Data-Driven Discovery and Design of Superionic Lithium Conductors for High Performance Lithium Ion Batteries

Aditi Krishnapriyan[1], Austin Sendek[1]

Abstract
We present a data-driven model for predicting the lithium ionic conductivity of a solid crystal from atomistic structure. We employ ordinary least squares (OLS) and locally-weighted least squares (LWLS) regression, coupled with manual dimension reduction and leave-one-out cross-validation (LOOCV), to identify linear hypotheses that can predict ionic conductivity within roughly two orders of magnitude on the siemens per centimeter scale. With a LOOCV error of 0.59σ, where σ is the training set ionic conductivity standard deviation, the OLS model approaches but does not break the conventional performance threshold of 0.5σ employed by the Quantitative Structure-Property Relationship (QSPR) community. The LWLS model has a LOOCV error of 0.41σ and thus breaks this threshold. We then screen 1,200 candidate lithium conductors with the OLS hypothesis to search for promising new lithium conductors, and employ the more accurate LWLS hypothesis for screened materials lying deep inside the region in data space defined by the training set.

[1] *Stanford University, Stanford, California*

Contents

Introduction

Many of the major technological developments of the young twenty-first century are increasingly reliant on the ability to locally store energy. Between the grid-scale penetration of renewable power systems, worldwide proliferation of mobile devices, and revolutions in hybrid and fully-electric vehicles, the future requires power on demand. However, progress in many of these diverse areas has been tempered by the same technological bottleneck: poor battery performance.

Lithium ion batteries (LIBs) are one of the most promising energy storage technologies for a wide array of applications, including mobile devices and transportation. However, LIBs are dogged by safety concerns: Boeing lost millions of dollars after it was forced to ground their entire fleet of 787s after an LIB pack caught fire, and news of a smoldering battery fire in a crashed Tesla Model S caused Tesla Motors' stock to drop by 6%. The root cause of these safety problems is the liquid electrolyte, which in most commercialized LIBs is a volatile and flammable organic solvent. Thus, solid-state lithium ion batteries hold promise as safer, longer-lasting alternatives to conventional liquid electrolyte batteries. The bottleneck for this technology is simply the dearth of quality solid electrolyte materials; solid electrolytes with higher ionic conductivities must be discovered before solid-state LIBs are competitive with liquid electrolyte LIBs.[3]

Over the past four decades, much effort has been given to searching for solid electrolyte materials that exhibit high ionic conductivity (these materials are known as "superionics"). Ionic conductivity is notoriously hard to intuit from crystal structure, and so identifying promising materials is largely a game of guess-and-check. However, checking guesses with experimental methods or electronic structure simulations can take days or weeks. Because of this, progress in the search for new superionics has been slow, and high conductivity materials are often discovered serendipitously. These is reason to be optimistic for progress, considering the tens of thousands of known, untested, lithium-containing materials, but exploring this space through guess-and-check experimentation is not tractable. Instead, we employ OLS and LWLS regression

techniques to build ionic conductivity predictors, thus facilitating time-efficient screening of these tens of thousands of materials and identification of the most promising candidates. Importantly, our features are based only on the atomistic structure (i.e. atom positions only, not electronic structure) of the crystal, so they can be calculated almost instantaneously.

1. Related Work

Using conventional theoretical methods for calculating ionic conductivity, the atomistic structure of a crystal is used to calculate the electronic structure (using e.g. Density Functional Theory [5]), and then the electronic structure is used to calculate ionic conductivity. However, calculating the electronic structure is time-consuming (oftentimes more time-consuming than actually synthesizing the material and measuring its ionic conductivity experimentally), and so knowing direct correlations between atomistic structure and ionic conductivity would be valuable. This connection between atomistic structure and ionic conductivity has been explored in recent years, though to our knowledge we present the first application of machine learning to the problem.

Reviews of superionics by Hull [3] and Avdeev, et al [4] mention several crystal characteristics that may be correlated to ionic conductivity, including the bonding character of the mobile ion (in our case, lithium), mobile ion concentration, and the polarizability of the "bottleneck" atoms blocking the passage of the mobile ion from one equilibrium site to another. However, Hull admits, "a number of plausible factors are likely to be influential, including the crystal structure... however, the relative importance of any one of these factors in promoting superionic behavior is difficult to assess, owing to the relatively small number of known superionic phases and the inherent interrelationship between the individual factors."[6] Our goal is to use machine learning techniques to identify these important factors and tease out the interrelationships.

Much work has been done by Adams and colleagues in the last fifteen years to develop a method for quickly calculating ionic conductivity from atomistic structure using the bond-valence (BV) method.[7] The BV method outputs a scalar measure of the amount of "empty space" in the crystal, which represents the free space for mobile ions to migrate through. This method has been cross-checked against experimental data and it has been noted that the cube root of this measure is linearly correlated with the base-10 log of ionic conductivity. However, this measure was only validated among glassy crystals of very similar stoichiometry, and so it remains to be seen whether it can accurately predict ionic conductivity across a set of diverse crystals. In the language of machine learning, this technique is simply a one-feature linear predictor, and thus it stands to reason that it is likely less accurate than a technique which incorporates several features.

Additionally, further hope for a connection between atomistic structure and ionic conductivity has been inspired by recent work out of the Ceder group.[9] Without using any learning techniques, a correlation between the arrangement of the anions in the sublattice and ionic conductivity was identified: the closer the arrangement of anions in the lattice represents a *bcc* lattice, the larger conductivity tends to be. Although this gives weight to the argument that atomistic structure can be directly tied to ionic conductivity, it does not address the fact that many fast ion conductors do not have this arrangement, nor was this trend discovered in a systematic, transferrable way. It is likely only one favorable arrangement of many, and machine learning represents an exciting way to discover a more generalized relationship from which this and other favorable arrangements may be found.

2. Dataset and Features

The data for the training set is composed of atomistic structures and ionic conductivity values for a range of materials. These ionic conductivity values are measured experimentally and reported in the scientific literature in units of Siemens per centimeter, S/cm. The corresponding structure files for the given conductivity measurements were downloaded from the Inorganic Crystal Structure Database (ICSD).[1] However, ionic conductivity data for crystalline materials is somewhat hard to come by and as a result, our data set consists of only 38 materials. Several of the training structures and their corresponding conductivities are listed in Table 1. The range of conductivities spans over 11 orders of magnitude (from 10^{-13} to 1.4×10^{-2} S/cm), so we take the base-10 log of the values. After taking the log, the mean ionic conductivity is μ = -5.93, and the standard deviation is σ = 2.49. Before running any regressions, we mean-center the data and normalize by the standard deviation. Given the small size of the training set, we do not set aside a separate test set (see the Methods section for more on how we prevent overfitting).

Table 1. Select training set structures and ionic conductivities (S/cm)

Formula	Ionic conductivity	Reference
$Li_{0.5}La_{0.5}TiO_3$	1.0×10^{-3}	[11]
$Li_{10}Ge(PS_6)_2$	1.4×10^{-2}	[10]
$Li_{14}ZnGe_4O_{16}$	1.0×10^{-6}	[12]
$Li_2Ca(NH)_2$	6.4×10^{-6}	[13]
$Li_2Ge_7O_{15}$	5.0×10^{-6}	[14]
Li_2NH	2.5×10^{-4}	[13]
Li_2S	1×10^{-13}	[15]
$Li_{3.4}Si_{0.7}S_{0.3}O_4$	6.0×10^{-7}	[16]

From the training set of crystal structures, we extract 19 features based on simple crystallographic and chemical relationships, including many of those elaborated by Hull and others.[3][4] This feature extraction is done automatically for each structure with a MATLAB script. Several of the 19 feature values for the fast lithium conductor lithium triphosphide (Li_3P) are given in Table 2. All feature values are mean-centered and normalized by their standard deviation to ensure equal weighting in the regression. The feature extraction takes

roughly ten seconds of wall time per structure.

The structure files for the high-throughput screening were taken from the Materials Project (MP) database, a publicly accessible database of materials.[2] We download the entire database of nearly 60,000 materials with permission through the API, and then filter this down to 1,283 non-metallic lithium-containing materials.

Table 2. Some atomistic features for lithium triphosphite

Average atomic volume	$11.09 Å^3$
Average Li bond ionicity	0.93
Average Li coordination	1.76
Li concentration	$0.013 Å^{-3}$
Average smallest Li-Li distance	$3.07 Å$
Lattice packing fraction	0.33
(Normalized) distribution of empty space	0.52
Ratio of Li bond ionicity to sublattice ionicity	2.33
Average Li coordination to sublattice coordination	1.27

3. Methods

To identify a model with high predictive capability, we employ both OLS and LWLS regression with an exhaustive search over all possible models. In the following sections we describe the regression algorithms in detail, followed by a description of our model search algorithm.

3.1 Ordinary least squares linear regression

For a given feature set $\{x_i, i = 0, 1, 2, ..., n\}$, where $x_0 = 1$, multiple OLS linear regression provides a parametric linear hypothesis of the following form:

$$y = \sum_{i=0}^{n} \theta_i x_i = \theta^T X$$

where θ is the vector of coefficients and X is the design matrix. In order to find the coefficients θ_i, we minimize the following cost function with respect to θ:

$$\theta = \text{argmin}_\theta J(\theta) = \text{argmin}_\theta \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2$$

Given the small size of our data set ($m = 38$), we are able to quickly find these coefficients exactly using MATLAB's mvregress function, which evaluates the OLS normal equation:

$$\theta = (X^T X)^{-1} X^T y$$

3.2 Locally-weighted least squares linear regression

The application of OLS regression assumes the dependence between the features and conductivity is linear. We have no reason to suspect that this dependence is linear *a priori*. Because of this, we also build a model using locally weighted linear regression. For LWLS linear regression, while the hypothesis has the same form as OLS linear regression, the cost function to minimize has the following form:

$$J(\theta) = \sum_{i=1}^{m} w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2,$$
$$\text{where } w^{(i)} = exp\left(- \frac{(x^{(i)} - x)^2}{2\tau^2} \right)$$

In this case, the weight terms w^i depend on the particular query point x. More weight is given to the training examples near the query point. Thus, θ is a function of the query point x, and the hypothesis is non-parametric. To find θ, we minimize the cost function accordingly, with the weighted term:

$$\theta = argmin_\theta J(\theta) = argmin_\theta \sum_{i=1}^{m} w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2$$

The bandwidth parameter τ defines what is "local" to the particular query point and controls the degree of smoothing over the training set. In practice, τ must be optimized numerically.

3.3 Exhaustive model search with LOOCV

Due to the small size of the training set (m = 38) and the high number of extracted features (n = 19), we take special care not to overfit the data. Thus, our task is to reduce the dimensionality of the training set and evaluate the predictive error through cross-validation as a function of the reduced set of features. However, rather than performing parameter regularization or a forward or reverse propagation (which we performed and found only marginal success with), we find the best performing model by actually performing cross-validation on every one of the 2^{19} - 1 = 524,287 possible models, a process that is enabled by the small training set.

For each model, we compute the training set error and the LOOCV error using both OLS and LWLS techniques. We then select the OLS model that gives the lowest LOOCV error of all OLS models, as well as the LWLS model that gives the lowest LOOCV error of all LWLS models. We use both of these models in the subsequent screening step. We note that the reduced set of features is not necessarily the same for the OLS and LWLS models.

We use LOOCV over k-fold cross-validation because of the small training set; we do not want to remove too much data from the training set at one time. For the same reason, we do not partition our data into a training and test set. Instead, we use the root-mean-squared LOOCV error as a proxy for the test set error.

4. Results

4.1 Ordinary least squares regression

In evaluating the mean squared training set and LOOCV error over all models using OLS, we identify a six-feature model with a mean-squared LOOCV error of 0.59σ and a training set error of 0.40σ. This LOOCV error corresponds to a predictive error of 1.9 orders of magnitude in S/cm, which, although fairly high, is still useful for identifying the best conductors on the 15+ orders of magnitude scale on which they live. In order to see the effect of under and overfitting, we plot in Figure 1

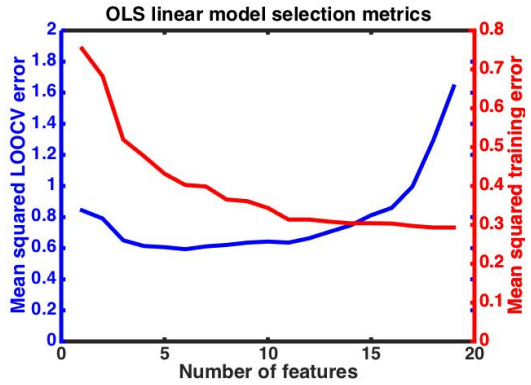

Figure 1. Blue curve: Mean squared LOOCV error for the lowest LOOCV x-feature model as a function of x. Red curve: Mean squared training set error for the lowest LOOCV x-feature model.

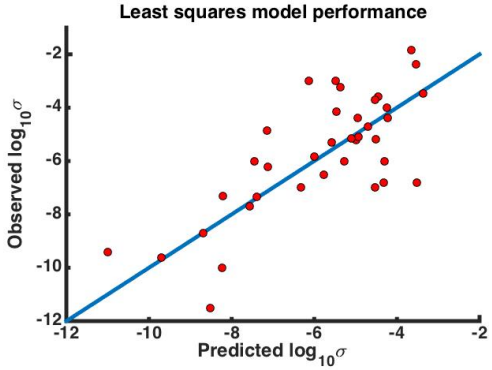

Figure 2. Observed vs. predicted log (conductivity) for the training set materials using OLS regression on the optimal six-feature subset, identified through LOOCV.

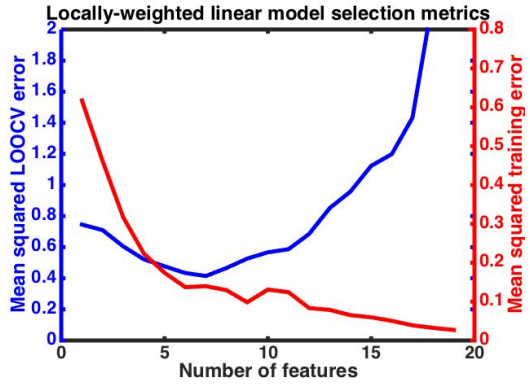

Figure 3. Blue curve: Mean squared LOOCV error for the lowest LOOCV x-feature LWLS model as a function of x. Red curve: Corresponding training set error.

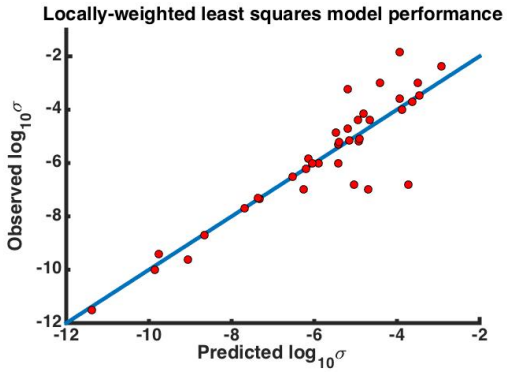

Figure 4. Observed vs. predicted log (conductivity) for the training set materials with the optimal LWLS model.

the mean squared LOOCV error for the lowest LOOCV x-feature model as a function of x. We see that, generally, the LOOCV error is higher for small x (high bias), minimized at intermediate x, and high again at large x (high variance). There is a minimum representing the optimal model at $x = 6$. Figure 2 shows the plotting of the observed vs. predicted log (conductivity) for the training set materials.

4.2 Locally-weighted least squares regression

Before a model can be chosen with LWLS, the bandwidth parameter τ must be optimized. To do this, we run the exhaustive model search for a range of values of τ and select the τ value which gives the lowest LOOCV error. However, we must take into account that we are searching over models of varying dimensionality, and the Euclidean distance between points separated by a given distance in all dimensions scale with the square root of the number of dimensions. Thus we choose $\tau = \alpha\sqrt{D}$, where D is the number of features. We perform the model search for varying values of α running from 0.01 to 30; we find that the most useful values lie in the range of 0.1 to 1 with an optimum at 0.5. These trends are plotted in Figure 5.

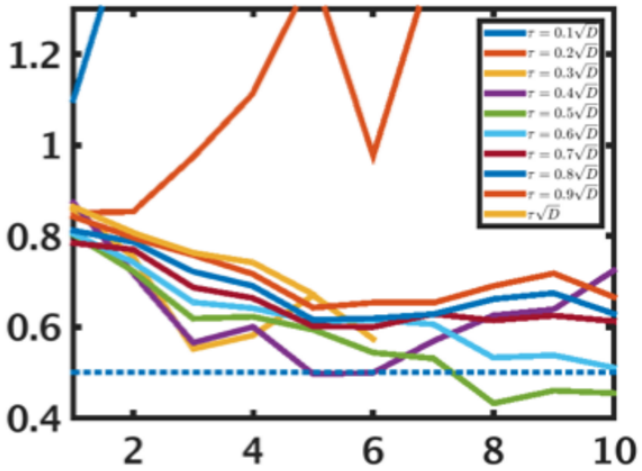

Figure 5. We plot the LOOCV of the lowest LOOCV x-feature model for $1 \leq x \leq 10$. We see the model is optimized (LOOCV error is minimized) at $\alpha = 0.5$.

Thus, the exhaustive model search technique with the optimized $\tau = 0.5\sqrt{D}$, we find the optimal model has seven features, a training error of 0.14σ and a mean squared LOOCV error of 0.41σ. This level of error corresponds to a predictive error of 1.6 orders of magnitude in S/cm, making the LWLS model more accurate than the linear model. The model selection and model performance data for this model is plotted in Figures 3 and 4 in the same style as the aforementioned OLS metrics.

4.3 Screening

After developing our OLS and LWLS hypothesis functions, we use them to predict conductivity to over 1,200 candidate materials from the Materials Project database.[2] This database contains the atomistic structures and DFT-computed electronic structure for over 60,000 materials; we download the entire database with permission through the API and then filter out all materials except those that contain lithium and are not metallic (i.e. those with a band gap greater than zero). For each of the remaining 1,283 materials, we extract the necessary features and evaluate the OLS hypothesis function. We choose to evaluate all structures with the OLS function first because it extrapolates better outside the training set region. The locally-weighted function requires training data nearby to give meaningful predictions. However, we do evaluate the locally-weighted hypothesis for those data points which have normalized feature values between -1 and 1 for all features. The predictions of the most promising materials are listed in Table 3. These predictions are calcluated with the OLS model and LWLS corrections are given if the point is within the training set region in space (all features must be within one standard deviation of the mean). Due to the potentially high-impact nature of this research and the likelihood of future publication, we have withheld the actual chemical formulas and simply replaced them by integer material indeces.

Table 3. Promising ionic conductivity predictions from the Materials Project structures from the OLS model and the LWLS model, if applicable (chemical formulas not displayed).

Material index	OLS	LWLS
1	2.3×10^{-2}	N/A
2	1.1×10^{-2}	1.0×10^{0}
3	8.3×10^{-3}	N/A
4	5.7×10^{-3}	1.8×10^{-3}
5	4.1×10^{-3}	9.9×10^{-4}
6	2.9×10^{-3}	N/A
7	1.2×10^{-3}	N/A
8	4.6×10^{-4}	8.1×10^{-4}

5. Conclusions and Future Work

The development of new approaches for identifying superionic lithium conductors is of critical importance for the advancement of solid-state lithium ion batteries. We develop a structure-property model for ionic conductivity using ordinary least squares and locally-weighted least squares regression, which are capable of predicting ionic conductivity within approximately two orders of magnitude on the S/cm scale. We use these hypotheses to then rapidly screen over 1,200 candidate lithium containing materials to search for new, undiscovered superionics and find many promising candidates for future investigation.

Future theoretical work may focus on finding more robust atomistic descriptors which can predict ionic conductivity with higher accuracy, or on applying new algorithms like support vector regression. A larger training set would be helpful, and so developing a method for automated data scraping of ionic conductivity values and structures from the literature would likely improve the model accuracy and robustness of the analysis manyfold.

Additionally, an in-depth study of the features and their coefficients that survive the model reduction is likely to be scientifically informative. From these features, we may be able to build new intuition on which classes of materials are unlikely to ever be highly conductive and which show great promise. This may enable many new systematic studies of various structure types such as that performed by Wang et al.[9]

Acknowledgments

The authors of this research paper would like to thank Professor Evan Reed, who helped conceive the ideas and guide the project. Prof. Reed also provided many valuable resources including texts on property prediction in molecules and materials. We would also like to thank Professor Yi Cui, who provided the motivation for the application to solid state batteries and provided valuable feedback on technologically feasible requirements. We thank our collaborators in the Reed group, Qian Yang and Karel-Alexander Duerloo, who helped guide the application of machine learning and facilitated the downloading of the Materials Project database, respectively. All data analysis, figure generation, and writing was done by the authors A.S. and A.K. We acknowledge funding from the Stanford University Office of Technology Licensing through the Stanford Graduate Fellowship program for supporting A.S.'s work, as well as support for A.K. through the Department of Energy's Computational Science Graduate Fellowship. Finally, we encourage the reader to look for a forthcoming paper in a peer-reviewed journal, where we will provide the sensitive information omitted here: the full training set, feature set, and most promising screened materials.

References

[1] ICSD Basic Search and Retrieve at <https://icsd.fizkarlsruhe.de/search/basic.xhtml;jsessionid=BCC2051D0E937E117C02FBEC1C958CF3>

[2] Jain, A. et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Mater. 1, 011002 (2013).

[3] Hull, S. Superionics: crystal structures and conduction processes. Rep. Prog. Phys. 67, 1233 (2004).

[4] Avdeev, M., Nalbandyan, V. B. and Shukaev, I. L. in Solid State Electrochemistry I (ed. Kharton, V. V.) 227-278 (Wiley-VCH Verlag GmbH and Co. KGaA, 2009). at <http://onlinelibrary.wiley.com/doi/10.1002/9783527627868.ch7/summary>

[5] P. Hohenberg and W. Kohn, Phys. Rev. 136, B864 (1964).

[6] Hull, S. and Berastegui, P. Crystal structures and ionic conductivities of ternary derivatives of the silver and copper monohalides—II: ordered phases within the (AgX)x-(MX)1-x and (CuX)x-(MX)1-x (M=K, Rb and Cs; X=Cl, Br and I) systems. J. Solid State Chem. 177, 3156–3173 (2004).

[7] Adams, S. and Swenson, J. Determining Ionic Conductivity from Structural Models of Fast Ionic Conductors. Phys. Rev. Lett. 84, 4144-4147 (2000).

[8] J. Swenson and St. Adams. Structure-conductivity relations in ion conducting glasses. Ionics 9, 1 pp 28-35 (2000).

[9] Wang, Y. et al. Design principles for solid-state lithium superionic conductors. Nature Materials 14, 1026-1031 (2015).

[10] Kamaya, N. et al. A lithium superionic conductor. Nat. Mater. 10, 682-686 (2011).

[11] Jimenez, R., Rivera, A., Varez, A. and Sanz, J. Li mobility in Li0.5-xNaxLa0.5TiO3 perovskites (0 ¡ x ¡ 0.5): Influence of structural and compositional parameters. Solid State Ion. 180, 1362-1371 (2009).

[12] Bruce, P. G. and West, A. R. The AC Conductivity of Polycrystalline LISICON, Li2 + 2x Zn1 - x GeO4, and a Model for Intergranular Constriction Resistances. J. Electrochem. Soc. 130, 662-669 (1983).

[13] Li, W., Wu, G., Xiong, Z., Feng, Y. P. and Chen, P. Li+ ionic conductivities and diffusion mechanisms in Li-based imides and lithium amide. Phys. Chem. Chem. Phys. 14, 1596-1606 (2012).

[14] Liebert, B. E. and Huggins, R. A. Ionic conductivity of Li4GeO4, Li2GeO3 and Li2Ge7O15. Mater. Res. Bull. 11, 533-538 (1976).

[15] Lin, Z., Liu, Z., Dudney, N. J. and Liang, C. Lithium Superionic Sulfide Cathode for All-Solid Lithium-Sulfur Batteries. ACS Nano 7, 2829-2833 (2013).

[16] Shannon, R. D., Taylor, B. E., English, A. D. and Berzins, T. New Li solid electrolytes. Electrochimica Acta 22, 783-796 (1977).