Author Response

Gyoung S. Na and Hyunju Chang
Korea Research Institute of Chemical Technology
Republic of Korea
ngs0@krict.re.kr, hjchang@krict.re.kr

July 7, 2022

Thank you for the valuable comments of the reviewers. Based on the review comments, we entirely revised the manuscript to clarify the contributions of the work, physical and chemical insights from the experiment results, and the readability of the manuscript. The point-by-point author response for the review comments is follows.

1 Reviewer #1

1.1 Contributions of the Paper

1.1.1 Journal Scope and Contributions

We would like to argue that npj Computational Materials is a method-oriented journal because it focuses on development of theoretical/simulation approaches and data-driven high-throughput techniques. ESTM data and the proposed method correspond to these objectives of npj Computational Materials.

1.1.2 Physical and Chemical Insights

The contributions of this paper can be summarized in physics and chemistry viewpoints as:

- **New experimental dataset:** We constructed a new public thermoelectric dataset containing various host materials and experimentally measured theremoelectric properties. Furthermore, we publicly open this dataset for further data-driven research in materials science. To the best of our knowledge, it is the first public dataset that was collected and validated for data-driven research of thermoelectric materials.
- **Distributional regression:** We adopted a promising approach called distributional regression to approximate input-to-target relationships based on noisy and heterogeneous data [1, 2]. Based on a physical viewpoint, we assumed that each material system forms a distribution with doped and alloyed materials from the pristine materials. We empirically demonstrated that we can approximate the entire material space more accurately by dividing the material space based on each material system. This approach is in line with a well-known "divide-and-conquer" [3] strategy.
- **False positive in high-throughput screening:** We conducted the high-throughput screening to discover novel and high-efficiency thermoelectric materials to provide an example application of the proposed method in materials science. The proposed method significantly reduced false positives in the high-throughput screening, and the low false positive is an important metric for efficient and reliable high-throughput screening in data-driven materials discovery.

In addition to these contributions, we conducted additional experiments and analyses to provide more general insights from the viewpoint of materials science as follows.

- **Hyper-parameter analysis for the number of training data:** Constructing training data is the main problem in machine learning for materials science. We conducted a hyper-parameter analysis for the different number of training data in Section 3.1, and we observed that the proposed method can improve the extrapolation capabilities of the prediction models by increasing the hyper-parameter K. This result will be helpful in the implementation of SIMD-based machine learning for real-world applications of materials science. The analysis results were added to Section 3.1 of the revised manuscript.
- **Case study for false positive samples:** The baseline model XGB_d made a lot of false positives in the high-throughput screening task in Section 2.4.2 because it incorrectly predicted the materials from the $\mathrm{Mg}_{1-x}\mathrm{Li}_x\mathrm{Ge}_{0.9}\mathrm{Si}_{0.1}$ system as the high-ZT materials. By contrast, SXGB_d based on SIMD greatly reduced the false positives from the $\mathrm{Mg}_{1-x}\mathrm{Li}_x\mathrm{Ge}_{0.9}\mathrm{Si}_{0.1}$ system. In Section 3.2 of the revised manuscript, we discussed the reason for the false positives of the high-throughput screening in the chemical space.

1.2 Related Work

As shown in paragraphs 3-5 of the Introduction section, we have already discussed state-of-the-art methods to predict the electronical and thermoelectric properties of the materials from their chemical compositions. Also, we pointed out the limitations of the existing machine learning methods. To supplement related work, we added the recommended references and discussed their main contributions.

1.3 Contributions in a Physical Chemistry Viewpoint

We supplemented the Discussion section (Section 3.1 and 3.2) and the Introduction section for chemical insights and related work, respectively.

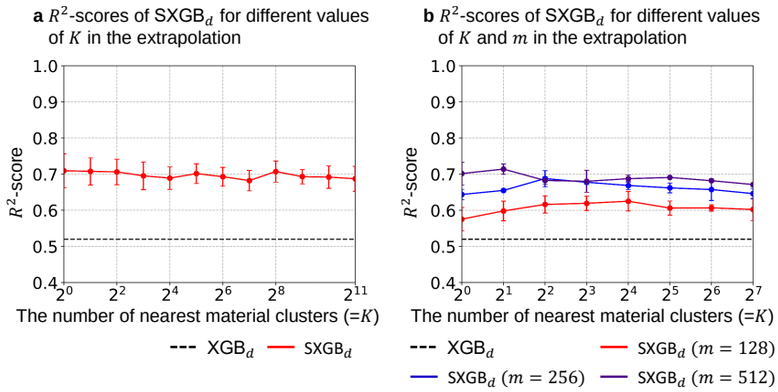

Figure 1: Measured R^2-scores of SXGB$_d$ in the extrapolation problem to predict ZTs for different values of the hyper-parameter and the data size. K is the number of nearest material clusters in Eq. (4) and (5). m is the number of material clusters in the training dataset. For the quantitative comparisons, we present the R^2-score of the baseline XGB$_d$ using the black dotted line.

2 Reviewer #2

2.1 Readability of the Paper

We revised the manuscript entirely to improve the readability. We removed redundant sentences and added Section numbers for each experiment result. In addition, we reorganized the sections of the manuscript to clarify the main contents of each section (Section 2.3.4 → Section 2.4, Section 2.5.1 → Section 2.4.1, Section 2.5.2 → Section 2.4.2, Section 2.5.3 → Section 2.4.3). The revised contents are marked in red color.

2.2 Readability of Figures and Tables

Also, the figures and tables were revised entirely to improve the readability of the manuscript. The revised figures and tables (Table 2, Figure 2, and Figure 7) are marked in red color.

2.3 Numerical Representation of Chemical Compositions

We added the data representation and the encoding process of the chemical compositions in Supplementary Information to clarify the data representation of the string variables in machine learning.

2.4 Extrapolation Results

We used FCNN and XGB to precisely validate the effectiveness of the proposed descriptor. Although several methods have been proposed to predict the materials properties from the chemical compositions [4], they are not designed for the doped materials or require specific data representation of the materials. For these reasons, we used the basic FCNN and the state-of-the-art XGB, which are generally applicable to the applications of materials science.

We also agree that the F1 score is not very high. However, in machine learning, the extrapolation is considered the most challenging problem because machine learning methods are not able to predict the data beyond the training distribution. A large number of methods try to achieve even a small improvement in the extrapolation problems [5, 6, 7, 8]. Therefore, we would like to argue that the improvement by 0.12 (0.49 → 0.61) in the F1-score is a remarkable achievement of the proposed method.

2.5 Hyper-Parameter Analysis of K

We conducted an additional experiment to measure R^2-scores of SXGB$_d$ for different values of K and the data size. Based on the results of the additional experiment, we entirely revised the hyper-parameter analysis section (Section 3.1) and Fig. 7. The modified figure is in Fig. 1 of this response letter. Please review the new results of the hyper-parameter analysis of K in the revised manuscript.

2.6 Code and Dataset Availability

Unfortunately, we cannot open ESTM dataset before the publishing process due to the policy of our research funding. However, we uploaded the source codes and the trained model without ESTM dataset at a public repository `https://github.com/ngs00/simd`. We hope the uploaded files are helpful in evaluating our work. Note that all source codes and datasets will be publicly released when the submission moves to the next process.

3 Reviewer #3

3.1 Prediction with Graph Neural Networks

As mentioned in the review comment, graph neural networks (GNNs) have shown state-of-the-art prediction accuracy in various applications of materials science. However, one rigid requirement to employ GNNs for the structure-based prediction is that a database containing the crystal structures of the materials must be prepared. In most cases, it is not feasible to calculate the crystal structures of thermoelectric materials due to the computational cost and structural complexity from the crystal structures containing dopants and alloys. In other words, GNNs including CGCNN, MEGNet, and ALIGNN are not applicable to the datasets containing the doped materials, such as our ESTM dataset. Although a graph-based approach for the chemical formulas was recently proposed [4], it is not applicable to the doped materials. The limitations of existing deep learning methods are described in the introduction section.

3.2 Comparision with Other Databases

There are several fragmented datasets of the thermoelectric materials. However, existing datasets contain a few data biased to specific host materials or do not provide the API service for bulk download [9, 10, 11]. Moreover, most datasets are not reliable because the source or experiment methods to generate the data are not described. Although Starry dataset provides a large number of thermoelectric materials (as we mentioned in Section 2.3.4), it is not suitable for machine learning due to the following two reasons.

- The experimentally collected and theoretically calculated thermoelectric materials and their properties are mixed without source labels in Starry dataset, which makes the prediction models unreliable.
- The invalid values and the parsing errors in the collected thermoelectric properties are inevitable in Starry dataset because the data was automatically collected by a parsing algorithm.

To overcome the limitations of the existing thermoelectric materials datasets, we constructed ESTM dataset by collecting the thermoelectric materials and their experimentally measured properties from published articles. Furthermore, we validated the collected data with domain experts and data validation algorithms to perform machine learning without labor-intensive data pre-processing. To the best of our knowledge, ESTM dataset is the first public large dataset for data-driven discovery and machine learning on the thermoelectric materials.

References

[1] Kneib, T., Silbersdorff, A. & Säfken, B. Rage against the mean–a review of distributional regression approaches. *Econometrics and Statistics* (2021).

[2] Imani, E. & White, M. Improving regression performance with distributional losses. In *International Conference on Machine Learning*, 2157–2166 (PMLR, 2018).

[3] Gu, M. & Eisenstat, S. C. A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem. *SIAM Journal on Matrix Analysis and Applications* **16**, 172–191 (1995).

[4] Goodall, R. E. & Lee, A. A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nat. Commun.* **11**, 1–9 (2020).

[5] Fort, S., Ren, J. & Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems* **34**, 7068–7081 (2021).

[6] Krueger, D. *et al.* Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, 5815–5826 (PMLR, 2021).

[7] Na, G. S., Jang, S. & Chang, H. Predicting thermoelectric properties from chemical formula with explicitly identifying dopant effects. *Npj Comput. Mater.* **7** (2021).

[8] Na, G. S., Jang, S. & Chang, H. Nonlinearity encoding to improve extrapolation capabilities for unobserved physical states. *Physical Chemistry Chemical Physics* **24**, 1300–1304 (2022).

[9] Mrl database. `http://www.mrl.ucsb.edu:8080/datamine/thermoelectric.jsp`.

[10] Thermoel database. `http://info.eecs.northwestern.edu/ThermoEl`.

[11] Ucsb thermoelectrics database. `https://citrination.com/data_views/322/summary`.