Energy Storage Materials 31 (2020) 434–450

Contents lists available at ScienceDirect

Energy Storage Materials

journal homepage: www.elsevier.com/locate/ensm

Machine learning assisted materials design and discovery for rechargeable batteries

Yue Liu [a,b,c], Biru Guo [a], Xinxin Zou [a], Yajie Li [d,e], Siqi Shi [d,e,*]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China
[b] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, 200444, China
[c] Shanghai Engineering Research Center of Intelligent Computing System, Shanghai, 200444, China
[d] State Key Laboratory of Advanced Special Steel, School of Materials Science and Engineering, Shanghai University, Shanghai, 200444, China
[e] Materials Genome Institute, Shanghai University, Shanghai, 200444, China

ARTICLE INFO

Keywords:
Machine learning
Rechargeable battery
Materials design and discovery
Feature engineering

ABSTRACT

Machine learning plays an important role in accelerating the discovery and design process for novel electrochemical energy storage materials. This review aims to provide the state-of-the-art and prospects of machine learning for the design of rechargeable battery materials. After illustrating the key concepts of machine learning and basic procedures for applying machine learning in rechargeable battery materials science, we focus on how to obtain the most important features from the specific physical, chemical and/or other properties of material by using wrapper feature selection method, embedded feature selection method, and the combination of these two methods. And then, the applications of machine learning in rechargeable battery materials design and discovery are reviewed, including the property prediction for liquid electrolytes, solid electrolytes, electrode materials, and the discovery of novel rechargeable battery materials through component prediction and structure prediction. More importantly, we discuss the key challenges related to machine learning in rechargeable battery materials science, including the contradiction between high dimension and small sample, the conflict between the complexity and accuracy of machine learning models, and the inconsistency between learning results and domain expert knowledge. In response to these challenges, we propose possible countermeasures and forecast potential directions of future research. This review is expected to shed light on machine learning in rechargeable battery materials design and property optimization.

1. Introduction

The development of energy storage and conversion devices is crucial to reduce the discontinuity and instability of renewable energy generation [1,2]. According to the global energy storage project repository of the China Energy Storage Alliance (CNESA) [3], as of the end of 2019, global operational electrochemical energy storage project capacity totaled 8239.5 MW (4.5% of the total global energy storage market of 183.1 GW). As a key component of electrochemical energy storage, rechargeable batteries are extremely vital for a broad range of applications, including new energy vehicles, consumer electronics, and aerospace. To meet the growing needs of these applications, the higher demands are being put forward for rechargeable batteries with higher energy density, higher power density, longer cycle life, higher safety and at an acceptable cost. Thus, it is urgent to develop key rechargeable battery materials including those for electrodes and electrolytes, to improve the performance of rechargeable batteries [4].

Traditional trial-and-error methods are too time consuming to keep pace with the rapid evolution of demand. Compared with trial-and-error methods, computational simulation is advantageous for providing useful experiments over which one has full control of the relevant variables. Since the 1980s, the crossing and integration of materials science, physics, and computational science, has resulted in the development of various computational simulation methods to accelerate the research on rechargeable battery materials [5]. The computational simulation methods have covered a wide range of spatial and temporal scales. These include microscale simulations (such as first-principles (FP) calculations, quantum mechanics (QM), molecular dynamics (MD) and Monte Carlo

* Corresponding author. State Key Laboratory of Advanced Special Steel, School of Materials Science and Engineering, Shanghai University, Shanghai, 200444, China.
E-mail address: sqshi@shu.edu.cn (S. Shi).

https://doi.org/10.1016/j.ensm.2020.06.033
Received 6 April 2020; Received in revised form 12 June 2020; Accepted 25 June 2020
Available online 2 July 2020

(MC) techniques), mesoscale simulations focused on the phase-field (PF) method and force-field (FF) approach, and macroscale simulations based on the finite element (FE) and finite difference (FD) methods. Multiscale calculation methods and some of the major applications at microscale, mesoscale, and macroscale in rechargeable batteries are shown in Fig. 1. In 1998, using the FP calculations, Ceder et al. [6] found that the novel cathode material $Li(Co,Al)O_2$, could not only increase the lithium battery voltage, but could also decrease the density and cost of the batteries. Subsequently, Yang and Tse (2011) [7] investigated the diffusion mechanisms of Li ions in the cathode material $LiFePO_4$, using MD calculations. By adopting QM calculations, Husch and Korth (2015) [8] assessed the thermodynamic effects of electrolyte materials. Moreover, to explore stable solvents for use in the design of novel high-voltage stable electrolytes, Pande and Viswanathan (2018) [9] applied FP to identify simple descriptors to determine the influence of solvation on the oxidative stability of various electrolyte components. In the same year, Fujie et al. [10] proposed a new energy estimation method for the MC procedure using the QM method. They applied this method to research the formation of solid electrolyte interphase film in lithium-ion batteries. In 2012, the FF approach was applied to study the mechanism of the conversion reaction by which lithium metal formed the nanomaterial FeF_2 in lithium ion batteries (Ma and Garofalini) [11]. Furthermore, in 2018, Cogswell and Bazant [12] simulated phase separation in realistic nanoparticle geometries for Li_xFePO_4 using a PF model. In addition, the temperature distribution and force distribution of lithium-ion batteries were predicted by using the FE method in 2010 [13] and 2012 [14], respectively. As seen above, the computational simulation methods have been successfully applied to the study of rechargeable battery materials at microscale, mesoscale, and macroscale.

Given the sophisticated requirements involved in understanding the basic physicochemical properties of rechargeable battery materials, the effectiveness experimental measurement and computational simulation in exploring unconstrained chemical spaces and/or complex real-world rechargeable battery materials remains limited. The resource consumption of a single experiment or calculation is usually large, and the number of experiments or calculations will increase by continuous trial and error only based on limited knowledge or experience will increase, resulting in the waste of a large number of resources. To accelerate the research and application process for new materials, the "Materials Genome Initiative" (MGI) was proposed in the United States in June 2011 [15]. The critical idea behind the MGI is the combination of "experiment", "calculation", and "data" [16], which means a new model of performing theoretical simulation and property prediction priority with experimental verification behind. In the past, extensive experimental materials databases were accumulated to provide rechargeable battery materials engineers with ready access to the properties of known materials. These were gathered into such as the inorganic crystal structure database (ICSD) [17], Cambridge Structure Database (CSD) [18] and Pauling file database [19]. Among the most fundamental properties of known rechargeable battery materials, is the crystal structure information. This includes atomic positions, space groups, lattice constants, and symmetry, and these are available in the above databases. In addition, MGI has given birth to many high-throughput material computing platforms and databases, such as the Materials Project (MP) [20], AFLOWLIB [21], the Open Quantum Materials Database (OQMD) [22], the Harvard Clean Energy Project (HCEP) [23], the Electronic Structure Project (ESP) [24], the Computational Materials Repository (CMR) [25], Novel Materials Discovery (NOMAD) [26], and NIST [27]. These databases provide MGI engineers with access to high quality data. As shown in Table 1, a huge collection of thermodynamic properties such as data on energy properties and structural properties have been accumulated by computational simulation and experimental measurement. These materials databases offer an opportunity for the emergence of the fourth paradigm of rechargeable battery materials science: data-driven materials discovery. The "data-driven materials discovery" model represents the core concept and development direction of MGI.

The high-throughput computational materials design is based on the combination of computational quantum-mechanical-thermodynamic approaches and a multitude of techniques rooted in database construction and intelligent data mining [28]. As shown in Fig. 2, since the launch of MGI in 2011, more than 2518 articles related to MGI have emerged, including 1788 articles related to high-throughput computing, and 730 articles about machine learning (ML) for materials discovery and property prediction. ML is being used as a powerful tool for finding patterns in high-dimensional data. It helps to reduce the amount of calculation needed and speeds up exploration of novel materials. The high-throughput databases provide opportunities for ML to exploit high-quality data for materials discovery. In the last decade, the literature number on materials design based on high-throughput computing and ML has been on the rise. In recent years, in particular, the number of academic papers has grown exponentially on "materials design by ML".

In recent years, some successful examples of ML in various materials

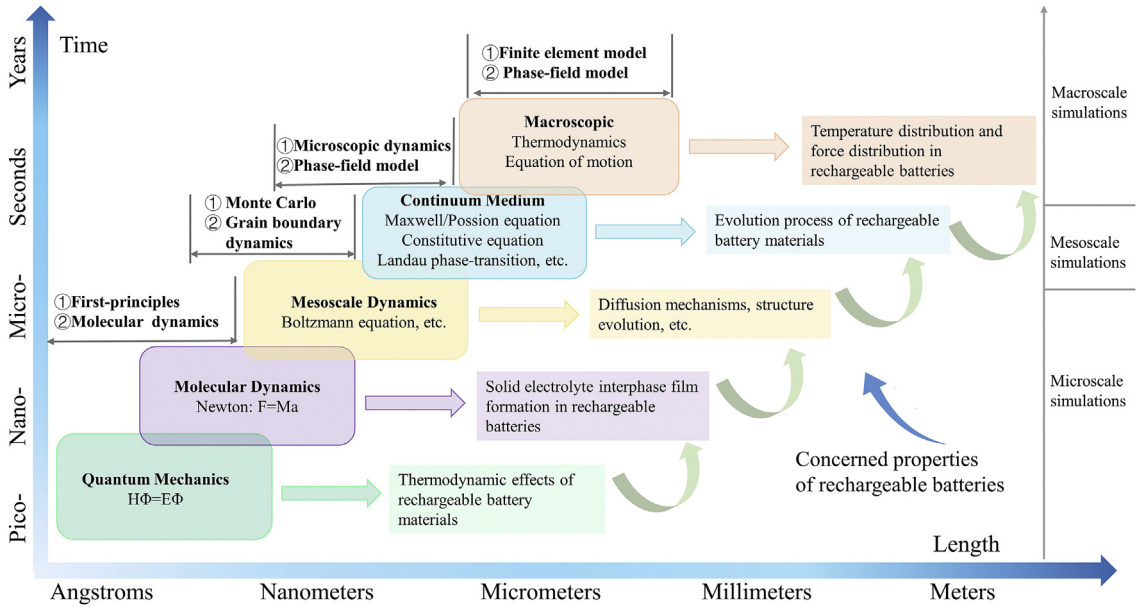

Fig. 1. Applications of multi-scale computation methods at microscale, mesoscale, and macroscale, in rechargeable battery materials.

Table 1
List of some notable material databases.

Database	URL	Category	Storage content	Data Sources
ICSD [17]	cds.dl.ac.uk/cds/datasets/crys/icsd/llicsd.html	Inorganic crystal	Structure properties	Experiment
CSD [18]	www.ccdc.cam.ac.uk/pages/Home.aspx	Organic crystal	Structural properties	Experiment
Pauling file [19]	paulingfile.com	Inorganic material	Constitution, structure, powder diffraction data, and physical properties	Experiment
MP [20]	www.materialsproject.org	All	Structure and energy properties	ICSD, computation
AFLOWLIB [21]	aflowlib.org	Alloy	Electronic structure, electromagnetic properties, etc.	Computation
OQMD [22]	oqmd.org	All	Structural, thermodynamic properties, etc.	ICSD, computation
HCEP [23]	cepdb.molecularspace.org	Organic solar cell material	Electronic structure, molecular information	Computation
ESP [24]	http://gurka.fysik.uu.se/ESP/	Inorganic compound	Electronic structure properties	Computation
CMR [25]	https://cmr.fysik.dtu.dk/	All	Physical, chemical properties, etc.	Computation
NOMAD [26]	https://nomad-coe.eu/	All	Structural, electronic properties, etc.	Computation
NIST [27]	http://webbook.nist.gov	All	Chemical, physical properties, etc.	Computation

fields have emerged, including composite materials [29–31], alloy materials [32–34], catalyst materials [35], and battery materials [36–43]. The most representative work about ML in rechargeable battery materials development follows. In 2014, Jalem et al. [38] combined an artificial neural network (ANN) and partial least squares (PLS) algorithm with density functional theory (DFT) to predict the diffusion barrier and cohesive energy of an olivine-type LiMXO_4. Using this approach, 15 promising lithium ion battery solid electrolyte materials were screened out (i.e., selected). Afterwards, they developed a Bayesian-driven approach to screen for fast-conducting Li-and Na-containing tavorite type compounds efficiently [39], which involved a search space of 318 $AMXO_4Z$ tavorites (A, M, X, and Z are sites for ionic substitution). The scheme only requires ~30% of the total DFT-based evaluations to recover the optimal compound ~90% of the time. In 2017, Sendek et al. [40] established a model by training on 40 samples using logistic regression (LR) to predict ionic conductivity. Twenty-one solid electrolyte materials with ionic conductivity greater than or equal to 10^{-4} S/cm were screened out (extracted) from the MP database. Furtherly, they performed DFT-MD calculations on the promising candidate materials, finding evidence of superionic Li conduction in eight and marginal Li conduction in two [41]. In 2018, Ahmad et al. [42] performed a computational screening of over 12,000 inorganic solids based on their ability to suppress dendrite initiation in contact with a Li metal anode. Using ML models, twenty interfaces with six solid electrolytes were predicted to be resistant to dendrite initiation. In 2019, Arghya et al. [43] proposed an approach to reversely design high-performance solid electrolyte interphase (SEI) using semi-supervised generative deep learning models, throughput synthesis, and laboratory testing together. With further development of MGI, ML will obtain the composition-structure-property relationships of rechargeable battery materials in a faster and more accurate manner and become an effective computational method for rechargeable battery materials.

Currently, several excellent review articles on materials informatics are available in the literatures [44–55]. Thereinto, references [52–55] demonstrate the research status of ML in key energy materials including catalysis, batteries, solar cells, and crystal discovery. However, the details about inner workings of ML algorithms in property prediction and application process for rechargeable battery materials have been rarely covered. With this review, we try to present a comprehensive overview of ML in rechargeable battery materials in the view of ML methods. We not only deeply analyze the successful experiences and the common existing problems, but also establish a new horizon for the discovery and design of rechargeable battery materials in the framework of ML. The remainder of this paper is structured as follows: Section 2 briefly introduces the paradigm of ML in rechargeable battery materials science. Then, Section 3 describes feature engineering in the step of sample construction, and

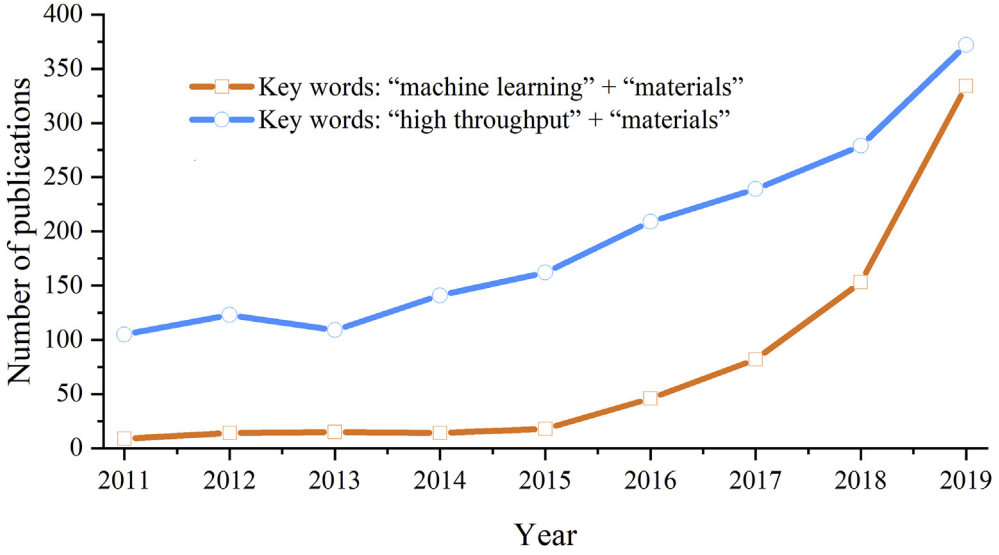

Fig. 2. Literature counts from 2011 to 2019. The information is obtained by searching on the "Web of Science" database using "high throughput" + "materials" and "machine learning" + "materials" as key words, respectively.

related feature selection methods. Section 4 demonstrates the research status with respect to the applications of ML in rechargeable battery materials. Section 5 discusses the limitations and challenges complicating the use of ML in the field of rechargeable battery materials science. Then, corresponding countermeasures for improvement are proposed. Finally, the conclusions are presented in Section 6.

2. Machine learning paradigm for use in rechargeable battery materials science

2.1. Machine learning and several related concepts

As a scientific endeavor, ML originated from the exploration of artificial intelligence [56]. In the 1950s, various symbolic methods are tried to solve the problem of acquiring knowledge by machine [57]. And later, some methods based on the connection principle such as neural network and perceptron have been widely studied [58]. Subsequently, support vector machines (SVM) and decision tree (DT) based on statistical learning theory (SLT) were proposed [59]. At present, some new ML methods, such as deep learning for big data analysis, are getting more attention in academia and industry.

ML is a discipline in which research is done to determine how to make computers learn automatically, acquire knowledge, and continuously improve their own performance without explicit programming [60]. Since the first ML seminar was held at Carnegie Mellon University of the United States in the summer of 1980, ML has become an independent discipline and has begun to take shape rapidly. As shown in Fig. 3, ML is also a cross discipline which has the close relationships with the current hot topics in computer science. ML is a branch of artificial intelligence and an important way for computer to acquire the knowledge. The development of ML originates from the neurocomputing. ML provides the methods for data mining and pattern recognition. Note that currently the most ML algorithms are based on the SLT.

ML exhibits good applicability in classification, regression and other issues related to the high-dimensional data. Aimed at extracting knowledge and finding insights hidden in data, ML learns from previous experiences to produce reliable, repeatable results. Thus, ML plays an important role in many fields, especially in speech recognition, image recognition, bioinformatics, information security, natural language processing (NLP), and materials science. The pioneering application of ML in rechargeable battery materials can be traced back to the 1990s, when the fuzzy logic methodology was employed to predict the state-of-charge and state-of-health of rechargeable battery systems [62]. Subsequently, ML was used to approach various kinds of topics in rechargeable battery materials science, such as new materials discovery and materials properties prediction.

2.2. Machine learning paradigm for use in rechargeable battery materials science

A classical definition of ML is as follows: $< P,T,E >$, where P , T , and E denote performance, task, and experience, respectively. The main interpretation is that a computer program learns from experience E with respect to task T and a performance measure P if its performance on task T, as measured by P, improves with experience E [63].

In rechargeable battery materials science, the task T mainly focuses on property prediction and discovery of new materials. In order to satisfy such tasks, the commonly used ML models involve regression and classification, such as linear models (PLS and LR), nonlinear models (SVM, ANN, and random forest: RF), and a small number of clustering models. The performance P is generally represented as the accuracy of these models. The experience E corresponds to the dataset of relevant materials in a specific task, which usually consists of a set of condition attributes that represent the characteristics of materials, and a decision attribute that reflects a certain property of materials. For example, to predict the ionic conductivity of solid-state lithium ion electrolytes, Beal et al. [64] established a radial basis function neural network (RBFNN) based on the data collected in the experiment as experience E. The model performance P was quantified by the average correlation coefficient (R^2) of 0.92.

The process of ML applied to rechargeable battery materials is shown in Fig. 4. First, some data can be collected directly from existing material databases, or the results of experimental measurements and simulation calculations can be used. Some data is usually obtained through further calculation, generally as the condition attributes of the ML models. Then, after the process of data cleaning and feature engineering (including feature extraction and selection), the original data can be converted to samples to train the ML model. Third, the mapping relationship between the conditional attributes and the decision attribute can be simulated by selecting the appropriate ML algorithm and tuning the optimal hyperparameter. Finally, experts and researchers can exploit these models to predict the properties of materials or to guide the discovery of new materials.

3. Feature engineering in rechargeable battery materials science

A huge amount of data has been accumulated by computational

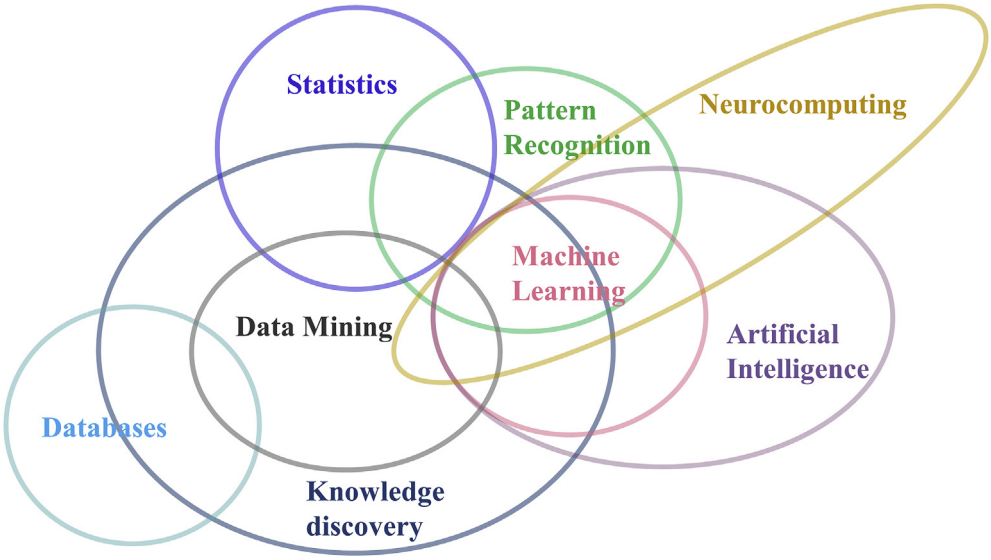

Fig. 3. The relationships among ML and current hot topics in computer science [61].

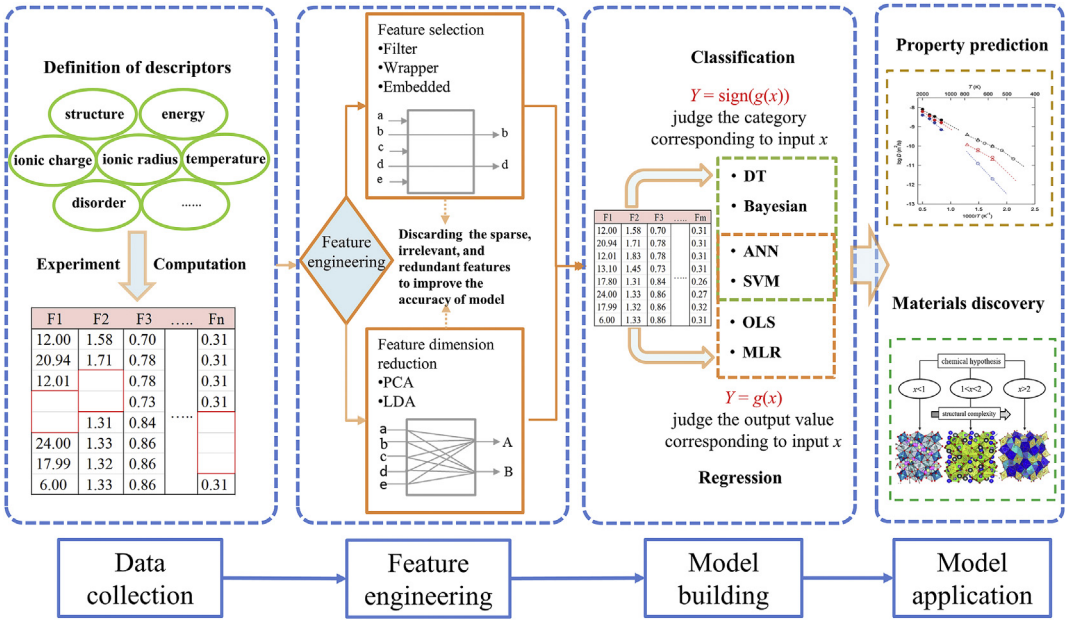

Fig. 4. Workflow of the application of ML in rechargeable battery materials, which includes four major steps: data collection, feature engineering, model building, and model application.

simulations and experimental measurements in the field of rechargeable battery materials. These data are typically incomplete, noisy, and inconsistent, and thus, data cleaning should be performed after collecting the original data. Specifically, the target property of rechargeable battery materials is usually described by many descriptors, including some basic descriptors such as atomic electronegativity, atomic radius and unit cell dimension which describe the elements or structural properties of compounds, and some derived descriptors defined by relevant formulas or empirical rules based on the above descriptors or the crystal structure itself, such as bottleneck size, ion radius radio, polyhedron volume composed of multiple atoms in unit cell.

When the basic descriptors are regarded as input factors, only the ML model can effectively preserve the specific physical, chemical and/or other properties of material, but the trained model is often unexplainable because of the extremely complex relationship between descriptors and the target property. The successful of the crystal graph convolutional neural network (CGCNN) lies in its ability to appropriately represent the crystal structure which preserves the structure topology by a series of basic descriptors, and the certain interpretability of the model [42,65,66]. Derived descriptors usually serve as input factors combined with basic descriptors, which can simplify the representation and improve the interpretability of the ML model, such as reference [40]. However, the selection of these descriptors often depends on the physical intuition of experts. For example, in the prediction of the capacity of lithium ion batteries, LASSO, KRR, RF, and SVR were used by Kauwe et al. [67], but the prediction results were not ideal because of the lack of appropriate descriptors. Anyway, it is still possible that some of descriptors are not really relevant to the target property nor do they contribute to the construction of the ML models. Therefore, feature engineering, which is being conducted to identify the appropriate features related to property prediction, is an important step in ML model building [68]. Feature selection (FS) [69], identification and ranking of the most relevant features, greatly affects the computational speed and predictive ability of the model. It is an important part of feature engineering. Several FS methods have been presented, which can generally be grouped into three categories that include filter, wrapper, and embedded methods [70]. Among them, the wrapper and embedded methods are mostly used in the field of rechargeable battery materials.

3.1. Wrapper feature selection applied in property prediction of rechargeable battery materials

Wrapper FS methods are the most widely used among all the FS methods in the property prediction of rechargeable battery materials. Wrapper FS methods are generally used in combination with a specific ML model and a meta-heuristic algorithm to identify the best subset of features without sacrificing prediction accuracy [71]. First, the wrapper generates several initial candidate feature subsets based on a predefined search strategy (e.g., exhaustive search: ES, genetic algorithm, random search, and sequential search: SS). Then, a specific ML model, such as SVM or ANN, is trained to evaluate each candidate feature subset. Some candidate feature subsets are retained and used to generate the next set of feature subsets. This process is performed iteratively until the selected feature subset meets the iteration stop condition, such as maximum stop times and prediction accuracy thresholds.

In order to find the appropriate features related to the property prediction, many researchers have introduced the wrapper methods into the property prediction (Table 2). For instance, Sendek et al. took special care to avoid overfitting the data including 38 samples with 19 features in predicting lithium ionic conductivity. They employed wrapper methods with an ES strategy and selected 7 features from 19 structured descriptors that depends on simple crystallographic and chemical relationships [72]. Then they trained a model with a mean-squared leave-one-out cross-validation (LOOCV) error of 0.41σ ($\sigma = 2.49$). Furthermore, in 2017 [40], Sendek et al. added two samples with one more feature to their previous work [72], and performed ES and LR to select the best subset from 20 features based on 40 samples. They selected five features such as average number of lithium neighbors for each lithium, the average sublattice bond ionicity, the average anion-anion coordination number in the anion framework, the average shortest lithium-anion distance in angstroms, and the average shortest lithium-lithium distance. According to the model coefficient, it was found that compared to the lithium-lithium bond number, the equilibrium lithium-anion distance has more influence on ionic conductivity. The best model got a cross-validated misclassification rate (*CVMR*) of 0.1. Additionally, in order to find the optimal subset of molecular descriptors of the electrical conductivity including temperature, anion-based

Table 2
Prediction results when using the wrapper method for feature selection.

Literature	Materials	Property	ML method	Original Features	Selected Features	Evaluation Index and Result
[40]	solid state electrolytes	lithium ionic conductivity	LR	20	5	*CVMR*: 0.1
[72]	solid state electrolytes	lithium ionic conductivity	LWLS	19	7	LOOCV error: 0.41σ
[73]	ionic liquids	electrical conductivities	LSSVM	unspecified	10	*AARD*: less than 1.9%
[74]	face-centered cubic (FCC) host systems	diffusion energy barriers	GKRR	111	23	*RMSE*: 0.15 eV

molecular descriptors, and cation-based molecular descriptors, Gharagheizi et al. [73] applied an SS strategy to successfully screen out 10 key descriptors and built a least square SVM (LSSVM) model with average absolute relative deviation (*AARD*) of less than 1.9%. Wu et al. [74] also leveraged this method to select 23 key descriptors from 111 features (more than 200 samples) containing elemental properties for the host and impurity atoms, as well as difference or ratio of properties between host and impurity atoms. A Gaussian kernel ridge regression (GKRR) model was developed to predict FCC solute diffusion barriers with root mean square error (*RMSE*) of 0.15 eV.

As seen from the above research studies, many data are not only of high dimensionality, but also of small sample size, which easily leads to overfitting of the ML model and reduces the generalization ability. This is another important reason why FS is required in rechargeable battery materials science, and this phenomenon is discussed in detail in Section 5.1. In addition, wrapper FS methods allow different search strategies and numerous ML models to be combined at will. This increases the difficulty of material experts without professional-level computer knowledge employing the method. Therefore, it is worthwhile to deeply study the automatic selection and design of the wrapper FS algorithm on the way of future research.

3.2. Embedded feature selection applied for performance prediction of rechargeable battery materials

Compared with wrapper FS methods, embedded FS methods perform FS during the process of model construction (of such as PLS, least absolute shrinkage selection operator: LASSO, RF, and Elastic net) [75,76]. In this case, the step of model construction and the FS part are synchronized, and the criterion for evaluating features is derived from the basic function of a particular class of regression or classification. This greatly simplifies the process of FS. Not only can embedded FS methods remove redundant and inappropriate features, but some can also rank features in order of importance. This makes sense for use in rechargeable battery materials science because experts can use it to make more targeted scientific research and materials design. Therefore, embedded FS methods are also employed frequently in rechargeable battery materials science.

In order to identify the major factors affecting the cathode volume of a Li-ion battery, Xiao et al. [77] employed PLS based on data of 28 cathodes to measure the importance of five types of descriptors including crystal structure, element, composition, local distortion and electronic level, with 34 factors in total. According to the variable importance in projection analysis, they found that the radius of the X^{4+} ion and the X octahedron descriptors strongly affect the volume changes. By using extremely randomized trees (ERT), Shandiz and Gauvin [78] predicted the crystal structure of Li-ion silicate cathode materials, and found that the volume of crystal was the most important feature. Experts can further design low-strain cathode materials by focusing on adjusting the values of these descriptors. In 2018, Li et al. [79] proposed a RF regression model for battery capacity estimation, which not only captured the dependency of the battery capacity and associative features, but also successfully evaluated the health status of different batteries under varied cycling conditions, and did so with *RMSE* of less than 1.3%.

Overall, embedded FS methods can determine the importance of each feature from the basics of a particular learner; thus, the FS results can be explained from the perspective of an algorithm. However, the embedded methods are represented by several ML models (e.g., RF, LASSO, and Elastic net), which makes its application in the field of rechargeable battery materials very limited. Moreover, the hyperparameters of the algorithm also need to be manually searched and optimized to achieve better performance.

3.3. Combination feature selection applied in performance prediction of rechargeable battery materials

The combination of multiple FS methods has become a hot topic in ML, because this approach allows processing of features from different perspectives such as relevance, sparsity, and redundancy. For example, in the novel hybrid FS method [80] in materials science, a candidate feature subset is first filtered out of the original feature set by performing a highly efficient filter procedure. Then, the feature subset is further tuned by performing a more accurate wrapper procedure.

In rechargeable battery materials science, we first developed a data-driven multi-layer FS method combined with domain expert knowledge (named DML-FSdek) [81]. This method combines filter and wrapper methods to eliminate sparse, irrelevant, and redundant features sequentially and automatically. The domain expert knowledge is integrated into the process of FS to eliminate the risk of crucial features being removed. We performed experiments to compare this method and two existing sparsity ones (LASSO, Elastic net) on four battery materials datasets [82–85]. As shown in Table 3, our method has lower *RMSE* than LASSO and Elastic net do, which shows better prediction performance. The number of features selected by our method is greater than that by the sparsity methods. This is because the introduction of expert knowledge allows important features to be retained. In general, the method can improve the prediction performance on the premise that the selected features are consistent with the domain expert knowledge.

4. Applications of machine learning in rechargeable battery materials science

ML is widely used in rechargeable battery materials science and its superiority has been proven in both time efficiency and prediction accuracy. The applications of ML in rechargeable battery materials mainly include property prediction and materials discovery. The essence of both is to establish the quantitative structure activity relationship (QSAR) between conditional attributes (descriptors) and decision attributes (the properties of interest) by using ML algorithms. The prediction of properties considers a wide range of material properties as decision attributes, such as migration energy, ionic conductivity, electrical conductivity, thermal conductivity, cathode volume and lattice constant, and typically applies regression analysis methods. In research on materials discovery, the QSAR between components or structural descriptors and stability-related properties such as formation energy, and classification and clustering algorithms are also applied.

4.1. Property prediction for rechargeable battery materials

As shown in Fig. 5, a large number of ML methods have been successfully applied to the property prediction of rechargeable battery materials. These include such as linear model LR and OLS, nonlinear model support vector regression (SVR), ANN, convolutional neural network (CNN), and deep neural networks (DNN). Details of ML applications that are successful in predicting the properties of rechargeable battery

Table 3
Comparison between LASSO and Elastic Net methods and ours on four battery materials datasets.

		Feature selection models					
		LASSO		Elastic net		Our method	
Dataset	Initial	Selected	*RMSE*	Selected	*RMSE*	Selected	*RMSE*
[82]	6	4	0.1546	3	0.1657	5	0.1428
[83]	5	2	0.1437	2	0.1437	3	0.1136
[84]	6	1	0.2166	1	0.2166	6	0.1301
[85]	6	2	0.1448	3	0.1373	5	0.1301

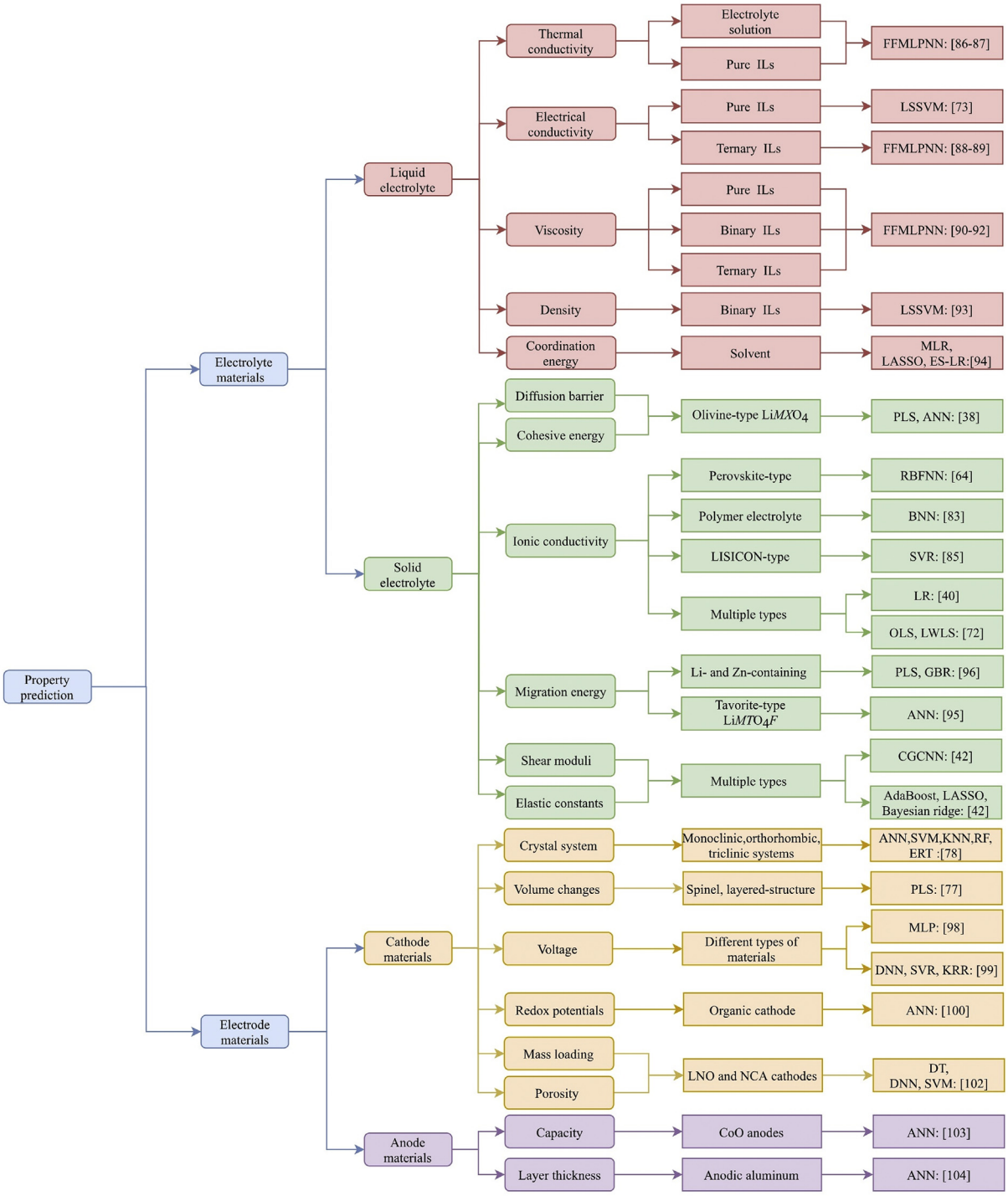

Fig. 5. Taxonomy of the literature for ML in rechargeable battery materials. Related methods and literatures for each performance prediction are marked.

materials will be discussed in the next three subsections, which include liquid electrolyte materials, solid electrolyte materials, and electrode materials.

4.1.1. Property prediction for liquid electrolytes

For liquid electrolytes, the thermal conductivity, electrical conductivity, viscosity, density, and the coordination energy of the ions to the solvent are always concerned. However, it is always difficult and expensive to estimate these properties by experiments and computation. Therefore, prediction of these properties by ML methods has attracted the attention of many researchers. In general, the thermal conductivity of a specific electrolyte solution is a function of temperature, concentration, and pressure. The traditional methods are usually suggested for a specific electrolyte solution and a limited range of temperatures and concentrations. Moreover, it is relatively difficult to characterize the thermal conductivity in the traditional way due to the long-range electrostatic effect. In order to overcome the aforementioned shortcomings, Eslamloueyan et al. [86] considered temperature, concentration, molecular weight, and the sum of electrons of the cation and anion as inputs, and trained a feed-forward multilayer perceptron neural network (FFMLPNN) model with only one hidden layer on 389 experimental data points to predict the thermal conductivity of an electrolyte solution at atmospheric pressure with mean square error (*MSE*) of 1.012×10^{-5}. Similarly, Hezave et al. [87] trained a same FFMLPNN model based on 209 data points from 21 different ionic liquids (ILs) to predict the thermal conductivity of ILs with mean square error (*MSE*) of 1.2×10^{-6}. The descriptors of their FFMLPNN model not only included temperature, molecular weight, but also added melting point and pressure as additional inputs. The above excellent works indicate that temperature, molecular weight, and pressure are significant descriptors for different types of electrolyte solution and the FFMLPNN is an appropriate model for prediction of the thermal conductivity.

As an important indicator to evaluate the performance of the ILs, electrical conductivity has attracted much attention. Generally, the electrical conductivity of ILs has a nonlinear behavior in terms of temperature. Therefore, it is difficult to develop a simple model to accurately predict the electrical conductivity in a wide range of temperature. In the prediction of electrical conductivity by ML methods, the nonlinear model such as LSSVM and FFMLPNN model often perform great superiority. For example, Gharagheizi et al. [73] first computed a series of molecular descriptors for 54 ILs collected form literatures from their chemical structure of anions and cations. Then, SS strategy (as described in section 3.1) and the LSSVM model are used for selecting the optimal subset of features containing molecular descriptors and temperature, and developing a nonlinear model by training on 783 samples. The LSSVM model ultimately obtained a low *AARD* of less than 1.9% by testing on 97 experimental data. In order to predict the electrical conductivity of ternary mixture ILs, Hezave et al. developed a FFMLPNN model [88] with only one hidden layer training on 78 data points. The melting point, the compositions, the molecular weight, and the temperature of the ternary system have been utilized as inputs. The model was then validated using 26 data points (test data) which indicated the good interpolative ability of the trained network with *AARD* of 1.44. Furtherly, mole percent of each component were selected as descriptors by Hosseinzadeh et al. [89] in addition to temperature, melting point, molecular weight of the compounds when a LSSVM model was employed to predict the electrical conductivity of ternary mixture ILs at various temperatures and atmospheric pressure. The LSSVM model was built on 179 samples and achieved R^2 of 0.999. By relevancy analysis, it can be concluded that the average melting point has the greatest impact on the ternary mixture electrical conductivities.

Viscosity is a representation of the internal friction of fluid flow, its measurement is meaningful for any chemical process of ILs. The descriptors for different types of ILs are different for prediction of viscosity using ML method, but most of them contain molecular properties such as molecular mass and molecular weight. And the most used ML method is FFMLPNN. For instance, Lashkarblooki et al. [90] trained a FFMLPNN model with one hidden layer based on 547 data points to predict the viscosity of ternary mixture ILs (R^2:0.9998), the molecular mass, boiling point of three components and compositions of non-ILs components were selected as descriptors. Fatehi et al. applied the same method to predict the viscosity of binary mixture ILs in 2014 [91] and the pure ILs in 2017 [92], respectively. The superior prediction was obtained by using the same descriptors including molecular weight and structure characteristics of the ILs (the number of rings, the mass of each ring group and the mass of the central atom of the cation), as well as the system conditions such as temperature and pressure. An advantage of this method is that the proposed network for the prediction of viscosity of ILs was able to predict the ternary viscosity with some easily obtained inputs.

Apart from the properties mentioned above, the density of ILs is one of the most significant physical properties required in rational design of ILs used as promising liquid electrolytes. By training the LSSVM model on 405 data points, Hemmati-Sarapardeh et al. [93] predicted the density of binary mixture ILs at different pressures by regarding pressure, temperature, and mole fraction as inputs (R^2: 0.9866). Moreover, in order to understand the tendencies in the ion transfer at the electrolyte/electrode interface, Ishikawa et al. [94] predicted the coordination energy of the ions to the solvent using multiple linear regression (MLR), LASSO, and ES with linear regression (ES-LR). In addition to basic element descriptors such as ionic radius, electronegativity and atomic charge, some computational descriptors calculated by DFT are exploited including energies of the highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO), dipole moment, natural bond orbital (NBO) charge of the O atom that coordinates to the metal ion, total energy and total dipole moment. By ES with Gaussian process, the radius of the metal ion, and the NBO charge of the O atom are considered to be the two most important parameters that affect the coordination energy, and good prediction accuracy for coordination energy of 0.016 eV was obtained.

In conclusion, temperature, molecular concentration and composition are commonly concerned descriptors in the prediction of different properties of liquid electrolytes. Because there is often a complex nonlinear relationship between the descriptors and the properties, SVM and ANN have been widely used and have achieved good results. However, limited by the small data, the ANN model used in above works only contains a hidden layer to simulate the non-linear relationships. With the increase of sample size, ANN is expected to show better prediction performance. Moreover, some methods corresponding to the small data will be discussed in section 5.1.

4.1.2. Property prediction for solid electrolytes

Solid electrolytes have attracted much attention in recent years. Compared with traditional batteries with liquid electrolytes, lithium batteries with solid electrolytes have better safety, higher energy density, and longer life [40]. Ionic conductivity, diffusion barrier, energy migration, shear moduli, and elastic constants are important indicators of whether a material can be used as a promising solid electrolyte. Many linear methods such as OLS, LR, PLS and nonlinear methods such as ANN and SVR are used to predict these properties of candidate compounds.

Many promising solid electrolytes have been reported, such as tavorite-type materials, olivine-type materials, LISICON-type materials, and perovskite-type materials. Their ionic transport properties have been widely predicted including diffusion barrier, migration energy and ionic conductivity. Generally, the microstructure and compositions of materials were always considered as important descriptors of ionic transport properties. In prediction of diffusion barrier and cohesive energy of an olivine-type Li*MX*O_4, Jalem et al. [38] developed a single hidden layer feed-forward network with two external attributes, which can increase the constraint for the model in contrast to the single response variable models. This model was built by training on 72 samples, which is characterized by some component descriptors such as the radius and electronegativity and structure descriptors such as lattice parameter and intra-atomic parameter. By comparing the results with PLS, the

feed-forward network model shown better prediction result for predicting diffusion barrier. Furtherly, this feed-forward network model by Jalem et al. [38] was also successfully applied to the prediction of migration energy of tavorite-type Li*MTO*$_4$*F* materials with the same descriptors [95]. Two works proved that there are some common descriptors derived by composition and structure no matter what the type of the material is. We can use them in different type of materials. Different from the above works, in which descriptors are gained mainly by theoretical calculating, some researchers introduce more experimental descriptors in the following narrative. For example, in order to find the optimal composition space of $Li_{3x}La_{2/3-x}TiO_3$, Beal et al. [64] used a high throughput physical vapor deposition system to synthesize 35 sample libraries and study the perovskite-type compounds. Limited by the small data, they utilized a RBFNN model which is more appropriate with smaller data sets than other network types over a wide range of parameter space (composition, thickness, deposition temperature, annealing temperature) to predict the total conductivity (buck conductivity and grain boundary conductivity). Additionally, by applying partition analysis, the composition, specifically the lithium content, and the annealing temperatures are revealed the important factors when producing a crystalline $Li_{3x}La_{2/3-x}TiO_3$ film. Moreover, Ibrahim and Johan [83] investigated the effects of chemical composition and temperature on the ionic conductivity of the polymer electrolyte system. Novel composite solid polymer electrolytes were synthesized successfully by solution-casting technique. They employed a Bayesian neural network (BNN) to predict the ionic conductivity of the polymer electrolyte system and revealed that the ionic conductivity of the polymer electrolyte system varies with different chemical compositions and temperatures. This was consistent with expert experience. Fujimura et al. [85] applied SVR to use theoretical data (diffusion coefficients at 1600 K, transition temperatures, average volume of disordered structures) and experimental data (conductivity measurements at different temperatures) in combination to predict low-temperature conductivities of the various compositions for LISICON-type. This method illustrates the potential for rational design of superior-ion conductors based on optimization of materials compositions through ML techniques. In the studies above, structure and composition descriptors are introduced in the work. It is interesting that Jalem et al. [38] introduced thermodynamic concept from literatures. It is natural to use process parameters as the descriptors when the work involved experimentation.

Actually, many works took multiple types of materials into consideration. For example, Sendek et al. established a ML model to predict the ionic conductivity of solid electrolytes which contains LISICON-type, NASICON-type, garnet-type and other types of compounds [40,72]. In Refs. [72], 19 structured descriptors based on simple crystallographic and chemical relationships were computed, and the linear relationship between the lithium ionic conductivity of electrolytes and these descriptors was found by using OLS and LWLS techniques. Furtherly, in order to obtain a more accurate prediction model, Sendek et al. added two additional samples into the train set, and computed one more feature [40]. Comparing with other linear models, LR is more suitable for small sample, LR model is established to predict ionic conductivity of solid electrolytes by training on 40 samples and achieved a *CVMR* of 10% [40]. Moreover, based on this model, they screened out 20 possible fast ionic conductor materials. In Ref. [96], PLS and gradient boosting regression (GBR) were used by Nakayama et al. to predict the migration energy of Li- and Zn-containing oxide (Li–Zn-X-O) compounds. The GBR technique provided better prediction than the PLS regression by regarding the electronegativity and interatomic distances as descriptors. In addition, in order to screen solid electrolyte materials that can suppress dendrite initiation in contact with a Li metal anode, Ahmad et al. [42] proposed a CGCNN model to predict the shear and bulk moduli of the crystalline solid electrolyte materials with *RMSE* of 0.1268 and 0.1013 (log (GPa)). Meanwhile, they also used AdaBoost, LASSO, and Bayesian ridge to predict the elastic constants of the cubic crystal compounds, yielding a total R^2 of 0.98, 0.85, and 0.69, respectively. As a result, over 20 mechanically anisotropic interfaces between Li metal and four solid electrolytes which can be used to suppress dendrite growth were screened out. The CGCNN is the multigraph representation of the crystal structure which encodes the atomic information and bonding interactions between atoms, and is expected to demonstrate good universality in the property prediction of mixed types of solid electrolyte materials, as it can directly understand materials properties from the connection of atoms in a crystal.

In summary, although there are differences among various different algorithms, most ML methods have been able to achieve excellent prediction performance in terms of acceptable temporal and spatial complexity. It is obvious that in the property prediction of solid electrolyte materials, a good prediction for Li-ion conductivity was achieved by using composition-derived descriptors such as ion radius, electronegativity, structure-derived descriptors such as bond lengths, bond angles whether for the specific type of materials or mixed types of materials. However, descriptors need to be analyzed in more detail for the specific problem.

4.1.3. Property prediction of electrode materials

Looking for suitable electrode materials with long-term stability is an important requirement for developing long-life lithium rechargeable batteries. The properties of electrode materials such as volume changes in lithium ion batteries, voltage, redox potential, capacity, layer thickness and so on are of wide concern in the property prediction of electrodes. Since the crystal system has a major effect on the physical and chemical properties of rechargeable battery electrodes, it is necessary to predict it in order to estimate other properties. By selecting space group, formation energy, energy above hull, band gap, number of sites, density, and volume of unit cell as descriptors, Shandiz et al. [78] applied five classification algorithms including ANN, SVM, K-nearest neighbors (KNN), RF, and ERT to predict the crystal systems of silicate-based cathodes with Li–Si-(Mn, Fe, Co)–O compositions, which include monoclinic, orthorhombic, and triclinic. Results show that ensemble methods including RF and ERT provided the highest accuracy of prediction among other classification methods, and the volume of crystal and number of sites contributed the most to determine the type of crystal system.

For the cathode of a rechargeable battery, the main properties of concern relate to volume changes, voltage and redox potentials. One efficient way to extend the cycle life is to design cathodes with small volume changes [97]. For this reason, Xiao et al. [77] developed the QSAR formulations of cathode volume changes of spinel structure LiX_2O_4 and layered-structure LiXO_2 by PLS model. As described in section 3.2, they found that the radius of X^{4+} ion, and the *X* octahedron descriptors make major contributions to the volume changes of cathode. The conclusion is expectedly applied to the virtual screening and combinatorial design of low-strain cathode materials for lithium ion batteries. High voltage cathode materials are also required for designing high energy density rechargeable batteries, and some researchers show that electronegativity of atoms in intercalation compounds has significant effect on the voltage. Thus, Sarkar et al. [98] predicted the voltage of different classes of lithium ion battery cathode materials by using Multi-Layer Perceptron (MLP) by selecting the electronegativity of central atom and the stronger electronegative elements as inputs. Due to the small data (31 samples), the MLP model is fixed with only one hidden. The prediction result was in good agreement with the known experimental results and DFT based simulation results. The main challenge in voltage prediction is lack of large data set, access to which is known to improve the accuracy of the ANN model. Furtherly, Joshi et al. [99] developed a tool based on ML models to predict voltages of electrode materials utilizing the features derived from the chemical properties of compounds and the properties of their elemental constituents including electronegativity. In this work, the training data were extracted from the MP database containing a total of 4250 data samples for 3580 intercalation-based electrode materials. Three ML algorithms including DNN, SVR, and kernel RR were applied, among which DNN performed

the best, and 5000 candidate electrode materials for Na- and K-ion batteries were identified. Moreover, to identify promising organic cathode materials, Allam et al. [100] developed an ANN method to establish the quantitative molecular structure-redox potentials relationships. Both the electronic properties and structural information, such as the numbers of oxygen atoms, lithium atoms, boron atoms, carbon atoms, hydrogen atoms, and aromatic rings, are considered as input variables for ML models. The ANN demonstrated a capability for accurately estimating the redox potentials with a R^2 of 0.9618. From the contribution analysis, the electron affinity is the greatest contributor to the redox potential, followed by the number of oxygen atoms, HOMO-LUMO gap, the number of lithium atoms, LUMO, and HOMO, in order.

It is a challenge to build the link between the preparation process of materials and the macroscopic properties of materials. In order to search for the interdependence between electrode characteristics and manufacturing parameters, Cunha et al. [101] used DT, SVM, and DNN methods to predict the to predict the mass load and porosity of electrodes using a series of slurry manufacturing parameters (active material mass content, viscosity and solid-to-liquid ratio). The DNN method obtained the same precision as SVM, but it could not easily analyze the influence of manufacturing parameters on the mass load and porosity of electrodes. As a result, several trends linking the electrode mass loading and porosity to the slurry characteristics were disclosed by the SVM model. Moreover, Eremin et al. [102] applied RR to find the structure-property relationship of configurational space in the $LiNiO_2$ (LNO) and $LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$ (NCA) cathode materials. By applying the sequential backward selection algorithm, it was concluded that the topology of Li layers and relative disposition of Li ions and dopants have the most significant effect on the energy balance.

There are highly non-linear relationships available in batteries devices for which there are no simple and accurate physical models available to predict the underlying complex phenomena, such as charge-discharge behavior and anodic oxide process. For the anode of a rechargeable battery, ANN was deployed to predict the charge and discharge capacity of lithium-ion batteries containing CoO anodes [103]. The best fit values corresponded to an error factor of less than 1% when the cycles of charge and discharge were used as input factors. The result showed that this model could predict the cycle life of a Li-ion battery with CoO anode and could be extended to a variety of alternate anodes. Michal et al. [104] investigated the influence of individual factors acting during the anodic oxide process and developed an ANN model to predict the layer thickness of anodic aluminum with reliability of 72.53% with inputs consists of composition of electrolyte and the individual operating conditions.

To sum up, the combination of appropriate descriptors and proper ML methods leads to the successful prediction. In these researches, some Embedded FS methods such as RF and ERT, and some correlation analysis methods such as sequential backward selection algorithm and contribution analysis are employed to obtain the most important factors affecting the properties, which has great significance to help materials experts to rationally design new materials. The selection of ML methods depends on the inherent characteristics of sample data. Some linear models such as PLS, LR are fast to model and easy to interpret. However, the relationship between the selected descriptor and the target property of rechargeable battery materials is complex and nonlinear in most cases. It is well accepted that ANN is a typical nonlinear learning model which simulates human brain procedures, but a great deal of data is always needed. In order to find the optimal prediction ML model of the corresponding problem, it is an effective solution strategy to use multiple ML algorithms for modeling.

4.2. Discovery of novel rechargeable battery materials

The purpose of discovery of novel materials is to find candidate materials with superior properties that can be synthesized so that researchers in the laboratory can perform targeted explorations and synthetic experiments. There are two key issues for the synthesis of new materials [105], one is that which chemical components likely form novel materials, another is that which structures likely match the composition and properties of novel materials. In rechargeable battery materials, various descriptors derived from composition and structure information can provide excellent support for the prediction of all kinds of properties as mentioned in section 4.1. Therefore, once the composition and structure which are likely to be able to synthesize new compounds are identified, it is highly possible to discover novel candidate materials with superior properties by coupling with ML models for the property prediction.

Hautier et al. [106] devoted to constructing a probabilistic model of components and structures by learning the potential matching patterns of components and structures from a large-scale data containing a large number of materials that have been proven to be synthesized. Then the mathematical principle of probability statistics is used to calculate the posterior probability of the synthesis of a new compound from a specific component combination and a specific structure. Finally, the composition and structure of candidate new compounds will be screen out through the probability threshold. However, the number of candidate new compounds is still a lot because of the enormous combination space of compositions and structures, and these candidate compounds still need to be verified by the FP calculation such as DFT, so it still takes a lot of time even if the calculations is achievable. In Ref. [106], Hautier et al. successfully screened 1126 candidate compounds from 2211 A-B-O systems by this method, and finally discovered 209 new ternary compounds with a limited computational budget, but the result was verified by 5546 DFT calculations. As shown in Table 4, many studies have successfully developed ML models only for component or structure prediction in the case of limiting structural search space, compositional search space or both in rechargeable battery materials, which are comparable to the error bars of DFT calculation against to experiments and sometimes lower. If these models are used to further screen the candidate compounds before the DFT calculation, it will be to further narrow the verification space and

Table 4
Application of ML in the discovery of new rechargeable battery materials.

Application description	Reference	ML method	Achievement
Finding nature's missing ternary oxide compounds	[106]	Bayesian	Finding 209 new compounds
Obtaining qualitatively useful guidance for a wide range of perovskite oxide stability	[107]	ERT	Predicting 15 new perovskite compounds accurately
Discovering elpasolites (with stoichiometry ABC_2D_6) crystals	[108]	KRR	90 unique structures were identified
Predicting the thermodynamic stability of solids	[109]	RR, RF, ERT and ANN	Speeding up considerably (by at least a factor of 5) high-throughput DFT calculations
Screening new materials in an unconstrained composition space	[110]	Bayesian	Screening out 4500 new stable materials
Predicting the formation energies by Voronoi tessellations	[111]	RF	*Obtaining MAE* of 80 meV/atom
Developing a tool for crystal structure prediction	[114]	PLS	Predicting the formation energy of 114 structures of binary alloy with 90% or higher of precision
Developing a tool for molecular structure prediction	[116]	BO	Reducing the number of searching trials required to find the global minimum structure by 30–40%
Discovery of new guanidinium ILs	[117]	ANN	Discovering six new guanidinium ILs

accelerate the discovery of new materials.

4.2.1. Component prediction

In the discovery of novel materials, thermodynamic stability is the essential concept which can characterize how difficulty compounds decompose (even in infinite time) into different phases or compounds [50], and it is one of goals in the discovery of novel materials to find compounds with high thermodynamic stability. The formation energy and the distance to the convex hull can quantitatively reflect the thermodynamic stability, and the latter is more accurate than that of the former and more complex to calculate. In most cases, the researchers of rechargeable battery materials realized the discovery of novel materials by constructing the QSPR between components or structures and the formation energy or the distance to the convex hull through ML technique.

Most of component predictions were carried out in the case of limiting the structural search space. For example, Li et al. [107] regarded 70 descriptors of component and structure as inputs of ML models for the prediction of thermodynamic phase stability of perovskite oxides. ML classification and regression models were successfully constructed based on the dataset of DFT-calculated energies above the convex hull of 1929 perovskite oxides. After 5-fold cross-validation, the optimal classification model is the ERT with *F1* of 0.881 ± 0.032 compared to LR, SVN, DT and ANN. The optimal regression model is KRR with *RMSE* of 28.5 ± 7.5 meV/atom compared to LR, DT, ERT and ANN. Finally, the methods accurately predicted the thermodynamic phase stability of 15 new perovskite compounds, and summarized 11 properties of elements which have great influence on the thermodynamic phase stability of perovskite compounds and will be used for guiding the future synthetic experiments of new perovskite compounds, such as number of unfilled valence orbitals, coefficient of thermal expansion and Mendeleev number. Elpasolite ($AlNaK_2F_6$) is a kind of potential low-temperature electrolyte materials. Faber et al. [108] generated an input vector x to represent the component and structure information of crystals and developed a KRR model with mean absolute error (MAE) of 0.1 eV/atom to predict formation energies of 2×10^6 pristine ABC_2D_6 elpasolite crystals. And x is a ($n \times$ 2-tuple) vector that encodes any stoichiometry within a given crystal prototype. For quaternary ($n = 4$) elpasolites, x is a ($4 \times$ 2-tuple) vector that each $x_{1\text{-}4}$ is consist of the row and column number of an element in the periodic table for, and $x_{1\text{-}4}$ is ordered by the Wyckoff sequence of the crystal. Experimental result shows that fluoride is best suited to fit the coordination of the D site, which lowers the formation energy whereas the opposite is found for carbon, and 90 unique stable stoichiometries were identified in 2×10^6 crystals. In addition, to evaluate the performance of all kinds of ML models more rationally, Schmidt et al. [109] performed a large scale benchmark to evaluate ML models for the prediction of the thermodynamic stability of solids by incorporating DFT. They constructed a dataset of DFT-calculated the distance to the convex hull of around 250000 cubic perovskite systems involving 64 elements, and the performance of RR, DT, RF, ERT and ANNs were evaluated by the benchmark. The result showed that the ERT trained on 20000 samples is the optimal model with *MAE* of 121 meV/atom on 230000 samples. They also surprisingly found that ML models could achieve superior performance by only using the position information of elements in the periodic as descriptors, and revealed the impact trend of different types of elements on the stability of cubic perovskite systems.

In order to eliminate the constrain of structural search space, compositional search space or both in the process of the discovery of novel materials, Meredig et al. [110] developed an approach that combined heuristic model based on physical mechanisms with ensembles of DTs method in 2014. The approach was used for predicting the formation energy of 1.6 million compositions for novel ternary compounds ($A_xB_yC_z$) with R^2 score of 0.9, and 8 new stable materials were verified by DFT calculation. The approach is structure-independent, that is, only component information is considered and no other input is required. However, a full DFT crystal structure search is necessary to validate 4500 compositions that were screened out, which consumes a lot of computing resources. And the final composition for DFT validation still need to be selected manually in combination with domain knowledge. Ward et al. [111] solved this problem from another perspective by extracting structural descriptors from Voronoi tessellations in 2017. These descriptors can represent different crystal structures, which can save the time in defining different descriptors for different structures. For a dataset of 435000 formation energies taken from OQMD, they successfully developed a RF model with *MAE* of 80 meV/atom in cross-validation by using these structural descriptors and atomic properties. The result showed that composition-dependent descriptors of elemental properties can provide enough information to construct a highly accurate ML model, but the prediction result for materials with large formation enthalpies. On the contrary, the descriptors from Voronoi tessellations can accurately identify the materials with large formation enthalpies, so the optimal strategy is to adopt both compositional and structural descriptors.

To sum up, it is clearly that ANNs were often used methods in component predictions, because ANNs can deal with more complex problems by making full use of large amounts of data. However, the performance of ANNs is susceptible to hyper-parameters and the scale of dataset, and it is rather hard and time consuming to find the optimal network configuration. In most of previous efforts, the selected ML models were ensemble models of DTs such as RF and ERT. There are two reasons why these models were so popular with researchers. One is that they can deal with more complex problems than other simple models such as RR and DT, and they can effectively avoid overfitting by integrating different DTs for the same problem or adding random factors. Another is that they need to set fewer hyper-parameters and the trained result is not susceptible to the variation of hyper-parameter values and the scale of dataset. Even so, the ability of ANNs to handle complex and diverse problems cannot be ignored, so how to balance the complexity and accuracy of ML models is an intractable issue and we will discuss in detail in section 5.2. In addition, the previous works also prove that the chemical and physical properties of elements or their positions in the periodic table and other component information can provide ML models with abundant training data, but structural descriptors still need to be fully considered in order to more comprehensively and accurately predict the formation energy and other properties of materials. Furthermore, it is important to develop more general descriptors for the component search space that can be divided into different subspace according to different structures.

4.2.2. Structure prediction

In 2006, Ceder et al. [112] highlighted the significance of data mining structure prediction (DMSP) combining data mining techniques and ab-initio methods. DMSP can overcome the disadvantages of structural search relying on limited heuristic rules based on physical mechanism in a large search space. Firstly, ML methods such as principal component analysis (PCA), ANNs, clustering schemes and so on are used to capture the underlying basic physical rules governing structural stability. Then, ab-initio methods are used to verify structures recommended by data mining. Fischer et al. [113] carried out a test on 3975 compounds appearing at least twice in the Pauling file database of binary metallic alloys, and proved that DMSP is very effective in predicting the true ground state, requiring the investigation of only five structures for a 90% chance of finding the true structure. At present, many structure prediction schemes using ML are combined with FP calculations. Furthermore, structure prediction is similar to component prediction, which is generally carried out in case of limiting compositional search space and is aimed at predicting the thermodynamic stability or properties of interest of candidate materials. A number of ML models have been developed for structure prediction, most of which have the potential to extended to any known structure prediction of any type of material.

There are two different schemes for the structure prediction by using ML. One is to select structures with high thermodynamic stability from a

pool of candidate structures by learning QSPR between structures and thermodynamic stability under the specific component space. For example, Curtarolo et al. [114] developed a tool for crystal structure prediction by combining dimension reduction algorithms and prediction models. Based on a library of ab-initio energies for 114 descriptors to represent different crystal structures in each of 55 binary metallic alloys, the dimension of structure space was reduced to 9 from 114 by using PCA, which can significantly shorten the DFT computation budget. They constructed a formation energy matrix based on the result of PCA, and used PLS to fit the formation energy of different structures. The experimental result showed that this method could successfully predict the formation energy of binary alloy with 90% or higher of precision. Goncalo et al. [115] developed a tool for molecular structure prediction using counter propagation neural networks (CPGNNs). In order to identify guanidinium salts with low mp values for ILs, they used CPGNN to construct QSPR between the mp values of guanidinium salts of four anionic families (Cl^-, Br^-, I^- *and* BPh_4^-) and the structural profile of guanidinium cations. A series of 92 molecular descriptors representing the structure of cations were used to generate input vectors for CPGNN, including 14 constitutional descriptors related to the number of bonds and number of specific atoms, 33 topological descriptors related to structural flexibility and symmetry and so on. The method has been proved to have completely acceptable predictive ability with a fivefold cross-validation procedure yielded R^2 of 0.742, and mp values of 6 new guanidinium salts consistent with experimental values were accurately predicted. Another scheme is to regard structure prediction as an optimization problem of finding the global optimal structure with high thermodynamic stability under a specific component space. Compared with the former scheme that performs preliminary screening before trails by calculating the thermodynamic stability of all candidate structures, the optimization algorithm can directly provide several candidate structures for trails. For example, Yamashita et al. [116] proposed a crystal structure prediction method based on the combination of Bayesian optimization (BO) and random search, and successfully applied it to known systems, including NaCl and Y_2Co_{17} systems. They exploited the fingerprint $F_{AB}(R)$ of Oganov and Valle [117] to describe different crystal structures, and used the fingerprint together with DFT-calculated total-energies as the input of BO. In this study, BO is used to find out the global minimum structure with a lower number of trials in large and complex systems. Implementing this scheme, the rock salt structure of the most stable of NaCl systems can be found in 800 structures only after 26 trails on average, and the most stable of Y_2Co_{17} can be found in 1000 structures only after 128 trails on average. Compared with random search, this method can reduce the average number of trails by more than 30%.

As mentioned above, each effort provides a complete workflow for both crystal structure prediction and molecular structure prediction, which made it possible to extend these methods to structure prediction of rechargeable battery materials. From these works, we can conclude two key issues that need to be considered for the structure prediction of different types of materials. One is how to encode different structures with different components. The available method is to represent the combination of different components and structures as discrete space vectors, or define descriptors to characterize and distinguish different structures. Another issue is how machine learning can be used to assess the thermodynamic stability of different structures or other properties of interest. The available method is to establish ML models for QSPR or to directly transform the structure prediction into the optimization problem. However, the structural prediction schemes mentioned above can only predict the known structures in the pool of candidate structures, but cannot predict unknown structures. It is still an urgent problem that how to develop methods incorporating domain knowledge of structure design summarized in experiments or implicit in heuristic rules for new-type structures prediction.

5. Challenges of machine learning in the application of rechargeable battery materials

The rechargeable battery material informatics database based on high-throughput calculations and experiments provides tremendous opportunities for ML in rechargeable battery materials. However, there still exist issues to be resolved such as the contradiction between high dimension and small sample, the complexity and accuracy of the ML model, as well as the ML results and domain expert knowledge. It is these contradictions that bring new challenges to the use of ML for rechargeable battery materials and it will no doubt be the subject of future work.

5.1. Contradiction and coordination between high dimension and small sample

Battery materials data are often characterized by having multiple sources (e.g., experimental data, computational data, production data, and literature data), being heterogeneous (e.g., structured, semi-structured, and unstructured data), and being small samples with high dimension (i.e., the dimension of data is much larger than the volume of data). Therefore, researchers using ML to study rechargeable battery materials may face conflict between the high dimension and the small size of the data [40,77,81]. To alleviate the impact of high-dimensional, small-size data on the predictive accuracy of models, the methods of dimensionality reduction (e.g., FS and extraction), sample augmentation, active learning, and ensemble learning can be employed during the ML process. The two most popular feature extraction methods are PCA [118] and linear discriminant analysis (LDA) [119], both of which have been successfully applied to various practical problems in materials design and discovery. In terms of sample augmentation, more and more published and shared data can be downloaded and acquired from open source websites. On the other hand, more samples can also be generated through generative models (e.g., autoencoder [120] and generative adversarial networks: GANs [121]). Some scholars have proposed virtual sample generation (VSG) methods [122–127], which can systematically produce virtual samples to fill in data gaps. Active learning is another method by which to construct a valid training set. Its purpose is to find effective samples by iterative sampling and to ensure that the model can obtain high accuracy with small sample data [128]. Gubernatis et al. [44] proposed a learning strategy based on an active learning framework to study small sample data in material studies. The learner constructs a certain query strategy to choose the most informative unlabeled samples actively, which are then assigned labels by the material experts. Sequentially, the newly labeled samples are combined with the original samples for training, so that the model can achieve higher accuracy when the size of a training set is small. Active learning has been successfully applied in the study of alloys [130,131] and ceramic materials [132–134], and thus is also expected to be applied to high-dimensional data learning of small samples of rechargeable battery materials. Ensemble learning accomplishes learning tasks through building and combining multiple weak learners, which can often achieve superior performance compared with a single learner when modeling the data with high dimension and small size [135]. For example, the self-sampling method proposed by Efron [136] can effectively solve not only the problem of small size and high dimension, but also the problem of imbalanced data by estimating the distribution of given data. Liu et al. [137] has researched the generalization ability, efficiency, and convenience of neural network ensembles and then used the proposed neural network ensemble methods to predict successfully the magnitude and time of earthquakes in mainland China. Therefore, the introduction of ensemble learning into rechargeable battery materials science could offer another effective solution to overcome the influence of the contradiction between high dimension and small sample in ML of rechargeable battery materials.

5.2. Conflict and compromise between complexity and accuracy of machine learning models

The original goal of ML is to extract interpretable knowledge from data and emphasize the interpretability when pursuing accuracy with an algorithm [135]. The primitive algorithms such as linear perceptron, DT, and nearest neighbors are typical masterpieces in this respect. In order to complete more complex learning tasks, multivariate linear and nonlinear models have been developed for ML. These ML algorithms focused on multivariate linear models, are mainly used to construct the linear relationships between multiple factors and targeted properties. They are simple to implement, and the learning results tend to be comprehensible. Therefore, these models are superior for constructing simple linear relationships between the structures and properties of rechargeable battery materials [40–72]. However, there often exists a complex nonlinear relationship between the microstructure and material properties of the battery material. This is due to the complex electrochemical behavior inside the rechargeable battery, which causes linear algorithms such as OLS and PLS to fail in constructing the nonlinear relationship between the microstructure of the battery material and its performance. Hence, nonlinear models such as ANN and SVR are widely used to predict performance of rechargeable battery materials because they can construct complex nonlinear relationships between various factors and target properties [85–87]. However, the researchers need to try repeatedly to optimize the hyperparameter configuration with trial and error to obtain optimal performance of models such as ANN and SVR, which is time-consuming and labor-intensive. To solve this problem, some hyperparameter optimization methods have been adopted to simplify the complex hyperparameter tuning process of a model. For example, Abdolhossein et al. [93] combined simulated annealing algorithm with LSSVM to predict the density of binary mixtures, which simplified the hyperparameter tuning process and achieved *RMSE* of 1.69%. However, they often ignore the complexity of the model selection process, which is still a challenging task in selecting suitable algorithms and models for battery material data. Although complex learning models have great ability to handle nonlinear relationships for rechargeable battery materials, they are less interpretable than linear models. For example, compared with a linear regression (e.g., OLS, PLS, LWLS) method, the learning result of a nonlinear regression model is a “black box”, which cannot provide multiple linear regression equations. In response to these problems, we can indirectly reduce the model complexity by reducing the usage complexity of the model and by improving the interpretability of the complex model. To reduce usage complexity, automated ML (Auto ML) attempts to reduce human intervention in model selection, optimization, and implementation using random search [138], evolutionary optimization [139], BO [140–142], meta-learning [143] (etc.), and builds optimal models automatically [142] for given data. Therefore, introducing Auto ML into the rechargeable battery materials field can make a model easy to construct and can reduce its usage complexity. For the interpretability of a complex model, rule extraction is one way to establish an interpretable mechanism of a “black box” ML model, the purpose of which is to express the implied knowledge in ML with a pattern easy to understand and to improve the interpretability of ML methods [144]. The rule extraction methods of ML can be divided into two types: model structure-based and model function-based [144]. The model structure-based rule extraction regards rule extraction as a search process, and maps the structure (e.g., network structure, weight, or support vector) of the trained neural network/SVM to a comprehensible if-then-else rule. The model function-based rule extraction is processed through some specific models with interpretability, such as DT, RF, and gradient boosting DT. The results of a tree learning model are then converted into a comprehensible if-then-else rule, which can reproduce the model functionally using the extracted rules (i.e., a set of rules that can replace the original model). Therefore, introducing rule extraction technology into battery material ML can make the ML easy to understand in terms of both model structure and prediction results.

5.3. Inconsistency and collaboration between learning results and domain expert knowledge

Data driven ML methods are widely used in recent studies for the prediction of new rechargeable battery materials properties and materials discovery. However, prior knowledge in specific domains is often not considered in them. The results learned by an ML method will sometimes conflict with domain expert knowledge. It is imperative to find an approach to combine ML with domain expert knowledge in the field of design and discovery of rechargeable battery materials. One effective way to resolve this issue is to incorporate expertise into the definition of problems, and then to integrate the domain knowledge of rechargeable battery material experts, including descriptor calculation, descriptor selection, etc. In addition, modeling by incorporating expert knowledge and ML methods is a promising research field, for which the typical algorithms include Bayesian network and fuzzy learning. A Bayesian network determines the network topology by combining training data with prior knowledge in the training process [145], while fuzzy learning integrates expert experience using a membership function [146]. For example, Andres and Moral [147] proposed an interactive approach integrating domain expert knowledge to identify the edges of a Bayesian network structure. This was to realize the active interaction between domain experts and the Bayesian network model, ultimately, to improve performance. Tang et al. [146] presented a fuzzy rule-based classification system into which was incorporated expert knowledge. Their experimental results shown that this method had significantly reduced classification ambiguity and improved classification accuracy. Designing new materials through both prior knowledge of rechargeable battery material experts and knowledge databases built from learning results is another capable approach. Martin et al. [148] proposed to combine domain expert knowledge and learning results to build a large knowledge database, which improved the reasoning ability of expert systems. We also incorporated the expert knowledge into FS when modeling materials with targeted properties [81]. Furthermore, introducing expert systems into ML in rechargeable battery materials science is expected to resolve the contradiction between learning results and domain expert knowledge.

While it will take some time to completely address all these challenges, ML has made progress in materials science, and data-driven methods will undoubtedly be a major area of rechargeable battery materials science research in the future.

6. Summary and prospects

Relevant studies have shown that ML has been widely used in the property prediction of rechargeable battery materials, especially for electrolyte and electrode materials, as well as the discovery of new materials. With the development of ML technology and the emergence of more novel problems in rechargeable battery material science, the scope of application of ML will gradually expand. Simultaneously, the application of ML in the rechargeable battery field also faces many severe challenges. It is a thorny problem to determine how to construct a large sample set with high quality. On the one hand, it can be solved by establishing large databases and high-throughput platforms. On the other hand, with the development of generative deep learning, it is possible to use ML to generate new samples. Moreover, active learning and semi-supervised learning are expected to help to solve this problem. In addition, because the applications of ML in the field of rechargeable battery materials involve different disciplines, we are having to consider relevant expert knowledge and the usability of a model in the process of model construction. The key to wider application of ML is to determine how to improve the traditional ML process to establish an accurate, efficient, easy-to-use, and explicable model. Finally, the ultimate purpose of the application of ML is to accelerate the discovery of new materials for use in rechargeable batteries. At present, most applications are still in the stage of establishing a property prediction model. The reverse design of

materials using ML allows more fully play to the initiative in the use of ML, so it can guide the final experiments in the future, thus saving resources and speeding up the discovery of new materials.

Data availability

The authors declare that all data supporting the finding of this study are available within the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Key R&D Program of China (Nos. 2017YFB0701502 and 2017YFB0701600) and the National Natural Science Foundation of China (Nos. 51622207, 11874254, U1630134), Shanghai Pujiang Program (No. 2019PJD016), Open Project of the State Key Laboratory of Advanced Special Steel, Shanghai University, China (No. SKLASS2018-01) and the Project of the State Key Laboratory of Advanced Special Steel, Shanghai University, China (No. SKLASS2019-Z023). We appreciate the High Performance Computing Center of Shanghai University, and Shanghai Engineering Research Center of Intelligent Computing System (No. 19DZ2252600) for providing the computing resources and technical support.

References

[1] Lindley David, Smart grids: the energy storage problem, Nature 463 (2010) 18–20, https://doi.org/10.1038/463018a.
[2] Z. Yang, J. Zhang, M.C.W. Kintner-Meyer, X. Lu, D. Choi, J.P. Lemmon, J. Liu, Electrochemical energy storage for green grid, Chem. Rev. 111 (2011) 3577–3613, https://doi.org/10.1021/cr100290v.
[3] China energy storage alliance. https://energystorage.org/industry-resource/china-energy-storage-alliance-cnesa/, 2020. (Accessed 26 March 2020).
[4] A. Manthiram, X. Yu, S. Wang, Lithium battery chemistries enabled by solid-state electrolytes, Nat. Rev. Mater. 2 (2017) 16103, https://doi.org/10.1038/natrevmats.2016.103.
[5] S. Shi, J. Gao, Y. Liu, Y. Zhao, Q. Wu, W. Ju, C. Ouyang, R. Xiao, Multi-scale computation methods: their applications in lithium-ion battery research and development, Chin. Phys. B 25 (2016), https://doi.org/10.1088/1674-1056/25/1/018212, 018212.
[6] G. Ceder, Y.M. Chiang, D.R. Sadoway, M.K. Aydinol, Y.I. Jang, B. Huang, Identification of cathode materials for lithium batteries guided by first-principles calculations, Nature 392 (1998) 694–696, https://doi.org/10.1038/33647.
[7] J. Yang, J.S. Tse, Li ion diffusion mechanisms in $LiFePO_4$: an ab initio molecular dynamics study, J. Phys. Chem. 115 (2011) 13045–13049, https://doi.org/10.1021/jp205057d.
[8] T. Husch, M. Korth, How to estimate solid-electrolyte-interphase features when screening electrolyte materials, Phys. Chem. Chem. Phys. 17 (2015) 22799–22808, https://doi.org/10.1039/c5cp03119b.
[9] V. Pande, V. Viswanathan, Descriptors for electrolyte-renormalized oxidative stability of solvents in lithium-ion batteries, J. Phys. Chem. Lett. 10 (2019) 7031–7036, https://doi.org/10.1021/acs.jpclett.9b02717.
[10] T. Fujie, N. Takenaka, Y. Suzuki, M. Nagaoka, Red Moon methodology compatible with quantum mechanics/molecular mechanics framework: application to solid electrolyte interphase film formation in lithium-ion battery system, J. Chem. Phys. 149 (2018), https://doi.org/10.1063/1.5034771, 044113.
[11] Y. Ma, S.H. Garofalini, Atomistic insights into the conversion reaction in iron fluoride: a dynamically adaptive force field approach, J. Am. Chem. Soc. 134 (2012) 8205–8211, https://doi.org/10.1021/ja301637c.
[12] D.A. Cogswell, M.Z. Bazant, Theory of coherent nucleation in phase-separating nanoparticles, Nano Lett. 13 (2013) 3036–3041, https://doi.org/10.1021/nl400497t.
[13] G. Guo, B. Long, B. Cheng, S. Zhou, P. Xu, B. Cao, Three-dimensional thermal finite element modeling of lithium-ion battery in thermal abuse application, J. Power Sources 195 (2010) 2393–2398, https://doi.org/10.1016/j.jpowsour.2009.10.090.
[14] A.F. Bower, P.R. Guduru, A simple finite element model of diffusion, finite deformation, plasticity and fracture in lithium ion insertion electrode materials, Model. Simulat. Mater. Sci. Eng. 20 (2012), 045004, https://doi.org/10.1088/0965-0393/20/4/045004.
[15] K. Feldman, S.R. Agnew, The materials genome initiative at the national science foundation: a status report after the first year of funded research, JOM 66 (2014) 336–344, https://doi.org/10.1007/s11837-014-0888-0.
[16] Widener Andrea, Materials genome initiative, Chem. Eng. News 91 (2013) 25–27, https://doi.org/10.1021/cen-09131-govpol1.
[17] G. Bergerhoff, R. Hundt, R. Sievers, I.D. Brown, The inorganic crystal structure data base, J. Chem. Inf. Comput. Sci. 23 (1983) 66–69, https://doi.org/10.1021/ci00038a003.
[18] F.H. Allen, The Cambridge Structural Database: a quarter of a million crystal structures and rising, Acta Crystallogr. Sect. B Struct. Sci. 58 (2002) 380–388, https://doi.org/10.1107/S0108768102003890.
[19] P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, T. Massalski, H. Okamoto, K. Osaki, A. Prince, H. Putz, S. Iwata, The pauling file, J. Alloys Compd. 367 (2004) 293–297. https://doi.org/10.4028/www.scientific.net/MSF.443-444.357.
[20] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: the Materials Project: a materials genome approach to accelerating materials innovation, Apl. Mater. 1 (2013), 011002, https://doi.org/10.1063/1.4812323.
[21] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L.W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, AFLOWLIB.ORG, A distributed materials properties repository from high-throughput ab initio calculations, Comput. Mater. Sci. 58 (2012) 227–235, https://doi.org/10.1016/j.commatsci.2012.02.002.
[22] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD), JOM 65 (2013) 1501–1509, https://doi.org/10.1007/s11837-013-0755-4.
[23] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R.S. Sanchez-Carrera, A. Gold-Parker, L. Vogt, A.M. Brockway, A. Aspuru-Guzik, The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid, J. Phys. Chem. Lett. 2 (2011) 2241–2251, https://doi.org/10.1016/j.commatsci.2012.02.002.
[24] C. Ortiz, O. Eriksson, M. Klintenberg, Data mining and accelerated electronic structure theory as a tool in the search for new functional materials, Comput. Mater. Sci. 44 (2009) 1042–1049, https://doi.org/10.1016/j.commatsci.2008.07.016.
[25] D.D. Landis, J.S. Hummelshoj, S. Nestorov, J. Greeley, M. Dulak, T. Bligaard, J.K. Norskov, K.W. Jacobsen, The computational materials repository, Comput. Sci. Eng. 14 (2012) 51–57, https://doi.org/10.1109/MCSE.2012.16.
[26] C. Draxl, M. Scheffler, NOMAD: the FAIR concept for big data-driven materials science, MRS Bull. 43 (2018) 676–682, https://doi.org/10.1557/mrs.2018.208.
[27] P.J. Linstrom, W.G. Mallard, The NIST chemistry WebBook: a chemical data resource on the internet, J. Chem. Eng. Data 46 (2001) 1059–1063, https://doi.org/10.1021/je000236i.
[28] S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design, Nat. Mater. 12 (2013) 191–201, https://doi.org/10.1038/NMAT3568.
[29] D.M.D.O. Zapiain, E. Popova, S.R. Kalidindi, Prediction of microscale plastic strain rate fields in two-phase composites subjected to an arbitrary macroscale strain rate using the materials knowledge system framework, Acta Mater. 141 (2017) 230–240, https://doi.org/10.1016/j.actamat.2017.09.016.
[30] J.G. Michopoulos, J.C. Hermanson, A. Iliopoulos, S.G. Lambrakos, T. Furukawa, On the constitutive response characterization for composite materials via data-driven design optimization, in: ASME International Design Engineering Technical Conferences, 2011, https://doi.org/10.1115/DETC2011-47740. Washington D.C, America, August 28-31.
[31] J.G. Michopoulos, J.C. Hermanson, A. Iliopoulos, S.G. Lambrakos, T. Furukawa, Data-driven design optimization for composite material characterization, J. Comput. Inf. Sci. Eng. 11 (2011), https://doi.org/10.1115/1.3595561, 021009.
[32] A. Agrawal, A. Choudhary, An online tool for predicting fatigue strength of steel alloys based on ensemble data mining, Int. J. Fatig. 113 (2018) 389–400, https://doi.org/10.1016/j.ijfatigue.2018.04.017.
[33] S. Srinivasan, S.R. Broderick, R.F. Zhang, A. Mishra, S.B. Sinnott, S.K. Saxena, J.M. LeBeau, K. Rajan, Mapping chemical selection pathways for designing multicomponent alloys: an informatics framework for materials design, Sci. Rep. 5 (2015) 17960, https://doi.org/10.1038/srep17960.
[34] J.H. Crews, R. Smith, K.M. Pender, J.C. Hannen, G. Buckner, Data-driven techniques to estimate parameters in the homogenized energy model for shape memory alloys, J. Intell. Mater. Syst. Struct. 23 (2012) 1897–1920, https://doi.org/10.1177/1045389X12453965.
[35] E. Ras, G. Rothenberg, Heterogeneous catalyst discovery using 21st century tools: a tutorial, RSC Adv. 4 (2014) 5963–5974, https://doi.org/10.1039/C3RA45852K.
[36] R.S. Gebhardt, P. Du, O. Wodo, B. Ganapathysubramanian, A data-driven identification of morphological features influencing the fill factor and efficiency of organic photovoltaic devices, Comput. Mater. Sci. 129 (2017) 220–225, https://doi.org/10.1016/j.commatsci.2016.12.020.
[37] L. Ghadbeigi, T.D. Sparks, J.K. Harada, B.R. Lettiere, Data-mining approach for battery materials, in: IEEE Conference on Technologies for Sustainability, 2015. Ogden, USA, July 30-August 01.
[38] R. Jalem, M. Nakayama, T. Kasuga, An efficient rule-based screening approach for discovering fast lithium ion conductors using density functional theory and artificial neural networks, J. Mater. Chem. 2 (2014) 720–734, https://doi.org/10.1039/C3TA13235H.
[39] R. Jalem, K. Kanamori, I. Takeuchi, M. Nakayama, H. Yamasaki, T. Saito, Bayesian-driven first-principles calculations for accelerating exploration of fast ion

conductors for rechargeable battery application, Sci. Rep. 8 (2018) 5845, https://doi.org/10.1038/s41598-018-23852-y.
[40] A.D. Sendek, Q. Yang, E.D. Cubuk, K.A.N. Duerloo, Y. Cui, E.J. Reed, Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials, Energy Environ. Sci. 10 (2017) 306–320, https://doi.org/10.1039/C6EE02697D.
[41] A.D. Sendek, E.D. Cubuk, E.R. Antoniuk, G. Cheon, Y. Cui, E.J. Reed, Machine learning-assisted discovery of solid Li-ion conducting materials, Chem. Mater. 31 (2019) 342–352, https://doi.org/10.1021/acs.chemmater.8b03272.
[42] Z. Ahmad, T. Xie, C. Maheshwari, J.C. Grossman, V. Viswanathan, Machine learning enabled computational screening of inorganic solid electrolytes for suppression of dendrite formation in lithium metal anodes, ACS Cent. Sci. 4 (2018) 996–1006, https://doi.org/10.1021/acscentsci.8b00229.
[43] A. Bhowmik, I.E. Castelli, J.M. Garcia-Lastra, P.B. Jorgensenb, O. Winthe, T. Vegge, A perspective on inverse design of battery interphases using multi-scale modelling, experiments and generative deep learning, Energy Storage Mater. 21 (2019) 446–456, https://doi.org/10.1016/j.ensm.2019.06.011.
[44] J.E. Gubernatis, T. Lookman, Machine learning in materials design and discovery: examples from the present and suggestions for the future, Phys. Rev. Mater. 2 (2018) 120301, https://doi.org/10.1103/PhysRevMaterials.2.120301.
[45] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, npj Comput. Mater. 3 (2017) 54, https://doi.org/10.1038/s41524-017-0056-5.
[46] Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning, J. Materiom. 3 (2017) 159–177, https://doi.org/10.1016/j.jmat.2017.08.002.
[47] A. Jain, G. Hautier, S.P. Ong, K. Persson, New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships, J. Mater. Res. 31 (2016) 1–18, https://doi.org/10.1557/jmr.2016.80.
[48] A. Agrawal, A. Choudhary, Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science, Apl. Mater. 4 (2016), https://doi.org/10.1063/1.4946894, 053208.
[49] J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, B. Meredig, Materials science with large-scale data and informatics: unlocking new opportunities, MRS Bull. 41 (2016) 399–409, https://doi.org/10.1557/mrs.2016.93.
[50] J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, npj Comput. Mater. 5 (2019) 83, https://doi.org/10.1038/s41524-019-0221-0.
[51] W. Lu, R. Xiao, J. Yang, H. Li, W. Zhang, Data mining-aided materials discovery and optimization, J. Materiom. 3 (2017) 191–201, https://doi.org/10.1016/j.jmat.2017.08.003.
[52] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, S.P. Ong, A critical review of machine learning of energy materials, Adv. Energy Mater. 10 (2020) 1903242, https://doi.org/10.1002/aenm.201903242.
[53] G.H. Gu, J. Noh, I. Kim, Y. Jung, Machine learning for renewable energy materials, J. Mater. Chem. 7 (2019) 17096–17117, https://doi.org/10.1039/c9ta02356a.
[54] D. Krishnamurthy, H. Weiland, A.B. Farimani, E. Antono, J. Green, V. Viswanathan, Machine learning based approaches to accelerate energy materials discovery and optimization, ACS Energy Lett. 4 (2019) 187–191, https://doi.org/10.1021/acsenergylett.8b02278.
[55] H. Wang, Y. Ji, Y. Li, Simulation and design of energy materials accelerated by machine learning, Wiley Interdiscipl. Rev.: Comput. Mol. Sci. 5 (2019) e1421, https://doi.org/10.1002/wcms.1421.
[56] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Posts and Telecom Press, 2003.
[57] G. Briscoe, T. Caelli, Compendium of Machine Learning I: A Symbolic Learning, Greenwood Publishing Group Inc, 1996.
[58] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (1958) 386–408, https://doi.org/10.1037/h0042519.
[59] Vladimir Vapnik, Statistical learning theory, Wiley-Interscience, http://citeseer.ist.psu.edu/showciting?cid=1633694, 1998.
[60] The science of caring. http://www.sas.com/en_us/home.html, 2020. (Accessed 26 March 2020).
[61] On machine learning and the next wave of innovation. http://www.linkedin.com/pulse/machine-learning-next-wave-innovation-eric-miley-mba, 2020. (Accessed 26 March 2020).
[62] A.J. Salkind, C. Fennie, P. Singh, T. Atwater, D.E. Reisner, Determination of state-of-charge and state-of-health of batteries by fuzzy logic methodology, J. Power Sources 80 (1999) 293–300, https://doi.org/10.1016/S0378-7753(99)00079-8.
[63] T.M. Mitchell, Machine learning and data mining, Commun. ACM 42 (1999) 30–36, https://doi.org/10.1145/319382.319388.
[64] M.S. Beal, B.E. Hayden, T.L. Gall, C.E. Lee, X. Lu, M. Mirsaneh, C. Mormiche, D. Pasero, D.C.A. Smith, A. Weld, C. Yada, S. Yokoishi, High throughput methodology for synthesis, screening, and optimization of solid state lithium ion electrolytes, ACS Comb. Sci. 13 (2011) 375–381, https://doi.org/10.1021/co100075f.
[65] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate, and interpretable prediction of material properties, Phys. Rev. Lett. 120 (2018) 145301, https://doi.org/10.1103/PhysRevLett.120.145301.
[66] T. Xie, J.C. Grossman, Hierarchical visualization of materials space with graph convolutional neural networks, J. Chem. Phys. 149 (2018) 147111, https://doi.org/10.1063/1.5047803.
[67] S.K. Kauwe, T.D. Rhone, T.D. Sparks, Data-driven studies of Li-Ion-Battery materials, Crystals 9 (2019) 54, https://doi.org/10.3390/cryst9010054.
[68] K. Roy, S. Kar, R.N. Das, A Primer on QSAR/QSPR Modeling: Fundamental Concepts, Springer, 2015.
[69] O. Uncu, I.B. Turkşen, A novel feature selection approach: combining feature wrappers and filters, Inf. Sci. 177 (2007) 449–466, https://doi.org/10.1016/j.ins.2006.03.022.
[70] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (2014) 16–28, https://doi.org/10.1016/j.compeleceng.2013.11.024.
[71] T.T. Erguzel, C. Tas, M. Cebi, A wrapper-based approach for feature selection and classification of major depressive disorder-bipolar disorders, Comput. Biol. Med. 64 (2015) 127–137, https://doi.org/10.1016/j.compbiomed.2015.06.021.
[72] A. Krishnapriyan, A. Sendek, Data-driven discovery and design of superionic lithium conductors for high performance lithium ion batteries. https://pdfs.semanticscholar.org/86e8/86a0105a91ae09fbcc39f9265608a9ae6070.pdf?_ga=2.171637795.283541546.1585472574-2072556946.1583779107, 2020. (Accessed 29 March 2020).
[73] F. Gharagheizi, M. Sattari, P. Ilani-Kashkouli, A. Mohammadi, D. Ramjugernath, D. Richon, A “non-linear” quantitative structure-property relationship for the prediction of electrical conductivity of ionic liquids, Chem. Eng. Sci. 101 (2013) 478–485, https://doi.org/10.1016/j.ces.2013.07.007.
[74] H. Wu, A. Lorenson, B. Anderson, L. Witteman, H. Wu, B. Meredig, D. Morgan, Robust FCC solute diffusion predictions from ab-initio machine learning methods, Comput. Mater. Sci. 134 (2017) 160–165, https://doi.org/10.1016/j.commatsci.2017.03.052.
[75] R. Genuer, J. Poggi, C. Tuleau-Malot, Variable selection using random forests, Pattern Recogn. Lett. 31 (2010) 2225–2236, https://doi.org/10.1016/j.patrec.2010.03.014.
[76] V.F. Rodriguez-Galiano, J.A. Luque-Espinar, M. Chica-Olmo, M.P. Mendes, Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods, Sci. Total Environ. 624 (2018) 661–672, https://doi.org/10.1016/j.scitotenv.2017.12.152.
[77] X. Wang, R. Xiao, H. Li, L. Chen, Quantitative structure-property relationship study of cathode volume changes in lithium ion batteries using ab-initio and partial least squares analysis, J. Materiom. 3 (2017) 178–183, https://doi.org/10.1016/j.jmat.2017.02.002.
[78] M.A. Shandiz, R. Gauvin, Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries, Comput. Mater. Sci. 117 (2016) 270–278, https://doi.org/10.1016/j.commatsci.2016.02.021.
[79] Y. Li, C. Zou, M. Berecibar, E. Nanini-Maury, J.C.W. Chan, P.V.D. Bossche, J.V. Mierlo, N. Omar, Random forest regression for online capacity estimation of lithium-ion batteries, Appl. Energy 232 (2018) 197–210, https://doi.org/10.1016/j.apenergy.2018.09.182.
[80] H.H. Hsu, C.W. Hsieh, M.D. Lu, Hybrid feature selection by combining filters and wrappers, Expert Syst. Appl. 38 (2011) 8144–8150, https://doi.org/10.1016/j.eswa.2010.12.156.
[81] Y. Liu, J. Wu, M. Avdeev, S. Shi, Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties, Adv. Theory Simul. 3 (2020) 1900215, https://doi.org/10.1002/adts.201900215.
[82] X. Liu, W. Lu, C. Peng, Q. Su, J. Guo, Two semi-empirical approaches for the prediction of oxide ionic conductivities in ABO_3 perovskites, Comput. Mater. Sci. 46 (2009) 860–868, https://doi.org/10.1016/j.commatsci.2009.04.047.
[83] S. Ibrahim, M.R. Johan, Conductivity, thermal and neural network model nanocomposite solid polymer electrolyte system (PEO-$LiPF_6$-EC-CNT), Int. J. Electrochem. Sci. 6 (2011) 5565–5587.
[84] C. Li, Y. Thing, Y. Zeng, C. Wang, P. Wu, Prediction of lattice constant in perovskites of $GdFeO_3$ structure, J. Phys. Chem. Solid. 64 (2003) 2147–2156, https://doi.org/10.1016/S0022-3697(03)00209-9.
[85] K. Fujimura, A. Seko, Y. Koyama, A. Kuwabara, I. Kishida, K. Shitara, C.A.J. Fisher, H. Moriwake, I. Tanaka, Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms, Adv. Energy Mater. 3 (2013) 980–985, https://doi.org/10.1002/aenm.201300060.
[86] R. Eslamloueyan, M.H. Khademi, S. Mazinani, Using a multilayer perceptron network for thermal conductivity prediction of aqueous electrolyte solutions, Ind. Eng. Chem. Res. 50 (2011) 4050–4056, https://doi.org/10.1021/ie101513z.
[87] A.Z. Hezave, S. Raeissi, M. Lashkarbolooki, Estimation of thermal conductivity of ionic liquids using a perceptron neural network, Ind. Eng. Chem. Res. 51 (2012) 9886–9893, https://doi.org/10.1021/ie202681b.
[88] A.Z. Hezave, M. Lashkarbolooki, S. Raeissi, Using artificial neural network to predict the ternary electrical conductivity of ionic liquid systems, Fluid Phase Equil. 314 (2012) 128–133, https://doi.org/10.1016/j.fluid.2011.10.028.
[89] M. Hosseinzadeh, A. Hemmati-Sarapardeh, F. Ameli, F. Naderi, M. Dastgahi, A computational intelligence scheme for estimating electrical conductivity of ternary mixtures containing ionic liquids, J. Mol. Liq. 221 (2016) 624–632, https://doi.org/10.1016/j.molliq.2016.05.059.
[90] M. Lashkarblooki, A.Z. Hezave, A.M. Al-Ajmi, S. Ayatollahi, Viscosity prediction of ternary mixtures containing ILs using multi-layer perceptron artificial neural network, Fluid Phase Equil. 326 (2012) 15–20, https://doi.org/10.1016/j.fluid.2012.04.017.
[91] M. Fatehi, S. Raeissi, D. Mowla, Estimation of viscosity of binary mixtures of ionic liquids and solvents using an artificial neural network based on the structure groups of the ionic liquid, Fluid Phase Equil. 364 (2014) 88–94, https://doi.org/10.1016/j.fluid.2013.11.041.

[92] M. Fatehi, S. Raeissi, D. Mowla, Estimation of viscosities of pure ionic liquids using an artificial neural network based on only structural characteristics, J. Mol. Liq. 227 (2017) 309–317, https://doi.org/10.1016/j.molliq.2016.11.133.

[93] A. Hemmati-Sarapardeh, M. Tashakkori, M. Hosseinzadeh, A. Mozafari, S. Hajirezaie, On the evaluation of density of ionic liquid binary mixtures: modeling and data assessment, J. Mol. Liq. 222 (2016) 745–751, https://doi.org/10.1016/j.molliq.2016.07.068.

[94] A. Ishikawa, K. Sodeyama, Y. Igarashi, T. Nakayama, Y. Tateyama, M. Okada, Machine learning prediction of coordination energies for alkali group elements in battery electrolyte solvents, Phys. Chem. Chem. Phys. 21 (2019) 26399–26405, https://doi.org/10.1039/c9cp03679b.

[95] R. Jalem, M. Kimura, M. Nakayama, T. Kasuga, Informatics-Aided density functional theory study on the Li ion transport of tavorite-type $LiMTO_4F$ (M^{3+}-T^{5+}, M^{2+}-T^{6+}), J. Chem. Inf. Model. 55 (2015) 1158–1168, https://doi.org/10.1021/ci500752n.

[96] M. Nakayama, K. Kanamori, K. Nakano, R. Jalem, I. Takeuchi, H. Yamasaki, Data-driven materials exploration for Li-ion conductive ceramics by exhaustive and informatics-aided computations, Chem. Rec. 19 (2019) 771–778, https://doi.org/10.1002/tcr.201800129.

[97] Z. Chen, L. Christensen, J.R. Dahn, Large-volume-change electrodes for Li-ion batteries of amorphous alloy particles held by elastomeric tethers, Electrochem. Commun. 5 (2003) 919–923, https://doi.org/10.1016/j.elecom.2003.08.017.

[98] T. Sarkar, A. Sharma, A.K. Das, D. Deodhare, M.D. Bharadwaj, A neural network based approach to predict high voltage li-ion battery cathode materials, in: 2014 2nd International Conference on Devices, Circuits and Systems (ICDCS), 2014, https://doi.org/10.1109/ICDCSyst.2014.6926140. Combiatore, India, March 06-08.

[99] R.P. Joshi, J. Eickholt, L. Li, M. Fornari, V. Barone, J.E. Peralta, Machine learning the voltage of electrode materials in metal-ion batteries, 11, 2019, pp. 18494–18503, https://doi.org/10.1021/acsami.9b04933.

[100] O. Allam, B.W. Cho, K.C. Kim, S.S. Jang, Application of DFT-based machine learning for developing molecular electrode materials in Li-ion batteries, RSC Adv. 8 (2018) 39414, https://doi.org/10.1039/c8ra07112h.

[101] R.P. Cunha, T. Lombardo, E.N. Primo, A.A. Franco, Artificial intelligence investigation of NMC cathode manufacturing parameters interdependencies, Batter. Supercaps 3 (2020) 60–67, https://doi.org/10.1002/batt.201900135.

[102] R.A. Eremin, P.N. Zolotarev, O.Y. Ivanshina, I.A. Bobrikov, Li, $(Ni,Co,Al)O_2$ cathode delithiation: a combination of topological analysis, density functional theory, neutron diffraction, and machine learning techniques, J. Phys. Chem. C 121 (2017) 28293–28305, https://doi.org/10.1021/acs.jpcc.7b09760.

[103] T. Parthiban, R. Ravi, N. Kalaiselvi, Exploration of artificial neural network [ANN] to predict the electrochemical characteristics of lithium-ion cells, Electrochim. Acta 53 (2007) 1877–1882, https://doi.org/10.1016/j.electacta.2007.08.049.

[104] P. Michal, A. Vagaska, M. Gombar, J. Kmec, E. Spisak, M. Badida, Prediction of the effect of chemical composition of electrolyte on the thickness of anodic aluminium oxide layer, Int. J. Math. Model Methods Appl. Sci. 8 (2014) 152–155.

[105] S.M. Woodley, R. Catlow, Crystal structure prediction from first principles, Nat. Mater. 7 (2008) 937–946, https://doi.org/10.1038/nmat2321.

[106] G. Hautier, C.C. Fischer, A. Jain, T. Mueller, G. Ceder, Finding nature's missing ternary oxide compounds using machine learning and density functional theory, Chem. Mater. 22 (2010) 3762–3767, https://doi.org/10.1021/cm100795d.

[107] W. Li, R. Jacobs, D. Morgan, Predicting the thermodynamic stability of perovskite oxides using machine learning models, Comput. Mater. Sci. 150 (2018) 454–463, https://doi.org/10.1016/j.commatsci.2018.04.033.

[108] F.A. Faber, A. Lindmaa, O.A.V. Lilienfeld, R. Armiento, Machine learning energies of 2 million elpasolite (ABC2D6) crystals, Phys. Rev. Lett. 117 (2016) 135502, https://doi.org/10.1103/PhysRevLett.117.135502.

[109] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M.A.L. Marques, Predicting the thermodynamic stability of solids combining density functional theory and machine learning, Chem. Mater. 29 (2017) 5090–5103, https://doi.org/10.1021/acs.chemmater.7b00156.

[110] B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J.W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning, Phys. Rev. B 89 (2014), https://doi.org/10.1103/PhysRevB.89.094104, 094104.

[111] L. Ward, R. Liu, A. Krishna, V.I. Hegde, A. Agrawal, A. Choudhary, C. Wolverton, Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations, Phys. Rev. B 96 (2017), 024104, https://doi.org/10.1103/PhysRevB.96.024104.

[112] G. Ceder, D. Morgan, C. Fischer, K. Tibbetts, S. Curtarolo, Data-mining-driven quantum mechanics for the prediction of structure, MRS Bull. 31 (2006) 981–985, https://doi.org/10.1557/mrs2006.224.

[113] C.C. Fischer, K.J. Tibbetts, D. Morgan, G. Ceder, Predicting crystal structure by merging data mining with quantum mechanics, Nat. Mater. 5 (2006) 641–646, https://doi.org/10.1038/nmat1691.

[114] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, G. Ceder, Predicting crystal structures with data mining of quantum calculations, Phys. Rev. Lett. 91 (2003) 135503, https://doi.org/10.1103/PhysRevLett.91.135503.

[115] V.S.M.C. Goncalo, C.B. Luis, A. Joao, A.M.A. Carlos, Exploration of quantitative structure property relationships (QSPR) for the design of new guanidinium ionic liquids, Tetrahedron 64 (2008) 2216–2224, https://doi.org/10.1016/j.tet.2007.12.021.

[116] T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, T. Oguchi, Crystal structure prediction accelerated by Bayesian optimization, Phys. Rev. Mater. 2 (2018), https://doi.org/10.1103/PhysRevMaterials.2.013803, 013803.

[117] A.R. Oganov, M. Valle, How to quantify energy landscapes of solids, J. Chem. Phys. 130 (2009) 104504, https://doi.org/10.1063/1.3079326.

[118] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1987) 37–52, https://doi.org/10.1016/0169-7439(87)80084-9.

[119] M. Loog, R.P.W. Duin, R. Haeb-Umbach, Multiclass linear dimension reduction by weighted pairwise Fisher criteria, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 762–766, https://doi.org/10.1109/34.935849.

[120] L. Meng, S. Ding, Y. Xue, Research on denoising sparse autoencoder, Int. J. Mach. Learn. Cybernetics 8 (2017) 1719–1729, https://doi.org/10.1007/s13042-016-0550-y.

[121] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. (2014) 2672–2680.

[122] T.C. Hung, K.Y. Chan, Uncertainty quantifications of Pareto optima in multiobjective problems, J. Intell. Manuf. 24 (2013) 385–395, https://doi.org/10.1007/s10845-011-0602-9.

[123] D.C. Li, I.H. Wen, A genetic algorithm-based virtual sample generation technique to improve small data set learning, Neurocomputing 143 (2014) 222–230, https://doi.org/10.1016/j.neucom.2014.06.004.

[124] Y. Xu, X. Li, J. Yang, D. Zhang, Integrate the original face image and its mirror image for face recognition, Neurocomputing 131 (2014) 191–199, https://doi.org/10.1016/j.neucom.2013.10.025.

[125] M. Gao, X. Hong, S. Chen, C.J. Harris, E. Khalaf, PDFOS: PDF estimation based over-sampling for imbalanced two-class problems, Neurocomputing 138 (2014) 248–259, https://doi.org/10.1016/j.neucom.2014.02.006.

[126] A. Berrones, E. Jimenez, M.A. Aracelia, F. Almaguer, B. Pena, Parameter inference of general nonlinear dynamical models of gene regulatory networks from small and noisy time series, Neurocomputing 175 (2016) 555–563, https://doi.org/10.1016/j.neucom.2015.10.095.

[127] B. Krawczyk, M. Galar, L. Jelen, F. Herrera, Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. 38 (2016) 714–726, https://doi.org/10.1016/j.asoc.2015.08.060.

[128] T. Lookman, P.V. Balachandran, D. Xue, R. Yuan, Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design, npj Comput. Mater. 5 (2019) 21, https://doi.org/10.1038/s41524-019-0153-8.

[130] D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nat. Commun. 7 (2016) 11241, https://doi.org/10.1038/ncomms11241.

[131] D. Xue, D. Xue, R. Yuan, Y. Zhou, P.V. Balachandran, X. Ding, J. Sun, T. Lookman, An informatics approach to transformation temperatures of NiTi-based shape memory alloys, Acta Mater. 125 (2017) 532–541, https://doi.org/10.1016/j.actamat.2016.12.009.

[132] D. Xue, P.V. Balachandran, R. Yuan, T. Hu, X. Qian, E.R. Dougherty, T. Lookman, Accelerated search for $BaTiO_3$-based piezoelectrics with vertical morphotropic phase boundary using Bayesian learning, Proc. Natl. Acad. Sci. Unit. States Am. 113 (2016) 13301–13306, https://doi.org/10.1073/pnas.1607412113.

[133] R. Yuan, Z. Liu, P.V. Balachandran, D. Xue, Y. Zhou, X. Ding, J. Sun, D. Xue, T. Lookman, Accelerated discovery of large electrostrains in BaTiO3-based piezoelectrics using active learning, Adv. Mater. 30 (2018) 1702884, https://doi.org/10.1002/adma.201702884.

[134] P.V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning, Nat. Commun. 9 (2018) 1668, https://doi.org/10.1038/s41467-018-03821-9.

[135] Z. Zhou, Machine Learning, Tsinghua University Press, 2016.

[136] R.W. Johnson, An introduction to the bootstrap, Teach. Stat. 23 (2001) 49–54, https://doi.org/10.1111/1467-9639.00050.

[137] Y. Liu, Y. Li, G. Li, B. Zhang, G. Wu, Constructive ensemble of RBF neural networks and its application to earthquake prediction, in: Second International Symposium on Neural Networks, 2005. Chongqing, China, May 30-June 01.

[138] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305.

[139] J. Huang, W. Sun, L. Huang, Deep neural networks compression learning based on multiobjective evolutionary algorithms, Neurocomputing 378 (2020) 260–269, https://doi.org/10.1016/j.neucom.2019.10.053.

[140] J. Bergstra, R. Bardenet, Y. Bengio, B. Kegl, Algorithms for hyper-parameter optimization, in: 25th Annual Conference on Neural Information Processing Systems (NIPS 2011), Granada, Spain, December 2011.

[141] F. Hutter, H.H. Hoos, K. Leyton-Brown, Sequential model-based optimization for general algorithm configuration, in: International Conference on Learning and Intelligent Optimization, 2011, https://doi.org/10.1007/978-3-642-25566-3_40. Berlin, Germany, January 18.

[142] A. Klein, S. Falkner, S. Bartels, P. Hennig, F. Hutter, Fast bayesian optimization of machine learning hyperparameters on large datasets, Proc. Mach. Learn. Res. 54 (2017) 528–536.

[143] M. Feurer, J.T. Springenberg, F. Hutter, Initializing bayesian hyperparameter optimization via meta-learning, in: 29th Association-For-The-Advancement-Of-Artificial-Intelligence (AAAI) Conference on Artificial Intelligence, 2015. Austin, America, January 25-30.

[144] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel, J. Hopfield, Large automatic learning, rule extraction, and generalization, Complex Syst. 1 (1987) 877–922.

[145] M.J. Flores, A.E. Nicholson, A. Brunskill, K.B. Korb, S. Mascaro, Incorporating expert knowledge when learning Bayesian network structure: a medical case

study, Artif. Intell. Med. 53 (2011) 181–204, https://doi.org/10.1016/j.artmed.2011.08.004.

[146] W. Tang, K.Z. Mao, L.O. Mak, G.W. Ng, Adaptive fuzzy rule-based classification system integrating both expert knowledge and data, in: IEEE 24th International Conference on Tools with Artificial Intelligence (ICTAI), 2012, https://doi.org/10.1109/ICTAI.2012.114. Athens, Greece, November 07-09.

[147] A.R. Masegosa, S. Moral, An interactive approach for Bayesian network learning using domain/expert knowledge, Int. J. Approx. Reason. 54 (2013) 1168–1181, https://doi.org/10.1016/j.ijar.2013.03.009.

[148] M. Mozina, M. Guid, J. Krivec, A. Sadikov, I. Bratko, Fighting knowledge acquisition bottleneck with argument based machine learning, in: 18th European Conference on Artificial Intelligence, 2008, https://doi.org/10.3233/978-1-58603-891-5-234. Patras, Greece, July 21-25.