Review‘s comments on “A Public Database of Thermoelectric Materials and System-Identified Material Representation for Data-Driven Discovery of High-Performance Thermoelectric Materials”:

After reading your rebuttal and revised paper, I appreciate that you have read a lot of my previous comments carefully and made targeted changes, but I think I still have several problems with the paper before it reaches publication:

1. As Reviewer #1 asked, perhaps in Section1. Introduction can be listed directly in the form of a list of motivations and contributions of the paper, which will help the reader to quickly grasp the point of the article, and in the process of reading the subsequent article easier to understand the point you want to make.
2. As you mentioned in your rebuttal, the F1 score of 0.61 is indeed not very satisfactory in terms of model accuracy, and does not deny the contribution of your paper to the ODD problem, but the SIMD method is very dependent on the previous KNN clustering model results, and there is no metric to evaluate this clustering algorithm in your paper, perhaps the error will accumulate to the final SIMD And can you try SOTA's clustering algorithm to do a comparison with the traditional KNN algorithm to find a better model.
3. About the last part of migration learning, the paper mentions that Starry dataset is not suitable for machine learning for two reasons: " (1) The thermoelectric data collected from actual experiments and theoretical calculations is not distinguished in Starry dataset, which makes the prediction models unreliable. (2) The parsing errors in the collected thermoelectric properties are inevitable in Starry dataset because the data was automatically collected by a parsing algorithm. " Perhaps a quantitative description of Starry can be made as well as evaluating the ESTM dataset to show "it is not suitable for machine learning". If the Starry dataset is not suitable for machine learning, why he is selected as the source domain for transfer learning, the quality of the source domain data is very important for the results of the target domain.
4. The apparent flaws of the source domain dataset are clearly pointed out in the paper. Is the direct migration without addressing the source domain dataset the reason for the poor F1 score of the SIMD method?