J Mater Sci

Computation & theory

Transferable and robust machine learning model for predicting stability of Si anodes for multivalent cation batteries

Joy Datta[1], Dibakar Datta[1], and Vidushi Sharma[1,*]

[1]Department of Mechanical and Industrial Engineering, New Jersey Institute of Technology, Newark, NJ 07103, USA

Received: 12 January 2023
Accepted: 17 June 2023

ABSTRACT

Data-driven methodology has become a key tool in computationally predicting material properties. Currently, these techniques are priced high due to computational requirements for generating sufficient training data for high-precision machine learning models. In this study, we present a support vector regression (SVR)-based machine learning model to predict the stability of silicon (Si)–alkaline metal alloys, with a strong emphasis on the transferability of the model to new silicon alloys with different electronic configurations and structures. We elaborate on the role of the structural descriptor in imparting transferability to the model that is trained on limited data (* 750 Si alloys) derived from the Material Project database. Three popular descriptors, namely X-ray diffraction (XRD), sine coulomb matrix (SCM), and orbital field matrix (OFM), are evaluated for representing Si alloys. The material structures are represented by descriptors in the SVR model, coupled with hyperparameter tuning techniques like Grid Search CV and Bayesian optimization, to find the best performing model for predicting total energy, formation energy and packing fraction of the Si alloy systems. The models are trained on Si alloys with lithium (Li), sodium (Na), potassium (K), magnesium (Mg), calcium (Ca), and aluminum (Al) metals, where Si–Na and Si–Al systems are used as test structures. Our results show that XRD, an experimentally derived characterization of structures, performs most reliably as a descriptor for total energy prediction of new Si alloys. The study demonstrates that by qualitatively selection of training data, using hyperparameter tuning methods, and employing appropriate structural descriptors, the data requirements for robust and accurate ML models can be reduced.

Handling Editor: Ghanshyam Pilania.

Address correspondence to E-mail: vs574@njit.edu; vidushis@ibm.com

https://doi.org/10.1007/s10853-023-08705-y

Published online: 01 July 2023

Introduction

Increasing demand for electric vehicles (EVs) has highlighted the energy storage limitations of commercial graphite anode-based lithium-ion batteries (LIBs). Energy storage in graphite with an intercalation mechanism offers low gravimetric energy densities of 372 mAh g^{1} [1]. Alternatively, energy storage in electrodes through a conversion mechanism can promise a tenfold improvement in energy densities. The most popular anode after graphite is Silicon (Si), which has a gravimetric energy density of 3572 mAh g^{1} [2]. Si reacts with incoming Li to form an alloying mixture of Li_xSi during battery charging [3]. Next, there is a pressing issue surrounding the scarcity of Li for LIBs. To meet the requirements of the future EV industry, we cannot rely solely on non-renewable Li [4]. Therefore, active research efforts are being made to develop advanced battery technologies beyond Li ion [5]. Due to the high capacity offered by Si–Li anode, Si has found applications in alkali earth metal batteries such as sodium (Na) ion batteries [6], magnesium (Mg) ion batteries [7], and calcium(Ca) ion batteries [8], to name a few. Similar to Si–Li system, Si anode reacts with Mg to form Mg_2Si phase with a gravimetric density of 3816 mAh g^{1} [9] and reacts with Ca to form Ca_2Si alloys with a maximum theoretical capacity of 3818 mAh g^{1} [8]. However, Si anode face challenges related to structural stability and volume expansion (* 300% for LIBs) that lead to premature fractures, capacity losses, and limited cycle life of batteries [10–12]. Therefore, before designing and experimenting such battery materials, it is imperative to study the stability and structural assessment of Si-metal anodes computationally.

Over the last two decades, numerous computational efforts have been dedicated to understanding the Si–Li microstructures in the alloy mixtures and structural integrity of Si-based anodes [13–16]. Fan et al. [17] studied the effect of increasing Li concentration on the electro-mechanical stability of amorphous Li_xSi anode by molecular dynamics (MD) simulations. They reported atomic bonding transitions from covalent to metallic bonds of a-Li_xSi under various loading conditions. The study details how mechanical property of a-Li_xSi changes during lithiation under different loading conditions. Similar atomistic simulation studies have also been conducted on Si anodes in alkali ion batteries [18, 19]. It is evident that Si anodes possess greater potential than their competitor anodes for future multivalent cation batteries and will remain a subject of extensive research in the years to come [8, 20].

Considering that the stability of Si-based anodes during electrochemical cycling is a primary concern for future batteries, it is necessary to study how Si anodes can maintain both low volume expansion and high capacity simultaneously. Exploring these possibilities of Si alloy anodes can be achieved by varying the compositions and stoichiometry ratio of Si and alloying multivalent cations. Prior to conducting experiments, it is essential to assess the synthesizability of any unknown structure based on its stability. Density functional theory (DFT) can be employed to predict the structural stability of materials using energy hull diagrams. This method involves calculating the formation energy of all possible stoichiometric ratios for a given composition using DFT. The structures with the lowest formation energy are considered the most stable, and an energy hull diagram can be created to illustrate the relative stability of different compositions. This approach has been successfully demonstrated in various studies [21, 22]. Another important characteristic of a structure is the packing fraction, which describes the degree of porosity within the material. This parameter is expressed as a dimensionless quantity and represents the ratio of the total volume occupied by the crystal atoms to the volume of the unit cell. A low packing fraction indicates highest stability of an unknown material [23].

Theoretical studies based on simulations have proven to be valuable in providing design insights and performance predictions for experimental design [24]. Lately, the efficiency of computational simulations has diminished with the increasing complexity of materials. Quantum mechanics-based DFT methods are limited to small atomic systems consisting of approximately 200 atoms or fewer due to their computational expense [25]. On the other hand, classical Newtonian simulations like MD require less computational power and can be a viable alternative to simulating larger atomic systems. However, these techniques significantly abate the thermodynamic accuracy of disordered structures [26]. Furthermore, there is a lack of available interatomic potentials in the literature for newer material combinations [27].

Machine learning (ML) techniques have been widely adopted in material modeling to predict energy and simulate materials more efficiently [28–31]. ML is actively being researched to tackle various material science challenges, including property prediction, material discovery, and system optimization [32–34]. Common ML models used in this field include regression [35, 36], Gaussian approximation [37, 38], graph neural networks (GNN) [28], and high-dimensional neural networks (HDNN) [39]. ML approaches in material systems can be categorized into graph-based and descriptor-based methods [40]. Graph-based models may not be well-suited to handle long-range inter-atomic interactions in condensed solid state systems such as Si alloys. On the other hand, the descriptor-based approach offers more flexibility for feature engineering, allowing researchers to incorporate a broader range of atomic interactions in both spatial and dimensional aspects. Jihang et al. [41] predicted molecular property by four descriptor-based ML models and four graph-based models. The results revealed that the graph-based models required higher computational costs and resources compared to the descriptor-based models.

The selection of appropriate descriptors is crucial and should be based on the specific task at hand including their compatibility with the targeted machine learning algorithm [42–44]. Descriptors are numeric vectors that characterize the atomic or molecular structure and serve as inputs to ML models [40]. Various descriptors have been proposed for materials, including the coulomb matrix (CM) [45], sine coulomb matrix (SCM) [46], atom-centered symmetry functions (ACSF) [47], smooth overlap of atomic orbitals (SOAP) [48], and orbital field matrix (OFM) [49], among others. Each descriptor has been designed to meet material science field requirements but may not have broad applicability across different domains. By combining a well-defined descriptor and an appropriate model, ML techniques can leverage the wealth of experimental and simulation data available in established databases such as Material Project Database (MPD) [50], Open Quantum Materials Database [51], AFLOW [52] and Inorganic Crystal Structure Database (ICSD) [53], to address some of the most pressing predictive and discovery challenges among materials.

In this study, we predict the stability and packing fraction (PF) of Si alloys using the Support Vector Regression (SVR) [54]-based ML model. SVR is a supervised learning approach well-suited for handling nonlinear regression problems [55]. It serves as an efficient alternative to more advanced methods like HDNN, which require extensive computational resources and data for training the model [56–58]. By coupling SVR with advanced statistical techniques, solid-state material properties can be predicted with high accuracy and minimal data requirements [59]. Thus, SVR is an ideal approach for rapidly predicting the stability of materials such as Si alloys, which have broad applications in the energy domain [60–64] but limited available data. We utilize the Si-based data from the MPD to train the SVM model for predicting the total energy per atom, formation energy per atom and PF of the Si-based alloy structures. The focus lies in selecting suitable structural descriptors that enable the model to achieve the best transferability, high prediction accuracies, and least dependence on data quantity. Three common descriptors are employed to convert atomic structures into ML inputs, and trained model's performance is evaluated on completely new Si alloys. We demonstrate that the SVR model exhibits high predictive accuracy for Si alloys compared to more advanced approaches that require extensive data generation through ab initio simulations for training [65, 66].

Methodology

Dataset preparation

The dataset used in this study consists of 745 inorganic structures comprising Si and A_xSi_y alloys, where A represents elements such as Li, Na, K, Mg, Ca, and Al. The stoichiometric ratios of A and Si are denoted by x and y, respectively. The dataset includes associated properties such as total energy per atom, packing fraction, and formation energy per atom. In MPD, researchers widely utilize DFT calculations based on the ground state energy to determine the total energy of compounds [67]. They evaluate the thermodynamic stability of a compound by considering Gibbs Free Energy (DG) while utilizing change in enthalpy (DH) as a practical approximation [68]. This simplification effectively equates DG and DH to the change in internal energy (DU), enabling the use of total internal energy as an approximation for thermodynamic stability at

absolute zero temperature (0 K) [68]. A compound's formation energy (DH_f) is determined by considering the ground state energies of its constituent elements and the combination itself [69]. It quantifies the difference between the total enthalpy of the compound and the sum of the enthalpies of the constituent elements, considering their stoichiometric fractions [68]. Notably, cohesive energy and formation energy share a relationship, where the cohesive energy represents energy release and the formation energy represents energy consumption, with opposite signs to reflect bond formation or breakup [69]. To retrieve the dataset, we utilized the MPD and accessed it through the application programming interface (API) using Matminer [70] and Python Materials Genomics (pymatgen) [71] libraries in python. Python dataset extraction implementation available on GitHub (https://github.com/joy1303125/Si-based-Material-stability-prediction/tree/main).

Figure 1 shows a sample structure from each A_xSi_y alloy, representing the Si alloys with different elements A. In the dataset, all the metals (A) are either monovalent or bivalent, except for the case of Al. We used this dataset to validate the transferability of the ML model to multivalent A_xSi_y alloys. Specifically, 12 Na_xSi_y and 3 Al_xSi_y structures are considered as the test dataset, while the remaining dataset as our training dataset. Additional information regarding the test structures can be found in Table S7.

Descriptors

The key to any ML model's success is the rightful representation of the atomic structures. The choice of the descriptors is sensitive to the learning labels and the model's paradigm. Traditionally, the process of selecting suitable descriptors involves a trial-and-error [44] approach. In this study, we employ multiple descriptors to represent atomic structures and evaluate their performance in predicting the stability metrics of the structures. The three descriptors compared in the study are X-ray diffraction pattern (XRD), sine coulomb matrix (SCM), and orbital field matrix (OFM). The primary characteristics of each of these descriptors have been detailed in supplementary section.

Support vector regression

SVR is a regression model that utilizes a statistical learning approach to forecast continuous values. It fits a hyperplane to the data in n-dimensional space and employs Vapnik's insensitive region approach to create a generalized model with high prediction

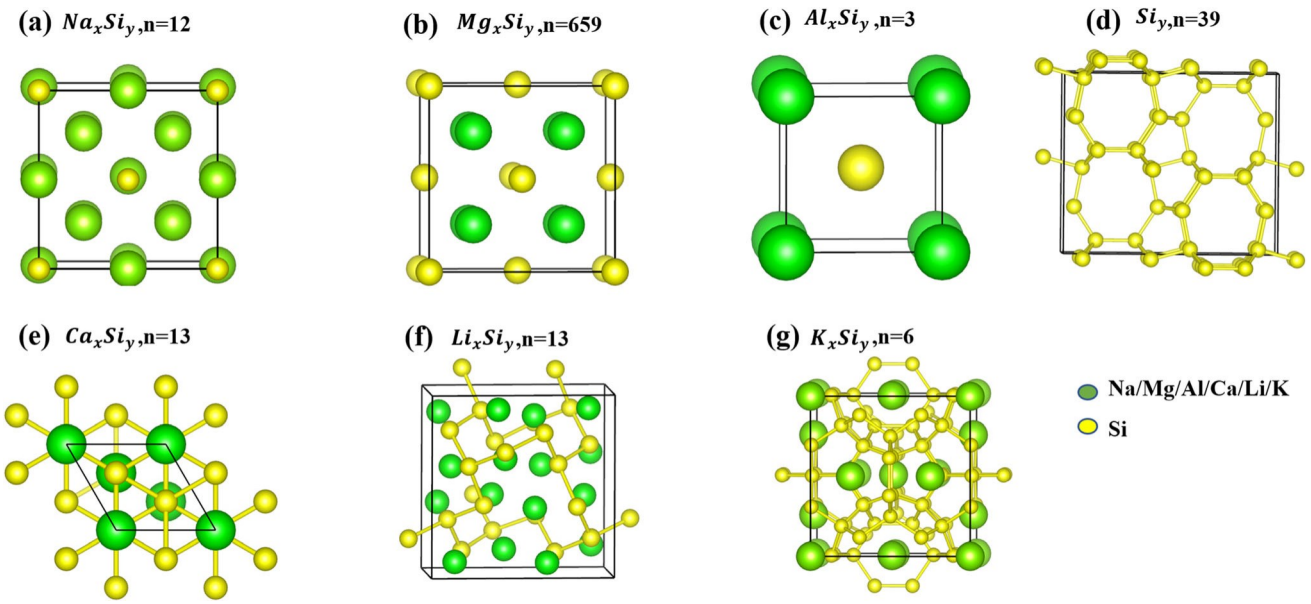

Figure 1 Representative atomic structures from the dataset consisting of inorganic crystal structures of Si and A_xSi_y alloys, where A = Li, Na, K, Mg, Ca and Al, x and y represent stoichiometric ratio of A and Si. List of Si-based metal anode structures are represented as a Na_3Si, b Mg_2Si, c AlSi, d Si, e $CaSi_2$, f LiSi, g K_4Si_{23} where n is the total number of sample structures present for each alloy category.

ability [72]. SVR has several hyperparameters to choose, such as the tube width (epsilon), suitable kernels (linear, polynomial, or radial basis function), the regularization parameter C, and gamma. The appropriate selection of these parameters requires applying additional hyperparameter tuning techniques. Two different types of hyperparameter tuning techniques are Grid Search CV and Bayes Search CV. Both techniques can be implemented in Python using the scikit-learn library [73]. The model can be evaluated using the repeated K-fold cross-validation technique to avoid overfitting. In this work, SVR algorithm has been implemented on python library called scikit-learn [73]. The training data is divided into a fivefold cross-validation dataset, where onefold is considered as a validation dataset, and the rest of the fold is going onto training the model. Each fold result is repeated 10 times to keep away from the noise in the predictions. Two different optimization techniques, Grid Search CV and Bayes Search CV, have been used for the best hyperparameters of our SVR model. Details of these two hyperparameters are detailed in supplementary section (see section 2.1 and 2.2).

Results and discussion

In this section, we discuss our model's exploratory data analysis and performance for the test dataset of Si alloys described in Section "Dataset preparation." Prediction results of system total energy and packing fraction (PF) for validation and test data are compared using root mean square error (RMSE) value. In addition, model's prediction ability is tested for 3 different types of structural descriptor detailed in supplementary section 1 (see Supporting Information section 1.1–1.3).

Exploratory data analysis

We analyzed the data distribution before fitting the model to interpret trends in mean, variance, frequency, and outliers. For interpreting the complete data, violin plots are used that come within seaborn python library [74]. The violin plot displays the inner interquartile range as a thick black box and the median value as a white dot. In Fig. 2, violin plots depict the data pattern of output labels in the A_xSi_y dataset, namely total energy/atom, PF, and formation energy/atom. From Fig. 2a, b, it is evident that K_xSi_y and Na_xSi_y structures exhibit the highest variance for both output labels, total energy/atom and PF, compared to the remaining data (Li_xSi_y, Mg_xSi_y, Ca_xSi_y; Al_xSi_y and Si_y). The protruding plots beyond the interquartile range for Na_xSi_y and M g_xSi_y structures in Fig. 2a, b suggest the presence of outliers in the data. In Fig. 2c, for the formation energy per atom data, we observe outliers in Mg_xSi_y structures compared to the rest of the dataset. We test all our results by removing outliers followed by the modified Z score method [75] (shown in Tables S4–S6). In regards to the improvement in the performance of the trained model, the cross-validation dataset exhibits better outcomes compared to the original data points. However, when evaluating the test dataset, we observe a degradation in performance (as depicted in Tables 1, 2 and 3, S4–S6).

Model performance

To access the prediction capability of the SVR model, the model is fitted on training data as described in Section "Dataset preparation." Each training data is further sectioned into training and validation datasets based on the fivefold cross-validation method, which separates 20% of the total training datapoints for the model validation during the training. The RMSE between the predicted and actual output values is employed to measure the model's accuracy. The model performance of the two trained models is evaluated by predicting the total energy/atom, formation energy/atom and PF values for Na_xSi_y and Al_xSi_y structures, respectively. The performance of trained SVR models is detailed in Sections "Test dataset study on Na_xSi_y and Al_xSi_y structures for total energy and PF prediction" and "Test dataset study on Na_xSi_y and Al_xSi_y structures for formation energy/atom prediction."

Test dataset study on Na_xSi_y and Al_xSi_y structures for total energy and PF prediction

In the first experiment, SVR is trained on 730 structures of A_xSi_y alloys, where A = Li, K, Mg, and Ca. Hyperparameter tuning for SVR is performed using both Grid Search CV and Bayes Search CV, resulting in different sets of hyperparameter values, which are tabulated in Tables 1 and 2. For each of these

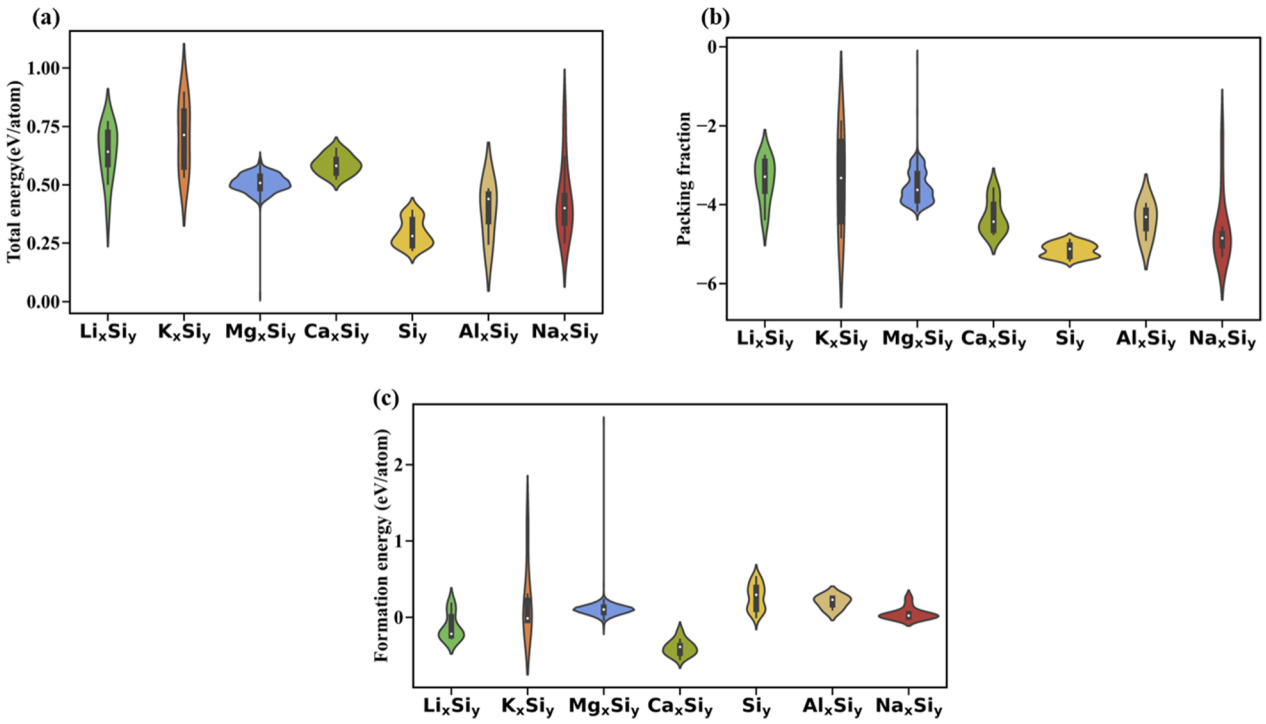

Figure 2 Data distribution of Si and A_xSi_y alloy structures based on a Total energy/atom (eV/atom) b Packing fraction c Formation energy/atom. Black bar represents the interquartile range, white dot shows the median value and data that falls outside the interquartile range is considered as outliers.

Table 1 SVR hyperparameters, descriptor used and associated results for Na_xSi_y and Al_xSi_y test dataset for total energy/atom prediction

Descriptor name	Hyperparameter tuning technique name	C	Gamma	Epsilon	Train RMSE (eV/atom)	Test RMSE (eV/atom)	Validation RMSE (eV/atom)
XRD	Grid search CV	51	0.0001	5.00E-05	0.17	0.28	0.23
	Bayes search cv	600	1.38E-06	1.56E-0.6	0.55	0.55	0.73
Sine	Grid search CV	21	0.05	0.1	0.1	1.12	0.21
	Bayes search cv	4.98	0.15	0.02	0.1	1.16	0.21
OFM	Grid search CV	11	0.01	0.0005	0.13	0.57	0.16
	Bayes search cv	276.76	0.009	0.008	0.08	0.65	0.15

Table 2 SVR hyperparameters, descriptor used and associated results for Na_xSi_y and Al_xSi_y test dataset for PF prediction

Descriptor name	Hyperparameter tuning technique name	C	Gamma	Epsilon	Train RMSE	Test RMSE	Validation RMSE
XRD	Grid search CV	1	0.1	0.005	0.004	0.1	0.04
	Bayes search cv	600	2.06E-05	3.00E-06	0.01	0.07	0.03
Sine	Grid search CV	1	0.1	0.005	0.02	0.2	0.04
	Bayes search cv	1.96	0.08	0.02	0.02	0.2	0.04
OFM	Grid search CV	1	0.01	5.00E-05	0.022	0.11	0.03
	Bayes search cv	0.43	7.00E-03	2.00E-04	0.02	0.11	0.03

Table 3 SVR hyperparameters, descriptor used and associated results for Na_xSi_y and Al_xSi_y test dataset for formation energy/atom prediction

Descriptor name	Hyperparameter tuning technique name	C	Gamma	Epsilon	Train RMSE (eV/atom)	Test RMSE (eV/atom)	Validation RMSE (eV/atom)
XRD	Grid search CV	1	0.01	0.05	0.06	0.09	0.13
	Bayes search cv	103.21	9.00E-04	0.07	0.04	0.08	0.12
Sine	Grid search CV	1	0.01	0.05	0.12	0.17	0.13
	Bayes search cv	1.73	5.00E-03	1.72E-05	0.12	0.17	0.13
OFM	Grid search CV	201	0.01	0.005	0.07	0.2	0.12
	Bayes search cv	5.64	0.5	0.007	0.06	0.09	0.12

hyperparameter sets, three SVR models are trained with different structural input descriptors, as detailed in Section "Descriptors." In total, six SVR models are trained, incorporating various combinations of hyperparameters and structural descriptor methods. The trained models are used to predict the total energy/atom and PF of 12 Na_xSi_y and 3 Al_xSi_y structures as test datapoints. Figure 3 describes the results obtained during the validation and testing of 6 models. The obtained RMSE value for total energy/atom prediction and PF prediction is plotted for validation (see Fig. 3a, b) and test dataset (see Fig. 3c, d).

From the histograms in Fig. 3a, b, it is visible that RMSE values for validation using Grid search and Bayes Search are nearly identical, except for the total energy/atom prediction using XRD descriptor in Fig. 3a, which shows validation results for total

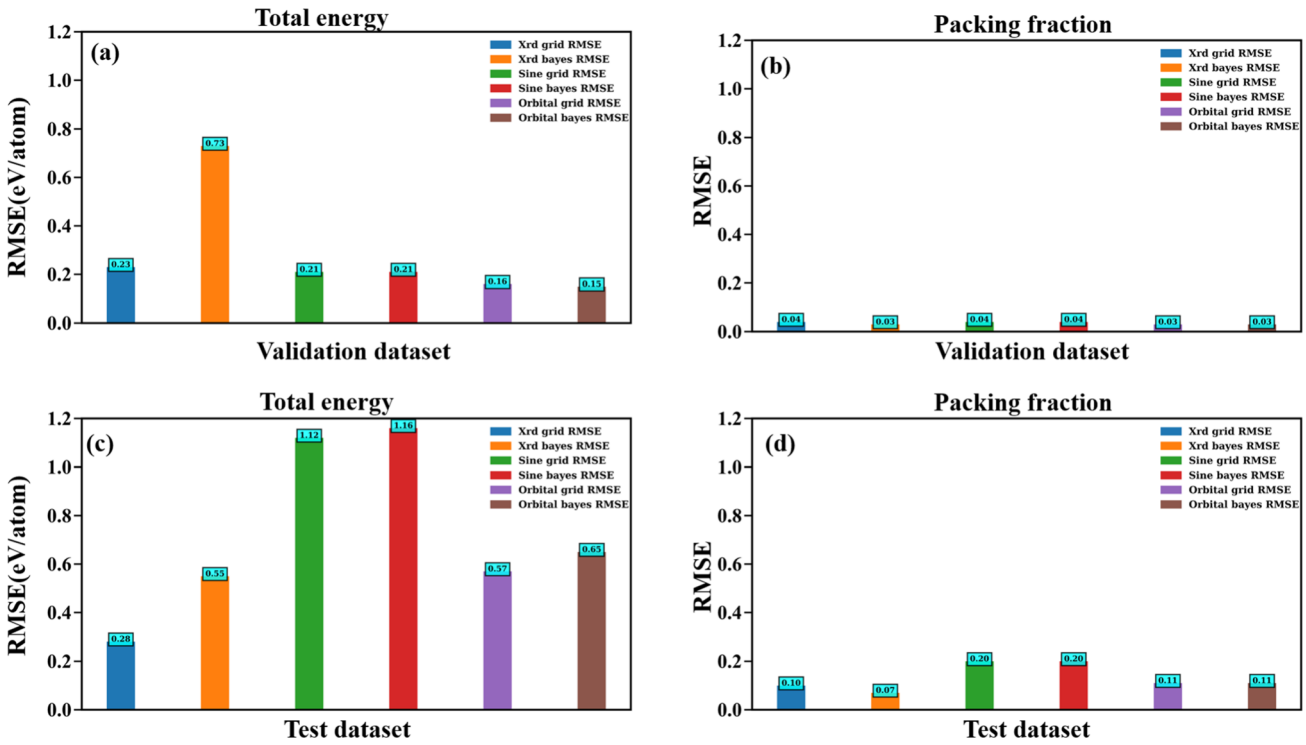

Figure 3 Performance comparison of hyperparameter tuning methods Grid Search CV and Bayes Search CV. a RMSE value of total energy/atom for validation dataset, b RMSE value of packing fraction for validation dataset, c RMSE value of total energy/atom for the test dataset (Na_xSi_y and Al_xSi_yÞ, d RMSE value of packing fraction for test dataset (Na_xSi_y and Al_xSi_yÞ. Three different structural descriptors with two different optimizations techniques generated 6 set of results labeled on top right side.

energy per atom prediction. Among the different descriptor methods, the highest RMSE of 0.73 eV/atom was obtained when using XRD descriptors with Bayes Search CV parameters. On the other hand, the lowest RMSE of 0.15 eV/atom was achieved when employing OFM descriptors with Bayes Search CV parameters. Regarding the PF prediction in validation dataset, the lowest RMSE of 0.03 was obtained when using both XRD and OFM descriptors. Conversely, the highest RMSE of 0.03 was observed with Sine descriptors (Fig. 3b).

In test datapoints, the XRD-Grid search-based model showed the lowest RMSE of 0.28 eV/atom for total energy per atom prediction (Fig. 3c). The highest RMSE of 1.16 eV/atom was observed with Sine descriptors. Regarding PF prediction in test datapoints, the XRD-Bayes model achieved the lowest RMSE of 0.07, while the highest RMSE of 0.20 was obtained with Sine descriptors (Fig. 3d). These results emphasize the importance of selecting appropriate descriptors and optimization strategies for accurate predictions of total energy per atom and PF.

A comprehensive comparison of the performance of hyperparameter tuning methods and structural descriptors for total energy/atom and PF is presented in Tables 1 and 2, respectively. Additionally, Fig. 1b illustrates that the datasets exhibit a high degree of skewness on Mg_xSi_y dataset. Consequently, 371 data points have been excluded by considering energy above hull greater than 20 eV/atom, as they fall outside the metastable material range [76]. However, the performance in terms of the test and validation datasets has decreased for all cases, as indicated in Tables S1 and S2. Therefore, we have decided to utilize the unfiltered dataset for our test and validation cases when predicting total energy per atom and PF.

Figure 4 illustrates the predicted total energy/atom and PF for test Na_xSi_y and Al_xSi_y structures, utilizing the SVR models based on Grid Search CV and Bayes Search CV. For comparison, the actual values of total energy/atom and packing fraction are depicted as red scatter plots in Fig. 4a–f. The total energy predictive supremacy of the XRD descriptor is evident in Fig. 4a. Regarding the packing fraction prediction, Fig. 4d demonstrates that the XRD-Bayes model outperforms the other descriptors (Fig. 4e, f). The RMSE for PF of test Na_xSi_y and Al_xSi_y structures ranges from 0.07 to 0.11 for all XRD and OFM models (Fig. 3d), as noted in Table 2. Hence, the findings from Fig. 4d–f and Table 1 provide compelling evidence of the superior performance of XRD and OFM descriptors over SCM descriptors in predicting both total energy per atom and PF.

Although the implementation of Bayes Search CV and Grid Search CV with SVR yields similar energy and structural predictions, we emphasize that Bayes Search CV is faster than the Grid Search CV. Grid Search CV iterates over complete permutated combinations of hyperparameters, which takes 144 h on 32 cores. In contrast, Bayes search CV completes the search in just 0.08 h on the same computational facility. Therefore, we can conclude that Bayes Search CV is nearly 1800 times more efficient than Grid Search CV based on the total execution time of the two approaches.

Test dataset study on Na_xSi_y and Al_xSi_y structures for formation energy/atom prediction

The histograms in Fig. 5a, b depict the RMSE values for both the validation and test datasets using Grid search and Bayes search, with the exception of the grid search CV parameter-based OFM test dataset case (see Fig. 5b). Figure 5a demonstrates that the best RMSE of 0.12 eV/atom is achieved when utilizing both XRD and OFM descriptors for the validation dataset. On the other hand, Fig. 5b reveals that the XRD-Bayes model achieves the lowest RMSE value of 0.08 eV/atom, while Sine descriptor performs poorly, yielding an RMSE value of 0.17 eV/atom. Table 3 provides a comprehensive comparison of the performance of hyperparameter tuning methods and structural descriptors for formation energy/atom.

Figure 6a, b, c further supports the evaluation of which descriptor correctly predicts the actual value of formation energy/atom. The errors between predicted values and actual values are lowest for formation energy/atom when structures are described by XRD-Bayes, as shown in Fig. 6a. Conversely, Fig. 6b illustrates the poor performance of the SCM descriptor in predicting actual formation energy/atom.

Due to heavy skewness toward Mg_xSi_y structure in the dataset (Fig. 1b), we excluded 371 data points with energy above hull [20 eV/atom [76]. All predictions for formation energy/atom were carried out using Bayes search. The performance on the validation dataset exhibited a consistent RMSE. However,

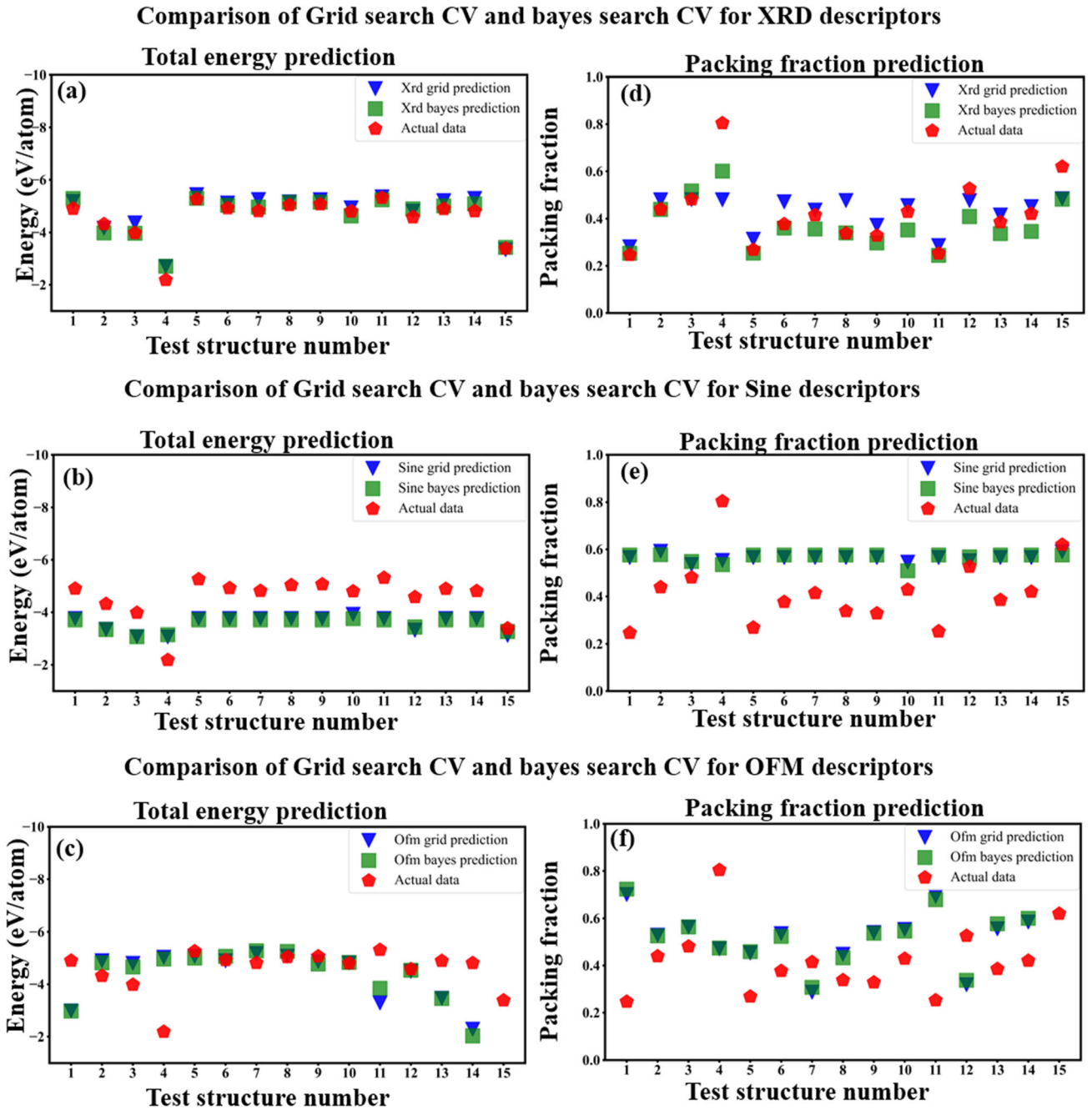

Figure 4 Comparing predictions using three structural descriptors with different hyperparameter search approaches for Na_xSi_y and Al_xSi_y test dataset. a, b, c Total energy/atom of the test structures, d, e, f Packing fraction of the test structures.

for the test dataset, the model's prediction ability deteriorates, resulting in an increased RMSE value of 0.11 eV/atom (see Table S3), where the best RMSE value obtained from Fig. 5b is 0.08 eV/atom. This represents a 37.5% increase in test dataset error compared to the RMSE value obtained with the actual data points. Therefore, original dataset has the advantage of using it as a training data point as they help cope with the limited availability of data. The hyperparameters and performance in terms of training, testing, and validation datasets for both the original and filtered datasets are provided in Tables 3 and S3.

Similar work has been done on predicting formation energy/atom, where the Kernel Ridge Regression (KRR) model is trained on 11,674 material

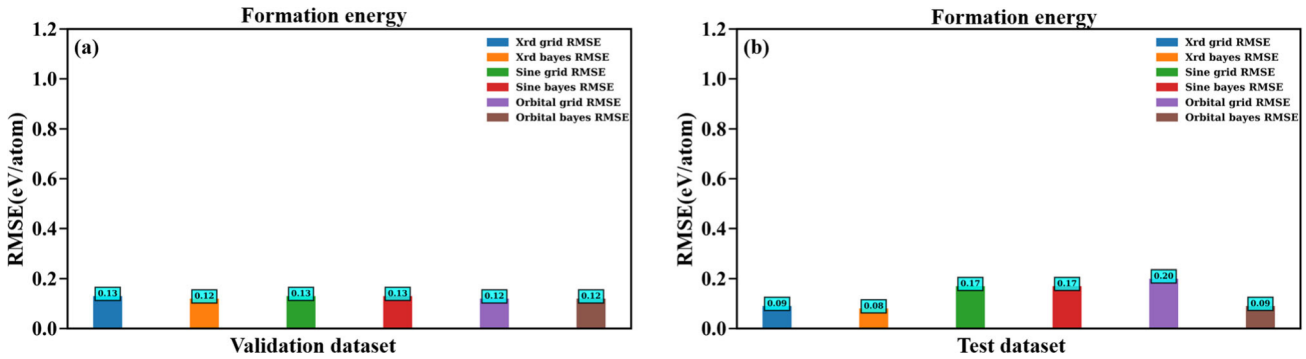

Figure 5 Performance comparison of hyperparameter tuning methods Grid Search CV and Bayes Search CV. RMSE value of formation energy/atom for (a) validation dataset, and (b) test dataset (Na_xSi_y and Al_xSi_yÞ. Three different structural descriptors with two different optimization technique generated 6 set of results labeled on top right side.

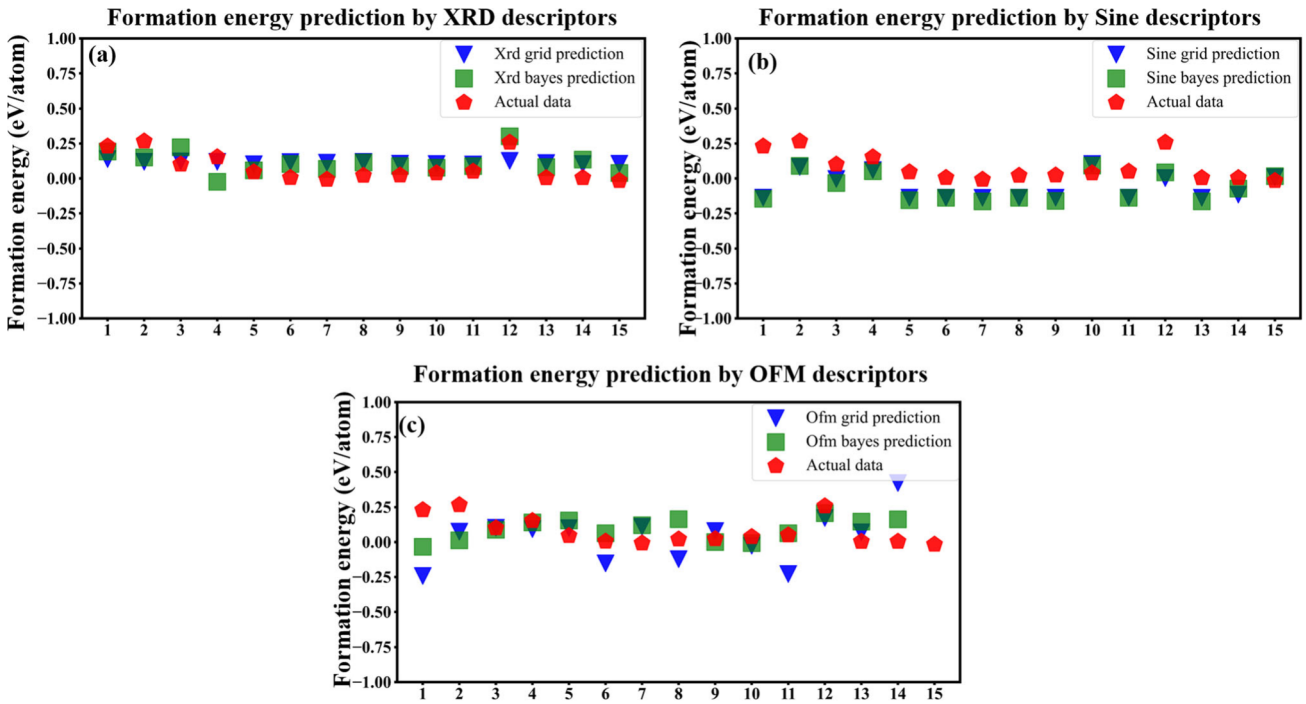

Figure 6 Comparing predictions using three structural descriptors with different hyperparameter search approaches for Na_xSi_y and Al_xSi_y test dataset. a, b, c Formation energy/atom of the test structures. XRD descriptors prediction value matches with the actual data points (see Fig. 6a). All the test structures and their properties are presented in supplementary information (see Table S7).

structures collected from MPD [70]. The study reported the RMSE of formation energy/atom prediction for validation datapoints to be 0.10 eV/atom. This performance is comparable to our current SVM-Bayes model, which achieved an RMSE of 0.12 eV/atom on the validation dataset when trained with only 730 structural data (see Fig. 7). Additionally, another similar study demonstrated that increasing the dataset size has a positive impact on prediction accuracy [77]. They employed a CNN-OFM-based model and observed that as the dataset size increased from 750 to 4,000, the RMSE for formation energy per atom on the validation dataset decreased from 0.18 to 0.10 eV/atom (see Fig. 7). Figure 7 compares the errors noted in previously reported energy prediction ML models and the training data used against the presented SVR-Bayes-XRD model. However, by excluding 21 outliers from the training data points

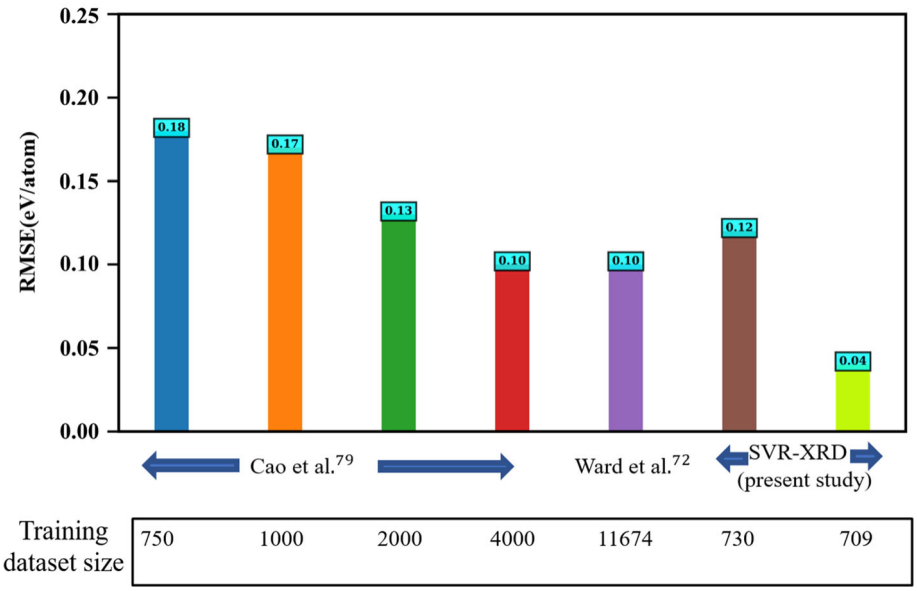

Figure 7 Performance comparison of the presented SVR-Bayes-XRD model with previously reported [70, 77] ML-based energy prediction models in terms of RMSE value and training dataset size. SVR-Bayes-XRD model shows better performance with limited training data size in comparison with previous reports.

following the modified Z score method [75], we can obtain an RMSE value of 0.04 eV/atom, shown in the last bar of Fig. 7 and Table S4. This comparison demonstrates that with the qualitative training data selection, hyperparameter tuning methods, and use of appropriate structural descriptors, ML models can overcome the need for extensive data requirements for training and accurate predictions.

Conclusion

In summary, we propose SVR-based machine learning method to speedily predict the thermodynamic and structural stability of Si alloying anodes before experimental design. The use of hyperparameter tuning methods such as Grid Search CV and Bayes Search CV, and the structural descriptors to convert atomic coordinates to comprehensive machine learning inputs have been elaborated. The predictive ability of three different types of descriptors has been studied for A_xSi_y atomic systems. XRD descriptor of the A_xSi_y structures as input data for the SVR model performed most reliably, especially considering the training structures were a mix of crystal, amorphous and different electronic configuration systems. While the OFM descriptor predicted total energy/atom, formation energy/atom and packing fraction with the lowest errors and highest accuracies for similar electronic configurations, OFM failed for the test cases where electronic configurations were slightly different from the training data (valency of cations). These results demonstrate that the choice of descriptor has more weight than the training data in making an ML model transferable to new systems. Moreover, the prediction accuracies were improved by the coupled use of SVR with Grid search CV method. In the two demonstrated experiments, hyperparameter selection by the Grid search CV method showed better predictions for the new structures. Though training and prediction times were shorter for SVR coupled with the Bayes search CV method, SVR-Bayes approach is suitable for predicting the stability of similar structures where transferability is not targeted. This study attempts to establish that the requirements of large datasets for machine learning-based approaches in material science domain can be overcome with the qualitative selection of training data, hyperparameter tuning methods, and appropriate structural descriptors.

Acknowledgements

The work is supported by National Science Foundation (NSF), Award Number # 2126180. Authors acknowledge Advanced Cyberinfrastructure Coordination Ecosystem: Service & Support (ACCESS) for the computational facilities (Award Number – DMR180013).

Author contributions

VS and JK conceived the project. JK performed all work and wrote the manuscript with VS and DD. All authors approved the final version of the manuscript.

Data availability

The data, pre- and post-processing code reported in this paper are available on GitHub (https://github.com/joy1303125/Si-based-Material-stability-prediction).

Declarations

Conflict of interest The authors have no conflicts of interest to declare. All authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

Ethical approval Not applicable.

Supplementary Information: The online version contains supplementary material available at https://doi.org/10.1007/s10853-023-08705-y.

References

[1] Nishi Y (2001) Lithium ion secondary batteries; past 10 years and the future. J Power Sources 100(1–2):101–106. https://doi.org/10.1016/S0378-7753(01)00887-4

[2] Shenoy VB, Johari P, Qi Y (2010) Elastic softening of amorphous and crystalline Li–Si phases with increasing Li concentration: a first-principles study. J Power Sources 195(19):6825–6830. https://doi.org/10.1016/j.jpowsour.2010.04.044

[3] Limthongkul P, Jang YI, Dudney NJ, Chiang YM (2003) Electrochemically-driven solid-state amorphization in lithium–silicon alloys and implications for lithium storage. Acta Mater 51(4):1103–1113. https://doi.org/10.1016/S1359-6454(02)00514-1

[4] Grosjean C, Herrera Miranda P, Perrin M, Poggi P (2012) Assessment of world lithium resources and consequences of their geographic distribution on the expected development of the electric vehicle industry. Renew Sustain Energy Rev 16(3):1735–1744. https://doi.org/10.1016/j.rser.2011.11.023

[5] Kubota K, Dahbi M, Hosaka T, Kumakura S, Komaba S (2018) Towards K-ion and Na-ion batteries as "beyond Li-ion." Chem Rec 18(4):459–479. https://doi.org/10.1002/tcr.201700057

[6] Arrieta U, Katcho NA, Arcelus O, Carrasco J (2017) First-principles study of sodium intercalation in crystalline na_x-$Si2_4$ (0 B x B 4) as anode material for Na-ion batteries. Sci Rep 7(1):1–8. https://doi.org/10.1038/s41598-017-05629-x

[7] Legrain F, Malyi OI, Manzhos S (2014) Comparative computational study of the energetics of Li, Na, and Mg storage in amorphous and crystalline silicon. Comput Mater Sci 94:214–217. https://doi.org/10.1016/j.commatsci.2014.04.010

[8] Ponrouch A, Tchitchekova D, Frontera C, Bardé F, Arroyo-de Dompablo ME, Palacín MR (2016) Assessing Si-based anodes for Ca-ion batteries: electrochemical decalciation of $CaSi_2$. Electrochem Commun 66:75–78. https://doi.org/10.1016/j.elecom.2016.03.004

[9] Zhang D, Fu J, Wang Z, Wang L, Corsi JS, Detsi E (2020) Perspective—reversible magnesium storage in silicon: an ongoing challenge. J Electrochem Soc 167(5):050514-050519. https://doi.org/10.1149/1945-7111/ab736b

[10] Beaulieu LY, Hatchard TD, Bonakdarpour A, Fleischauer MD, Dahn JR, Soc JE, A-a P, Beaulieu LY, Hatchard TD, Bonakdarpour A, Fleischauer MD (2003) Reaction of Li with alloy thin films studied by in situ AFM service reaction of Li with alloy thin films studied by in situ AFM. J Electrochem Soc 150(11):A1457. https://doi.org/10.1149/1.1613668

[11] Lee S, Lee J, Chung S, Lee H, Lee S, Baik H (2001) Stress effect on cycle properties of the silicon thin-film anode. J Power Sources 97:191–193

[12] Wang W, Kumta PN (2007) Reversible high capacity nanocomposite anodes of Si/C/SWNTs for rechargeable Li-ion batteries. J Power Sources 172:650–658. https://doi.org/10.1016/j.jpowsour.2007.05.025

[13] Kim H, Chou C, Ekerdt JG, Hwang GS (2011) Structure and properties of Li–Si alloys: a first-principles study. J Phys Chem C 115:2514–2521

[14] Wan W, Zhang Q, Cui Y, Wang E (2010) First principles study of lithium insertion in bulk silicon. J Phys Condens Matter 22(41):415501-415510. https://doi.org/10.1088/0953-8984/22/41/415501

[15] Chevrier VL, Zwanziger JW, Dahn JR (2010) First principles study of Li–Si crystalline phases: charge transfer, electronic structure, and lattice vibrations. J Alloys Compd 496(1–2):25–36. https://doi.org/10.1016/j.jallcom.2010.01.142

[16] Chiang HH, Lu JM, Kuo CL (2016) First-principles study of the structural and dynamic properties of the liquid and

amorphous Li–Si alloys. J Chem Phys 144(3):034502-034512. https://doi.org/10.1063/1.4939716

[17] Fan F, Huang S, Yang H, Raju M (2013) Mechanical properties of amorphous Li_xSi alloys: a reactive force field study. Model Simul Mater Sci Eng 21:074002. https://doi.org/10.1088/0965-0393/21/7/074002

[18] Johari P, Qi Y, Shenoy VB (2011) The mixing mechanism during lithiation of Si negative electrode in Li-ion batteries: an ab initio molecular dynamics study. Nano Lett 11(12):5494–5500

[19] Lee S, Ko M, Jung SC, Han YK (2020) Silicon as the anode material for multivalent-ion batteries: a first-principles dynamics study. ACS Appl Mater Interfaces 12(50):55746–55755. https://doi.org/10.1021/acsami.0c13312

[20] Niu J, Zhang Z, Aurbach D (2020) Alloy anode materials for rechargeable Mg ion batteries. Adv Energy Mater 10(23):2000697

[21] Mandal S, Haule K, Rabe KM, Vanderbilt D (2019) Systematic beyond-DFT study of binary transition metal oxides. npj Comput Mater 5(1):1–8

[22] Li W, Walther CFJ, Kuc A, Heine T (2013) Density functional theory and beyond for band-gap screening: performance for transition-metal oxides and dichalcogenides. J Chem Theory Comput 9(7):2950–2958

[23] Olson J, Priester M, Luo J, Chopra S, Zieve RJ (2005) Packing fractions and maximum angles of stability of granular materials. Phys Rev E Stat Nonlinear Soft Matter Phys 72(3):1–6. https://doi.org/10.1103/PhysRevE.72.031302

[24] He Q, Yu B, Li Z, Zhao Y (2019) Density functional theory for battery materials. Energy Environ Mater 2(4):264–279

[25] Deringer VL (2020) Modelling and understanding battery materials with machine-learning-driven atomistic simulations. J Phys Energy 2:041003-041011

[26] Deringer VL, Bernstein N, Bartók AP, Cliffe MJ, Kerber RN, Marbella LE, Grey CP, Elliott SR, Csányi G (2018) Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics. J Phys Chem Lett 9(11):2879–2885. https://doi.org/10.1021/acs.jpclett.8b00902

[27] He X, Zhu Y, Epstein A, Mo Y (2018) Statistical variances of diffusional properties from ab initio molecular dynamics simulations. npj Comput Mater 4(1):1–9

[28] Xie T, Grossman JC (2018) Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett 120(14):145301-145306. https://doi.org/10.1103/PhysRevLett.120.145301

[29] Sanyal S, Balachandran J, Yadati N, Kumar A, Rajagopalan P, Sanyal S, Talukdar P (2018) MT-CGCNN: integrating crystal graph convolutional neural network with multitask learning for material property prediction, https://doi.org/10.48550/arXiv.1811.05660

[30] Karamad M, Magar R, Shi Y, Siahrostami S, Gates ID, Farimani AB (2020) Orbital graph convolutional neural network for material property prediction. Phys Rev Mater 4(9):93801

[31] Laws KJ, Miracle DB, Ferry M (2015) A predictive structural model for bulk metallic glasses. Nat Commun 6:8123. https://doi.org/10.1038/ncomms9123

[32] Zeng S, Zhao Y, Li G, Wang R, Wang X, Ni J (2019) Atom table convolutional neural networks for an accurate prediction of compounds properties. npj Comput Mater 5(1):1–7. https://doi.org/10.1038/s41524-019-0223-y

[33] Bartel CJ, Trewartha A, Wang Q, Dunn A, Jain A, Ceder G (2020) A critical examination of compound stability predictions from machine-learned formation energies. npj Comput Mater 6(1):1–11. https://doi.org/10.1038/s41524-020-00362-y

[34] Natarajan AR, Van der Ven A (2018) Machine-learning the configurational energy of multicomponent crystalline solids. npj Comput Mater 4(1):1–7. https://doi.org/10.1038/s41524-018-0110-y

[35] Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, Rühl S, Wolverton C (2015) The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. npj Comput Mater 1:1–15. https://doi.org/10.1038/npjcompumats.2015.10

[36] Shapeev AV (2016) Moment tensor potentials: a class of systematically improvable interatomic potentials. Multiscale Model Simul 14(3):1153–1173

[37] Bartók AP, Payne MC, Kondor R, Csányi G (2010) Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. Phys Rev Lett 104(13):136403

[38] Fujikake S, Deringer VL, Lee TH, Krynski M, Elliott SR, Csányi G (2018) Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures. J Chem Phys 148(24):241714

[39] Behler J, Parrinello M (2007) Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys Rev Lett 98(14):146401

[40] Zhang J, Lei Y-K, Zhang Z, Chang J, Li M, Han X, Yang L, Yang YI, Gao YQ (2020) A perspective on deep learning for molecular modeling and simulations. J Phys Chem A 124(34):6745–6763

[41] Jiang D, Wu Z, Hsieh CY, Chen G, Liao B, Wang Z, Shen C, Cao D, Wu J, Hou T (2021) Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based

models. J Cheminform 13(1):1–23. https://doi.org/10.1186/s13321-020-00479-8

[42] Faber FA, Hutchison L, Huang B, Gilmer J, Schoenholz SS, Dahl GE, Vinyals O, Kearnes S, Riley PF, Von Lilienfeld OA (2017) Prediction errors of molecular machine learning models lower than hybrid DFT error. J Chem Theory Comput 13(11):5255–5264

[43] Yao K, Herr JE, Toth DW, Mckintyre R, Parkhill J (2018) The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. Chem Sci 9(8):2261–2269

[44] Himanen L, Jäger MOJ, Morooka EV, Federici Canova F, Ranawat YS, Gao DZ, Rinke P, Foster AS (2020) DScribe: library of descriptors for machine learning in materials science. Comput Phys Commun 247:106949. https://doi.org/10.1016/j.cpc.2019.106949

[45] Rupp M, Tkatchenko A, Müller KR, Von Lilienfeld OA (2012) Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett 108(5):1–5. https://doi.org/10.1103/PhysRevLett.108.058301

[46] Faber F, Lindmaa A, Von Lilienfeld OA, Armiento R (2015) Crystal structure representations for machine learning models of formation energies. Int J Quantum Chem 115(16):1094-1101. https://doi.org/10.1002/qua.24917

[47] Behler J (2011) Atom-centered symmetry functions for constructing high-dimensional neural network potentials. J Chem Phys 134(7):74106

[48] Bartók AP, Kondor R, Csányi G (2013) On representing chemical environments. Phys Rev B 87(18):184115

[49] Pham TL, Kino H, Terakura K, Miyake T, Tsuda K (2017) Machine learning reveals orbital interaction in materials. Sci Technol Adv Mater 6996(November):1–2. https://doi.org/10.1080/14686996.2017.1378060

[50] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G (2013) Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater 1(1):11002

[51] Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C (2013) Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). Jom 65(11):1501–1509

[52] Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, Taylor RH, Wang S, Xue J, Yang K, Levy O (2012) AFLOW: an automatic framework for high-throughput materials discovery. Comput Mater Sci 58:218–226

[53] Bergerhoff G, Hundt R, Sievers R, Brown ID (1983) The inorganic crystal structure data base. J Chem Inf Comput Sci 23(2):66–69

[54] Awad M, Khanna R (2015) Support vector regression. Efficient learning machines. Springer, p 67–80

[55] Balabin RM, Lomakina EI (2011) Support vector machine regression (LS-SVM)—an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data? Phys Chem Chem Phys 13(24):11710–11718

[56] Kondati Natarajan S, Behler J (2017) Self-diffusion of surface defects at copper-water interfaces. J Phys Chem C 121(8):4368–4383

[57] Behler J, Parrinello M (2007) Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys Rev Lett 98(14):1–4. https://doi.org/10.1103/PhysRevLett.98.146401

[58] Behler J (2021) Four generations of high-dimensional neural network potentials. Chem Rev 121(16):10037–10072

[59] Graser J, Kauwe SK, Sparks TD (2018) Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons. Chem Mater 30(11):3601–3612

[60] Jung H, Park M, Yoon Y-G, Kim G-B, Joo S-K (2003) Amorphous silicon anode for lithium-ion rechargeable batteries. J Power Sources 115(2):346–351

[61] Ohara S, Suzuki J, Sekine K, Takamura T (2004) A Thin film silicon anode for Li-ion batteries having a very large specific capacity and long cycle life. J Power Sources 136(2):303–306

[62] Chang KC, Nuhfer NT, Porter LM, Wahab Q (2000) High-carbon concentrations at the silicon dioxide-silicon carbide interface identified by electron energy loss spectroscopy. Appl Phys Lett 77(14):2186–2188. https://doi.org/10.1063/1.1314293

[63] Pujahari RM (2021) Crystalline silicon solar cells, Elsevier. https://doi.org/10.1016/B978-0-12-823710-6.00004-2

[64] Guha S, Yang J, Nath P, Hack M (1986) Enhancement of open circuit voltage in high efficiency amorphous silicon alloy solar cells. Appl Phys Lett 49(4):218–219. https://doi.org/10.1063/1.97176

[65] Yanxon H, Zagaceta D, Wood BC, Zhu Q (2020) Neural network potential from bispectrum components: a case study on crystalline silicon. J Chem Phys 153(5):54118

[66] Comin M, Lewis LJ (2019) Deep-learning approach to the structure of amorphous silicon. Phys Rev B 100(9):94107

[67] Jain A, Hautier G, Moore CJ, Ping Ong S, Fischer CC, Mueller T, Persson KA, Ceder G (2011) A high-throughput infrastructure for density functional theory calculations. Comput Mater Sci 50(8):2295–2310. https://doi.org/10.1016/j.commatsci.2011.02.023

[68] Peterson GGC, Brgoch J (2021) Materials discovery through machine learning formation energy. J Phys Energy 3(2):022002. https://doi.org/10.1088/2515-7655/abe425

[69] Chen WC, Vohra YK, Chen CC (2022) Discovering superhard B-N-O compounds by iterative machine learning and

evolutionary structure predictions. ACS Omega 7(24):21035–21042. https://doi.org/10.1021/acsomega.2c01818

[70] Ward L, Dunn A, Faghaninia A, Zimmermann NER, Bajaj S, Wang Q, Montoya J, Chen J, Bystrom K, Dylla M, Chard K, Asta M, Persson KA, Snyder GJ, Foster I, Jain A (2018) Matminer: an open source toolkit for materials data mining. Comput Mater Sci 152(April):60–69. https://doi.org/10.1016/j.commatsci.2018.05.018

[71] Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier VL, Persson KA, Ceder G (2013) Python materials genomics (Pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci 68:314–319

[72] Awad M, Khanna R (2005) Support vector regression. Effic Learn Mach 2007:67–80. https://doi.org/10.1007/978-1-4302-5990-9_4

[73] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

[74] Waskom ML (2021) Seaborn: statistical data visualization. J Open Source Softw 6:3021

[75] Aggarwal V, Gupta V, Singh P, Sharma K, Sharma N (2091) Detection of spatial outlier by using improved Z-score test. In: 2019 3rd international conference on trends in electronics and informatics (ICOEI). IEEE, pp 788–790

[76] Kim S, Noh J, Gu GH, Aspuru-Guzik A, Jung Y (2020) Generative adversarial networks for crystal structure prediction. ACS Cent Sci 6(8):1412–1420

[77] Cao Z, Dan Y, Xiong Z, Niu C, Li X, Qian S, Hu J (2019) Convolutional neural networks for crystal material property prediction using hybrid orbital-field matrix and magpie descriptors. Cryst 9(4):1-15 https://doi.org/10.3390/cryst9040191