Multi-task Attention for Doped Thermoelectric Properties Prediction

Leng Ze Tang [1] **Trupti Mohanty** [2] **Sterling G. Baird** [2 3] **Leonard W. T. Ng** [1] **Taylor D. Sparks** [2]

[1] *School of Materials Science and Engineering, Nanyang Technological University, 639798, Singapore* [2 2] *Department of Materials Science & Engineering, University of Utah, Salt Lake City, Utah 84108, USA* [3] *Acceleration Consortium, University of Toronto, 80 St George St, Toronto, ON M5S 3H6, Canada*. Correspondence to: Taylor D. Sparks sparks@eng.utah.edu.

1. Introduction

Given the potential of thermoelectric (TE) materials in sustainable energy generation, waste heat recovery and refrigerant-free cooling, improving their transport properties is critical for their widespread adoption [1, 2, 3, 4, 5]. A common technique is impurity doping, which is defined as the intentional introduction of small concentrations of foreign elements into the host material. Given the diversity of TE families, exhaustive experimental synthesis and characterization are impractical, especially when multiple dopants are introduced [6, 7]. Similarly, numerical simulations such as Density Functional Theory (DFT) [8] are expensive when modeling doped systems, which require large supercells to model dilute dopant concentrations.

To resolve this, machine learning (ML) can complement existing simulation and experimental frameworks as an initial screening tool due to its computational efficiency. ML models can be trained on existing experimental data, which typically contain compositional information but largely lack structural or processing conditions [9, 10]. This fundamentally limits ML model performance, particularly when applied to doped materials. Typical composition-based features, such as average atomic mass [11, 12], are based on elemental prevalence statistics. Since impurities are introduced in small amounts, doped and undoped materials have nearly identical features, limiting the model's ability to capture doping effects on transport properties [13, 14].

Several approaches have been attempted to address these limitations. Na et al. introduced DopNet [13], a deep learning model that separately featurizes doping elements and host materials using a predefined atomic fraction threshold. However, to our knowledge, no formal threshold separates dopants from alloys, which means this parameter requires dataset-specific tuning. Antunes et al. [15] developed a multi-output model inspired by CrabNet [14] , trained on DFT data [16] to predict S, σ, and PF. However, this approach has limited experimental applicability as it is trained on DFT datasets that do not specify doping elements and is constrained to predictions at discrete temperatures and doping levels [17].

To improve doped materials prediction, we introduce multi-task CrabNet (MT-CrabNet), a direct modification of CrabNet. First, we incorporate temperature as a continuous input to enable the attention mechanism to focus on doping elements despite their small concentrations, allowing the model to implicitly learn complex dopant-host interactions and temperature-dependent effects on specific elements. Second, we implement multi-task learning (MTL) to simultaneously predict seven transport properties: the total, electronic, and lattice thermal conductivities (κ, κ_e, and κ_l, respectively), electrical conductivity (σ), Seebeck coefficient (S), power factor (PF), and thermoelectric figure of merit (zT). As these properties are interrelated, MTL allows the model to leverage insights from one property to improve predictions of others, leading to overall performance gains.

2. Substantial section

2.1 Methods Overview

MT-CrabNet was compared with its single task variant (ST-CrabNet), DopNet and Random Forest (RF). All models were trained on the Systematically Verified Thermoelectric (sysTEm) dataset, which contains over 8,000 data points and more than 1,400 unique compositions, developed from a previous work [18]. A 5-fold cross validation over five different seeds was performed to evaluate model performance, with each fold performed along mutually exclusive composition splits, which we termed CompositionKFold. This is essential to ensure the test set compositions remain unseen and to evaluate how well the models screen for new compositions. Three different methods to incorporate external features like temperature were explored, two of which were passed to the attention input heads by concatenation (`concat_at_input`) or tiling (`tile_at_input`), while another was concatenated to the residual block (`concat_at_output`) after the transformer layers. This allowed us to compare whether temperature has separate effects on each element, which likely necessitates the attention mechanism, or whether an overall effect on the entire composition is sufficient. For MTL, we used a selector vector approach inspired by previous works [19, 20]. Specifically, we perform one-hot encoding for each property and add the corresponding vector to either each attention head or one of the residual layers. This allows us to handle data points with partially missing transport properties while introducing minimal architectural changes. To ensure a fair comparison, we also performed hyperparameter optimization on the baseline DopNet and RF models.

2.2 *Primary Results*

MT-CrabNet showed consistent top performance across all properties, in particular for the electronic thermal conductivity (κ_e) and the electrical conductivity (σ), depicted in terms of R^2 in Figure 1. For κ_e, baseline models achieved R^2 scores of 0.43–0.44, whereas all MT-CrabNet variants achieved approximately 0.60. Between the two baselines, RF and DopNet showed similar performance for most properties, with the exceptions of κ and σ. For σ, RF exhibited particularly poor performance. It was found that RF severely overestimated many σ values. This highlights the fundamental difficulty decision tree models face when extrapolating to unseen compositions.

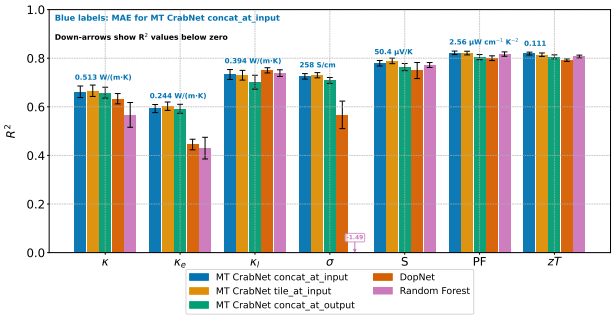

Fig. 1: Average R^2 scores of MT-CrabNet versus baseline DopNet and Random Forest. The MAE scores for the `concat_at_input` variant are annotated above the plots. The error bars represent the standard error.

Comparing MT-CrabNet with ST-CrabNet in Figure 2, it is evident that the performance improvement for κ_e is largely due to MTL, where the R^2 of MT-CrabNet (average 0.60) is roughly 1.5 times greater than that of ST-CrabNet models (average 0.39). We note that there is a general scarcity of κ_e and κ_l data points (numbering around 2,000 values), which is around one-quarter the size of other transport property datasets. This limited training data typically results in lower model performance. However, for κ_e, MTL enables significant performance gains by leveraging its close correlation with σ through the Wiedemann-Franz Law, where L is the Lorenz number:

$$\kappa_e = L\sigma T \tag{1}$$

This relationship enables positive transfer, where insights from the more abundant σ data are used to better inform predictions for κ_e. Additionally, the results indicate that the three different methods used to encode temperature perform comparably. While concatenating temperature information to the output of the attention block shows a slight decrease in mean R^2 for several properties (κ, κ_l, σ, S, and PF), these differences are typically within the standard errors and thus not statistically significant. This suggests that the choice of temperature encoding strategy has minimal practical impact on model performance for this dataset and that learning the overall effect of temperature on composition is sufficient for predicting TE properties.

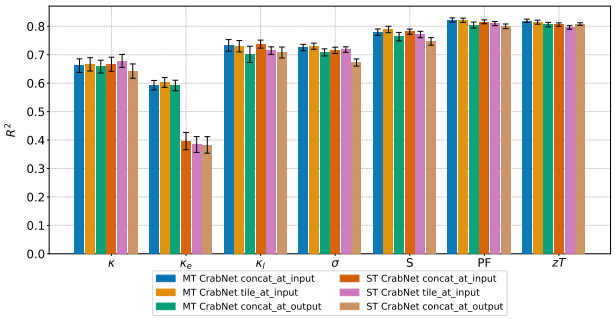

Fig. 2: Performance comparison between MT-CrabNet and ST-CrabNet, its single-task variant. There is significant performance improvement for κ_e when multi-task learning is used.

2.3 *Supporting Results*

Analyzing the weights of the first attention layer (Figure 3) relative to the most common element in the dataset, Te, we observe that despite each head learning a distinct mapping between elements, certain heads show stronger attention weights between Te and doping elements (heads 1 and 2 in this case). Head 2 is more focused on lanthanides, which form doped compounds such as $Ge_{0.98}Sc_{0.02}Te$ and $Lu_{0.2}Bi_{1.8}Te_3$. This supports the finding that CrabNet can learn dopant-host interactions more successfully than typical composition-based methods.

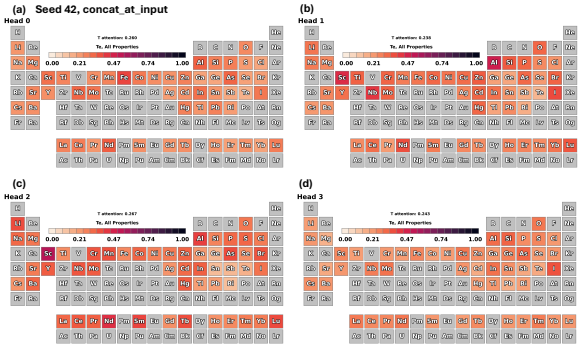

Fig. 3: Normalized attention weights at the first attention layer, using the `concat_at_input` variant of MT-CrabNet. Each figure represents the weights of one of the four attention heads (a-d).

Overall, this work supports the finding that self-attention mechanisms can implicitly learn dopant-host interactions and that MTL of related properties leads to improvements in performance despite using only simple compositional information. Beyond TE, MT-CrabNet can also be used for other dopant property prediction tasks.

Acknowledgments

L.Z.T. is grateful for the support from the CN Yang Scholars Programme and the School of Materials Science and Engineering at Nanyang Technological University Singapore that enabled his visit to the University of Utah, where the bulk of the work took place. He acknowledges the contributions by Andrew Falkowski for modifications made to the CrabNet model between the original work and this work. Additionally, he would like to thank Professor Prashun Gorai for his advice, particularly related to thermoelectric materials theory and simulations. T.M. and T.D.S. acknowledge support from the Army Research Office Materials Design, under Contract No. W911NF-23-1-0333.

Appendix A. Code Availability

The data and code that support the findings are openly available at the following link: `https://github.com/tankylz/multi-task_crabnet`. The general nature of MT-CrabNet allows researchers of different domains to utilize the model for their specific property prediction and screening tasks.

References

[1] G. D. Mahan. Introduction to thermoelectrics. *APL Materials*, 4(10):104806, June 2016.

[2] Daniel Sanin-Villa. Recent Developments in Thermoelectric Generation: A Review. *Sustainability*, 14(24):16821, January 2022.

[3] Rakesh Singh, Surya Dogra, Saurav Dixit, Nikolai Ivanovich Vatin, Rajesh Bhardwaj, Ashok K. Sundramoorthy, H. C. S. Perera, Shashikant P. Patole, Rajneesh Kumar Mishra, and Sandeep Arya. Advancements in thermoelectric materials for efficient waste heat recovery and renewable energy generation. *Hybrid Advances*, 5:100176, April 2024.

[4] Matteo d'Angelo, Carmen Galassi, and Nora Lecis. Thermoelectric Materials and Applications: A Review. *Energies*, 16(17):6409, January 2023.

[5] H. Julian Goldsmid. Theory of thermoelectric refrigeration and generation. In *Introduction to Thermoelectricity*, pages 9–24. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

[6] Dmitry Chernyavsky, Jeroen van den Brink, Gyu-Hyeon Park, Kornelius Nielsch, and Andy Thomas. Sustainable Thermoelectric Materials Predicted by Machine Learning. *Advanced Theory and Simulations*, 5(11):2200351, 2022.

[7] Nikhil K Barua, Sangjoon Lee, Anton O Oliynyk, and Holger Kleinke. Recent strides in artificial intelligence for predicting thermoelectric properties and materials discovery. *Journal of Physics: Energy*, 7(2):021001, March 2025.

[8] Lenz Fiedler, Normand A. Modine, Steve Schmerler, Dayton J. Vogel, Gabriel A. Popoola, Aidan P. Thompson, Sivasankaran Rajamanickam, and Attila Cangi. Predicting electronic structures at any length scale with machine learning. *npj Computational Materials*, 9(1):115, June 2023.

[9] Gyoung S. Na and Hyunju Chang. A public database of thermoelectric materials and system-identified material representation for data-driven discovery. *npj Computational Materials*, 8(1):1–11, October 2022.

[10] Michael W. Gaultois, Taylor D. Sparks, Christopher K. H. Borg, Ram Seshadri, William D. Bonificio, and David R. Clarke. Data-Driven Review of Thermoelectric Materials: Performance and Resource Considerations. *Chemistry of Materials*, 25(15):2911–2920, August 2013.

[11] Nikhil K. Barua, Evan Hall, Yifei Cheng, Anton O. Oliynyk, and Holger Kleinke. Interpretable Machine Learning Model on Thermal Conductivity Using Publicly Available Datasets and Our Internal Lab Dataset. *Chemistry of Materials*, 36(14):7089–7100, July 2024.

[12] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, August 2016.

[13] Gyoung S. Na, Seunghun Jang, and Hyunju Chang. Predicting thermoelectric properties from chemical formula with explicitly identifying dopant effects. *npj Computational Materials*, 7(1):1–11, July 2021.

[14] 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*, 7(1):1–10, May 2021.

[15] Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Predicting thermoelectric transport properties from composition with attention-based deep learning. *Machine Learning: Science and Technology*, 4(1):015037, April 2023.

[16] Francesco Ricci, Wei Chen, Umut Aydemir, G. Jeffrey Snyder, Gian-Marco Rignanese, Anubhav Jain, and Geoffroy Hautier. An ab initio electronic transport database for inorganic materials. *Scientific Data*, 4:170085, July 2017.

[17] Qihong Xiong, Guang Han, Guoyu Wang, Xu Lu, and Xiaoyuan Zhou. The Doping Strategies for

Modulation of Transport Properties in Thermoelectric Materials. *Advanced Functional Materials*, 34(52):2411304, 2024.

[18] Leng Ze Tang, Layla Purdy, Trupti Mohanty, Leonard W. T. Ng, and Taylor D. Sparks. Systematically Verified Experimental Thermoelectric Dataset For Data-driven Approaches, August 2025.

[19] Christopher Kuenneth, Arunkumar Chitteth Rajan, Huan Tran, Lihua Chen, Chiho Kim, and Rampi Ramprasad. Polymer informatics with multi-task learning. *Patterns (New York, N.Y.)*, 2(4):100238, April 2021.

[20] Christopher Kuenneth, William Schertzer, and Rampi Ramprasad. Copolymer Informatics with Multitask Deep Neural Networks. *Macromolecules*, 54(13):5957–5961, July 2021.