

解 説

高分子インフォマティクス

石切山 一彦

株式会社 東レリサーチセンター

(受取日: 2024 年 3 月 27 日, 受理日: 2024 年 5 月 9 日)

Polymer Informatics

Kazuhiko Ishikiriya

Toray Research Center, Inc.

(Received Mar. 27, 2024; Accepted May 9, 2024)

This study investigates quantitative structure-property relationships for polymers using machine learning techniques, leveraging the ATHAS data bank. Developed by Wunderlich et al., this data bank contains specific properties of polymers. These properties encompass essential information, including the glass transition temperature (T_g^0) of fully amorphous polymers, the heat capacity difference at T_g^0 , the equilibrium melting temperature, the heat of melting point at 100 % crystallinity, Θ temperature, the number of skeletal vibrations, and the temperature dependence of heat capacities, etc. Our methodology focuses on establishing correlations between these polymer-specific physical properties and the repeating structural units within the polymers. To achieve this, we employed various regression models, including artificial neural networks with equivalent and damping structures, Lasso, Random Forest, and XGBoost regression. Model validations were conducted using 5-Fold cross-validation. Our results indicate that the artificial neural network model exhibited the lowest error. Furthermore, we showcase the practical application of quantitative structure-property relationships in predicting the properties of two polymers not yet included in the ATHAS data bank.

Keywords: simplified molecular input line entry system, artificial neural network, quantitative structure-property relationships

1. はじめに

材料の研究開発では、物性・品質向上のため、種々の物性のトレードオフを解消し、物性発現メカニズムや物性低下・劣化等の本質原因究明が求められている。まずは、情報を整理し仮説を立て、その仮説に基づいて材料を試作し、仮説の妥当性を検証する。その結果に基づき仮説を修正して、再度妥当性を検証する、という PDCA サイクルを高速回転させることによってこれまで材料開発を加速させてきた (Fig.1)。その仮説検証型の課題解決プロセスでは、物性測定や分析評価がキーになることが多く、赤外分光、ラマン分光、核磁気共鳴等の分光分析をはじめ、X 線回折・散乱や、電子顕微鏡観察による微細構造分析・観察手法がよく活用されている。それらの手法により課題が解決すること多いが、ときには局所的情報のみに基づき仮説が構築され、必ずしもマクロな物性向上に直結しないことがある。一方、熱測定や熱分析等の熱力学的アプローチは、古典的手法とみられがちであるが、熱力学に反した事象は起こりえず、課題解決の重要なヒントが得られることも多い。

仮説検証型プロセスにより、本質的な原因究明がなされ、抜本的な課題解決は可能になるが、PDCA サイクルを回すのにこれまで時間がかかっていた。しかし最近、計算機の高性能化、大量のデータ保存、深層学習等のアルゴリズムに基づき、第3次人工知能 (Artificial Intelligence, AI) ブームが到来し、効率的な課題解決が図れる、第四のパラダイムと呼ばれるデータ駆動型科学に基づく新規な材料開発の手法であるマテリアルズ・インフォマティクス (MI) が登

場し、脚光を浴びている。^{1,2)}

MI の研究プロジェクトは、2011 年に米国の Materials Genome Initiative に始まり、³⁾ その後、2014 年にスイスの Materials Revolution Computational Design and Discovery of Novel Materials、⁴⁾ 2015 年に欧州連合の Novel Materials Discovery (NOMAD) Laboratory、⁵⁾ 国内でも 2014 年に戦略的イノベーション創造プログラム (SIP) マテリアルインテグレーション、⁶⁾ 2015 年に情報統合型物質・材料開発イニシアティブ、⁷⁾ 2018 年に統合型材料開発システムによるマテリアル革命、⁸⁾ 等が実施され、新機能材料の発見に繋がりはじめています。2023 年には GNoME という無機結晶の安定構造を生成するプログラムも開発されている。^{9,10)} また高分子の繰り返し構造単位に対して自然言語処理を適用した polyBERT が開発され、29 種の高分子物性を高精度かつ高速で予測できることが報告されている。¹¹⁾

量子化学に基づくシミュレーションは演繹的な方法であり、分子構造・組成から物性への順方向予測はできるが、その逆方向の予測、すなわち目標物性に対する分子構造・組成の予測は難しい。一方、MI は、機械学習により分子構造・組成と物性との相関 (定量的構造物性相関) を求めて解析する手法であり、順方向予測だけでなく、その逆方向予測も可能である (Fig.1)。本質原因を究明せずとも目標物性を有する分子構造・組成を予測することができることから、材料開発の効率化という観点で注目されている。

すでに低分子有機物については、量子化学データベースを用いて、機械学習により分子構造と物性との相関を得た結果が報告されている。¹²⁾ 高分子については、繰り返し高

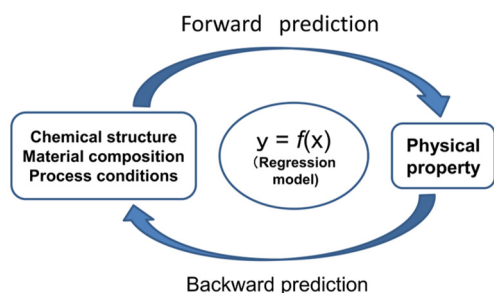


Fig.1 Conceptual diagram illustrating the principles of Materials Informatics.

分子構造単位のフィンガープリント（高分子ゲノム）を回帰モデルへの入力（説明変数）に用いた方法が報告されており、¹³⁻¹⁵ 高分子ゲノムから高分子物性を予測するプラットフォームが開発されている。³⁾ 高分子インフォマティクスでは説明変数をどうするかという課題はあるものの、すでに樹脂コンパウンド工程の条件最適化については企業で活用されはじめており、高熱伝導率の高分子構造を予測し、実際にその高分子を合成して物性検証した研究、¹⁶⁾ 固体高分子電解質、¹⁷⁾ 高屈折率高分子、¹⁸⁾ ガス分離高分子膜、¹⁹⁾ 新規網目状高分子の設計、²⁰⁾ 誘電率の周波数依存性予測、²¹⁾ 形状記憶高分子への適用、²²⁾ 等が報告されている。また、屈折率^{14,23)} や気体透過率、¹⁴⁾ ガラス転移温度、^{24,25)} 窒素中あるいは空気中での熱分解温度を機械学習させて高耐熱なポリイミドの分子構造を予測した研究成果²⁶⁾ が報告されている。

2. 高分子のデータベース

高分子インフォマティクスでは、回帰モデルの出力は物性値になるので、その出力値と期待値との誤差が最小になるように回帰モデルのパラメータを最適化する。そのため、期待値として与える物性データベース (DB) の選択が鍵となる。高分子の物性 DB はいくつか構築されている。¹⁴⁾ なかでも、18,958 以上 (2024 年 3 月現在) のホモポリマーの物性値が登録されている PoLyInfo がよく活用されている。²⁷⁾ ただし、ホモポリマーでも物性値に分布があり、例えば、ポリエチレンテレフタレートでは、ガラス転移温度 (T_g) は $-13^\circ\text{C} \sim 150^\circ\text{C}$ 、融点 (T_m) は $135^\circ\text{C} \sim 340^\circ\text{C}$ 、融解熱量 (ΔH_m) は $0 \sim 0.035 \text{ kcal g}^{-1}$ の値が登録されており、どの値を用いればよいのか迷うことがある。物性値の分布は、i) 分子量分布が異なる、ii) 結晶化度や中間相等の高次構造が異なる、iii) 添加剤の量や種類が異なる、iv) 結晶サイズや結晶の不完全性が異なる、などの高分子特有の理由によって生じているものと考えられる。

一方、高分子の DB として ATHAS (Advanced Thermal Analysis System) data bank がある。²⁸⁻³⁴⁾ これは Wunderlich らによって構築されたもので、文献値との整合性を調べ、実測も併用して注意深く物性値を取捨選択して収集された高品質な DB である。ATHAS data bank には、ポリオレフィン、アクリルポリマー、ポリエステル、ポリアミド、ポリオキサイド、ポリビニル、ポリアミノ酸、ポリイミド、イオウ含有ポリマー、ハロゲン元素含有ポリマー、フッ素系ポリマー、ポリシロキサン、ポリシラン等の他、低分子（炭化水素 C3~C70、ビスフェノール A、エポキシモノマー）も含む、合計約 200 の物質についてのデータが収録されている。³²⁾

一般に、高分子の融点は、Gibbs-Thomson 式にしたがい結晶サイズの依存性がみられ、しかも昇温中に再結晶化・

再組織化が生じて融点が上昇することから、高次構造だけでなく測定条件にも依存する。^{28,32,35)} そのためホモポリマーでもさまざまな値が観測される。これに対して、ATHAS data bank には、測定条件に依存しない高分子固有の平衡融点 (T_m^0) が収録されている。 T_m^0 は Hoffman-Weeks プロットや Gibbs-Thomson プロットから決定される。^{32,36)} また、高分子の T_g も結晶化度や冷却速度に依存するが、³²⁾ ATHAS data bank には、それらに依存しない、完全非晶物のガラス転移温度 (T_g^0) が収録されている。さらに結晶化度 100 % の融解熱量 (ΔH_m^0) や T_g^0 での熱容量差 (ΔC_p^0) が登録されている。

Fig.2(a) にポリエチレンテレフタレートの典型的な DSC 曲線を示す。ガラス転移領域から ΔC_p と融解領域から ΔH_m (DSC の昇温過程で冷結晶化がみられる場合にはその熱量を差し引いて補正) を読み取りプロットすると、**Fig.2(b)** (模式図) が得られる。結晶と非晶の 2 相モデルが成立する場合には、 ΔC_p は非晶量に比例し、 ΔH_m は結晶量に比例するので、結晶が多いほど非晶が少なくなることから、実測値は C、D、E のプロットのように直線上に乗る。これに対して、結晶、非晶、中間相（剛直非晶）の 3 相モデルが成立する場合には、F、G のプロットのように下に凸の曲線上に乗る。^{32,37)} したがって、 ΔC_p や ΔH_m は高分子の高次構造が関係しているため、高分子の繰り返し単位構造だけでは一義的には決まらず、C~G を使って機械学習させても予測誤差は大きくなる。高分子の高次構造は、機械特性をはじめ、他の高分子物性にも関係することから、C~G の試料の物性を機械学習しても、やはり予測誤差は大きくなることが予想される。

一方、ATHAS data bank には高分子固有の A 点 (ΔC_p^0) や B 点 (ΔH_m^0) の値が登録されているため、高次構造を考慮することなく、高分子の繰り返し単位構造と物性の関係を求める可能となる。

高分子の ΔH_m^0 と ΔC_p^0 が決定できれば、次式により結晶化度 (X_c)、可動非晶量 (X_{ma})、剛直非晶量 (X_{ra}) を求めることができる。^{28,32)}

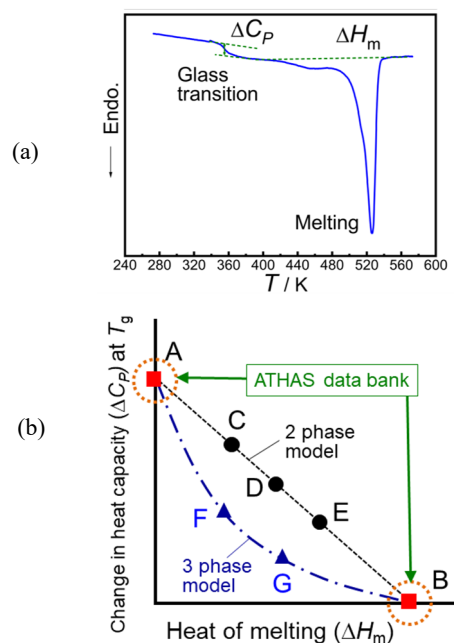


Fig.2 (a) A representative differential scanning calorimetry (DSC) heating trace for polyethylene terephthalate (PET); (b) Schematic depicting the relationship between enthalpy change (ΔH_m) and heat capacity change (ΔC_p) at the glass transition temperature (T_g) for a polymer.

$$X_c = 100 \times (\Delta H_m - \Delta H_{cc}) / \Delta H_m^0 \quad (1)$$

$$X_{ma} = 100 \times \Delta C_P / \Delta C_P^0 \quad (2)$$

$$X_{ra} = 100 - X_c - X_{ma} \quad (3)$$

これらは力学物性などにも関係するため,³⁷⁾ ΔH_m^0 と ΔC_P^0 の値は重要であるが、その決定には時間を要する。通常、様々な結晶化度の異なる試料について測定して、**Fig.2(b)**のプロットを行うと、2 相モデルなら直線関係になるので、その直線を外挿して求めた切片が ΔH_m^0 と ΔC_P^0 になる。結晶構造が未知の場合、第一原理に基づくシミュレーションでは ΔH_m^0 の算出は困難であるが、機械学習により短時間で予測可能になれば、課題解決のスピードアップが図れる可能性がある。

本稿では、ATHAS data bank を用いて高分子の定量的構造物性相関を求めた結果について報告する。以下では、**Fig.1** に示す、入力層、出力層、回帰モデルの順に述べる。回帰モデルについては、人工ニューラルネットワーク (Artificial Neural Network, 以下 ANN)、Lasso、XGBoost、Random Forest を用いて検討した。得られた最適モデルを用いて、ATHAS data bank に未登録の高分子に関して機械学習によって得られた予測値と文献値とを比較する。

3. ANN による定量的構造物性相関

Fig.1 の回帰モデルには、数値配列しか入力できないことから、高分子の繰り返し構造単位や化学組成は、フィンガープリントなどの数値配列に変換する必要がある。高分子インフォマティクスでは、繰り返し構造単位の他、オリゴマーの構造単位のフィンガープリントや、グラフィカルな 3 次元構造から求める記述子 (説明変数) が提案されている。^{14,25,38)} また、反応確率、モノマー組成、分子量分布を考慮した BigSMILES という記述子も検討されている。^{39,40)}

以下では、高分子の繰り返し構造単位のフィンガープリントを説明変数として、回帰モデルに入力した結果を示す。フィンガープリントとして、サーキュラー型の ECFP4 (結合直径 4 の Extended-Connectivity Fingerprints) と呼ばれるモルガンフィンガープリントを用いた。^{41,42)} それは、特定の原子に結合している 2 つ先までの原子を含む部分構造から分類するものである。**Fig.3** に、ポリプロピレン (PP) の部分構造の一例を示す。まず、繰り返し構造単位を SMILES (Simplified Molecular-Input Line-Entry System) 表記の文字列で記述し、接続部位を [*] で表すと、^{3,11,14,15)} PP の場合は [*]C(C)C[*] となる。その文字列から、Python のライブラリの RDKit⁴³⁾ を使うと、ハッシュ関数による識別値 (ID) が得られる。その識別値は部分構造の種類を数値化したもので、**Fig.3** に PP の一例を示す。

次に繰り返し構造単位と同じ部分構造 (ID) が何個あるかを数えて表にする (**Table 1**)。最後に、その個数のみを回帰モデルに入力して、機械学習させる。通常、ECFP4 のフィンガープリントとして、2 進数の 2048 桁あるいは 4096 桁の固定長がよく使われているが、情報欠落により予測精

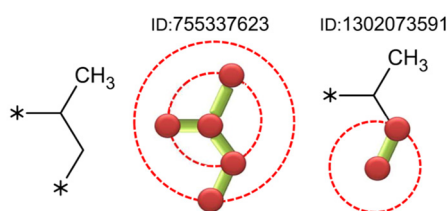


Fig.3 Depiction of the repeating structural unit of polypropylene (PP) and examples of the identification value (ID) and partial structure in PP fingerprint.

度が低下する。⁴⁴⁾ そこで、以下では情報欠落の少ない、**Table 1** に示す個数配列のフィンガープリントを用いて、ATHAS data bank に収録されている、 T_g^0 , ΔC_P^0 , T_m^0 , ΔH_m^0 , 絶対零度での残余エントロピー (S_0), 150~300 K の各温度における定積熱容量 (C_P), さらに式(4)で表される骨格振動の定積熱容量 (C_{skeletal}) に対する Tarasov 関数^{28,32,44,45)} の特性温度 (θ_1 , θ_3) と骨格振動の個数 N_s に対して機械学習を実施した。⁴⁴⁾ ただし、ATHAS data bank に登録されている低分子のデータは除外して、高分子のデータのみを用いた。

$$C_{\text{skeletal}} / N_s R = D_1(\theta_1/T) - (\theta_3/\theta_1)[D_1(\theta_3/T) - D_3(\theta_3/T)] \quad (4)$$

ここで、 T は温度、 R は気体定数、 D_i は i 次のデバイ関数である。

$$D_i(\theta_i/T) = i(T/\theta_i)^i \int_0^{\theta_i/T} \frac{(\theta/T)^{i+1} e^{\theta_i/T}}{[e^{(\theta/T)} - 1]^2} d(\theta/T) \quad (5)$$

ANN の隠れ層 (中間層) の構造によっては、物性予測誤差が異なる可能性があるため、今回、2 種類の隠れ層の構造について予測誤差を求め、比較した。一つは、隠れ層のノード数 (ニューロン数) を全て同じにした等価構造 (**Fig.4**),

Table 1 Number of identification values (IDs) in repeating structural units of polymers.

Abbreviations for polymer	SMILES	ID				...
		755337623	1302073591	1307459754	2245273601	
PP	[*]C(C)C[*]	1	1	1	1	...
PMMA	[*]C(C)C(C)[*]C(=O)OC	0	1	0	0	...
PMAA	[*]C(C)C(C)[*]C(=O)O	0	1	0	0	...
PMA	[*]C(C)[*]C(=O)OC	0	1	1	1	...
PIBMA	CC([*])C(C)[*]C(=O)OC(C)C	0	1	0	1	...
PIBA	CC(C)COC(=O)C([*])C[*]	0	1	1	2	...
...	...	...	...	...	...	...

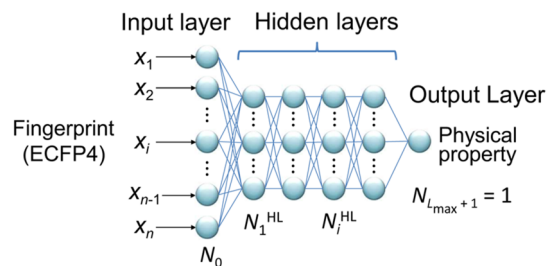


Fig.4 The artificial neural network (ANN)-equivalent structure of hidden layers in machine learning (ML) for modeling the physical properties of polymers. Here, L_{max} represents the total number of hidden layers, and N_i^{HL} represents the number of nodes in the i -th hidden layer.

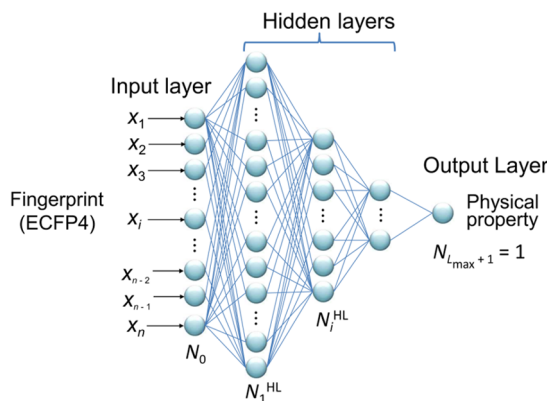


Fig.5 The ANN-damping structure of hidden layers in ML for the physical properties of polymers. Here, N_1^{HL} denotes the number of the first hidden layer.

もう一つは、1層目のノード数を最大とし、2層目以降は順次減少させるという減衰構造 (Fig.5) である。⁴⁴⁾ 等価構造 (Fig.4) と減衰構造 (Fig.5) の入力層に Table 1 の個数配列を入力して、出力層の期待値として、ATHAS data bank を与えて、機械学習における活性化関数の重みとバイアスを最適化した。

$$y_k^{m+1} = \varphi \left(\sum_{i=1}^{N_m^{\text{HL}}} w_i^m x_i^m + b_k^{m+1} \right) \quad (6)$$

ここで、 x_i^m は、 i 番目の隠れ層における m 番目のノードの値を表し、 w_i^m はその重み、 y_k^{m+1} は $m+1$ 番目の隠れ層における k 番目のノードの値、 b_k^{m+1} はそのバイアス、 N_m^{HL} は m 番目の隠れ層のノード数を表す。活性化関数 φ には、式(7)で示す非線形の Relu 関数を用いた。

$$\varphi(x) = \begin{cases} 0 & (x < 0) \\ x & (x \geq 0) \end{cases} \quad (7)$$

Figs.4, 5 に示す ANN の出力から各物性(i)に対する予測値 (P_i^{pred}) が得られるので、ATHAS data bank の登録値 (P_i^{ATHAS}) との、式(8)で示す平均二乗誤差 (MSE) が最小になるように、式(6)中の各パラメータを最適化した。

$$\text{MSE} = \sum_{i=1}^n (P_i^{\text{ATHAS}} - P_i^{\text{pred}})^2 / n \quad (8)$$

最適化の収束回数は 1000 とし、機械学習における分割計算のバッチ数は小さいほど、速く収束するが、計算結果が不安定になることから、事前にバッチ数の最適値 ($N_{\text{batch}}^{\text{opt}}$) を決定した (Table 2)。⁴⁴⁾

ここで、ATHAS data bank に収録されている物質の数は高々 200 程度と少ないことから、5-Fold cross-validation (CV) を適用した。まず、データを 5 分割する。各分割は Fold と呼ばれる。Fig.6 に示すように、1 つの Fold をテスト用、残りの 4 つの Fold を訓練 (学習) 用にする。訓練データで得られた回帰モデルを用いてテストデータに対して、式(8)の平方根から計算される Root Mean Squared Error (RMSE) を求める。同様に、テストデータがすべての Fold に適用されるように、訓練用とテスト用の各データを準備して、それぞれのデータセットに対する RMSE を算出した。

ANN の隠れ層の最適構造は以下の手順で決定した。まず、Figs.4, 5 で示す、各隠れ層の数 (L_{max}) と 1 層目のノード数 (N_1^{HL}) をメッシュ状に変え、その隠れ層の構造を用いて P_i^{pred} を算出し、RMSE を求め、その等高線図を作成して、その最小点から最適な隠れ層の構造を決定した。 T_g^0 の等価構造に対する RMSE の等高線図を Fig.7 に示す。これから T_g^0 の最適な等価構造は、 $N_1^{\text{HL}} = 400$, $L_{\text{max}} = 8$ と求まる。⁴⁴⁾ 同様な方法で、各物性に対する最適な等価構造と最適な減衰構造を求めた。その結果を Table 2 に示す。等価構造と減衰構造の最適な ANN の構造は物性ごとに全く異なるにもかかわらず、両者の誤差 (RMSE) は物性ごとにほぼ等しいことがわかる。すなわち、最適化さえすれば、Fig.4 の等価構造と Fig.5 の減衰構造のどちらの隠れ層の構造を用いてもよいことを表している。

最適化された等価構造を用いて物性を予測し、ATHAS data bank の登録値 (実測値) と比較した結果を Fig.8 に示す。各横軸は ATHAS data bank の登録値 (すなわち実測値)、各縦軸は機械学習による予測値である。Fig.8 中の ■印は訓練データ、○印はテストデータであり、 R^2 (train) と R^2 (test) は、それぞれ 5-Fold CV の訓練とテストのデータから式(9)によって得られた値である。

$$R^2 = 1 - \sum_{i=1}^n (P_i^{\text{ATHAS}} - P_i^{\text{pred}})^2 / \sum_{i=1}^n (P_i^{\text{ATHAS}} - \overline{P^{\text{ATHAS}}})^2 \quad (9)$$

ここで、 $\overline{P^{\text{ATHAS}}}$ は P_i^{ATHAS} の平均値、 n は登録数である。 Θ を除くと実測値と予測値はよく一致していることから、高

分子の繰り返し構造単位だけで、これらの物性が予測可能であることが示唆される。 Θ の R^2 が低いのは、測定難易度の高い 50 K 以下の熱容量から Θ が決定されるため、³²⁾ その値の不確かさが大きいことに起因しているものと推測される。ただし、100K 以上の実用温度域では、 Θ の値は熱容量の算出には影響しないので、支障はないであろう。 Θ と T_m^0 の R^2 がやや低いのも、測定の難易度が高いことに起因しているのかもしれない。

4. 回帰モデルによる比較

Table 2 には、Lasso, Random Forest, XGBoost の各回帰モデルを用いて 5-Fold CV による RMSE (平均値) を求めた結果を併記した。各回帰モデルでのハイパーパラメータの最適化には Optuna のライブラリを用いた。⁴⁶⁾ ほぼ全ての物性で、人工ニューラルネットワークが最も予測誤差が小さく、原理は単純であるが、効果が大いことがわかる。Fig.7 に示したような、隠れ層の構造の最適化には計算時間がかかるが、一旦最適な隠れ層の構造が求めれば、予測計算は瞬時に完了する。

高分子の物性予測として、Van Krevelen によって提唱されている原子団寄与法が広く知られている。⁴⁷⁾ これに対して、上記の ECFP4 を説明変数とした ANN による機械学習モデルは、500 種類以上の部分構造の寄与を考慮した非線形回帰モデルであり、Van Krevelen の方法よりも予測精度の向上が期待される。

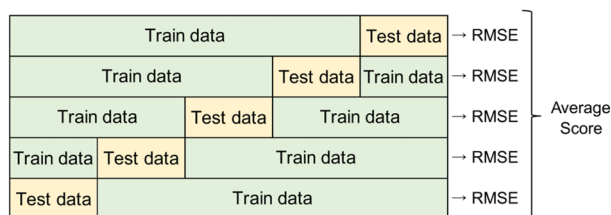


Fig.6 The conceptual diagram illustrating 5-Fold cross-validation.

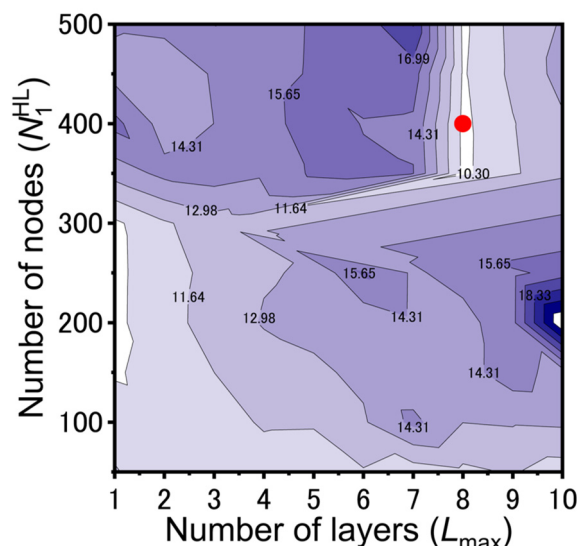


Fig.7 Contour plots illustrating the Root Mean Squared Error (RMSE) between the registered glass transition temperature (T_g^0) in the ATHAS data bank and the predicted values obtained using ANN-equivalent structures with respect to hidden layers. The closed circles on the plot indicate the minimum RMSE points.

Table 2 Comparison of RMSEs calculated from various regression models, including ANN-equivalent structures, ANN-damping structures, Lasso, XGBoost, and Random Forest. These RMSEs were obtained from 5-Fold cross-validation using count-based ECFP4 descriptors for the repeating polymeric structural units and data from the ATHAS database.

Physical property	Artificial neural network							Lasso	XGBboost	Random forest
	$N_{\text{batch}}^{\text{opt}}$	Equivalent structure			Damping structure					
		N_1^{HL}	L_{max}	RMSE	N_1^{HL}	L_{max}	RMSE			
T_{g}^0	40	400	8	10	500	1	10	61	36	33
ΔC_P^0	40	250	3	4	1000	2	4	13	12	13
T_{m}^0	40	50	8	34	500	4	37	75	55	59
ΔH_{m}^0	32	500	1	1	1500	1	1	5	7	7
θ_1	32	150	3	38	500	2	38	104	108	98
θ_3	40	150	5	23	500	4	23	27	22	23
S_{o}	24	450	1	2	750	2	4	6	8	11
N_{s}	32	50	1	1	250	1	1	4	4	5
C_{p} (150 K)	32	100	3	4	250	2	6	30	23	30
C_{p} (200 K)	32	200	1	4	250	5	5	34	30	38
C_{p} (250 K)	40	50	7	6	250	1	6	38	37	60
C_{p} (300 K)	32	400	1	6	500	1	6	46	42	66

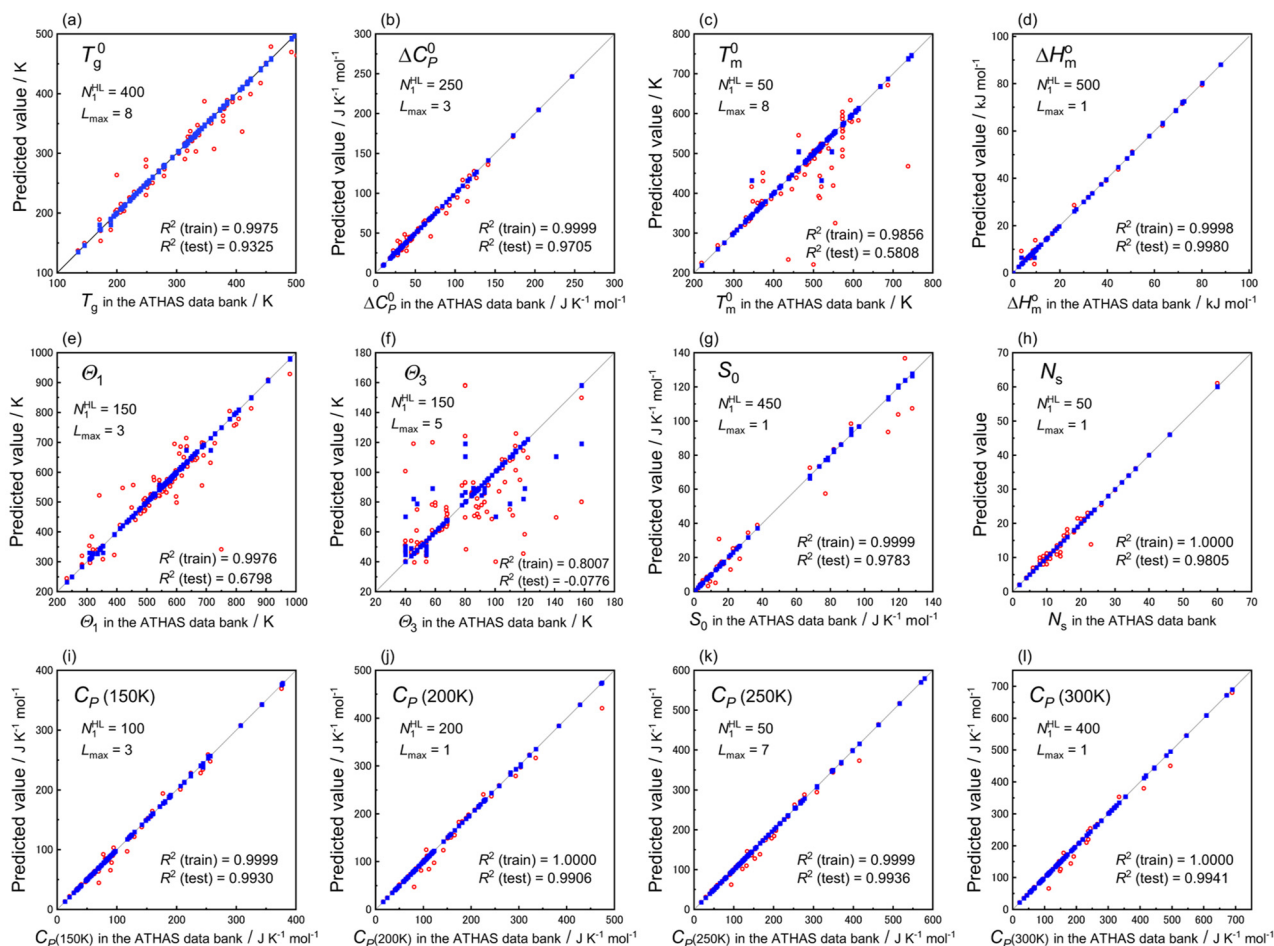


Fig.8 Relationships between registered properties of polymers in the ATHAS data bank and predicted properties using the optimal ANN-equivalent structures. (a) T_g^0 : Glass transition temperature for fully amorphous polymers. (b) ΔC_P^0 : Heat capacity difference at T_g^0 . (c) T_m^0 : Equilibrium melting temperature. (d) ΔH_m^0 : Heat of melting for 100%-crystallinity polymers. (e) Θ_1 : Characteristic temperature defined by Eq. (4). (f) Θ_3 : Characteristic temperature defined by Eq. (4). (g) S_0 : Residual entropy for fully amorphous polymers at absolute zero. (h) N_s : Number of skeletal vibrators. (i) C_P (150 K), (j) C_P (200 K), (k) C_P (250 K), and (l) C_P (300 K): Heat capacity at constant pressure at 150 , 200 , 250 , and 300 K, respectively. Blue closed squares and red open circles represent train and test data, respectively.

5. ATHAS data bank 未登録高分子への適用

次に, ATHAS data bankに未登録の2種類の高分子について, 3.で述べた方法により得られた最適な人工ニューラルネットワークの等価構造を用いて, 機械学習により予測値を求めて実測値と比較した。一つは外科用縫合糸に使われているpoly(p-dioxanone), PPDO ($[\text{CH}_2\text{CH}_2\text{OCH}_2\text{COO}]_n$)⁴⁸⁾であり, もう一つは生体適合性高分子として使われている, poly(N-vinylpyrrolidone), PVP ($[\text{C}_6\text{H}_9\text{NO}]_n$)⁴⁹⁾である。Fig.6で示すテストデータにそれらの繰り返し構造単位を与え, 5つのFoldで予測された物性値の平均値とその標準偏差を算出した。

PPDOの繰り返し構造単位から各物性値を予測した結果をTable 3とFig.9に示す。Fig.8(f)で示すように, R^2 が低い θ_3 を除くと, Table 3から機械学習による予測値は文献値(実測値)⁴⁸⁾と比較的よく一致する。 ΔH_m^0 については, Fig.8(d)で示す通り, R^2 は良好であるものの, 予測値の相対誤差は37%と大きい。 ΔH_m^0 の予測値を利用する場合には, このような予測誤差が存在することを考慮して解析する必要がある。同様にPPDOの定圧熱容量について予測した結果をFig.9に示す。機械学習による予測値($\bullet$, C_P^{ML})と文献値($C_P^{\text{cal, ref}}$, 青色実線)の相対誤差は, 50~350 Kでは1%以内, 400 K以上では2%以内で一致する。既報⁴⁸⁾では, 交差検証が不十分なため, 予測値と文献値の相対誤差が-8%~-18%と大きい結果が得られたが, 本稿では, 5-Fold CVにより, 既報よりも良好な結果が得られている。これは, ATHAS data bankにPPDOと類似の部分構造を有する, Polyoxy-ethylene, (POE, $[\text{CH}_2\text{CH}_2\text{O}]_n$)やPolyglycolic acid, (PGL, $[\text{CH}_2\text{COO}]_n$)の値が収録されており, ATHAS data bankの内挿によって予測が行われたためと推測される。

Fig.9には, 機械学習で予測された, Table 3に示す, θ_1 , θ_3 , T_m^0 の値を用いて振動解析⁴⁵⁾と組み合わせて C_P を算出した結果を▲印で示す。実用温度域では, $\pm 1\%$ 以下の相対誤差で熱容量を予測できている。 C_P^{cal} は, 通常, 熱容量の極低温測定を実施して, 式(4)に示すTarasov関数とグループ振動数に基づく熱容量解析により算出され, DSCのベースラインに相当する。⁴⁵⁾ 機械学習により, 極低温測定や振動解析をせずに, 熱容量の温度依存性が瞬時に予測できるのであれば, 新規高分子の解析の際に有用な情報となろう。

一方, PVPの予測結果をTable 4に示す。PVPでは, 予測値(C_P^{ML})と文献値(C_P^{cal})との相対誤差は-30%から-37%と, C_P^{ML} の方が明らかに低い。これは, PVPの部分構造であるピロリドン環を有する高分子がATHAS data bankに一つも登録されておらず, 機械学習による予測がデータの外挿になっているためと考えられる。すなわち, 機械学習の説明変数のフィンガープリントにピロリドン環構造が含まれていないため, その部分構造による熱容量寄与分が全く考慮されておらず, 熱容量が低く見積もられたと推測される。ATHAS data bankにPVPの文献値⁴⁹⁾を含めて機械学習すれば, ピロリドン環を有する類似高分子についてもPPDOと同様, 予測精度が向上するかもしれない。

6. おわりに

本稿では, 高分子インフォマティクスの基礎と応用について述べた。該手法は, 高分子の熱的性質の予測, 測定結果の妥当性検証, 効率的な材料スクリーニングなど, 高分子の研究開発に役立つものと思われる。

最近, Transformerのアルゴリズム⁵⁰⁾に基づく大規模言語モデルによる生成AIが登場し, その分野の進化・発展は著しい。ただし, 大規模言語モデルは, インターネット情報に基づき構築されており, 言語情報が主で, サイエンス

Table 3 Comparison of reported properties for poly(p-dioxanone)⁴⁸⁾ and those predicted by ML from 5-Fold cross validation using the optimal ANN-equivalent structures.

Physical property	Reported value	ANN-equivalent structure		
		Predicted value	Standard deviation	Relative error / %
T_g / K	264.0	252.9	1.3	-4.2
$\Delta C_P^0 / \text{J K}^{-1} \text{mol}^{-1}$	69.9	53.7	0.1	-23
T_m^0 / K	400.0	392.3	1.4	-1.9
$\Delta H_m^0 / \text{kJ mol}^{-1}$	14.4	19.7	0.0	37
θ_1 / K	478.7	487.2	1.8	1.8
θ_3 / K	50.4	109.6	0.4	117
$S_0 / \text{J K}^{-1} \text{mol}^{-1}$	13.4	13.7	0.2	2.0
N_s	10.0	9.9	0.0	-1.0
$C_P (150 \text{ K}) / \text{J K}^{-1} \text{mol}^{-1}$	74.7	74.5	0.3	-0.3
$C_P (200 \text{ K}) / \text{J K}^{-1} \text{mol}^{-1}$	91.5	91.5	0.2	0.0
$C_P (250 \text{ K}) / \text{J K}^{-1} \text{mol}^{-1}$	108.7	108.8	0.6	0.1
$C_P (300 \text{ K}) / \text{J K}^{-1} \text{mol}^{-1}$	126.8	126.0	0.1	-0.6

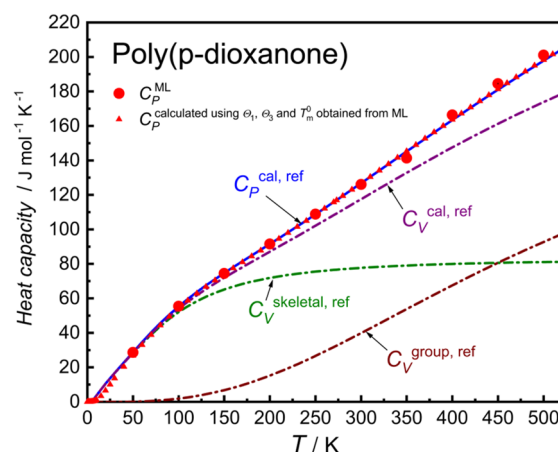


Fig.9 Comparison of heat capacities predicted by ML (C_P^{ML}) and reference data⁴⁸⁾ for solid poly(p-dioxanone). Each solid line at C_P^{ML} represents a range of standard deviations.

Table 4 Heat capacities (C_P) for poly(N-vinylpyrrolidone) as a function of temperature. C_P^{exp} represents the experimental heat capacity,⁴⁹⁾ while C_P^{cal} s were obtained based on vibrational analysis reported in the same reference. C_P^{ML} s were derived from ML using the optimal ANN-equivalent hidden-layer structures via 5-Fold cross-validation.

T K	C_P^{exp} $\text{J K}^{-1} \text{mol}^{-1}$	C_P^{cal} $\text{J K}^{-1} \text{mol}^{-1}$	C_P^{ML} $\text{J K}^{-1} \text{mol}^{-1}$	Standard deviation $\text{J K}^{-1} \text{mol}^{-1}$	Relative error %
50.0	31.34	29.51	19.91	1.20	-32.5
100.0	56.52	54.87	38.67	1.38	-29.5
150.0	76.18	74.32	49.59	3.12	-33.3
200.0	94.65	93.45	60.79	1.01	-35.0
250.0	114.45	114.13	74.48	7.36	-34.7
300.0	135.04	136.32	87.45	1.03	-35.8
350.0	156.95	159.17	100.57	1.05	-36.8
400.0	180.83	181.73	123.28	6.13	-32.2
450.0	—	203.37	138.78	1.48	-31.8
500.0	—	223.78	153.61	1.56	-31.4

のデータが少なく, 情報源に偏りがあると言われている。そのため, 数値情報を含むマルチモーダル基盤モデルの構築が検討されはじめています。今まで以上に物性値を含むデータ収集・実測が必要とされている。機械学習による予測は, 統計的な, 確率的手法に基づくため, 間違える可能性

があり、検証のためにも実測は必要になっている。

高分子物性には特有の課題があり、単純な機械学習では良好な結果は得られない。データサイエンティストだけでなく、熱測定、熱分析、高分子物性の専門家が歩み寄り、この分野を切り拓いてゆく必要があろう。

文 献

- 1) J. M. Rickman, T. Lookman, and S. V. Kalinin, *Acta Mater.* **168**, 473–510 (2019).
- 2) K. M. Tolle, D. S. W. Tansley, and A. J. G. Hey, in *The fourth Paradigm: Data-intensive scientific discovery* vol.99 1334–1337 (2011).
- 3) C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad, *J. Phys. Chem. C* **122**, 17575–17585 (2018).
- 4) *Materials Revolution Computational Design and Discovery of Novel Materials, MARVEL*, <https://nccr-marvel.ch/> (2014).
- 5) *Novel Materials Discovery (NOMAD)*, <https://www.go-fair.org/implementation-networks/overview/nomad/> (2015).
- 6) *Materials Integration* Preprint at <https://www.jst.go.jp/sip/k03/sm4i/en/project/project-d.html> (2014).
- 7) *Material Research by Information Integration Initiative, MI2I* <https://www.nims.go.jp/MI2I-I/en/> (2015).
- 8) *Materials Integration for Revolutionary Design System of Structural Materials* at <https://www.jst.go.jp/sip/en/p05/index.html> (2023).
- 9) A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, *Nature* **624**, 80–85 (2023).
- 10) N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng, et al., *Nature* **624**, 86–91 (2023).
- 11) C. Kuenneth, and R. Ramprasad, *Nat. Commun.* **14**, 4099 (2023).
- 12) G. A. Pinheiro, J. Mucelini, M. D. Soares, R. C. Prati, J. L. F. Da Silva, and M. G. Quiles, *J. Phys. Chem. A* **124**, 9854–9866 (2020).
- 13) F. Cravero, S. A. Schustik, M. J. Martínez, C. D. Barranco, M. F. Díaz, and I. Ponzoni, *Chemometrics Intellig. Lab. Syst.* **191**, 65–72 (2019).
- 14) L. Chen, G. Pilania, R. Batra, T. D. Huan, C. Kim, C. Kuenneth, and R. Ramprasad, *Mater. Sci. Eng. R Rep.* **144**, 100595 (2021).
- 15) H. Doan Tran, C. Kim, L. Chen, A. Chandrasekaran, R. Batra, S. Venkatram, D. Kamal, J. P. Lightstone, R. Gurnani, P. Shetty, M. Ramprasad, J. Laws, M. Shelton, and R. Ramprasad, *J. Appl. Phys.* **128**, (2020).
- 16) S. Wu, Y. Kondo, M. aki Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa, and R. Yoshida, *npj Computational Materials* **5**, 1–33 (2019).
- 17) M. Liu, C. Clement, K. Liu, X. Wang, and T. D. Sparks, *Comput. Mater. Sci.* **187**, 110108 (2021).
- 18) V. Venkatraman, and B. K. Alsberg, *Polymers* **10**, 1–16 (2018).
- 19) J. W. Barnett, C. R. Bilchak, Y. Wang, B. C. Benicewicz, L. A. Murdock, T. Bereau, and S. K. Kumar, *Sci. Adv.* **6**, 1–8 (2020).
- 20) S. Shibayama, and K. Funatsu, *Bull. Chem. Soc. Jpn.* **94**, 112–121 (2021).
- 21) L. Chen, C. Kim, R. Batra, J. P. Lightstone, C. Wu, Z. Li, A. A. Deshmukh, Y. Wang, H. D. Tran, P. Vashishta, G. A. Sotzing, Y. Cao, and R. Ramprasad, *npj Computational Materials* **6**, 30–32 (2020).
- 22) A. Shafe, C. D. Wick, A. J. Peters, X. Liu, and G. Li, *Polymer* **242**, 124577 (2022).
- 23) S. A. Schustik, F. Cravero, I. Ponzoni, and M. F. Díaz, *Comput. Mater. Sci.* **194**, 110460 (2021).
- 24) G. Chen, L. Tao, and Y. Li, *Polymers* **13**, (2021).
- 25) L. Tao, V. Varshney, and Y. Li, *J. Chem. Inf. Model.* **61**, 5395–5413 (2021).
- 26) M. Yu, Y. Shi, X. Liu, Q. Jia, Q. Wang, Z. -H. Luo, F. Yan, and Y. -N. Zhou, *Chem. Eng. J.* **465**, 142768 (2023).
- 27) S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, and M. Yamazaki, *Proceedings – 2011 International Conference on Emerging Intelligent Data and Web Technologies, EIDWT 2011* 22–29 (2011) doi:10.1109/EIDWT.2011.13.
- 28) B. Wunderlich, *Thermal Analysis. Thermal Analysis*, (Academic Press, Inc., 1990).
- 29) M. Varma-Nair and B. Wunderlich, *J. Phys. Chem. Ref. Data* **20**, 349–404 (1991).
- 30) B. Wunderlich, *J. Macromol. Sci. Part A Pure Appl. Chem.* **67**, 1019–1026 (1995).
- 31) B. Wunderlich, *Thermochim. Acta* **300**, 43–65 (1997).
- 32) B. Wunderlich, *Thermal analysis of polymeric materials. Thermal analysis of polymeric materials* 1–894 (Springer Berlin Heidelberg, 2005). doi:10.1007/b137476.
- 33) M. Pyda, *The Advanced THERMAL Analysis System (ATHAS) Databank – Polymer Thermodynamics* Preprint at <https://materials.springer.com/> (2014).
- 34) J. Pionteck, and M. Pyda, *Part 2: Thermodynamic Properties–pVT–Data and Thermal Properties, Subvolume A: Polymer Solids and Polymer Melts. Part 2: Thermodynamic Properties–pVT–Data and Thermal Properties, Subvolume A: Polymer Solids and Polymer Melts* (Springer, 2014).
- 35) M. Todoki and T. Kawaguchi, *J Polym Sci Polym Phys Ed* **15**, 1507–1520 (1977).
- 36) A. Toda, K. Taguchi, and K. Nozaki, *Cryst. Growth Des.* **19**, 2493–2502 (2019).
- 37) K. Ishikiriyama, *Netsu Sokutei* **46**, 155–162 (2019).
- 38) M. Aldeghi and C. W. Coley, *Chem. Sci.* **13**, 10486–10498 (2022).
- 39) M. E. Deagen, B. Dalle-Cort, N. J. Rebello, T. -S. Lin, D. J. Walsh, and B. D. Olsen, *Macromolecules* **57**, 42–53 (2024).
- 40) L. Schneider, D. Walsh, B. Olsen, and J. de Pablo, *Digital Discovery* **3**, 51–61 (2024).
- 41) D. Rogers and M. Hahn, *J. Chem. Inf. Model.* **50**, 742–754 (2010).
- 42) T. Minami and Y. Okuno, *MRS Advances* **3**, 2975–2980 (2018).
- 43) *The RDKit Documentation*, <https://www.rdkit.org/docs/> (2023).
- 44) K. Ishikiriyama, *Thermochim. Acta* **708**, 179135 (2022).
- 45) K. Ishikiriyama, *Netsu Sokutei* **48**, 114–121 (2021).
- 46) T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, *Optuna: A next-generation hyperparameter optimization framework*, <https://github.com/pfnet/optuna/> (2019).
- 47) D. W. Van Krevelen, *Properties of Polymers: Fourth Edition. Properties of Polymers: Fourth Edition* (Elsevier, 2009).
- 48) K. Ishikiriyama, M. Pyda, G. Zhang, T. Forschner, J. Grebowicz, and B. Wunderlich, *J. Macromol. Sci. Pt. B: Phys.* **37**, 27–44 (1998).
- 49) K. Ishikiriyama, K. Kondo, Y. Miyazaki, Y. Sasada, K. Sawada, R. Endo, N. Man, M. Nakano, and Y. Nakazawa, *Thermochim. Acta* **722**, 179456 (2023).
- 50) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Adv. Neural Inf. Process. Syst.* **30**, (2017).



石切山 一彦
Kazuhiko Ishikiriyama
E-mail: kazuhiko.ishikiriyama.s8@trc.toray