I}(a_{ij}=1)} \cdot (1-\psi_{jkk'})^{(g)\mathbb{I}(a_{ij}=0)} \tag{B.2}$$

$$p(w_{im} \mid H^{(g)}, \Theta^{(g)}, \Phi^{(g)}) = \sum_{k=1}^{K}\sum_{l=1}^{L} \eta_{ik}^{(g)} \cdot \theta_{kl}^{(g)} \cdot \phi_{lw_{im}}^{(g)}. \tag{B.3}$$

参 考 文 献

Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008). Mixed membership stochastic blockmodels, *Journal of Machine Learning Research*, **9**(SEP), 1981–2014.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of Machine Learning Research*, **3**(4-5), 993–1022.

Bouveyron, C., Latouche, P. and Zreik, R. (2018). The stochastic topic block model for the clustering of vertices in networks with textual edges, *Statistics and Computing*, **28**(1), 11–31.

Chang, J. and Blei, D. M. (2010). Hierarchical relational models for document networks, *The Annals of Applied Statistics*, **4**(1), 124–150.

Daudin, J.-J., Picard, F. and Robin, S. (2008). A mixture model for random graphs, *Statistics and Computing*, **18**(2), 173–183.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics, *Proceedings of the National Academy of Sciences*, **101**(Supplement 1), 5228–5235.

Handcock, M. S., Raftery, A. E. and Tantrum, J. M. (2007). Model-based clustering for social networks, *Journal of the Royal Statistical Society. Series A*: *Statistics in Society*, **170**(2), 301–354.

Hoff, P. D., Raftery, A. E. and Handcock, M. S. (2002). Latent space approaches to social network analysis, *Journal of the American Statistical Association*, **97**(460), 1090–1098.

Igarashi, M. and Terui, N. (2020). Characterization of topic-based online communities by combining network data and user generated content, *Statistics and Computing*, DOI: http://dx.doi.org/10.1007/s11222-020-09947-5.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks, *Physical Review E — Statistical, Nonlinear, and Soft Matter Physics*, **83**(1), 016107.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E. and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models, *Social Networks*, **31**(3), 204–213.

Latouche, P., Birmelé, E. and Ambroise, C. (2012). Variational Bayesian inference and complexity control for stochastic block models, *Statistical Modelling*: *An International Journal*, **12**(1), 93–115.

Liu, Y., Niculescu-Mizil, A. and Gryc, W. (2009). Topic-link LDA: Joint models of topic and author community, *Proceedings of the 26th International Conference on Machine Learning, ICML 2009*, 665–672.

Saldaña, D. F., Yu, Y. and Feng, Y. (2017). How many communities are there?, *Journal of Computational and Graphical Statistics*, **26**(1), 171–181.

Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure, *Journal of Classification*, **14**(1), 75–100.

Wang, Y. J. and Wong, G. Y. (1987). Stochastic blockmodels for directed graphs, *Journal of the American Statistical Association*, **82**(397), 8–19.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research*, **11**, 3571–3594.

Zhu, Y., Yan, X., Getoor, L. and Moore, C. (2013). Scalable text and link analysis with mixed-topic link models, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 473–481.

A Model for Social Network Analysis Considering Text Information on Social Media —Extended Model Considering Node Degree Heterogeneity—

Mirai Igarashi and Nobuhiko Terui

Graduate School of Economics and Management, Tohoku University

In the study of social networks, it has become increasingly important to consider not only network information but also text information generated by people on social media in order to deeply understand the community structure. By taking text information into account, it is possible to analyze social networks with complex community structures, which have multiple clusters of people depending on their interests in a single network with densely connected edges. In this study, we extend the existing model that simultaneously consider network and text information, and propose a model with degree heterogeneity to represent that the probability of edge generation varies for each node. In empirical analysis using Twitter dataset, we compare it with comparative models not using text information or degree heterogeneity, and show it's better predictive performance.

Key words: Social network analysis, community detection, text analysis, topic modeling, Bayes inference, node heterogeneity.

2018 年 3 月改訂

「統計数理」投稿規程

1. 「統計数理」は，統計科学の深化と発展，そして統計科学を通じた社会への貢献を目指すものである．投稿原稿は，統計科学に関連した内容を持つもので，和文の原稿に限る．
2. 投稿原稿は次の 6 種とする．
   a. 原著論文 (Paper)
      統計科学の発展に貢献すると考えられる研究結果.
   b. 総合報告 (Review Article)
      特定の主題に関する一連の研究およびその周辺領域の発展を著者の見解に従って総括的，かつ体系的に報告したもの.
   c. 研究ノート (Letter)
      研究速報，新しい発想，提言，問題提起，事例報告など研究上，記録にとどめておく価値があると認められるものや，既発表の論文等に対するコメントで，研究上，記録にとどめておく価値があると認められるもの.
   d. 研究詳解 (Research Review)
      特定の研究領域における理論的あるいは応用的成果を，最近の結果や知見を加えてわかりやすく説明したもの.
   e. 統計ソフトウェア (Statistical Software)
      有用な計算法や解析法に関する短いプログラムおよびサブルーチンのリスト，利用手引き，実行例など.
   f. 研究資料 (Research Archives)
      歴史的なデータ，入手困難なデータや統計的手法の比較検討のために有用なデータ，あるいは，歴史的文献の翻訳や解説など.
   いずれも原則として，未発表のものに限る.
3. 投稿された原稿は，編集委員会が選定・依頼した査読者の審査を経て，掲載の可否を決定する.
4. 投稿原稿は電子投稿査読システム https://www.editorialmanager.com/toukei/ より投稿するものとする．原稿は pdf ファイルとし，必要なフォントはすべて埋め込み，原稿全体を一つのファイルにまとめることとする．論文が採択になった場合，著者は最終稿のソースファイルとハードコピーを提出するものとする.
5. 著作権
   (1) 掲載される論文等の著作権はその採択をもって統計数理研究所に帰属するものとする．統計数理研究所は，紙媒体の「統計数理」のほか電子媒体などを通じて論文等を公表することができる．特別な事情がある場合は，著者と本編集委員会との間で協議の上措置する.
   (2) 投稿原稿の中で引用する文章や図表の著作権に関する問題は，著者の責任において処理する.
   (3) 著者が自分の論文等を複製，転載，翻訳，翻案等の形で利用するのは自由である．この場合，著者は掲載先に出典を明記する.
6. 原稿は次の執筆要項に従って作成する.

「統計数理」執筆要項

1. 原稿は A4 用紙に 1 行 36 字から 40 字で 1 行おき，1 頁あたり 22 行程度とする．原稿の長さは原則として表・図を含めて 30 頁相当以内とし，各ページにページ番号を付す．図表は別紙にまとめ，本文中には挿入箇所のみを指定する．LaTeX で原稿を作成する場合は，「統計数理」スタイルファイルの使用を推奨する.
   https://www.ism.ac.jp/editsec/toukei/
2. 原稿は以下の順に書くものとする.
   ［第 1 頁］ 標題，著者名，所属名，和文要旨 (500 字程度，文献の引用および数式は原則として避ける)，和文キーワード (6 語以内).
   ［第 2 頁］ 英語による標題，著者名，所属名，Abstract (450 ワード程度)，Key words (6 words and phrases 以内)．Abstract は，問題の所在と得られた結果等がそれだけで理解できるようなものとする.

［第 3 頁以降］

① 本文：章，節の番号は，第 1 章にあたるものは，“1.”, 第 1 章 第 1 節にあたるものは，“1.1” というようにつける．また，式の番号は，章ごとに (2.1), (2.2) のようにし，式の左側に配置する．

② 数式：数式は簡明さを心がけ，添字にさらに添字をつけるのはなるべく避ける．

③ 参考文献：書き方は本要項 第 4 項を参照.

④ 表：一枚の用紙に一つの表を書く．表の番号は論文中に現れる順に従って，表 1，表 2,... または，Table 1, Table 2,... のようにする．

⑤ 図：一枚の用紙に一つの図を描く．図はそのまま写真製版できる鮮明なものを用意する．大きさは印刷出来上がりの 1〜2 倍とし，トレースが必要な場合は原則として著者が行うものとする．図の番号は論文中に現れる順に従って，図 1，図 2,... または，Fig. 1, Fig. 2,... のようにする．図は原則としてモノクロ印刷とするが，カラー印刷を必要とする場合は編集委員会に相談すること．

⑥ 注：本文中の注釈は極力避ける．やむを得ず注釈をつける場合は脚注とせず，論文末尾に後注とする．後注は，順番に “1, 2,...” の番号を付け，本文中では上付きで示す.

3. 本文中での参考文献の引用は，著者名 (出版年) とする．たとえば，Efron (1982), 清水・湯浅 (1984), Cox and Snell (1981), 坂元 他 (2004), Nakano et al. (2000).

4. 参考文献の書き方

① 雑誌の場合：

著者名 (出版年)．標題，雑誌名，巻，ページ〔始–終〕. (雑誌名は省略しないものとする.)

【例】 Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically, *Journal of the American Statistical Association*, **68**, 361–368.

② 叢書の中の一巻の場合：

著者名 (出版年)．書名 (編集者名)，叢書名，発行所名，発行地名.

【例】 Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1983). *Akaike Information Criterion Statistics*, Mathematics and Its Applications, Reidel, Dordrecht.

③ 単行本等の場合：

著者名 (出版年)．書名，発行所名，発行地名.

【例】 Cressie, Noel (1993). *Statistics for Spatial Data*, Wiley, New York.

④ 編集書の中の一部の場合：

著者名 (出版年)．標題，編集書名 (編集者名)，巻，ページ，発行所名，発行地名.

【例】 Akaike, H. (1980). Likelihood and the Bayes procedure, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 143–166, University Press, Valencia, Spain.

なお，同じ著者によるものが同一年に複数個現れる場合には，(1980a), (1980b) などとして区別する．文献は，日本人も含め，著者名のアルファベット順に並べる.

5. 著者校正は原則として一回とする．その際，印刷上の誤り以外の字句や図版の訂正，挿入，削除等は原則として認めない.