ACS central science

http://pubs.acs.org/journal/acscii

Research Article

Machine-Learning Approach for Predicting the Discharging Capacities of Doped Lithium Nickel–Cobalt–Manganese Cathode Materials in Li-Ion Batteries

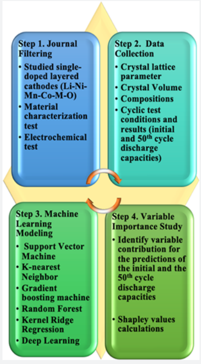
Guanyu Wang, Tom Fearn, Tengyao Wang,* and Kwang-Leong Choy*

Cite This: *ACS Cent. Sci.* 2021, 7, 1551–1560

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Understanding the governing dopant feature for cyclic discharge capacity is vital for the design and discovery of new doped lithium nickel–cobalt–manganese (NCM) oxide cathodes for lithium-ion battery applications. We herein apply six machine-learning regression algorithms to study the correlations of the structural, elemental features of 168 distinct doped NCM systems with their respective initial discharge capacity (IC) and 50th cycle discharge capacity (EC). First, a Pearson correlation coefficient study suggests that the lithium content ratio is highly correlated to both discharge capacity variables. Among all six regression algorithms, gradient boosting models have demonstrated the best prediction power for both IC and EC, with the root-mean-square errors calculated to be 16.66 $mAhg^{-1}$ and 18.59 $mAhg^{-1}$, respectively, against a hold-out test set. Furthermore, a game-theory-based variable-importance analysis reveals that doped NCM materials with higher lithium content, smaller dopant content, and lower-electronegativity atoms as the dopant are more likely to possess higher IC and EC. This study has demonstrated the exciting potentials of applying cutting-edge machine-learning techniques to accurately capture the complex structure–property relationship of doped NCM systems, and the models can be used as fast screening tools for new doped NCM structures with more superior electrochemical discharging properties.


Step 1. Journal Filtering
• Studied single-doped layered cathodes (Li-Ni-Mn-Co-M-O)
• Material characterization test
• Electrochemical test
Step 2. Data Collection
• Crystal lattice parameter
• Crystal Volume
• Compositions
• Cyclic test conditions and results (initial and 50th cycle discharge capacities)
Step 3. Machine Learning Modeling
• Support Vector Machine
• K-nearest Neighbor
• Gradient boosting machine
• Random Forest
• Kernel Ridge Regression
• Deep Learning
Step 4. Variable Importance Study
• Identify variable contribution for the predictions of the initial and the 50th cycle discharge capacities
• Shapley values calculations

1. INTRODUCTION

The unprecedented increase in the demand for clean energy has accelerated the research for discovering new lithium-ion batteries with higher energy density, higher power density, and more steady cyclic performance. Cathodes, in particular, have received a considerable amount of attention due to their current high cost, arising from the use of expensive cobalt metals, and the limited capacity that cannot fulfill the current demand.[1]

Among the various cathode candidates, layered cathodes have achieved tremendous market success due to their high practical capacity and wide operating voltage window. Quinary oxides (e.g., $LiNi_xCo_yMn_zO_2$) are currently the state of the art layered cathode materials as they integrates the superior properties of all three fundamental layered materials: $LiCoO_2$ (high kinetics), $LiNiO_2$ (high capacity), and $LiMnO_2$ (high safety). The nature of its broad compositional space has enabled scientists to discover new and robust electrochemical compounds such as $LiNi_{0.33}Co_{0.33}Mn_{0.33}O_2$ (NCM333), $LiNi_{0.50}Co_{0.20}Mn_{0.30}O_2$ (NCM523), $LiNi_{0.60}Co_{0.20}Mn_{0.20}O_2$ (NCM622), and $LiNi_{0.80}Co_{0.10}Mn_{0.10}O_2$ (NCM811).[2−4] It is important to note that the different transition metals in these compounds play different roles during electrochemical reactions: the nickel ion acts as the main active component during redox reactions, as it has the most diverse range of oxidation states among all of the atoms used. Manganese helps to stabilize the overall structure, while cobalt can effectively prohibit the cation mixing effect between Li ions and Ni ions. Furthermore, the mixing ratio of each transition metal (TM) in the material can bring different benefits to the cathode's properties. A higher concentration of nickel can greatly improve the overall capacity, as opposed to the benefits of higher kinetics and better safety from increasing the respective concentrations of cobalt and manganese.[4]

A common bottleneck issue is encountered during the selection of the optimal mixing ratio of these TMs to reach all desirable cathode properties (i.e., high kinetics, high stability, high capacity). The underlying reasons are the compositional space being too broad to be explored experimentally and the unavoidable benefit tradeoffs from TM substitution. A wide range of studies has been conducted in doping the quinary oxide system with a trace amount of cation atoms to enhance the cathode's electrochemical capability with minimal disturbance to the properties of the original crystal structure. Several successful cases have been made using various doping elements such as Al,[5] Fe,[5] Cu,[6] Cr,[7,8] Mg,[9] Mo,[10] K,[11] Pb,[12] Ti,[13] Si,[14] and Sn.[15] In general, two major benefits can be attained through the doping method. The first benefit can be seen from the hindering of the migration of Ni^{2+} into the Li^+

Received: June 9, 2021
Published: August 20, 2021

layer to reduce the anionic mixing during the intercalation reactions. The second benefit is to increase the strength of the TM−O bond to improve the overall structural stability and reduce the oxygen release during charge−discharge cycling. Nevertheless, the diverse available doping sites (Li, Ni, Co, Mn) and the large compositional space have inevitably increased the difficulty of identifying the most suitable dopant for each NCM-derived cathode material. The conventional approach to characterize the electrochemical properties of a new doped system is through conducting repetitive experiments, which is costly and time-consuming. Another approach based on first-principles computational modeling is also hindered by the great computing cost for studying very large supercell systems. To conquer these shortcomings, this paper reports the use of the robust data learning and analyzing features of machine learning to investigate the linkages among various doping factors and the experimental cyclic performance of doped NCM cathodes.

Machine learning (ML) methods have become increasingly popular across different fields of research currently. Min et al.[16] implemented seven different algorithms to predict the cycling properties of a Ni-rich NCM cathode from the corresponding synthesis parameters and reached an average prediction score of R^2 = 0.833. Houchins et al.[17] implemented DFT-based neural network models to predict the structure energy and forces of various forms of NCM materials (e.g., 111, 532, 811, and 622) and achieved a promising prediction accuracy of 3.7 meV/atom and 0.13 eV/Å, respectively. Allam et al.[18] constructed a deep learning model and attained a prediction error of 3.54% for predicting the redox potential of organic materials. From these works, data quality is frequently reported as an influential factor for model performance. Although databases such as the Inorganic Crystal Structure Database (ICSD)[19] and Materials project[20] are widely accessible for ML training, there is still a lack of an established large database in experimentally measured material properties: in particular, the measured discharging properties of various cathode materials in Li-ion batteries. From our previous work,[21] we had successfully curated a data set of 102 doped spinel cathodes containing the elemental and structural information and discharge performance. In addition, small prediction errors of 11.90 mAhg^{-1} and 11.77 mAhg^{-1} were achieved by the gradient-boosting machine models for the prediction of the initial and 20th cycle discharge capacities. These promising results further encouraged us to curate a more high quality discharge performance data set for the layered NCM cathode and implement ML to reveal the complex structure−property relationship.

In this work, 168 distinctive doped NCM systems are collected carefully with strict selection rules, as described in Figure 1. The data set contains 3696 data entries that cover 20 variations of dopants for all doped NCM-derivative material classes (NCM-333, NCM-523, NCM-622, and NCM-811). First, a Pearson correlation coefficient study was performed to investigate the collinearity of every variable pair. Furthermore, six nonlinear algorithms, including gradient-boosting machine, random forest, kernel ridge regression, feedforward deep learning, *k*-nearest neighbors, and support vector machine, were implemented with the design given in Table 1 to predict the initial discharge capacities and 50th cycle discharge capacities of the doped NCMs on the basis of 20 covariates (e.g., material characterization results, experimental parameters, elemental properties). By a comparison of their electrochemical performance against a hold-out test set, the best models can be identified for each discharge capacity prediction task. Furthermore, a variable-importance study was performed with the best-performing model to reveal the key doping features that governed the accurate predictions of discharge performance of the doped NCM systems. These insights would greatly enhance the current understanding of the doping effects and facilitate the design of future experimental work (e.g., independent variable selection and the doping concentrations).

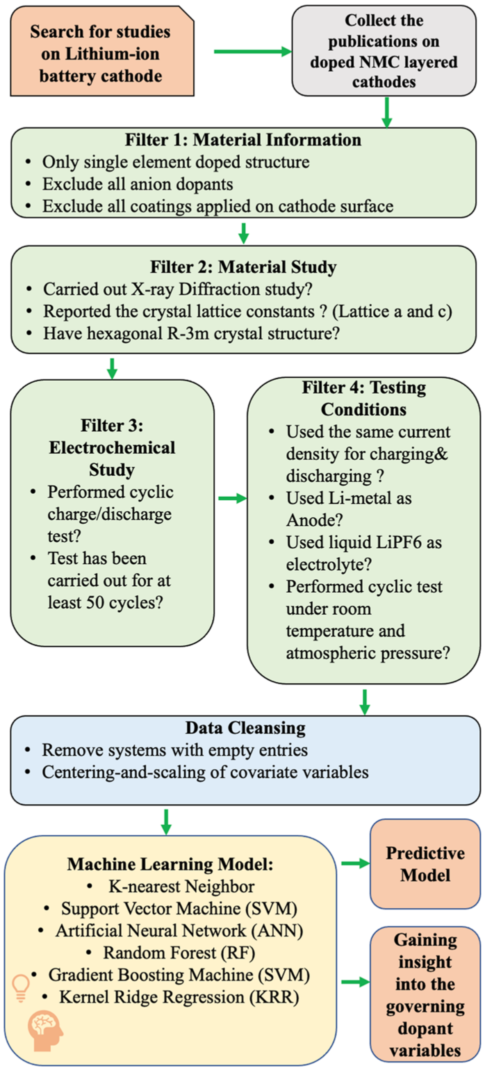

Search for studies on Lithium-ion battery cathode
Collect the publications on doped NMC layered cathodes
Filter 1: Material Information
• Only single element doped structure
• Exclude all anion dopants
• Exclude all coatings applied on cathode surface
Filter 2: Material Study
• Carried out X-ray Diffraction study?
• Reported the crystal lattice constants ? (Lattice a and c)
• Have hexagonal R-3m crystal structure?
Filter 3: Electrochemical Study
• Performed cyclic charge/discharge test?
• Test has been carried out for at least 50 cycles?
Filter 4: Testing Conditions
• Used the same current density for charging& discharging ?
• Used Li-metal as Anode?
• Used liquid LiPF6 as electrolyte?
• Performed cyclic test under room temperature and atmospheric pressure?
Data Cleansing
• Remove systems with empty entries
• Centering-and-scaling of covariate variables
Machine Learning Model:
• K-nearest Neighbor
• Support Vector Machine (SVM)
• Artificial Neural Network (ANN)
• Random Forest (RF)
• Gradient Boosting Machine (SVM)
• Kernel Ridge Regression (KRR)
Predictive Model
Gaining insight into the governing dopant variables

Figure 1. Overview of the data collection process with the demonstration of every filter applied in each publication selection stage.

2. METHODS

2.1. Data Collection of the Doped NCM Layered Materials. The data set consists of 168 different doped spinel systems with 20 dopant variations (e.g., Al, Ce, Cr, Cu, Cs, Eu, Fe, La, Mo, Mg, Nd, Na, Nb, Ru, Rb, Sn, Ti, V, Y, and Zr) and was curated from over 59 reliable journals published from 1998 through 2020 (given as Table S3 in the Supporting Information). During the journal selection, strict rules were applied to ensure a highly consistent standard of the collected data: NCM materials should be (i) singly doped with cation ions since the multiply doped systems are hard to fabricate and

Table 1. Proposed ML Model Architecture of This Study, Including the Names and Abbreviations of the Covariate Variables and Response Variables

Covariate Variables			
Publication Results		Elemental Properties	
Name	Abbreviation	Name	Abbreviation
The ratio of **lithium**, **nickel**, **cobalt**, **manganese**, **dopant** in the material formula	Li, Ni, Co, Mn, M - dopant	Material molar mass	Mr
Crystal lattice constants "*a*" and "*c*"	LC_a, LC_c	**Dopant**'s molar mass	Mr_dopant
Crystal Volume	CV	**Dopant**'s number of electrons	No_electron_dopant
Experimental current density	CD	**Dopant**'s electronegativity	EN_dopant
Minimum and maximum cyclic voltage	V_min, V_max	**Dopant**'s number of isotopes	No_iso_dopant
		Dopant's first ionization energy	E_ionisation_dopant
		Dopant's electron affinity	EA_dopant
		Dopant's atomic radius	AR_dopant
		Dopant's ionic radius	IR_dopant
Response Variables			
Name	Abbreviation	Name	Abbreviation
Initial discharge capacity	IC	50th cycle end discharge capacity	EC

are more costly, (ii) be single phase, (iii) have a space group of $R\bar{3}m$, and (iv) have no surface coating. Furthermore, the electrochemical testing should also fulfill the following criteria to meet the data collection requirements: (i) at least 50 cycles were performed for charging/discharging cyclic tests, (ii) lithium foil was used as the anode and nonaqueous $LiPF_6$ as the electrolyte, (iii) a constant current density was applied for charging and discharging the battery, (iv) the cyclic tests were carried out under the atmospheric conditions (i.e., temperature 25 ± 5 °C, pressure 1 atm). It is also important to note that the 50th cycle discharge capacity has been chosen, as it is the most performed test cycle among all studies.

The electrolyte plays a significant role in bridging the two contrasting electrodes and in facilitating the formation of a solid−electrolyte interface layer to protect the electrode from any unwanted side reactions. Electrolytes are often a mixed system with a solvent and additives that could lead to different performances if not standardized. The types of electrolyte systems from our collected studies are summarized in Figure S3. In our data set, nearly 71% of the investigating electrochemical tests were performed from either a mixture of ethylene carbonate and dimethyl carbonate (1/1 v/v) or the ethylene carbonate, dimethyl carbonate, and ethyl methyl carbonate (1/1/1 v/v). These systems have similar dielectric constants (Table S2), which should result in similar electrochemical performance. Only seven of the material systems have been tested with the addition of fluoroethylene carbonates, and these were used to improve the battery operation safety and hence should not influence the overall data quality by a considerate amount.

2.2. Model Training.

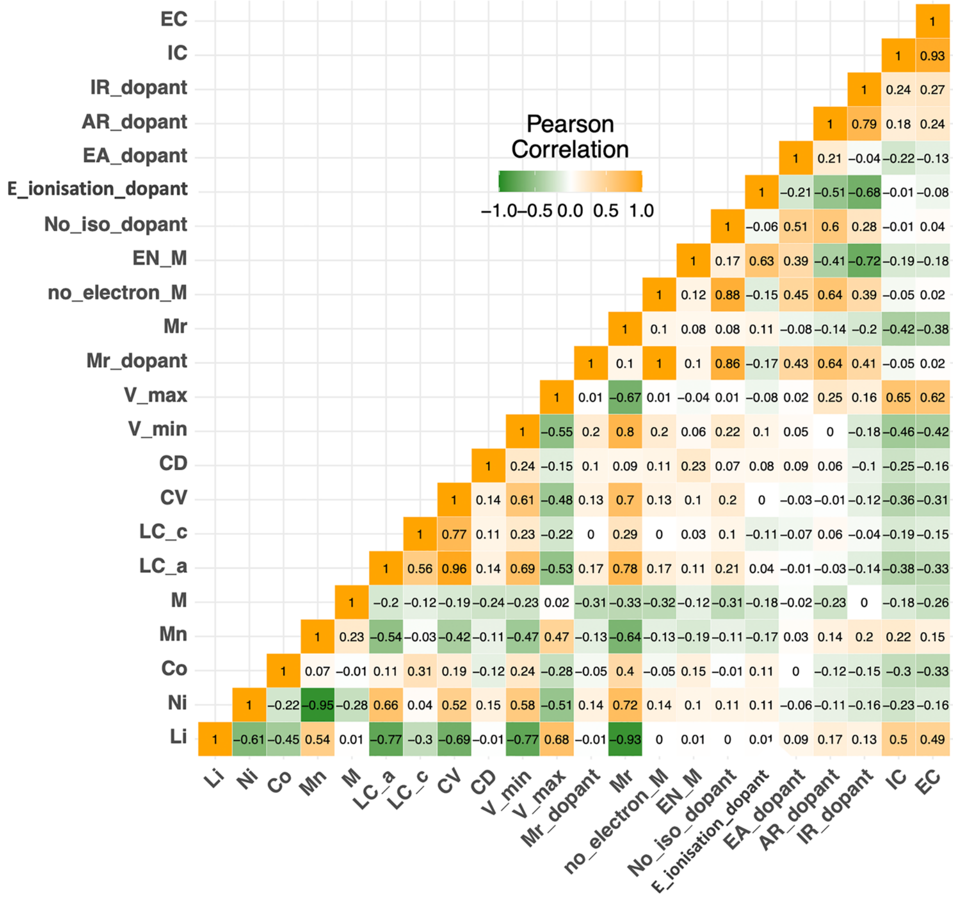
The ML models used in this work were trained using the Python programming language and with its relevant ML libraries (Sciki-learn, Pandas). Within the model, 20 covariate variables were selected to predict the initial and 50th cycle discharge capacities of each material. These cover the experimental results such as the crystal lattice constants (*a* and *c*), the formula ratio of lithium, nickel, manganese, cobalt, and dopant in the material formula (Li, Ni, Mn, Co, M), the material molar mass, the volume of the unit cell (CV), and cyclic parameters such as the charge/discharge current density (CD) as well as the upper and lower operating voltage limits (V_min, V_max). In addition, seven dopant elemental properties were chosen as covariate variables to reveal their correlations with the discharging properties. These include the dopant molar mass, the number of electrons, the electronegativity, the electron affinity, the first ionization energy, the atomic radius, and the ionic radius. In this work, six nonlinear algorithms were implemented, including artificial neural network (ANN), random forest (RF), gradient-boosting machine (GBM), support vector machine (SVM), kernel ridge regression (KRR), and *k*-nearest neighbors (KNN). The whole data space was randomly split into a ratio of 4:1, corresponding to the model training set and test set, respectively. Model hyperparameters were optimized using 5-fold cross-validation during model training, as there are 134 sets in the training set, and the optimized hyperparameters are given in Table S1.

2.2.1. *Model Evaluation Metrics.*

The model performance was evaluated through the calculation of the root-mean-square error (RMSE) and the coefficient of determination (R^2) from the predictions against the training and test sets. The calculation methods are given as eqs 1 and 2

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$



	Li	Ni	Co	Mn	M	LC_a	LC_c	CV	CD	V_min	V_max	Mr_dopant	Mr	no_electron_M	EN_M	No_iso_dopant	E_ionisation_dopant	EA_dopant	AR_dopant	IR_dopant	IC	EC
EC																						1
IC																					1	0.93
IR_dopant																				1	0.24	0.27
AR_dopant																			1	0.79	0.18	0.24
EA_dopant																		1	0.21	−0.04	−0.22	−0.13
E_ionisation_dopant																	1	−0.21	−0.51	−0.68	−0.01	−0.08
No_iso_dopant																1	−0.06	0.51	0.6	0.28	−0.01	0.04
EN_M															1	0.17	0.63	0.39	−0.41	−0.72	−0.19	−0.18
no_electron_M														1	0.12	0.88	−0.15	0.45	0.64	0.39	−0.05	0.02
Mr													1	0.1	0.08	0.08	0.11	−0.08	−0.14	−0.2	−0.42	−0.38
Mr_dopant												1	0.1	1	0.1	0.86	−0.17	0.43	0.64	0.41	−0.05	0.02
V_max											1	0.01	−0.67	0.01	−0.04	0.01	−0.08	0.02	0.25	0.16	0.65	0.62
V_min										1	−0.55	0.2	0.8	0.2	0.06	0.22	0.1	0.05	0	−0.18	−0.46	−0.42
CD									1	0.24	−0.15	0.1	0.09	0.11	0.23	0.07	0.08	0.09	0.06	−0.1	−0.25	−0.16
CV								1	0.14	0.61	−0.48	0.13	0.7	0.13	0.1	0.2	0	−0.03	−0.01	−0.12	−0.36	−0.31
LC_c							1	0.77	0.11	0.23	−0.22	0	0.29	0	0.03	0.1	−0.11	−0.07	0.06	−0.04	−0.19	−0.15
LC_a						1	0.56	0.96	0.14	0.69	−0.53	0.17	0.78	0.17	0.11	0.21	0.04	−0.01	−0.03	−0.14	−0.38	−0.33
M					1	−0.2	−0.12	−0.19	−0.24	−0.23	0.02	−0.31	−0.33	−0.32	−0.12	−0.31	−0.18	−0.02	−0.23	0	−0.18	−0.26
Mn				1	0.23	−0.54	−0.03	−0.42	−0.11	−0.47	0.47	−0.13	−0.64	−0.13	−0.19	−0.11	−0.17	0.03	0.14	0.2	0.22	0.15
Co			1	0.07	−0.01	0.11	0.31	0.19	−0.12	0.24	−0.28	−0.05	0.4	−0.05	0.15	−0.01	0.11	0	−0.12	−0.15	−0.3	−0.33
Ni		1	−0.22	−0.95	−0.28	0.66	0.04	0.52	0.15	0.58	−0.51	0.14	0.72	0.14	0.1	0.11	0.11	−0.06	−0.11	−0.16	−0.23	−0.16
Li	1	−0.61	−0.45	0.54	0.01	−0.77	−0.3	−0.69	−0.01	−0.77	0.68	−0.01	−0.93	0	0.01	0	0.01	0.09	0.17	0.13	0.5	0.49

Figure 2. Results matrix of Pearson coefficient correlations for every pair of variables in the data set, including covariate variables, Li, Ni, Co, Mn, M, LC_a, LC_c, CV, V_min, V_max, CD, Mr, Mr_dopant, No_electron_M, EA_dopant, No_iso_dopant, AR_dopant, IR_dopant, E_ionisation_dopant, and EN_M, and two response variables, IC and EC. The estimated correlation values are distributed within the range of −1 to +1, with a number reaching either end value implying a more perfect negative correlation and positive correlation, respectively.

Where n is the number of values, y_i is the observed variable, $\hat{y}_i$ is the predicted values and $\bar{y}$ is the average of the observed values.

The SHAP summary plots for the variable correlation and importance ranking are generated using the SHAP python package[22] and the further instruction are available in https://github.com/slundberg/shap.

2.3. Safety Statement. This work is performed wholly on the machine learning computational indicated and hence no unexpected or unusually high safety hazards were encountered.

3. RESULTS AND DISCUSSION

3.1. Pearson Coefficient Correlation Study. To gain initial insight into the underlying variable correlations, a Pearson correlation coefficient study was performed for every pair of variables retrieved in the data set. Figure 2 shows the matrix of correlation values (R) calculated for the 20 covariates and the two response variables. The extent of correlation between every pair is color-coded, with darker orange indicating a strong positive correlation and dark green a strong negative correlation.

From the computed R values, it can be seen that there are more strong correlations ($R > 0.75$) observed between covariates variables than for the covariate variables with either of the two response variables. However, some of these strong correlations observed between covariates might be misleading and do not provide any intuitive insights. For instance, the high correlations of lithium content ratio with minimum operating voltage ($R = -0.77$) and maximum operating voltage ($R = 0.68$) do not imply that any change in lithium content would influence the value of operating voltages. These operating voltage values are often preset for the experiments on the basis of the specifications of the testing machine. In addition, a decrease in the lithium ratio seems to increase the molar mass of the material ($R = -0.93$), and this is because there are more available crystal lattice sites for the occupancies of heavier-weighted TM and dopant elements. Similarly, the manganese molar ratio appears to have a correlation value of −0.95 with the nickel molar ratio in the formula, which is potentially due to a direct TM crystal site substitution. In addition, a high correlation ($R = 0.79$) is also identified for the pair of ionic radius of the dopant ion (IR_dopant) and the atomic radius of the dopant atom (AR_dopant). Both radii are measurements of the distance away from the central nucleus despite the fact that one is for the neutral state and the other is for the charged state and therefore their values might have a high linearity correlation with each other. For the model construction, it is important to find the linkage of covariate variables to the electrochemical properties. First, no strong correlations are observed between covariate variables and response variables, which might be due to the presence of nonlinear correlations. The maximum cyclic voltage is found to have a relatively high correlation with both IC and EC at 0.65 and 0.62, respectively. In addition, an increase in the Li content ratio in the formula

Table 2. Comparisons of the Mean RMSE Values during the 5-Fold Cross-Validation and for Testing against the Holdout Test and the R^2 Test Score Computed by Six Nonlinear Models, for the Prediction of Initial Discharge Capacity and the 50th Cycle End Discharge Capacities of Doped NCM Cathode Systems

ML techniques	Initial Discharge Capacity (mAhg^{-1})			50th Cycle End Discharge Capacity (mAhg^{-1})		
	Cross-validated RMSE mean	RMSE on the test set	R^2 scores on the Test Set	Cross-validated RMSE mean	RMSE on the test set	R^2 scores on the Test Set
GBM	20.26	16.66	0.76	21.53	18.59	0.64
RF	21.98	17.00	0.59	21.86	19.25	0.42
SVM	22.94	21.11	0.37	22.00	19.38	0.41
KRR	20.65	17.28	0.58	21.77	19.13	0.43
KNN	23.57	18.98	0.49	25.03	21.51	0.28
ANN	34.15	22.39	0.29	33.93	24.58	0.05

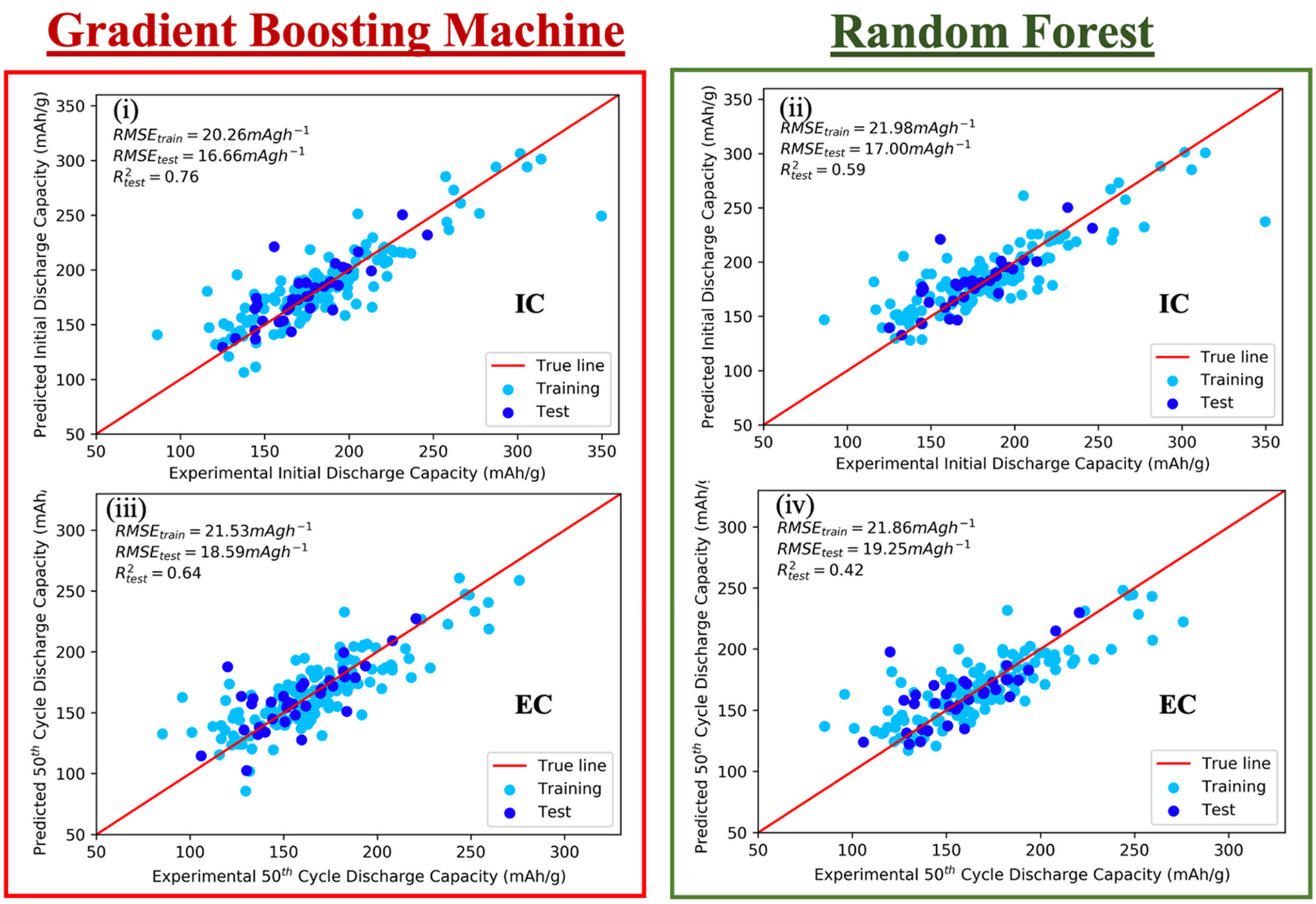


Figure 3. Scatter plots of the experiment values against the predicted values for the prediction of initial discharge capacity and the 50th cycle end discharge capacity of doped NCM cathode systems computed by gradient-boosting models (i.e. (i) and (iii), respectively(, and random forest (i.e. (ii) and (iv), respectively).

seems to suggest a partial increase in both IC and EC, as their correlation values are calculated to be positive: 0.5 and 0.49, respectively. This seems to agree with the latest results on higher discharging performance being obtained from the lithium-rich layered cathode (~200 mAhg^{-1}) than from the normal NCM-111 compounds (~165 mAhg^{-1}).[23,24]

3.2. Model Performance Comparisons. To build accurate prediction models, six different nonlinear regression algorithms have been trained and validated against a holdout test set for their prediction powers. Table 2 shows the RMSE values computed during the training and predicting the holdout test set. R^2 values are also calculated to demonstrate

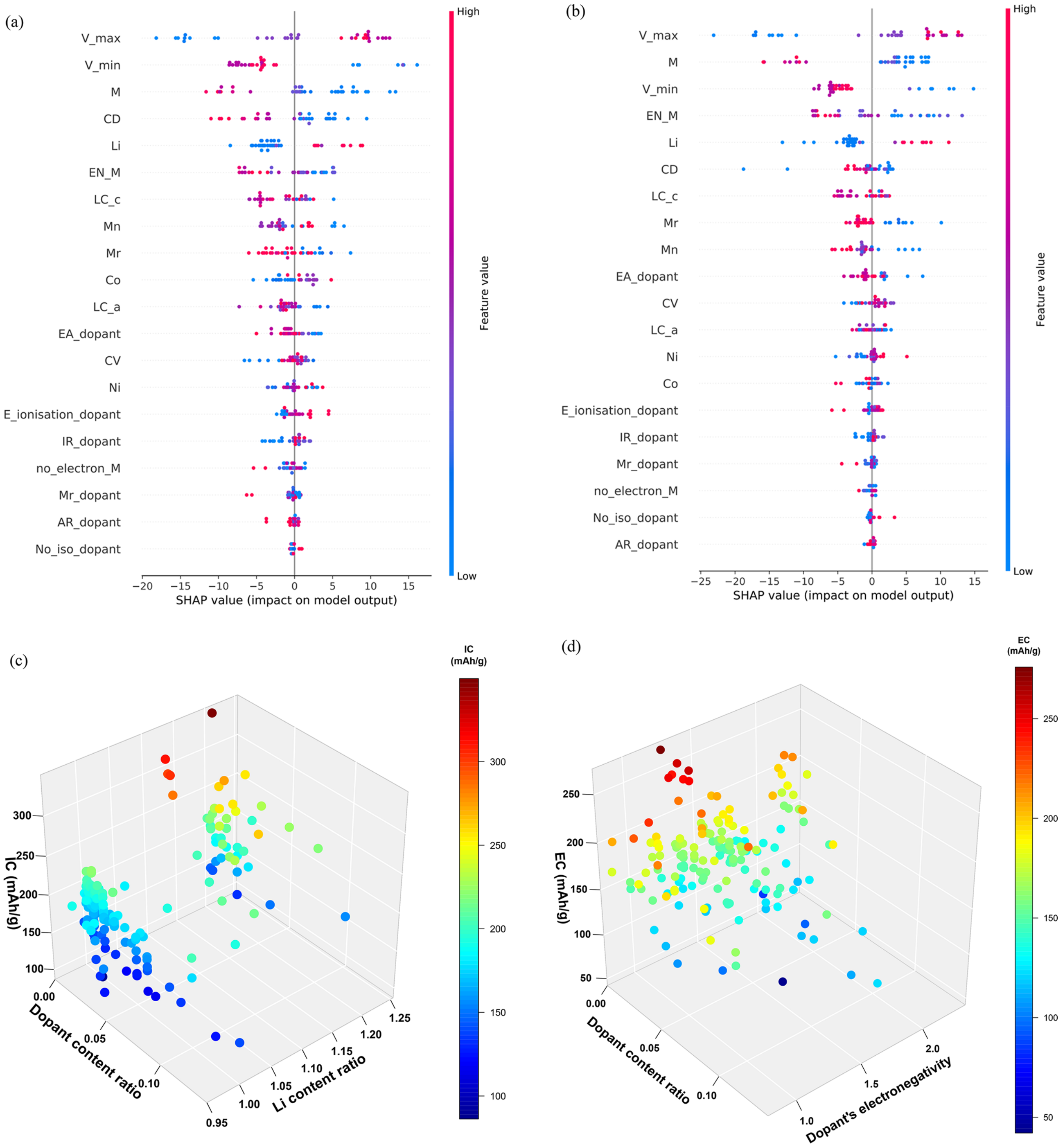


Figure 4. Summary plots for the feature contribution of 20 covariate variables in the test-set prediction of (a) IC generated on the basis of the GBM-IC model and (b) EC generated on the basis of the GBM-EC model. The *y*-axis indicates the feature importance of variables ranked in descending order. The *x*-axis shows the scale of the Shapley values for every feature and indicates their contribution to the prediction. The figure legends are given as a heat map showing the values of the respective response feature variable. The 3D plots give insights into the intercorrelations of (c) IC with the two most important variables (Li content ratio and dopant content ratio) and (d) EC with the two most important variables (dopant content ratio and dopant electronegativity) in the whole data set.

the proportion of variation in the test set being accurately captured by the model. In general, the validated test-set RMSE and R^2 values are more insightful for selecting the best-performing model as the data are not involved in the training process and hence contain less prediction bias. First, the ANN models are shown to have the worst performance with the lowest test-set RMSE among all. This is because both ANN models are embedded with many model parameters (see Table S1) and would require a much larger sample size to estimate these well. Furthermore, the tree-based ensemble methods generally have much lower test-set RMSE values in comparison to other nonlinear models such as SVM and KNN in addition to the KRR models. To help with visualizing the prediction mechanism of the tree-based ensemble methods, we have included diagrammatic illustrations for the random forest and gradient-boosting algorithms and they are given in Figures S1 and S2, respectively, in the Supporting Information. In addition, a plot of one decision tree generated from the random forest model for the prediction of the 50th cycle discharge capacity is given to illustrate the predicting process and this is given in Figure S4, in the Supporting Information. Overall, it is seen that the gradient-boosting machine (GBM) has the best prediction performance for both tasks, with the test set RMSE values being the lowest among all, at 16.66 mAhg^{-1} and 18.59 mAh g^{-1} respectively.

In the efforts to resolve the capacity-fading issues faced by a layered cathode material, a wide range of mathematical modeling-based studies have been conducted to understand the time-series-based changes in NCM capacity with the loss of active materials.[25,26] Although several insights were gained in these studies, the inconsistent change in the capacity for different material compositions and the influence of other essential testing conditions such as current density remain unresolved. It is estimated that an experimental cyclic test of 25 cycles for a newly assembled Li-ion battery with a discharge rate of C/10 (1 C denotes the discharge current density that would full discharge the battery within 1 h) can take up to 20 days to complete.[27] Hence, the establishment of a highly accurate predictive model would greatly reduce the time required for the testing of electrochemical properties and shorten the timespan for discovering new and robust cathode materials. The construction of a highly accurate model for predicting discharge capacities normally requires a large amount of experimental data with high diversity in material compositions and a good consistency in experimental factor controls. This has prompted us to implement strict selection rules for the journals and use high-quality data to train and build predictive models that would best describe the changes in discharging capacities for various NCM compounds at the initial and the 50th cycle.

Gradient-boosting machine (GBM) algorithms[28] have been known to be robust in describing the nonlinear correlations across the wide variable space. GBM has previously seen successful applications in the prediction of the bulk and shear moduli of zeolites,[29] the classification of metal and insulators of inorganic crystals,[30] and prediction of the band gap of new hybrid (organic + inorganic) perovskites.[31]

Figure 3 shows scatter plots of the predicted and experimental values of the initial and 50th cycle end discharge capacities during the training and testing stages for the optimal GBM models and RF models. The values of the R^2 scores and RMSE values for the test set prediction are highlighted in the graph along with the mean training RMSE values which are averaged across the 5-fold cross-validation. First, all models have shown good ability in generalizing the training set, as all of the 134 training points are shown to be close to the red 45° line. For the given test data set, the GBM models have much higher R^2 scores for both prediction tasks than the RF models and this suggests superior prediction power in capturing the variations in the new data set. These high correlation scores are shown to be consistent with the low test-set RMSE for all GBM models. However, a few outliers can be identified from the training and test sets at ca. 225 mAhg^{-1} and 250 mAhg^{-1}, respectively, from the GBM-EC graph and this would potentially affect the R^2 scores.

Nevertheless, the correlation scores from the GBM-IC (R^2 = 0.76) and GBM-EC (R^2 = 0.64) models have both exceeded the benchmark value of $R^2 > 0.6$ for a model to be considered as predictive.[32] These high correlation values have indicated that structural and elemental parameters such as the crystal lattice dimension and dopant ionic radius can predict the discharge capacity of a layered doped NCM cathode as accurately as the synthesis parameters variables used in the work of Min et al.[16] On the basis of the above results, GBM models for both the IC and EC predictions were chosen for further analysis.

3.3. Variable-Importance Studies.

The covariate variable importance can be estimated through the calculation of the Shapley values from the best-performing model predictions of the holdout test set. Shapley values come from the coalitional game theory, where each of the covariate variables is treated as an individual "player" and the values estimate the covariate variables' contribution to the final prediction of a response variable instance. This method is more desirable than the traditional permutation method for an easier interpretation of the variable correlation with the response variable. In this project, the treeSHAP (Shapley additive explanations) method, proposed by Lundberg et al.,[33] is used to gain insight into the importance of all covariate variables and their feature effect on the prediction. Figure 4a,b shows the summarized Shapley values for all 20 covariate variables during the predictions of IC and EC in the test set through the GBM-IC model and the GBM-EC model, respectively. The Shapley values measure the effect of that covariate variable on the model prediction, with a more positive or negative value implying a larger overall influence. The *y* axis of each graph gives the list of covariates in the order of their contribution to the overall prediction, with the most important one being at the top and the less important ones at the lower ranks.

To begin with, the minimum cutoff voltage, maximum cutoff voltage, and current density are ranked within the top 10 important variables. These covariate variables are all of the experimental conditions for cycling and thereby are expected to have great influences on a material's discharging performance as a cathode.[34] After excluding these experimental setting variables, one can see that the dopant content ratio and the lithium content ratio are ranked as the third and the fifth most important features, respectively, for the IC test-set prediction from Figure 4a.

A negative correlation is identified for the dopant content ratio and the IC, as an increase in the corresponding Shapley values leads to a decrease in the IC feature values (shown in the sequence of red to blue). In contrast, the Shapley values of lithium content ratio are shown to be positively correlated with the IC values, with the color of the data plot shown to be blue to red. Figure 4c shows a 3D plot of the dopant content ratio and Li content ratio correlating with the respective IC values for the entire data set (train + test). First, two clusters of data can be identified, with one characterized as having a lower Li content ratio with different dopant content ratios and the other having a higher Li content ratio and lower dopant content ratio. Observations can be made such that the IC values increase (change from blue to green) as the dopant content ratio is reduce in the first cluster. Moreover, the latter cluster has much higher average IC values in comparison to the first cluster, which implies that a higher Li content is generally more desirable for obtaining a high IC value. This indicates that a higher Li content ratio ($x > 1.20$) coupled with a lower dopant content ratio ($y < 0.02$) can reach a higher IC. Further key insights can be also gained from Figure 4a in that a doped NCM cathode material formula with a lower dopant electronegativity (EN_M), shorter lattice constants a (LC_a) and c (LC_c), a smaller formula molar mass, less manganese content, and more cobalt content can lead to higher IC values.

Figure 4b shows that the dopant content ratio and its electronegativity value are ranked as the second and the fourth most important for the predictions of EC, respectively. Electronegativity measures the dopant element's ability to attract electron pairs toward itself. A dopant's EN controls the bonding strength with the surrounding TMs and oxygen atoms and influences the structural stability as well as the overall crystal structure density. During a long cyclic charging and discharging performance, the overall crystal structure often becomes unstable, which then triggers significant lattice collapses and leads to severe capacity fading.[35] The involvement of dopant content can greatly improve the structure stability by forming stronger bonds, while the strategy with doping with a small amount can ensure that no second material phase is formed and also that the whole crystal structure is not modified significantly to disturb the Li-ion intercalation/deintercalation mechanisms. Both the content ratio and the electronegativity of dopant are demonstrated to be negatively correlated with the EC feature value, as the color of the trend changes from red to blue (left to right). Figure 4d displays the 3D intercorrelation of the two dopant-related covariate variables with the respective EC in the whole data set. A clear trend is observed for EC decreasing with a decrease in the dopant content ratio (from blue to red). In addition, high-EC data are observed to be at the lower range between 1.25 and 1.5 for the dopant's electronegativity when the dopant content ratio is kept low ($x > 0.02$) and this corresponds to the magnesium (1.31) and zirconium (1.33) dopants in the collected data set. From these phenomena, it is suggested that doping the atom with an electronegativity of closer to 1.5 and with doping with a smaller amount can lead to higher EC values. Other observations can be made from Figure 4b that a smaller material molar mass with lower manganese and higher nickel content can lead to a higher EC value for using doped NCM materials as the cathode, which shares a great deal of similarities with the previous findings in the IC variable correlations. Interestingly, the dopant ratio in the material formulas has been shown to be the most influential factor, as it is ranked the third and the second most important for IC and EC, respectively, and this is much higher in comparison to other material properties such as the dopant's electronegativity value and the lithium content ratio in the material formula. This suggests that the doping amount might play a much more important role in influencing the discharge capacities in comparison other material systematic properties.

To conclude, our results have demonstrated that the materials that constitute both high IC and EC share the common characteristics of high Li content ratio, small dopant ratio, small manganese ratio, and doping with atoms of low- to middle-range electronegativity and low electron affinity. In addition, the design of a doped NCM material with low formula molar mass is also encouraged, as it is inversely related to both discharge capacities.

3.4. Overall Discussions.

Although some of the obtained correlations were known qualitatively, our model gives new insights by providing a quantitative prediction of IC and EC using these features for any new cathode materials that practitioners want to experiment on. These quantitative correlations were identified through the use of the Shapley value method developed from the coalitional game theory, which to our knowledge was the first in the field to implement this theory into analyzing the contribution of the doping features for the predictions of the LIB discharge capacities.

In addition, the results of our research are novel in that they give an estimation of the importance of each of the material property related features for each capacity property. For instance, despite the fact that a higher lithium content is more favorable for achieving higher IC and EC as identified in the paper, its importance as given in Figure 4a,b is shown to be much less than that of the dopant content feature. This could be a suggestion for the experimentalist to consider the factor of optimal dopant ratio first before considering the lithium content ratio in the formula in the design of experiments.

During the selection of covariate variables (input variables) for the machine-learning model, two major criteria have been used to guide this process: (i) the relevance of the feature as reflected in the material properties and the performance properties and (ii) whether such data is widely reported or collectable. We selected the variables that can best describe the properties of synthesized materials to reflect wholly the differences in synthesis methods andraw materials used across different research groups. For example, properties including the crystal volume and crystal structure lattice constant of the materials can reflect the conditions of the cathode materials as the host for the Li ions. These properties are completely dependent on what the authors have reported in their publications, and therefore we have not introduced any bias in the selection of these. In addition, we included the elemental properties related to the dopant atoms used in the studies from the NCM material data set. As indicated in the initial results of the Pearson correlation coefficient matrix, no strong linear correlations have been identified for the pairs of covariate and response variables, which indicate that the correlations between the selected covariate variables cannot possibly be explained by a simple linear model. This suggests that the selection process of the covariate variables in our project contains little bias.

Simple correlations of the structure and properties for an NCM material can be observed if the investigating material system is fixed. For example, the researcher could be investigating the effects of one dopant with a different concentration on the discharge performance of the NCM material. On the other hand, the interpretation of a large data set containing different doped NCM material systems is extremely hard to achieve through simple human intuition. The novelty of our work lies in investigating a much wider range of doped NCM materials with 168 different compositions and 20 different dopant elements. We introduced this machine-learning method to gain much broader insights into the overall variable correlations of different types of doped NCM materials to promote a much broader understanding of the doping effects on the NCM materials' electrochemical performance and the relevant governing variables in each case.

3.5. Remaining Challenges and Future Improvements.

Data quality is essential for building highly predictive ML models. In this section, the data collection challenges for this work are highlighted and discussed along with the recommendation made for future potential research. First, the doped NCM materials involved in this project are all composite materials with the variations seen in the mixing ratio of the remaining two components: namely, the conductive additive and binder. The lack of standardization in the conductive additive and binder to be used has led to a large variation in material use across different research teams. The roles of these materials are to stabilize the overall cathode structure and to promote the Li-ion mobility within the structure, which are considered essential for long cycle discharging. Furthermore, since the active material is considered to be the major component (75−90%) of this composite, the information on the conductive additives and binder is assumed to be standardized for all collected data and further research could be done in investigating their effects. Second, the effects of the microstructural properties of the material (e.g. particle sizes) and the morphological features on the material discharging properties have been extensively studied.[36−40] Such information, however, is very difficult to collect, due to the reporting of various particle reporting scales (e.g., D10, D50, D90) as well as the general lack of conducting cathode surface studies. For an electrochemical test, information such as the surface area of the cathode material and volume of the electrolyte is often misreported and, since the elemental composition might be unevenly distributed in the whole of the cathode composite system, this could lead to unreasonable fluctuations in the capacity loading.

Despite a great deal of effort being devoted to establishing strict journal filters and selecting suitable journals (Figure 1) in this work, there is still room for improvement to be made as discussed above. To fully unleash the power of ML for the application of predicting the futuristic discharge performances of an NCM cathode, the following points are worthy of consideration for future experimental and modeling research in the NCM cathode.

(1) Fully report the cathode material information such as the surface compositions, primary and secondary particle sizes, and pore size.

(2) For the full cell electrochemical test, report key information such as the surface area of the cathode and anode materials.

(3) Conduct comparative studies on the changes in microstructure and crystal structures for the cathode material after long cycle discharging.

(4) For a small data set (less than 500 rows), implement tree-based algorithms such as random forest and gradient boosting first before constructing an artificial neural network, as this construction has been shown to be time-consuming and less efficient in predicting capacities.

4. CONCLUSIONS

Analyzing the past experimental results is a crucial step to better understand the complex correlations of the doped NCM system properties and their discharging performance, and additionally, the outcome of this project demonstrated the feasibility of using machine-learning techniques in doing so. Six various nonlinear machine-learning algorithms have been trained and validated with the manually curated experimental results of 168 doped NCM materials. The models are built from 13 material physical properties and 7 dopant elemental properties as covariate variables to predict the initial (IC) and 50th cycle (EC) discharge capacities of each material structure. First, a Pearson coefficient correlation study has indicated that no strong linear correlations are captured for any pairs of covariate variables and two response variables. In addition, gradient-boosting models have been proven to hold the best prediction power against the holdout test set for having the lowest root-mean-square errors of 16.66 mAhg^{-1} and 18.59 mAhg^{-1} and the highest R^2 scores of 0.76 and 0.64 during IC and EC predictions, respectively. Further insights are gained into the governing material features for each discharging property. NCM materials with higher lithium content, smaller dopant content, and doping with a lower electronegativity value atom seem to give higher values in both IC and EC. From these promising results, we expect that these machine-learning models can be used as a guide to estimate the discharging properties of any singly doped NCM material and potentially discover new cathode materials with more advanced electrochemical properties.

ASSOCIATED CONTENT

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author. The in Supporting Information. The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscentsci.1c00611.

> Data set used for constructing these machine learning algorithms is available on the GitHub page (https://github.com/thepowerligand/NCM-ML/blob/main/NMC_numerical_new.csv) and the references for these selected journals are given (PDF)

AUTHOR INFORMATION

Corresponding Authors

Tengyao Wang − *Department of Statistical Science, University College London, London WC1R 7HB, United Kingdom;* Email: tengyao.wang@ucl.ac.uk

Kwang-Leong Choy − *Institute for Materials Discovery, Faculty of Maths and Physical Sciences, University College London, London WC1E 7JE, United Kingdom;* orcid.org/0000-0002-5596-4427; Email: k.choy@ucl.ac.uk

Authors

Guanyu Wang − *Institute for Materials Discovery, Faculty of Maths and Physical Sciences, University College London, London WC1E 7JE, United Kingdom;* orcid.org/0000-0003-1736-5797

Tom Fearn − *Department of Statistical Science, University College London, London WC1R 7HB, United Kingdom;* orcid.org/0000-0003-2222-6601

Complete contact information is available at: https://pubs.acs.org/10.1021/acscentsci.1c00611

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

All authors contributed equally to this work. This research received no grants from any funding agencies. The codes used for creating the machine-learning models in this paper are given in https://github.com/thepowerligand/NCM-ML.

REFERENCES

(1) Manthiram, A. A reflection on lithium-ion battery cathode chemistry. *Nat. Commun.* **2020**, *11*, 1−9.

(2) Xu, J.; Lin, F.; Doeff, M. M.; Tong, W. A review of Ni-based layered oxides for rechargeable Li-ion batteries. *J. Mater. Chem. A* **2017**, *5*, 874−901.

(3) Schipper, F.; et al. Review—Recent Advances and Remaining Challenges for Lithium Ion Battery Cathodes. *J. Electrochem. Soc.* **2017**, *164*, A6220−A6228.

(4) Chakraborty, A.; et al. Layered Cathode Materials for Lithium-Ion Batteries: Review of Computational Studies on LiNi1- x-yCoxMnyO2 and LiNi1- x- yCoxAlyO2. *Chem. Mater.* **2020**, *32*, 915−952.

(5) Lee, K. K.; Yoon, W. S.; Kim, K. B.; Lee, K. Y.; Hong, S. T. Characterization of LiNi0.85Co0.10M0.05O2 (M = Al, Fe) as a cathode material for lithium secondary batteries. *J. Power Sources* **2001**, *97−98*, 308−312.

(6) Sa, Q.; Heelan, J. A.; Lu, Y.; Apelian, D.; Wang, Y. Copper Impurity Effects on LiNi1/3Mn1/3Co1/3O2 Cathode Material. *ACS Appl. Mater. Interfaces* **2015**, *7*, 20585−20590.

(7) Nisar, U.; et al. Synthesis and electrochemical characterization of Cr-doped lithium-rich Li1.2Ni0.16Mn0.56Co0.08-xCrxO2 cathodes. *Emergent Mater.* **2018**, *1*, 155−164.

(8) Sun, Y.; Xia, Y.; Noguchi, H. The improved physical and electrochemical performance of LiNi0.35Co0.3-xCrxMn0.35 O2 cathode materials by the Cr doping for lithium ion batteries. *J. Power Sources* **2006**, *159*, 1377−1382.

(9) Jin, Y.; Xu, Y.; Ren, F.; Ren, P. Mg-doped Li 1.133 Ni 0.2 Co 0.2 Mn 0.467 O 2 in Li site as high-performance cathode material for Li-ion batteries. *Solid State Ionics* **2019**, *336*, 87−94.

(10) Breuer, O.; et al. Understanding the Role of Minor Molybdenum Doping in LiNi $_{0.5}$ Co $_{0.2}$ Mn $_{0.3}$ O $_{2}$ Electrodes: from Structural and Surface Analyses and Theoretical Modeling to Practical Electrochemical Cells. *ACS Appl. Mater. Interfaces* **2018**, *10*, 29608−29621.

(11) Yang, Z.; et al. K-doped layered LiNi0.5Co0.2Mn0.3O2 cathode material: Towards the superior rate capability and cycling performance. *J. Alloys Compd.* **2017**, *699*, 358−365.

(12) Zhang, X.; Xiong, Y.; Dong, M.; Hou, Z. Pb-Doped Lithium-Rich Cathode Material for High Energy Density Lithium-Ion Full Batteries. *J. Electrochem. Soc.* **2019**, *166*, A2960−A2965.

(13) Markus, I. M.; Lin, F.; Kam, K. C.; Asta, M.; Doeff, M. M. Computational and experimental investigation of Ti substitution in Li1(NixMnxCo1−2x-yTiy)O2 for lithium ion batteries. *J. Phys. Chem. Lett.* **2014**, *5*, 3649−3655.

(14) Na, S. H.; Kim, H. S.; Moon, S. I. The effect of Si doping on the electrochemical characteristics of LiNi xMnyCO(1-x-y)O2. *Solid State Ionics* **2005**, *176*, 313−317.

(15) Qiao, Q. Q.; Qin, L.; Li, G. R.; Wang, Y. L.; Gao, X. P. Sn-stabilized Li-rich layered Li(Li0.17Ni0.25Mn0.58)O2 oxide as a cathode for advanced lithium-ion batteries. *J. Mater. Chem. A* **2015**, *3*, 17627−17634.

(16) Min, K.; Choi, B.; Park, K.; Cho, E. Machine learning assisted optimization of electrochemical properties for Ni-rich cathode materials. *Sci. Rep.* **2018**, *8*, 1−7.

(17) Houchins, G.; Viswanathan, V. An accurate machine-learning calculator for optimization of Li-ion battery cathodes. *J. Chem. Phys.* **2020**, *153*, 054124.

(18) Allam, O.; Cho, B. W.; Kim, K. C.; Jang, S. S. Application of DFT-based machine learning for developing molecular electrode materials in Li-ion batteries. *RSC Adv.* **2018**, *8*, 39414−39420.

(19) Zagorac, D.; Muller, H.; Ruehl, S.; Zagorac, J.; Rehme, S. Recent developments in the Inorganic Crystal Structure Database: Theoretical crystal structure data and related features. *J. Appl. Crystallogr.* **2019**, *52*, 918−925.

(20) Jain, A.; et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.

(21) Wang, G.; Fearn, T.; Wang, T.; Choy, K.-L. Insight Gained from Using Machine Learning Techniques to Predict the Discharge Capacities of Doped Spinel Cathode Materials for Lithium-Ion Batteries Applications. *Energy Technol.* **2021**, *9*, 2100053.

(22) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, 4766−4775.

(23) Toney, M. F. Li gradients for Li-rich cathodes. *Nature Energy* **2019**, *4*, 1014−1015.

(24) Xu, L.; et al. A Li-rich layered-spinel cathode material for high capacity and high rate lithium-ion batteries fabricated via a gas-solid reaction. *Sci. China Mater.* **2020**, *63*, 2435−2442.

(25) Carnovale, A.; Li, X. A modeling and experimental study of capacity fade for lithium-ion batteries. *Energy AI* **2020**, *2*, 100032.

(26) Plattard, T.; Barnel, N.; Assaud, L.; Franger, S.; Duffault, J.-M. Combining a Fatigue Model and an Incremental Capacity Analysis on a Commercial NMC/Graphite Cell under Constant Current Cycling with and without Calendar Aging. *Batteries* **2019**, *5*, 36.

(27) Kauwe, S.; Rhone, T.; Sparks, T. Data-Driven Studies of Li-Ion-Battery Materials. *Crystals* **2019**, *9*, 54.

(28) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189−1232.

(29) Evans, J. D.; Coudert, F. X. Predicting the Mechanical Properties of Zeolite Frameworks by Machine Learning. *Chem. Mater.* **2017**, *29*, 7833−7839.

(30) Isayev, O.; et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **2017**, *8*, 8.

(31) Lu, S.; et al. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **2018**, *9*, 1−8.

(32) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.

(33) Lundberg, S. M.; Erion, G. G.; Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv* (**2018**).

(34) Wu, Y.; Keil, P.; Schuster, S. F.; Jossen, A. Impact of Temperature and Discharge Rate on the Aging of a LiCoO 2 /LiNi 0.8 Co 0.15 Al 0.05 O 2 Lithium-Ion Pouch Cell. *J. Electrochem. Soc.* **2017**, *164*, A1438−A1445.

(35) Li, W.; Asl, H. Y.; Xie, Q.; Manthiram, A. Collapse of LiNi1- x-yCoxMnyO2 Lattice at Deep Charge Irrespective of Nickel Content in Lithium-Ion Batteries. *J. Am. Chem. Soc.* **2019**, *141*, 5097−5101.

(36) Tang, T.; Zhang, H. L. Synthesis and electrochemical performance of lithium-rich cathode material Li-[Li0.2Ni0.15Mn0.55Co0.1-xAlx]O2. *Electrochim. Acta* **2016**, *191*, 263−269.

(37) Gao, S.; Zhan, X.; Cheng, Y.-T. Structural, electrochemical and Li-ion transport properties of Zr-modified LiNi0.8Co0.1Mn0.1O2 positive electrode materials for Li-ion batteries. *J. Power Sources* **2019**, *410−411*, 45−52.

(38) Xue, L.; et al. Effect of Mo doping on the structure and electrochemical performances of LiNi0.6Co0.2Mn0.2O2 cathode material at high cut-off voltage. *J. Alloys Compd.* **2018**, *748*, 561−568.

(39) Lim, S. N.; et al. Rate capability for Na-doped Li1.167-Ni0.18Mn0.548Co0.105O2 cathode material and characterization of Li-ion diffusion using galvanostatic intermittent titration technique. *J. Alloys Compd.* **2015**, *623*, 55−61.

(40) Kim, U. H.; Myung, S. T.; Yoon, C. S.; Sun, Y. K. Extending the Battery Life Using an Al-Doped Li[Ni0.76Co0.09Mn0.15]O2 Cathode with Concentration Gradients for Lithium Ion Batteries. *ACS Energy Lett.* **2017**, *2*, 1848−1854.