Computational Materials Science 246 (2025) 113344

Contents lists available at ScienceDirect

Computational Materials Science

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

Full length article

Predicting battery applications for complex materials based on chemical composition and machine learning

Zixin Zhuang, Amanda S. Barnard *

School of Computing, Australian National University, 145 Science Road, Acton, 2601, ACT, Australia

ARTICLE INFO

Dataset link: https://next-gen.materialsproject.org/batteries, https://github.com/ShuHuang/batterydatabase/tree/master/chemdataextractor_batteries, https://doi.org/10.1038/s41597-020-00602-2, https://doi.org/10.6084/m9.figshare.11888115.v2, https://doi.org/10.1021/ja508154e

Keywords:
Machine learning
Classification
Anode
Cathode
Charge
Discharge
Batteries

ABSTRACT

Materials informatics uses machine learning to predict the properties of new materials, but generally requires extensive characterisation and feature extraction to describe the input data, which can be time consuming and expensive. Predicting properties or classes of materials based on minimal input information, such as a chemical formula, can be a useful first step to identify which materials are promising candidates before investing resources. This is particularly desirable when working with complex compounds containing a large variety of elements, such as materials for battery applications. In this paper we show how to classify battery compounds into either charge or discharge formulas, or identify suitable anode or cathode materials, based exclusively on the chemical formulas of materials available in online repositories. Without any structural information, we train high-performing classifiers that can be used to rapidly screen hypothetical materials and assign potential applications. The models are applied to a total of 471 materials from the literature, and deliver a 96% success rate over 80% probability. These methods are general and the workflow can be applied to any complex crystalline materials to predict end-uses in advance of synthesis or simulation, opening up the opportunity for machine learning to use used for research planning, in addition to prediction or inference.

1. Introduction

Machine learning (ML) is a powerful method for predicting the properties of new materials based on patterns learnt from prior observations (data) [1–5]. Trained models enable researcher to infer materials for particular applications once they have been characterised and the model applied. However, in many cases it is desirable to infer the applications directly, or identify candidature materials before they are synthesised, in order to save time and resources [6,7]. Predicting the possible applications for hypothetical materials based exclusively on a chemical formula, in advance of any experimentation or simulation, casts ML as a research planning tool, in addition to a tool for post-hoc analysis [8].

Structure-free encoding provides a way of doing this, and is particularly useful for complex crystalline compounds. A variety of structure-free encoding are offered by materials informatics platforms [9–14], and a recent comparison has shown that superior performance can be achieved using only the chemical formula for unsupervised [15] and supervised classification tasks [16]. Using three data sets of complex materials and four different unsupervised learning methods, inclusive of six algorithms with four evaluation metrics, it was shown that structure-free Mendeleev encoding out-performed simple one-hot encoding and more sophisticated representations. This capability would be particularly valuable for complex materials such as those proposed for energy storage applications [17–19] such as batteries [20–27], where the combinatorial problem grows exponentially as more elements are included.

ML has been used to predict the properties of known battery materials [28] and devices [29]. Common problems include predicting charging performance [30,31], lifetimes and energy levels [32–35], properties under different conditions [36,37], design of new materials or electrolytes [38,39] and improving safety [40]. For example, supervised ML has been used to predict the electrochemical properties for new cathode materials and establish quantitative structure-redox potentials relationships that improve stability [41,42]. Studies such as these involve multiple structural characteristics (features), such as the concentration of various chemical elements and/or bonding configurations such as aromatic rings, with the ultimate aim of predicting functional properties (labels) such as voltage, capacity or charge [43,44]. However, this assumes that the physical structure is known, that the material has been correctly assigned to one or more battery components, and that the time and effort has already been expended to

* Corresponding author.
E-mail address: amanda.s.barnard@anu.edu.au (A.S. Barnard).

https://doi.org/10.1016/j.commatsci.2024.113344
Received 2 May 2024; Received in revised form 7 August 2024; Accepted 3 September 2024
Available online 9 September 2024

measure the properties of interest. At this point it is too late to make small changes to the materials that could potentially have a big impact on the economic and energy efficiency.

Pre-screening based on minimal information in advance of research has numerous advantages, particularly if reliable predictions can be made based on existing data from a variety of sources. However, to be able achieve this goal with confidence, ML much be able to reliably reproduce domain knowledge under these data-constrained conditions. If data-constrained ML cannot predict facts that we know to be true, how can we rely of the predictions when we do not know the outcome in advance? One could argue that this is a given when models are trained on data with high veracity, but it is not clear that ML models can be relied upon when based on minimal information that contains no properties or structure at all.

In this paper we aim to determine if ML can accurately reproduce domain knowledge under the data-constraint of using only the chemical formula, with no structural information, environmental information (such as pairings or working ions), or measured properties. To do this we use structure-free Mendeleev encoding to represent complex compounds used for battery applications, to train accurate classifiers predicting the appropriate end-uses of the materials. Both computational and experimental data sets are used, predicting charge or discharge, and anode or cathode formulas respectively. The classifiers are validated using a total of 471 materials from the literature that were not included in the training or testing sets, resulting in a 100% success rate over 90% probability (96% with 80% probability), using only the chemical formula as input. This is the first demonstration of this capability that can be extended to other applications or used with other sets of complex compounds based exclusively on the materials chemistry.

2. Machine learning methods

In this study we have used three data sets of battery compounds with different complexity to challenge the concept of predicting the suitability of a material using the chemical formula alone based on three separate tasks. The materials have been encoded using structure-free Mendeleev encoding to train random forest classifiers. The models have been evaluated using five evaluation methods and interpreted using feature importance profiles.

2.1. Data

The computational data set consists of 10129 instances of battery compounds retrieved from using both the legacy and the new APIs of the materials project [45,46]. In this set each compound contributes at least one “voltage pair”, containing the charge and discharge formulas, and the working ion of a single electrochemical reaction step. Most battery compounds contain one voltage pair each, but some contribute several voltage steps to the data set, since each voltage pair represents a unique data instance. The working ions are Li, Na, K, Rb, Cs, Mg, Ca, Al, Zn, and Y. Post-processing has been applied to remove redundant metadata, and this data set will be used to classify compounds as either suitable for charge or discharge materials. This results in 10129 × 2 labelled data instances for this task. An electronic notebook outlining this procedure is included in the Supporting Information, and can be downloaded along with the original data and axillary files at https://github.com/ZoraZhuang/battery_datasets.

The experimental data set consists of 265481 data instances of battery compounds, refined from the work of Huang and Cole [47]. In this set each data instance represents a battery cell that contains one or more battery component, with each component expressed as a chemical formula. The set was originally 311716 chemical formulas auto-extracted from the literature, but due to inaccurate document digitisation and error-prone chemical language processing, significant data clearing and processing was required. We developed a set of sophisticated rules to inspect the formulas, and data instances are either retained, corrected, or discarded accordingly, to minimise errors. An electronic notebook outlining this procedure is included in the Supporting Information, and can be downloaded along with the original data and axillary files at https://github.com/ZoraZhuang/battery_datasets. This data set will be used to classify compounds as either suitable for anode or cathode materials. Of the remaining instances only 8013 have the anode/cathode label (4649 cathode and 3364 anode), limiting the size of the set for this task.

The application data set used for inference consists of 360 MXenes simulated by Eames and Islam [48]. It was suggested by the authors that these materials would be suitable for battery applications, and this will be put to the test in the present study by inputting the chemical formulas into the final model, and predicting the probability that a particular MXene is suitable as a charge or discharge material, in agreement or disagreement with the original report. All data instances are use for this task. The aim of this exercise is to challenge the models, and confirm whether the classifiers can reproduce scientific domain knowledge, in the absence of any information about the structure, working ions or charge pairings.

2.2. Structure-free encoding

Recent research has shown that complex battery compounds can be accurately represented using *Mendeleev* encoding [15], which is chemically intuitive, structure-free and capable of separating compounds in classification tasks [16]. In Mendeleev encoding the categorical compounds are expanded into a $1 \times N$ matrix (vector) with entries that reflect the chemical composition. This encoding can accommodate non-stoichiometric chemical formulas (provided the fractional value of constituents, usually denoted by x, is assigned a value). More sophisticated versions of this encoding, containing more structure-free information from the periodic table (referred to as *Mendeleev+*) can reduce model complexity, at the expense of interpretability. Mendeleev+ encoded results will be included here for comparison in Supporting Information, which incorporate extra features including statistics minimum, maximum, range, mean and standard deviation, of properties molar volume, density, atomic weight and volume, the atomic number and Mendeleev number, the column, row, block and group of the atom, and the number of valence electrons. Mendeleev encoding has chosen for presentation in the main text has it has been show to be superior for a wide range of unsupervised learning and classification tasks in previous studies [15,16], but there are still cases where multiple materials may have the same encoding (such as graphite and diamond) and can only be differentiated by the property label. In the case of battery compounds, diamond is unsuitable and would not be represented in the training set, so all final predictions obviously relate to graphite.

2.3. Classification

Since Mendeleev encoded battery compounds are high dimensional tabular sets, we have used tree-based random forests (RF) ensemble classifiers [49]. Decision tree is a non-parametric explanatory method that makes no assumption on the structure of the model but exposes interpretable feature importances [50], making it ideal for predicting structure/property relationships. RFs use a large number of decision trees and combines bootstrap aggregation (also known as bagging [51]), where each decision tree in the forest is trained on a different data sample with random subsets of features, and “votes” determine the prediction that best resembles the limited scope at the point of prediction. The forest is trained on random individual trees to predict the mode of the class for classification, and the diversity among the ensemble members helps to avoid over-fitting [52]. In this study, we adopt the RF implementation of scikit-learn [53], and all model parameters are contained in the Supporting Information. RF is chosen because it has been shown to perform well for high dimensional

Table 1
Binary classification report using a random forest classifier for the computational battery compounds data set and the experimental battery compounds data set.

Set	Computational			
Data	Training Set		Testing Set	
Formula	Discharge	Charge	Discharge	Charge
Support	8596	8597	1518	1517
Precision	0.904	0.939	0.823	0.858
Recall	0.942	0.900	0.866	0.813
F1-score	0.922	0.919	0.844	0.835
AUC–ROC	0.985		0.935	
Set	Experimental			
Data	Training Set		Testing Set	
Formula	Anode	Cathode	Anode	Cathode
Support	2859	3952	505	697
Precision	0.925	0.969	0.926	0.962
Recall	0.959	0.944	0.949	0.945
F1-score	0.942	0.956	0.937	0.954
AUC–ROC	0.989		0.984	

structured data sets in the past, and affords us the opportunity to investigate the architecture of the model via feature importance profiles, but in principle any classier could be used with these encodings and workflow [16].

2.4. Evaluation

The training of the models in this study have been evaluated using learning curves to identify any bias error (under-fitting), variance error (over-fitting) and convergence with respect to the number of training instances (data sufficiency), with k-fold cross-validation. The performance of the classifiers on unseen data in the testing set is evaluated using classification reports, based on the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The reports include the precision (positive predictive value, TP/[TP+FP]) and the recall (sensitivity, TP/[TP+FN]). Accuracy (measured here using the F1-score) is simply a ratio of correctly predicted observation to the total observations, such that: F1-score = 2 × ([precision × recall]/[precision+ recall]). In addition to classification reports, we also present the absolute number of TP, TN, FP, FN as a confusion matrix, cross referencing the prediction with the ground truth label. Receiver Operating Characteristic (ROC) are also used to evaluate the sensitivity of the classifiers. ROC curves use the FP rate and TP rate as the horizontal and vertical axes respectively, and the area under the curve (AUC) measures the successful classification TP and TN rates of the ground truth classes. In this study we provide the AUC–ROC curves in the Supporting Information.

3. Results and discussion

Following the processing and encoding of each data set, two classifiers were trained to assign materials to either the charge or discharge side of a voltage pair, or as an anode or cathode, respectively. The performance of each classifier will be evaluated, as well as the characteristics of the classes, to determine their reliability for assigning appropriate used for unknown materials. The results are summarised in Table 1.

3.1. Charge/discharge

Charge and discharge assignments are predicted using a RF classifier retained on the computational data set of 20258 materials, noting that the same material formula can be the charge formula in one voltage pair and the discharge formula in another. Pairing of anode and cathode electrodes is related to the reduction potential, where compounds with relatively higher chemical potentials are used as cathodes [54]. In a metal-ion battery, materials that exclude the working ion are anodes, while materials with some concentration of the metal-ion can be uses as cathodes or anodes, depending on the counterpart electrode, which leads to come ambiguity in the data set. For example, when Li^+ is the working ion, LiC_{12} can be the discharge formula if the charge formula is C. LiC_{12} can also be the charge formula if the discharge formula is LiC_6. Hence, materials such as these are ambiguous in nature, and this ambiguity is handled by including them in each category, rather than removing them entirely. Removal is both disingenuous, and reduces the number of instances which compromises model training, so we retained the materials before splitting and some (randomly) appear in both training and testing sets. The RF was optimised by sampling 1000 random hyperparameter combinations, and trained using a with 7-fold cross-validation. The model parameters are provided in the Supporting Information.

Table 2
Contingency table of testing formulas.

	Ambiguous	Unambiguous	Total
Misclassified	415	72	487
Correctly classified	326	2222	2548
Total	741	2294	3035

The classification results are included in Table 1, confirming the new model classifies charge/discharge formulas with high accuracy (AUC–ROC score 93.5%). All testing results are over 80%, but the precision of the discharge prediction is lower than that of charge prediction, while the recall of charge is lower than that of discharge. This indicates that the model is biased towards discharge, and (although rarely) will more frequently misclassify charge as discharge than the other way around. This is corroborated by the confusion matrices shown in Fig. 1, which quantifies greater number of FNs than FPs, and more TNs than TPs (from the charge perspective), despite the fact that the data set is class balanced. This is likely due to ambiguous materials.

From the learning curve shown in Fig. 2(a) we can also see that this classifier is slightly over-fitting, and the model is not yet converged. This indicates that more data could potentially help improve both model performance and over-fitting, but at the cost of increasing under-fitting (train) unless more suitable hyperparameters could be identified. The validation curve (val) is also rather unstable, though the instability generally decreases as the number of data samples increases. Once again this could be attributed to the random allocation of the ambiguous formulas. The classifier performance on the testing data based on formula ambiguity is shown in Table 2, where we can see that over 85% of the misclassifications can be attributed to ambiguous formulas, and over 50% percent ambiguous formulas are misclassified. Only 3% of unambiguous materials are misclassified.

A benefit of using a RF is that it is interpretable and exposes feature importance profiles that rank the features by how frequently they are used to make a decision in the model architecture. The feature importance profile displaying the top 15 most important features/elements are shown in Fig. 2(b) for the charge/discharge classifier. As expected, Li is the most influential feature, followed by O, with metals (alkali, alkaline earth and some transition metal) also important to the predicting. These metals are accompanied by non-metals that they commonly bond with (P, O and F). In particular, the importance of F indicates that underlying chemical reactions involve the transfer of F (potentially involving F_2 or F^-) which is generally undesirable. The important metals in the charge/discharge predictions include the working ions, but this is not evenly distributed. The discharge formulas almost always contain working ions, but the charge formulas do not necessarily contain them. In general, the model primarily distinguishes between the charge and discharge formulas based on the concentration of s-group elements.

There is correlation between feature importance and the mean feature value difference between charge and discharge (shown in Fig. 3),

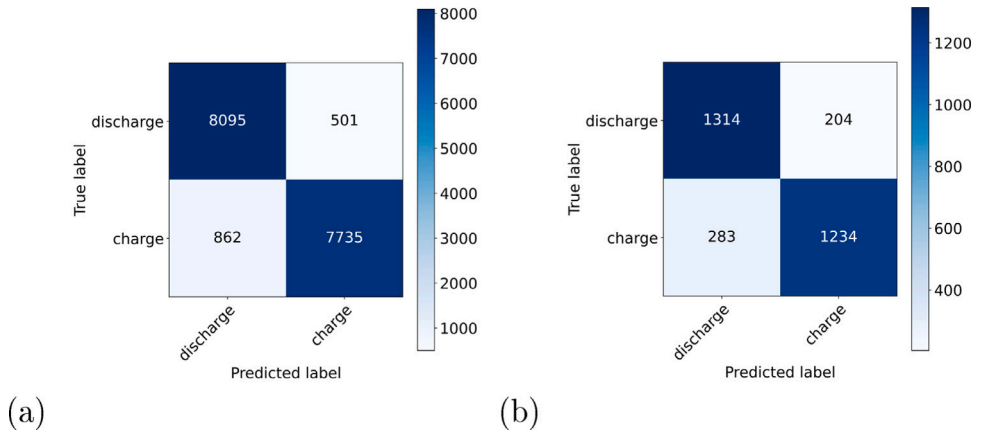

Fig. 1. Confusion matrices of the prediction results on the (a) training and (b) testing data sets.

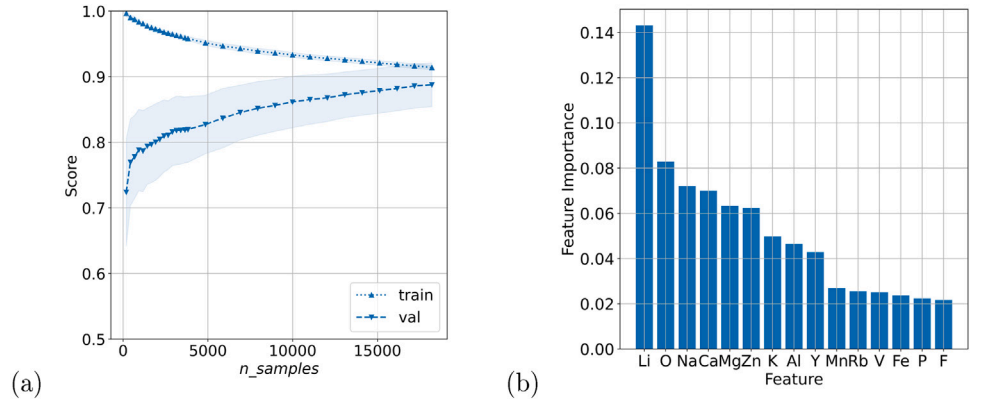

Fig. 2. (a) Learning curve of the accuracy of the random forest tree model, and (b) feature importance profile describing the relative importance of the top 15 most influential features for predicting charge and discharge materials in the computational data set. Learning curves identify model underfitting, overfitting and instability, as well confirming sufficiency of training data; in this case *n_samples* i the number of training instance, *Score* is the coffin of determination, 'train' and 'val' are the training and validation loss, respectively.

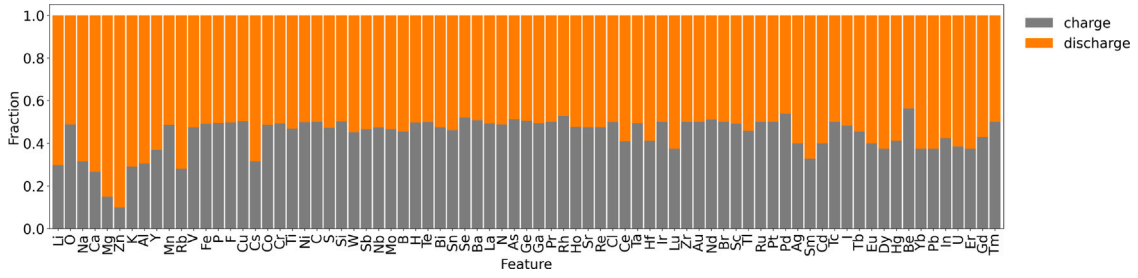

Fig. 3. Relative average feature values of charge and discharge formulas in computational data set. The order left to right reflects the order of importance in the RF classifier.

but the relationship is not simple or linear. In this histogram, higher feature importance does not necessarily correspond to larger difference, as the decision process of RF is complicated. For example, O has a high feature importance, but the mean feature values of charge and discharge are extremely close. However when the difference of mean values is large (usually potential working ions), feature importance tends to be higher. While the average feature values, which are elemental concentrations when using Mendeleev encoding, are similar for charge and discharge formulas for most materials, many important alkali, alkaline earth and some transition metals are in greater supply in discharge.

To further illustrate the concentration elements in this data set, the distribution of some important elements is highlighted using t-distributed stochastic neighbour embedding (t-SNE) in Fig. 4, and the remainder are provided in the Supporting Information. Here we can see that Li (Fig. 4(a)) forms local groupings, with local gradients, but is distributed throughout the set. The materials are not characterised by Li alone. O, commonly exists with high concentration (Fig. 4(b)), and materials can be characterised more easily by their omission of O. Materials containing F are unique in forming one major local grouping with a high concentration, and a few low concentration cases dispersed in the feature space. These elements are important to the model because they are complementary. Both P and F form distinct local groupings, with local gradients, but in lower concentrations. These elements are important to the model because they are more common and capture more material diversity.

In general, we expect this model to perform better for assigning discharge formulas. This classifier is put to the test in Section 4.

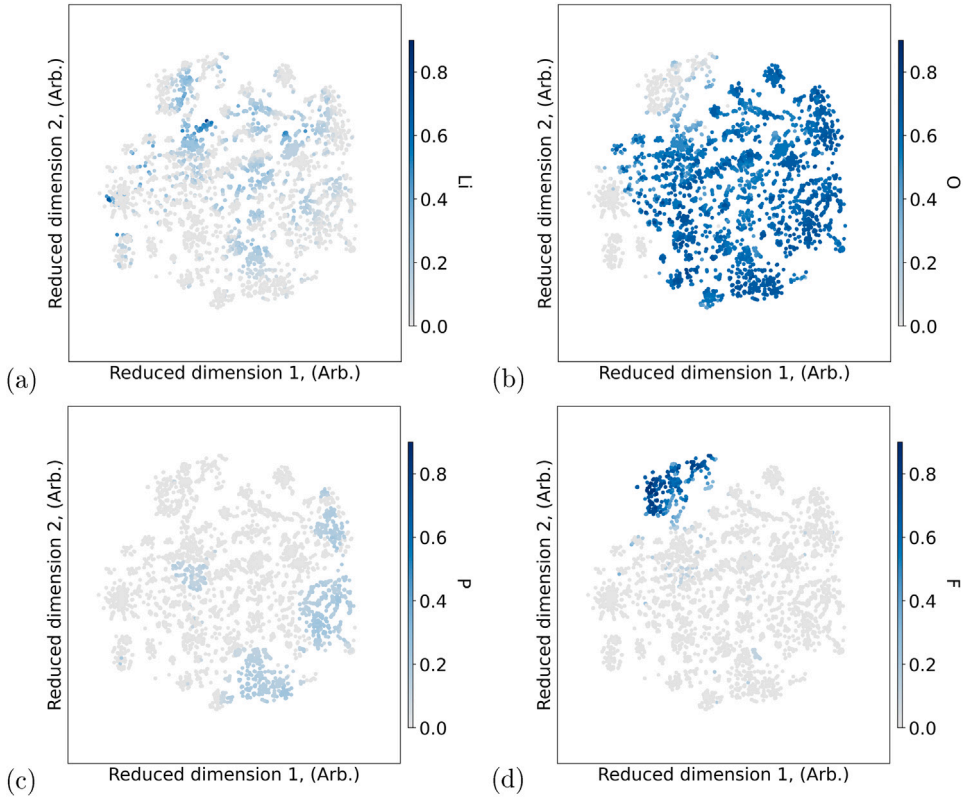

Fig. 4. Feature value distribution of (a) Li (b) O (c) P (d) F in the reduced feature space related to charge and discharge materials in the computational data set.

3.2. Anode/cathode

The experimental data set containing 8013 battery materials with electrode information are used for the classification of anodes or cathodes. A new RF classifier was optimised by sampling 1000 random hyperparameter combinations, and trained with 7-fold cross-validation. The classification results are summarised in Table 1, which shows an increase in model performance relative to the computational data set, attributed to the absence of ambiguous formulas. This is also evidenced by the confusion matrices, shown in Fig. 5 where there are a low number of FNs and FPs.

This classifier is slightly biased towards anode materials, even though the anode category has fewer support instances, but generalises well on the testing data set. Looking to the learning curve in Fig. 6(a) we can see evidence of minor over-fitting following convergence at approximately 2000 instances, and instability of the validation set. This sensitivity to sampling indicates that there may be selection bias in this data set, which is unsurprising and common in materials data sets scraped from literature and known to suffer from reporting bias [55].

There are some similarities between the feature importance profiles for the RF classifier of this experimental data set and that of the computational data set. Li and O are still the top 2 most important elements, but Ti and Sn are noticeably more important in this task, along with C and Si. This could be due to the inclusion of more Ti, Sn, C and Si in the experiments than in the original simulations (selection bias) or that they are more evenly distributed between the classes in the computational data set. A comparison between Figs. 3 and 7 confirms that it is the latter. In general, the classifier primarily distinguishes between the anode and cathode based on the concentration of transition metals.

Fig. 7 also shows that the compositions of the anode and cathode materials are very different. This barplot is correlated with feature importance profile, and we can see that there are elements (Ge, Re, Cd and As) that only appear in anode materials, or almost only appear in anode material (Sn, Zn, Sb, Mg and Ga). Conversely there are elements that only appear in cathode materials (I, Ta, Pr, Nd, Rh). Few elements could equally appear in anodes or cathodes (Co, Cr, Ca and Pd), which was most common in the charge/discharge materials simulated in the computational data set (see Fig. 3).

For the same selected elements (Li, O, P and F), the visualisation of the elemental distributions using t-SNE are shown in Fig. 8. Here we can see that there is a much lower concentration of P, F and Li in most materials in the anode and cathode data, the qualitative patterns are similar to the charge and discharge formulas. This imbalance is likely due to researchers avoiding elements for practical reasons (such as availability or toxicity), rather than inherent unsuitability. We can also see that F is important because the materials are highly similar and quite unique (grouped) while Li, O and P are important because they are common to many materials and therefore more representative.

In general, we expect this model to perform equally well for anode and cathode materials. To put this to the test the classifier has been applied to 15 anode and 6 cathode materials found in the literature [56, 57]. Among this set 3 anode materials are misclassified as anodes, and all cathode materials were correctly identified (see Table 3). This is a success rate of 86% using nothing more than the chemical formula as the input. In this table only two "true anode" materials contain Zn; ZnC_2O_4 and $ZnMn_2O_4$ can be considered as anode or cathode depending on the chemical potential of the counterpart electrode. It is therefore clear that ML classifies these materials correctly, even without details of the working ions, counterpart electrodes, or structural information.

4. Application to MXenes

To demonstrate the utility of these models in practice, we apply the charge/discharge classifier trained on the computational data set to an independent data set of MXenes that were previously proposed for energy storage applications [48]. In this data set M represents an

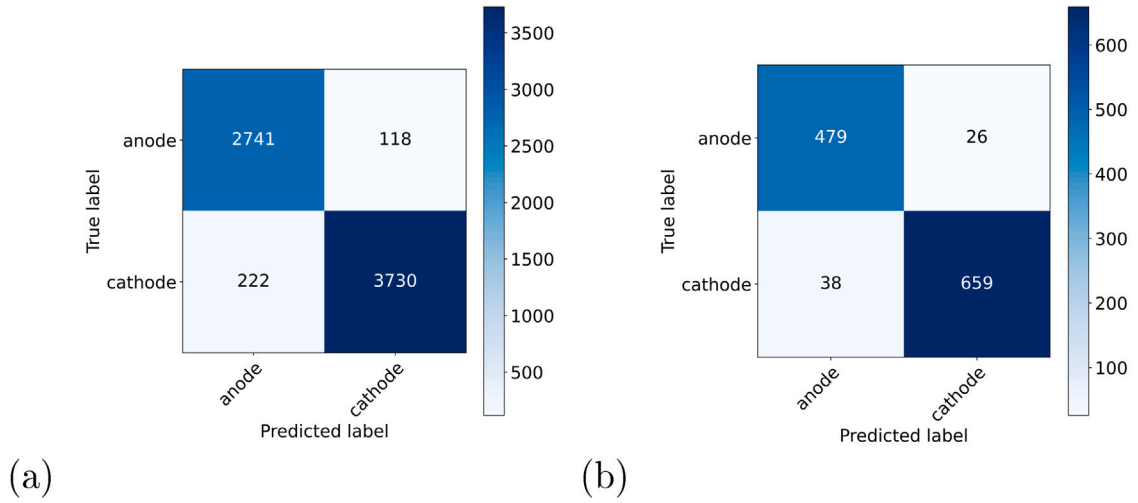

Fig. 5. Confusion matrices of the anode and cathode prediction results on the (a) training and (b) testing data sets.

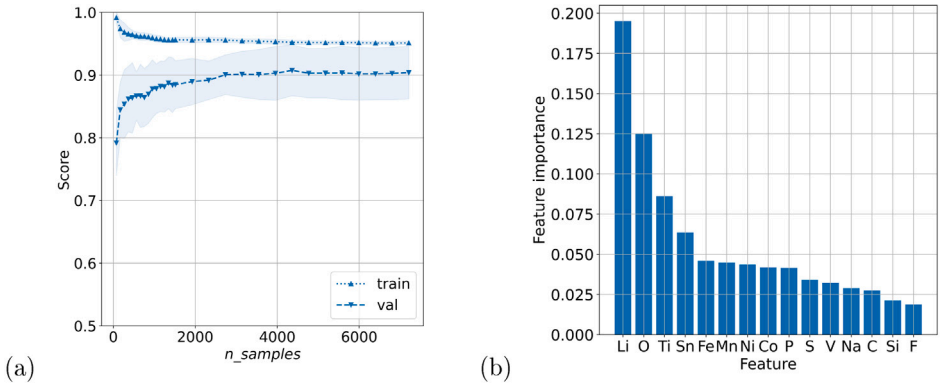

Fig. 6. (a) Learning curve of the accuracy of the random forest tree model, and (b) feature importance profile describing the relative importance of the top 15 most influential features for predicting anode and cathode materials. Learning curves identify model underfitting, overfitting and instability, as well confirming sufficiency of training data; in this case *n_samples* i the number of training instance, *Score* is the coffin of determination, 'train' and 'val' are the training and validation loss, respectively.

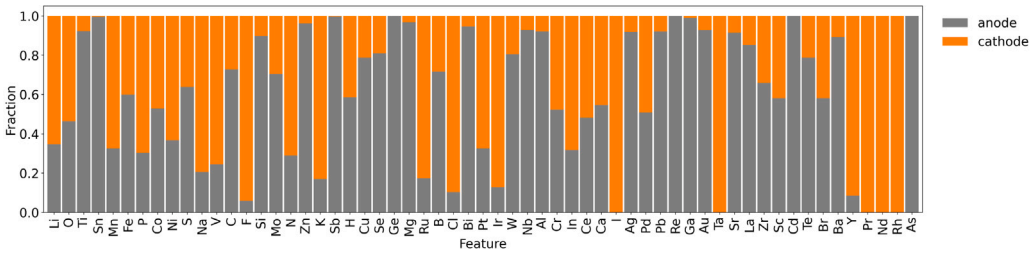

Fig. 7. Relative average feature values of anode and cathode formulas in experimental data set. The order left to right reflects the order of importance in the RF classifier.

early transition metal (Sc, Ti, V, Cr, Zr, Nb, Mo, Hf, and Ta), and X represents either C or N. Additionally the induced charge and capacity were modulated by intercalating different ions and ionic molecules (Li^+, K^+, K^+ and Mg^{2+}), and functional groups supporting metallic conductivity and hydrophilicity were represented by T (O, F, OH and 'Null'). This results in a set of organic anions (C, CF_2, CH_2, CO_2, COH, N, NF_2, NH_2, NO_2 and NOH). The 360 materials are all originally designated as discharge formulas, which correspond to 90 distinct charge formulas and 4 distinct working ions. These results are in agreement with experimental observations [58,59].

We apply the classifier to each of these formula and calculated the probability of being classified as 'charge' or 'discharge', with a threshold of 0.9. If a material has a probability of <90% then it is not assigned to that class.

For the 90 formulas that were originally designated as 'charge', the classifier delivers 100% accuracy. All the charge formulas have a near 100% probability, which can be interpreted as a correct (original) designation, or as agreement between the prediction and the original ground truth. This is visualised in Fig. 9(a), where the groupings have been previously shown to reflect the different M [15].

For the 360 formulas originally designated as 'discharge', the classifier also classifies all formulas correctly at the given threshold (100% success rate). However, if the threshold is reduced to 0.8, the classifier would find 14 misclassifications, including Li_2Zr_2C, Li_2Sc_2C, Li_2Cr_2C, Li_2Hf_2C, Li_2Nb_2C, Li_2Ta_2C, Li_2Mo_2C, $Li_2Sc_2CH_2$, $Li_2Hf_2CH_2$, $Li_2Ta_2CH_2$, $Li_2Zr_2CH_2$, Mg_2Ti_2N, $Mg_2Ti_2NH_2$, Mg_2Ti_2NOH. This is visualised in Fig. 9(b) as low probabilities, where the four 'fireworks' groupings have been previously shown to correspond to the four working ions [15].

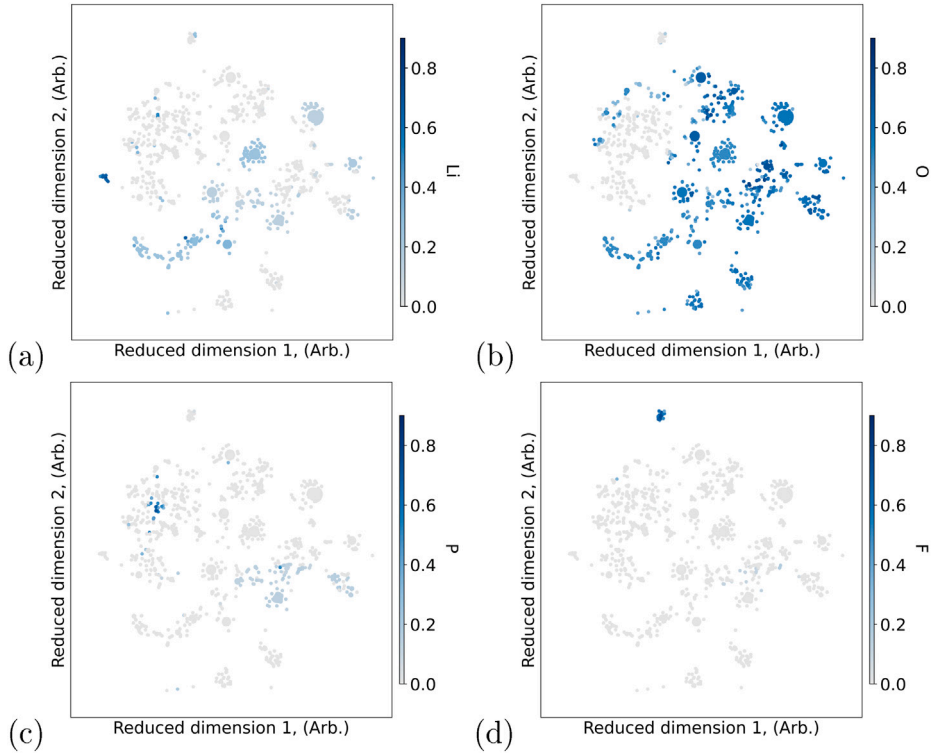

Fig. 8. Feature value distribution of (a) Li (b) O (c) P (d) F in the reduced feature space, in anode and cathode materials in the experimental data set.

Table 3
Application of the anode/cathode classifier to materials from the literature, comparing the predicted designation with the true designation from the original Reference. Only MnO_2, CoS_2 and $Ni(OH)_2$ have been misclassified, representing an 86% success rate.

Reference [56]			Reference [57]		
Material	True	Predicted	Materials	True	Predicted
Fe_3O_4	Anode	Anode	$LiFePO_4$	Cathode	Cathode
FeS_2	Anode	Anode	$LiCoO_2$	Cathode	Cathode
MnO_2	Anode	Cathode	$LiMn_2O_4$	Cathode	Cathode
SnS	Anode	Anode	$Li_3V_2(PO_4)_3$	Cathode	Cathode
CoS_2	Anode	Cathode	$Na_3V_2(PO_4)_3$	Cathode	Cathode
ZnC_2O_4	Anode	Anode	$Na_3V_2(PO_4)_2F_3$	Cathode	Cathode
CoC_2O_4	Anode	Anode			
MnC_2O_4	Anode	Anode			
NiC_2O_4	Anode	Anode			
FeC_2O_4	Anode	Anode			
$TiNb_2O_7$	Anode	Anode			
$Ti_2Nb_10O_29$	Anode	Anode			
$ZnMn_2O_4$	Anode	Anode			
$Ni(OH)_2$	Anode	Cathode			
$Cu(OH)_2$	Anode	Anode			

These misclassifications could be because all formulas originally designated as ‘charge’ do not contain a working ion, which disqualifies them from being discharge formula. However from a physico-chemical perspective, having a working ion in the formula does not necessarily mean that is cannot be a discharge formula. An alternative conclusion is that the original designation is not supported by the data and possibly incorrect. The classifier is known to be slightly biased towards discharge formulas, which will be compensating this effect to some degree, so the issue could be more significant than it appears.

A more detailed visualisation of the results help to clarify this issue. Fig. 10 plots the discharge formula encoded by the working ion (a), the metals, M (b) and the organic anions (c). By comparing with Fig. 9(b) the misclassifications are visually discernible. Formulas containing K and Na are more likely to be classified as discharge than those containing Li and Mg. (Fig. 10(a)). Formula containing Ti and Hf are more likely to be classified as charge, and formula with anions containing N (e.g. N, NH_2, NOH, NF_2, NO_2) are more likely to be classified as discharge than those containing C (e.g. C, CH_2, COH, CF_2, CO_2).

These three trends are invaluable for assigning new MXenes to potential battery components, since they are based entirely on the chemical formula, and do not require preliminary synthesis, characterisation or quantum mechanical simulations.

5. Conclusion

In this study we have developed high performing machine learning classifiers, with only minor under-fitting, over-fitting and instability, using just the Mendeleev-encoded computational and experimental data sets of battery materials. With these classifiers we can confidently separate charge and discharge formulas (with a slight bias towards discharge) and anode and cathode materials, both with high accuracy, precision and recall. This was possible even with imbalanced classes and ambiguous formulas that can technically be assigned to either class. The results from area under the receiver operating characteristic curve (AUC–ROC) show excellent sensitivity. This novel predictive ability, using only chemical formula as input, demonstrates how machine learning models can be used as research planning tools.

We have also shown how the models can be used for research evaluation and inference. When each of these classifiers was applied to battery materials proposed in the literature, they correctly identified all 6 cathode materials and 90 charge formulas with 90% probability. The anode/cathode classifier also delivered 86% success rate on the 15 anode materials and the change/discharge classifier delivered a 100% success rate for the 360 discharge formulas with 90% probability (or a 96% success rate with 80% probability), which is remarkable given the ambiguity between charge and discharge formulas in the training data. These results are in agreement with experiment [58,59], and future

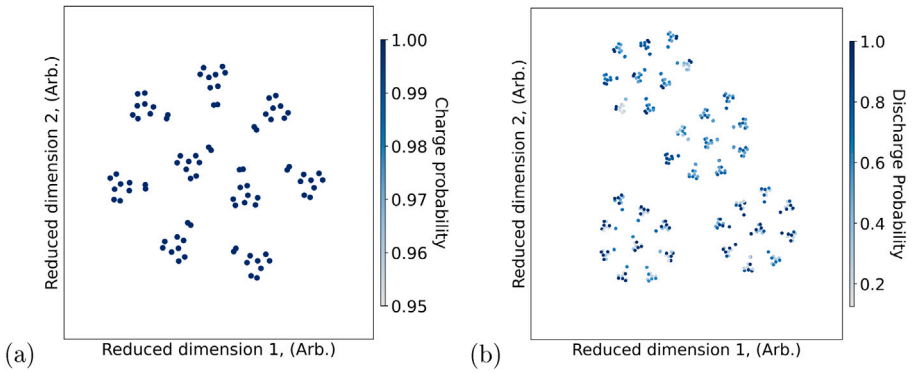

Fig. 9. Predicted probability of (a) the 90 materials originally designated as charge formula being classified as 'charge' (b) the 360 materials originally designated as discharge formula being classified as 'discharge', in the MXene data set.

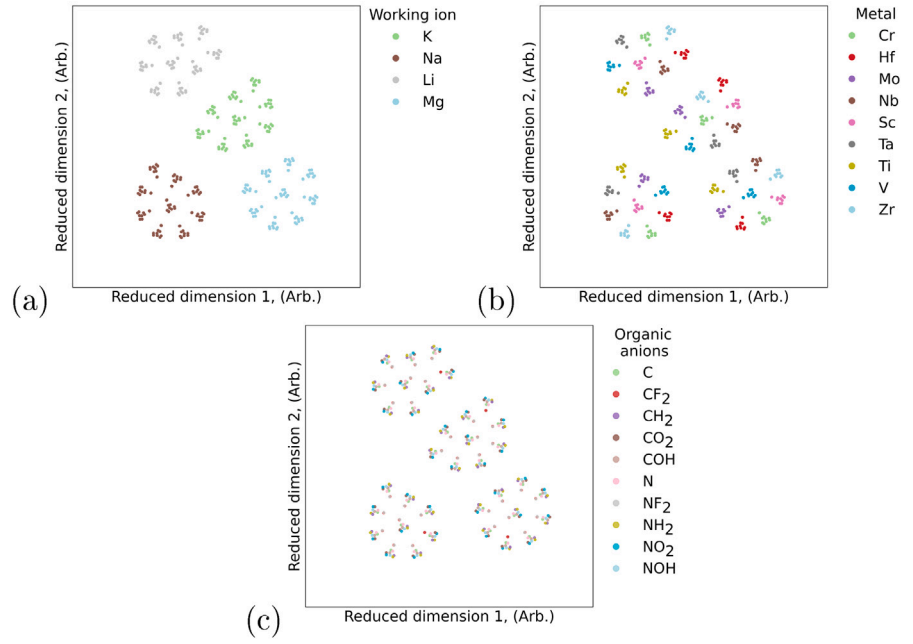

Fig. 10. Discharge MXene formula encoded by (a) the working ion (Z), (b) metal cation (M), and (c) the organic anion of MXene (T).

work is underway to assess performance of the structure-free encodings in regression tasks, or predict specific materials properties.

This study (supported by previous methodological studies [15,16]) demonstrates that structural information is not necessary to train a high performing classifier, even though as scientists we are trained to place high importance on information such as symmetry, space group or thermochemical stability. While structural information forms part of the inputs during annotation by the original authors contributing to the Materials Project, it is not used by the final charge/discharge predictive models at all. The experimental data set includes no structural information at all; not in the original annotation or in the anode/cathode predictive model. The extra (structural) science is important to human learning, but is not necessarily important to machine learning.

The materials and applications included in this study are an exemplar. The methods, workflow and code used in this project are general and openly available, making this an approach that can be easily applied to other complex compounds to train comprehensive models capable of confidently assigning hypothetical materials to specific applications ahead of extensive synthesis, simulation and characterisation. A practical application of this approach would be to rapidly screen millions of hypothetical chemical formulas to identify applications, to quickly estimate the impact of strategically adjusting relative concentrations, or replacing toxic or expensive elements with more environmentally or economically acceptable alternatives. The same approach could be used for other types of complex compounds or other applications, to expand this capability and assign a wide range of possible applications based exclusively on materials chemistry, saving time and identifying candidates worth researching in more detail.

CRediT authorship contribution statement

Zixin Zhuang: Writing – review & editing, Visualization, Validation, Software, Investigation, Formal analysis, Data curation. **Amanda S. Barnard:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

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.

Data availability

No data was generated as part of this work. All data sets used in this work are publicly available, and referenced in the bibliography. The computational data set in Section 3.1 is available from Materials Project at https://next-gen.materialsproject.org/batteries; the experimental data set in Section 3.2 was originally generated using ChemExtractor from https://github.com/ShuHuang/batterydatabase/tree/master/chemdataextractor_batteries as described in https://doi.org/10.1038/s41597-020-00602-2 is available from https://doi.org/10.6084/m9.figshare.11888115.v2 and cleaned using the code in the Supporting Information; the MXene data set in Section 4 was generated in https://doi.org/10.1021/ja508154e and is included in the Supporting Information.

Acknowledgements

Computational resources for this project have been supplied by the National Computational Infrastructure (NCI) national facility under Grant p00.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.commatsci.2024.113344.

References

[1] 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) 1–36.
[2] S. Axelrod, D. Schwalbe-Koda, S. Mohapatra, J.R. Damewood, K.P. Greenman, R. Gómez-Bombarelli, Learning matter: Materials design with machine learning and atomistic simulations, Accounts Mater. Res. 3 (2022) 343–357.
[3] P. Xu, H. Chen, M. Li, W. Lu, New opportunity: Machine learning for polymer materials design and discovery, Adv. Theory Simul. 5 (2022) 2100565.
[4] P. Xu, X. Ji, M. Li, W. Lu, Small data machine learning in materials science, npj Comput. Mater. 9 (2023) 1–15.
[5] J.R. Damewood, J. Karaguesian, J.R. Lunger, A.R. Tan, M. Xie, J. Peng, R. G'omez-Bombarelli, Representations of materials for machine learning, Annu. Rev. Mater. Res. 53 (2023) 399–426.
[6] G. Huang, Y. Guo, Y. Chen, Z. Nie, Application of machine learning in material synthesis and property prediction, Materials 16 (2023) 5977.
[7] M.H. Mobarak, M.A. Mimona, M.A. Islam, N. Hossain, F.T. Zohura, I. Imtiaz, M.I.H. Rimon, Scope of machine learning in materials research—A review, Appl. Surf. Sci. Adv. 18 (2023) 100523.
[8] D. Jha, L.T. Ward, A. Paul, W. keng Liao, A.N. Choudhary, C.M. Wolverton, A. Agrawal, ElemNet: Deep learning the chemistry of materials from only elemental composition, Sci. Rep. 8 (2018) 17593.
[9] L.T. Ward, A. Agrawal, A.N. Choudhary, C.M. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, arXiv: Mater. Sci. (2016) arXiv:1606.09551.
[10] L.T. Ward, A. Dunn, A. Faghaninia, N.E.R. Zimmermann, S. Bajaj, Q. Wang, J.H. Montoya, J. Chen, K. Bystrom, M.T. Dylla, K. Chard, M. Asta, K.A. Persson, G.J. Snyder, I.T. Foster, A. Jain, Matminer: An open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69.
[11] A.Y.-T. Wang, S.K. Kauwe, R. Murdock, T.D. Sparks, Compositionally restricted attention-based network for materials property predictions, npj Comput. Mater. 7 (2020) 1–10.
[12] R.E. Goodall, A.A. Lee, Predicting materials properties without crystal structure: Deep representation learning from stoichiometry, Nature Commun. 11 (2020) 6280.
[13] S. Li, A.S. Barnard, Inverse design of MXenes for high-capacity energy storage materials using multi-target machine learning, Chem. Mater. 34 (2022) 4964–4974, http://dx.doi.org/10.1021/acs.chemmater.2c00200.
[14] A. Alsaui, S.M. Alqahtani, F. Mumtaz, A.G. Ibrahim, A. Mohammed, A.H. Muqaibel, S.N. Rashkeev, A.A. Baloch, F.H. Alharbi, Highly accurate machine learning prediction of crystal point groups for ternary materials from chemical formula, Sci. Rep. 12 (2022) 1577.
[15] Z. Zhuang, A.S. Barnard, Structure-free Mendeleev encodings of material compounds for machine learning, Chem. Mater. 35 (2023) 9325–9338, http://dx.doi.org/10.1021/acs.chemmater.3c02134.
[16] Z. Zhuang, A.S. Barnard, Classification of battery compounds using structure-free Mendeleev encodings, J. Cheminformatics 16 (2024) 47, http://dx.doi.org/10.1186/s13321-024-00836-x.
[17] Z. Shen, H. Liu, Y. Shen, J. Hu, L.-Q. Chen, C.-W. Nan, Machine learning in energy storage materials, Interdiscip. Mater. 1 (2022) 175–195.
[18] Z. Wang, M. Yang, X. Xie, C. Yu, Q. Jiang, M. Huang, H. Algadi, Z. Guo, H. Zhang, Applications of machine learning in perovskite materials, Adv. Compos. Hybrid Mater. 5 (2022) 2700–2720.
[19] Y. Liu, X. Tan, J. Liang, H. Han, P. Xiang, W. Yan, Machine learning for perovskite solar cells and component materials: Key technologies and prospects, Adv. Funct. Mater. 33 (2023) 2214271.
[20] B. Kang, G. Ceder, Battery materials for ultrafast charging and discharging, Nature 458 (2009) 190–193.
[21] V. Palomares, M. Casas-Cabanas, E. Castillo-Martínez, M.H. Han, T. Rojo, Update on Na-based battery materials. A growing research path, Energy Environ. Sci. 6 (2013) 2312–2337.
[22] N. Nitta, F. Wu, J.T. Lee, G. Yushin, Li-ion battery materials: Present and future, Mater. Today 18 (2015) 252–264.
[23] J.-Y. Hwang, S.-T. Myung, Y.-K. Sun, Sodium-ion batteries: Present and future, Chem. Soc. Rev. 46 (2017) 3529–3614.
[24] S.K. Kauwe, T.D. Rhone, T.D. Sparks, Data-driven studies of Li-ion-battery materials, Crystals 9 (2019) 54.
[25] V.L. Deringer, Modelling and understanding battery materials with machine-learning-driven atomistic simulations, J. Phys.: Energy 2 (2020) 041003.
[26] S.R. Hui, W.W. Ho, A new generation of universal contactless battery charging platform for portable consumer electronic equipment, IEEE Trans. Power Electron. 20 (2005) 620–627.
[27] J.O. Besenhard, Handbook of Battery Materials, John Wiley & Sons, 2008.
[28] L.T. Ward, S.J. Babinec, E.J. Dufek, V. Viswanathan, M. Aykol, D.A. c. Beck, B. Blasizk, B.-R. Chen, G.W. Crabtree, V.D. Angelis, P. Dechent, M. Dubarry, E.E. Eggleton, D.P. Finegan, I.T. Foster, C.B. Gopal, P.K. Herring, V.W. Hu, N.H. Paulson, Y. Preger, D.U. Sauer, K.A. Smith, S. Sripad, T.R. Tanim, L. Tao, Principles of the battery data genome, Joule 6 (2021) 2253–2271.
[29] O. Furat, D.P. Finegan, D.R. Diercks, F.L.E. Usseglio-Viretta, K.A. Smith, V. Schmidt, Mapping the architecture of single lithium ion electrode particles in 3D, using electron backscatter diffraction and machine learning segmentation, J. Power Sources 483 (2021) 229148.
[30] M.R. Kunz, E.J. Dufek, Z. Yi, K.L. Gering, M.G. Shirk, K.A. Smith, B.-R. Chen, Q. Wang, P. Gasper, R.L. Bewley, T.R. Tanim, Early Battery Performance Prediction for Mixed Use Charging Profiles Using Hierarchal Machine Learning, vol. 4, 2021, pp. 1186–1196.
[31] Z. Wei, Q. He, Y. Zhao, Machine learning for battery research, J. Power Sources 549 (2022) 232125.
[32] S. Kim, Z. Yi, M.R. Kunz, E.J. Dufek, T.R. Tanim, B.-R. Chen, K.L. Gering, Accelerated battery life predictions through synergistic combination of physics-based models and machine learning, Cell Rep. Phys. Sci. 3 (2022) 101023.
[33] T. Jayakumar, N.M. Gowda, R. Sujatha, S.N. Bhukya, G. Padmapriya, S. Radhika, V. Mohanavel, M. Sudhakar, R. Sathyamurthy, Machine learning approach for prediction of residual energy in batteries, Energy Rep. 8 (2022) 756–764, http://dx.doi.org/10.1016/j.egyr.2022.10.027, The 2022 International Conference on Energy Storage Technology and Power Systems.
[34] P. Gasper, N. Collath, H.C. Hesse, A. Jossen, K.A. Smith, Machine-learning assisted identification of accurate battery lifetime models with uncertainty, J. Electrochem. Soc. 169 (2022) 080518.
[35] W. Zhang, P. Sayavong, X. Xiao, S.T. Oyakhire, S.B. Shuchi, R.A. Vilá, D.T. Boyle, S.C. Kim, M.S. Kim, S.E. Holmes, Y. Ye, D. Li, S.F. Bent, Y. Cui, Recovery of isolated lithium through discharged state calendar ageing, Nature 626 7998 (2024) 306–312.
[36] A. Fly, B. Wimarshana, I. Bin-Mat-Arishad, M. Sarmiento-Carnevali, Temperature dependency of diagnostic methods in lithium-ion batteries, J. Energy Storage 52 (2022) 104721.
[37] P. Gasper, A. Schiek, K.A. Smith, Y. Shimonishi, S. Yoshida, Predicting battery capacity from impedance at varying temperature and state of charge using machine learning, Cell Rep. Phys. Sci. 3 (2022) 101184.
[38] T. Yu, C. Wang, H. Yang, F. Li, Machine learning in metal-ion battery research: advancing material prediction, characterization, and status evaluation, J. Energy Chem. 90 (2024) 191–204.
[39] J. Li, M. Zhou, H. Wu, L. Wang, J. Zhang, N. Wu, K. Pan, G. Liu, Y. Zhang, J. Han, X. Liu, X. Chen, J. Wan, Q. Zhang, Machine learning-assisted property prediction of solid-state electrolyte, Adv. Energy Mater. (2024).
[40] D.P. Finegan, J. Zhu, X. Feng, M. Keyser, M. Ulmefors, W. Li, M.Z. Bazant, S.J. Cooper, The application of data-driven methods and physics-based learning for improving battery safety, Joule 5 (2021) 316–329.
[41] Y. Liu, B. Guo, X. Zou, Y. Li, S. Shi, Machine learning assisted materials design and discovery for rechargeable batteries, Energy Storage Mater. 31 (2020) 434–450.
[42] G. Houchins, V. Viswanathan, An accurate machine-learning calculator for optimization of Li-ion battery cathodes, J. Chem. Phys. 153 (5) (2020) 054124.
[43] 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, IEEE, 2014, pp. 1–3.

[44] 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 (69) (2018) 39414–39420.
[45] 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, http://dx.doi.org/10.1063/1.4812323.
[46] F. Zhou, M. Cococcioni, C.A. Marianetti, D. Morgan, G. Ceder, First-principles prediction of redox potentials in transition-metal compounds with LDA+u, Phys. Rev. B 70 (2004) 235121, http://dx.doi.org/10.1103/PhysRevB.70.235121.
[47] S. Huang, J.M. Cole, A database of battery materials auto-generated using ChemDataExtractor, Sci. Data 7 (2020) 260.
[48] C. Eames, M.S. Islam, Ion intercalation into two-dimensional transition-metal carbides: Global screening for new high-capacity battery materials, J. Am. Chem. Soc. 136 (2014) 16270–16276.
[49] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[50] L. Breiman, R.A. Olshen, Points of significance: Classification and regression trees, Nat. Methods 14 (2017) 757–758.
[51] L. Breiman, Bagging predictors, Mach. Learn. 24 (2004) 123–140.
[52] R. Duroux, E. Scornet, Impact of subsampling and tree depth on random forests, Esaim: Probab. Stat. 22 (2018) 96–128.
[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. P., R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[54] C. Liu, Z.G. Neale, G. Cao, Understanding electrochemical potentials of cathode materials in rechargeable batteries, Mater. Today 19 (2016) 109–123.
[55] A.S. Barnard, B. Motevalli, A.J. Parker, J.M. Fischer, C.A. Feigl, G. Opletal, Nanoinformatics, and the big challenges for the science of small things, Nanoscale 11 (2019) 19190–19201, http://dx.doi.org/10.1039/C9NR05912A.
[56] P. Nzereogu, A.D. Omah, F.I. Ezema, E.I. Iwuoha, A.C. Nwanya, Anode materials for lithium-ion batteries: A review, Appl. Surf. Sci. Adv. 9 (2022) 100233.
[57] A. Manthiram, A reflection on lithium-ion battery cathode chemistry, Nature Commun. 11 (2020) 1550.
[58] M. Greaves, S. Barg, M.A. Bissett, Mxene-based anodes for metal-ion batteries, Batteries Supercaps 3 (2019) 214–235.
[59] H. Kaland, J. Hadler-Jacobsen, F.H. Fagerli, N.P. Wagner, Z. Wang, S.M. Selbach, F. Vullum-Bruer, K. Wiik, S.K. Schnell, Are MXenes suitable as cathode materials for rechargeable mg batteries? Sustain. Energy Fuels 4 (2020) 2956–2966.