论文主要贡献：

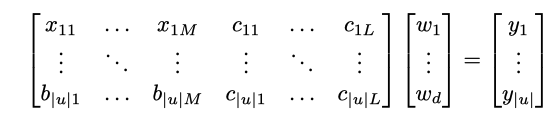
1. 搜索文献，建立热电材料(包含热电特性)实验公开数据库ESTM，但是文章中并没有列出数据获取网站。

2. 从数据库的化学成分信息与材料特性之间建议回归模型R2>0.9（3折的结果）

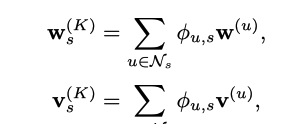
3. 基于热电材料化学成分设计新的描述表示SIMD方法，在新的数据上利用迁移学习(迁移特征表示)R2从0.13提高到0.71

SIMD方法总结：

1. 首先构建一个很大的数据集，包括源数据集（Starry，作者说这个数据集也是根据文献搜索建议，中间有基于理论计算的，也有实验测试的，数量大，质量不如ESTM，所以作为源数据集）和目标域上的训练集，**注意测试集中的数据材料组都是训练集中未出现过的，所以是Out Data Detector(ODD)的问题。**
2. 根据下面的公式，使用最小二乘法计算W，x是成分特征，c是实验环境特征，y就是目标值：

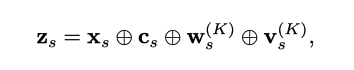


1. 使用KNN对材料组进行聚类，再根据下面的公式得到同一个材料组的统一system vector表示Ws和Vs：





1. 我理解是一个材料组有一个统一的Ws和Vs的向量表示，然后通过下面的公式进行向量拼接，得到SIMD方法的特征输入，**注意测试集不参与Ws和Vs的计算，但是通过KNN算法对其进行聚类，同一类的使用的Ws和Vs是同一个，所以测试数据也能得到最后的SIMD特征表示**：



1. 最后使用XGBoost模型来学习，论文中使用的模型普遍很简单，和XGB对比的就是FNN（前馈神经网络，SVR，Ridge因为效果差没有进行SIMD方法的对比）

**我的理解相当于是通过聚类的方法，得到了未知数据的特征表示，所以迁移的是经过学习的特征向量，并且这个特征向量是根据源域和训练集中Y值得到的，相当于迁移特征的时候吧源域和训练集中的Y值也利用起来，思想是比较新且有意思的。但同时问题也比较多，首先就是文章的结构比较混乱，一开始看前面的段落很容易不知道文章在讲什么，感觉前后矛盾，看到很后面才有具体的数据解释，阅读体验不是很好；对比的算法是否太少了一些，应该尝试更先进的模型；图表的表示也有改进的点，比如R2和MSE的表示非常不直观，观感不好；最后Discussion中讨论了K值的鲁棒性，我觉得将近3W条数据，你K的取值从1-10是否太少了一些，因为没看到具体的数据，所以我也不能具体说你这3W数据中聚类达到多少合适，我觉得应该跨度再大点可能能看到K值对最后模型的影响。**

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

The entry point of this paper is very novel, as the paper first searches the literature for the properties of Thermoelectric materials to build an ESTM dataset, which can greatly facilitate efficient data-driven discovery of novel thermoelectric materials.

The SIMD algorithm proposed in the paper can migrate features in the source domain and training set data to improve the accuracy of the ODD (Out Data Detection) problem. The algorithm idea is very novel, and it also changes the traditional method of using only X features in migrating features, and innovatively introduces the features System Vector (W) generated by X and Y at the same time and uses only the Y and only the statistical features (V) of Y. And based on SIMD algorithm, the R2 improvement of 0.57 is obtained in ODD problem, which is a very big accuracy improvement. Meanwhile, SIMD was successfully applied to discover high-ZT thermoelectric materials based on highthroughput screening. However, the paper still has some issues that need to be addressed：

1. The structure of the paper needs to be improved. For example, in the first chapter Introduction, R2's 0.52 to 0.71 and 57.14% improvement is not explained very clearly until the specific explanation in chapter 2.3, which can easily cause misunderstanding to the reader.
2. The presentation of the charts also needs to be improved, for example, the presentation of MSE and R2 in Table 2 could be clearer.
3. The component features of the materials in the ESTM dataset are strings in the presentation of Table 1, which cannot be directly used for machine learning. It is not stated until Section 2.3.3 that they are digitally encoded and entered into the model, should it be stated earlier, or should the component features of the materials be encoded when the dataset is presented.
4. The effect of the value of K on the model is discussed in the Discussion, which illustrates the robustness of the model to K, whether it is really necessary and the effect of the comparison, whether the value of K should be proportional to the number of material groups in the dataset, and whether the range of K can be increased to see how the model performs.
5. In the paper the address of the data and code acquisition is empty, can it be filled in completely.（因为论文中数据和代码的地址都为——，是否是因为论文未公开发表所以不能公开的原因，如果是就删除这条意见。）

Overall, this paper requires more work to be in a publishable state. The final recommendation for this paper is to \_\_\_\_\_\_\_\_.

1. Accept
2. Accept after minor revision
3. Probably acceptable after major revision with re-review
4. Unacceptable as is, but worth reconsideration if extensively revised
5. Reject