入門講座

インフォマティクス入門－13

MIPHA and shinyMIPHA for Use in Materials Characterization

材料組織・特性解析に用いる材料情報統合システム

名古屋大学 大学院工学研究科
材料デザイン工学専攻
助教 Zhi-Lei Wang

名古屋大学 大学院工学研究科
材料デザイン工学専攻
教授 足立吉隆
Yoshitaka Adachi

名古屋大学 大学院工学研究科
材料デザイン工学専攻
講師 小川登志男
Toshio Ogawa

Abstract:

As the data of materials science is rapidly increasing yearly, the data source has changed from the conventional paper-based to online-based. Under such an environment, machine learning is drawing increasing attention for finding certain rules from data in complex systems. As a result, materials informatics is proposed in materials research field. Similar with the traditional mathematics, physics, and chemistry, machine learning is a basic discipline that researchers need to master in the future. Especially for the materials researchers, it is a net increase. In order to lower the learning hurdle as much as possible and actually utilize machine learning by materials researchers, machine-learning systems with excellent operability are necessary. Recently, programming languages such as Python and R that can perform machine learning easily have appeared, which makes machine learning familiar to materials researcher.

This paper introduces two materials informatics integration systems, called MIPHA and shinyMIPHA, which can perform various image processing and machine learning at a practical level. In detail, object detection, 2D/3D feature extraction, mathematical feature extraction, sparse study, direct analysis, and inverse analysis will be described for demonstrating the two systems.

要旨

材料科学のデータが年々急激に増加しており、そのデータの提供方法も従来の紙ベースからオンラインに変更されつつある。この環境の中で、複雑系におけるデータから一定のルールを見出す手法として機械学習に注目が集まっている。この研究分野は今日ではマテリアルズインフォマティクスと呼ばれている。機械学習は、従来の数学、物理、化学と同様に、今後研究者が修得しておく必要がある基礎学問と言え、材料工学を学ぶ者にとってはこれを修得することは純増である。その学習のハードルを少しでも下げ、実際に材料研究者が活用するためには、操作性に優れるシステムが用意されていることが必要と思われる。昨今ではpythonやRといった比較的容易に機械学習が行えるプログラミング言語が登場しており、材料研究者にも機械学習が身近になってきている。

本稿では、実用レベルで様々な画像処理、機械学習を行える二つの材料情報統合システムMIPHAとshinyMIPHAについて紹介する。具体的には、画像の物体検出、2D3D特徴量抽出、数学的特徴量抽出、スパース学習、順解析、逆解析を行うシステムについてその特徴を説明する。

1 Introduction

Scientific data is doubly increased every year, which drives the evolution of scientific methods from traditional paper notebooks toward enormous online databases. As data volumes increase, the ability to efficiently extract knowledge from the huge amount of data becomes increasingly important. Machine learning, which is an artificial intelligence approach to analyzing data and making predictions and decisions based on a huge data volume through various models and algorithms, has already been successfully applied in many scientific fields.

Because of the staggering compositional and configurational degrees of freedom in materials, the chemical space of materials is far from being exhausted; an enormous number of new materials with useful properties are yet to be discovered. Therefore, machine learning is now attracting increasing attention in the materials research field to explore unknown information about materials and thus accelerate advances in materials discovery. One proposed approach is known as materials informatics, which is scientific and technical and seeks to establish processing–structure–property relationships in a high-throughput, statistically robust, and physically meaningful manner using computational science.

This paper presents two independently developed machine learning tools involved in the previous papers[1-3] called Materials Genome Integration System Phase and Property Analysis (MIPHA) and shiny Materials Genome Integration System Phase and Property Analysis (shinyMIPHA). The frameworks, characteristics, and functions of MIPHA and shinyMIPHA as well as their applications in materials characterization are demonstrated in this paper.

2 MIPHA[4] and shinyMIPHA[5]

2.1 Framework of MIPHA

Fig.1 shows the of functions and characteristics of MIPHA, integrating image recognition, image processing, 2D/3D microstructure analysis, and direct and inverse analyses. In terms of microstructure analysis, MIPHA focuses materials' metallurgical feature with image-engineering-based machine learning approach, where deep learning and Trainable Weka Segmentation (TWS) techniques are installed for image recognition and processing functions, respectively[1]. An artificial neural network (ANN) and genetic algorithm (GA) are prepared in direct and inverse analyses, respectively, used for property predictions and inverse design[3].

2.2 Framework of shinyMIPHA

Fig.2 shows the framework of shinyMIPHA, which is summarized as five function divisions: image analysis,

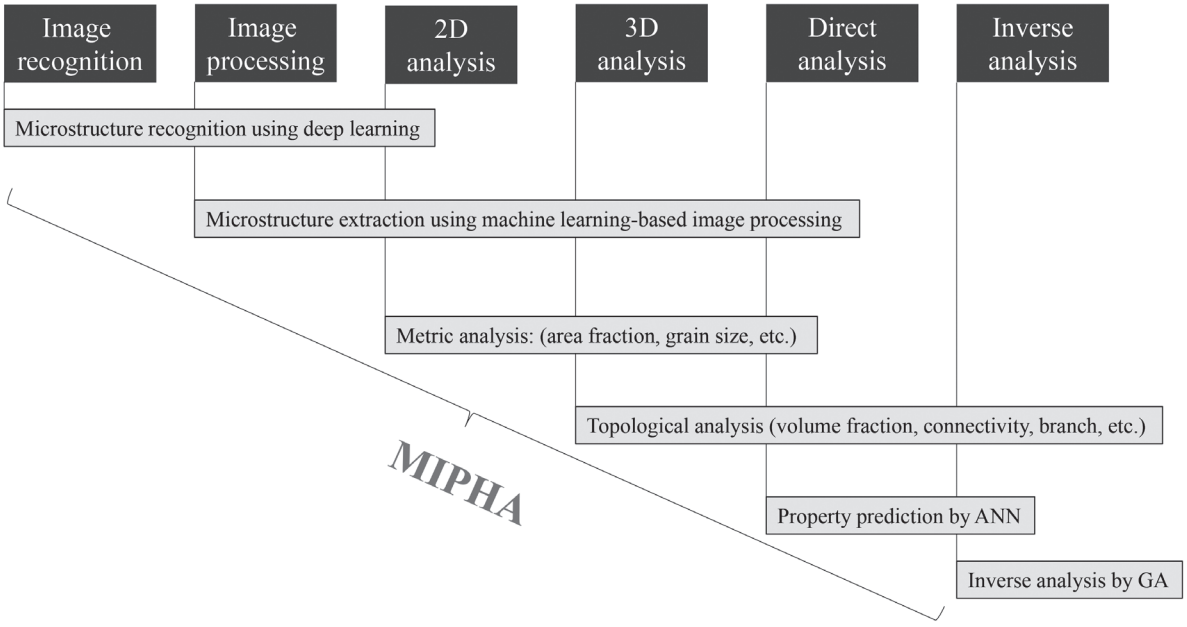


Fig.1 Functions and characteristics of MIPHA. (reprinted from Ref. 4))

sparse analysis, direct analysis, inverse analysis, and options. On the basis of image natural properties such as brightness, shinyMIPHA readily characterizes topological features of materials' microstructures tracking with image similarity analysis, two-point correlation statistic, persistent homology analysis, and mean (H) –Gauss (K) curvature analysis[2]. The sparse analysis comprises the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the least absolute shrinkage and selection operator (LASSO) for variable selection; principal component analysis (PCA), kernel PCA, and Autoencoder for dimension reduction; and K-means and self-organizing map algorithms for cluster analysis[3]. In the direct analysis function, the models of multiple regression, Gaussian process regression, ANN, SVR, and RF with hyper-parameter Bayesian optimization (BO) are used for property predictions. On the basis of direct analysis models, inverse analysis can be performed using BO, genetic algorithm (GA), or particle swarm optimization (PSO) to explore the potential materials' properties as well as their corresponding microstructure and processing variables[3]. The options division provides accessibilities for creating 2D/3D random data used for persistent homology analysis, modifying data after variable selection, random sampling from a large-sized dataset, and image processing features such as resizing, cropping, binarization, conversion, rotation, and plotting of 2D/3D/4D graphs.

3 Application of MIPHA and shinyMIPHA

3.1 Design of high-performance steels with MIPHA[4,6]

The commercial demands for highly strong and flexible steels are growing. However, traditional experiment-based materials research is becoming insufficient for meeting such demands. This section demonstrates a machine-learning-based property-to-microstructure-to-processing inverse analysis approach used for designing high-performance steels.

Cold-rolled (CR) low-carbon steels were studied in this section. The chemical compositions of the raw materials and processing parameters are detailed in Table1[4]. Fig.3 illustrates the 3D microstructure of sample A10-01 reconstructed by MIPHA, which intuitively and proximately present the real microstructure of the sample. The quantitative microstructure information in terms of count fraction (CF: count/total volume) and volume fraction (VF) of the identified phase components (polygon ferrite (PF), Widmanstatten ferrite (WF), pearlite (P), degenerated pearlite (DP), bainite (B) and martensite (M)) is given in Table2. The mechanical properties of tensile strength (TS) and total elongation (tEL) estimated from the stress-strain curves are also included.

The data contained in Table2 comprise the dataset used for regression analysis and inverse analysis, for which ANN and GA were employed here, respectively. An ANN model

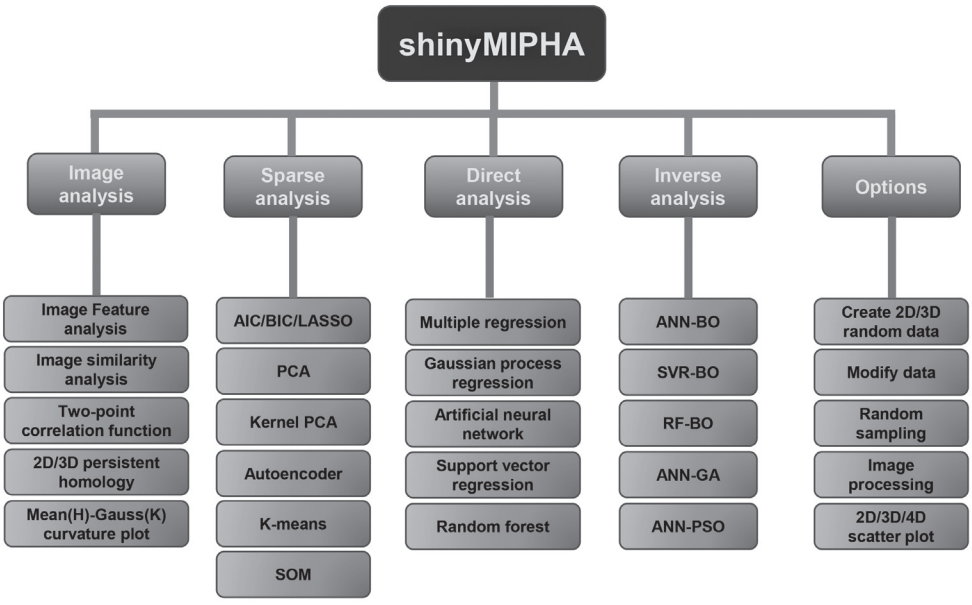


Fig.2 Functions and characteristics of shinyMIPHA. (reprinted from Ref. 5))

Table1　Chemical compositions and processing conditions of the used steels.（reprinted from Ref. 4））

Steel	Chemical composition (wt.%, N, O: ppm)	Process
A10-01	0.152C-0.015Si-1.51Mn-0.007P-0.0016S-0.027Al-18N-28O	CR→ annealed at 1000°C for 5 s→ cooling at 1°C/s
A10-03	0.152C-0.015Si-1.51Mn-0.007P-0.0016S-0.027Al-18N-28O	CR→ annealed at 1000°C for 5 s→ cooling at 3°C/s
A10-10	0.152C-0.015Si-1.51Mn-0.007P-0.0016S-0.027Al-18N-28O	CR→ annealed at 1000°C for 5 s→ cooling at 10°C/s
A10-30	0.152C-0.015Si-1.51Mn-0.007P-0.0016S-0.027Al-18N-28O	CR→ annealed at 1000°C for 5 s→ cooling at 30°C/s
A14-01	0.152C-0.015Si-1.51Mn-0.007P-0.0016S-0.027Al-18N-28O	CR→ annealed at 1400°C for 5 s→ cooling to 1000°C at 50°C/s → cooling at 1°C/s
A14-03	0.152C-0.015Si-1.51Mn-0.007P-0.0016S-0.027Al-18N-28O	CR→ annealed at 1400°C for 5 s→ cooling to 1000°C at 50°C/s → cooling at 3°C/s
A14-10	0.152C-0.015Si-1.51Mn-0.007P-0.0016S-0.027Al-18N-28O	CR→ annealed at 1400°C for 5 s→ cooling to 1000°C at 50°C/s → cooling at 10°C/s
A14-30	0.152C-0.015Si-1.51Mn-0.007P-0.0016S-0.027Al-18N-28O	CR→ annealed at 1400°C for 5 s→ cooling to 1000°C at 50°C/s → cooling at 30°C/s
B10-01	0.151C-0.013Si-1.53Mn-0.007P-0.002S-0.193Mo-0.028Al-21N-21O	CR→ annealed at 1000°C for 5 s→ cooling at 1°C/s
B10-03	0.151C-0.013Si-1.53Mn-0.007P-0.002S-0.193Mo-0.028Al-21N-21O	CR→ annealed at 1000°C for 5 s→ cooling at 3°C/s
B10-10	0.151C-0.013Si-1.53Mn-0.007P-0.002S-0.193Mo-0.028Al-21N-21O	CR→ annealed at 1000°C for 5 s→ cooling at 10°C/s
B10-30	0.151C-0.013Si-1.53Mn-0.007P-0.002S-0.193Mo-0.028Al-21N-21O	CR→ annealed at 1000°C for 5 s→ cooling at 30°C/s

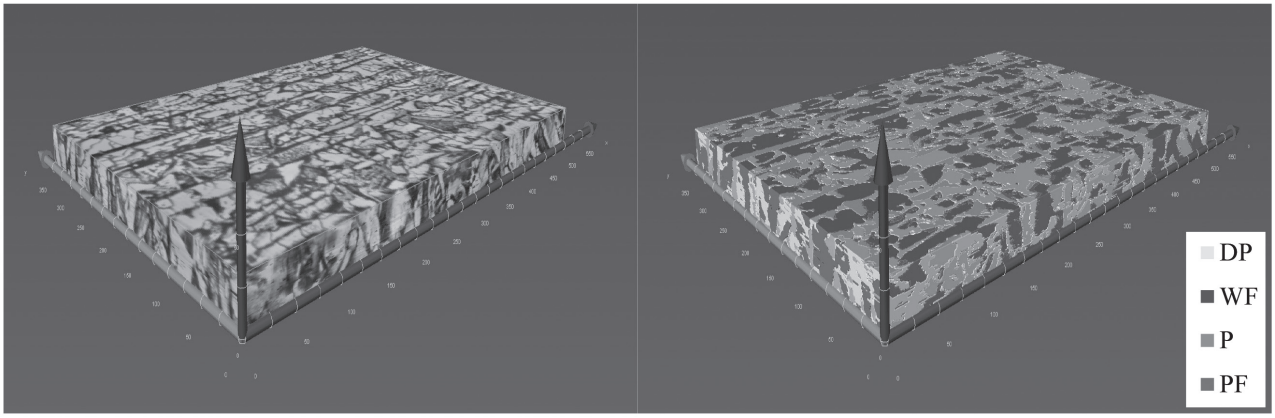


Fig.3　Reconstructed 3D microstructure of sample A10-01.（reproduced from Ref. 4））

Table2　Mechanical properties and material genomes of the used steels.（reprinted from Ref. 6））

Steel	YS(MPa)	TS(MPa)	tEL(%)	CFPF	CFP	CFWF	CFDP	CFB	CFM	VFPF	VFP	VFWF	VFDP	VFB	VFM
A10-01	323	481	80.6	3.30E-05	5.16E-05	0.000165	1.83E-05	0	0	0.4047	0.2005	0.0845	0.3104	0	0
A10-03	308	489	76.4	5.43E-05	9.07E-05	1.28E-06	9.24E-05	0	0	0.2608	0.118	0.5537	0.0674	0	0
A10-10	390	591	71.1	5.17E-05	0.000136	0.000126	0	1.54E-06	0	0.1836	0.0452	0.1414	0	0.6297	0
A10-30	444	663	63.9	8.63E-05	0	0.00027	0	9.26E-07	4.42E-05	0.1576	0	0.0842	0	0.5765	0.1817
A14-01	353	516	64.4	7.67E-05	5.15E-05	4.04E-06	3.21E-05	5.30E-06	0	0.1573	0.0212	0.3938	0.0379	0.3897	0
A14-03	412	561	67.5	4.40E-05	4.67E-05	3.12E-05	5.94E-05	5.93E-06	0	0.0808	0.0143	0.2572	0.1232	0.5245	0
A14-10	521	688	61.5	0	2.27E-05	0	0	1.08E-05	1.73E-05	0	0.0094	0	0	0.6249	0.3657
A14-30	620	807	60.7	0	0	0	0	0	3.00E-05	0	0	0	0	0	1
B10-01	375	550	70.4	0.002244	0.000637	0.001005	0	0	0	0.373652	0.06523	0.561117	0	0	0
B10-03	434	600	66.3	0.00325	0.000344	0.005683	0	0.000477	0	0.109254	0.006947	0.022508	0	0.861291	0
B10-10	483	691	61.5	0.000185	0.000215	0	0	2.76E-05	0	0.118045	0.006877	0	0	0.875078	0
B10-30	489	725	58.4	0	0	0	0	7.43E-05	1.52E-07	0	0	0	0	0.160882	0.839118

with two objective variables of TS and tEL was established, as schemed in Fig.4 [6]. A maximum search of TS × tEL was subsequently conducted, and the explored results are shown in Table3. The results indicate that a potential balanced property TS × tEL value of 65105.7 requires higher phase fractions of M, B and WF. Moreover, the inverse analysis also semi-quantitatively provides the chemical composition and processing conditions corresponding to the potential properties. Fig.5 clearly shows the relationships between the experimental design and data-driven design.

3.2 Design of high-performance thermoelectric materials shinyMIPHA[7]

The traditional experimental research is generally carried out using a trial-and error method, by which a material is designed from given chemical composition and processing conditions, followed by evaluation of microstructure and properties. In addition to being time-consuming, such procedures highly depend on the experiences and knowledge of the researchers, which can easily underestimate the materials' characteristics. Since substantial progress tends to require a combination of chemical intuition and serendipity, traditional experiment-based methods appear to be increasingly insufficient for designing new materials with desired properties. In this section, the machine-learning-based data-driven approach was applied to hot-extruded $Cu_xBi_2Te_{2.85+y}Se_{0.15}$ thermoelectric materials, where the relationships among composition, processing, microstructure, and properties were further understood.

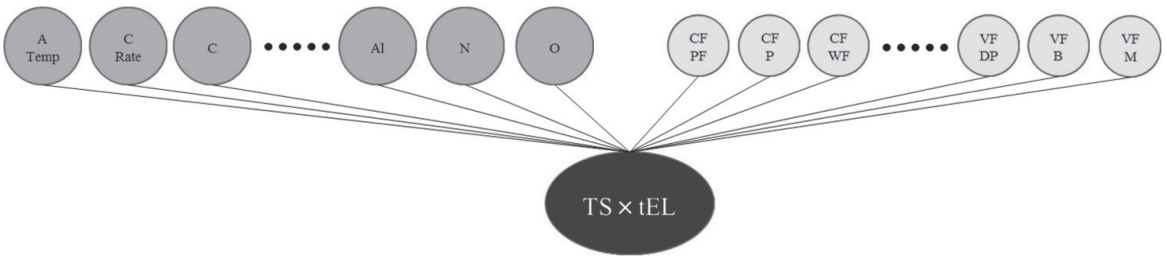


Fig.4 Schematic of the properties-to-microstructure-to-processing inverse analysis. ATemp and CRate denote the austenizing temperature and cooling rate. (reproduced from Ref. 6))

Table3 Inversely explored properties, microstructure, and processing. (reprinted from Ref. 6))

ATemp (°C)	CRate (°C/s)	C	Si	Mn	S	Mo	Al	N	O	CFPF	CFP	CFWF
1344	21.30	0.1519	0.0146	1.5132	0.00169	0.00772	0.02786	18.8	22.1	0.00234	0.00056	0.00215
CFDP	CFB	CFM	VFPF	VFP	VFWF	VFDP	VFB	VFM	TS (MPa)	tEL (%)	TS × tEL	
2.03E-05	0.00044	4.33E-05	0.03958	0.08736	0.21455	0.09936	0.38128	0.17784	808.278	80.5486	65105.7	

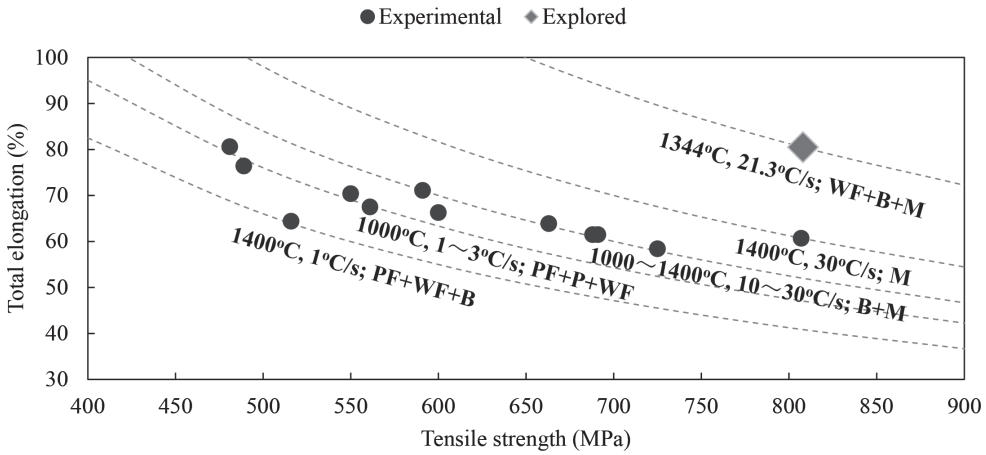


Fig.5 Relationships between the experimental design and data-driven design. (reproduced from Ref. 6))

Materials design often involves multiple covariant variables in terms of composition and processing, which is a challenge for the experiment-based method to thoroughly understand the materials paradigm. Here, the sparse algorithm of principal component analysis (PCA) was employed to characterize the influence of multiple composition and processing variables. The raw data tracking with the processing, composition, microstructure, and properties of $Cu_xBi_2Te_{2.85+y}Se_{0.15}$ materials used in this section are provided in Ref. 7). Fig.6 shows the PCA map expressed by PC1 and PC2, demonstrating the relationships among temperature, Te content, Cu content, and Cu particle size, microstructure, and properties. According to the Euclidean distance, temperature and Cu content were shown to have remarkable influences on the microstructure and properties, whereas the influences of Cu particle size and Te content were small. These results suggest that temperature and Cu content are two priority parameters in processing design of $Cu_xBi_2Te_{2.85+y}Se_{0.15}$ thermoelectric materials.

Regression analysis was conducted using ANN, followed by predictions of property index of themoelectric materials (figure of merit ZT), as shown in Fig.7. The results demonstrated that the experimental and predicted ZT well matched with each other, suggesting a capability of the present model to describe underlying data. On the basis of the ANN model, an inverse analysis was performed

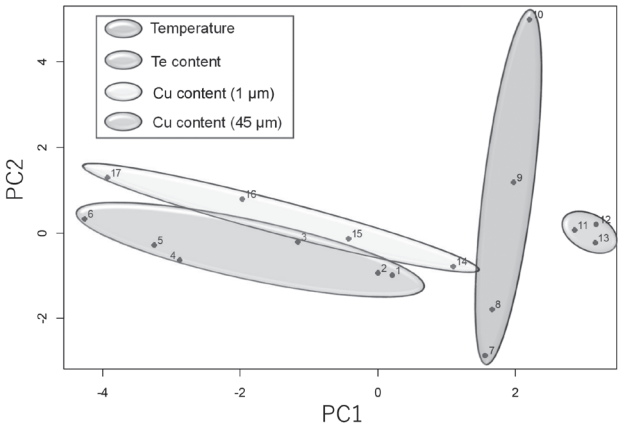


Fig.6 A principle component map demonstrating the primary variance of the observations by their PC1 and PC2. (reprinted from Ref. 7))

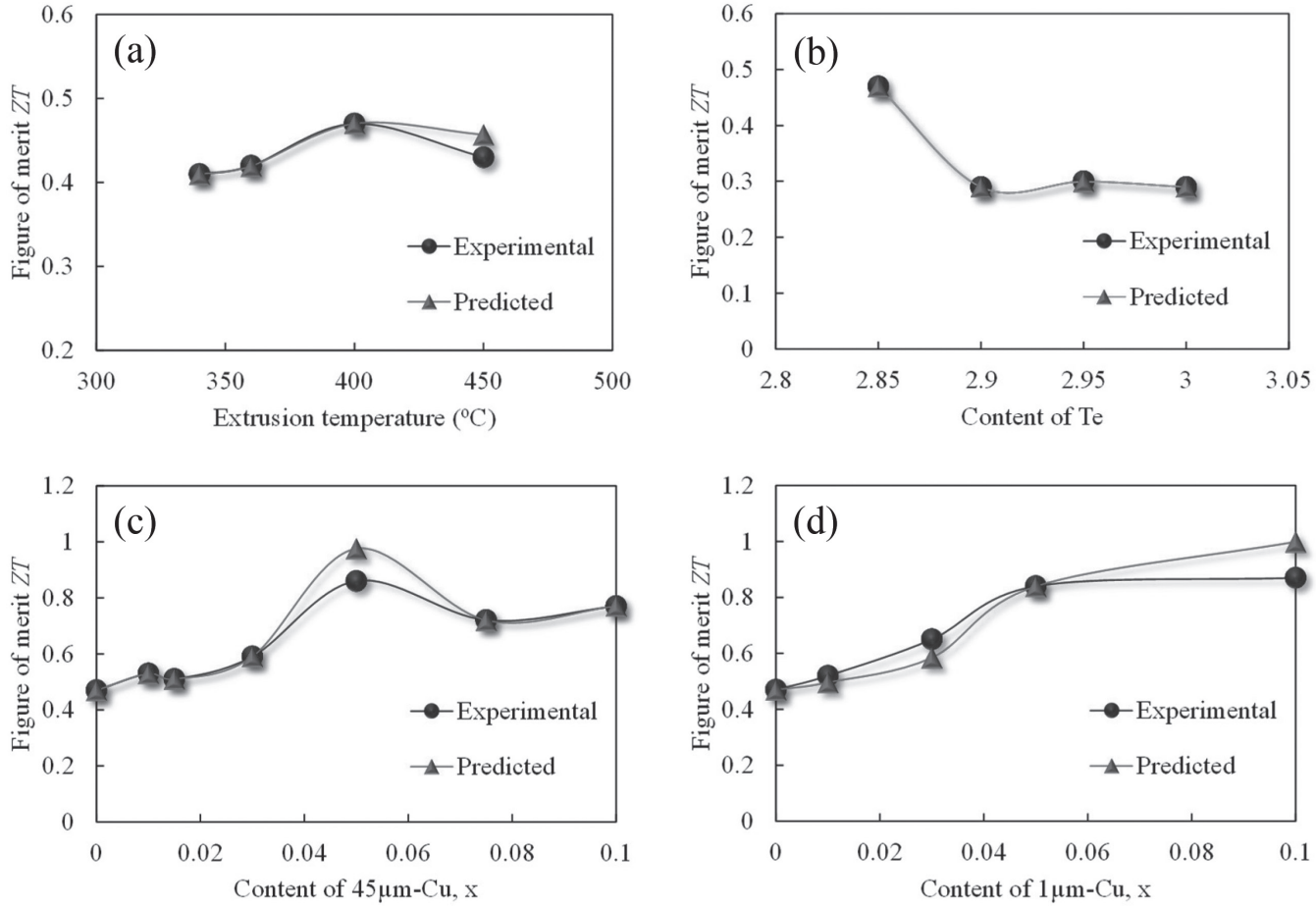


Fig.7 Experimental and predicted ZT values. The predictions were made by ANN model under different processing variables of (a) extrusion temperature, (b) Te content, (c) Cu content (45 μm), and (d) Cu content (1 μm). (reprinted from Ref. 7))

using GA with a maximum ZT value search. Table4 gives the inversely explored results with respect to the potential properties and their corresponding microstructure, composition and processing. The inverse analysis indicates that the $Cu_xBi_2Te_{2.85+y}Se_{0.15}$ materials have a potential best *ZT* value of 1.15, which is 1.32 times larger than the best experimental value. In materials design, the optimal *ZT* requires processing variables of higher extrusion temperature and larger Cu content and microstructure variables of higher density and larger average grain size.

3.3 Property predictions using persistent homology analysis with shinyMIPHA[8)]

The current microstructural descriptors tracking with properties of interest are primarily in terms of metallurgical features, *e.g.*, grain size, texture, and area/volume fraction, which often ignore the complexities of the microstructure's geometry and thus easily underestimate materials' properties. In addition, the materials' microstructure is generally quantified using stereological measurements, which highly rely on the prior metallurgical knowledge of an expert to recognize and identify certain key microstructural features in advance. Such characterizations often result in significant bias and individual errors. In this section, persistent homology was demonstrated to characterize topological microstructure features of the DP steel samples, followed by predictions of stress－strain curves using a machine-learning model of ANN. In addition, the correlations between stress and microstructure descriptor of persistent images are estimated using sensitivity analysis, Bayesian information criterion (BIC), and the least absolute shrinkage and selection operator (LASSO) respectively.

Fig.8 illustrates the persistent homology analysis

Table4 Inversely explored results by GA, where *T*, *Cu*, *CuSize*, *D*, *d*, μ, α, and κ denote the temperature, Cu content, Cu particle size, relative density, average grain size, mobility, Seebeck coefficient, and thermal conductivity, respectively. (reprinted from Ref. 7))

	T	*Cu*	*CuSize*	*D*	*d*	μ	α	κ	*ZT*
GA	437	0.07	-9.1	1.046	0.94	204.702	-219.108	0.772	1.15
Exp.	400	0.05	45	0.931	0.75	156.6	-177	0.93	0.86

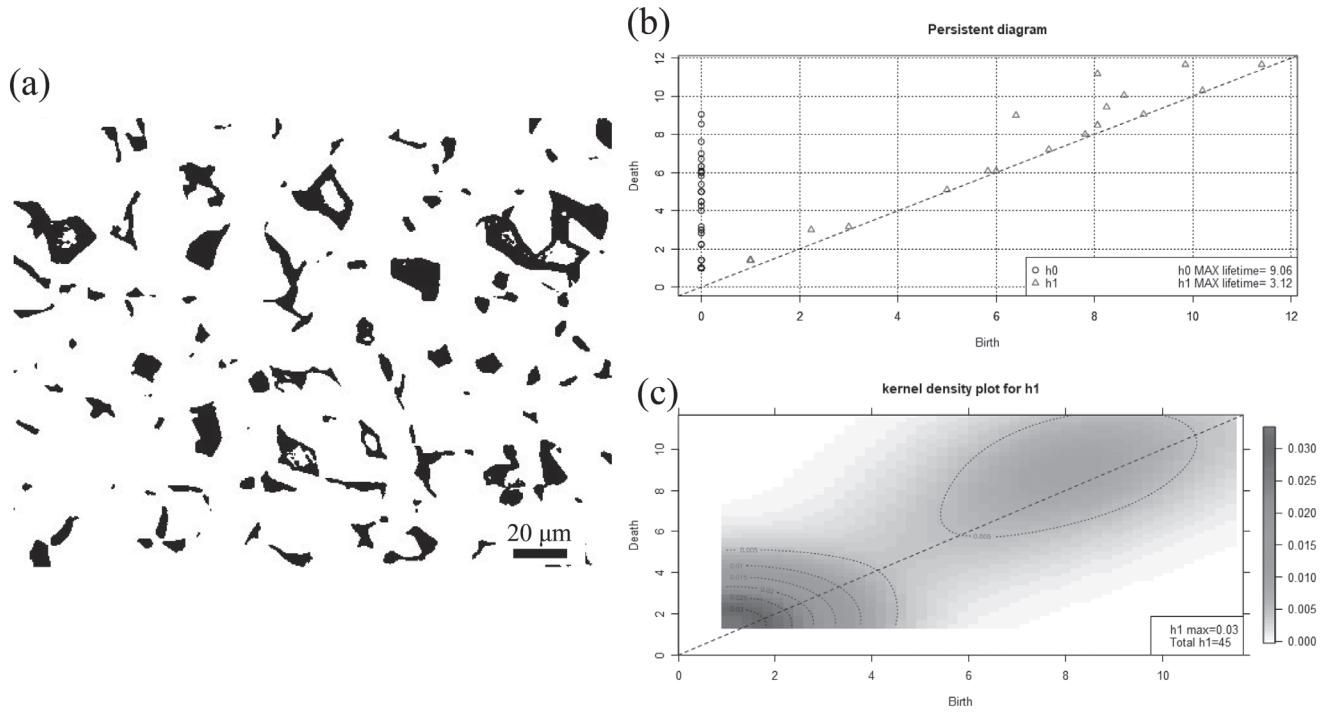


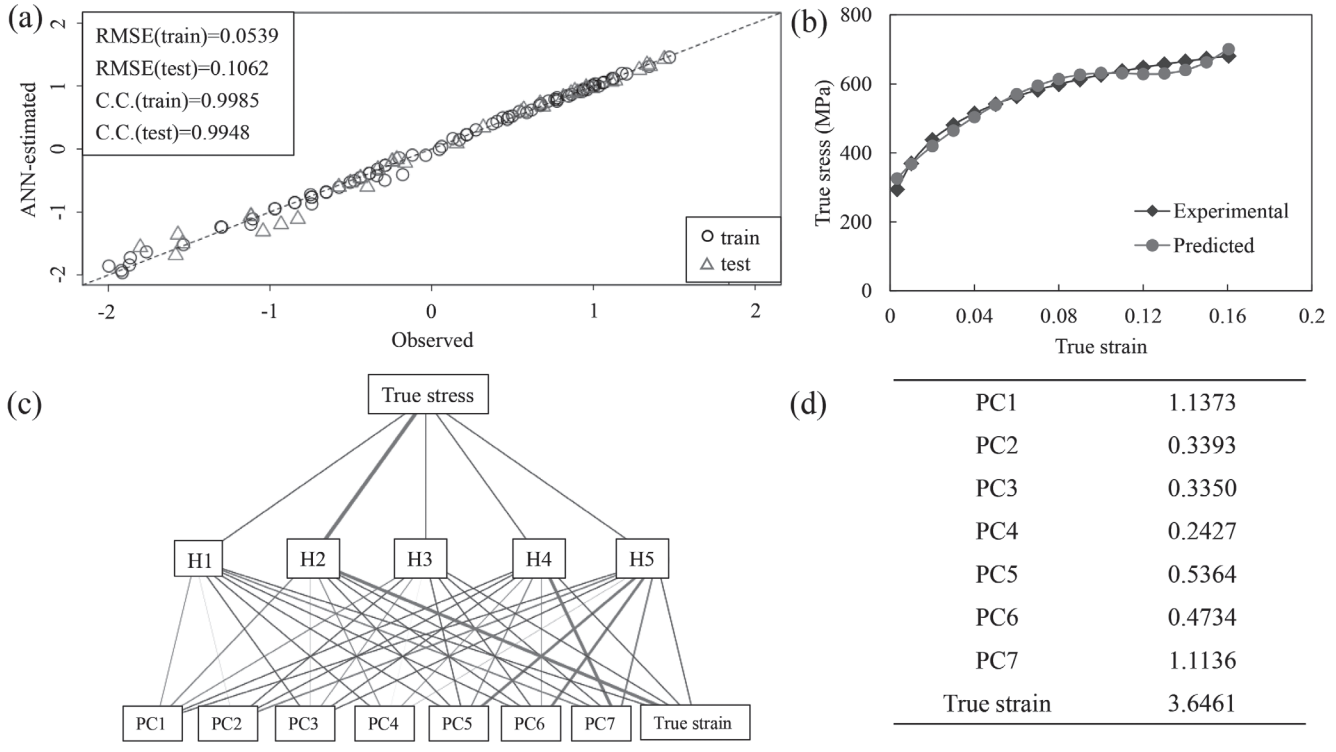
Fig.8 Persistent homology analysis for a DP sample: (a) a binary image with 480 × 360 pixels; (b) a persistent diagram estimated from a space containing 50 × 37.5 pixels resized from (a); and (c) the kernel density map for the h_1 feature in a persistent diagram. (reprinted from Ref. 8))

results of a DP sample. Fig.8 (a) shows the binary image, where the ferrite and martensite are highlighted by white and black contrast, respectively. Fig.8 (b) shows the persistent diagram for the martensite. Two features were identified: h_0 is a ring existing in a certain martensite island, and h_1 is a potential ring among the martensite islands. A unique quality of persistent homology is that it can capture meaningful underlying topological features. Thus, the h_1 ring feature was estimated to accounting for the microstructure. Fig.8 (c) shows the kernel density map demonstrating the distribution of the h_1 feature in the persistent diagram. The microstructure descriptor of persistent image (PI) was estimated based on the persistent diagram and kernel density[2].

Since the source data of PI (given in Ref. 8)) possesses a dimension of 2500, PCA was thereby employed to reduce the dimension of the dataset, by which 7 PCs were identified sufficient for interpreting the original observations. Thus, 7 PCs, strain, and stress comprise the the dateset for regression analysis by ANN. Fig.9 shows the regression analysis results. The fitted ANN model exhibited satisfactory accuracies for both the training and testing datasets, as shown in Fig.9 (a). Fig.9 (b) illustrates experimental and ANN-predicted stress－strain curves. The experimental and predicted curves nearly coincide, indicating a good prediction performance of the present model.

Sensitivity analysis[3] was conducted to identify the correlation between stress and PI based on the neural network shown in Fig.9 (c). Red and blue colors express positive and negative sensitivity, respectively, and a wider connection line expresses a high degree of sensitivity. The quantitative sensitivity degrees of the objective variable to each explanatory variable are given in Fig.3 (d). The results show that true strain is the most sensitive factor to true stress with a sensitivity degree of 3.6464, whereas the total sensitivity degree of PCs reaches 4.1777, suggesting a strong correlation between true stress and PI. In addition, PC1 and PC7 exhibit relatively high degrees of sensitivity, indicating that true stress is most sensitive to the microstructure information contained in these two components, followed by PC5, PC6, and the weak factors of PC2, PC3, and PC4.

LASSO and BIC[3] were further carried out to identify the correlations between the objective and explanatory variables. As shown in Fig.10, LASSO estimation



PC1	1.1373
PC2	0.3393
PC3	0.3350
PC4	0.2427
PC5	0.5364
PC6	0.4734
PC7	1.1136
True strain	3.6461

Fig.9 Regression analysis using ANN: (a) accuracy of the fitted model; (b) experimental and ANN-predicted stress–strain curves; (c) network of the ANN model; and (d) quantitative sensitivity degree of the explanatory variables. (reprinted from Ref. 8))

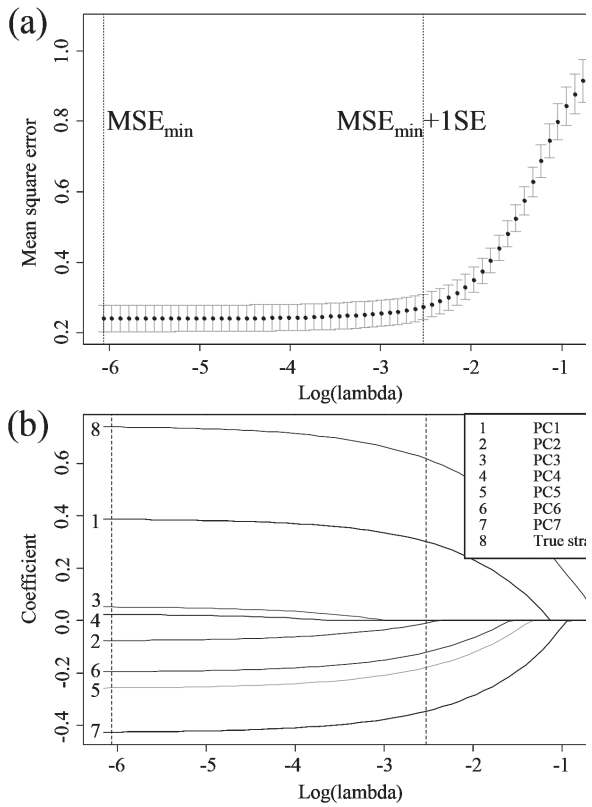


Fig.10 Regularization parameter (*log*(*λ*)) dependences of the (a) mean square error and (b) regression coefficient estimated by LASSO. (reprinted from Ref. 8))

demonstrates that true strain exhibits the largest absolute value of coefficient, followed by PC7, PC1, PC5, and PC6 at the threshold (right dashed line), and the coefficients of weak correlations PC2, PC3, and PC4 are constrained to 0. BIC estimation identifies a relationship between the stress and the explanatory variables as BIC_{min}*true stress* ~ *0.7405true strain + 0.4320PC7 + 0.4024PC1 – 0.2597PC5 – 0.2002PC6*.

The above three sparse studies demonstrate similar correlations between the objective and explanatory variables, indicating that the microstructure descriptor PI is capable of interpreting properties. Here, present persistent homology presents a route for characterizing materials' microstructure in geometry, which is capable of complementing the deficiencies in the metallurgical-feature-based microstructure characterization. Furthermore, combined with image similarity analysis[2], an inverse analysis approach based on persistent homology is under development, so as to explore the microstructure and processing conditions that track with a desired property. The proposed approach is aimed to reduce the dependence of the aforementioned stereological measurements and thus accelerate materials discovery process.

4 Summary

In response to increasing demand for the highly efficient design of new materials, friendly and efficient machine learning facilities are becoming critical for applying artificial intelligence to materials research community. This paper introduces two independently developed machine learning tools, whose frameworks and functions have been demonstrated in terms of property predictions and inverse design. The developed machine learning tools and related work involved are expected to provide new persepective for promoting the materials research.

References

1) Y. Adachi, Z.-L. Wang and T. Ogawa：ふぇらむ, 25 (2020), 569.
2) Y. Adachi, Z.-L. Wang and T. Ogawa：ふぇらむ, 25 (2020), 628.
3) Y. Adachi, Z.-L. Wang and T. Ogawa：ふぇらむ, 25 (2020), 695.
4) Z.-L. Wang and Y. Adachi：Mater. Sci. Eng. A, 744 (2019), 661.
5) Z. L. Wang, T. Ogawa and Y. Adachi：Advanced Theory and Simulations, (2019), 1900177.
6) Z.-L. Wang, T. Ogawa and Y. Adachi：ISIJ Int., 59 (2019), 1691.
7) Z.-L. Wang, Y. Adachi and Z.-C. Chen：Advanced Theory and Simulations, (2019), 1900197.
8) Z.-L. Wang, T. Ogawa and Y. Adachi：Advanced Theory and Simulations, (2020), 1900227.

(2021年2月19日受付)