Symbolic Regression as a Tool for Solid Polymer Electrolyte Conductivity Estimation

1. Introduction

The global demand for batteries is projected to increase fourteen times from its 2018 figure by the year 2030. It is estimated that approximately 89% of this increase in demand will be attributed to electric mobility including electric bikes, cars, buses and planes with the rest being due to energy storage and consumer electronics [1]. Despite the growing need for better batteries, the fundamental technology in consumer level products has not significantly changed since the 1990s. In recent years, solid state batteries (SSB) have regained attention throughout academia due to their high theoretical capacity and energy density. The major difficulty in developing SSBs is finding a capable electrolyte. Current lithium ion batteries are beneficial because liquid electrolytes have high ionic conductivity (>10^{-3} S/cm) over a broad range of temperatures and good compatibility with different electrodes. The issue lies in the fact that liquid electrolytes react violently with lithium metal which restricts their use to intercalation-type cells, therefore limiting performance. Additionally liquid electrolytes have special packing needs and contain highly flammable organic solvents subject to thermal runaway. For these reasons, researchers have been looking towards a few alternatives including ceramic, gel and polymer electrolytes. Solid polymer electrolytes (SPE) are a promising alternative because of their low cost, flexibility, good interfacial compatibility, safety and ease of manufacture.

The most common polymer used in SPEs is polyethylene oxide (PEO). PEO is a very good candidate for an electrolyte material because it complexes well with Li salts, but it suffers from low ionic conductivity at room temperature. Development of PEO based SPE's is rather slow due to the complexity of the underlying ionic conduction mechanisms. PEO is a semicrystalline polymer which has a highly conductive amorphous phase attributed to increased chain mobility. Therefore, increasing ionic conductivity is a matter of increasing amorphous phase content in the electrolyte system. There are a few ways of approaching this task, the first being to introduce filler particles in the PEO/Li-Salt matrix. Introducing conductive fillers, typically ceramics containing lithium, or non-conductive fillers such as silica and alumina into the polymer system has proven to be a suitable means of increasing conductivity. An alternative method is to use a plasticizer which is generally a low molecular weight organic.

Although progress has been made in the field of developing highly conductive SPEs, current knowledge is limited by how many experiments are run and how much time is devoted to exploration of individual SPE systems. Recently, researchers have looked towards implementing machine learning for a better understanding of these systems and prediction of properties. In 2020, Yanming Wang et al. [2] used Bayesian optimization to assist in coarse grained molecular dynamics studies of the PEO-LiTFSI system. Using inputs such as molecular size and intermolecular interaction strength they were able to calculate ionic transport properties and optimize the CGMD parameters to achieve the peak ionic conductivity. In a less intensive approach, Marianne Liu et al. [3] used data from literature to create a random forest model which predicted ionic conductivity with increasing salt content of the same system, PEO-LiTFSI. The authors were able to predict ionic conductivity for 40% LiTFSI with an R^2 of 0.8578.

In this study they also predicted activation energies and correlated the decrease in activation energy with increases in conductivity for different salt contents. These approaches prove that machine learning can work for simple systems, but SPE systems are often more complex than a single polymer and salt.

In 2011, researchers utilized a neural network to predict the conductivity of a plasticized SPE using weight percent of each component and temperature as a composition [4]. The lack of metrics in the paper makes it difficult to assess the success of their study and similar approaches are scarce in the literature. A drawback of using a neural network for this task is that there is often not a large amount of data available for conductivities at different temperatures and compositions. This study aims to explore the efficacy of a different approach, symbolic regression, in predicting conductivities of a complex SPE systems as a function of composition and temperature on a small dataset.

2. Methodology

2.1. Data acquisition, curation and splitting

Machine learning generally requires large datasets which can often be found on open source websites, however there is not a lot of data for solid polymer electrolytes readily available. Due to the lack of data and the scope of the task, data was compiled from a series of conductivity measurements taken on a particular SPE comprised of a PEO host matrix, LiTFSI, halloysite nanotubes (HNT) filler particles and two plasticizers which will be referred to as A and B. Prior studies on the PEO/LiTFSI/HNT system found an optimal EO:Li ratio of 15:1 and HNT content of 10 wt% [5]. The current work involves varying plasticizer contents of A and B to study the effect on temperature therefore, each set of samples contained the same amount of PEO, LiTFSI and HNT. The original dataset contained 83 rows and 4 columns. The columns were the weights of A and B, temperature and the logarithm of ionic conductivity.

The input features were all scaled by column using a min-max scaler to obtain values between 0 and 1. The columns were scaled individually due to the different magnitude of weight and temperature data. The logarithm of ionic conductivity was left unscaled since the values were already small and did not have a large variation. The data was then grouped such that all points corresponding to the same composition of A and B had to be together in either the train or test sets to assure fair scoring. Due to the small amount of data there was no validation set for the majority of this work. For comparison of regression models, the data was split into train and test sets with an 80:20 split. For Bayesian optimization of the symbolic regressor training, validation and test splits were made with a 70:10:10 split.

2.2. Symbolic Regressor

Symbolic regression is a tool which uses a large set of equations to fit the given data. This method was chosen because of the small amount of data acquired for regression. Classical

models such as linear, lasso, ridge and random forest typically do not do well on small datasets because they essentially try to tune parameters of a pre-defined model to the data. However, symbolic regressors pick a model that fits the data well which means that it has more potential for performing well on small datasets than the classical models.

For the work done in this study a symbolic regressor using genetic programming was utilized from gplearn. The key components of genetic programming are that the regressor can breed, mutate and evolve to find the best result. In the case of symbolic regression, the starting point is a set of naïve equations which are generated by the regressor and specified by an input population size. From this initial set, the equations breed in tournaments which output the best equations for future generations. This is the general process which allows for evolution of the equation and it carries out until the number of generations or a stopping criterion is met. During this process mutation can also occur in three forms: subtree, hoist and point mutation. Subtree mutation replaces a random subtree within the tournament winner and replaces it with a new randomly generated subtree to produce offspring in the next generation. Hoist mutation takes a subtree of the tournament winner and replaces it with a subtree of that subtree to produce offspring in the next generation. Point mutation replaces random nodes within a tournament winner tree. The probability of these types of mutation can all be specified and help to control bloat, a phenomenon which describes the evolution of large programs with minimal increases in fitness. Crossover is a similar concept to mutation, but rather than randomly generating a subtree to replace a tournament winner, a second tournament is created and a subtree from that tournament winner is used to replace the original subtree [6].

2.3. Implementation

The initial part of this study focused on comparing symbolic regression with classical models. This was carried out by using five classical models: linear, lasso, ridge, random forest and support vector regression. For each of these regressors hyperparameter tuning was done using a grid search and the best parameters were used to fit the training data. The symbolic regressor was roughly manually tuned on the training data for the first portion of this study. The exact parameters tuned can be seen in the code on github. Mean absolute error (MAE), R^2 and root mean squared error (RMSE) metrics were recorded on each model with the test predictions to assess model performance. After the metrics were recorded for each of the models, hyperparameter tuning was performed on the symbolic regressor using Bayesian optimization with the Ax platform.

3. Results

Initial findings revealed that the symbolic regressor outperformed all five classical models in all three metrics. The results of the grid searches and manually tuned symbolic regression are shown in Figure 1. The simple classical models (linear, lasso, ridge) performed poorly with R^2 scores in the range of 0.62-0.65. Random forest performed better with an R^2 of

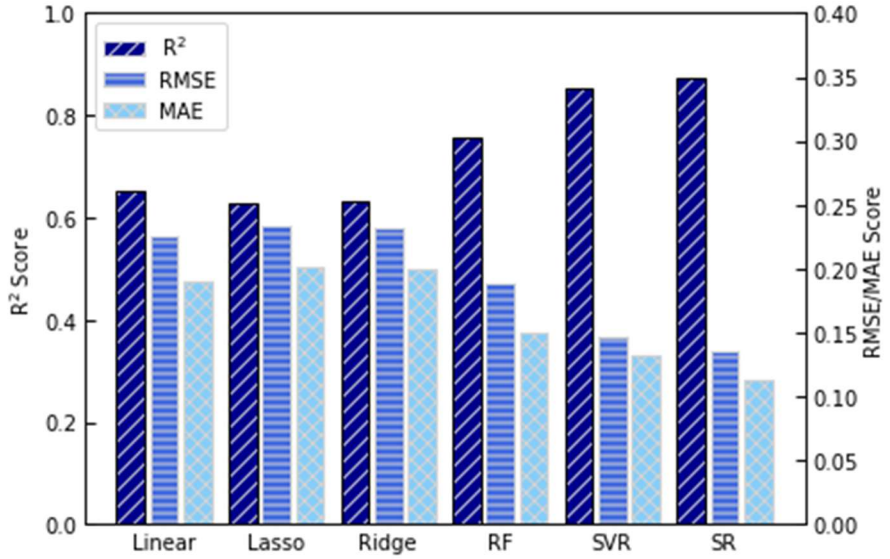

Figure 1. R^2, RMSE and MAE metrics recorded for each regression type. The color of the bars indicates the metric, and the score can be seen by the corresponding y axis.

0.7568. The manually tuned symbolic regressor only marginally outperformed the support vector machine with an R^2 of 0.8738, MAE of 0.1127 and RMSE of 0.1355 as compared to an R^2 of 0.8529, MAE of 0.1325 and RMSE of 0.1463. The metrics increased considerably after hyperparameter tuning of the symbolic regressor. Side by side parity plots are provided in Figure 2 for direct comparison between the manually tuned and Bayesian optimized hyperparameter tuned symbolic regressors. It is clear to see that the linear fit is aligned much better for the Bayesian optimized model and the MAE and RMSE scores are significantly lower than all of the other models.

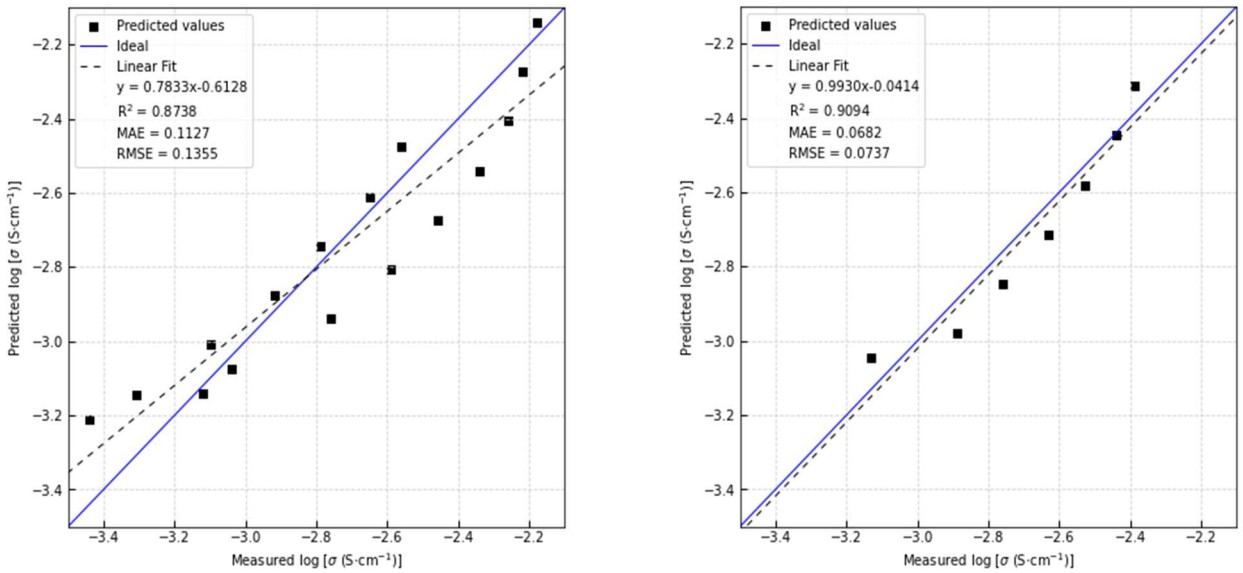

Figure 2. Parity plots of the manually tuned symbolic regressor (left) and Bayesian optimized hyperparameter tuned symbolic regressor (right). The Bayesian optimized model has less predicted values due to the implementation of a validation set for scoring during hyperparameter tuning. The set is therefore half the size.

To gain a better understanding of the application of this work a standard log conductivity plot was constructed for the optimized symbolic regressor and is shown in Figure 3 along with a surface plot of the resultant equation used for prediction. Equation 1 shows the logarithm of conductivity prediction based off of the three input parameters generated by the optimized symbolic regressor.

$$\log(\sigma) = A + T - B - 3.6376 \qquad \text{(Eq. 1)}$$

It can be seen that the predictions don't quite capture the curvature of the measured value trend due to the simplicity of the best fit equation. However, the model predicts the downward trend and the values are fairly close to the measured values. According to the best fit equation, the logarithm of conductivity increases linearly with temperature and the amount of A in the electrolyte but decreases linearly with the amount of B added.

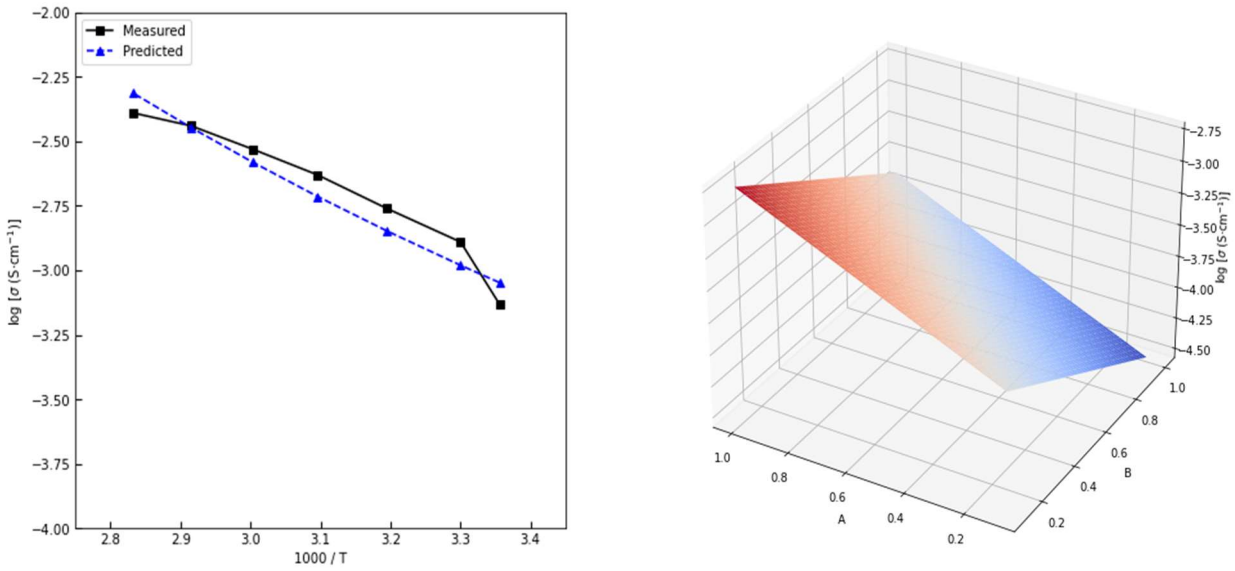

Figure 3. Logarithm of conductivity predicted by the optimized symbolic regressor plotted against 1000/T (left). Surface plot of the best fit equation generated from hyperparameter tuning (right). The x and y axes correspond to the amount of A and B and the z axis corresponds to the logarithm of conductivity. This plot was generated for room temperature (T = 25°C). The color scale does not have any numerical meaning, it is used simply to make the surface more readable.

Conclusions

The symbolic regressor proved to be significantly better at fitting the small dataset than almost all of the classical models it was compared against even without exhaustive hyperparameter tuning. After Bayesian optimization of hyperparameters the symbolic regressor was able to achieve an R^2 score above 0.9 and had significantly lower MAE and RMSE scores than the previously comparable support vector regressor. Although this work was not very interpretable due to the fact that the model is specific to the SPE chemistry studied here, the approach could be adopted for other chemistries. This work proved that symbolic regression

could be a useful tool for predicting SPE properties based on compositional and temperature features on small datasets.

The implication of this work is that symbolic regression could be used to help optimize SPE chemistries when it is too costly and time intensive to perform extensive research into different compositions. In this study this was not exactly seen because the dataset did not include other features which are influenced by the composition such as specific capacity of a battery made with the SPEs, charge transfer resistances or film strength. For symbolic regression to be truly useful, more data would need to be acquired which account for mechanical and electrochemical properties other than lithium-ion conductivity. Continued work will be carried out to add these features to the dataset as the data becomes available. It will likely be necessary to perform multi-objective optimization with these different desirable properties to figure out the best composition. With more points (~150) the model would probably do a better job at fitting the curve rather than using a linear approximation.

Code/Data Availability

The code and data used to complete this work is available on github: https://github.com/pmuller17/MSE-7050-Final

Acknowledgments

Classical model choice and grid searches performed on the classical models were adapted from the paper produced by Marianne Liu et al [1]. This work was made possible due to the symbolic regressor provided by gplearn and the Adaptive Experimentation (Ax) platform.

References

[1] "A Vision for a Sustainable Battery Value Chain in 2030," *World Economic Forum*, Sep-2019. [Online]. Available:https://www3.weforum.org/docs/WEF_A_Vision_for_a_Sustainable_Battery_Value_Chain_in_2030_Re port.pdf. [Accessed: 23-Feb-2022].

[2] Y. Wang, T. Xie, A. France-Lanord, A. Berkley, J. A. Johnson, Y. Shao-Horn, and J. C. Grossman, "Toward designing highly conductive polymer electrolytes by machine learning assisted coarse-grained molecular dynamics," *Chemistry of Materials*, vol. 32, no. 10, pp. 4144–4151, 2020.

[3] M. Liu, C. Clement, K. Liu, X. Wang, and T. D. Sparks, "A data science approach for advanced solid polymer electrolyte design," *Computational Materials Science*, vol. 187, p. 110108, 2021.

[4] M. R. Johan and S. Ibrahim, "Optimization of neural network for ionic conductivity of nanocomposite solid polymer electrolyte system (PEO–lipf6–EC–CNT)," *Communications in Nonlinear Science and Numerical Simulation*, vol. 17, no. 1, pp. 329–340, 2012.

[5] Y. Lin, X. Wang, J. Liu, and J. D. Miller, “Natural halloysite nano-clay electrolyte for advanced all-solid-state lithium-sulfur batteries,” *Nano Energy*, vol. 31, pp. 478–485, Nov. 2016.

[6] “API reference¶,” *API reference - gplearn 0.4.2 documentation*. [Online]. Available: https://gplearn.readthedocs.io/en/stable/reference.html. [Accessed: 03-May-2022].