統計数理(2021)
第 69 巻 第 1 号 35–47
©2021 統計数理研究所

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 $\epsilon$ は平均 0，分散 $\tau$ 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 $\boldsymbol{\beta}$，$\tau$ 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 $\tau$ 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 $\tau$ 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 $\tau$ 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，$\tau$ 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，$\tau$ 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 $\nu_Y$ 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

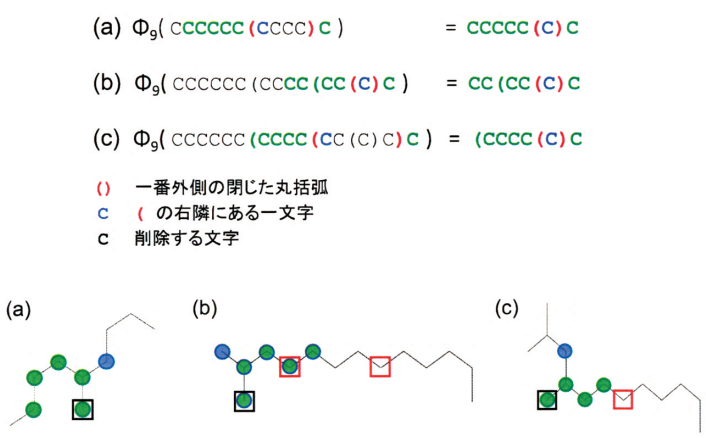


図 1. $\phi_{n-1}(\cdot)$ による部分文字列選択.

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 $\phi$ 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 $\kappa$(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 $\eta$(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

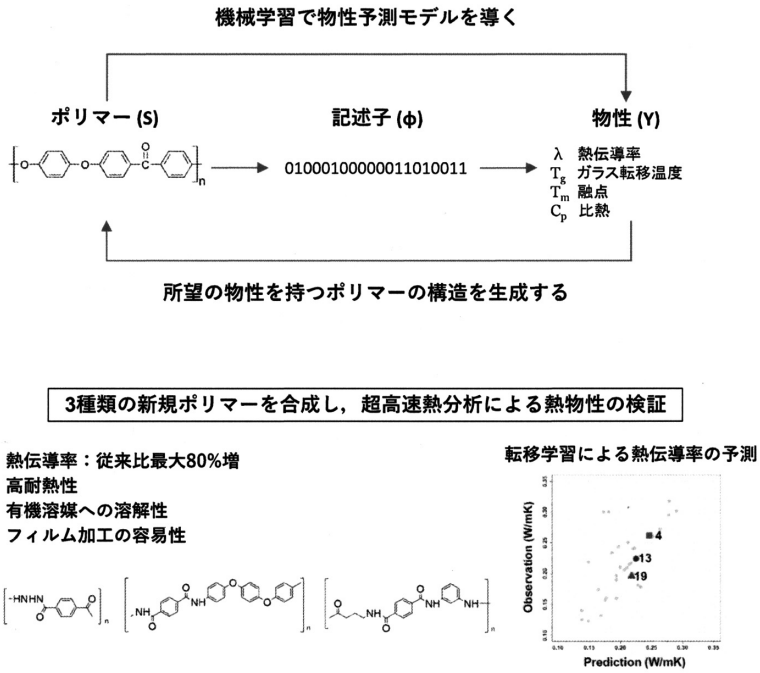


図 2．iQSPR-X を用いた高熱伝導率を持つ熱可塑性樹脂の発見．

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 ($\lambda$) | 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

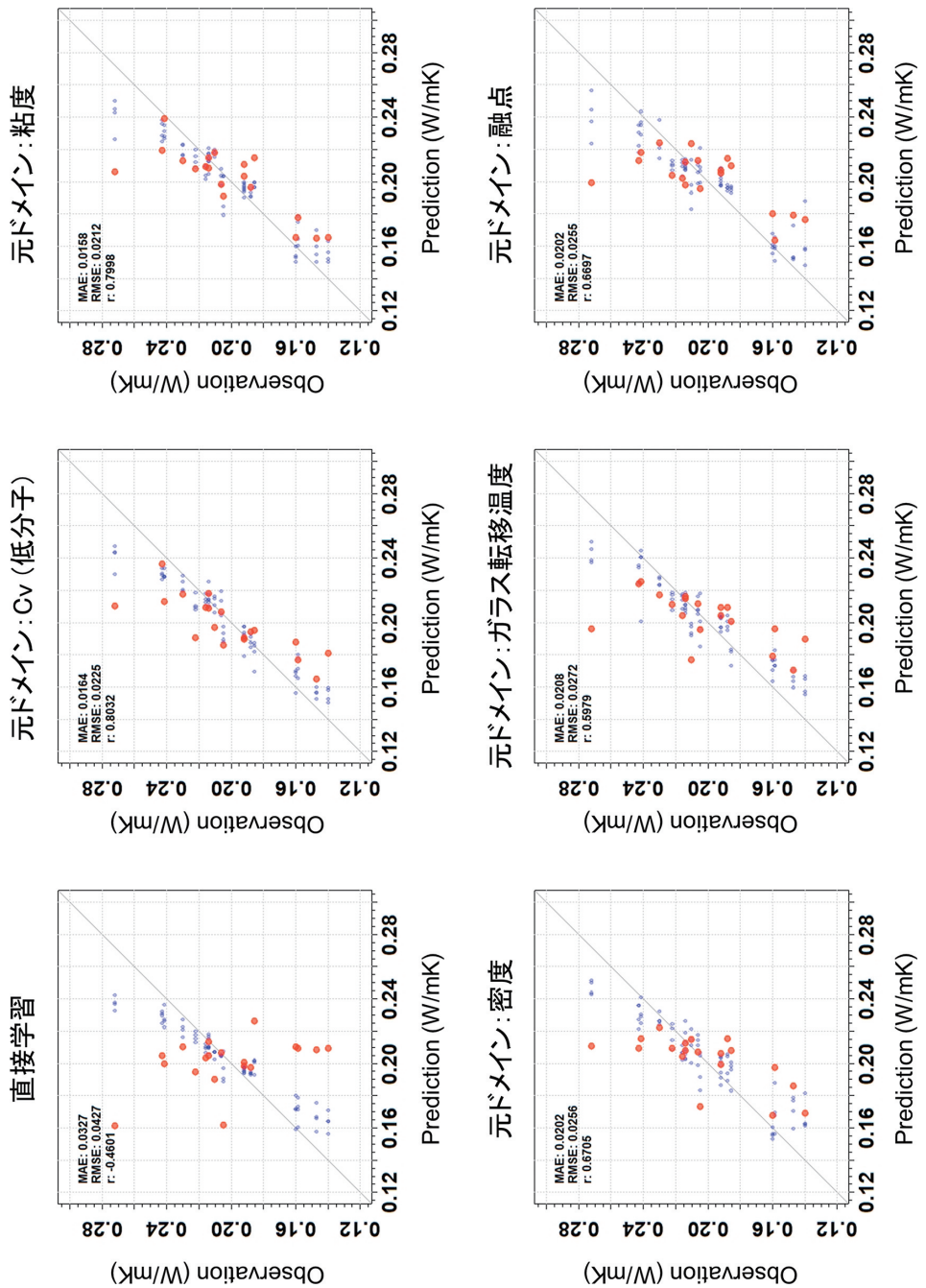


図 3．様々な元ドメイン（高分子のガラス転移温度，融点，密度，粘度，低分子化合物の定容比熱（Cv））から高分子熱伝導率への転移と転移学習を介さない通常の機械学習の 10 分割交差検証の結果．横軸は交差検証の予測値，縦軸は実測値．青点は訓練，赤点は交差検証の結果を表す．

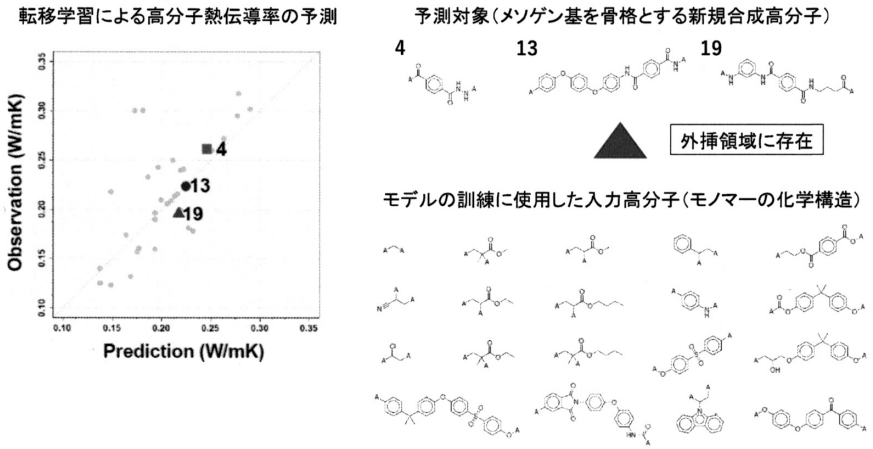


図 4. 左：3 種類の新規ポリマーに対する転移モデルの予測値と実測値．右：新規高分子のモノマーと転移学習で用いた訓練データの化学構造．

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.