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

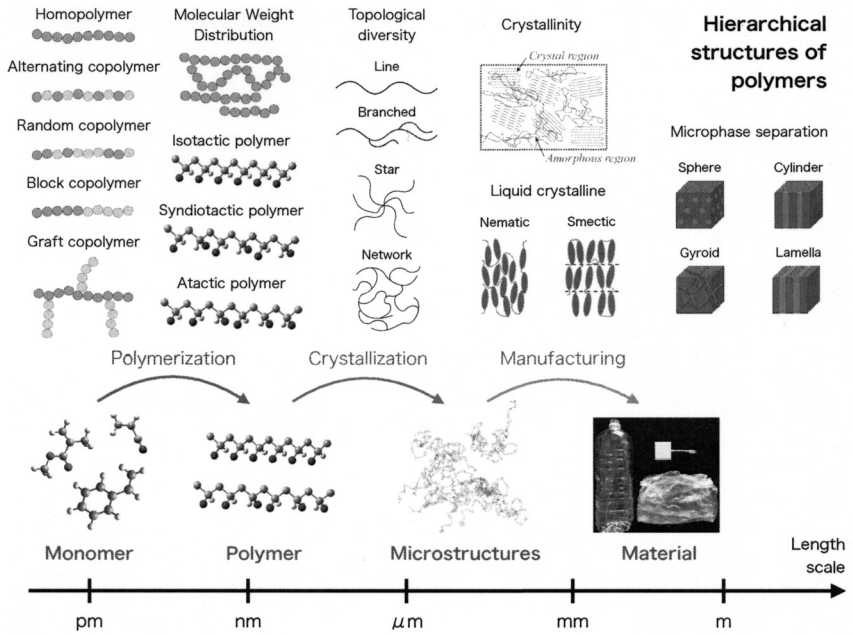


図 1．様々なスケールにおけるポリマーの構造．

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,

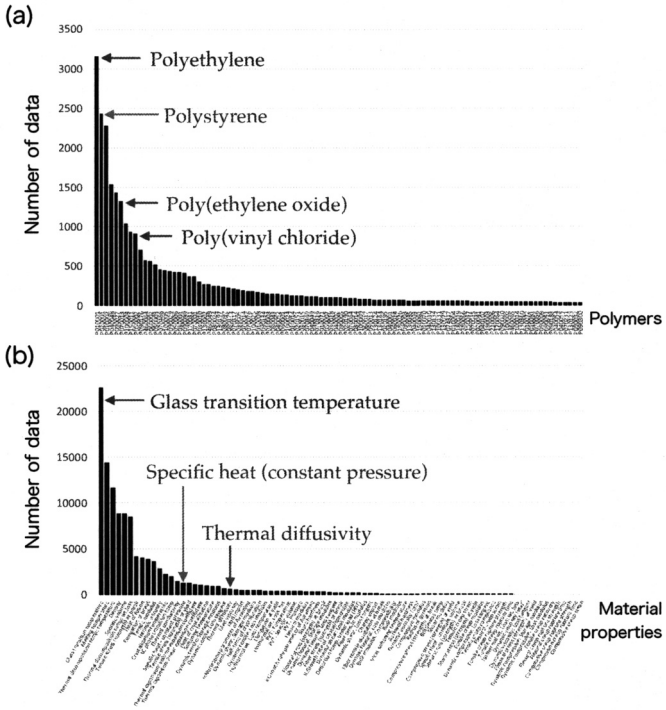


図 2. PoLyInfo の 54,151 件のデータの内訳(2016 年 4 月時点)．(a)データ数が多い上位 100 ポリマーの頻度を降順にプロットしている．(b)83 種類の熱物性のデータ数を降順にプロットしている．

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 $\chi$ 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

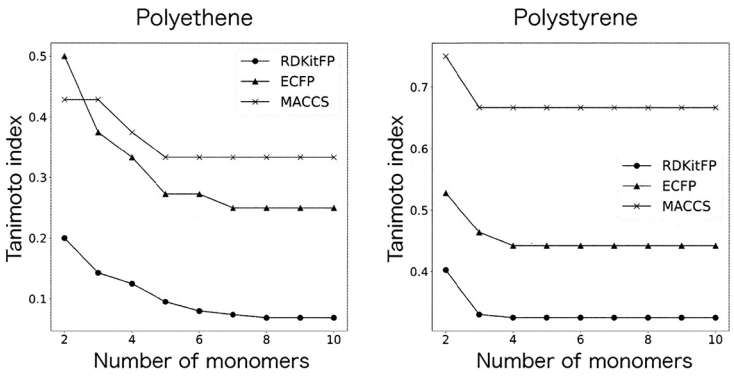


図 3. オリゴマー化におけるモノマー数を変えたときのフィンガープリントの変化．Tanimoto 類似度を用いてオリゴマーとモノマーのフィンガープリントの差を評価した．XenonPy に実装されている三つの RDKit フィンガープリントを使用した．“RDKitFP” は標準フィンガープリント，“ECFP” は Morgan フィンガープリント，“MACCS” は MACCS キーを表す．

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．$\Delta E$ は原子化エネルギー，$\epsilon_{gap}$ はバンドギャップ，$\kappa$ は誘電率，$\rho$ は密度，HOMO は最高占有分子軌道，LUMO は最低未占有分子軌道，$\epsilon_{opt}$ は光学的ギャップ，$\eta$ は屈折率，$\delta$ 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 |
|---|---|---|---|---|---|---|---|---|
| $\Delta E$ (Kim et al., 2018) | 392 | Mix | GP | CV-5 | 0.01 | 0.01 | 0.999 | eV/atom |
| $\epsilon_{gap}$ (Wu et al., 2016) | 155 | ICD | SVM | Split-20 | — | — | 0.88 | eV |
| $\epsilon_{gap}$ (Kim et al., 2018) | 382 | Mix | GP | CV-5 | 0.3 | 0.23 | 0.971 | eV |
| $\epsilon_{gap}$ (Jørgensen et al., 2018) | 3,989 | Str | VAE | CV-5 | — | 74 | — | meV |
| $\kappa$ (Wu et al., 2016) | 155 | ICD | SVM | Split-20 | — | — | 0.96 | — |
| $\kappa$ (Kim et al., 2018) | 384 | Mix | GP | CV-5 | 0.48 | 0.32 | 0.815 | — |
| $\rho$ (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 |
| $\epsilon_{opt}$ (Jørgensen et al., 2018) | 3,989 | Str | VAE | CV-5 | — | 70 | — | meV |
| $\eta$ (Kim et al., 2018) | 384 | Mix | GP | CV-5 | 0.08 | 0.05 | 0.892 | — |
| $\eta$ (Khan et al., 2018) | 221 | D&P | PLS | Split-30 | — | 0.004 | 0.899 | — |
| $\eta$ (Lightstone et al., 2020) | 527 | Mix | GP | Select-27 | 0.05 | — | 0.88 | — |
| $\delta$ (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

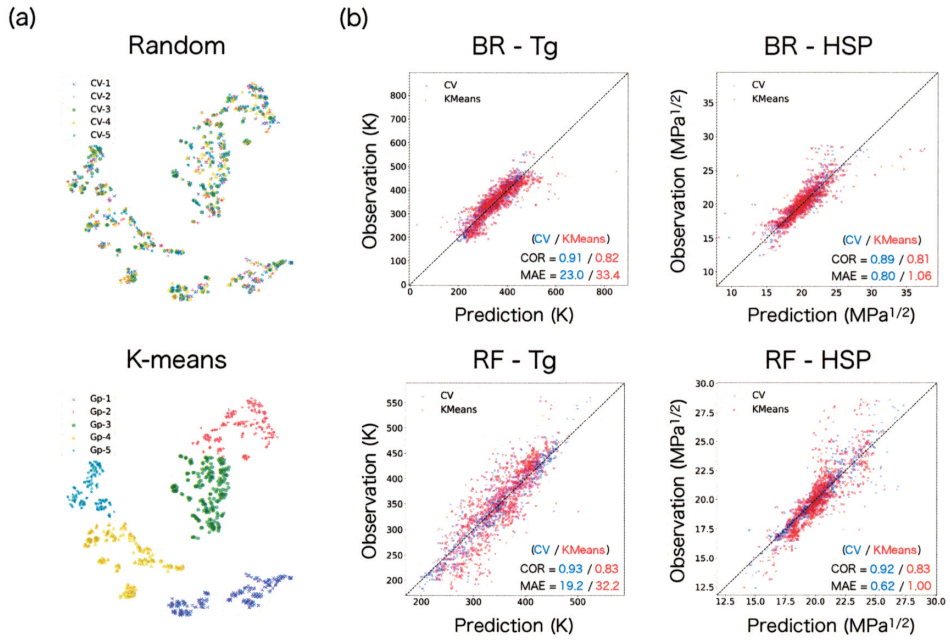


図 4. Polymer Genome (Kim et al., 2018)のデータを用いた機械学習モデルの外挿性能の検証．(a)t-SNE(perplexity=30)を適用して，RDKit の 200 次元記述子ベクトルを 2 次元空間に投影した．上図はこのデータセットをランダムに 5 分割した結果，下図は K-means による 5 分割の結果を表す．ガラス転移温度(Tg)と Hildebrand 溶解度パラメータ(HSP)の交差検証の際，分割したデータを訓練とテストデータセットに分配した．(b)ベイズリッジ回帰(BR)とランダムフォレスト(RF)に基づく Tg と HSP の予測値と観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