統計数理(2021)
第 69 巻 第 1 号 49–63

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]

材料研究における転移学習の応用

劉 暢[1]・山田 寛尚[1,2]・ウ ステファン[1,3]

(受付 2020 年 11 月 6 日;改訂 2021 年 5 月 10 日;採択 5 月 12 日)

要　　旨

材料研究のデータ量は,機械学習の他の応用分野に比べると圧倒的に少ない.このスモールデータの壁を乗り越えるために,転移学習を活用する.化学的特性,電気的特性,熱力学的特性,機械特性など,材料特性の間には物理化学的な依存関係が存在している.限られたデータからある特性を予測するために,十分なデータが利用可能な代理特性のモデルを事前に学習し,このモデルを目標物性の予測に利用する.このように他のドメインから獲得したモデルや特徴表現を目標ドメインの予測に利用することで,非常に少ないデータでも高い予測能力を持つモデルを構築できることがある.本稿では,高分子や無機材料を含む様々な問題設定で転移学習を適用し,その潜在的能力をデモンストレーションする.特に,転移学習を適用することで,訓練データ集合の分布の範囲を大きく逸脱した領域において予測能力を獲得した事例を報告する.

キーワード:転移学習,物性予測,スモールデータ,有機材料,ポリマー,無機材料.

1. はじめに

従来の材料研究では,候補材料の特性を評価するために分子動力学計算や第一原理計算のような物理モデルに基づく計算機実験が活用されてきた.しかしながら,一般にシミュレーションは膨大な計算コストを要するため,網羅的な材料スクリーニングへの適用は難しい.そこで,計算コストが小さい統計モデルに特性評価の計算を代替させて,膨大な数の候補材料のスクリーニングを実現しようという研究が様々な材料系を対象に進行している(Carrete et al., 2014; Seko et al., 2015; Gómez-Bombarelli et al., 2016; Hansen et al., 2016; Oliynyk et al., 2016; Sumita et al., 2018; Matsumoto et al., 2018; Wu et al., 2019; Liu et al., 2021).多いときで,仮想ライブラリを含む数億オーダーの候補材料を対象にスクリーニングが実施される.機械学習の目的は,候補材料 S の特性 Y を計測した実験や物理シミュレーションのデータ集合を用いて,予測モデル $Y = f(S)$ を導くことである.問題形式は単純な教師あり学習である.

データ駆動型材料研究に立ちはだかる最も大きな壁は,限られたデータ量とデータの多様性の不足の問題である.画像認識や自然言語処理などのデータ科学の他の応用分野と比べると,材料研究に利用可能なデータの量は圧倒的に少ない.例えば,本稿で示す無機材料の熱伝導率の研究では,データ数がたったの 45 個しかない.データが少ない主な原因として,次の三点が考えられる.

[1] 統計数理研究所:〒 190–8562 東京都立川市緑町 10–3
[2] 東京薬科大学 薬学部:〒 192–0392 東京都八王子市堀之内 1432–1
[3] 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

- 実験や計算機実験のコストが高い．
- 材料組成の選択や材料作製のプロセス設計(温度依存性，添加物・溶媒選択)など，材料特性の決定には非常に多くの因子が関与する．したがって，一般に設計空間が極めて広大になる．さらに，個々の研究者の研究対象が大きく異なるため，社会全体でコモンデータを創出しようという動きが起こりにくい．
- 科学的成果と産業応用の垣根が低いため，競合相手に対する情報秘匿の意識が高く，データ公開に対するインセンティブが研究者に働きにくい．

以上の理由により，コミュニティが協力してコモンデータを創出しようという動向は極めて低調である．さらに，研究対象が先端領域に近づくにつれ，スモールデータの傾向はより顕著になる．少なくとも中長期的には，大学の研究室や一企業で生産可能なデータがマテリアルズインフォマティクスの標準的な解析対象になると予想される．

本稿では，転移学習という方法論を用いて材料研究のスモールデータの問題にアプローチする（Agrawala and Choudhary, 2016; Hutchinson et al., 2017; Oda et al., 2017; Jalem et al., 2018; Yonezu et al., 2018; Kailkhura et al., 2019; Segler et al., 2018; Cubuk et al., 2019; Li et al., 2018; Kaya and Hajimirza, 2019)．転移学習はあるドメイン(元ドメイン)の学習モデルを別のドメイン(目標ドメイン)に利活用するための方法論である．目標ドメインのデータ量が不足している場合，一旦，十分な量のデータを利用できる元ドメインのモデルを構築する．この訓練済みモデルの特徴量や推定されたパラメータを目標ドメインのスモールデータで改変して最終的なモデルを導く．データ量が少なくてフルスクラッチでの学習は難しいが，関連する元ドメインの訓練済みモデルを適切に利用することでデータ量の不足を補う．このようなアプローチがスモールデータの壁を乗り越える有効な手段になりうることが，様々な分野で実証されつつある．本稿では，有機高分子や無機化合物の物性予測など，様々な材料研究における転移学習の成功事例を紹介する．特に転移学習が有する外挿的な予測能力をデモンストレーションする．なお，本稿で示す解析結果の詳細な説明については，著者らの原著論文 Yamada et al. (2019) を参照せよ．

2. 転移学習

2.1 ニューラルネットワークを用いた教師あり転移学習

本稿はニューラルネットワークを用いた教師あり転移学習に焦点を絞る．特別な工夫は何も施さず，教科書的なテクニックのみを用いる．一般にニューラルネットワークの学習では，入力層に近い下層のニューロンが一般的な特徴量を表し，出力層に近づくにつれてドメイン固有の特徴量に変換されていく．ニューラルネットワークの転移学習はこの性質を利用する．

ここで，転移元のドメインを元ドメイン(source task)，転移先の目標ドメイン(target task)と呼ぶことにする．元ドメインの教師データを用いて，モデル $Y_s = f_s^L(X)$ を構築する．Y_s と X は系の元ドメインの出力変数と入力変数，$f_s(X) = f_L \circ f_{L-1} \ldots \circ f_1(X)$ は L 層のニューラルネットワークを表す．この訓練済みモデルの第 K 層($K < L$)までの部分モデル $\phi(X) = f_K \circ f_{K-1} \ldots \circ f_1(X)$ を目標ドメインのモデルの記述子として利用する方法を特徴抽出による転移学習と呼ぶ．すなわち，目標ドメインの教師データを用いて $Y_t = f_t \circ \phi(X)$ という形式のモデルを学習する．ここで，Y_t は目標ドメインの出力変数，f_t は任意のモデルである．図 1 は，転移学習のワークフローを模式的に表したものである．元ドメインの学習過程で，ニューラルネットワークは Y_s の予測に有用な特徴量 $\phi(X)$ を獲得する．元ドメインと目標ドメインの間に共通のメカニズムが存在すれば，この特徴量は Y_t の予測にも活用できることが期待される．$\phi(X)$ の次元を十分に小さくとることができ，かつ線形モデルのような簡素な

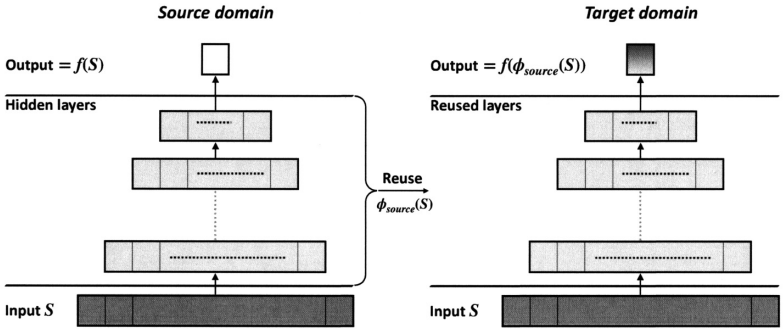

図 1. 元ドメインの訓練済みモデルの一部(この例では出力層を除いた部分モデル)を目標ドメインの特徴抽出器として活用するタイプの転移学習.

モデルで f_t を記述できる系であれば，目標ドメインのモデルをフルスクラッチで学習するよりも，少ないデータ量で高い予測精度のモデルを構築できるかもしれない.

次にファインチューニング(fine-tuning)と呼ばれる転移学習の方法を解説する．ファインチューニングは訓練済みモデルの重みを初期値とし，目標ドメインのデータセットを用いて再学習する．目標ドメインの学習の際は，パラメータ更新を早期に停止し，重みを大きく変化させずに，訓練済みモデルを微修正する．適切な早期停止を実施するために，目標ドメインのデータを訓練用と検証用に分割した上で，訓練用データを用いて低い学習率でパラメータを更新しながら，検証用データの予測精度の変化をモニタリングし，精度が最も高くなるタイミングでパラメータ更新を停止する．理論的な根拠は明確ではないが，ファインチューニングは経験的に有効なアプローチであることが知られており，特に画像認識や機械翻訳等の訓練済みモデルを対象に広く用いられている.

2.2 訓練済みモデルライブラリ XenonPy.MDL

本グループは，XenonPy (https://xenonpy.readthedocs.io/)という Python パッケージを開発している．XenonPy は，様々な材料を対象に材料設計のワークフローを構築するために必要なモジュール群を実装している．本稿では XenonPy の解説は行わないが，ここでは，サブモジュールの一つである XenonPy.MDL という物性予測タスクを対象とする訓練済みモデルライブラリを紹介する．このライブラリには 2020 年時点で，低分子，高分子，無機材料の 45 種類の特性を予測する約 140,000 個の訓練済みモデルが実装されている(表 1)．ユーザーは API (Application Programming Interface)を用いて訓練済みモデルを取得し，XenonPy を経由して材料設計の様々なワークフローを構築できる．脳の学習メカニズムとの対比で言えば，多様且つ包括的な訓練済みモデル群を実装することは，多くの経験から記憶の集合体を獲得することに相当する．モデルの多様性が増すほど強力な転移学習を実現できる可能性が高まる.

3. 転移学習の適用例

ここからは，重要物性である熱伝導率と屈折率を対にした転移学習の三つの適用例を取り上げ，転移学習の有効性を示す.

3.1 無機化合物の熱伝導率

熱伝導率は，伝熱，対流，放熱等の熱の移動を考える上で，材料の断熱性能を表す指標であ

表 1. XenonPy.MDL に収録されている訓練済みモデルの抜粋．低分子，高分子，無機材料の 45 種類の特性を予測するモデルを実装している（2020 年 10 月時点）．

Material type	Database	Property	Model type	Model parameters	No. of models	Best model correlation	No. of descriptors	Descriptor type
Organic	PoLyInfo (polymer)	Glass transition temperature	RF-R	RF setup 1	1,000	0.950	max 500*	rcdk-all
			GB-R	GB setup	1,000	0.950	max 500*	rcdk-all
			EN-R	EN setup	1,000	0.920	max 500*	rcdk-all
			NN-R	NN setup 1	1,000	0.950	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.955	2,048	RDKit-5
		Density	NN-R	NN setup 1	1,000	0.910	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.859	2,048	RDKit-5
		Viscosity	NN-R	NN setup 1	1,000	0.890	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.613	2,048	RDKit-5
		Melting temperature	NN-R	NN setup 1	1,000	0.880	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.885	2,048	RDKit-5
		Heat capacity (const. pressure)	NN-R	TL setup 1	25,000	0.992	max 400-600#	rcdk-all
		Thermal conductivity	NN-R	TL setup 1	25,000	1.000	max 400-600#	rcdk-all
	QM9 (small molecule)	Heat capacity at constant volume	NN-R	NN setup 1	∼500	0.900	max 400-600#	rcdk-all
		LUMO	NN-R	NN setup 1	∼500	0.950	max 400-600#	rcdk-all
		HOMO-LUMO gap	NN-R	NN setup 1	∼500	0.940	max 400-600#	rcdk-all
		Zero point vibrational energy	NN-R	NN setup 1	∼500	0.940	max 400-600#	rcdk-all
		Internal energy at 0 K	NN-R	NN setup 1	∼500	0.920	max 400-600#	rcdk-all
		Enthalpy at 298.15 K	NN-R	NN setup 1	∼500	0.910	max 400-600#	rcdk-all
		Free energy at 298.15 K	NN-R	NN setup 1	∼500	0.910	max 400-600#	rcdk-all
		HOMO	NN-R	NN setup 1	∼500	0.880	max 400-600#	rcdk-all
		Internal energy at 298.15 K	NN-R	NN setup 1	∼500	0.880	max 400-600#	rcdk-all
		Isotropic polarizability	NN-R	NN setup 1	∼500	0.870	max 400-600#	rcdk-all
		Electronic spatial extent	NN-R	NN setup 1	∼500	0.800	max 400-600#	rcdk-all
		Dipole moment	NN-R	NN setup 1	∼500	0.740	max 400-600#	rcdk-all
	Polymer Genome (polymer)	Bandgap	RF-R	RF setup 2	1,000	0.964	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.985	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.983	2,048	RDKit-5
		Dielectric constant	RF-R	RF setup 2	1,000	0.965	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.982	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.958	2,048	RDKit-5
		Ionic dielectric constant	RF-R	RF setup 2	1,000	0.898	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.934	max 400-600#	rcdk-all
		Electronic dielectric constant	RF-R	RF setup 2	1,000	0.930	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.947	max 400-600#	rcdk-all
		Refractive index	RF-R	RF setup 2	1,000	0.953	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.985	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.981	2,048	RDKit-5
		Atomization energy	RF-R	RF setup 2	1,000	0.974	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.986	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.992	2,048	RDKit-5
		Density	RF-R	RF setup 2	1,000	0.961	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.982	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.989	2,048	RDKit-5
		Ionization energy	RF-R	RF setup 2	1,000	0.922	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.962	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.940	2,048	RDKit-5
		Electron affinity	RF-R	RF setup 2	1,000	0.955	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.978	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.987	2,048	RDKit-5
		Cohesive energy	RF-R	RF setup 2	1,000	0.839	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.943	max 400-600#	rcdk-all
		Melting temperature	RF-R	RF setup 2	1,000	0.920	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.94	max 400-600#	rcdk-all

表 1.（つづき）

Material type	Database	Property	Model type	Model parameters	Num. of models	Best model correlation	Num. of descriptors	Descriptor type
Organic	Polymer Genome (polymer)	Glass transition temperature	RF-R	RF setup 2	1,000	0.937	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.962	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.931	2,048	RDKit-5
		Hildebrand solubility parameter	RF-R	RF setup 2	1,000	0.951	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.962	max 400-600#	rcdk-all
			NN-Py	NN setup 2	500	0.879	2,048	RDKit-5
		Molar heat capacity	RF-R	RF setup 2	1,000	0.989	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.991	max 400-600#	rcdk-all
		Molar volume	RF-R	RF setup 2	1,000	0.965	max 1,500-3,000#	rcdk-all
			NN-R	NN setup 1	1,000	0.984	max 400-600#	rcdk-all
	PHYSPROP	Boiling point	NN-R	NN setup 1	1,000	0.782	max 400-600#	rcdk-all
	MD database	Solvation free energy	NN-R	NN setup 1	1,000	0.94	max 400-600#	rcdk-all
	Jean-Claude Bradley	Melting temperature	NN-R	NN setup 1	1,000	0.84	max 400-600#	rcdk-all
Inorganic	Materials Project	Volume	NN-Py	NN setup 3	3,600%	0.997	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.606	N/A	N/A
		Formation energy per atom	NN-Py	NN setup 3	3,600%	0.997	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.977	N/A	N/A
		Total energy per atom	NN-Py	NN setup 3	3,600%	0.996	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.963	N/A	N/A
		Density	NN-Py	NN setup 3	3,600%	0.994	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.996	N/A	N/A
		Fermi energy	NN-Py	NN setup 3	3,600%	0.968	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.933	N/A	N/A
		Magnetization	NN-Py	NN setup 3	3,600%	0.923	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.723	N/A	N/A
		Bandgap	NN-Py	NN setup 3	3,600%	0.910	290/150	XenonPy
			CGCNN-Py	CNN setup	324	0.936	N/A	N/A
	Citrination datasets id:152062	Total dielectric constant	NN-Py	NN setup 3	3,600%	0.565	290/150	XenonPy
		Electronic dielectric constant	NN-Py	NN setup 3	3,600%	0.504	290/150	XenonPy
		Refractive index	NN-Py	NN setup 3	3,600%	0.762	290/150	XenonPy
	Shiomi data	Scattering phase space	NN-Py	NN setup 3	~1,200	0.912	290/150	XenonPy
		Lattice thermal conductivity	NN-Py	NN setup 3	~1,200	0.998	290/150	XenonPy
			NN-Py	TL setup 2	~200	0.999	290	XenonPy

訓練済みモデルの概要．RF-R，GB-R，EN-R，NN-R はそれぞれ，ランダムフォレスト(ranger)，勾配ブースティング(xgboost)，Elastic Net 回帰(glmnet)，ディープニューラルネットワーク(MXnet)を表す．ここで括弧内のシンボルは R のパッケージ名を表す．NN-Py と RF-Py はそれぞれ，ディープニューラルネットワーク(PyTorch)，ランダムフォレスト(scikit-learn)を表す．括弧内のシンボルは Python のパッケージ名である．CGCNN-Py は crystal graph convolution neural network(PyTorch)を表す．RF setup 1 は回帰木の数(nTree)を 100–800，特徴量の数(mTry)を 20–100 の範囲でランダムに選んでいる．RF setup 2 の場合は，nTree が 50–500，mTry が 50–500 である．GB setup は学習率(*eta*)を 0.1–1，決定木の深さの最大を 3–10，学習回数(nround)を 50–200 の範囲に設定している．EN setup は Elastic Net の正規化パラメータ(λ)をランダムに選択し，*alpha* を 0–1 の範囲に設定している．NN setup 1 は，ニューラルネットワークの学習のエポック数を 3,000–4,000 の範囲に設定．隠れ層を 3 もしくは 4 に設定し，最初の隠れ層のニューロン数の最大値を 400，隠れ層の最終層のニューロン数を 10–30 の範囲に設定している．NN setup 2 は最初の隠れ層のニューロン数の最大値を 1,640 に設定し，他のパラメータは NN setup 1 と同様に設定している．NN setup 3 はエポック数を 1,000–3,000，隠れ層の数を 3–6 の範囲に設定し，最初の隠れ層のニューロン数は 348 に固定している．隠れ層の最終層のニューロン数の最小値を 5 に設定している．TL setup 1 はランダムフォレストの入力を元ドメインの隠れ層の最終層のニューロンを利用する．nTree と mTry は訓練データ数と訓練データ数の半分の範囲に設定している．TL setup 2 は RF-Py の入力に SPS のベスト訓練済みモデルの全ての隠れ層を連結し，全隠れ層からランダムに選択したニューロンを利用している．nTree は 200 に設定し，選択されたニューロン数の最大値は隠れ層の総ニューロン数の平方根を取った値に設定している．rcdk-all は rcdk で利用可能なフィンガープリント(standard, extended, graph, hybridization, maccs, estate, pubchem, kr, circular)を連結したものを利用したものを表す．RDKit-5 は Atom-Pair，Topological-Torsion フィンガープリント，Morgan フィンガープリント(特徴量ベースの不変量の有・無)，RDKit に含まれる基本的なフィンガープリントを利用したものを表している．XenonPy は XenonPy パッケージに搭載されている化学組成と RDF 記述子を利用したものを表す．(*) 11,106 bits の全記述子の中で 90% 以上が 0 であるフィンガープリントを取り除き，残った記述子の中からランダムに選択(最大 500 個)．(#) *と同様の処理を行い，400–600 もしくは 1,500–3,000 の範囲でランダムに選択．(%) compositional 記述子モデル；1,200，RDF for stable 記述子モデル；1,200，compositional for unsable 記述子モデル；1,200．

る．我々のグループの過去の研究から，転移学習により高い熱伝導率を持つ無機化合物の同定に成功した例を紹介する（Ju et al., 2021）．45 化合物の格子熱伝導率（LTC: lattice thermal conductivity）を予め第一原理計算で算出し，このデータを用いて化学組成から LTC を予測するモデルを構築する．データは 45 件しかないため，何の工夫もない機械学習では予測精度を引き出すことは難しい．そこで，散乱位相空間（SPS: scattering phase space）という中間物性を元ドメインに定め，転移学習で問題解決を図る．LTC に比べると SPS の第一原理計算のコストは圧倒的に低く，320 化合物に対する物性データを用意した．図 2(a)に示すように，SPS と LTC の間には弱い負の相関が存在する．

転移学習によるモデル構築とスクリーニングの手順および結果は，以下の通りである．

(1) XenonPy の *xenonpy.descriptor.Compositions* モジュールで算出した 290 次元の組成記述子を入力とする（記述子の詳細については，Liu et al., 2021 を参照）．
(2) ニューラルネットワークの最大層数を 5 とし，ランダムに 100 個のネットワーク構造を生成し，SPS のモデルを訓練する．ネットワーク構造は，入力層（290 ニューロン）から出力層（1 ニューロン）にかけてニューロンの数が単調減少するように制約する（ピラミッド型）．ハイパーパラメータの詳細については，XenonPy に組み込まれているサンプルコードを参照してほしい．
(3) 100 個の訓練済みモデルを，45 件のデータを用いて LTC の予測モデルに転移する．このとき 10 分割交差検証（クロスバリデーション）を実施し，検証用データセットに対して平均絶対誤差（MAE: mean absolute error）が最も小さい転移モデルを抽出する（図 2(b)）．
(4) 転移モデルで Materials Project の約 140,000 化合物の LTC を予測し，14 個の化合物を同定（選択基準の詳細については，Ju et al., 2021 を参照）．
(5) 第一原理計算で 14 化合物の LTC を検証．

同定された 14 個の化合物の LTC 値に対する検証結果を図 3 に示す．14 個の化合物の LTC は最高で 3,000 W/mK を超える水準に到達している．一方，訓練に使用した 45 個の化合物の LTC は 400 W/mK に満たない領域に分布していることがわかる（図 3 のヒストグラム）．機械学習は「入力が近ければ，出力も近い」という原理に従い予測を行うため，一般的にモデルは訓

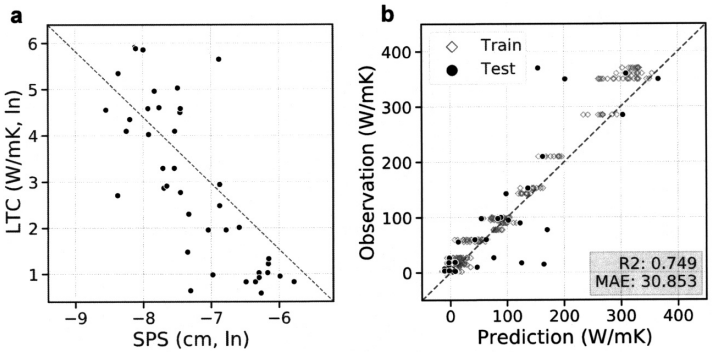

図 2. SPS の訓練済みモデルから LTC への転移学習．(a) SPS（元ドメイン）と LTC（目標ドメイン）の同時分布．SPS と LTC は自然対数の値をプロット．(b) 検証用データに対する SPS（元ドメイン）と転移モデルによる LTC（目標ドメイン）の予測（10 分割交差検証）．横軸と縦軸は予測値と実測値を表す．ダイアモンド（白）とサークル（黒）はそれぞれ訓練データとテストデータを表す．

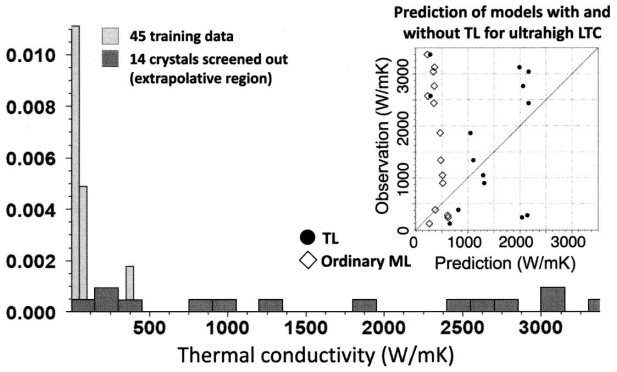

図 3. 転移モデルによる高熱伝導化合物のスクリーニング結果を第一原理計算で検証した結果．ヒストグラム：45 点の訓練データ（灰色）とスクリーニングで同定された 14 化合物の LTC の分布（黒）．LTC は第一原理計算で算出した．散布図：直接訓練したモデル（Ordinary ML）と転移モデル（Transfer Learning: TL）を用いた 14 化合物の予測値の比較．

練データの分布の近傍でのみ予測能力を有する．実際，45 個のデータのみを用いて構築したニューラルネットワーク（直接訓練）は 14 個の化合物の LTC を全く予測できないことがわかる．一方，SPS を経由した転移学習のモデルは，14 個の化合物の LTC をある程度予測できることがわかる．転移学習の予測モデルには，本事例のように外挿性が備わっているケースがしばしば観測されている（Yamada et al., 2019）．元ドメインの 320 個のデータに汎用的な特徴量の獲得に寄与する何らかの情報が含まれており，この特徴抽出器を再利用することで訓練データの水準を大きく超えた領域においても予測性能を有するモデルを構築できた．確かなことは言えないが，これがこの結果に対する自然な解釈である．どのような状況で外挿性が発現するのかは分からない．

3.2 結晶性ポリマーと無機結晶化合物の屈折率

次に結晶性ポリマーと無機結晶化合物における転移学習の事例を紹介する．ここでの予測対象は，有機分子と無機化合物の屈折率である．

Polymer Genome（Mannodi-Kanakkithodi et al., 2018）という高分子物性のデータベースには，第一原理計算で算出した 853 個の高分子の屈折率が収録されている．無機化合物の屈折率については，Citrination（https://citrination.com/）というデータベースから抽出した 1,056 件のデータを使用した．高分子，無機化合物ともに構造情報を一切利用せず，モデルの入力変数は組成のみとし，XenonPy の 290 次元組成記述子を用いて屈折率を予測した．

図 4 は，全サンプルの記述子行列と物性値の関係を可視化したヒートマップである．この図から高分子と無機化合物の各々の記述子と物性の相関パターンが読み取れるが，高分子と無機化合物の間に共通性はほとんどないことが分かる．図 5（b）は，t-SNE（van der Maaten and Hinton, 2008）という次元圧縮の手法を用いて 290 次元の記述子ベクトルを二次元平面に配置した結果である．この図からも分かるように，高分子と無機化合物の記述子ベクトルは，特徴空間上でかなり離れた位置に分布しており，無機化合物の組成パターンのわずか一部にのみ，高分子とのオーバーラップがみられる．

無機化合物から高分子への転移学習と高分子から無機化合物への転移学習を行った．4 層の隠れ層からなるピラミッド型のニューラルネットワークで元ドメインの訓練済みモデルを構築

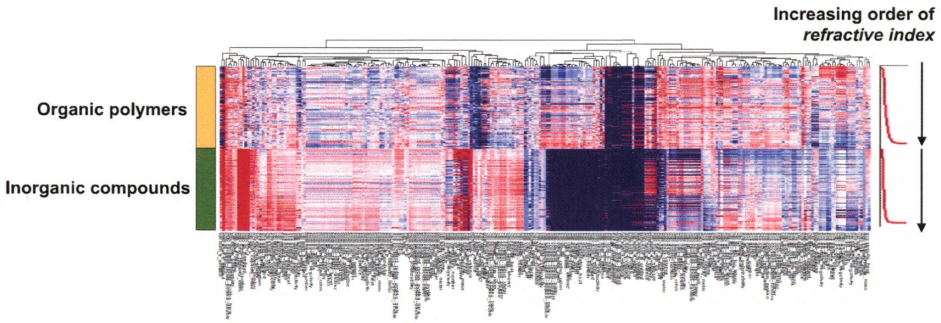

図 4. 1,056 個の無機化合物と 853 個の高分子の組成記述子(横軸)のヒートマップ表示．屈折率の昇順にサンプルを上から下に並び替えている．屈折率と相関の高い記述子には，何らかの特徴的なパターンがみられる．高分子と無機化合物の間には，ほとんど共通性がないことが分かる．

し，最上位の隠れ層を記述子に用いて．ランダムフォレストで目標ドメインのモデルに転移した．ハイパーパラメータ等の詳細は，論文 Yamada et al. (2019)を参照せよ．元ドメインの訓練には全データを使用し，転移の際は全データの 80% を訓練，残りをテストデータに使用した．

まずは，無機化合物から高分子への転移学習の結果をみてみる．図 5(a)に示すように，無機のデータで訓練したモデルに高分子の組成を入力しても屈折率をほとんど予測できない(図 5(a)上図)．一方，無機化合物のモデルを高分子に転移したモデルは，高分子の屈折率を高い精度で予測できる(図 5(a)下図)．図 4 や図 5(a)で示したように，一見すると無機化合物と高分子の屈折率の間には共通性はほとんどなさそうである．それにもかかわらず，転移学習の結果は両者の間に何らかの共通性が存在することを示唆しているが，この結果は極めて非直感的である．

一方，高分子から無機化合物に転移したモデルは，無機化合物の屈折率を全く予測できないことがわかる(図 6(a)(b))．この転移の不可逆性は，転移学習の本質の一つを捉えている．高分子の組成データには計 17 種類の元素しか含まれておらず，C, H, O, Cl の元素に偏っている．一方，無機化合物の組成データは遷移金属を含む計 63 種類の元素から構成されている．したがって，構成元素の包含関係としては，高分子の組成は無機化合物の部分集合ということになる．したがって，高分子の訓練済みモデルには 17 種類以外の元素に対する表現能力が備わっておらず，無機化合物の入力組成は外挿領域に存在すると考えられる．

3.3 高分子の熱伝導率

高分子物性データベース PoLyInfo (Otsuka et al., 2011)に収録されている 19 個のアモルファスポリマーの熱伝導率のデータを使用する．データの選定と前処理の方法については，論文 Wu et al. (2019)を参照せよ．

高分子のガラス転移温度，融点，定圧比熱容量，粘度に加えて，低分子化合物の定積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のデータは QM9 (Ramakrishnan et al., 2014; Ruddigkeit et al., 2012)という第一原理計算の物性データベースから抽出した．後述の手順で各々の元ドメインに対して 1,000 個の異なるモデルを構築し，これらのモデルを 19 個のデータを用いて熱伝導率の予測モデルに転移した．ここで 5 分割交差検証を適用して転移学習モデルの汎化性能を評価し，MAE が最小のモデルを抽出する．

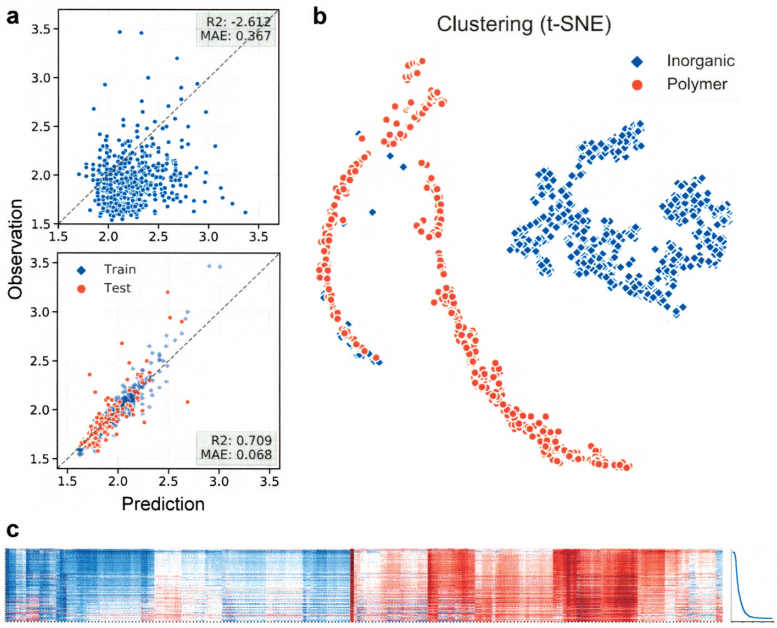

図 5. 無機化合物から高分子屈折率への転移学習．(a)上図は無機のデータで訓練したモデルに高分子の組成を入力した際の屈折率の予測値を表す．下図は，無機化合物の訓練済みモデルを高分子のモデルに転移した結果．訓練データとテストデータはダイアモンド(青色)とサークル(橙色)で表される．両図とも，横軸は予測値，縦軸は観測値を表す．(b)t-SNE による無機化合物と高分子の組成記述子の二次元平面への可視化．無機結晶化合物と結晶性ポリマーはダイアモンド(青色)とサークル(橙色)で表される．(c)無機化合物の訓練済みモデルに 853 個の高分子の組成を入力した際の隠れ層(特徴量)をヒートマップで可視化した結果．853 個のサンプルは屈折率の大きさ順に並び替えている．転移学習では，これらの特徴量を記述子とした．

モデルの入力にはモノマーの化学構造のみを用いた．RDKit に実装されている 9 種類のフィンガープリント記述子(ECFP, FCFP, MACCS など)を連結し，11,106 次元の記述子ベクトルを構築した．その中からランダムに抽出した 400~600 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てニューロン数と層の数をランダムに選択した．各々の元ドメインにおいて，このような訓練済みモデルをランダムに 100 個作り，最終隠れ層を記述子としてランダムフォレストでモデルを構築した．100 個の転移モデルの内，目標ドメインの 5 分割交差検証の平均 MAE を最小にする転移モデルを選定した．

図 7 に，各々の元ドメインからの転移の結果を示している．それぞれ，目標ドメインの MAE が最も小さかった転移モデルの 5 分割交差検証の結果を示している．さらに，図 7 には，転移学習を適用せずに直接訓練したモデルの 5 分割交差検証の結果も示している．転移学習を経由しない通常のモデルの汎化性能は極めて低い．一方，全ての元ドメインにおいて，転移モデルは通常のモデルの汎化精度を大きく上回った．

Wu et al. (2019)では，低分子化合物の定圧比熱容量からの転移モデルを選択し，モデルの逆問題を解き，高い熱伝導率に達すると予想されるモノマー分子を設計した．最終的に 3 種類

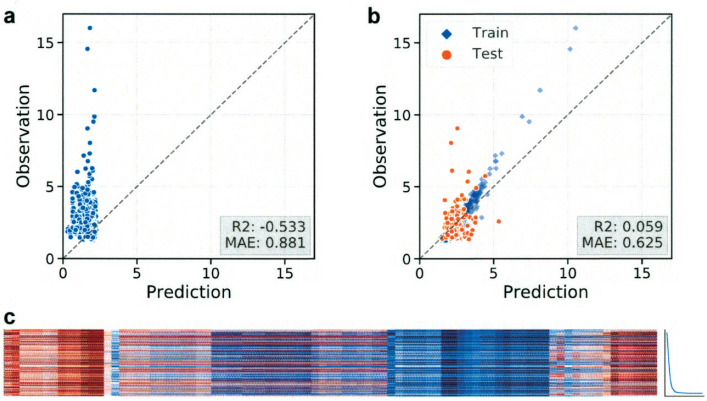

図 6. 高分子から無機化合物への転移学習．(a)高分子のデータで訓練したモデルに無機化合物の組成を入力した屈折率の予測値を表す．横軸は予測値，縦軸は観測値を表す．(b)高分子の訓練済みモデルを無機化合物のモデルに転移した結果．訓練データとテストデータはダイアモンド(青色)とサークル(橙色)で区別される．(c)高分子屈折率の訓練済みモデルに無機化合物の組成を入力した際の隠れ層(特徴量)をヒートマップで可視化した結果．サンプルは屈折率の大きさ順に並び替えている．転移学習では，これらの特徴量を記述子とした．

の芳香族ポリアミドに候補を絞り込み，ポリマー合成と物性測定を実施した．合成された高分子の一つは，熱伝導率が 0.41 W/mK に達することが確認された(詳細は Wu et al., 2019 を参照)．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．図 8 に示すように，熱伝導率の実験値は転移モデルの予測値と概ね一致している．ここで注目すべきは，合成した高分子と類似する化学構造は 19 個の訓練データにほとんど含まれていない点である．無機化合物の熱伝導率のケースと同様に，転移学習の外挿性が観察された．

4. まとめ

本稿では，材料研究の複数のタスク(スモールデータに基づく低分子・高分子・無機化合物の物性予測)を例に取り上げ，転移学習が有する潜在的な予測性能を実験的に示した．材料研究のスモールデータの限界を乗り越える上で，転移学習の活用は有望な解決方策を与える．本研究では，低分子，高分子，無機化合物の 45 種類の特性を対象に約 140,000 個の機械学習の予測モデルを開発し，訓練済みモデルライブラリ XenonPy.MDL に実装した．XenonPy のユーザーは，API を用いてライブラリから転移元モデルの候補を取得し，転移されたモデルを用いて材料設計のワークフローを構築できる．優れた研究者が過去の経験から大量且つ多様な記憶を獲得しているのと同様に，転移学習の成功の鍵は，包括的な訓練済みモデルライブラリを実装することにあると考えている．

本稿では特に転移学習が有する外挿性獲得のメカニズムに注目した．一般に革新的な材料の周辺にはデータが存在しない．しかしながら，機械学習のモデルは訓練データとテストデータの類似性に基づいて予測を行うため，周辺にデータが存在しない外挿領域では予測能力を失う．一方，本稿で示したように，転移学習を巧みに適用することで，極めて少ない訓練データでも時に外挿的といえる予測モデルを構築できる．物理化学的な関連性を持つ元ドメインにおいて，広い物質空間に分布する大量のデータを用いて事前学習を行う．この過程で広い物質

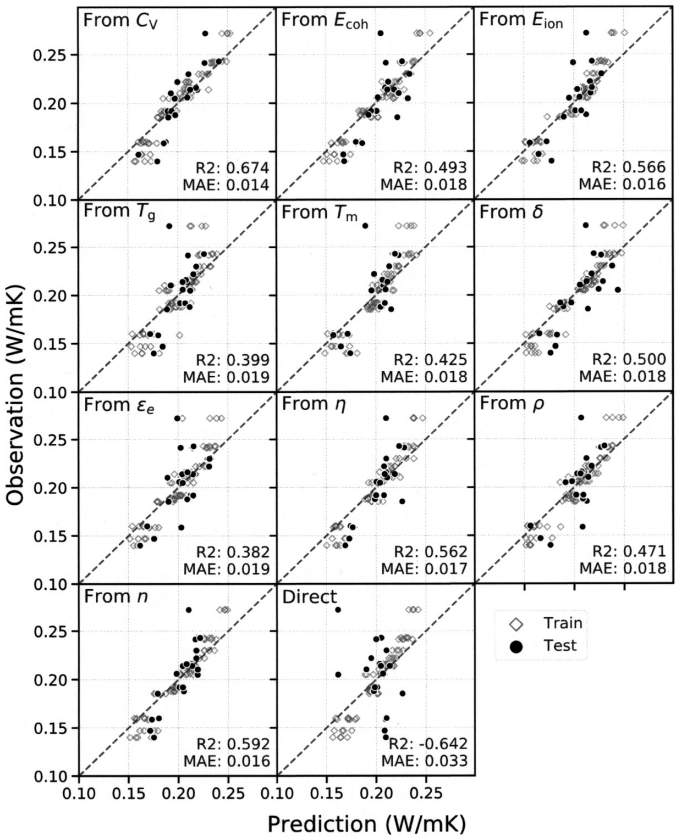

図 7. 様々な元ドメイン(高分子ポリマーの凝集エネルギー E_{coh}，イオン化エネルギー E_{ion}，ガラス転移温度 T_{g}，融点 T_{m}，溶解度パラメーター δ，誘電率 ϵ_e，粘度 η，密度 ρ，屈折率 n ，低分子化合物の定圧比熱 C_{V})から高分子熱伝導率への転移と通常の教師あり学習(Direct)の 5 分割交差検証の結果．横軸は予測値，縦軸は観測値を表す．ダイアモンド(白)とサークル(黒)は，それぞれ訓練データとテストデータを表す．

空間で適用可能な汎用的な特徴表現を獲得できれば，これを目標ドメインに転移することで，データの範囲外での予測能力を有するモデルを構築できる．このことが実験的に観測された．しかしながら，転移学習の外挿性獲得のメカニズムはほぼ未解明である．外挿性の発現には元ドメインの選択や訓練データの分布のパターンが関与していることが示唆されているが，この現象を説明できる理論はまだ確立されていない．

謝　辞

本研究は科研費 18K18017 の助成を受けて遂行されたものです．また，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吉田亮教授と野口瑶特任研究員から有益な助言をいただきました．データの共有については，東京大学の塩見淳一郎教授，東京工業大学の森川淳子教授，物質・材料研究機構の小山幸典主幹研究員から支援を受けました．この場を借りて深く御礼申し上げます．

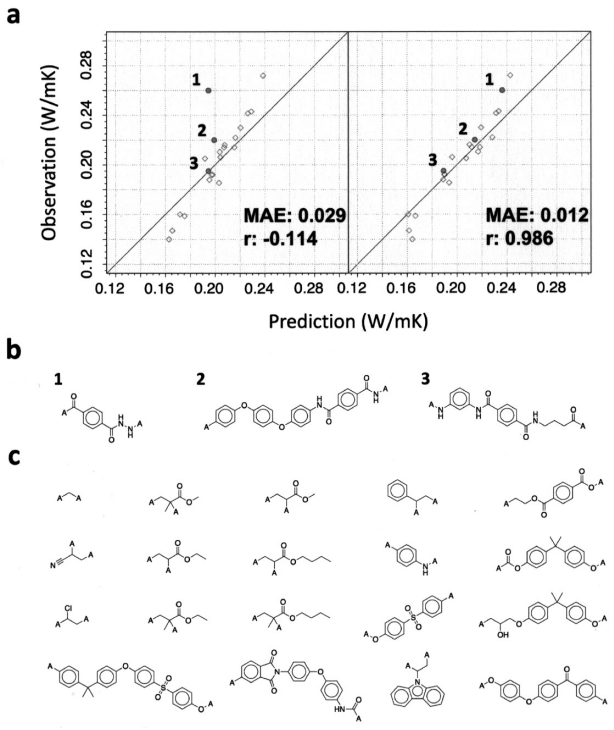

図 8. 転移モデルによる熱伝導率の予測精度．(a)新たに合成した 3 種類の高分子(サークルで表示，番号を付与)に対する通常の教師あり学習のモデル(左)と転移モデル(右)の予測値と実測値．(b)合成した 3 種類の芳香族ポリアミドのモノマー．(c)転移学習に用いた 19 個の訓練データの化学構造．

参 考 文 献

Agrawala, A. and Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science, *APL Materials*, **4**(5), p.053208, DOI: http://dx.doi.org/10/gd7d53.

Carrete, J., Li, W., Mingo, N., Wang, S. and Curtarolo, S. (2014). Finding unprecedentedly low-thermal-conductivity half-heusler semiconductors via high-throughput materials modeling, *Physical Review X*, **4**(1), 011019–011019, DOI: http://dx.doi.org/10/gbfqzz.

Cubuk, E. D., Sendek, A. D. and Reed, E. J. (2019). Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data, *The Journal of Chemical Physics*, **150**, p.214701, DOI: http://dx.doi.org/10/gf3k4k.

Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T. D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M. A., Chae, H. S., Einzinger, M., Ha, D.-G., Wu, T., Markopoulos, G., Jeon, S., Kang, H., Miyazaki, H., Numata, M., Kim, S., Huang, W., Hong, S. I., Baldo, M., Adams, R. P. and Aspuru-Guzik, A. (2016). Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach, *Nature Materials*, **15**(10), 1120–1127, DOI: http://dx.doi.org/10/f859z9.

Hansen, E. C., Pedro, D. J., Wotal, A. C., Gower, N. J., Nelson, J. D., Caron, S. and Weix, D. J. (2016). New ligands for nickel catalysis from diverse pharmaceutical heterocycle libraries, *Nature Chemistry*, **8**(12), 1126–1130, DOI: http://dx.doi.org/10/f9dvx3.

Hutchinson, M. L., Antono, E., Gibbons, B. M., Paradiso, S., Ling, J. and Bryce Meredig (2017). Overcoming data scarcity with transfer learning, arXiv:1711.05099.

Jalem, R., Kanamori, K., Takeuchi, I., Nakayama, M., Yamasaki, H. and Saito, T. (2018). Bayesian-driven first-principles calculations for accelerating exploration of fast ion conductors for rechargeable battery application, *Scientific Reports*, **8**(1), p.5845, DOI: http://dx.doi.org/10/gdfwfv.

Ju, S., Yoshida, R., Liu, C., Wu, S., Hongo, K., Tadano, T. and Shiomi, J. (2021). Exploring diamond-like lattice thermal conductivity crystals via feature-based transfer learning, *Physical Review Materials* (in press).

Kailkhura, B., Gallagher, B., Kim, S., Hiszpanski, A. and Han, T. Y.-J. (2019). Reliable and explainable machine learning methods for accelerated material discovery, arXiv:1901.02717.

Kaya, M. and Hajimirza, S. (2019). Using a novel transfer learning method for designing thin film solar cells with enhanced quantum efficiencies, *Scientific Reports*, **9**, p.5034, DOI: http://dx.doi.org/10/gjw4n7.

Li, X., Zhang, Y., Zhao, H., Burkhart, C., Brinson, L. C. and Chen, W. (2018). A transfer learning approach for microstructure reconstruction and structure-property predictions, *Scientific Reports*, **8**, p.13461.

Liu, C., Fujita, E., Katsura, Y., Inada, Y., Ishikawa, A., Tamura, R., Kimura, K. and Yoshida, R. (2021). Machine learning to predict quasicrystals from chemical compositions (preprint, in review), DOI: http://dx.doi.org/10.21203/rs.3.rs-240290/v1.

Mannodi-Kanakkithodi, A., Chandrasekaran, A., Kim, C., Huan, T. D., Pilania, G., Botu, V. and Ramprasad, R. (2018). Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond, *Materials Today*, **21**(7), 785–796, DOI: http://dx.doi.org/10/gd7q4v.

Matsumoto, R., Hou, Z., Hara, H., Adachi, S., Takeya, H., Irifune, T., Terakura, K. and Takano, Y. (2018). Two pressure-induced superconducting transitions in SnBi2Se4 explored by data-driven materials search: New approach to developing novel functional materials including thermoelectric and superconducting materials, *Applied Physics Express*, **11**(9), 093101–093101, DOI: http://dx.doi.org/10/gjw4nw.

Oda, H., Kiyohara, S., Tsuda, K. and Mizoguchi, T. (2017). Transfer learning to accelerate interface structure searches, *Journal of the Physical Society of Japan*, **86**(12), p.123601, DOI: http://dx.doi.org/10/gjv2v9.

Oliynyk, A. O., Antono, E., Sparks, T. D., Ghadbeigi, L., Gaultois, M. W., Meredig, B. and Mar, A. (2016). High-throughput machine-learning-driven synthesis of Full-Heusler Compounds, *Chemistry of Materials*, **28**(20), 7324–7331, DOI: http://dx.doi.org/10/f88n5s.

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, DOI: http://dx.doi.org/10/fhvjd8.

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), p.140022, DOI: http://dx.doi.org/10/gdq9k4.

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, DOI: http://dx.doi.org/10/f4d9mt.

Segler, M. H. S., 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, DOI: http://dx.doi.org/10/gcwpxd.

Seko, A., Togo, A., Hayashi, H., Tsuda, K., Chaput, L. and Tanaka, I. (2015). Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization, *Physical Review Letters*, **115**(20), 205901–205901, DOI: http://dx.doi.org/10/f8d2ww.

Sumita, M., Yang, X., Ishihara, S., Tamura, R. and Tsuda, K. (2018). Hunting for organic molecules with artificial intelligence: Molecules optimized for desired excitation energies, *ACS Central Science*, **4**(9), 1126–1133, DOI: http://dx.doi.org/10/gfcpxs.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using T-SNE, *Journal of Machine Learning Research*, **9**, 2579–2605.

Wu, S., Kondo, Y., aki Kakimoto, M., 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), DOI: http://dx.doi.org/10/gf6mkg.

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**, 1717–1730, DOI: http://dx.doi.org/10/ggrbd7.

Yonezu, T., Tamura, T., Takeuchi, I. and Karasuyama, M. (2018). Knowledge-transfer-based cost-effective search for interface structures: A case study on Fcc-Al [110] tilt grain boundary, *Physical Review Materials*, **2**(11), p.113802, DOI: http://dx.doi.org/10/gjw4nx.

Application of Transfer Learning in Materials Research

Chang Liu[1], Hironao Yamada[1,2] and Stephen Wu[1,3]

[1]The Institute of Statistical Mathematics
[2]School of Pharmacy, Tokyo University of Pharmacy and Life Sciences
[3]Department of Statistical Science, School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies, SOKENDAI

The digital transformation of materials research has resulted in a broad array of materials property databases; however, the available databases do not include advances realized in machine learning. Transfer learning is a machine learning framework with potential to break the barrier and identify various properties that are physically interrelated. For a given target property to be predicted from a limited supply of training data, models on related proxy properties are pre-trained using enough data to capture the common features relevant to the target task. Repurposing such machine-acquired features for a target task results in an outstanding predictive power even with exceedingly small data. We demonstrate transfer learning in various real-world applications, including property prediction of polymers and inorganic materials. In particular, we show several examples in which transfer learning is applied to obtain a predictive capability in a domain that greatly deviates from the training data distribution.

Key words: Transfer learning, novel material design, crystalline, molecular, polymer.