Discovering Thermoelectric Materials Using Machine Learning: Insights and Challenges

Mandar V. Tabib[1(✉)], Ole Martin Løvvik[2(✉)], Kjetil Johannessen[1], Adil Rasheed[1], Espen Sagvolden[2], and Anne Marthine Rustad[1]

[1] SINTEF Digital, Mathematics and Cybernetics, Trondheim, Norway
Mandar.Tabib@sintef.no
[2] SINTEF Industry, Sustainable Energy Technology, Oslo, Norway
OleMartin.Lovvik@sintef.no

Abstract. This work involves the use of combined forces of data-driven machine learning models and high fidelity density functional theory for the identification of new potential thermoelectric materials. The traditional method of thermoelectric material discovery from an almost limitless search space of chemical compounds involves expensive and time consuming experiments. In the current work, the density functional theory (DFT) simulations are used to compute the descriptors (features) and thermoelectric characteristics (labels) of a set of compounds. The DFT simulations are computationally very expensive and hence the database is not very exhaustive. With an anticipation that the important features can be learned by machine learning (ML) from the limited database and the knowledge could be used to predict the behavior of any new compound, the current work adds knowledge related to (a) understanding the impact of selection of influence of training/test data, (b) influence of complexity of ML algorithms, and (c) computational efficiency of combined DFT-ML methodology.

Keywords: Machine learning · Density functional theory
Thermoelectric · Material screening · Discovery

1 Introduction

Thermoelectric (TE) materials are receiving wide attention due to their potential role in mitigating global greenhouse effects as they enable conversion of waste heat energy directly to electrical energy. Currently, the three approaches to find better thermoelectric material involve: (a) traditional experimental approach, (b) physics based computational approach like Density Functional Theory (DFT), and (c) recent machine learning (ML) based data-driven approach. Amongst these, the machine learning approach has shown some success in finding new chemistries (that are capable of being thermoelectric) but it is a nascent application area with limited published work. There are certain limitations with

V. Kůrková et al. (Eds.): ICANN 2018, LNCS 11139, pp. 392–401, 2018.
https://doi.org/10.1007/978-3-030-01418-6_39

all the approaches, like: (a) The traditional experimental approaches are not efficient way of exploring new unknown chemistries and they focus mostly on modifying known material compounds by doping and nano-structuring to make these known thermoelectric materials better, while, (b) high fidelity physics based models like DFT are computationally prohibitive to use, and (c) for ML, obtaining bountiful data is an expensive process. ML models need to be able to generalize well, and learn patterns well enough from a small pool of available training data to be able to search for new potential materials in the vast expense of search-space of unknown materials. The current work aims to contribute to the field of machine learning and material screening by understanding influence of limited dataset, and whether it can be mitigated by studying: (a) influence of training-test split in model development, (b) influence of model selection and (c) by applying a framework combining data-driven machine learning models with physics-based density functional theory (DFT) to identify potential thermoelectric materials using a metric called 'figure of merit'. DFT enables generation of training data for ML, and a trained ML is expected to save time in finding potential material in the vast material search-space. The main objectives of this work can be enumerated as:

1. In the limited dataset scenario: understand the influence of training/test compound selection on ML predictions.
2. Combine data-driven models with physics-driven models to mitigate limited dataset scenarios, and understanding efficiency of this approach in identifying potential thermoelectric materials.
3. Compare the performance of the two ML algorithms: Random Forest (RF) and Deep Neural Network (DNN) for the limited dataset scenario.

2 Methodology and Data

This is treated as a regression problem, where the ML model learns to predict the *figure of merit* (ZT) values of a given compound at a given temperature and at a given chemical potential state. The performance of a material as a thermoelectric material is evaluated using this ZT. A material with a high ZT is supposed to be a good thermoelectric material. The ZT is a function of Seebeck coefficient, temperature, electrical conductivity, the electronic thermal conductivity, and lattice thermal conductivity. Previous research on thermoelectric materials involving machine learning did not use ZT as a characteristics, instead, it used the key properties in a stand-alone way (i.e. band gap, Seebeck coefficient, etc.). The three key components needed for developing the methodology are described next: (a) Data: data for model development (cross-validation/training data), for model testing (hidden test data) and for model application (search-space data to look for potential materials), (b) Descriptors (features), and (c) Choice of ML algorithms. These three components are discussed next:

2.1 Descriptors

Descriptors (known as features in ML community) are the characteristics of materials (e.g., crystal structure, chemical formula, etc.) that might correlate with material's properties of interest (ZT). Here, we use 50 features (descriptors or independent variables) for a given data-point. The features involve both numerical variables and categorical variables (crystal shape). The list of 50 features used are: temperature, chemical potential - eV, elements in cell, mean and variance of atomic mass, atomic radius, electronegativity, valence electrons, a set of features related to periodic table (group numbers, row numbers,electronic configurations), 6 one-hot encoded features for crystal shape ('tetragonal', 'trigonal', 'orthorhombic', 'cubic', 'monoclinic', 'triclinic', 'hexagonal').

2.2 Data

Limited Data Scenario: The dataset is deemed limited in this work because based on the available training dataset of just 115 compounds (having about 87,975 instances/data points with known ZT values), the trained ML model has to learn to predict potential compounds (i.e. ZT values) in a vast chemical search-space of 4800 compound (having 2,40,312 data-points). The compounds in training dataset will be different than the compounds in the chemical search-space.

Data Generation and DFT: It is time-consuming to generate dataset using experiments. Here, the database is generated using high-fidelity physics-driven DFT followed by semi-classical Boltzmann theory. The DFT is a computational quantum mechanical modeling method used to investigate the electronic structure (principally the ground state) of many-body systems, in particular atoms, molecules, and the condensed phases. Using this theory, the properties of a system can be determined by using functionals, i.e. functions of the spatially dependent electron density. Boltzmann theory helps to estimate the Boltzmann transport properties of candidate materials (like, Seebeck Coefficient, thermal conductivity, electrical conductivity) based on DFT-predicted band structures. The ZT for each compound is then computed using these transport properties. The ZT values of about 115 materials (compounds) have been generated. A database of about 87,975 instances (datapoints) comprising of 115 compounds materials has been created, as each compound material is studied over 15 temperature levels and over 51 chemical potential states. Thus, the number of datapoints are $115 \times 51 \times 15 = 87,975$. Each instance (or data-point) has 50 features associated with it. Thus, the input data matrix for building ML model is $87,975 \times 50$ - which is to be divided into training data (training and validation sets) and test data set.

Uniqueness in Splitting the Training and Test Dataset: We do not randomly split the 87,975 datapoints into training and test dataset. The dataset is

split so that ML model is trained on certain compounds and the model is tested on unseen compounds. About 85% of data-set (about 98 compounds - a dataset of $74,970 \times 50$) is used for model building through both training and validation sets, and 15% of dataset (about 17 compounds - a dataset of $13,005 \times 50$) is to test the model. Since, the purpose is to test the generalization ability of the ML model to discover new chemical species - so, we looked at whether the ML model trained on 98 compounds can help to predict the ZT values of the unseen 17 compounds. Hence, sensitivity of selection of compounds into training and test data needs to be checked. This is checked by creating 3 cases of train/test split data:

1. Case 1. Test/train split. Randomly selecting 17 compounds in test (corresponding to 13,005 datapoints) and 98 compounds in train (corresponding to 74,970 datapoints) (with random seed 0.2).
2. Case 2. Test/train split. Randomly selecting 17 compounds in test and 98 compounds in train (with random seed 0.4). A different random selection gives different sets of compounds in train/test than case 1.
3. Case 3. Deterministically selecting Test and train compound. Out of the 115 compound database, a chunk of 17 compounds lying in the middle have been selected as test data. These 17 compounds in the middle do not possess extreme characteristics (like either being too simple compound or too complex compound, which are represented in the values of features associated with the compound), while the training data encompasses all types of compound. Here, by *complex compounds*, we refer to compounds with more than 3 elements.

Search-Space Data: For screening and discovering potential thermoelectric materials, the trained machine learning model has been applied on database of silicides (silica based compounds). This database is extracted from the material science project, and is called chemical search-space data set in this work. The search-space data-matrix size is: 2,40,312 data instance $\times$ 50 features.

2.3 Choice of Algorithms

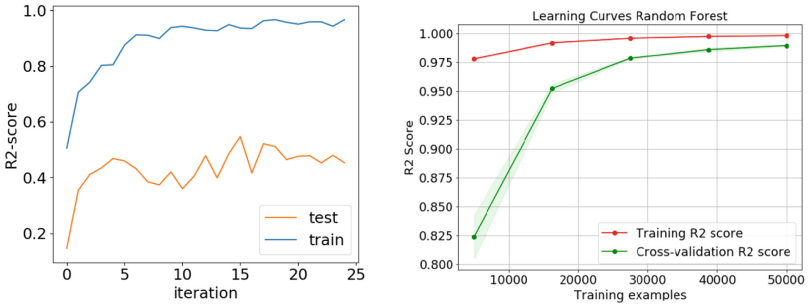
Here, two different algorithms have been tested: Random Forest [1] and a more complex Deep Neural Network [2]. This work is intended to understand whether with the limited dataset, a complex model can perform well or not.

2.4 Model Selection - Cross Validation and Learning Curve

The two machine learning models have been compared using the cross-validation (CV) method. CV is a model validation technique for assessing the generalization ability of a machine learning algorithm to an independent data set. In our work, we split the original dataset into the 'training' and the 'test' dataset. Here, we have selected a *3-fold CV* procedure, where the 'training set' is split further into

3 different smaller sets. The model prediction is learned using 2 of these 3 folds at a time, and the 3rd fold that is left out is used for validation (called validation set). The average R2 (coefficient of determination) score from 3-fold CV is used as performance measure accuracy. Best possible R2 score is 1.0 suggesting a model with high accuracy and the score can be negative if the model performs badly. The learning curve helps to obtain the best parameter sets for the two models using the above CV process. In Fig. 1, we use CV procedure to obtain a learning curve. The curve shows the variation of average R2 score with training data and validation data (for RF) and variation of average R2 score with increasing epochs (iteration) for DNN. These curves help in understanding the bias-variance tradeoff. The learning curve (in Fig. 1) is shown for only case (case 3), and for only the best parameter sets of case 3 (for brevity). For case 3, the best parameter sets are: *RF*: Maximum number of trees - 30. The maximum depth of the tree is 20. *DNN*: The network used in this work comprises of an input layer (with 50 neurons representing the 50 input feature), an output layer and six hidden layers (comprising of following number of units in each successive layer: 43; 20; 20; 15; 10; 5 respectively). A combination of ReLU and Tanh activation functions are used in this work.

The learning curve (in Fig. 1) suggests some over-fitting for both the models; which is more dominant in the case of DNN compared to the RF model. This could be attributed to the need for larger data needed by DNN models. The R2 score on training data for both RF and DNN are in the range of 0.95–1, while, for the validation data (called test in DNN figure here), the R2 scores fall drastically in case of DNN to R2 = 0.45, while, the R2 scores falls slightly to 0.985 for RF. The overfitting (variance errors) is seen in other cases too (case 1 and case 2, but these learning curves are not shown here for brevity). The influence of 3 different train-test split on the performance of two ML models is considered next. It needs to be seen whether proper selection of training compound-test compound split can mitigate the overfitting and improve generalization ability of ML models.



(a) CASE 3 . DNN . (b) CASE 3 . RF .

Fig. 1. Judging bias (underfitting) vs variance (overfitting) errors for RF and complex DNN models for the two cases

3 Results and Discussion

Material screening is challenging in the sense that using the available limited database of known chemistry, the trained ML model should have learned the ability to find new potential material characteristics in new unseen chemistry in the vast material search-space. It is important to understand whether the way to split the limited material database into training dataset (training and validation dataset) and testing dataset (of unseen compounds) will influence the performance of the two machine learning models (simple RF or complex DNN).

3.1 Sensitivity Study: Influence of Training and Testing Dataset Selection

Figure 2 shows the influence of splitting the training/test data on the performance of models for the three cases. For each case, the Fig. 2 shows the predicted ZT values vis-a-vis the actual ZT values for the compounds in training and test data by the two models (RF and DNN). Results for the 3 cases show:

Case 1 and Case 2 (Comparing R2 Scores on Train and Test Data by the Two Models): Both cases have randomly generated but different sets of 98 compounds for training and 18 compounds in test.

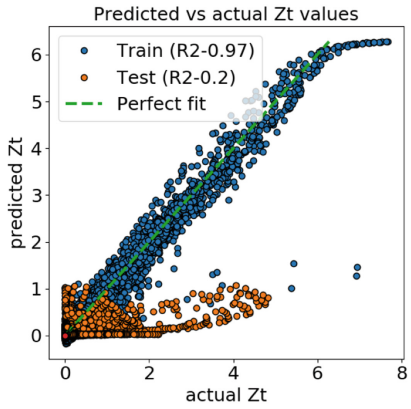
DNN Performance: R2 score for case 1 drops to 0.2; while, the corresponding case 1 train R2 score is 0.97. Similarly, case 2 test R2 score drops to −0.14; while, the corresponding case 2 train R2 score is 0.97. The large drop in R2 scores for test indicates poorer generalization ability for DNN.

RF Performance: In case of RF too, R2 scores drop for the two test dataset, but its performance is much better than the DNN. For RF, the Case 1 test R2 score is 0.82; while the corresponding case 1 train R2 score is 0.99. Similarly, Case 2 test R2 score drops to 0.23; while the corresponding case 2 train R2 score of 0.99.

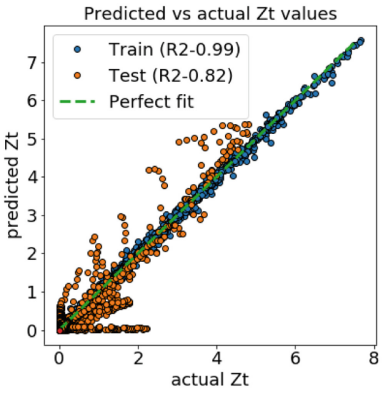
Thus, for both RF and DNN, as the split of train/test varies, the generalization ability is influenced (despite selecting the best parameter set of the respective model for that database during CV). The reason for lower R2 scores in case 2 test dataset (for both the models) as compared to their case 1 test scores is that the 98 randomly selected compounds in case 2 training dataset with their features (a dataset of $74{,}970 \times 50$) do not provide similar pattern characteristics (i.e. variation of ZT with features) as in the 17 compound case2-test dataset (a dataset of $13{,}004 \times 50$).

Case 3 (Comparing R2 Scores on Train and Test Data): Case 3 involves 98 training compounds that encompasses both simple and extreme compounds, and hence the models trained on it are able to capture the pattern to enable determination of ZT values of data-points pertaining to the 17 unseen test compounds. That is why we see improved predictions by the DNN and RF model on the case 3-test dataset: *DNN* shows a case 3-test R2 score of 0.45; while corresponding case 3 train R2 score is 0.96.

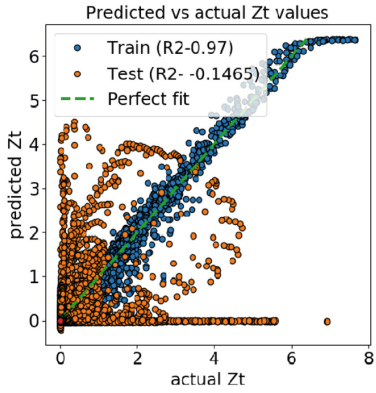
RF shows a case 3 test R2 score of 0.76; while corresponding case 3 train R2 score of 0.99.



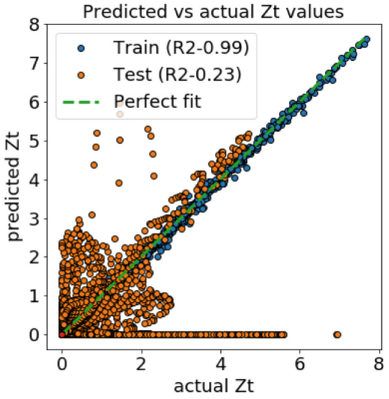
(a) CASE 1. DNN.



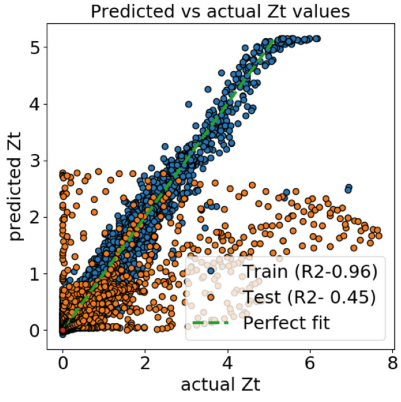
(b) CASE 1. RF.



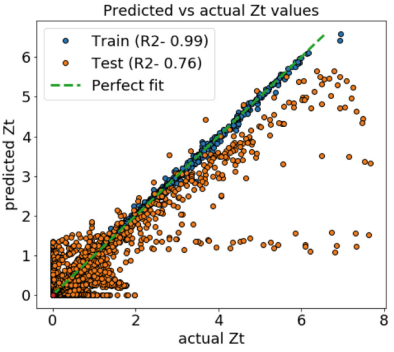
(c) CASE 2. DNN.



(d) CASE 2. RF.



(e) CASE 3. DNN.



(f) CASE 3. RF.

Fig. 2. Predicted vs actual ZT (with R2-score) for DNN and RF on training and unseen test data for the three cases.

Next, we check whether the improvements in generalization ability (better test R2 scores) brought about by balanced training-test split leads to better predictions of material in both models?.

3.2 Comparison of RF vs. DNN Models: Material Screening and Efficiency

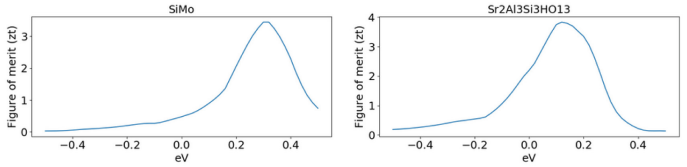
Searching for Potential Thermoelectric in New Search-Space: Figure 3 shows the best two thermoelectric materials identified in a new chemical search-space of silicide materials of 4800 compounds for the 3 cases. For brevity, only top two are shown in Fig. 3 but the results explained are beyond the best two predicted. This chemical search-space has not been exposed to the ML models during their training/validation/testing phase. In all the figures, the predicted *figure of merit* (ZT) is plotted against one of the most influential features (chemical potential - eV). These six compounds below have the highest predicted ZT values as obtained by DNN and RF.

The RF is mostly predicting comparatively simpler compounds than the DNN with maximum value of ZT in the range of 3–3.6. RF has predicted only simple compounds (such as Li2MgSi, SrMgSi, BeSilr2, SiP2O7, VSiPt) as potential thermoelectric silicides in its top two predictions. While, DNN is predicting complex compounds (with more than 3 elements) in about 66% of the top two predictions (with compounds such as Sr2AI3Si3HO13 in case 1, LiCoSiO4 in case 2, and Na3CaAI3Si3SO16 and Na3VSiBO7 in case 3) with higher maximum value of ZT in range 4–5. Both DNN and RF have identified a common thermoelectric silicide (BeSilr2) as potential candidate but predict a different maximum ZT value (RF predicts ZT of 3.5, while DNN predicts around $ZT = 4.5$).

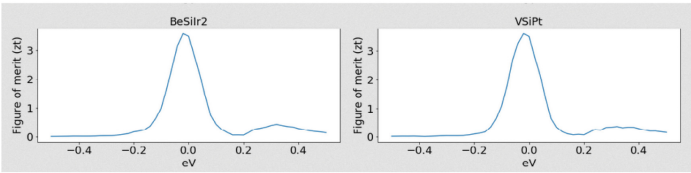
DNN is learning complex patterns than RF and predicting higher ZT values due to overfitting (higher variance error) as observed in previous fits in Fig. 2. Further, DNN is predicting erroneous profile of Zt as a function of chemical potential (Fig. 3(c) left, and (e) both) as they are not physically realistic. Thus, the split in training data is not benefitting DNN. The solution for overfitting in DNN is to either build artificial neural network (ANN) models with simpler architecture or to generate a larger training dataset.

Since the intention of this paper was to gain knowledge about possible behavior of DNN in current material screening applications (where most have limited dataset), so simpler ANN models were not shown in this work. DNN despite being the most popular model today does not work when dataset is limited.

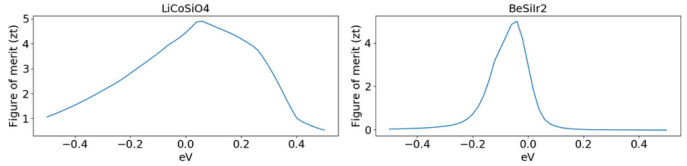
Validation of Selecting Training/Test Dataset and Model Selection: In the literature, currently the materials of the form Mg2LiSi are under investigation [3]. Li2MgSi is the closest form that has been predicted by RF in the balanced Case 3 training/test dataset. This work shows the importance of balancing training/test dataset when the dataset is limited and when, the trained model has to have good generalization ability so as to find materials in new chemical space. Most of the complex compounds predicted by DNN are not possible to test experimentally in lab, but the overfitting seen in DNN performance



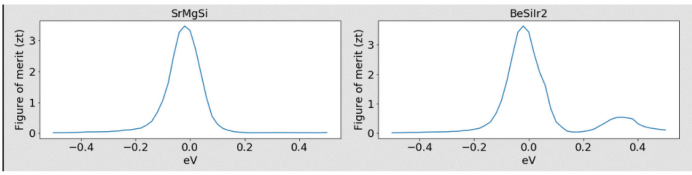
(a) CASE 1. DNN.



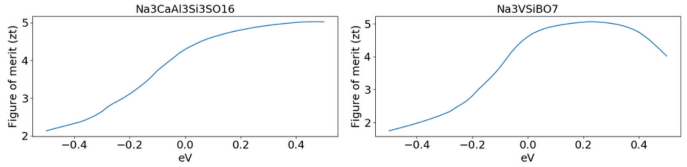
(b) CASE 1. RF.



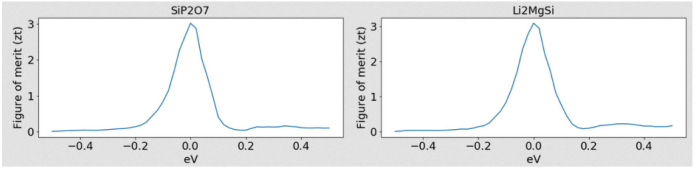
(c) CASE 2. DNN.



(d) CASE 2. RF.



(e) CASE 3. DNN.



(f) CASE 3. RF.

Fig. 3. DNN vs RF (shaded) predicted best two thermoelectric materials for the three cases. eV refers to chemical potential on the horizontal axis. DNN suggests more complex compounds as compared to the Random Forest.

suggests that it is better not to pursue those complex models (as the results may not be reliable).

Computational Efficiency: For DFT alone, the CPU consumption is between 25 and 1500 h to evaluate Z_T value of a composition (compound), and the average CPU time per compound is 85 h for finding Zt of material. It would take around 4,08,000 CPU hrs for discovering the material with best ZT amongst the 4800 compound chemical search-space. For ML step alone, the computation cost for obtaining Zt values of about 4800 compounds, after getting trained on dataset of 115 compounds is: 132 s for DNN and 80 s for RF. The cost of preparing training base for these 115 compounds from DFT could be around = 85 h per compound $\times$ 115 compounds =9775 h. Thus, we can neglect the 132 s from DNN and 80 s of RF with respect to the 9775 h required to generate the training database. Thus, the total cost for evaluating Zt using ML approach for 4800 compounds is **just 2% of time** needed by the DFT-alone method.

4 Conclusions

1. In limited dataset scenario: RF has lesser variance error than DNN and is seen to predict potentially simpler compounds from the search-space data than the DNN model. DNN predicts complex compounds from search-space data (that are difficult to make in lab and verify). Further, DNN sometimes shows physically unrealistic Zt profile prediction due to overfitting and the solution to this is that only more data can make the DNN better.
2. Significant influence of training-test split on the model is seen despite using CV procedure to select the best model parameters for generalization. Hence, when dataset is limited - this aspect should be checked. Amongst the three cases (two random and one deterministic train-test split), the variances error lowered for the case where training data could encompass compounds with extreme features. The RF model also provided the 'verifiable' predicted potential thermochemical in search-space (Li2MgSi) from this balanced deterministic train-test dataset, but this strategy did not benefit DNN.
3. Combined DFT and machine learning approach with RF is computationally efficient than an approach involving DFT alone.

Acknowledgment. We would like to thank SINTEF Foundation for the internal SEP funding for enabling the methodology development.

References

1. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
2. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
3. Nieroda, P., Kolezynski, A., Oszajca, M., Milczarek, J., Wojciechowski, T.: Structural and thermoelectric properties of polycrystalline p-type Mg2-x LixSi. J. Electron. Mater. **45**, 3418 (2016)