Article

Machine learning prediction of activation energy in cubic Li-argyrodites with hierarchically encoding crystal structure-based (HECS) descriptors

Qian Zhao [a], Maxim Avdeev [b,c], Liquan Chen [a,d], Siqi Shi [a,e,*]

[a] Materials Genome Institute, Shanghai University, Shanghai 200444, China
[b] Australian Nuclear Science and Technology Organization, New Illawarra Rd, Lucas Heights, NSW 2234, Australia
[c] School of Chemistry, The University of Sydney, Sydney 2006, Australia
[d] Key Laboratory for Renewable Energy, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China
[e] School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China

ARTICLE INFO

ABSTRACT

Rational design of solid-state electrolytes (SSEs) with high ionic conductivity and low activation energy (E_a) is vital for all solid-state batteries. Machine learning (ML) techniques have recently been successful in predicting Li$^+$ conduction property in SSEs with various descriptors and accelerating the development of SSEs. In this work, we extend the previous efforts and introduce a framework of ML prediction for E_a in SSEs with hierarchically encoding crystal structure-based (HECS) descriptors. Taking cubic Li-argyrodites as an example, an E_a prediction model is developed to the coefficient of determination (R^2) and root-mean-square error ($RMSE$) values of 0.887 and 0.02 eV for training dataset, and 0.820 and 0.02 eV for test dataset, respectively by partial least squares (PLS) analysis, proving the prediction power of HECS-descriptors. The variable importance in projection (VIP) scores demonstrate the combined effects of the global and local Li$^+$ conduction environments, especially the anion size and the resultant structural changes associated with anion site disorder. The developed E_a prediction model directs us to optimize and design new Li-argyrodites with lower E_a, such as $Li_{6-x}PS_{5-x}Cl_{1+x}$ (<0.322 eV), $Li_{6+x}PS_{5+x}Br_{1-x}$ (<0.273 eV), $Li_{6+x}PS_{5+x}Br_{0.25}I_{0.75-x}$ (<0.352 eV), $Li_{6+(5-n)y}P_{1-y}N_yS_5I$ (<0.420 eV), $Li_{6+(5-n)y}As_{1-y}N_yS_5I$ (<0.371 eV), $Li_{6+(5-n)y}As_{1-y}N_ySe_5I$ (<0.450 eV), by broadening bottleneck size, invoking site disorder and activating concerted Li$^+$ conduction. This analysis shows great potential in promoting rational design of advanced SSEs and the same approach can be applied to other types of materials.

© 2021 Science China Press. Published by Elsevier B.V. and Science China Press. All rights reserved.

1. Introduction

Concerted efforts continue to be made in the search for lithium-ion solid state electrolytes (SSEs) with high ionic conductivity and low activation energy (E_a), which is vital for high performance all-solid-state rechargeable Li ion batteries [1–3]. However, only a limited number of potential materials have been obtained to date and there is an urgent need to accelerate the discovery of new advanced SSEs [4,5]. Over the past few years, successful predictions of several novel SSEs were achieved by first-principles computation [6–10]. First-principles calculations can be done at a fraction of the cost of equipment and human time and have better accuracy in lower temperatures domains, where vibrational and electronic excitations play a minor role. However, they are still too computationally expensive for high-throughput automatic unsupervised search. Recently, alternative ways to predict ion conduction in SSEs started gaining popularity, e.g., the machine learning (ML) technique. Thanks to its capability of recognizing complex patterns in data, ML has the potential to notably improve the efficiency of material discovery by developing prediction or classification models of target physical properties [11–17]. One of the critical aspects in controlling the performance of a ML model is what "descriptors", or "features", are used to represent the compounds in the data set [18,19]. The diversity and complexity of ionic conduction mechanisms in various SSEs challenge the efforts to relate conduction properties to appropriate descriptors and develop high accuracy prediction model.

Recently, several researchers have developed ML models for predicting ion conduction property of selected SSEs with various descriptors as summarized in Table 1. For instance, Fujimura et al. [20] used Li-ion diffusion coefficients, average volumes of disordered structures, free energies of solid solution formation and transition temperatures as descriptors to develop an ML model

* Corresponding author.
  E-mail address: sqshi@shu.edu.cn (S. Shi).

Table 1

Examples of ML predicting ion conduction property of selected SSEs with various descriptors.

Classes of SSEs	Descriptors	Source
LiSICON-type materials	Li-ion diffusion coefficients, average volumes of disordered structures, formation free energies of solid solutions and transition temperatures	[20]
Olivine	Lattice parameters, cation dynamic charge parameters, intra-polyhedron parameters, inter-polyhedron parameters	[21, 22]
Tavorite	Lattice parameters, cation dynamic charge parameters, intra-polyhedron parameters, inter-polyhedron parameters	[23]
Garnet	Synthesis information, geometry-based parameters, interrelations of constituent element radii with geometric principles of ion packing, etc.	[24]
Multiple type SSEs	20 Descriptors derived from local atomic arrangements and chemistry of the crystals	[25]
Sodium and lithium-based superionic conductors.	Chemical composition, atom site occupancies, unit cell parameters, unit cell volume, ionic radii, electronegativity	[26]
Li-containing materials	Modified X-ray diffraction (mXRD)	[27]

for prediction of Li ion conductivity values of LiSICON-type materials by first principles calculations combined with support vector regression. Jalem et al. extracted crystal structure information, i.e., lattice parameters, cation dynamic charge parameters, intra-polyhedron parameters, inter-polyhedron parameters, etc. as descriptors to generate robust models for the prediction of the Li-ion hopping energy to screen olivine-type [21,22] and tavorite-type [23] materials for SSEs applications by a combination of DFT computations with ML methods. Kireeva et al. [24] used several types of parameters related to different categories including synthesis information, geometry-based parameters, interrelations of constituent element radii with geometric principles of ion packing, etc. as descriptors to develop a Li-ion conductivity prediction model for virtual screening of new garnet-type SSEs by ML techniques. Sendek et al. [25] extracted information on local atomic arrangements and chemistry of the crystals as descriptors to develop an ionic conductivity classification model for several types of SSEs. Xu et al. [26] used elemental descriptors derivable from tabulated information on the unit cell to design a logistic regression-based model capable of distinguishing between poor and good SSEs across sodium and lithium-based superionic conductors. Zhang et al. [27], using modified X-ray diffraction (mXRD) as a descriptor, developed an unsupervised model for new SSEs discovery, which succeeded in distinguishing fast Li-conducting and poor Li-conducting materials and revealing unique correlations between anion lattice and Li$^+$ conduction over wide materials space.

Provided that there are large differences between the descriptors supplied for ML in different systems, it is inconvenient to directly transfer to other systems. It is noted that a broad range of descriptors, ranging from simple geometric and lattice information to complex energy information, were used in prior successful ML models for different SSEs respectively. Of these, even descriptors suitable for multiple systems [25] are not enough to delineate systems with similar structures, e.g., cubic Li-argyrodites containing consistent external frameworks and diverse internal structures. To accelerate the development of ML models for SSEs, one strategy is to reduce the time spent on developing a set of descriptors for a particular system. As part of achieving that goal, in view of the

diversity and particularity of characteristics for SSEs, it is necessary to combine ML and a set of descriptors to generalize SSEs especially with similar structures.

In this paper, we extend the previous efforts and build the ML framework relying on hierarchically encoding crystal structure-based (HECS) descriptors derived from unit cell information to predict E_a of SSEs. By integrating the information on global and local effects of ion conduction environment, we constructed the generic HECS-descriptors for Li$^+$ conduction encompassing composition, structure, conduction pathway, ion distribution and special ions, which are conveniently available for crystalline SSEs. Then the framework of ML prediction for E_a in SSEs with HECS-descriptors is established, as shown in Fig. 1. We take cubic Li-argyrodites, one of the large families of SSEs, as an example to elaborate on it.

2. Methods

2.1. Dataset for cubic Li-argyrodites

#### 2.1.1. Li$^+$ conduction property for cubic Li-argyrodites

50 Eligible cubic Li-argyrodites, with the general formula $Li_{7-x+y}^+(M_{1-y}^{5+}N_y^{4+}Y(1)_4^{2-})Y(2)_{2-x}^{2-}X_x^-$ where Y = S^{2-}, Se^{2-}; X = Cl$^-$, Br$^-$, I$^-$; M = P^{5+}, As^{5+}; N = Si^{4+}, Ge^{4+}, Sn^{4+}; $0 \leq x \leq 2$, $0 \leq y \leq 1$ (Fig. 2a), are identified using our high-throughput screening platform for solid electrolytes (SPSE) [28]. Their Li$^+$ conduction property is characterized by E_a. To preserve the experimentally determined structure information and achieve high-efficiency calculation, E_a is uniformly calculated by the Bond-Valence Site Energy (BVSE) method embedded in the SPSE [28] instead of density functional theory (DFT) method. The complete dataset of the E_a values are given in the Table S1 (online).

#### 2.1.2. Descriptors for Li$^+$ conduction in cubic Li-argyrodites

The application of ML techniques requires a set of descriptors which accurately delineate the diverse crystal structures and can be expected to be related to the Li$^+$ conduction property. Fig. 2a depicts the crystal structure pattern of cubic Li-argyrodites. Anions (X or Y) at Wyckoff site 4a are labeled as A, which form a face-centered cubic (FCC) close-packed lattice with PE_4^{3-} tetrahedron as the backbone. The PE_4^{3-} tetrahedron is formed by anions on the Wyckoff site 16e, labeled as E, and P^{5+} on the Wyckoff site 4b. The free anions (X or Y), in half of the tetrahedral sites/voids on the Wyckoff site 4c (or 4d in alternative space group setting), are labeled as C. The lithium ions are randomly distributed over the remaining tetrahedral cavities (Wyckoff site 48 h), forming Li$^+$ cages around the free anions on the Wyckoff site 4c. To delineate the crystal structures of cubic Li-argyrodites, we construct a list of HECS-descriptors that characterize the global and local Li$^+$ conduction environment, encompassing composition, structure, conduction pathway, ion distribution, and special ions (Fig. 2b). All can be obtained from the structure file stored in the CIF format. The full list of the 32 descriptors can be found in Table S2 (online).

2.2. Machine learning methods

Partial least squares (PLS) is a technique that reduces the independent variables to a smaller set of uncorrelated components and performs least squares regression on these components. PLS is especially useful when the multicollinearity is present among independent variables. Fig. S1 (online) gives a schematic outline of the PLS method. In PLS modeling, first, the collinear data matrix of independent variables is transformed into a latent variable matrix with orthogonal vectors [29], that can be used to reduce the number of independent variables. The number of components
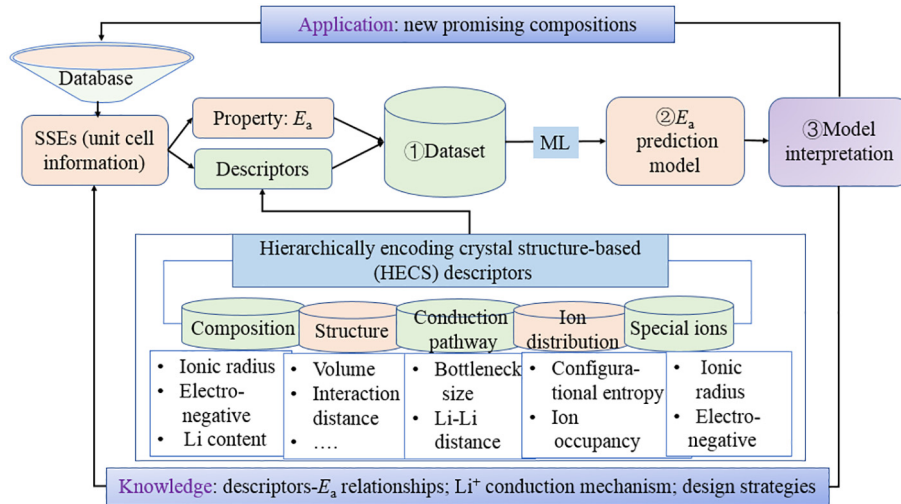
Fig. 1. (Color online) The framework of ML prediction for E_a in SSEs with HECS-descriptors derived from unit cell information.
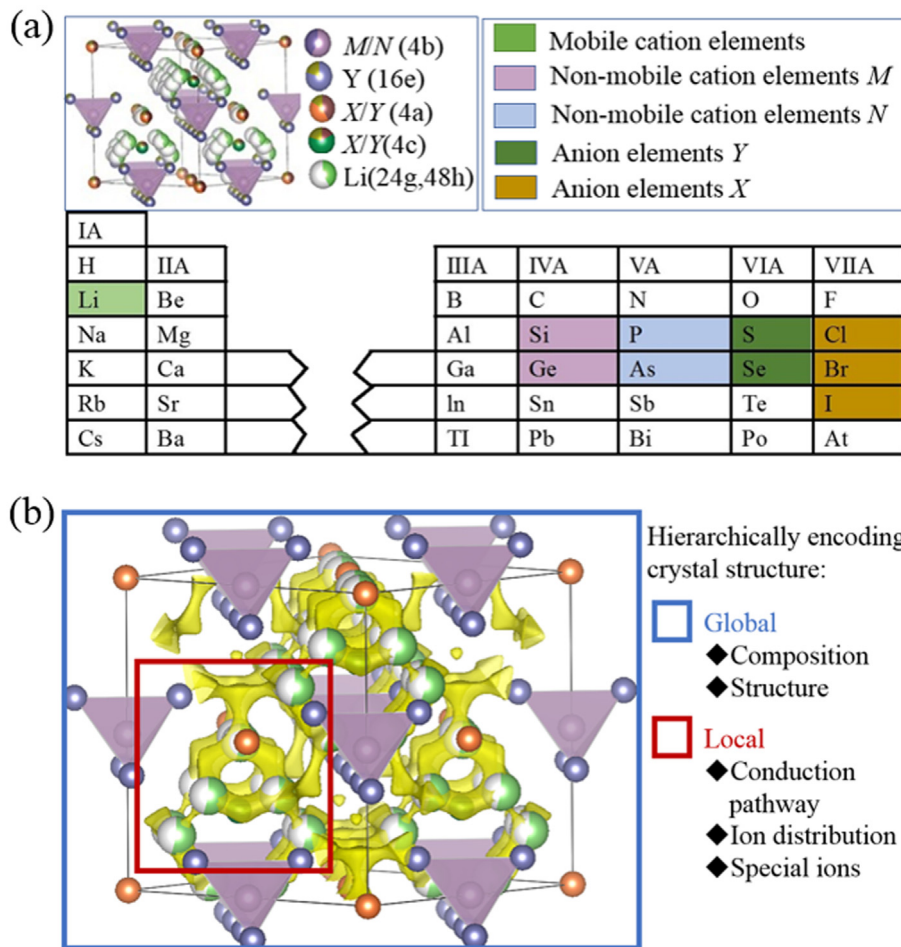
Fig. 2. (Color online) (a) Sampling scheme for the cubic Li-argyrodite $Li_{7-x+y}^+(M_{1-y}^{5+}N_y^{4+}Y(1)_4^{2-})Y(2)_{2-x}^{2-}X_x^-$ compounds. (b) Conceptual scheme of the descriptors construction by hierarchically encoding crystal structure as composition, structure, conduction pathway, ion distribution and special ions.

can be decided from the predicted residual error sum of squares (PRESS) against the leave-one-out cross validation (LOOCV) for each regression with the number of components. The PRESS with LOOCV can be expressed as:

$$\text{PRESS} = \sum_{i=1}^{n} (\widehat{y}_l - y_i)^2, \tag{1}$$

where n represents sample size, y_i is the actual value of i-th sample, and $\widehat{y}_i$ is the prediction value of the i-th sample obtained by the model through fitting all samples except the i-th sample. Generally, the optimum number of components corresponds to the minimum PRESS, as the smaller PRESS value represents the smaller prediction error and the better predictive ability of the model. Additionally, the Q^2, defined as 1-PRESS/ SSY (SSY is the sum of squared deviation from the mean), also can be used to help us to determine the number of component and understand the quality of the regression model. PRESS/SSY indicates proportion of prediction error, which is the variation unexplained by the regression model. Thus, the larger the Q^2, the better prediction ability of the model is. Then the regression model with the optimal set of components is developed. Moreover, the influence of the independent variables on the PLS regression model can be reflected by variable importance in projection (VIP) scores [29,30]. The VIP score for the i-th variable can be expressed as

$$\text{VIP}_i = \sqrt{\sum_{j}^{J} \text{w}_{ij}^2 \text{SSY}_j I / \text{SSY}_{\text{total}} \cdot J}, \tag{2}$$

where w_{ij} indicates the weights of the i-th variable and j-th component; SSY_j is the sum of the squares of y explained by the j-th component; $\text{SSY}_{\text{total}}$ is the total sum of the squares of y explained by the all components; I is the number of independent variables x; J is the total number of components. Values of SSY_j and $\text{SSY}_{\text{total}}$ are given as follows

$$\text{SSY}_j = (b_j^{\text{PLS}})^2 t_j^{\text{T}} t_j, \tag{3}$$

$$\text{SSY}_{\text{total}} = (b^{\text{PLS}})^2 T^{\text{T}} T, \tag{4}$$

where b_j^{PLS} is the j-th component of the PLS regression coefficient b^{PLS}. More details on VIP scores can be found in Ref. [31]. The PLS regression analysis is performed by JMP software [32,33]. The cut-off line for VIP plot is at 0.8, which is the default cut-off value of the JMP software used in this work [32]. The variables with VIP scores greater than 0.8 can be regarded as important variables for building the PLS regression model, while those with small VIP scores are less significant for the PLS regression model.

2.3. Criteria for model assessment

Two metrics are used to evaluate quality of the fitting, i.e., the coefficient of determination (R^2), and root mean square error ($RMSE$). R^2 is the proportion of the variability in a dataset that is accounted for by the statistical model, and $RMSE$ is a measure of the difference between the values predicted by the model and the actual values. The definitions are given by

$$R^2 = 1 - (\sum_{i=1}^{n} (y_i - y_i')^2 / \sum_{i=1}^{n} (y_i - \bar{y})^2), \tag{5}$$

$$RMSE = [(\sum_{i=1}^{n} (y_i - y_i')^2)/n]^{1/2}, \tag{6}$$

where y_i is the ith calculated/simulated E_a values, y_i' represent the ith predicted E_a values, and $\bar{y}$ is the mean/average E_a values for n samples.

3. Results and discussion

3.1. E_a prediction model building by PLS

Since some of the descriptors have high correlations (Fig. S2 online), we select the PLS regression to construct the E_a prediction model, where the E_a calculated by BVSE method (Table S1 online) is used as the response variable, and the 32 descriptors (Table S2 online) extracted from unit cell information are used as independent variables. To train and verify the PLS model, the original dataset is divided into a training dataset (80%) and a test dataset (20%). Fig. S3 (online) shows PRESS and Q^2 against the number of components in the PLS model. It is found that the minimum PRESS statistic (0.58415) and the maximum Q^2 value (0.65877) are obtained when the number of the component is four; hence the optimum PLS regression model with four components is adopted in the study. Then the model is used to predict the E_a of test dataset that is not involved in the model training. Fig. 3 shows the fit quality of the PLS model by comparing E_a values from BVSE calculation and from PLS prediction for training dataset and test dataset respectively. It can be seen that the prediction abilities are fairly good and there is no overfitting, with R^2 and $RMSE$ values of 0.887 and 0.02 eV for training dataset, and 0.820 and 0.02 eV for test dataset, respectively. Table S1 (online) gives the respective E_a values and corresponding residuals. The residuals are about 0.01 eV for most of the compounds. The highly accurate model for predicting E_a proves the prediction power of HECS-descriptors, that in turn enables us to uncover the inherent link between HECS-descriptors and Li$^+$ conduction property.

3.2. Interpretations of the model and ionic conduction mechanism insights

In order to get the physical insights from the developed model, we analyze the involved descriptors. Fig. 4 shows the resulting VIP scores, which reflect the importance of each descriptor in PLS regression model. Notably, r_{anion} is the significantly important descriptor among them, which is consistent with previous studies [34–38]. However, the anion size cannot determine E_a alone. The total 19 descriptors with VIP scores greater than 0.8, encompassing four compositional descriptors (anion substitution characteristic: r_{anion}, EN_{anion}, PL_{anion}; cation substitution characteristic: EN_{cation}), six structural descriptors (a, v, $V_{\text{[BE4]}}$, $D_{\text{Li-B}}$, $D_{\text{Li-C}}$, $D_{\text{Li-E}}$), two conduction pathway descriptors (BN_{intra}, BN_{inter}), three ion distribution
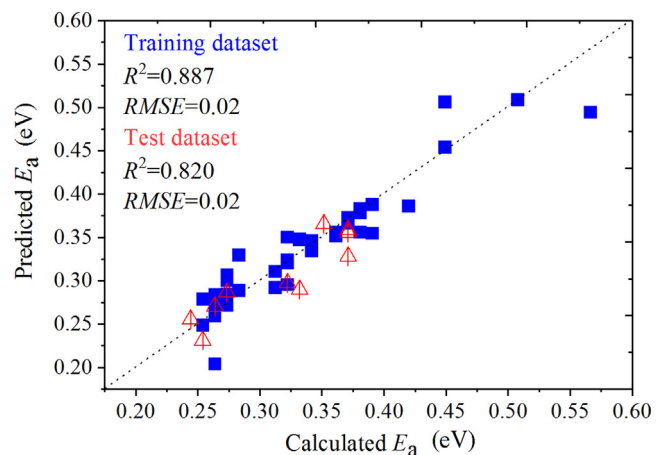
Fig. 3. (Color online) The scatter plot of BVSE calculated E_a versus PLS predicted E_a for training dataset and test dataset, respectively. The dash line is the ideal straight line where the calculated and predicted data are equal.
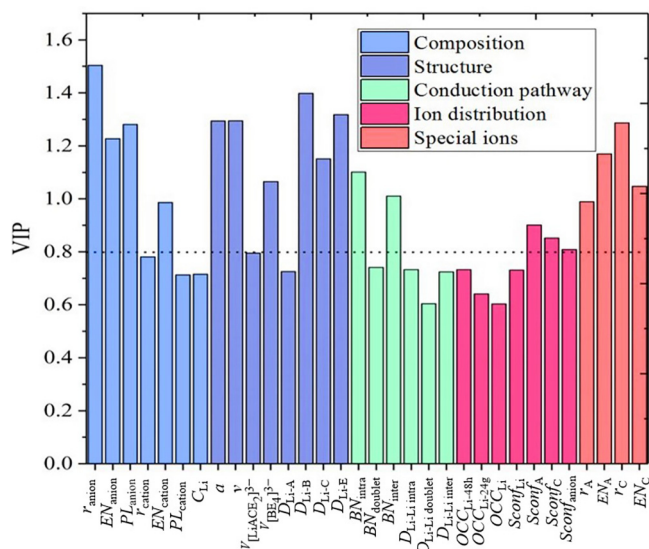
Fig. 4. (Color online) VIP scores of descriptors (explanatory variables) in the PLS model. Five different color areas correspond to five categories. The dotted line at 0.8 represents the default cut-off line of the JMP software used in this work.

descriptors ($Sconf_A$, $Sconf_C$, $Sconf_{anion}$), four special ions descriptors (r_A, EN_A, r_C, EN_C) are all important for the model. Their relationships can be substantiated by the contour plots.

Fig. 5 illustrates some two-dimensional projections of the variation of E_a as a function of selected representative descriptor pairs. Fig. 5a shows the dependence of E_a on the compositional descriptors, r_{anion} and EN_{cation}, both of which offer opportunities for tuning E_a. Within these projections, there are two distinct regions of decreasing or increasing E_a separated by a sharp boundary at about r_{anion} = 1.9 Å. The shape of the E_a surface in the low-energy partition with r_{anion} < 1.9 Å suggests that the substitution of small anions should lead to promising compositions, which results from the competing effects between the lattice space and bottleneck size controlled by framework site disorder. Fig. 5b shows the E_a as a function of the unit cell volume and local polyhedron volume. The data suggest that the lattice expansion on substitution does not lead to lower E_a, which is a counterintuitive phenomenon resulting from the narrower bottleneck size. Fig. 5c shows the E_a as a function of the distances between Li and surrounding ions and the result is consistent with Fig. 5b. Fig. 5d suggests that there is an optimal combination of BN_{inter} and BN_{intra} to achieve lower E_a. Fig. 5e shows the positive effect of larger configurational entropy $Sconf_A$ and $Sconf_C$. Fig. 5f demonstrates that smaller r_A and r_C
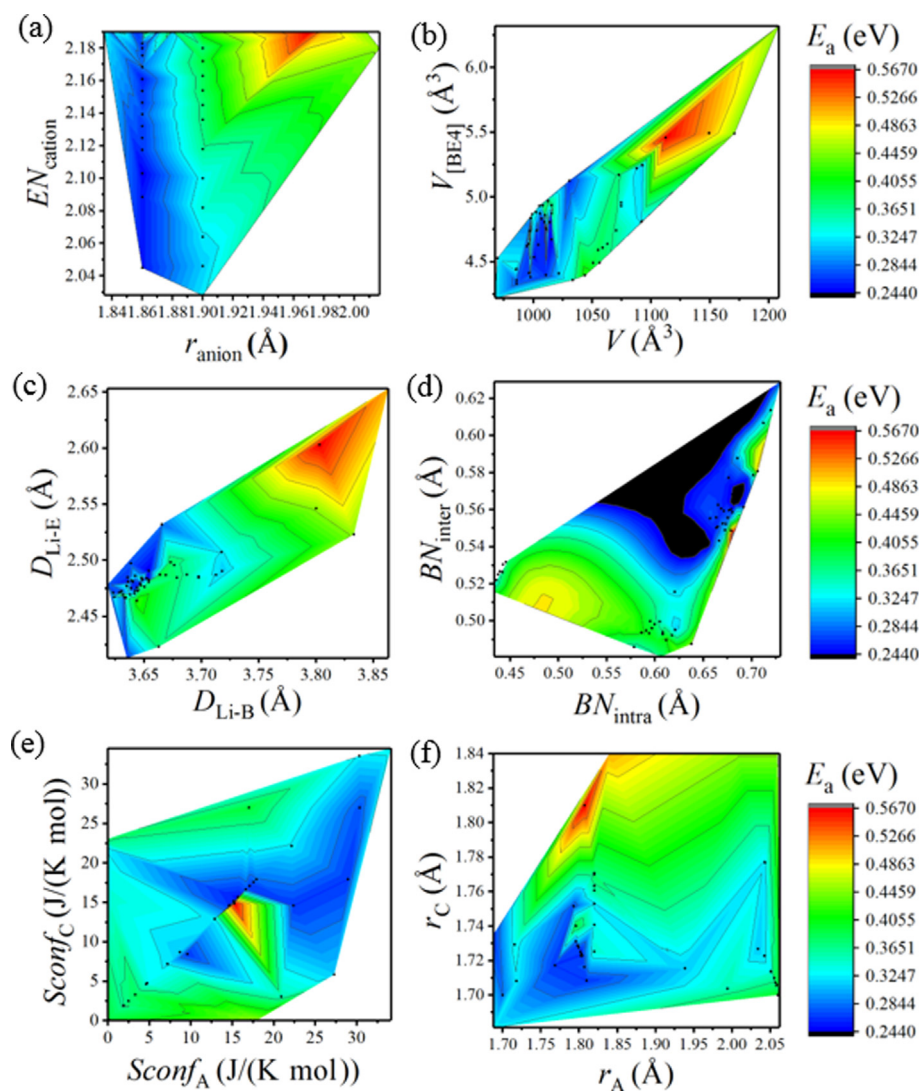
Fig. 5. (Color online) Contour plots showing the variation of E_a as a function of representative descriptors belonging to different categories.
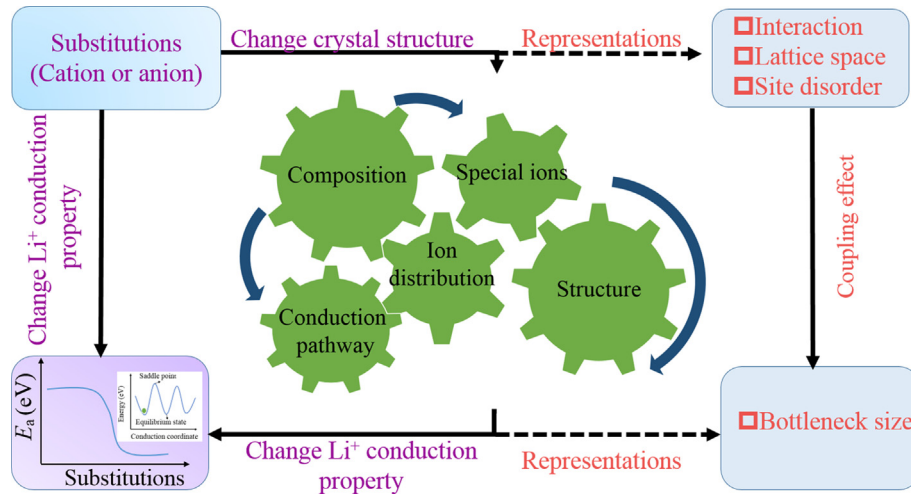
Fig. 6. (Color online) The mechanism of interactions between the HECS-descriptors and their combined effects on E_a.
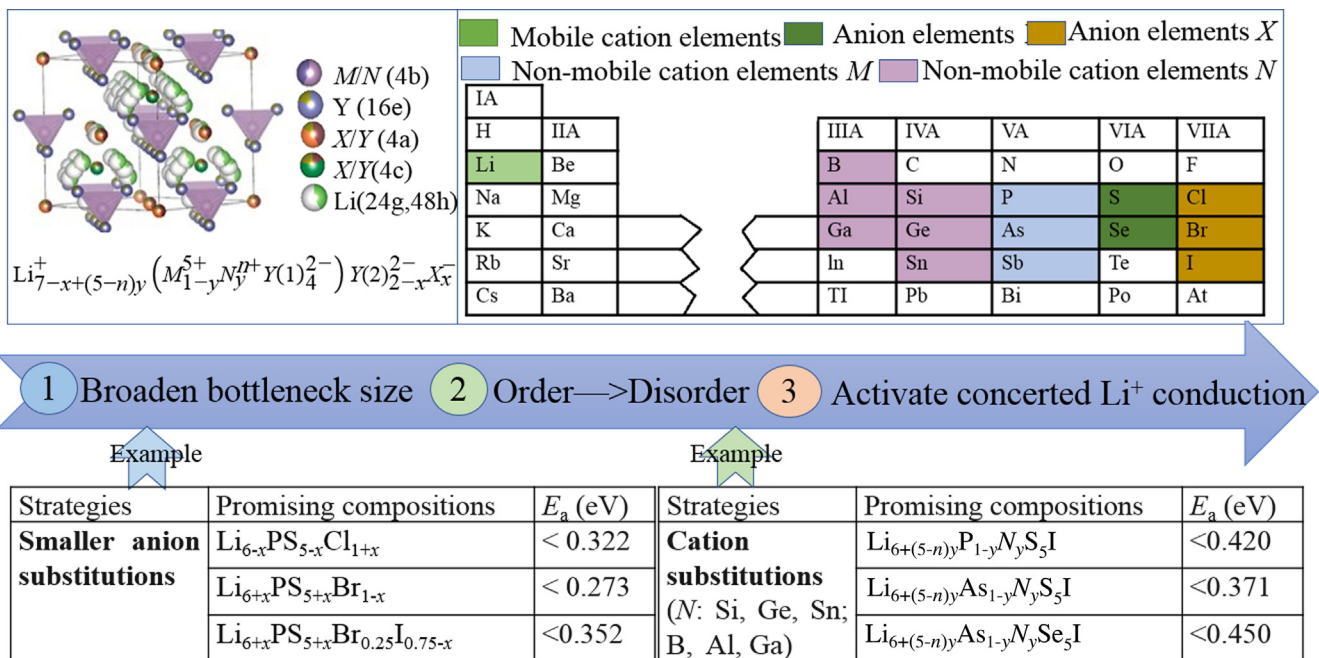

Fig. 7. (Color online) Summary of the design strategies, implementation paths and examples of promising compositions.

correspond to lower E_a. Overall, the E_a change is the result of collective effect of the aforementioned descriptors. This inspires us to get more insights into the Li$^+$ conduction mechanism in SSEs, which is vital for various engineering and industrial applications.

Fig. 6 shows the mechanism of interactions between the HECS-descriptors and their combined effects on E_a. The corresponding processes and causations are as follows:

(1) Cation and anion size. Cation or anion substitutions induce the change of composition and structure. Ions with the certain radii result in the appropriate distances in the structural framework and overall unit cell volume. The corresponding six structural descriptors are a, v, $V_{[BE4]}$, D_{Li-B}, D_{Li-C}, D_{Li-E}.

(2) Cation and anion type and distribution. Li-argyrodite structure type is very flexible in terms of ion mixing. The type of the substituted ions and their distribution over the crystal framework play significant role. The descriptors corresponding to the ion distribution are $Sconf_A$ and $Sconf_C$ and the descriptors corresponding

to the chemical type of the substitutions are r_A, EN_A, r_C, EN_C. In turn, these parameters influence the conduction pathway (BN_{intra}, BN_{inter}) and ultimately affect the E_a. Moreover, the bottleneck of the inter-cage conduction narrows when site disorder is reduced, which then hinders the ionic conduction. Overall, with chemical substitutions at the 4a and 4c sites, the competing effects of lattice space and bottleneck size become crucial for controlling E_a. The action mechanism can explain most of Li$^+$ conduction phenomena in cubic Li-argyrodites.

3.3. Further applications of the model

The developed E_a prediction model enables rational design and optimization of Li-argyrodite SSEs with new compositions, as shown in Fig. 7. For instance, based on Fig. 5a the substitution of small anions should be prioritized for achieving promising compositions, which is the result of the competing effects between the

lattice space and bottleneck size controlled by framework site disorder. Other similar guiding recommendations can be made using the contour plots showing the variation of E_a as a function of r_{anion} and other important descriptors (Fig. S4 online). Additionally, cation substitution can also lead to lower E_a by inducing the local disorder of the anions, which has been demonstrated by previous successful examples [39]. Moreover, despite the fact that lithium content has minor influence on E_a in the regime of uncorrelated ion conduction, increasing the Li$^+$ content to a sufficiently high level can activate concerted ion conduction and thus lead to lowering of the E_a [6,40,41].

Based on these findings, we formulate some promising compositions, as shown in Fig. 7, which can be further verified experimentally, although it should be noted that some substitutions may be challenging due to the limited solid solution ranges and may require special synthesis conditions [42]. At the same time, the solubility itself can be factored into the ML predictive model, which will be the focus of further work.

4. Conclusion

In summary, a framework of HECS-descriptors based E_a prediction by ML is proposed and demonstrated by the example of cubic Li-argyrodites. The excellent performance of the E_a prediction model by PLS (training set: R^2, 88.7%; $RMSE$, 0.02 eV; test set: R^2, 82.0%; $RMSE$, 0.02 eV) allows for a comprehensive mapping of the E_a on HECS-descriptors. The VIP scores show that global and local environment of Li$^+$ conduction, hierarchically characterized as composition, structure, conduction pathway, ion distribution and special ions, jointly affect the E_a of cubic Li-argyrodites. In particular, smaller anion substitutions and the structural changes associated with anion site disorder dictate the Li$^+$ conduction. This analysis provides a fruitful insight into the Li$^+$ conduction mechanism at atomic scale. Based on these findings, we formulate design strategies, implementation paths and propose examples of promising compositions. It is expected that the HECS-descriptors based E_a prediction can be extended to a wide range of SSEs and promote their rational design and optimization.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

Author contributions

Qian Zhao contributed to the conceptualization, methodology, data curation and writing-original draft of the study. Maxim Avdeev contributed to the writing-review & editing. Liquan Chen contributed to the management of the project. Siqi Shi contributed to the conceptualization, writing-review & editing and management of the project.

Appendix A. Supplementary materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.scib.2021.04.029.

References

[1] Gao Z, Sun H, Fu L, et al. Promises, challenges, and recent progress of inorganic solid-state electrolytes for all-solid-state lithium batteries. Adv Mater 2018;30:1705702.
[2] Liang L, Sun X, Zhang J, et al. Sur-/interfacial regulation in all-solid-state rechargeable Li-ion batteries based on inorganic solid-state electrolytes: advances and perspectives. Mater Horiz 2019;6:871–910.
[3] Zhang Q, Cao D, Ma Y, et al. Sulfide-based solid-state electrolytes: synthesis, stability, and potential for all-solid-state batteries. Adv Mater 2019;31:1901131.
[4] Ye T, Li L, Zhang Y. Recent progress in solid electrolytes for energy storage devices. Adv Funct Mater 2020;30:2000077.
[5] Chen R, Qu W, Guo X, et al. The pursuit of solid-state electrolytes for lithium batteries: from comprehensive insight to emerging horizons. Mater Horiz 2016;3:487–516.
[6] He X, Zhu Y, Mo Y. Origin of fast ion diffusion in super-ionic conductors. Nat Commun 2017;8:15893.
[7] Nolan AM, Zhu Y, He X, et al. Computation-accelerated design of materials and interfaces for all-solid-state lithium-ion batteries. Joule 2018;2:2016–46.
[8] Ong SP, Mo Y, Richards WD, et al. Phase stability, electrochemical stability and ionic conductivity of the Li$_{10\pm1}$MP_2X_{12} (M = Ge, Si, Sn, Al or P, and X = O, S or Se) family of superionic conductors. Energy Environ Sci 2013;6:148–56.
[9] Richards WD, Wang Y, Miara LJ, et al. Design of Li$_{1+2x}$Zn$_{1-x}$PS$_4$, a new lithium ion conductor. Energy Environ Sci 2016;9:3272–8.
[10] Wang Y, Richards WD, Ong SP, et al. Design principles for solid-state lithium superionic conductors. Nat Mater 2015;14:1026–31.
[11] Butler KT, Davies DW, Cartwright H, et al. Machine learning for molecular and materials science. Nature 2018;559:547–55.
[12] Wang A, Zou Z, Wang D, et al. Identifying chemical factors affecting reaction kinetics in Li-air battery via *ab initio* calculations and machine learning. Energy Storage Mater 2021;35:595–601.
[13] Liu Y, Wu J, Yang G, et al. Predicting the onset temperature (T_g) of Ge$_x$Se$_{1-x}$ glass transition: a feature selection based two-stage support vector regression method. Sci Bull 2019;64:1195–203.
[14] Liu Y, Zhao T, Ju W, et al. Materials discovery and design using machine learning. J Materiomics 2017;3:159–77.
[15] Schmidt J, Marques MRG, Botti S, et al. Recent advances and applications of machine learning in solid state materials science. Npj Comput Mater 2019;83:1–36.
[16] Shi S, Gao J, Liu Y, et al. Multi-scale computation methods: their applications in lithium-ion battery research and development. Chin Phys B 2016;25:018212.
[17] Liu Y, Guo B, Zou X, et al. Machine learning assisted materials design and discovery for rechargeable batteries. Energy Storage Mater 2020;31:434–50.
[18] Ghiringhelli LM, Vybiral J, Levchenko SV, et al. Big data of materials science: critical role of the descriptor. Phys Rev Lett 2015;114:105503.
[19] Seko A, Hayashi H, Nakayama K, et al. Representation of compounds for machine-learning prediction of physical properties. Phys Rev B 2017;95:144110.
[20] Fujimura K, Seko A, Koyama Y, et al. Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms. Adv Energy Mater 2013;3:980–5.
[21] Jalem R, Aoyama T, Nakayama M, et al. Multivariate method-assisted *ab initio* study of olivine-type LiMXO$_4$ (main group M^{2+}–X^{5+} and M^{3+}–X^{4+}) compositions as potential solid electrolytes. Chem Mater 2012;24:1357–64.
[22] Jalem R, Nakayama M, Kasuga T. An efficient rule-based screening approach for discovering fast lithium ion conductors using density functional theory and artificial neural networks. J Mater Chem A 2014;2:720–34.
[23] Jalem R, Kimura M, Nakayama M, et al. Informatics-aided density functional theory study on the Li ion transport of tavorite-type LiMTO$_4$F (M^{3+}–T^{5+}, M^{2+}–T^{6+}). J Chem Inf Model 2015;55:1158–68.
[24] Kireeva N, Pervov VS. Materials space of solid-state electrolytes: unraveling chemical composition–structure–ionic conductivity relationships in garnet-type metal oxides using cheminformatics virtual screening approaches. Phys Chem Chem Phys 2017;19:20904–18.
[25] Sendek AD, Yang Q, Cubuk ED, et al. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. Energy Environ Sci 2017;10:306–20.
[26] Xu Y, Zong Y, Hippalgaonkar K. Machine learning-assisted cross-domain prediction of ionic conductivity in sodium and lithium-based superionic conductors using facile descriptors. J Phys Commun 2020;4:055015.
[27] Zhang Y, He X, Chen Z, et al. Unsupervised discovery of solid-state lithium ion conductors. Nat Commun 2019;10:5260.
[28] He B, Chi S, Ye A, et al. High-throughput screening platform for solid electrolytes combining hierarchical iontransport prediction algorithms. Sci Data 2020;7:151. (Or go directly to the SPSE web: https://matgen.nscc-gz.cn/solidElectrolyte/).
[29] Tenenhaus M, Vinzi VE, Chatelin Y-M, et al. PLS path modeling. Comput Stat Data An 2005;48:159–205.
[30] Farrés M, Platikanov S, Tsakovski S, et al. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. J Chemometr 2015;29:528–36.
[31] Helland IS. Partial least squares regression and statistical models. Scand J Stat 1990;17:97–114.
[32] Institute S. JMP® 13 Multivariate Methods. Cary, NC: SAS Institute Inc.; 2017.

[33] Cox I, Gaudard M. Discovering Partial Least Squares with JMP®. Cary, North Carolina (USA): SAS Institute Inc.; 2013.

[34] Deiseroth HJ, Kong ST, Eckert H, et al. Li_6PS_5X: a class of crystalline Li-rich solids with an unusually high Li^+ mobility. Angew Chem Int Ed 2008;47:755–8.

[35] Kraft MA, Culver SP, Calderon M, et al. Influence of lattice polarizability on the ionic conductivity in the lithium superionic argyrodites Li_6PS_5X (X = Cl, Br, I). J Am Chem Soc 2017;139:10909–18.

[36] Rao RP, Sharma N, Peterson VK, et al. Formation and conductivity studies of lithium argyrodite solid electrolytes using *in-situ* neutron diffraction. Solid State Ionics 2013;230:72–6.

[37] Wang H, Yu C, Ganapathy S, et al. A lithium argyrodite $Li_6PS_5Cl_{0.5}Br_{0.5}$ electrolyte with improved bulk and interfacial conductivity. J Power Sources 2019;412:29–36.

[38] Wang Z, Shao G. Theoretical design of solid electrolytes with superb ionic conductivity: alloying effect on Li^+ transportation in cubic Li_6PA_5X chalcogenides. J Mater Chem A 2017;5:21846–57.

[39] Kraft MA, Ohno S, Zinkevich T, et al. Inducing high ionic conductivity in the lithium superionic argyrodites $Li_{6+x}PGeSI$ for all-solid-state batteries. J Am Chem Soc 2018;140:16330–9.

[40] Zhou L, Assoud A, Zhang Q, et al. A new family of argyrodite thioantimonate lithium superionic conductors. J Am Chem Soc 2019;141:19002–13.

[41] Zhang Z, Zou Z, Kaup K, et al. Correlated migration invokes higher Na^+-ion conductivity in NaSICON-type solid electrolytes. Adv Energy Mater 2019;9:1902373.

[42] Zhang Z, Sun Y, Duan X, et al. Design and synthesis of room temperature stable Li-argyrodite superionic conductors via cation doping. J Mater Chem A 2019;7:2717–22.

Qian Zhao is a Ph.D. student at Materials Genome Institute of Shanghai University under the supervision of Professors Liquan Chen and Siqi Shi. Her current research interest focuses on the ionic conduction problems related to solid electrolytes for lithium-ion battery and application of machine learning in materials science.

Siqi Shi obtained his B.S. and M.S. degrees from Jiangxi Normal University in 1998 and 2001, respectively. He finished his Ph.D. degree from Institute of Physics, Chinese Academy of Sciences in 2004. After that, he joined National Institute of Advanced Industrial Science and Technology of Japan and Brown University of USA as a postdoctor and then joined Shanghai University as a professor in 2013. His current research interest focuses on the fundamentals and multiscale calculation of electrochemical energy storage materials and application of machine learning in materials science.

Supplemental information

Machine learning prediction of activation energy in cubic Li-argyrodites with hierarchically encoding crystal structure-based (HECS) descriptors

Qian Zhao[a], Maxim Avdeev[b,c], Liquan Chen[a,d], Siqi Shi[a,e*]

[a]*Materials Genome Institute, Shanghai University, Shanghai 200444, China*

[b]*Australian Nuclear Science and Technology Organization, New Illawarra Rd, Lucas Heights, NSW 2234, Australia*

[c]*School of Chemistry, The University of Sydney, Sydney 2006, Australia*

[d]*Key Laboratory for Renewable Energy, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China*

[e]*School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China*

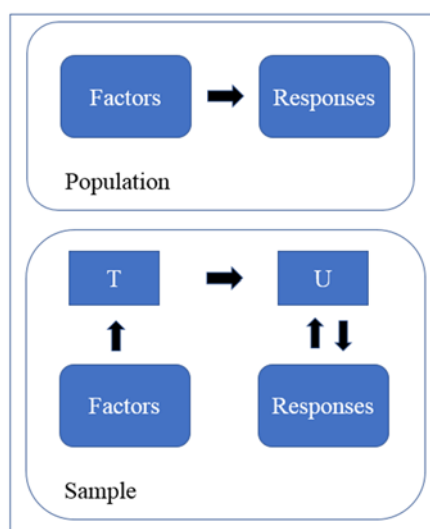[*]*Email: sqshi@shu.edu.cn (Siqi Shi)*

S1 Methods and Details

Fig. S1. A schematic outline of the PLS method. The overall goal is to use the factors to predict the responses in the population. This is achieved indirectly by extracting latent variables T and U from sampled factors and responses, respectively [1].

S2 Data preparation and Results

Table S1. Values of the activation energy (E_a) for 50 cubic Li-argyrodite compounds obtained by BVSE calculations and PLS predictions respectively.

No.	E_a (eV) by BVSE calculation	E_a (eV) by PLS prediction	Residuals
Training dataset			
1	0.5078125	0.508758499	-0.000945999
2	0.2734375	0.30056762	-0.02713012
3	0.263671875	0.204264241	0.059407634
4	0.390625	0.354807146	0.035817854
5	0.263671875	0.27167222	-0.008000345
6	0.380859375	0.37843286	0.002426515
7	0.37109375	0.37321306	-0.00211931
8	0.2734375	0.284675523	-0.011238023
9	0.2734375	0.271633505	0.001803995
10	0.25390625	0.24880127	0.00510498
11	0.283203125	0.329719261	-0.046516136
12	0.419921875	0.386198483	0.033723392
13	0.44921875	0.454282071	-0.005063321
14	0.341796875	0.346080404	-0.004283529
15	0.33203125	0.348385469	-0.016354219
16	0.3125	0.292111488	0.020388512
17	0.37109375	0.366820776	0.004272974
18	0.361328125	0.355685505	0.00564262
19	0.283203125	0.288869363	-0.005666238
20	0.263671875	0.268698096	-0.005026221
21	0.322265625	0.321107289	0.001158336
22	0.263671875	0.284116925	-0.02044505
23	0.44921875	0.506045077	-0.056826327
24	0.322265625	0.350274293	-0.028008668
25	0.361328125	0.351798254	0.009529871
26	0.390625	0.388121793	0.002503207
27	0.33203125	0.34734903	-0.01531778
28	0.341796875	0.334600259	0.007196616
29	0.25390625	0.279046043	-0.025139793
30	0.263671875	0.263454334	0.000217541
31	0.56640625	0.494567833	0.071838417
32	0.322265625	0.295727359	0.026538266
33	0.2734375	0.306696834	-0.033259334

34	0.361328125	0.3530054	0.008322725
35	0.380859375	0.356088449	0.024770926
36	0.380859375	0.382840723	-0.001981348
37	0.322265625	0.323748192	-0.001482567
38	0.3125	0.310952562	0.001547438
39	0.263671875	0.275065821	-0.011393946
40	0.263671875	0.25968542	0.003986455
Test dataset			
1	0.33203	0.289972659	0.042058591
2	0.37109	0.328329524	0.042764226
3	0.37109	0.355668239	0.015425511
4	0.37109	0.358579469	0.012514281
5	0.35156	0.365457605	-0.013895105
6	0.32227	0.297539692	0.024725933
7	0.27344	0.286794422	-0.013356922
8	0.26367	0.270696109	-0.007024234
9	0.24414	0.25558438	-0.011443755
10	0.25391	0.23124641	0.02265984

Table S2. 32 HECS-descriptors with the corresponding descriptions and types for the

ML modeling.

No.	Descriptors	Descriptions	Type
1	r_{anion}	Average ionic radius (r) for the anions	Composition
2	EN_{anion}	Average Pauling electronegativity (EN) for the anions	Composition
3	PL_{anion}	Average ionic polarizability (PL) for the anions	Composition
4	r_{cation}	Average ionic radius (r) for the cations	Composition
5	EN_{cation}	Average Pauling electronegativity (EN) for the cations	Composition
6	PL_{cation}	Average ionic polarizability (PL) for the cations	Composition
7	C_{Li}	Lithium content	Composition
8	a	lattice parameter	Structure
9	v	unit cell volume	Structure

10	$V_{[LiACE_2]^{3-}}$	Volume (V) for Frank−Kasper polyhedral $[LiACE_2]^{3-}$	Structure
11	$V_{[BE_4]^{3-}}$	Volume (V) for tetrahedra $[BE_4]^{3-}$	Structure
12	D_{Li-A}	interatomic distance (D) between Li and nearest neighboring ions on Wyckoff $4a$	Structure
13	D_{Li-B}	interatomic distance (D) between Li and nearest neighboring ions on Wyckoff $4b$	Structure
14	D_{Li-C}	interatomic distance (D) between Li and nearest neighboring ions on Wyckoff $4c$	Structure
15	D_{Li-E}	interatomic distance (D) between Li and nearest neighboring ions on Wyckoff 4d	Structure
16	BN_{intra}	bottleneck size (BN) for the intra-cage pathway	Conduction pathway
17	$BN_{doublet}$	bottleneck size (BN) for the doublet pathway	Conduction pathway
18	BN_{inter}	bottleneck size (BN) for the inter-cage pathway	Conduction pathway
19	$D_{Li-Li\ intra}$	interatomic distance (D) between Li and nearest neighboring Li on the intra-cage pathway	Conduction pathway
20	$D_{Li-Li\ doublet}$	interatomic distance (D) between Li and nearest neighboring Li on the doublet pathway	Conduction pathway
21	$D_{Li-Li\ inter}$	interatomic distance (D) between Li and nearest neighboring Li on the inter-cage pathway	Conduction pathway
22	OCC_{Li-48h}	Lithium ion occupancy on Wyckoff $48h$ site,	Ion distribution
23	OCC_{Li-24g}	Lithium ion occupancy on Wyckoff $48h$ site, Wyckoff $24g$ site	Ion distribution

|			
24	OCC_{Li}	Lithium ion occupancy on all site	Ion distribution
25	$Sconf_{Li}$	configurational entropy ($Sconf$) for the mobile Lithium ions	Ion distribution
26	$Sconf_A$	configurational entropy ($Sconf$) for the ions on Wyckoff $4a$ site	Ion distribution
27	$Sconf_C$	configurational entropy ($Sconf$) for the ions on Wyckoff $4c$ or $4d$ site	Ion distribution
28	$Sconf_{anion}$	configurational entropy ($Sconf$) for the anions	Ion distribution
29	r_A	Average ionic radius (r) for the ions on Wyckoff $4a$ site	Special ions
30	EN_A	Average Pauling electronegativity (EN) for the ions on Wyckoff $4a$ site	Special ions
31	r_C	Average ionic radius (r) for the ions on Wyckoff $4c$ or $4d$ site	Special ions
32	EN_C	Average Pauling electronegativity (EN) for the ions on Wyckoff $4c$ or $4d$ site	Special ions
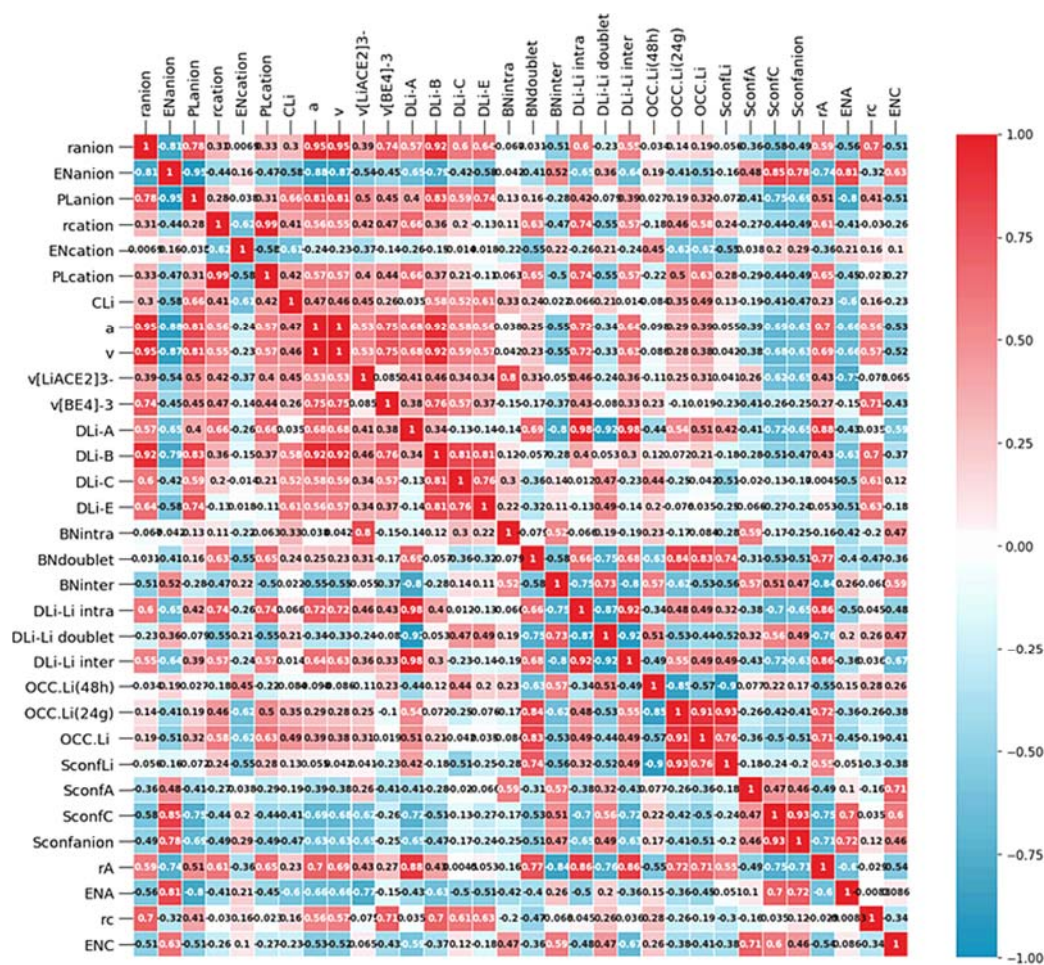
Fig. S2. Pearson correlation coefficient heatmap of descriptors.
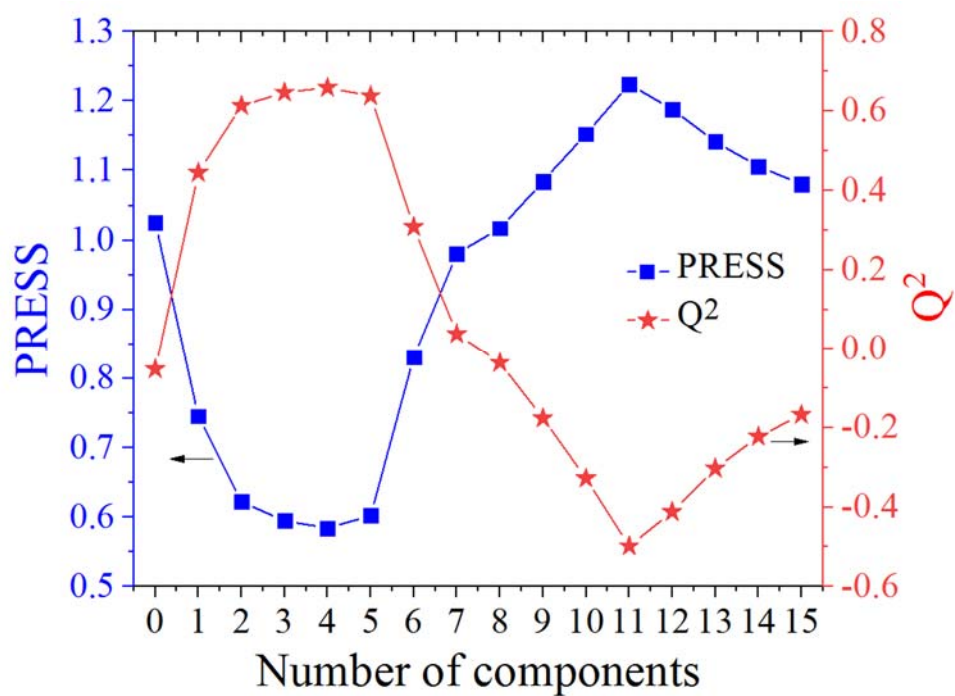
Fig. S3. Plot of PRESS and Q^2 against the number of components in the PLS regression model. The minimum PRESS value (0.58415) and the maximum Q^2 value (0.65877) occur at four components.
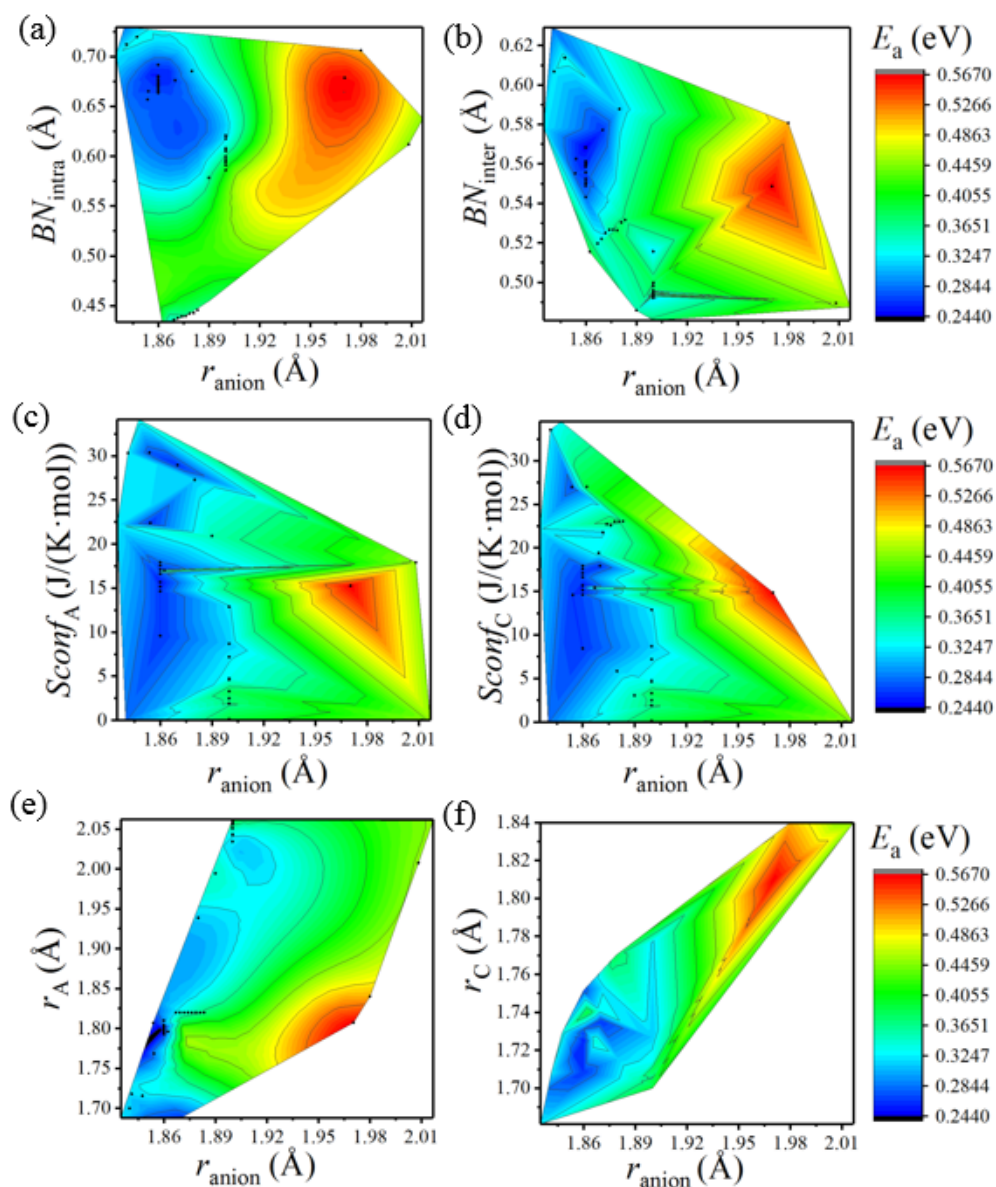
Fig. S4. Contour plots showing the variation of E_a as a function of r_{anion} and other important descriptors.

[1] R.D. Tobias, SAS Institute Inc., Cary, NC, 1995.