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

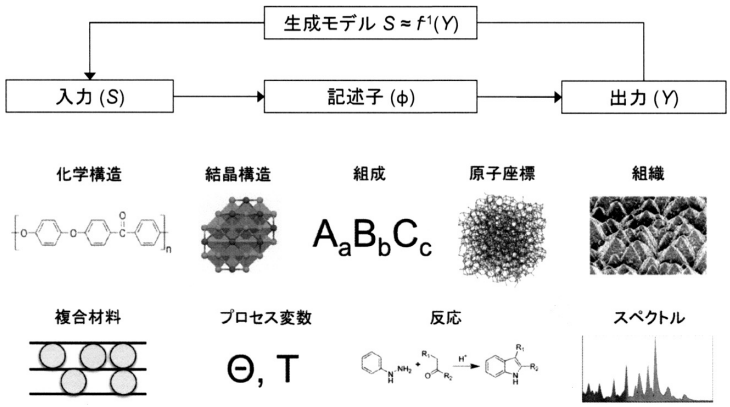


図 1. MI の基本的なワークフロー．入力 S(例えば，化学構造)から出力 Y(物性)の順方向の予測モデルを導き，出力 Y の目標値を近似的に達成する入力 S を逆向きに予測する．

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．

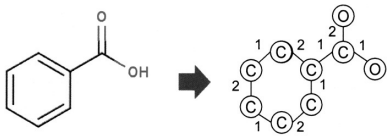
2. 物質・材料の表現

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.

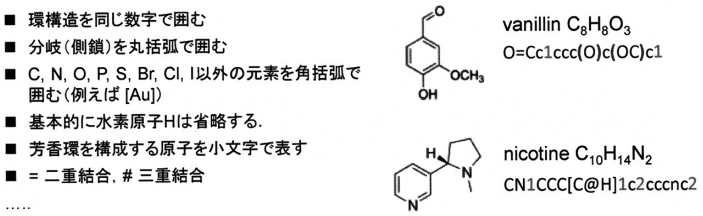


図 3. SMILES による化学構造の文字列表現.

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

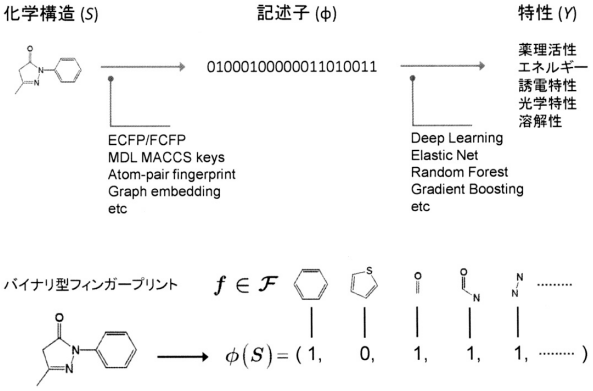


図 4. フィンガープリント記述子に基づく構造物性相関分析.

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

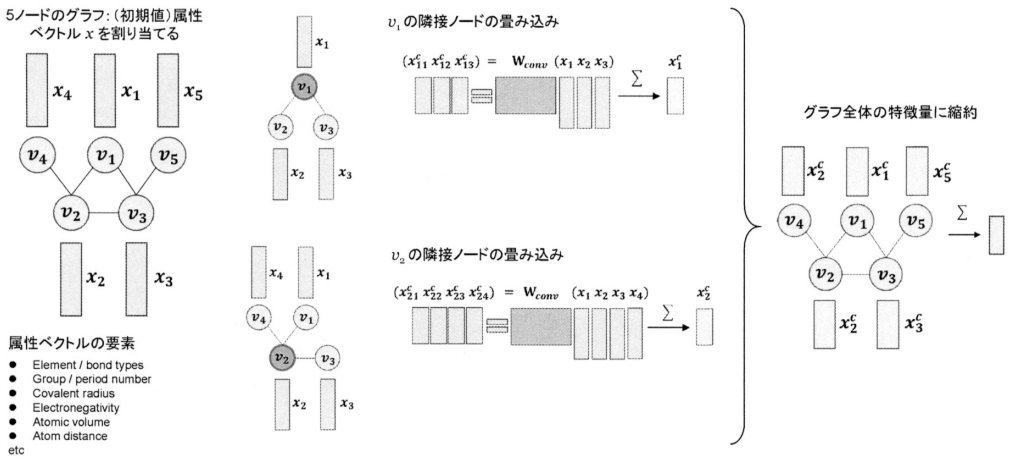


図 5．グラフニューラルネットワークの畳み込み層の計算．

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

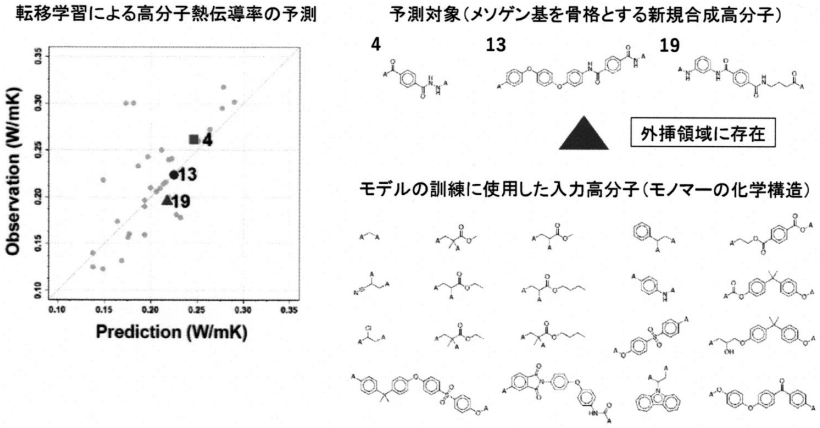


図 6. 転移学習による高分子熱伝導率の予測．左：3 種類の新規高分子に対する転移モデルの予測値と実測値．右：新規高分子のモノマーと転移学習で用いた訓練データの化学構造．

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

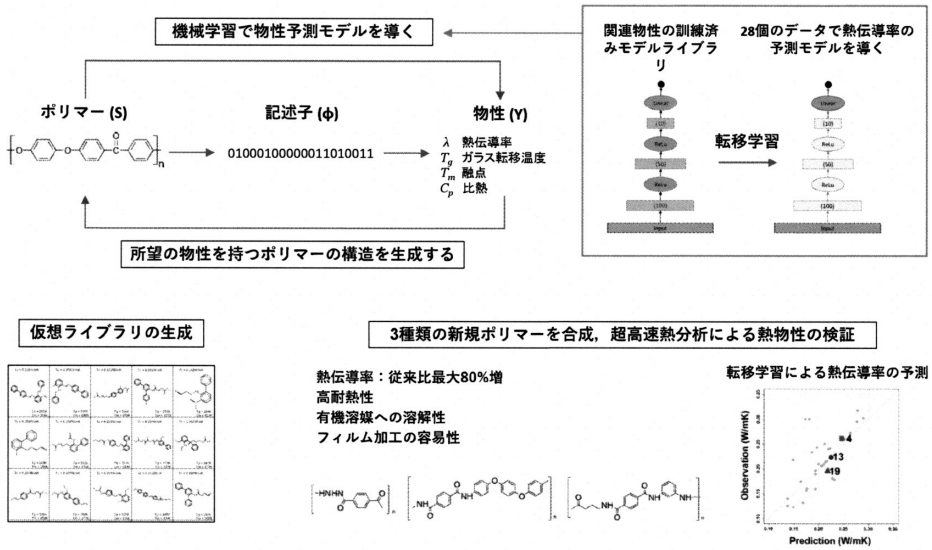


図 7．ベイズ推論に基づく高熱伝導性高分子の設計．

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，

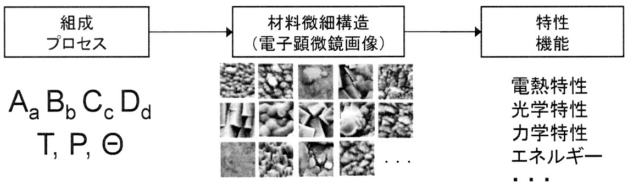


図 8. プロセスと組成からの材料微細組織の予測と組織から特性の予測．材料組織を電子顕微鏡画像で表現することで，前者のタスクは画像生成，後者のタスクは画像認識に帰着する．

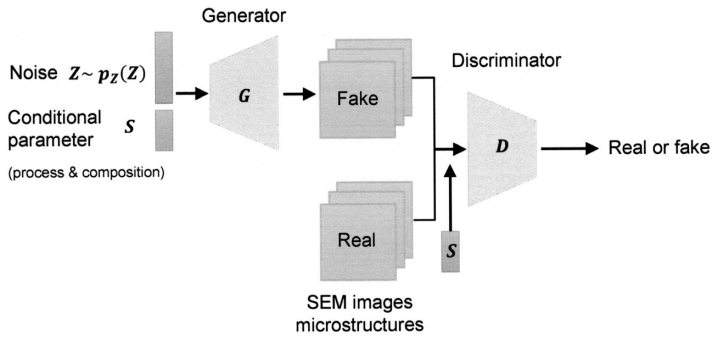


図 9. Conditional GAN のネットワークダイアグラム.

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

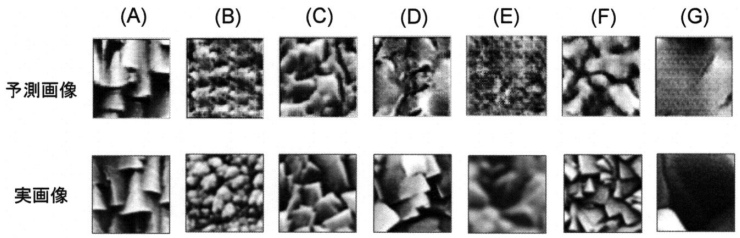
出力は材料組織の SEM 画像である．学習用の元データの数は 123 個である．ここから，128 ピクセル ×128 ピクセルの部分画像をランダムに 128 個抽出し，計 123×128 枚の画像をモデルの学習に使用する．

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，

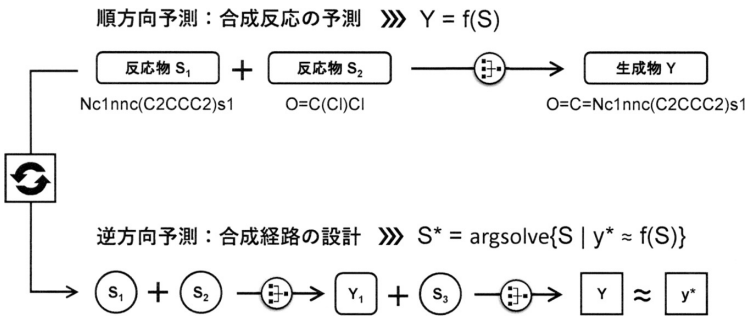


図 11．SMILES 表現に基づく合成反応予測と逆方向予測による合成経路の設計．

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

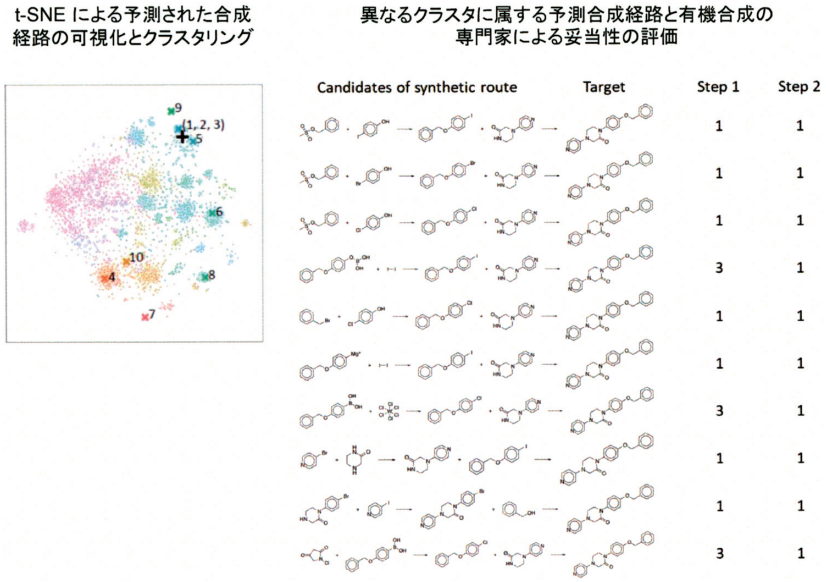


図 12. 機械学習による合成経路の設計．左：標的生成物に対する 6,613 個の 2 段階合成経路を予測し，t-SNE を用いて二次元空間にマッピングした結果．+ はデータベースにある既知の反応経路を示す．X-means クラスタリング (Dau Pelleg, 2000) で予測された経路を 98 個のグループに分類し，同定されたクラスタを色分けして表示している．シンボル × は、右図に示す 10 個の候補経路の位置を表す．右：標的生成物に対する 10 個の候補経路．反応の各ステップに対し，有機合成の専門家が三段階評価 (1: 可，2: 不明，3: 不可) を実施した．

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．

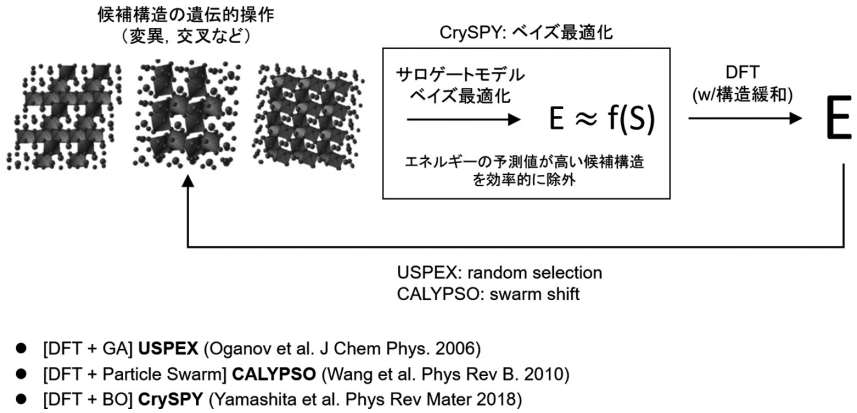


図 13．結晶構造探索アルゴリズム（USPEX, CALYPSO, CrySPY）のワークフロー．

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

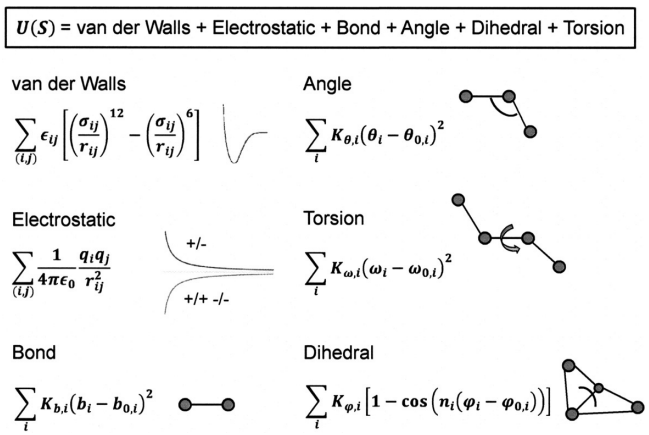


図 14. 古典分子動力学法のポテンシャル関数の例．

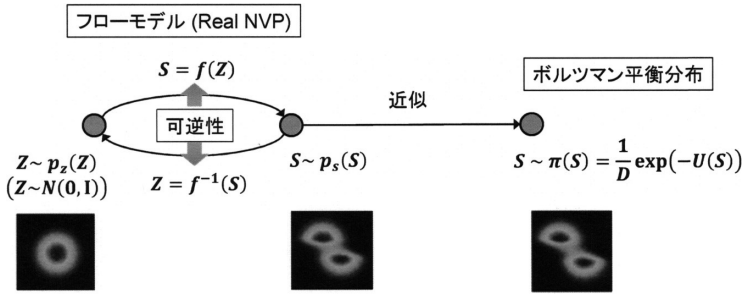


図 15. フローモデルによるボルツマン分布の近似.

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.