超伝導材料の転移温度予測における事例間の繋がりを考慮した知識グラフの有効性の調査

吉野草太 [1]　旭良司 [2]　三輪誠 [1]　佐々木裕 [1]
[1] 豊田工業大学　[2] 名古屋大学
{sd20109,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp
ryoji.asahi@chem.material.nagoya-u.ac.jp

概要

近年，データ駆動で材料開発を行うマテリアルズ・インフォマティクスの発展に伴い，属人的な知見への依存を減らし，より高性能に物性値を予測するための研究が進められている．超伝導材料は高い応用可能性により開発が急がれているが，膨大な探索範囲から目的に見合った転移温度を持つ材料を発見するのは困難である．そのため，組成式や構造情報から転移温度を予測し，材料探索を加速化するアプローチが研究され，その 1 つにデータベースから作成した知識グラフを用いて転移温度予測を行う手法がある．しかし，既存手法の知識グラフは事例間の繋がりを考慮できていない．そこで，本研究では材料が組成式・構造名・論文題名の頂点を介して繋がる知識グラフを用いた転移温度予測モデルを提案する．実験により，SuperCon データセットにおいて既存手法と比較して RMSE が 0.072 減少することが明らかとなった．さらに，材料間の繋がりを考慮した知識グラフの転移温度予測に対する有効性を確認した．

1　はじめに

近年，データ駆動で材料開発を行うマテリアルズ・インフォマティクス（Materials Informatics; MI）の発展に伴い，属人的な知見への依存を減らした効率的な材料開発が活発化している．MI による材料開発のアプローチの 1 つとして，機械学習を用いて材料の物性値を予測する研究がある．材料の組成式，構造情報などを入力として材料の物性値を予測する手法が提案されており，より高性能に物性値を予測するための研究が進められている [1, 2, 3, 4, 5, 6].

材料研究において，注目されている材料の一つに超伝導材料がある．超伝導材料はある温度（転移温度）以下で電気抵抗がゼロとなる特性から，工学・医学などの幅広い分野で応用されており，その開発は急務である．超伝導材料の用途やコストは転移温度に依存するため，目的に見合った転移温度を持つ材料の開発が求められているが，膨大な材料組成の探索範囲から所望の超伝導材料を発見するのは困難である．そのため，組成式や構造情報から転移温度を予測し，所望の転移温度を記録した組成式や構造情報をもとに材料を作製するアプローチが研究されており，超伝導材料データベース (SuperCon [7], 3DSC [8], Superconducting Research Database [9] など）を用いた予測モデル [10, 11, 12, 13] が提案されている．

材料が似た組成式を持つ・同じ構造を持つなどの関係があるとき，その材料の転移温度は関連を持つ可能性がある．また，同じ論文で報告されている材料は，何らかの観点で似た性質の材料である可能性がある．このような材料間の繋がりを考慮するための方法の 1 つとして，データを頂点と辺で表現する知識グラフに変換し，グラフニューラルネットワーク (Graph Neural Network; GNN) で学習する方法が考えられる．材料分野で GNN を利用した手法の 1 つとして，Hatakeyama ら [14] はテーブルデータベースに記録されている材料の各事例を，組成式・構造・文字を頂点で，それぞれの頂点の関係を辺で表現する知識グラフに変換して，GNN を学習することで，材料分野の物性値予測において高い予測性能を示した．しかし，この方法では，事例ごとに知識グラフを作成するため，事例間を繋ぐ辺が無く，材料間の繋がりを考慮できない．さらに，転移温度予測に対して，知識グラフを利用した例はなく，その有効性は未知のままである．

そこで本研究では，超伝導材料の転移温度予測に

おける知識グラフの有効性の調査を目的として，材料の組成式・構造・論文題名からなる知識グラフを入力とした転移温度予測モデルを提案する．知識グラフにより，組成式・構造・論文において事例間で繋がりを持つ材料が互いに及ぼし合う影響を捉えることができれば，より高性能な予測ができると期待できる．本研究の貢献は以下の通りである．

- 超伝導材料分野における知識グラフを利用した転移温度予測モデルの提案
- データベースの事例間の繋がりによる転移温度予測への有効性の確認

2 関連研究

材料の組成式や構造を入力として特定の物性値を予測する物性値予測が広く研究されている．物性予測は様々な情報を入力として用いる手法 [2, 3] と組成式のみを入力に用いる手法 [1, 5, 6] に大別される．様々な情報を入力として用いる手法は一般に高性能だが，材料に対する情報量に偏りがあり不均一である．一方，組成式のみを入力に用いる手法は組成式だけを必要とするため，運用が手軽だが性能は低い．

Wang ら [6] は組成式から注意機構を用いて物性値を予測する CrabNet を提案した．CrabNet は，組成式について元素と比率をそれぞれ符号化した入力を，Transformer [15] のエンコーダ部を参考にしたモデルに与えて物性値を予測する．組成式の比率の符号化において，比率を線形変換したものと対数変換したものを符号化することで，組成式の比率の違いに敏感な材料，たとえばドーピングの影響を受けやすい材料に対応し，様々なデータセットで高い予測性能を示している．

Hatakeyama ら [14] はテーブルデータベースを変換した知識グラフに GNN を適用し，物性予測を行う手法を提案した．テーブルデータベースでは，異なるデータベースの統合や実験手順などの記録が難しい，といった問題点が存在する．そのため，テーブルデータをグラフに変換して 47 種類の物性値を 1 つのモデルで予測し，高い予測性能を実現した．しかし，このグラフデータベースは 1 つの事例を 1 つのグラフに変換して作成されるため，事例間の繋がりが無い．また，超伝導材料分野において，テーブルデータをグラフで表す有効性は明らかになっていない．

3 提案手法

超伝導材料を対象に，知識グラフを入力とする転移温度予測モデルを提案する．3.1 節で超伝導材料データベースからの知識グラフの作成について説明し，3.2 節で知識グラフを入力とする転移温度予測モデルについて説明する．提案手法の概要を図 1 に示す．

3.1 知識グラフの作成

超伝導材料データベースに登録されている全ての事例から知識グラフを作成する．処理前組成式・処理後組成式・構造名・論文題名を知識グラフの頂点とし，処理後組成式が材料の中心的な属性と考えて，処理前組成式・構造名・論文題名から処理後組成式へそれぞれ異なる関係の有向辺を張る．このようなグラフの構造にすることで，複数の処理後組成式が処理前組成式・構造名・論文題名を介して次のように知識グラフ上で繋がることで，学習する際に，共通点を持った処理後組成式の表現同士が影響し合うことを期待する．例えば，処理後組成式の表現の単なる線形結合ではなく，介す頂点が処理前組成式・構造名・論文題名のどれであるかによってそれぞれ異なる表現が伝わって混ざり合う．

まず，処理前組成式に様々な処理を施したものが処理後組成式であるため，大体の組成は一致している．この 2 つを辺で繋ぐことで，似た組成の処理後組成式が同じ処理前組成式を介して繋がる．次に，処理後組成式とその構造名を辺で繋ぐことで，同じ構造を持つ処理後組成式が構造名を介して繋がる．最後に，処理後組成式とそれが報告された論文題名を繋ぐことで，同じ論文で報告された処理後組成式が論文題名を介して繋がる．図 2 に具体的な値を用いた知識グラフの作成手順を示す．

3.2 転移温度予測モデル

3.1 節で作成した知識グラフを用いて転移温度予測を行う．知識グラフのそれぞれの頂点の表現を初期化した後，Relational Graph Convolutional Networks (RGCN) [16] と全結合層により転移温度と不確実度を出力し，CrabNet と同様に RobustL1 損失で学習する．

RGCN に入力する知識グラフの頂点である処理前組成式・処理後組成式・構造名・論文題名の表現はそれぞれ異なる方法で初期化する．処理前組成

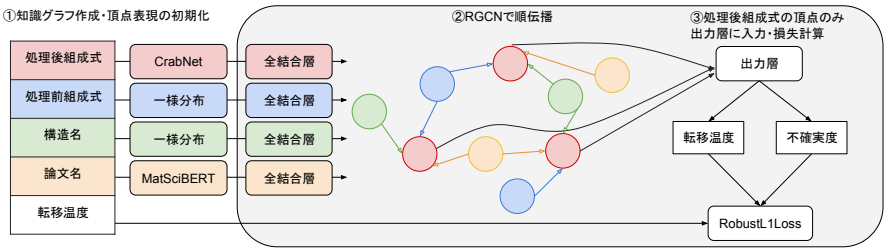

図 1　提案モデルの全体像

超伝導材料データベース

処理後組成式	処理前組成式	構造名	論文題名
A1.9B3.1	A2-xB3+x	X	xxx
A1.7B3.3	A2-xB3+x	Y	yyy
A1B1C3	(A,B)2C3	Y	xxx

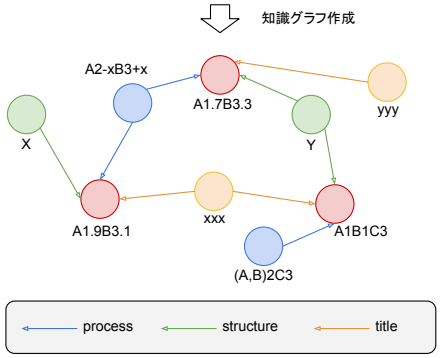

図 2　知識グラフの具体的な作成手順

式と構造名は一様分布の乱数（1,024 次元），処理後組成式は CrabNet の表現（1,024 次元），論文題名は MatSciBERT [17]（768 次元）により，それぞれ初期化する．ここで用いた CrabNet と MatSciBERT は出力のみ用いているため，提案手法のネットワークには繋がっていない．ここで処理前組成式は，処理後組成式と異なり括弧や未定義元素が使用されており CrabNet では初期化できないため，一様分布の乱数で初期化する．また，CrabNet の表現は組成式の最大元素数を 8 で固定することで (8，128) 次元の行列となっているが，RGCN に入力するため 1,024 次元のベクトルに変形して使用する．

ここでいう CrabNet の表現とは，CrabNet 内の ResidualNetwork の最終層から出力される表現である．CrabNet では ResidualNetwork の後に，(8, 128) 次元の表現を (8, 3) 次元にする全結合層と，(8, 3) 次元を 2 次元（転移温度，不確実度）にする物性値予測層が存在する．以降では，この全結合層と物性値予測層を合わせて出力層と呼ぶ．

頂点の表現を初期化した知識グラフを RGCN に入力し，順伝播後の処理後組成式の頂点の表現を出力層に入力し，転移温度を出力する．知識グラフを RGCN に入力する前に，処理前組成式・処理後組成式・構造名・論文題名の頂点の表現をそれぞれ全結合層に入力し，次元を 512 次元に揃える．頂点の表現の次元を揃えた知識グラフを，頂点の表現の次元を保ちながら，2 層の RGCN に入力する．RGCN から出力された知識グラフのうち，処理後組成式の頂点の表現のみ使用する．処理後組成式の頂点の表現は 1,024 次元であるが，CrabNet の出力層に入力するため (8,128) 次元に変形する．変形後の処理後組成式の頂点の表現を出力層に入力することで転移温度と不確実度が出力される．

このように RGCN を導入することにより，処理後組成式と処理前組成式・構造名・論文題名の辺を区別しつつ，頂点の表現同士を影響させ合うことができる．処理後組成式の頂点の表現のみを全結合層に与える理由は，処理後組成式と転移温度が 1 対 1 に対応しているためである．知識グラフの中に無い処理後組成式に対しては転移温度を予測することができないため，開発データ・テストデータの処理後組成式も含めて知識グラフを作成し，トランスダクティブな学習を行う．

上述した転移温度予測モデルを数式で表す．作成した知識グラフを G，次元を揃える前のそれぞれの頂点の表現を H_i，$i \in \{$ 処理前組成式，処理後組成式，構造名，論文題名 $\}$ とすると，転移温度予測モデルは次のように転移温度 $output$ と不確実度 σ を予測する．

$$H_i' = \mathrm{FC}_i(H_i) \tag{1}$$

$$X = \mathrm{Concat}(H_i') \tag{2}$$

$$X' = \mathrm{RGCN}(G, X) \tag{3}$$

$$X''_{\text{処理後組成式}} = \mathrm{Reshape}(X'_{\text{処理後組成式}}) \tag{4}$$

$$X''' = \mathrm{FC}(X''_{\text{処理後組成式}}) \tag{5}$$

$$output, \sigma = \mathrm{CrabNetOutput}(X''') \tag{6}$$

表 1 転移温度予測の結果 [K]．太字は列における最高のスコアを示す

Method	MAE(↓)		RMSE(↓)	
	評価	テスト	評価	テスト
CrabNet [6]	9.259 ± 0.092	**7.202 ±0.178**	14.617 ± 0.153	17.048 ± 0.204
提案手法	**9.001 ± 0.119**	7.396 ± 0.176	**14.118 ± 0.306**	**16.976 ± 0.164**

ここで予測された転移温度 $output$ について，CrabNet で用いられている以下の RobustL1 関数を用いて転移温度の正解 $target$ と比較することで，損失を計算する．

$$L = \sqrt{2}\exp(\sigma)|output - target| + \sigma \tag{7}$$

4 実験と考察

4.1 実験設定

超伝導材料データセットとして，MDR SuperCon[7] を用いる．超伝導材料に対応する 33,407 件の事例それぞれについて記録されている 200 種類以上の属性の中から，処理前組成式として name 列，処理後組成式として element 列，構造名として str3 列，転移温度として tc 列，論文題名として title 列，論文出版年として year 列を取り出す．この際，処理前組成式・処理後組成式・転移温度・論文出版年に欠損値を含む事例を削除する．それぞれ 1913 年〜2003 年の事例を訓練データ，2004 年〜2013 年の事例を開発データ，2014 年〜2021 年の事例をテストデータとして，おおよそ 7:2:1 の割合で分割した．

データの分割とは別に，SuperCon の全事例から知識グラフを構築した（付録 A）．全事例から知識グラフを作成するため，開発データやテストデータの処理前組成式・処理後組成式・構造名・論文題名の頂点の表現も訓練時に参照する．開発・テスト時において，RGCN までは訓練時と同様に順伝播が行われるが，その後は開発データ・テストデータに含まれる処理後組成式のみを全結合層に入力し，転移温度を出力する．本研究で提案するモデルは知識グラフにある処理後組成式の転移温度を予測するため，新たな処理後組成式の追加は想定していない．また，知識グラフの頂点の表現を初期化する際に使う CrabNet は訓練データで事前に学習しておく．

4.2 実験結果

SuperCon に記録されている超伝導材料のうち，評価データとテストデータに対して転移温度予測を行った結果を表 1 に示す．SuperCon から構築した知識グラフを用いることで，評価データにおいては提案手法が CrabNet の性能を上回る結果となったが，テストデータにおいては CrabNet に比べて MAE が約 0.2 大きく，RMSE が約 0.1 小さい結果となった．

表 2 アブレーションの結果．太字は列における最高スコアを示す．

Method	MAE(↓)	RMSE(↓)
w/o 処理前組成式	9.055 ± 0.062	14.163 ± 0.153
w/o 構造名	9.095 ± 0.094	14.304 ± 0.202
w/o 論文題名	9.066 ± 0.105	14.249 ± 0.204
提案手法	**9.001 ± 0.119**	**14.118 ± 0.306**

4.3 考察

提案手法による性能向上が知識グラフによって事例間の繋がりを考慮したことによるのかを調べるために追加実験を行った結果を表 2 に示す．具体的には，提案手法から処理前組成式・構造名・論文題名の頂点をそれぞれ除外したアブレーションを行った．すべての場合において提案手法の性能より下回っており，事例間の繋がりの有効性を示唆している．性能の低下が最も大きいのは構造名を抜いた場合であった．

5 おわりに

本研究では，超伝導材料の転移温度予測に対する知識グラフの有効性の確認を目的として，超伝導材料データベースから作成した知識グラフを入力とする転移温度予測モデルを提案した．SuperCon を用いて提案手法の学習・評価を行った結果，転移温度の RMSE が 0.072 減少した．また，提案手法における性能向上は，材料に関する多様な項目を考慮することに依ると確認できた．

今後は，括弧や変数を含む処理前組成式と構造名の適切に初期化する，SuperCon 以外のデータベースを追加するなどの改善を行い，転移温度予測のさらなる精度向上や，学習した頂点の表現の他タスクへの活用を目指す．

謝辞

本研究は JSPS 科研費 23K11237 の助成を受けたものです.

参考文献

[1] Dipendra Jha, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. Elemnet: Deep learning the chemistry of materials from only elemental composition. **Scientific Reports**, Vol. 8, No. 1, p. 17593, Dec 2018.

[2] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. **The Journal of Chemical Physics**, Vol. 148, No. 24, 2018.

[3] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. **Physical review letters**, Vol. 120, No. 14, p. 145301, 2018.

[4] Dipendra Jha, Logan Ward, Zijiang Yang, Christopher Wolverton, Ian Foster, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Irnet: A general purpose deep residual regression framework for materials discovery. In **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, KDD '19, p. 2385–2393, New York, NY, USA, 2019. Association for Computing Machinery.

[5] Rhys EA Goodall and Alpha A Lee. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. **Nature communications**, Vol. 11, No. 1, p. 6280, 2020.

[6] Anthony Yu-Tung Wang, Steven K Kauwe, Ryan J Murdock, and Taylor D Sparks. Compositionally restricted attention-based network for materials property predictions. **Npj Computational Materials**, Vol. 7, No. 1, p. 77, 2021.

[7] National Institute for Materials Science Materials Database Group. Mdr supercon datasheet ver.220808. https://doi.org/10.48505/nims.3837, 12 2022. Accessed: 2023-07-17.

[8] Timo Sommer, Roland Willa, Jörg Schmalian, and Pascal Friederich. 3dsc - a dataset of superconductors including crystal structures. **Scientific Data**, Vol. 10, No. 1, p. 816, Nov 2023.

[9] Ivan K. Schuller. Discovery of new superconductors. Technical Report AD1103101, University of California San Diego United States, 3 2020. APPROVED FOR PUBLIC RELEASE.

[10] Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. **Computational Materials Science**, Vol. 154, pp. 346–354, 2018.

[11] Kaname Matsumoto and Tomoya Horide. An acceleration search method of higher tc superconductors by a machine learning algorithm. **Applied Physics Express**, Vol. 12, No. 7, p. 073003, jun 2019.

[12] Tomohiko Konno, Hodaka Kurokawa, Fuyuki Nabeshima, Yuki Sakishita, Ryo Ogawa, Iwao Hosako, and Atsutaka Maeda. Deep learning model for finding new superconductors. **Phys. Rev. B**, Vol. 103, p. 014509, Jan 2021.

[13] Yutaka Sasaki Kento Mitsui and Ryoji Asahi. Automatic knowledge acquisition from superconductivity information in literature. **Science and Technology of Advanced Materials: Methods**, Vol. 3, No. 1, p. 2206532, 2023.

[14] Kan Hatakeyama-Sato and Kenichi Oyaizu. Integrating multiple materials science projects in a single neural network. **Communications Materials**, Vol. 1, No. 1, p. 49, Jul 2020.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[16] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, **The Semantic Web**, pp. 593–607, Cham, 2018. Springer International Publishing.

[17] Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. **npj Computational Materials**, Vol. 8, No. 1, p. 102, May 2022.

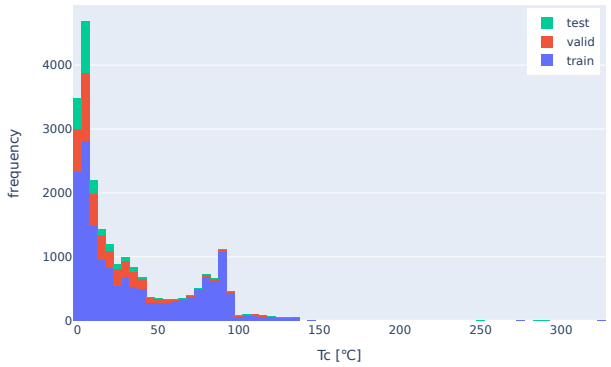

図 3　SuperCon 内の転移温度の分布

A　統計情報

SuperCon に記録されている転移温度のヒストグラムを図 3 に示す．SuperCon から構築した知識グラフについての統計を表 3 と 4 に示す．

表 3　知識グラフの頂点の統計

頂点の種類	
処理前組成式	6,464
処理後組成式	14,780
構造名	404
論文題名	5,479
計	27,127

表 4　知識グラフの辺の統計

辺の種類	
処理前組成式→処理後組成式	15,981
処理後組成式→処理後組成式	14,780
構造名→処理後組成式	11,316
論文題名→処理後組成式	19,364
計	61,441

B　実験環境

実装には，Python 3.10.8 を用いた．転移温度予測モデルを実装するために，PyTorch 1.13.1，DGL 1.1.2+cu116，Transformers 4.35.2，CrabNet 2.0.8，scikit-learn 1.3.2 を用いた．転移温度予測モデルはハイパーパラメータチューニングを行い，最終的に表 5 に示すハイパーパラメータで学習を行った．転移温度予測モデルの学習に用いた計算機の詳細は表 6 に示す．

表 5　ハイパーパラメータチューニングの結果

ハイパーパラメータ	値
RGCN 層の次元	512
バッチサイズ	64
学習率	4e-2
base lr (CyclicLR)	3e-3
max lr (CyclicLR)	9e-3
ドロップアウト	3e-2

表 6　計算機の詳細

項目	値
OS	Ubuntu 20.04.3 LTS
CPU	Intel(R) Xeon(R) W-3225
GPU	NVIDIA RTX A6000