Using Artificial Intelligence to Predict the Discharge Performance of Cathode Materials for Lithium-ion Batteries Applications

Guanyu Wang

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Institute for Materials Discovery

University College London

April 4, 2022

I, Guanyu Wang, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

A comprehensive understanding of the composition-structure-property relationships for doped cathode materials used in lithium-ion batteries remains lacking which delays the progress of developing new cathode materials. This thesis proposes that machine learning (ML) techniques can be used to predict the discharge capacities of the cathode materials whilst revealing these underlying relationships. To achieve this, the data for three different doped cathodes are curated from the publications, namely, the doped spinel cathode, $LiM_xMn_{2-x}O_4$, the M-doped nickel-cobalt-manganese layered cathode, $LiNi_xCo_yMn_zM_{1-x-y-z}O_2$, and the carbon-coated and doped olivine cathode, $C/LiM_1M_2PO_4$ (M_1, M_2 denote different metal ions). Several linear and non-linear ML models are trained with the data and compared for the power of predicting initial and higher cycle discharge capacity.

Gradient boosting models have shown the best prediction power for predicting the initial and 20^{th} cycle end discharge capacity of 102 doped spinel cathode and the initial and 50^{th} cycle discharge capacity of 168 doped nickel-cobalt-manganese layered cathodes. For the doped spinel cathode, higher discharge capacities at both cycles can be achieved through increasing the material formula mass, reducing the crystal lattice constant and using dopants with smaller electronegativity. For the doped layered cathodes, it is revealed that the higher lithium content, lower formula molar mass, small doping content and doped with low electronegativity dopant are more likely to possess greater capacities at both cycles.

Bayesian ridge regression and gradient boosting model are shown to have the highest prediction power over the initial and the 20^{th} cycle discharge capacity of carbon-coated and doped olivine cathode. In addition, the olivine systems with lower

dopant content, higher base-metal content and smaller unit cells are shown to be more likely to possess higher capacities at both cycles.

Finally, future research directions are presented including the suggestion of involving other new input variables and using principal component analysis and feature selection algorithms to use to improve the model performance.

Impact Statement

Sustainable energy sources such as wind and solar generate power intermittently depending on the season, time of day, and weather conditions. This generation is not consummate with energy demand. To circumvent this, energy storage can be used to store the excess electric energy to be released during periods of high demand. Among all battery technologies, lithium-ion batteries are promising for this application as they have high specific energy and power density and high cell nominal voltage up to 4.2V. This is much higher than other alternative rechargeable batteries such as Nickel-Metal-Hydride (1.2V) and Lead-Acid battery (2.1V). Although LIB has been successfully applied to many energy storage applications, they also have several disadvantages such as high cost, short lifetime and risk of spontaneous combustion due to dendrite formation in the liquid electrolyte. These issues inhibit the application of lithium-ion batteries for electric vehicles or other large industrial electricity loading devices. Lithium-ion batteries have three principal components: an anode, cathode and electrolyte material. Among these, the cathode is a key target for the improvement of battery working potential, capacity and cost, as it has a much lower average storage capacity than the contrasting anode electrode whilst its stand-alone cost is approximately 25% of the cost of a typical lithium-ion battery. Several studies have investigated the effects of various cathode synthesis methods, precursors, doping treatment and coating application, on electrochemical performance properties such as the initial discharge capacity, capacity fading rate and working potential. Since the relation between the crystal structure and a material property in some cases are not very well understood, an approach of trial-and-error experimenting has been adopted which is time-consuming, expensive

and largely dependent on the researcher's intuition, insight and knowledge, which, being fallible can lead research to the wrong direction.

Data-driven approaches integrate experimental and modelling results with sophisticated statistical modelling tools to predict properties of materials or devices, thereby guiding and informing experimental campaigns. These approaches have been successfully applied to predict the properties of cathode materials including the crystal structure, Li-storage capacity and optimal synthesis parameters. The success of these initial studies shows the applicability of this approach to investigate possible new cathode materials. However, there are still many opportunities to use these techniques to further investigate the relationship between cathode material properties and their respective discharge capacity.

This work demonstrates the potentials of using the data-driven method, and, machine learning algorithms, to predict the initial discharge capacity and a higher cycle discharge capacity of the doped material systems using a variety of their structural and elemental properties. The underlying variable correlations will also be analysed to give valuable insights into the governing variable for each capacity property. The results of this work is an essential milestone towards shortening the time needed for cyclic performance tests, improving the efficiency of developing new cathodes and eventually improve the performance of Li-ion batteries for broader applications.

List of Publications

1. Wang, G., Fearn, T., Wang, T. and Choy, K.-L. (2021) Insight gained from using machine learning techniques to predict the discharge capacities of doped spinel cathode materials for lithium-ion batteries applications. Energy Technol., DOI: 10.1002/ENTE.202100053.

2. Wang, G., Fearn, T., Wang, T. and Choy, K.-L. (2021) Machine learning approach for predicting the discharging capacities of doped lithium nickel-cobalt-manganese cathode materials in Li-ion batteries. ACS Cent. Sci., DOI: 10.1021/ACSCENTSCI.1C00611.

3. Machine learning prediction of the discharging capacities of carbon coated and doped Olivine cathode systems for Lithium-ion battery applications. In Preparation.

Acknowledgements

The essence of life is to constantly push forward the boundaries that define your comfort zone, and navigate your way through with great curiosity and ambitions. Undertaking this PhD whilst being surrounded by the uncertainties of COVID-19 has been tough. But I am glad that I have completed this journey with the supports of my family, research team and friends. First of all, I would like to pass my sincere appreciation to my primary supervisor, Professor Kwang-Leong Choy, for giving me this wonderful opportunity to undertake research in this novel field. To both of my amazing secondary supervisors, Prof. Tom Fearn and Dr. Tengyao Wang from UCL's Department of Statistical Science, thank you for introducing me to the field of data science and for giving valuable support during my research. Without all of your support, I wouldn't have made it this far.

Secondly, I would like to thank my parents, Liyan Gao, Dahui Wang; my grandparents, Shulan Du, Guangjia Wang, Guiqin Liu, Shuyou Gao; my aunties, Yueying Wang, Lixin Gao, Lina Gao. Thank you for the endless love and patience you all have given me along this way. You all have shaped me into who I am today and I hope that I do you all proud.

To Dr. Prospero Taroni Junior, thank you for being a loyal friend, a good listener, whenever I am going through a deep breakdown both in life and research.

To Dr. Matthew Jackson, thank you for always being there like a big brother to encourage me to think big of the research topics and bring ideas to alive.

To Dr. German Martinez, your encouragements have fuelled me for undertaking this research and thank you for the support you have given me along this way.

To Dr. Yzu-Ling Lin, thank you for always being there to listen and support me whenever I needed it. You are truly the sun that brings the light after the darkness.

To Minghou and Pieter, thank you for all the emotional support you both have given me the past few years. Meeting you both and having you both by my side during this tough journey, was the luckiest thing that has happened to me.

To Sagarika, Sutandra, Sudeshna, Puja, Udit, thank you for all the supports and kindness you have given to me during my PhD. You have certainly made my journey in London feel more homey.

I feel extremely fortunate to be a part of the amazing research team at Institute for Materials Discovery; Dr. Mingqing Wang, Dr. Mengyan Nie Dr. Oliver Walk, Dr. Antonio Ruiz Gonzalez, Dr. Andres Nistal Gonzalez, Dr. Yanbei Hou, Dr. Manni Yang, Ahmet Nazligul, Mariam Marzook, Quanyao Lin, Haobo Dong, Siyu Xiong, Yinan Lu, Pinyi Zhao. Thank you all for the amazing memories we shared and the valuable support we had given to each other.

To my extremely supportive colleagues/friends from the Euston house office, Gianna, Julia, Emily, Janina, thank you for all the reassurance and the kind words during the last part of my PhD marathon. You all have enlightened my way. I must thank the people who have given me a lot of encouragement and support during my PhD, Ashleigh, Matthew, Dr. Juanze Wang, Sienna, Gisela, Shilpa, Prash, Marzena, Manu. You all have made my PhD journey easier and more enjoyable.

Without any further ado, let's begin with the thesis ...

Contents

List of Figures

List of Tables

List of Abbreviations

Active_r	Ratio of the active component in cathode mixture
Adam	Adaptive Momentum Estimation
ANN	Artificial Neural Network
AR_dopant	Dopant atom's atomic radius
Avg_E_ion	Averaged first ionisation energy of the based-metal and the dopant metal
Avg_EN	Averaged electronegativity of the based-metal and the dopant metal
Avg_IR	Averaged ionic radius of the based-metal and the dopant metal
B	The molar ratio of base-metal (Fe and Mn) in olivine material
Binder_r	Ratio of the binder in cathode mixture
BRR	Bayesian Ridge Regression
C_ratio	Carbon Content ratio
CBD	Carbon/Binder Domain
CD	Charge and Discharge Current Density
Conduct_r	Ratio of the Conductive component in cathode mixture
CV	Volume of the unit cell of crystal structure
D/G	Ratio of the Disordered carbon content to Graphitized carbon content

DFT Density Functional Theory

DMC Dimenthyl Carbonate

DNN Deep neural network

DT Decision Tree

E_ion_dopant Dopant's first ionisation energy

EA_dopant Dopant's electron affinity

EC End discharge capacity (Cycle number depends on the project)

EGDMA Polyethylene glycol-dimethacrylate

EMC Ethyl methyl carbonate

EPDM Ethylene propylene diene monomer

FT-IR Fourier Transform Infrared Spectrum

GBM Gradient Boosting Machine

HOMO Highest occupied molecular

IC Initial discharge capacity

IR_dopant Dopant's ionic radius

JTD Jahn-Teller Distortion Effect

KNN K-Nearest Neighbour

KRR Kernel ridge regression

LASSO Least Absolute Shrinkage and Selection Operator

LC_a Lattice constant "a" of the unit cell in crystal lattice

LC_b Lattice constant "b" of the unit cell in crystal lattice

LC_c Lattice constant "c" of the unit cell in crystal lattice

LCO Lithium cobalt oxide ($LiCoO_2$)

LFP Lithium Iron Phosphate ($LiFePO_4$)

Li Lithium or the lithium molar ratio in the formula

LIB Lithium-ion Batteries

LMO Lithium manganese oxide ($LiMn_2O_4$)

LMP Lithium Manganese Phosphate (LiMnPO4)

LNO Lithium nickel oxide ($LiNiO_2$)

LR Least Absolute Shrinkage and Selection Operator Regression

LUMO Lowest unoccupied molecular orbital

M Dopant/Metal identity or its respective molar ratio

M_EN The electronegativity of the dopant

MGI Materials Genome Initiatives

ML Machine Learning

Mn Manganese or the manganese molar ratio in the material formula

Mr Molar mass

Mr_dopant Molar mass of the dopant

MSE Mean Squared Error

NCM Lithium nickel cobalt manganese oxide

No_electron_M Dopant atom's number of electrons

No_iso_dopant Dopant's number of isotopes

OLS Ordinary Least Squares

OOB Out-of-Bag Technique

PED Portable Electronic Devices

PEI Polyethyleneimine

PEO	Polyethylene oxide
PMMA	Polymethylmethacrylate
PVC	Polyvinyl chloride
PVDF	Polyvinylidenefluoride
RBF	Radial Basis Function
ReLu	Rectified Linear Unit
RF	Random Forest
RMSE	Root Mean Squared Error
RR	Ridge Regression
SEI	Solid Electrolyte Interphase
SHAP	SHapley Additive exPlanations
SSE	Sum of Squared Error
SVM	Support Vector Machine
T-n	The n^{th} time of hyperparameter optimisation
V_max	Maximum operating voltage
V_min	Minimum operating voltage
XRD	X-ray Diffraction

Chapter 1

Introduction

1.1 Background of Batteries Research

The development of new battery technology of higher energy and power density is vital to facilitate the transition from fossil fuels to renewable energy sources and to satisfy the portable energy demands of modern technological advancements. Renewable energy sources have the advantages that they are sustainable due to the unlimited raw materials, have lower carbon footprints, and lower futuristic costs relative to fossil fuels. **Figure 1.1-a** demonstrates the cost of energy for various fossils fuels and renewable energy sources. From the diagram, it is seen that the energy price per megawatt-hours generated through solar panels and wind is reducing dramatically across the ten years. This trend is in contrast to the fossil fuels options such as coal and natural gas where the respective energy cost continuously rising. In addition, the share of contribution from renewable energy sources is observed to be increasing from 1950 to 2019 (**Figure 1.1-b**). This demonstrates the great potentials of using a renewable energy source to replace a large portion of fossil fuel consumption in the near future. Nevertheless, solar cells and wind turbines are often limited to specific geographical locations (requiring the environment to be either rich in sunshine or windy) and have intermittent power generation which cannot be used to follow demand [32, 33, 34, 35, 36]. To circumvent this, Battery technologies can be applied to store the excess energy generated from renewable energy systems and release it during demand peak time [37]. A recent example of this can be seen from

the large scale implementation of lithium-ion batteries by Tesla in South Australia as the storage system for the wind farms and yielded promising results for meeting the domestic energy demand [38].

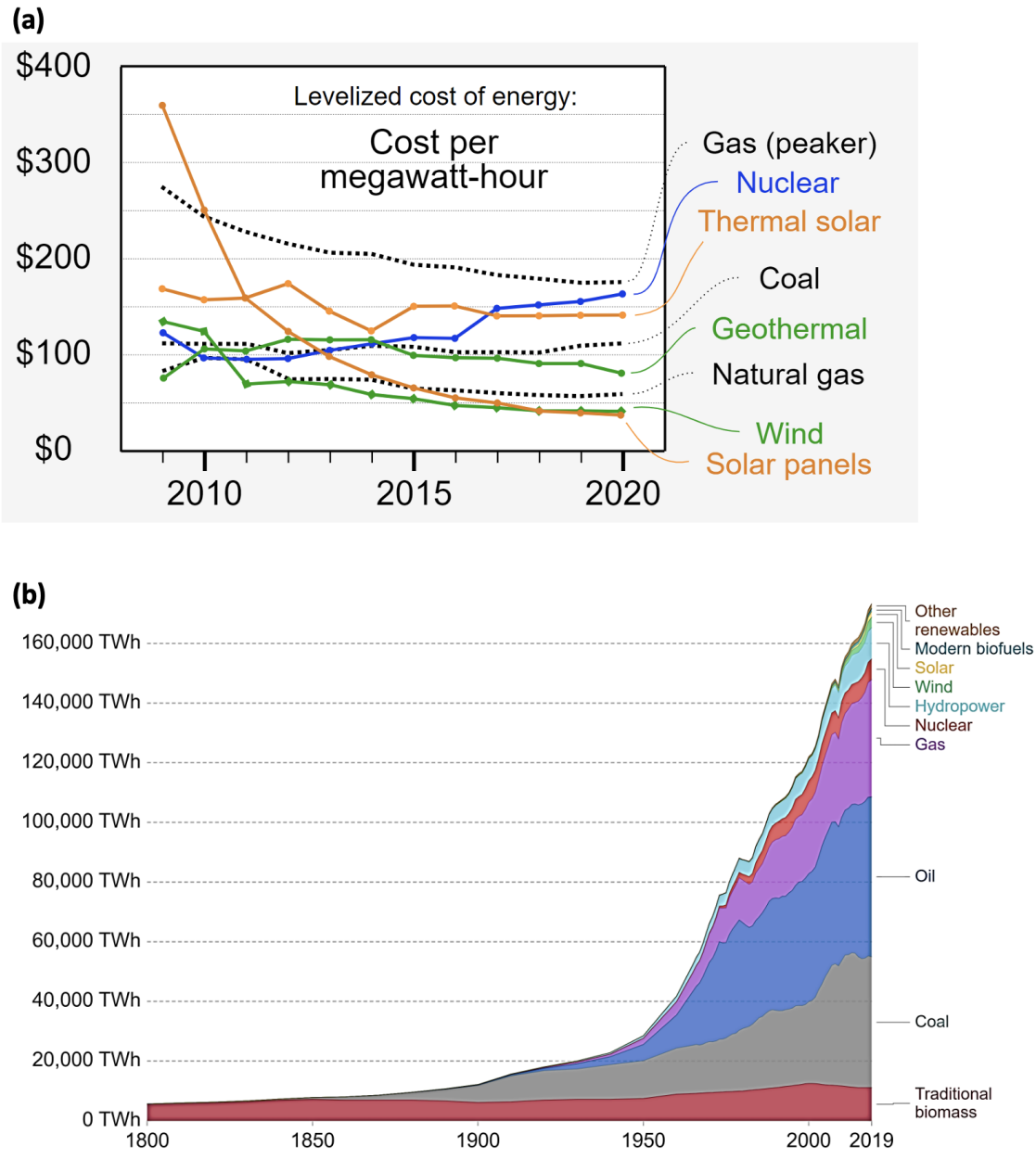


Figure 1.1: (a) The Levelized cost of energy (USD dollars per megawatt-hours) estimated from various energy sources (fossil fuels + renewable energy sources) from 2009 to 2020. The image was taken with permission from [1] (b) The consumption of the energy from various energy sources from 1800 to 2019. The image was taken with permission from [2].

A battery can be understood as a container that consists of one or more internal

components, cooperating to generate electric power. The typical battery cell designs are the cylindrical cell, button cell, prismatic cell and pouch cell. They all consist of a positive electrode(cathode) and a negative electrode(anode) that are responsible for generating electrical power and a separator that separates these components to ensure operating safety. A graphical illustration of the internal configuration of a prismatic cell is given in **Figure 1.2**.

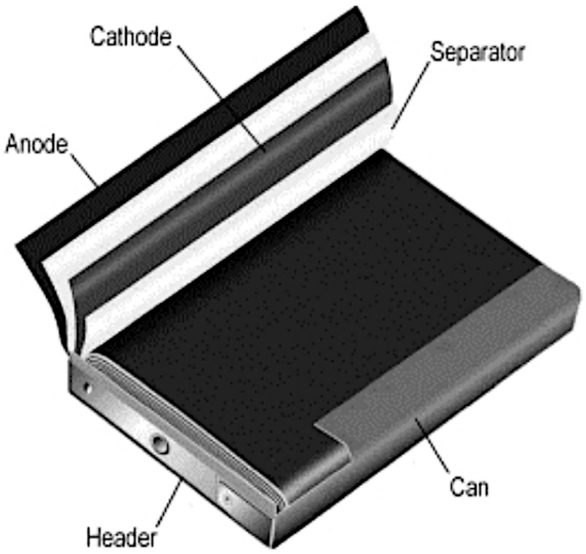


Figure 1.2: Diagrammatic illustration of the internal design of a prismatic cell. Reproduced with permission from [3].

The journey of battery research initiated from the dead-frog-leg experiment conducted by Luigi Galvani in 1780 where muscle twitching was observed after clipping two different metal pieces to the frog's leg and spinal cord. This phenomenon was later revealed as being the interaction of two metals by an experimentalist physicist named Volta and leads to the creation of the first battery known as "voltaic pile" in 1780 [39]. Since then, various research was done to expand the horizons of these metallic-based electric inducing devices. In 1859, the physician Gaston Planté invented the first rechargeable battery known as the Lead-acid battery [40]. Shortly after 1899, Waldmar Jungner introduced the Nickel-Cadmium battery and received a lot of market success due to the battery being capable of working under harsh working conditions [41]. This battery model was used as a base model for the later invention of the nickel-metal hydride battery in the late 1980s and yielded

similar market success [42].

On the other hand, the concept of intercalation electrodes was introduced by *Whittingham et al* [43] in 1979 and these have the capability to store a type of guest ions within the internal crystal lattices and are available for withdrawal efficiently to generate electricity whilst giving a better cyclability. Following this proposal, a major breakthrough was obtained by *John Goodenough et al* [44] of which the Li-ions have shown great promises for the use as guest ions during the transfer between the $LiCoO_2$ cathode material and the lithium metal. This rechargeable battery model was then referred to as lithium-ion batteries. In addition, the transition metal oxides are found to be suitable for this intercalation reaction feature, in particular, the $LiCoO_2$ material can provide a high voltage at 4.35V against the lithium metal [45]. Comparing to the rechargeable metallic lithium batteries, the Li-ion batteries (LIB) are seen as a relatively safe option as they are less suspectable to fire upon misusage [46].

In 1991, the first commercialised LIB was made by Sony corporations with the use of carbon being the contrasting electrode and since then many new manufacturers have followed their lead [47]. Another popular type of rechargeable battery was known to be lithium polymer batteries, of which the original design was made back in the 1970s and majorly differed by its use of dry and solid polymer as the electrolyte instead of the flammable liquid.

Among all types of rechargeable batteries, LIBs remain one of the most promising energy storage devices. **Figure 1.3** demonstrates the mass and volumetric power densities of rechargeable batteries. LIBs have much higher energy density than most of the other battery types, with nearly twice that of Ni-Cd and more than three times more than the lead-acid battery. Furthermore, LIBs are also favoured as they are relatively easy to manufacture because of the one-cell design and do not have the safety issues associated with Li metal designs. As such, several reviews have projected the LIB to dominate the renewable battery market going forward [48].

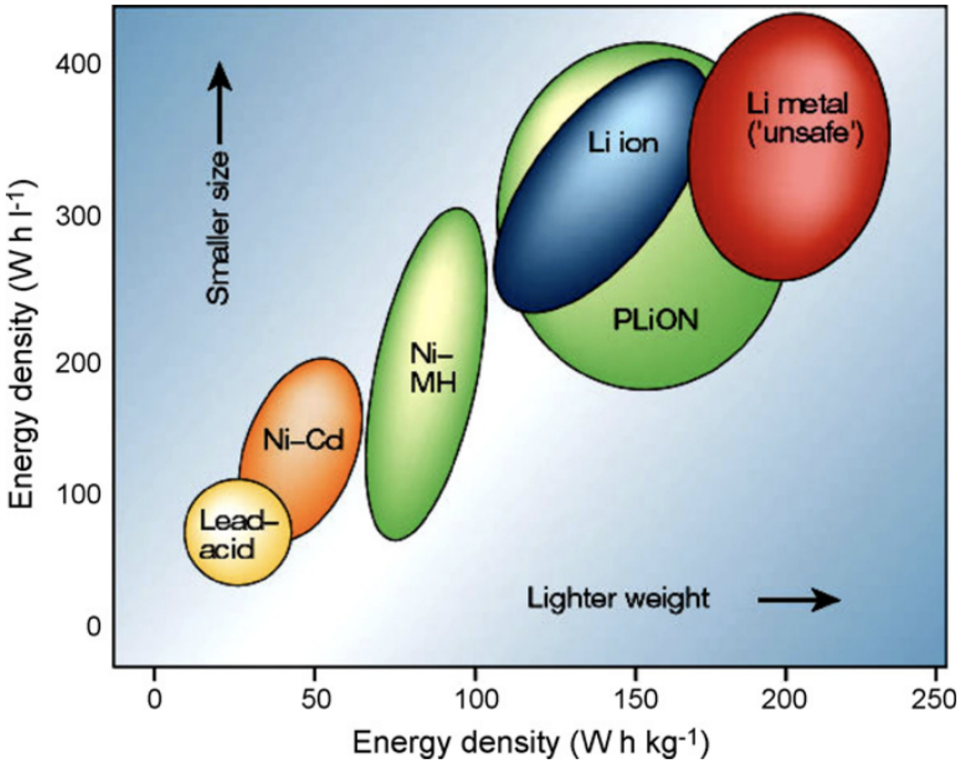


Figure 1.3: The comparisons of the gravimetric energy density (x-axis) and volumetric energy density (y-axis) of various rechargeable battery technologies. Reproduced with permission from [4].

1.2 Review of Lithium-ion Batteries (LIB)

In this section, a theoretical background of the LIB is given including its operating principles and key performance indicators. In addition, detailed reviews on the material selection of each internal component are examined to give a full picture of the current research directions for LIB.

1.2.1 Configuration and Operating Principles

Lithium-ion batteries, as the name suggests, use the oxidation and reduction of lithium ions to store and release chemical energy. A typical commercial LIB is made from four key components; a lithium oxide-based cathode, a graphite anode that can accommodate lithium ions, an ionically conductive non-aqueous electrolyte solution and a polymer membrane separator. A graphical illustration of this structure is given in **Figure 1.4** where the lithium cobalt oxide ($LiCoO_2$) is used as cathode and graphite (C_n) is used as the anode.

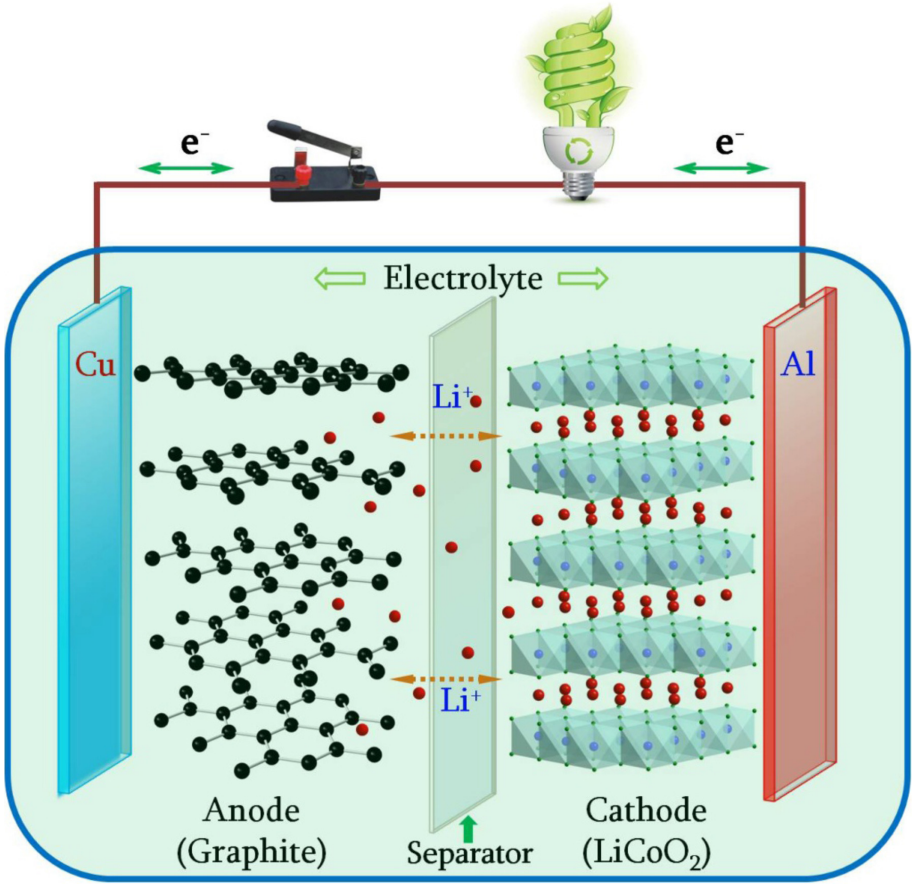


Figure 1.4: A schematic illustration of the Li-ion battery inner structure with graphite used as the anode materials, LiCoO2 used as the cathode materials. Reproduced with permission from [5].

The underlying operating principle of LIB is driven by the motion of the Li-ions between the oppositely charged electrodes through the electrolyte, as summarised in the equations:

$$\text{Cathode} : LiCoO_2 \leftrightarrow Li_{1-x}CoO_2 + xe^- + xLi^+ \tag{1.1}$$

$$\text{Anode} : xe^- + xLi^+ + C_n \leftrightarrow Li_xC_n \tag{1.2}$$

During charge, an electric potential is applied to extract lithium ions from the $LiCoO_2$ cathode host, and this oxidation reaction can be represented from left to right in **Equation 1.1**. The Li-ions then travel through the separator and electrolyte to intercalate into the empty layers of the anode material, initiating a reduction reaction shown from left to right in **Equation 1.2**.

During discharge, the electrochemical reactions are reversed where Li-ion are transferred from anode electrode to cathode, indicated from the right to left reactions in **Equation 1.1** and **Equation 1.2**. The oxidation reaction at the anode induces a free flow of current which are transferred in an external current collector (made from Copper and Aluminium as shown in **Figure 1.4**) and can be used to power other systems before entering back into the cathode structure. It is important to note that the magnitude of this external current flow is in balance with the internal ionic current flow to maintain systematic charge neutrality.

1.2.2 Underlying Electrochemical Theory

The extent of the oxidation or reduction reactions indicated in **Equation 1.1** and **Equation 1.2** are largely controlled by the change in the Gibbs energy and the voltage (electromotive force). The change in Gibbs free energy determines the spontaneity of the reactions with the sign of the value indicating the likelihood for the reaction to proceed. The more negative value indicating less energy required for initiating the reactions and thereby are more likely to occur. Similarly, the potential difference between the anode and cathode can indicate the tendency for the reaction to occur within the cell. These two parameters are essentially measuring the likelihood for the cell reaction to occur and can be quantitatively related using the Nernst equation given as **Equation 1.3** below.

$$\Delta G = -nE_{cell}F \tag{1.3}$$

Where ΔG is the change in Gibbs free energy, n is the number of electrons involved in the reaction, F is the Faraday constant (96785 $Cmol^{-1}$) and E_{cell} is the cell potential.

The working voltage or open circuit voltage is an intrinsic property that varies for the use of anode, cathode, and electrolyte material in the LIB system. To calculate it, one can find the difference between the electrochemical potentials of anode and cathode as shown in **Equation 1.4**.

$$V_{oc} = \frac{\mu_A - \mu_C}{e} \tag{1.4}$$

Where V_{oc} is the open-circuit voltage, μ_A and μ_C are the electrochemical potentials of anode and cathode, respectively and e is the magnitude of the elementary electronic charge.

1.2.3 Key Performance Metrics

Parameters such as the specific energy, capacity, life cycle and C-rate are often used during the examination of the performance level of the whole Li-ion battery and for the constituent electrodes.

The capacity of a battery ($C_{battery}$) is defined as the number of hours that a new battery can deliver under 1A of current. This is an extensive property that is varied and controlled by the concentration of the lithium ions and electrons involved in the electrochemical reactions. Furthermore, this property can be further converted into energy by multiplying with the average voltage ($V_{average}$) during discharge, following Faraday's law as indicated in **Equation 1.5**.

$$E\ (Wh) = V_{average}(V) \times C_{battery}(Ah) \tag{1.5}$$

From the calculated energy, one can estimate the volumetric and gravimetric energy densities of the cell, simply dividing it by the mass and the volume of the battery and the units for these parameters are given as $mAhL^{-1}$ and $mAhg^{-1}$, respectively.

In terms of the electrode materials, their theoretical capacities are often calculated using **Equation 1.6**.

$$C_{electrode} = \frac{nF}{3.6M} \tag{1.6}$$

Where $C_{electrode}$ is the specific capacity of the electrode material with the common unit of $mAhg^{-1}$, n is the number of available electrons or the Li-ions for withdrawn or extraction per material unit, F is the Faraday constant

($\approx 96485.33 \mathrm{Cmol}^{-1}$), M is the molar mass of the materials.

The C-rate is the principal measure of the current density rate used for charge and discharge. This rate can be quantitively related to the set time for charge and discharge as well as the properties of the electrodes as shown in **Equation 1.7**.

$$\mathrm{C-rate\ (Current\ density)} = \frac{\mathrm{C_{th}} \times \mathrm{m}}{\mathrm{T}} \tag{1.7}$$

Where m is the mass (in grams) of the active material in the electrode, T is the time required for charge or discharge in hours. 1C represents that a battery with 5Ah can be fully discharged/charged within 1 hour using the 5A discharge/charge rate. Testing current density is normally chosen to match with the application requirement, such as the average discharge rate of an electric car is normally around 3 hours which stands for the C/3 discharge rate.

Another key metrics is the life cycle, which is the number of charge/discharge cycles that a battery can support before the capacity drops below 80% of the original capacity. This is a direct parameter for estimating how many times the battery can be recharged and whether the performance will be maintained at a similar level after long and repetitive usage. Similar metrics are capacity retention rate and fading rate per cycle which gives a detailed analysis of the degradation of battery performance.

1.2.4 Battery Materials Challenges

A common issue of solid electrolyte interphase (SEI) is reported on the surface of various electrodes during the first cycle, and this is due to the reactions between the electrolyte and the electrodes. **Figure 1.5** illustrates the relative electronic energy of the cathode, anode, and liquid electrolyte in the LIB system under the thermodynamic equilibrium state.

The energy gap ($\mathrm{E_g}$) of the electrolyte is illustrated as the difference between the levels of the highest occupied molecular level (HOMO) and the lowest unoccupied molecular orbital (LUMO). Besides, the likelihood of SEI formation is largely dependent on the electrochemical potentials of the anode and cathode material. The two electrode materials, cathode, and anode have their electrochemical potentials

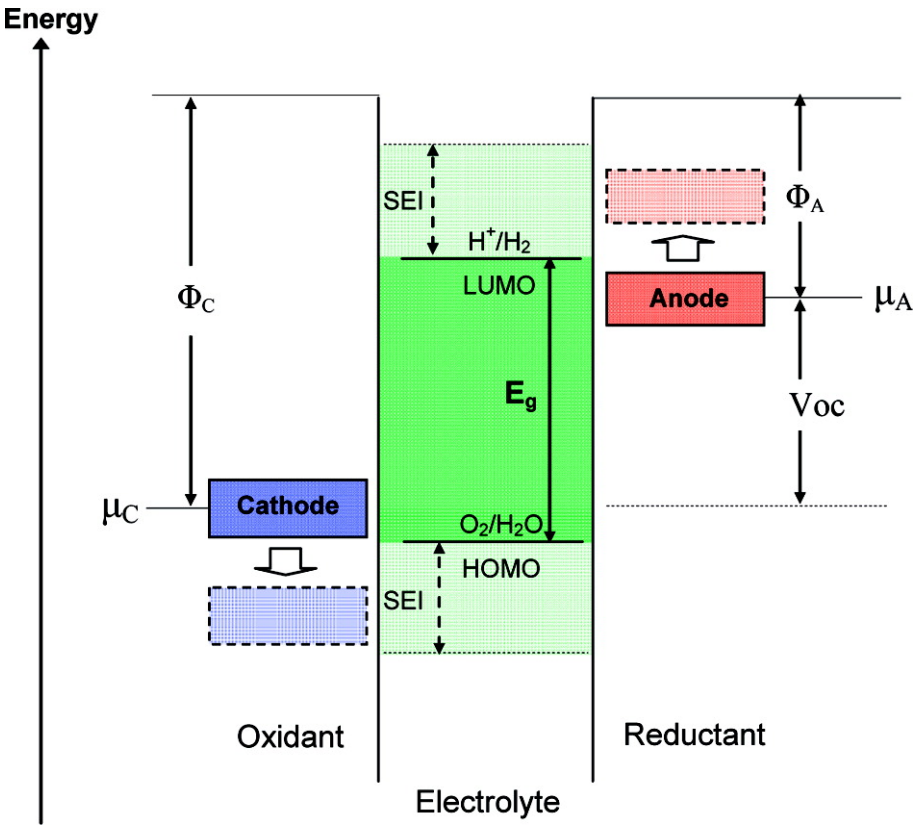


Figure 1.5: A semantic illustration of the open circuit diagram for relative electron energy of anode, cathode, the aqueous electrolyte in the LIB system, Φ_C and Φ_A denotes for the work function of cathode and anode, SEI stands for the solid-electrolyte interphase,μ_C and μ_A are the electrochemical potentials of cathode and anode, E_g is the energy gap. Reproduced with permission from [6].

labelled as μ_C and μ_C, respectively in **Figure 1.5**. Under the absence of SEI to create an electron transfer barrier between the electrode and electrolyte interphase, the aqueous electrolyte will be either oxidized if the selected cathode had a lower μ_C value than the HOMO energy level or reduced if the selected anode has a higher μ_A than the LUMO energy level of the electrolyte. The formation of SEI on the electrode surface will increase the internal resistance of the LIB and additionally lead to irreversible capacity loss. On the other hand, some studies have also shown that SEI can provide extra kinetic stability to the system if the chosen electrode materials create a large V_{oc} whilst providing that it doesn't have a large energy difference with the E_g of the electrolyte [49]. Various results have shown that the damage in SEI would further limit cell life by promoting the lithium dendrite growth on the anode and causing severe short-circuit problems [50].

To create a thermodynamically stable LIB system, the electrode materials should be selected that both μ_A and μ_B should be within the E_g of the electrolyte and conventionally and therefore the final operating voltage V_{oc} should be smaller than the E_g.

As such, the selection of the LIB materials should be strictly considered to fulfil the requirements that: 1) the electrolyte should have a large E_g for breaking the limits of electrochemical voltage window 2) the anode and cathode should be selected to suit this electrochemical window whilst minimising the likelihood of SEI formation after the first circuit.

1.2.5 Research of Component Materials and the Challenges

The development of new LIB with higher electrochemical performance is inseparable from the advancement in each of its components. In this section, the material selection and current research for each component are reviewed.

1.2.5.1 Electrolyte

The Electrolyte serves as the electronically insulating ion diffusion pathway between the two electrodes during charge/ discharge. An ideal electrolyte should have high Li-ion conductivity, low electrical conductivity, low cost, eco-friendliness, good chemical stability and compatibility with the other electrode materials. The current research of electrolytes is mainly categorised into organic liquid, gels, and solid polymer.

The majority of LIB use organic liquid mixtures of lithium- salt ($LiPF_6$, $LiBF_4$, $LiClO_4$), organic solvent (Ethylene carbonate (EC), Dimethyl carbonate (DMC), Ethyl methyl carbonate (EMC)) and carbonate-based additives as the electrolyte because of their wide electrochemical windows, the low activation energy for Li-ion diffusion and high ionic conductivity (10^{-3} $\sim 10^{-2}$ S cm^{-1}). It is also important to note that the inclusion of carbonate additives in the electrolyte can help to build a passivating SEI protection layer on the carbon-based anode and hence stabilize the electrode/electrolyte interfaces. Nevertheless, the highly flammable carbonate-based solvent and the high reactivity of the Li-salt decomposition prod-

ucts (PF_6^-) towards the small water content in the electrolyte do raise potential safety hazards. As an alternative, solid polymer electrolytes have shown promising features in being non-flammable and can reduce the dendrite growth on the electrode surface [51]. Several types of polymer materials have been researched and they have demonstrated their promising mechanical strengths and electrochemical performance. These include the polyethylene oxide (PEO) [52, 53, 54, 55], polymethyl methacrylate (PMMA) [56, 57], polyvinyl chloride (PVC) [58] and polyvinylidene fluoride (PVDF) [59, 60]. Regardless, the real application of solid polymer electrolytes in LIB is still constrained due to the poor ionic conductivity ($< 10^{-5}$ Scm^{-1}) of these materials. This is the chief challenge in developing new materials for this application.

Next, gel-typed polymer electrolyte is seen as an alternative option between the electrolyte system of liquid-free and full liquid. It has become increasingly popular as it combines the advantages of both the liquid-free system and the full-liquid electrolyte system; good ionic conductivity, great mechanical strength and the separator-free feature (replaced by host solid polymer matrix) [61]. The mainstream research of gel-based electrolytes is divided into the physical gel system and the chemical gel system. The physical gel system is essentially a hybrid system where the liquid electrolyte is contained in the polymer matrix on the conditions that they do not interact chemically with each other. In addition, the chemical gel system uses a cross-linker chemical that forms a bond with the polymer matrix. Since there are different components in the gel-based electrolyte, research has been done separately into the host polymer matrix materials as previously mentioned for the solid-state electrolyte materials and the cross-linker regents. The widely researched cross-linking reagents are the inorganic compounds such as the TiO_2 [62], SiO_2 [63] and the polymer-based compounds such as the polyethyleneimine (PEI) [64] and polyethylene glycol-dimethacrylate (EGDMA) [65]. Although the ionic conductivity of both gel-based systems has shown to be greater than the liquid-free solid electrolyte, they both still suffer from the problems of high instability of the interfaces with both the lithium metal anode and with the cathode and the lack of

thermal stability, which remain as challenges to be solved.

1.2.5.2 Separator

The role of a separator in a LIB system is to eliminate the possibility of physical contact between the two electrodes to prevent short-circuiting. Although the separator does not take parts in the electrochemical reactions, its structural properties have been shown to have a considerable impact on the cyclic performance of LIB [66]. A suitable separator should be chemically compatible (stable) with the electrodes and the electrolyte and have high mechanical strength and toughness. Moreover, the separator should have uniformly distributed porosity (commonly $> 40\%$) to allow for Li-ion diffusion, with the pore sizes smaller than the electrode particle sizes and good wettability for the liquid electrolyte. Another important feature of a separator is also its emergency shutdown capability at elevated temperatures (> 130 ℃) to prevent thermal runaway and ultimately combustion [67].

The most used separators are the microporous membrane materials consisting of various polymeric materials such as polyethylene, polypropylene or polyvinylidene fluoride (PVDF). New research has been done into modifying the structures of these membranes with plasma or electron-beam radiations treatments to enhance the wettability and the retention rate of electrolytes [68, 69]. The current state-of-art separators are fibre-based non-woven mat separators, [70, 71] and composite membrane materials [72]. These materials have demonstrated great improvements in ionic conductivity as well as thermal stability relative to standard microporous membranes.

1.2.5.3 Anode (Negative Electrode)

An ideal anode should fulfil the requirements of having high specific capacities, lower potential than the cathode, low cost, robust crystal structure during charge and recharge and should be eco-friendly for sustainable usage. In theory, lithium metal is the most ideal anode material as it has the highest possible high theoretical specific capacity (3860 mAhg^{-1}) as it has no other elements that do not take part in the reduction/oxidation reaction. However, the growth of unwanted lithium dendrites is observed on the anode surface, leading to the cell short-circuiting, and

possibly combusting [73]. As such, although it cannot be used itself, Li-metal is used as the performance benchmark for candidate anode materials. Most modern anode research can be classified as either carbon-based materials or non-carbon-based materials.

Carbon-based anodes such as graphite are widely implemented in the LIB industry due to their excellent electric conductivity, low cost, low density and great safety. The widely used carbon compound LiC_6 has a high experimental specific capacity of 372 $mAhg^{-1}$ whilst research has been conducted for other stoichiometric forms such as Li_2C_6 [74], Li_3C_6 [75] with high specific capacities of 780 $mAhg^{-1}$ and 1116 $mAhg^{-1}$ respectively. Furthermore, *Yoo et al* [76] improved the graphene specific capacity from 540 $mAhg^{-1}$ to 700 $mAhg^{-1}$ by enlarging their interlayer spacing using various nanotubes and fullerenes (C_{60}). Nevertheless, graphene materials have high surface energy which leads to immediate triggering side-reactions with aqueous electrolyte before the initial formation of SEI as protection. This would lead to irreversible capacity loss encountered during the initial charge/discharge process and hence reduce its cyclability.

Research into non-carbon-based anodes is mainly focused on metal oxide and alloy materials. Among all the metal oxides, titanium-based oxides ($Li_4Ti_5O_{12}$ and TiO_2) are leading materials as they have a lower operating voltage (0.8V) than graphene and have comparatively less SEI formation on the anode surface. However, the applicability of this material is limited due to the high cost for processing, poor electronic conductivity, and the additional safety concerns of gas release from their reactions with the electrolyte.

Recent research has focused on developing graphene supported metal oxide materials to combine the high surface area and electrical conductivity of graphene with the low SEI formation and operating voltage of the oxide materials [77]. Furthermore, alloy materials such as Li-Si, Li-Ge and Li-Sn are of interest for their high specific capacities (multiple times more than graphene) and their reasonably low operating voltage. However, the challenges of capacity loss and the safety issues such as the dendrite growth remain since the alloy material undergoes volume

expansion during the charge and discharge process [78].

1.2.5.4 Cathode (Positive Electrode)

The currently researched intercalation cathode materials are all transition metal oxide (M-O) based and can be classified according to their underlying structures, namely, layered materials ($LiMO_2$), spinel materials (LiM_2O_4) and olivine materials ($LiMPO_4$) as shown in **Figure 2.2**. The cathode acts as a Li-ions host in LIB and thereby should meet several requirements to best facilitate the insertion and extraction reactions. Firstly, an ideal cathode should have a suitable structure for fast Li-ion diffusion during the Li-ion intercalation/deintercalation process whilst being able to store as many Li-ions as possible (high specific capacities). In addition, the cathode crystal structure should be stable during the charge/discharge to avoid any irreversible capacity loss. As opposed to the anode material, the cathode should have a high operating voltage to reduce the likelihood of SEI formation after the initial cycle. Other desirable features of the ideal cathode materials are low cost, easy processing, and low toxicity.

In addition to the metal oxide, most cathodes have additives of polymer binders (polyvinylidene fluoride (PVDF), ethylene propylene diene monomer (EPDM)) for binding the cathode mixture to improve the mechanical stability of the material and the carbon-based conductive additives to improve the electrical conductivity. The mixture ratio of these components varies, but a typical ratio of 8:1:1 (active: additives: binder) is used. Small weight percentages of the binders and conductive additives are added to the electrode preparation method as these components are electrochemical inactive. In addition, the mixing of carbon conductive and binder would lead to the formation of a new mixed-phase called carbon binder domain (CBD). The latest research has demonstrated that the morphology of CBD has a great impact on the kinetics of the Li-ion transport from the pore space as it can cause the potential blockage of the electrochemically active sites of the cathodes [79, 80, 81].

Each class of active cathode materials has limited electrochemical performance due to the limitations of its structural properties. For instance, the layered materi-

als $LiCoO_2$ suffers from low practical capacities because of the weak structure, the spinel cathode $LiMn_2O_4$ faces severe capacity fading issues due to the undesired Mn-ion dissolution reaction and the $LiFePO_4$ olivine materials have extremely low electronic conductivity and slow Li diffusion. As such, research has been done to enhance their electrochemical potentials by changing their physical or chemical properties with varying degrees of success. These optimisation methods include the synthesis of the composite cathode system that integrates multiple active components, [82, 83, 84] doping with alien atoms [85, 86, 87], changing the crystallite size and the particle size, [88, 89, 90] and finally, applying surface coatings [91, 92].

Among all components that have been discussed above, the cathode component is inevitably the most influential component for the development of new LIB with better performance; they have a large influence on the overall energy and power density of the LIBs; its crystal structural properties are crucial for determining the LIB charge/discharge life cycle due to the repetitive insertion/extraction of the Li-ions; it typically contributes to nearly 25% of the overall LIB cost. In the following sections, detailed reviews will be given on the chemistries, challenges and the ongoing research progress for three classes of current state-of-art cathode materials, namely, the layered structure, the spinel structure and the olivine structure.

1.3 Thesis Objectives and Outline

To facilitate the discovery of new cathode materials with better electrochemical performance, a deep understanding of the relationship between composition, structure and materials property is required. However, this is difficult to achieve through conventional means due to the high complexity of experimental variable correlations and the large time taken to perform performance tests, limiting the rate of experimental iteration. Discharge capacities of a battery is a vital performance metric to examine on the amount of electric power that a battery can deliver at any point of the cyclic performance under the discharging conditions. In addition, this reveals the battery health conditions such as the remaining useful life and help to distinguish the quality of the battery materials. Nevertheless, a full 25 cycle ca-

pacities test of a Li-ion battery containing a new cathode material would take up to 20 days if the battery was discharged at the rate of C/10 [93]. As a result, this significantly increases the time required to investigate the performance of the new cathode materials.

The benefits of doping on cathode materials' electrochemical performance and the overall battery discharge capacities have been discussed in **subsubsection 2.1.2.3**, however, the research community still have a lack of understanding of the detailed correlations due to the complex variable space. In this regard, the primary goal of this thesis centres on investigating the potential of using machine learning techniques to uncover variable correlations of the material systematic properties and the discharge capacities of three classes of doped LIB cathode materials.

The three doped material systems that are investigated in this thesis are referred to as the doped spinel system ($LiM_xMn_{2-x}O_4$, M is the dopant), the doped layered system ($LiNi_xMn_yCo_zM_{1-x-y-z}O_2$, M is the dopant), the carbon-coated and doped olivine system (C/LiM_1M_2PO4, where M1 and M2 are transition metals and C, is the carbon coating). The data used for training the ML models for each cathode material type is collated manually from accessible publications. Since the quality of the data determines the accuracy of the prediction model, data sources such as the DFT simulation data were avoided due to their large deviation from the experimental results. Instead, experimental publications were manually selected with rules implemented on the material type, experimental conditions and the reported experimental results for each material. The pursuits of each material system investigation are present in **Chapters 5, 6, 7**, respectively.

In Chapter 5, the relationship between the material properties and discharge performance of the doped lithium manganese oxide spinel cathode, $LiM_xMn_{2-x}O_4$ is investigated. A suitable dataset was collected, with a range of ML algorithms applied to predict the initial and 20^{th} cycle end discharge capacity of several cathode materials, and thereby uncover the features of the most predictive materials and analysed variable correlations with each discharge capacity property.

The relationships between the material properties and the discharge

performance of four types of the doped lithium nickel-cobalt-manganese based (NCM) cathode systems including, $LiNi_{0.33}Co_{0.33}Mn_{0.33}O_2$ (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) are investigated in **Chapter 6**. K-nearest neighbour, support vector machine, artificial neural network, random forest, gradient boosting machine, kernel ridge models are applied to predict the initial and the 50^{th} cycle discharge capacities of the doped NCM systems. The correlations of various dopant-related elemental features and the material systematic properties such as crystal lattice dimension, volume with discharge capacity are explored.

Finally, in **Chapter 7**, the relationships between the material properties and the discharge performance of single doped, carbon-coated olivine systems, including the iron-based system ($C/LiFe_{1-x}M_xPO_4$) and the manganese-based system ($C/LiMn_{1-x}M_xPO_4$) are investigated. In particular, several ML algorithms are applied to a manually collected data set to predict the initial and the 20^{th} cycle discharge capacity of the iron-based and manganese-based systems. The variable correlations of the selected input variables such as the carbon coating content, crystal lattice dimension with each of the response discharge capacity variables are explored, and a model with the best prediction power is applied to predict the initial and the 20^{th} cycle discharge capacity of the cobalt-based olivine system.

Please note that the material presented in **Chapter 5** and **Chapter 6** have been adapted from [94] and [95] respectively. This action has been issued the licenses by the publisher with the terms being agreed with the first author of these publications, which is also the author of this thesis (Guanyu Wang). For further details, please refer to the license agreements attached in the Appendix section, given as **Figure A.17** and **Figure A.18**.

Chapter 2

State of the Art of Cathode Materials Research

2.1 Cathode Materials

Research into economically feasible cathode materials with high storage capability has been carried out for many years due to the vital role that cathode plays in the overall energy density of LIBs. The cathode should have a strong crystal structure that remains volumetrically stable during Li insertion and removal and have high ionic conductivity for Li-ions. The major types of intercalation cathodes are classified by their available pathways for Li-ions diffusion, namely, one-dimensional (1D), two-dimensional (2D) and three-dimensional (3D), as shown in **Figure 2.1**.

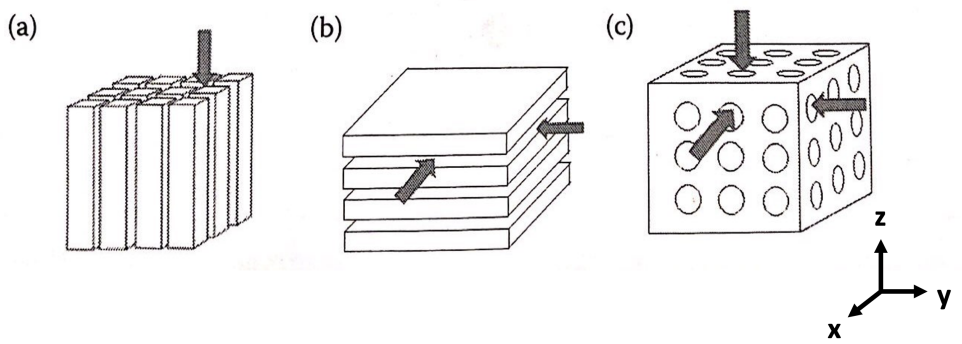


Figure 2.1: Schematic illustration of the possible Li-ion intercalation pathways for cathode materials in a) one dimension, b) two-dimension, c) three-dimension. Reproduced with permission from [7].

In general, the perpendicular top-downward direction of inserting and extracting Li-ions would leave the material structure undergoing extreme anisotropic expansion or shrinkage and experiencing structural exfoliation. On the other hand, the two-dimension intercalation structure (**Figure 2.1-b**) is more desirable as it offers alternative pathways (along the x and y-axis) and has higher structural stability in terms of maintaining the host matrix [7].

2.1.1 Material Selection and Chemistry

The three main cathode materials that have been researched widely across the globe in recent years are layered compounds, spinels and olivines. This section will review these three classes of materials in terms of their chemistry, modifications and suitability for use as cathodes. The crystal structures of these three classes of compounds are shown in **Figure 2.2**.

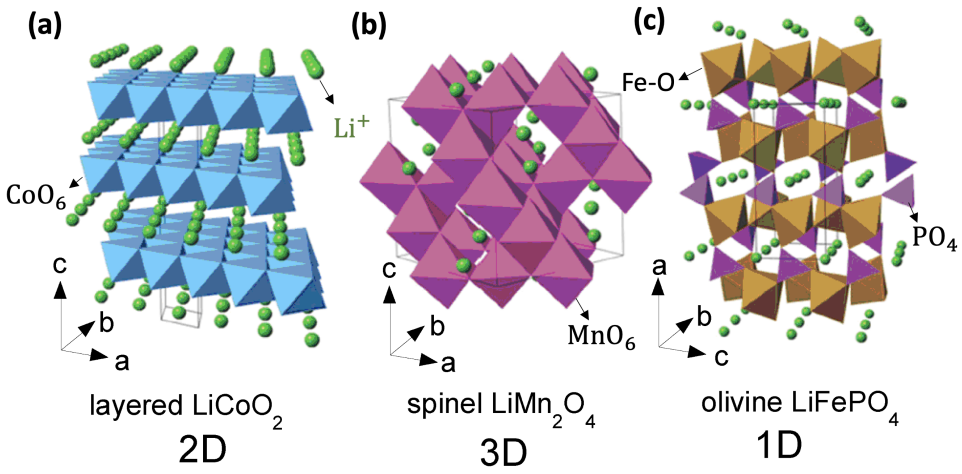


Figure 2.2: A schematic illustration of the crystal structures of the representative (a) layered material $LiCoO_2$, (b) spinel material $LiMn_2O_4$ and (c) olivine material $LiFePO_4$.Reproduced with permission from [8].

2.1.1.1 Layered Materials

Layered materials have a general formula of $LiMO_2$ (M is a transition metal) and play a significant role in the current LIB market. The electrochemical potentials of $LiCoO_2$ (LCO) layered cathode material was initially proposed by the Goodenough group [96] and investigated experimentally by the *Plichta et al* [97]. It was concluded that layered cathodes with a Li metal anode could provide a much higher

operating voltage of 4V than the leading commercialised battery at the time (Ni-based 1.2V, [98]).

LCO has a prototypical α-$NaFeO_2$ layered structure with a space group of $R\bar{3}m$ as shown in **Figure 2.2-a**. Layers of Li^+ are separated by sheets of CoO_2^- that are formed by the edge-sharing CoO_6 octahedra. Upon the Li-ion insertion or withdrawal, the oxidation state of the cobalt will switch in-between +3/+4. Furthermore, the withdrawal of Li-ion from the structure $Li_{1-x}CO_2$ is limited to be $x < 0.5$, which leads to the practical capacity being approximately half of the theoretical (from 274 $mAhg^{-1}$ to 140 $mAhg^{-1}$). Any further withdrawal of Li-ions leads to a phase change from monoclinic to hexagonal phase, which is accompanied by severe structural shrinkage and lattice disorder (Li-ions travel from 3a to the 3b Co lattice) which degrades its cyclic performance. In addition, the bonds between metal oxide layers are maintained by the weak van der Waal's forces which makes the whole structure unstable at high lithium loading of $x > 0.5$ of the formula unit. This leads to an irreversible phase transformation in the host structure and greatly deteriorates future cyclic performance [99, 100, 101]. Another drawback of LCO is the use of cobalt metal which is both expensive and toxic to the environment. As such, it has been proposed to substitute cobalt for other metal elements to find alternative layered materials that are cheaper, has greater structural stability, and have better cyclic performance. Manganese and nickel metals are popular substituting options as they are both earthly abundant and have good electrochemical potentials. This proposal has led to the discovery of new layered compounds including $LiNiO_2$(LNO) [102], $LiMnO_2$ [103] and $LiNi_{1-x-y}Co_xMn_yO_2$ (NCM) [104]. Among these, $LiNiO_2$ is popular for its high practical capacity (150 $mAhg^{-1}$) and non-toxicity, however, it suffers from poor Li-ion diffusion and poor cyclic performance due to the cation disorder of the similarly sized Ni^{2+} and Li^+ ions and is found difficult to synthesis [105]. $LiMnO_2$ is the cheapest of the candidate cathodes, however tends to undergo a phase change to the spinel structure during the charge/discharge resulting in rapid, irreversible capacity fading.

NCM compounds were developed to overcome the shortcomings of the Co,

Ni and Mn-based structures, combing the three to overcome the individual drawbacks whilst maintaining a good level of performance. Further, the transition metal ratios can be tuned according to the requirements of cost and electrochemical performance. During the cycling, Ni is the main redox component with its diversified oxidation states (II, III, IV) in the redox reaction whilst Co reduces the overall cation mixing effect and the Mn is used to stabilise the overall structure from its inertness to the reactions. Various NCM compounds with different compositions have been discovered to have high energy density. In particular, the nickel-rich NCM (nickel content > 0.8 per formula unit) is the current state-of-art material with its high electronic conductivity, similar operating voltage range (3.0-4.3V) to LCO and high capacity (203 $mAhg^{-1}$).

2.1.1.2 Spinel Materials

Spinel cathodes were introduced as a more eco-friendly and cheaper option than layered $LiCoO_2$ cathodes. The most representative spinel compound is $LiMn_2O_4$ with an Fd-3m space group as shown in **Figure 2.2-b**. Manganese atoms occupy octahedral sites at 16d whilst the lithium ions occupy the tetrahedral sites at 8a. This arrangement allows 3-D intercalation/deintercalation pathways for lithium ions with fast diffusion kinetics. The oxidation state of manganese can switch reversibly between +3 and +4 during the charge/discharge process. The crystal structure of $LiMn_2O_4$ is robust, enabling a full unit of Li-ion delithiation per formula unit, thus the practical capacity (120 $mAhg^{-1}$ at 0.1C) of $LiMn_2O_4$ is close to its theoretical capacity of 148$mAhg^{-1}$.

The major shortcoming of $LiMn_2O_4$ is the rapid capacity fading observed upon numerous cycling at 4V against the Li metal anode [96]. For this, two of the major root causes are thought to be the manganese dissolution into the electrolyte and the Jahn-Teller distortion effects from d-electrons by the redox reactions. **Equation 2.1** illustrates the disproportional reactions of Mn^{3+} on the surface of the $LiMn_2O_4$ cathode which initiate the dissolution of Mn^{2+} into the electrolyte and causing significant loss of the active material [106].

$$2Mn^{3+} \rightarrow Mn^{2+} + Mn^{4+} \tag{2.1}$$

The Jahn-teller distortion effects are mainly caused by the unique electronic configuration (t_{2g}^3, e_g^1) of Mn^{3+} in octahedral complexes with the oxygen as shown in **Figure 2.3**. As a result of this effect, a structural transition (cubic to tetragonal) is observed as the Mn-O bonds are elongated and this creates a new cathode phase which reduces the Li-ion diffusivity and leads to severe capacity decay [107]. Furthermore, the capacity fading issue is more pronounced when the Li-ion diffusion rate is greater in the electrolyte than in the electrode (caused by high discharge rate) which then leads to the formation of an Mn^{3+}-rich region on the $LiMn_2O_4$ surface [108].

In response to these issues, doping treatment and surface coatings have shown to be effective in suppressing the Jahn-teller distortion effects. The rationale behind the doping methods (both bulk and surface doping) is to increase the average valence number of Mn and therefore inhibit the Mn^{+3} Jahn-teller distortion effect. Dopants like aluminium (Al^{3+}) [109], erbium (Er^{3+}) [110], cerium (Cr^{3+}) [111], scandium (Sc^{3+}) [112] have been investigated in the literature with the great enhancement to the observed cyclic performance. In addition, a bulk doping approach with the electrochemically active Ni element has led to the discovery of $LiNi_{0.5}Mn_{1.5}O_2$, which has a high operating voltage (4.7V against Li/Li^+) and a high energy density (650 $Whkg^{-1}$) [113, 114, 115].

Cathode coatings can effectively reduce the contact between the organic electrolyte and the $LiMn_2O_4$ and hence suppress the Mn dissolution. Coating materials such as Al_2O_3 [116, 117], CeO_2 [118], $FePO_4$ [119], AlF_3 [120] and LaF_3 [121, 122] have been widely reported to reduce the capacity fading issues. Furthermore, the effect of morphology and crystallinity and particle size on the electrochemical performance has also been investigated and shown to influence the overall cycle performance of the spinel materials [123, 124, 125].

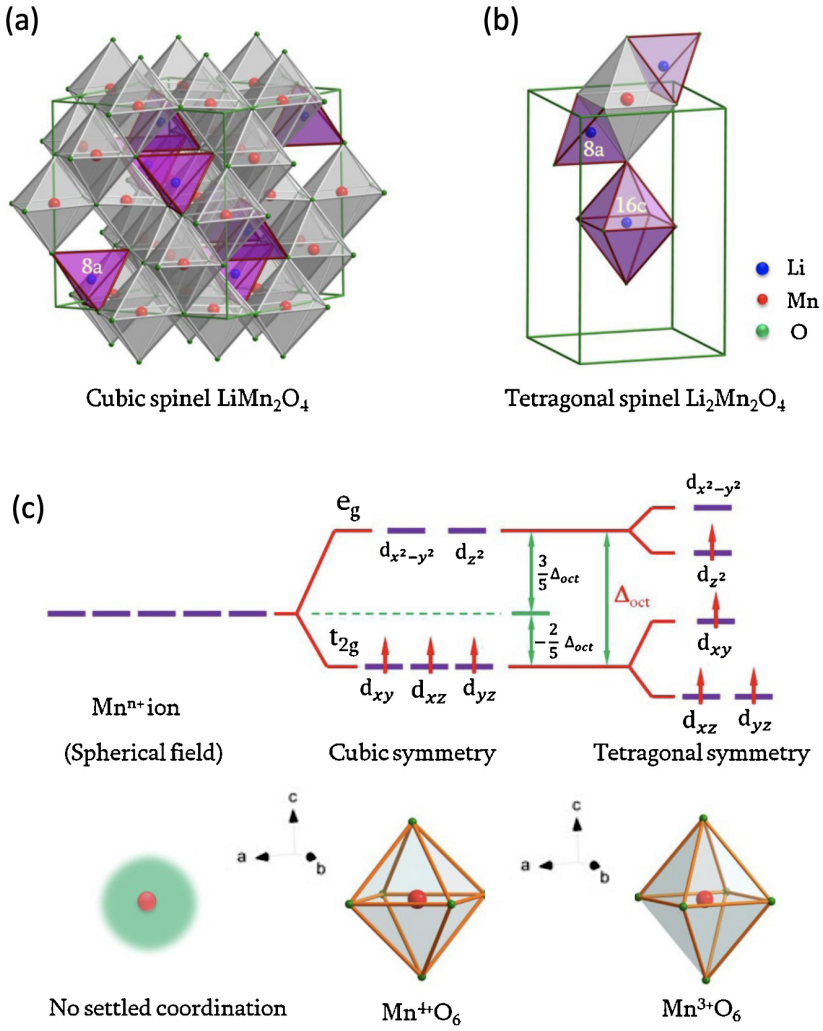


Figure 2.3: The crystal structures of (a) the cubic spinel $LiMn_2O_4$ and (b) the tetragonal spinel $LiMn_2O_4$ and the change in the 3d orbitals diagrams of the Mn ions as the structure elongation occurs along the c-axis and the structure transforms from the cubic symmetry to the tetragonal symmetry. Reproduced with permission from [5].

2.1.1.3 Olivine Materials

Olivine materials were first proposed by *Radhi et al* [126] for their advantages of non-toxicity, and low manufacturing cost. The olivine material itself consists of transition metal elements accompanying a phosphate group and often can be represented by the general formula $LiMPO_4$ (where M is a transition metal). The commonly used transition metals are iron [127], manganese [128] and cobalt [129], with the respective theoretical capacities of 170 $mAhg^{-1}$, 171 $mAhg^{-1}$ and 167 $mAhg^{-1}$. All three compounds belong to the *Pnma* space group with the hexagonal close-packed structure (see **Figure 2.2-c**). The structure consists of FeO_6 octahedra groups sharing oxygen atoms with the PO_4 tetrahedra group whilst the Li sit in interstitial voids in the structure. One benefit of this structure is that the strong covalent bonding between phosphate polyhedra stabilises the structure allowing it to perform under extreme operating conditions ($> 200℃$) [130]. On the other hand, this leads to Li^+ diffusivity and electronic conductivity ($10^{-8} \sim 10^{-10}$ Scm^{-1}). In addition, this material is susceptible to the growth of a second phase, $FePO_4$, on the surface of the electrode which inhibits Li-ion diffusion and result in low rate capability [131].

Other olivine-based materials like $LiCoPO_4$ and $LiMnPO_4$ have their advantages and disadvantages. $LiCoPO_4$ has a higher operating potential (4.8V versus Li^+/Li) than $LiFePO_4$ (3.4V versus Li^+/Li), however, is also more expensive and toxic due to the use of cobalt. In comparison, $LiMnPO_4$ has a similar operating potential of 4.1 V, and its cost is much lower, however, it has low ionic and electronic conductivity and therefore is not suitable for high charge/discharge rate applications such as EVs.

Having explored currently researched potential olivine cathode materials, it can be surmised that three approaches are being actively pursued to improve their electrochemical performance. Coating the $LiFePO_4$ with some highly conductive materials such as carbon, copper or silver has been reported to improve the overall electronic conductivity and thus reaction kinetics [132, 133, 134]. In addition, the rate performance and the reversible charge/discharge capacity can be improved

through the modification of the particle size of the olivine materials and the partial doping methods [135, 136, 137, 138].

2.1.2 Elemental and Structural Design Parameters

To improve on the materials outlined in **subsection 2.1.1**, the relationship of the properties of potential substitution/dopant elements, as well as crystal structure and microstructure with the key performance metrics of the cathode including energy density and cost should be well understood [139]. This section examines these relationships in more detail.

2.1.2.1 Elemental Selection

Potential substitutions in the cathode formula should be considered carefully as they directly influence the materials' physical and chemical properties.

Transition metal elements are typically desirable as dopants/substitutes on anion sites as they can have multiple valence states, which allows for the intercalation/deintercalation of the Li-ions as they are reduced/oxidised. In transition metal-based cathode compounds, the transition metal is treated as the complex centre and normally bonded to surrounding ligand groups (oxygen, polyanion). To ensure that the structure is robust and stable during the cycle, the strength of metal-ligand bonds should be considered, and this relates to the electronegativity difference of the atoms at each end of the bond. A larger electronegativity difference indicates that the bond is more ionic and is usually shorter and stronger than a covalent bond, thereby increasing structural stability.

The polyanionic ligands carry different electronegativities due to the diverse internal electronic interactions and can form more ionic bonds between the metal centre and oxygen than the single-element ligands (oxygen ion). Due to the more apparent ionic character of the metal and oxygen bond, the separation between their bonding orbital and the anti-bonding orbital can be reduced. Moreover, the energy of the lowest unoccupied 3d orbitals of the metal centre will also be lowered as illustrated in **Figure 2.4**, leading to a downward shifting of the cathode potential indicated in **Figure 1.5** and thereby increasing the cell potential. For example, the

polyanionic-base compound $LiNiPO_4$ has a higher operating voltage (5.1V versus Li^+/Li) [140] than $LiNiO_2$ (4.4V versus Li^+/Li) [141].

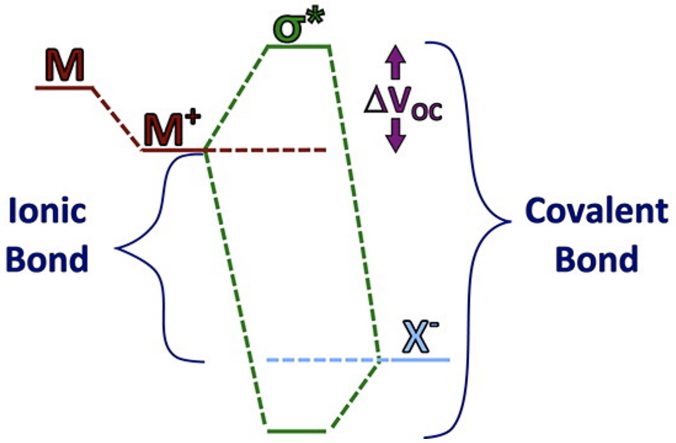


Figure 2.4: A schematic illustration of the difference in the separation between the bonding orbital and anti-bonding orbital for a more ionic M-X bond and a more covalent M-X bond and ΔV_{oc} denotes the change in the open-circuit voltage as the result of this bonding character modification. Reproduced with permission from [9].

A high capacity can be theoretically achieved by selecting an electrode material that has both low molar mass and a greater number of available electrons per unit Li, as suggested in **Equation 1.6**. Taking the example of $LiCoO_2$, it has an overall molar mass of 97.87 $gmol^{-1}$, which is nearly half of the molar mass of $LiMn_2O_4$ (M = 180 $gmol^{-1}$), and therefore its theoretical capacity is much higher at 274 $mAhg^{-1}$ than the later 148 $mAhg^{-1}$. In practice, however, $LiCoO_2$ has high internal resistance and only approximately half of the Li can be removed before an irreversible phase change is induced, which leads to a much lower practical capacity (180 $mAhg^{-1}$ versus Li^+/Li), unlike $LiMn_2O_4$ which can undergo full delithiation and has experimental capacity close to the theoretical value (120 $mAhg^{-1}$ [142]).

Finally, the cost and toxicity of the material should be taken into consideration. Metals such as Os, Hg, TI and Cd are highly toxic as shown in **Figure 2.5** and therefore should not be considered. In addition, some metallic elements in the 4d and 5d periods such as Au, Ru, Rh are expensive and all have high molecular weight. On the other hand, some studies have shown improvements in cyclic performance by doping trace amounts of these heavy and expensive metals into the original ma-

terial [143], as discussed in **subsubsection 2.1.1.2** for treating the spinel cathode's capacity fading issue.

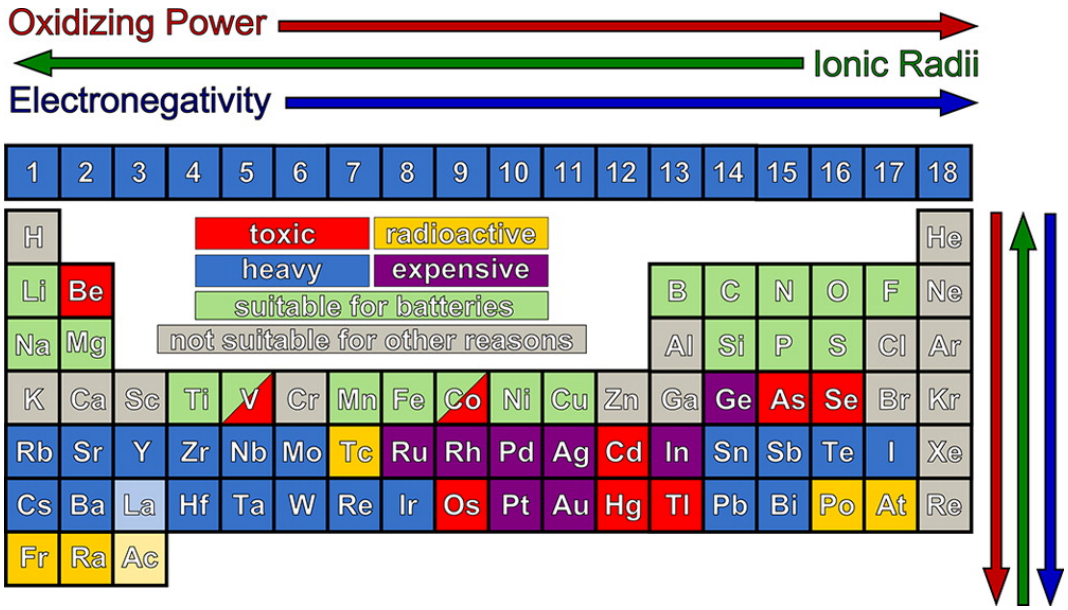


Figure 2.5: A Summary of the potential elements for the design of new cathode materials in the periodic table format. Image reproduced with permission from [10]

2.1.2.2 Crystal Structure Effects

The crystal structure of the cathode can have significant impacts on material performance. One important feature to consider is the Li lattice site energy as it effectively determines how easy it is to extract Li-ions from the host structure. This is defined as $\frac{dH}{dn}$, where H is the enthalpy change of the reaction and n is the number of electrons/Li-ions transferred. The higher the site energy, the less energy is required to remove one Li-ion from the structure. If there are multiple symmetrically distinct lattice sites, each site may have different site energy.

Both the spinel and the olivine materials are formed of the strong and densely packed anion-O polyhedron, which stabilise the crystal structure and provide the 3-D diffusion (**Figure 2.1**) channel for Li-ions. As a result of this, the site energy for Li-ion is maintained during the intercalation/deintercalation process and hence providing a more durable performance than other layered materials. Layered materials, on the other hand, are formed of loosely packed layers bound with weaker bonds (**Figure 2.1**) which is vulnerable to structural distortion during in-

tercalation/deintercalation. **Figure 2.6** shows the crystal structures of three classes of the cathode and their respective electrochemical performance. Firstly, it is seen that $LiNi_{0.5}Mn_{1.5}O_4$ exhibits a much higher operating potential at 4.7 V than the $LiFePO_4$ Olivine (3.4 V) and they both have flat potential curves due to phrase transition reactions [144, 145]. $LiCoO_2$ has a sloping potential curve as its structure being distorted from the rising site energy in the host structure. The evolved layered materials such as the standard NCM and Li-rich NCM have shown a higher operating potential and greater capacities than $LiCoO_2$ due to the mixing of electrochemically active Nickel ions and the structure stabilizer Mn^{4+}.

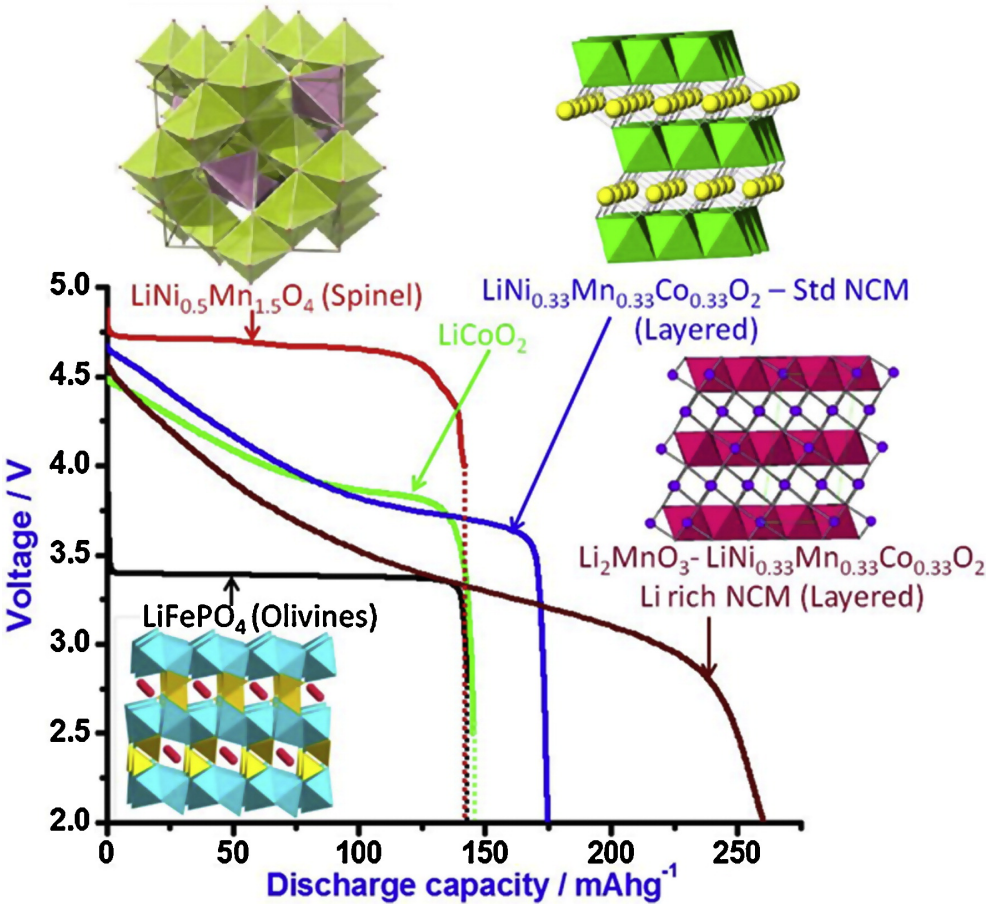


Figure 2.6: The comparisons of the three classes of intercalation classes with their crystal structures, chemical formula and the voltage profiles (versus Li^+/ Li). Reproduced with permission from [11].

2.1.2.3 Doping

Doping is an effective method to improve the overall electrochemical performance of cathode materials. This is achieved by introducing a low concentration of an additional element into the structure, usually as a substitution rather than interstitial. If the substitutional element has a different charge state to the host site, this can create additional vacancies for the anions or cations. In addition, the chemical pressure, crystal field and distortion field from the original structure will be modified to change the Li-ion site energies and influence the rate of insertion/extraction reactions. Doping can also improve structural stability and enhance both the electrical and ionic conductivity of the material. To reduce capacity fading, *Duan et al* [146] doped the spinel cathode with 3% nickel which led to an improvement in the 2000^{th} cycle capacity retention rate from 31% to 56%. In addition, significant improvements in the long-cycle capacity for spinel $LiMn_2O_4$ have also reported from the application of other dopants such as Al^{3+} [147, 148], B^{-} [149], Cr^{2+} [150, 151], Ce^{+3} [152], Sc^{3+} [112], Mg^{2+} [153, 154] . Furthermore, *Bakierska et al* [155] observed two times higher Li-ion diffusivity in 1% K & S doped $LiMn_2O_4$ relative to the undoped material. These results demonstrate the utility of doping to improve the overall electrochemical performance of these materials, however, there are still many potential dopants that remain unexplored.

A key challenge is to select a suitable dopant and concentration to maximise the material's electrochemical performance. *Xiong et al* [156] observed the same level of discharge capacity at the 70^{th} cycle for $LiMn_{1.9}Ti_{0.1}O_4$ and $LiMn_{1.6}Ti_{0.4}O_4$ despite the latter have given much lower discharge capacities in the initial 50 cycles. Similarly, the amount of doping with electrochemically inactive atoms onto the crystal lattice sites should be carefully selected as it would reduce the theoretical capacities of the original cathode. In general, there is a general lack of understanding in the correlation of the dopant selection and how the electrochemical performance would consequently change. The doping strategy to reduce capacity fading for spinel cathodes is based on selecting lower valence atoms (<3) to lower the overall Mn valence state in the formula and suppress the overall Jahn-Teller Dis-

tortion effects. On the other hand, several studies have demonstrated that using a cation dopant with higher valence (>3) could also improve the electrochemical performance which contradicts the theoretical assumption [157]. This highlights the complex role that doping plays in the cathode system and should be carefully studied using advanced computational tools. Furthermore, various studies have been conducted to investigate the effectiveness of dual-doping or multiple-cation doping on enhancing cyclic capacities[158, 159, 160, 161].

While some success has been achieved with this method, in order to reach the maximum benefits one would need to gain a more thorough understanding of the dopant distribution in the host lattice structure, interactions with the original atoms and the respective synthesis process to achieve the targeted composition. Further, the number of potential dopants, dopant combinations and concentrations represent a vast, many-dimensional compositional space, which is impossible to fully explore using expensive, time-consuming manual experimentation. Data-driven machine learning techniques have the potential to direct the experimental search into promising directions. The discussion of this cutting-edge research method will be provided in **subsection 2.2.2**.

2.1.2.4 Particle Size

The microstructure, in particular the primary particle size, dictates the length of the Li diffusion pathway. This will in turn determine the reversibility of intercalation reactions as well as the charge/discharge rate. The microstructure of the cathode material can be altered using different synthesis methods and precursor materials [123]. *Drezen et al.* [162] investigated the electrochemical potentials of $LiMnPO_4$ with a range of primary particle sizes. The results presented in **Figure 2.7** demonstrate that smaller particle sizes (140 nm) have a higher initial capacity and capacity retention rate at the 9^{th} cycle relative to larger particle sizes (160 nm, 200 nm, 270 nm, 830 nm), most likely due to the shorter Li diffusion pathway in the relatively low (relative to the electrolyte) ionic conductivity material. In addition, the change in particle size affects the specific surface area of the materials. *Xia et al* [163] observed a better charge/discharge rate for $LiMnPO_4$ with a larger specific surface

area (24.1 m^2g^{-1}) than the smaller ones (14.2 m^2g^{-1}, 4.6 m^2g^{-1}). This is caused by the larger contact area with the electrolyte which increases the rate of Li-ion insertion/extraction.

The current state-of-the-art is nanoscale active cathode materials with a particle size less than 100 nm [164, 165, 166, 167, 168]. Nevertheless, this introduces issues such as low packing density and are more susceptible to interface reactions with the electrolyte which limit their commercial viability [169]. Moreover, the morphology of the nanoparticle can alter its electrochemical performance, with nanotubes (1D structure) [148], nanoplates (2D structures) [170, 171, 172] and hierarchical nanostructures (3D structure) [173, 174] all having been investigated.

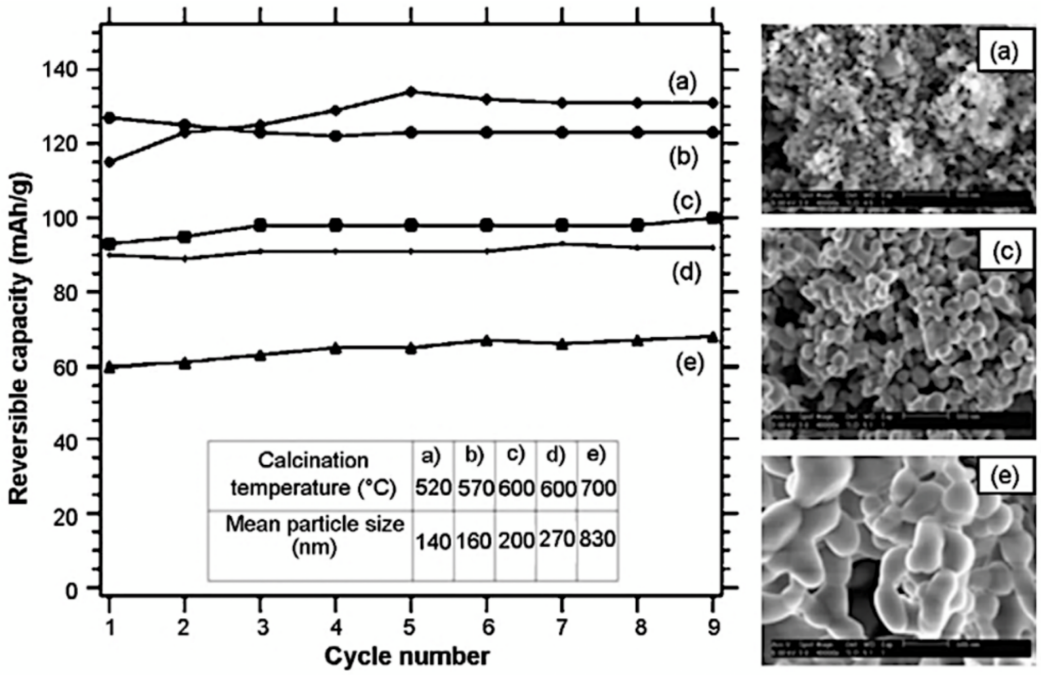


Figure 2.7: . The cyclic performance of $LiMnPO_4$ cathode with different particle sizes from 520nm to 700nm against the lithium metal anode and the SEM pictures for three materials with different particle sizes. Reproduced with permission from [12].

2.1.3 Material Discovery Progress and Limitations

The development of new cathode materials with higher storage capability is the major key to unlock the full electrochemical potentials of LIB. Despite a few new promising material formulas that have been discovered in recent decades, the pro-

cess of material discovery, in general, suffers from a great time span. **Figure 2.8** illustrates the whole process of material research from the initial discovery of new material through development until commercial deployment. Often, material discovery is the most time-consuming step with a projected time of 15 years on average to complete. A thorough understanding of the relationship between composition, crystal/microstructure and material properties is an essential prerequisite for the design and discovery of new functional cathode materials. There are in general two conventional methods being used to gain these insights: theoretical computation using techniques such as DFT and traditional experimental methods.

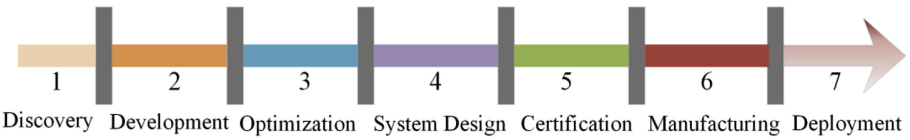


Figure 2.8: A schematic illustration of the material discovery process from the conventional research method. Produced with permission from [13].

Experimental research consists of material synthesis, fabrication and characterisation. For most cathode synthesis, the correlation between the synthesis route, precursor, and the final products' properties such as the particle size and morphology are still unclear. For instance, the cheap solid-state reactions synthesis route for $LiMn_2O_4$ offers little control over the particle distribution and particle size of the final material despite its wide use in large-scale production [175]. In addition, synthesis routes such as the hydrothermal reactions and the sol-gel process require niche, expensive equipment [176]. Furthermore, the high possibility of failure due to the unsuitable control of processing variables should not be neglected as this greatly increases the material cost over the long term and delays the discovery process. Electrochemical property tests, and in particular whole cycle tests, can be extremely expensive and time-consuming, with the latter for up to 25 cycles with a C/10 discharge rate taking nearly 20 days to complete [93]. The prediction of the material properties can also be done through various theoretical simulation methods such as density functional theory (DFT), Monte Carlo, molecular dynamics and the phase-field method. These computational methods offer the advantages of being

much cheaper and faster than experimental methods. DFT is widely seen in the research for different cathode classes such as NCM and $LiMn_2O_4$ and the studies range from the prediction of the crystal structure, Li-ion diffusion energy and the average voltage during cycling [177, 178]. Nevertheless, an average deviation percentage of 1-2% is observed between the DFT predicted lattice constant results and the experimentally observed lattice constants [179]. In most cases, a prerequisite for DFT simulation is an approximation of the crystal structure. This can be limiting, as in the case of NCM where each cation can partially occupy different 3b lattice sites, requiring the modeller to perform thousands of simulations to get a reasonable statistical sample of all possible atomic arrangements to determine the most stable structure with the lowest free energy [180]. In addition, the applications of the molecular dynamics and the Monte Carlo methods are limited by the need for empirically derived molecular force fields (DFT, although universal becomes computationally intractable for more than a few hundred atoms) before performing the systematic potential energy calculations. In general, computational simulation methods are mostly used as a guide tool to direct experimental research. The commonly shared limitations of these methods can be summarized as 1) the requirement of a high computing-power platform, 2) require specific input information about the material such as the full details of atom arrangements and crystal lattice parameters and restricting the size of the molecules, 3) lack of flexibility in stimulating various material length scale such as DFT cannot be used to model the grain boundaries which are at the macroscale level. The drawbacks of both experimental and simulated research can be addressed in part by using data-driven approaches, which will be explored in the subsequent chapter.

2.2 Review of Data-Driven Research

2.2.1 Introduction

Ever since the Materials Genome Initiatives (MGI) was announced in 2011, more and more research attention has been drawn towards the use of complementary data-driven methods [181] which have been enabled by widescale materials data collec-

tion, curation and digitisation. Using this data, advanced statistical methods can be used to uncover the underlying correlations between multi-dimensional variables and provide a much faster prediction of material properties than either experimental or simulation methods.

Although this concept was introduced very early, many obstacles were faced such as the lack of computing power and limited data resources and hence, it found limited industrial applications. The 2010s saw an explosion in the development of data-driven approaches due to increased computing power and more established online data-sharing platforms, which attracted worldwide interest. It has now become a key contributor to the global economy and more especially in the technology industry. The current size of data science market revenue is seven times more than for 2011 as shown in **Figure 2.9**. This high growth rate is predicted to continue, with revenue of 103b$ in 2027.

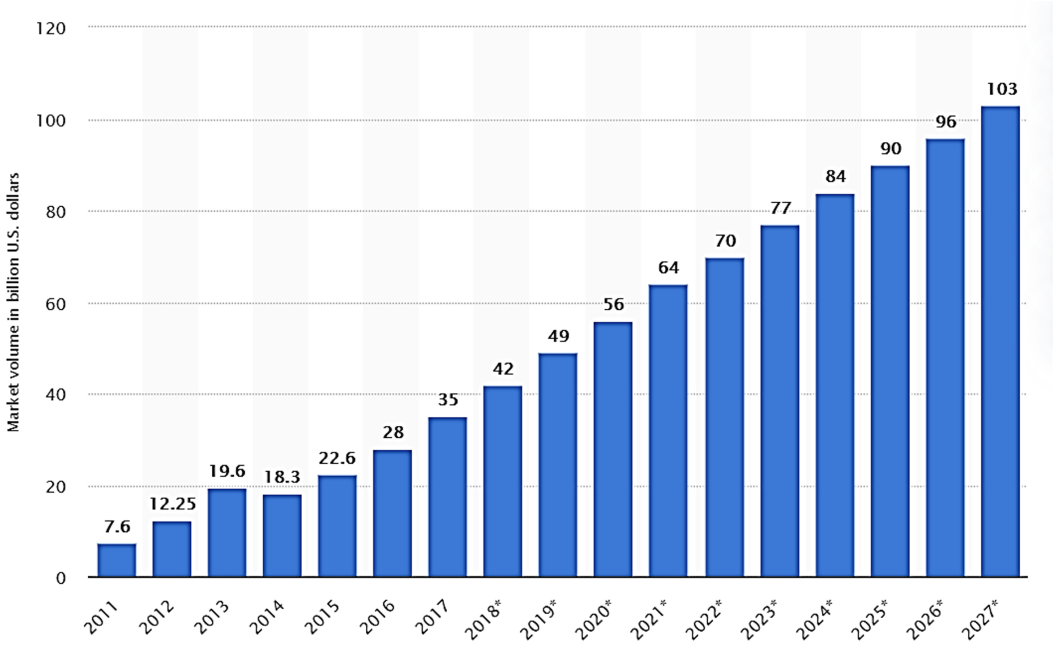


Figure 2.9: The Forecast for the worldwide revenue of the data science market from 2011 to 2027. Image reproduced with permission from [14].

Applications range from IT to drug discovery. For example, google Deep Mind [182] developed an ML algorithm to play GO, which defeated the world GO champion. Other examples include models for facial recognition [183], sound classifica-

tion [184] (currently used by Amazon - Alexa), as well as the recently developed autonomous car [185].

Data-driven methods can broadly be divided into supervised and unsupervised learning. The supervised machine learning method is seen as a much more mature and useful technique in giving underlying insights into the correlations of material structure properties. Unsupervised machine learning technique is normally used to investigate basic trends from the dataset such as clustering different types of materials based on their shared systematic properties and therefore can provide fewer insights. In the supervised learning method, a well-labelled dataset consists of input and output variables are collected either from the available database, experiments or computational simulations. **Figure 2.11** illustrates the essentials steps involved in a project using machine learning to perform material science research. At first, a clear outlook of the machine learning model should be established including the identifications of the covariate variables and the response variables. The covariate variables are referred to as the input variable to the model and a machine learning algorithm automatically try to generate a mathematical function that can link the input information with the target response variables which is also known as the model output variable. For the applications in material science research, the covariate variables are divided into three main categories, namely, the material composition information, structural properties and the chemical and physical properties of the elements such as the ionic radius and melting points [186]. The compiled dataset was then carried out with data cleaning (removing the empty entries) and being centred and scaled to ensure that the dataset is high quality and suitable for the later ML training.

Before the ML training, the dataset is normally split into a training set and a test set with a user-specified ratio. A rule of 8/2 for the training/test splitting ratio is typically used in ML research as the size of the training set is large enough for the ML model to capture enough signals whilst the remaining test-set is large enough to test the model performance. The next step is followed by the selection of the underlying algorithm. the selection of the underlying machine learning algorithms should be carefully selected as each of them have their unique advantages

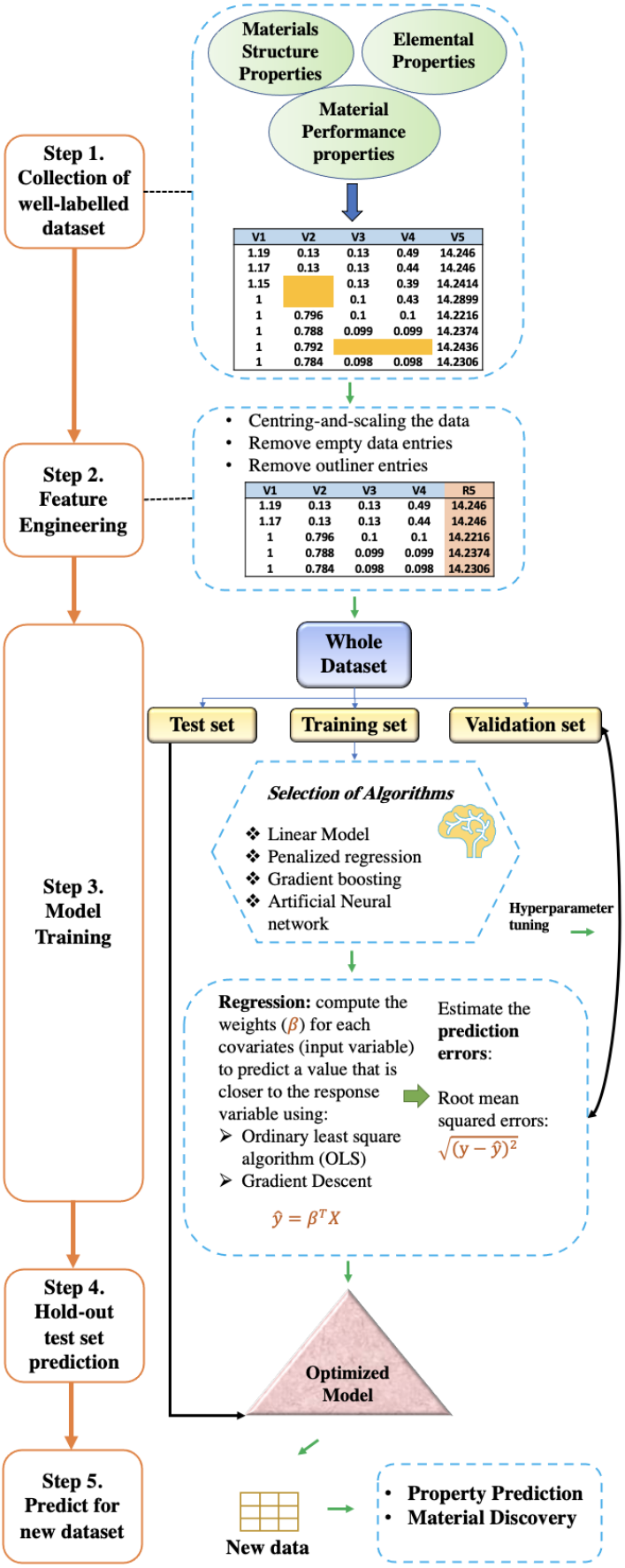


Figure 2.10: Schematic illustration of the supervised learning process.

and disadvantages which can greatly impact the performance of the model [187]. Such comparisons of different types of algorithms have been given in **Table 4.4** and **Table 4.3**. Broadly speaking, there are four categories of algorithms being used to conduct data-driven research with each tailored to their unique application. **Figure 2.11** illustrate the common machine learning algorithms under each algorithm category. Regression, classification and, less frequently, clustering algorithms are mainly used for material property prediction. In addition, the probability estimation algorithm is mainly used for predicting the likelihood of new materials with new crystal structure or component combination to have suitable functionality and the successful candidates are normally selected for further investigation with experiments or DFT simulations.

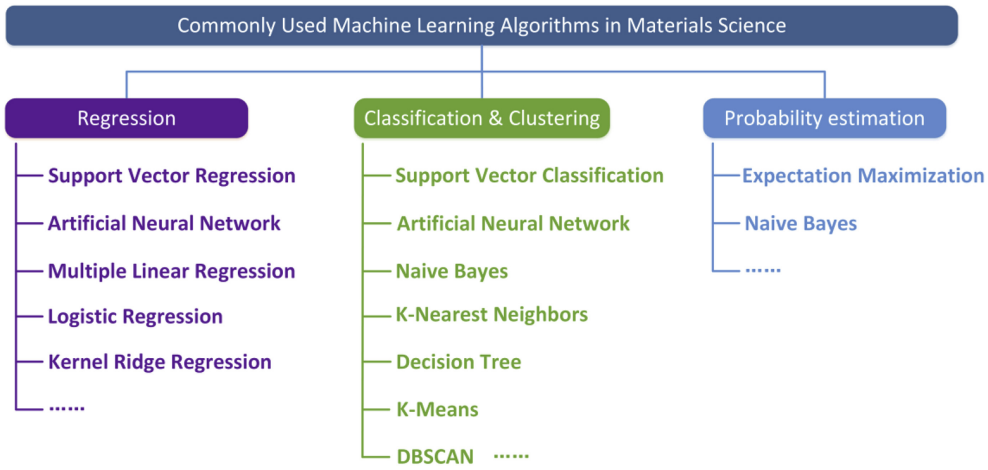


Figure 2.11: A list of the machine learning algorithms that are commonly used for material science research. Reproduced with permission from [13].

During the training of the ML model, overfitting issue normally occurs as the results of the model overlearn the dataset and capture more noise than the underlying signals. This would result in a poor prediction performance over the new dataset. As a common strategy to tackle the issues of ML model overfitting to the training set data, the training set is normally split further into N equal-sized sub-datasets, and these are gradually fed into the model. In the case of regression models such as linear regression, the training process aims to construct a mathematical formula consists of the suitable coefficients of each covariate and can predict a value closer to the response variable. This process can be achieved using weight optimisation al-

gorithms such as gradient descent, or the ordinary least square method as illustrated in **Step 3** in **Figure 2.10**. These weight updates are driven by the goal of minimising the difference between the final predicted value and the target response variable value. Furthermore, another essential goal of ML training is to seek the optimal model hyperparameters that would grant the model the best prediction power. These hyperparameters can be seen as the model settings which affect the length and the complexity of the learning process. A clear example can be seen from the needed specifications of the number of hidden neurons and the number of hidden layers for a neural network model and these hyperparameters could be tuned by computing a wide range of parameter combinations and select the pair that generates the lowest prediction error. It is important to note that the prediction error denotes the difference between the predicted value and the real response variable value and can be calculated through the root mean square error formula indicated in **Figure 2.10**. Finally, further assessments of the prediction power of the optimised ML model can be made by predicting against the hold-out test set. This hold-out test set was split from the main training set and therefore would result in less bias towards the final confirmation of the ML prediction power. Different ML algorithms embed different underlying principles for learning and predicting and therefore their respective performance can vary greatly. A common research strategy is to construct and compare different algorithm-based ML models under the same test dataset (controlled by setting random seed) and select the model with the lowest prediction error for further analysis. In the end, the material scientist can utilise these constructed ML models to predict the material properties or to facilitate the discovery of new materials.

2.2.2 Implementation in Material Science Research

ML-assisted research methods have shown great promise in assisting the recent material science research due to the increasing volume of available databases and the boosted computing power in recent years. Several high qualities review papers have highlighted the recent significant research results obtained on several energy materials such as solar cell material, catalysis and battery materials [186, 188, 189, 190, 191, 192]. The types of ML-assisted research can also be split

into the themes of the prediction of material properties and the prediction of new materials. Material property prediction concerning the lithium-ion battery materials is dominated by the prediction of the stability of the materials (formation energy, cohesive energy), material properties (crystal structure lattice constant, crystal systems, elastic properties) and the performance metrics related properties (voltage, Li-ion diffusion, discharge capacities). On the other hand, the prediction of new materials is made through predicting which crystal structures that might be stable (or meta-stable) for a given composition or predicting which chemical composition might have desirable material properties. In the following section, further discussions will be given on ML-assisted research in each research domain with the illustrations of some examples.

2.2.2.1 Material Property Prediction

The fast prediction of the material properties such as lattice constant, the bandgap can give researchers valuable insight into the final materials before carrying out any expensive experiments. This is particularly important for new cathode material research considering that both material properties influence both the electronic conductivity of the material and hence impact the rate of Li-ion intercalation/deintercalation during the electrochemical performance. Attempts have been made with using the Gaussian process regression [193, 194], artificial neural network (ANN) [195] random forest algorithm [196] to predict the material's lattice constant with predictions having comparable accuracy to the experimental results. DFT with the general gradient approximation functional often underestimates this value and the use of higher complex hydride functionals are more computationally expensive [197]. To explore the potential of the data-driven method, the prediction of band gaps for inorganic materials are seen using various algorithms such as support vector regression [198], kernel ridge regression and randomised forest [199], cubist regression [200] which all models have given comparable predictions of the bandgaps with the DFT method, with a much lower computational cost. ANN algorithm is inspired by the biological structures of the neural network in our brain and construct various neuron points to form the final prediction. One major disad-

vantage associated with this type of algorithm is the requirement for a much larger size of the dataset due to many present internal parameters. This in turn would lead the model to have less prediction power over the unknown dataset and thereby less applicable in guiding the material research. On the other hand, the tree-based algorithm random forest is less likely to overfit the training data and have shown to have higher prediction power in general.

One crucial condition for achieving a high-rate capability is that the cathode material should have good diffusion properties to enable the fast or slow withdraw/insert of the Li-ions. The computation of electronic conductivity can be achieved using Nudged elastic band calculations, however, this calculation method is known to be difficult and considerably costly to perform. In recent years, promising results have been seen from the examples of implementing machine learning methods to predict the Li-ion conductivity of inorganic battery materials with comparable results from the simulation results [201, 202, 203, 204, 205]. Similarly, the examples of predicting other relevant properties have also been attempted such as the migration and the hopping energies [206, 207, 208], formation energy [209] and the elastic properties such as the elastic tensors, bulk moduli and shear moduli constants [210, 211, 212, 213].

The lack of available datasets is often reported as the major hindrance for ML methods, particularly for neural network models. On the other hand, ensemble-averaged ML methods such as random forest or gradient boosting have been suggested to have robust prediction power over smaller datasets than the non-ensemble models [214]. The selection of a suitable ML method should be made according to the problem on hand. It is commonly known that no “best” ML approach would solve all material related issues according to the “no free lunch” theorem [215]. These promising examples have demonstrated that the machine learning methods have a strong ability to uncover the complicated variable correlations whilst providing effective predictions for the properties that are interested.

2.2.2.2 Materials Discovery

The design and discovery of new materials have also been a key driving force in material science research. The traditional material discovery methods consist of elemental substitution and perturbing known crystal structures, and these are found to be both time-inefficient and computationally expensive. The data-driven material discovery method trains the algorithms with the molecular and structural behaviours of the existing compounds and is used in the prediction of new materials with a new composition or crystal structure.

The essential aim is to find a composition/crystal structure with a spatial configuration of atoms that is thermodynamically stable. In general, the component prediction method is more commonly seen than the crystal structure method because the formula's components and composition are easier to modify. **Table 2.1** lists examples of both methods. It is noted that tree-based models such as decision trees, random forests have been frequently applied in predicting new material such as mixed-anion conductors and Co-V-Zr metallic glasses [216]. Besides the high prediction power, these tree-based models are applicable for small datasets, and this is more advantageous than the artificial neural network where a large dataset is normally required to achieve a sufficient prediction force and the relative computing time for finding the optimal internal parameters is much higher. Although having mentioned it, the ability of the ANN algorithm for handling many complex problems should not be neglected especially when the model is fully optimised. In addition, there are many cases of implementing a Naïve Bayesian algorithm (NB) to predict the likelihood for certain compositions or structures from having the stability of the desired properties. Similar to the tree-based models, the NB models are much easier and quicker to train (small dataset) whilst providing a high prediction power. However, its major shortcomings can also be seen from its presumption of covariate variables being independent of each other and this is rarely the case in the real life.

Table 2.1: Examples of data-driven material research in crystal structure prediction and component prediction.

Reference	Research Purpose	ML method	Achievement	Data Source
Crystal Structure Prediction				
Amsler et al [217]	Discover new mixed-anion semiconductors	Decision Tree	Identified 21 new compounds with I-42d space group	simulation
Liu et al [218]	Design new microstructure of Fe-Ga alloy materials	Decision Tree	Discovered 5 new alloy microstructure that fulfils the property requirements	Stimulation
Component Prediction				
Ren et al [216]	Design new Co-V-Zr system metallic glasses	Random Forest	Identified 3 new metallic systems, validated experimentally	Experimental
Jalem et al [207]	Discover new ionic substituted compounds	Bayesian	Identified 20 new ionic substitutions	Stimulation
Hautier et [219]	Discover new ternary oxide compounds	Bayesian	Identified 209 new compounds	Experimental

2.2.3 Review of Data Sources for Material Science Research

Data-driven research is predicated on the availability of relevant data. Ever since the first open-access journal website named arXiv was established in 1991, the concept of data sharing has been widely promoted in the research community. Nowadays, it is promising to see that the strict requirements are placed by many journals and funding agencies on making data assessable to the readers. In material science, the raw data used for training or testing can be collected either through conducting the physical experiment or performing first principle-based computational stimulations. After years of continuous effort in collecting and collating high-quality data, many data infrastructures are seen to be containing a high volume of data and increase the chance of conducting data-driven material research.

Table 2.2 summarises a few publicly assessable datasets that can be used to conduct inorganic material research. In general, the available datasets can be split

into the types of experimental data and computational data. The inorganic crystal structure database is one of the largest repository sites that contains over 190,000 entries of data on the reported experimental crystal structures. Structural properties such as the bond length, crystal structure positions are included in the database. These can be used to perform DFT stimulation or to be trained on as the variables in the machine learning model. In addition, Material projects stores nearly four folds of material data that are estimated by the DFT calculations. This web-based dataset offers more diverse information on the material compounds such as the phase diagrams, elastic properties, and crystal structure information. Different to other databases, the Material project includes calculated battery performance properties for the battery materials including the average discharge voltage and the estimated volume change during discharge. These features are essential in screening suitable battery materials that exhibit good cycling properties and increase the chance of discovering new functional battery materials.

The quality and the quantity of the data are the two major factors that govern the prediction performance of the machine learning models. One common issue with the experimentally collected data is the high likelihood of human errors or measurement bias during the collection process and the lack of sample uniformity such as the material impurities and the different product phases. The lack of reporting standards in the measurement of the essential material properties can also lead to unreproducible data and hence increasing the uncertainty of the data accuracy. For the large databases in **Table 2.2** follow a strict data management plan and practise regular maintenance work such as clearing out the data entries that are reported to be faulty. Finally, failed experimental data can also be useful, for example, *Raccuglia et al* [227] used a support vector machine model trained on failed reaction data to attain 89% accuracy for the prediction of reaction products.

2.2.4 Data-driven Research for LIB Cathode Materials

Leveraged from the large volume of publications made in LIB materials research ($>$7000) every year [93], machine learning is becoming a more commonly used tool by the research community, especially for the investigation of cathode materials.

Table 2.2: A summary of the available databases for accessing the structure and properties of inorganic materials.

Dataset Name	Description	Type of Data
Inorganic Crystal Structure Database (ICSD) [220]	Contains more than 210,000 experimentally synthesised inorganic crystal structures for binary, ternary, quaternary and quintenary compounds	Experimental
ChemSpider [221]	A chemical structure database established by the royal society of chemistry containing more than 100 million structure records and containing various experimental and computationally estimated properties	Experimental and computational
Crystallography Open Database [222]	Crystal structure dataset containing the calculated structure properties for over 1,561,023 compounds (organic, inorganic, metal-organic, mineral)	Computational
Citrination [223]	AI-powered data-sharing platform containing the experimental and computed properties of over 4,023,176 compounds	Experimental and Computational
Material Project [224]	Chemical and physical properties of over 16,000 known and hypothetical compounds calculated by DFT	Computational
Open Quantum Materials Database [225]	Covering the DFT calculated results for 815,654 materials' structural and thermodynamic properties such as the phase diagrams	Computational
AFLOWLIB [226]	Ab-initio computed structure and properties of over 3,400,000 inorganic materials	Computational

Primarily, ML is used to predict material properties of new candidate materials, such as the material crystal system, configuration energy and electrochemical properties such as the voltage, charge/discharge capacities and redox potential.

ML models based on DFT data have been commonly used to predict the formation energy and lattice parameters of candidate materials. *Li et al* [228] trained a support vector regression model using the DFT calculated formation energy and lattice parameters of 5329 spinel oxide (AB_2O_4) and attained a prediction accuracy of 0.81 and 0.961 in the R^2 scores respectively (R^2 scores range from 0 – 1). Furthermore, 451 new spinel structures are predicted to be stable and were recommended for further experimental validation. *Eremin et al* [229] applied

ridge regression models to predict the configurational energy of various forms of $LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$ and $LiNiO_2$ from their respective DFT optimised atomic arrangement, achieving a prediction error of less than 2 meV per atom relative to the DFT results.

Attarian Shandiz et al [230] compared the prediction power of six different machine learning models for classifying the crystal systems of 339 silicate-based cathodes (general formula of Li-Si-Mn/Fe/Co-O). These models take in seven DFT computed microscopic properties including the energy of formation, gap energy, energy above the hull, density, number of sites and the crystal volume to make the crystal system prediction. Among all seven machine learning methods, the random forest model has yielded the best prediction power with their overall average accuracy calculated to be 75% and 76% respectively. In addition, the results also revealed that the crystal volume was the most important variable that influences the classification of the crystal system.

Other research takes this one step further, using ML to replicate the forces predicted by DFT and thereby create ML – interatomic potentials that can be used to simulate much larger atomic systems (and therefore material features such as dislocations and grain boundaries) than DFT. *Li et al* [15] developed artificial neural network potentials (NNP) using the MD-stimulated Li-ion movement with their underlying forces and energies being described by the DFT simulations. The initial training set for the NNP includes information such as the snapshots of the MD-material structure under various temperatures (300K to 4000K) and the defected material structure with one missing Li-atom. These optimised potentials were then used in an MD simulation to predict Li-ion diffusion behaviour in amorphous Li_3PO_4. **Figure 2.12** compares the prediction results of using NNP and DFT for the prediction of Li-ion vacancy formation energy in $Li_{12}P_4O_{16}$ and of the Li-diffusion energy barrier, with small deviations of 0.029 eV and 0.01 eV observed respectively. These errors are comparatively smaller than the intrinsic DFT errors. The NNP simulation was 4x faster than DFT and can therefore be applied to stimulate much larger atomic systems.

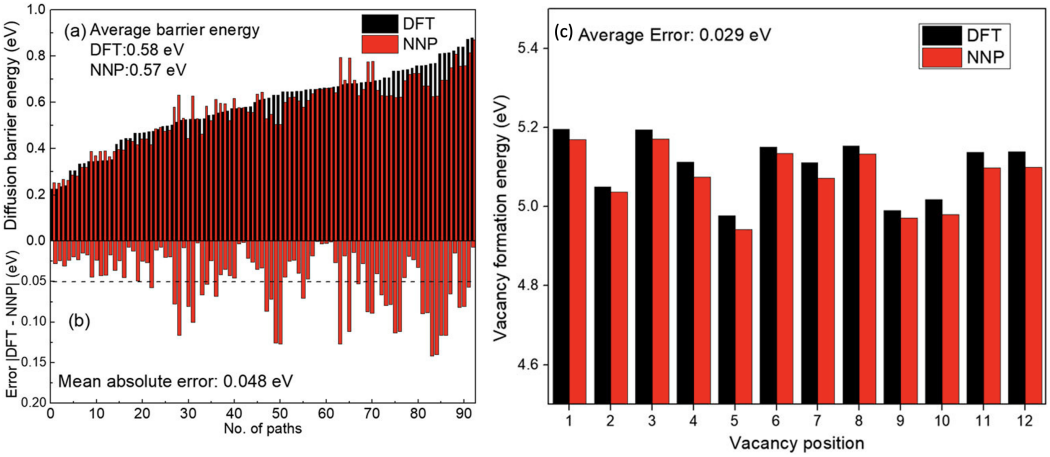


Figure 2.12: The comparison of the results computed using the neural network potentials and using the traditional DFT methods for estimations of (a) barrier energy of the vacancy diffusion path in Li_3PO_4, (b) their prediction differences and for calculating (c) Li-vacancy formation energy in $Li_{12}P_4O_{16}$. Reproduced with permission from [15].

The organic cathode material is also another choice due to its good redox potential, cheaper price and highly sustainable and steady performance after repetitive charging and discharging. *Allam et al* [231] built deep neural networks to predict the redox potential for quinone molecules from variables such as HOMO, LUMO energy, electron affinity, HOMO-LUMO gap and the design of the model are given in **Figure 2.13**. From this study, electron affinity was determined to be the most influential parameter for the prediction of redox voltage and an average prediction deviation of 3.54% was obtained after validating with the existing literature results.

In addition to intrinsic material properties that can be reasonably predicted based on DFT data sets, overall device performance, including the electrochemical performance of the cathodes during cycling is also of interest. This can be captured from their voltage profile and in some cases, the respective property prediction can lead to the discovery of new cathodes. *Josh et al* [232] trained deep neural network (DNN), support vector regression (SVR) and kernel ridge regression (KRR) to predict the average voltage using the information of the active metal ion concentration, crystal lattice types and the space group of 3580 cathodes for six different types of metal-ion batteries (Li, Mg, Ca, Al, Zn, Y). Out of the three models constructed, DNN shows the best prediction power with R^2 accuracy of 0.81 when predicting

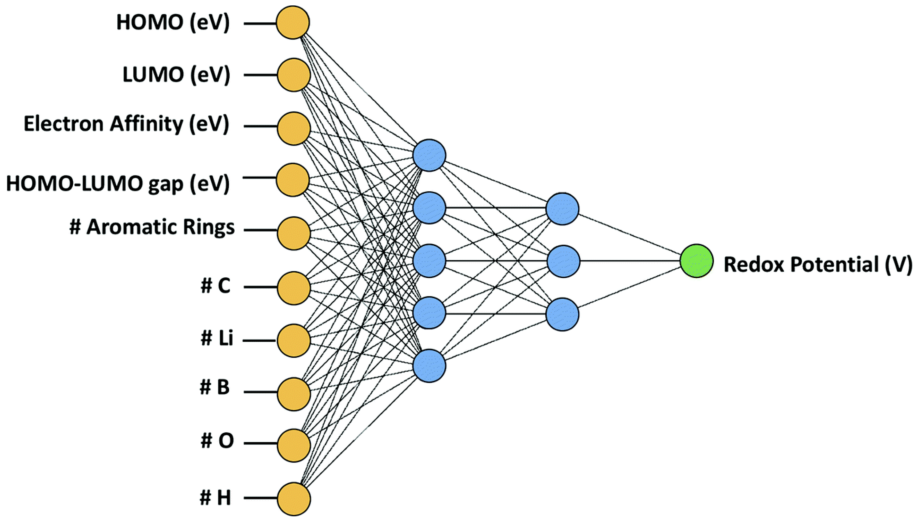


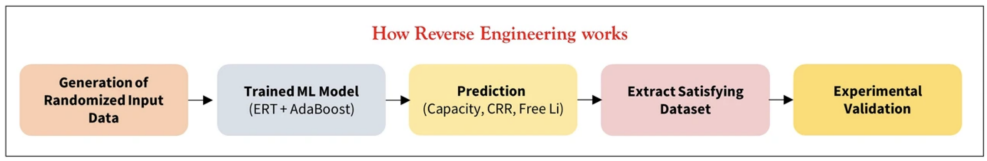
Figure 2.13: A schematic illustration of the deep neural network for the prediction of redox potentials of organic molecules. Image reproduced with permission from [16].

for the published experimental average voltage. In addition, 5000 new electrode candidate materials have been identified for Na and K-ion batteries. *Sarkar et al* [16] built a three-layered feed-forward neural network model using the electronegativity values of the component elements of 31 different LIB cathodes to predict their average voltage. The model results in an average prediction error of 0.04 V when validating against the experimentally measure voltage. Similarly, *Houchins et al* [233] trained a Neural Network model with the DFT calculated Behler-Parrinello symmetry functions of five phases of $LiNi_xMn_yCo_{(1-x-y)}O_2$ compounds and used it to predict their average voltage at the temperature of 300K. In the end, the model resulted in an average deviation of 0.042V from the cluster expansion stimulated results.

The charge and discharge behaviours of LIB are rather difficult to be explained by any simple physical models because of the presence of a non-linear relationship and the multi-dimensional variable space (electrode properties, test conditions, etc.). *Min et al* [17] implemented various non-linear machine learning techniques (support vector regression, neural network, etc.) to study the correlations between the synthesis parameters and the capacities of nickel-rich cathode $LiNi_xCo_{1-x-y}O_2$, where x $>$0.8. Among all ML models, the AdaBoost model was found to have the highest predictive power. It identified the synthesis temperature, Ni-content and

particle size as the most influential variables for both the initial discharge capacity and the capacity retention rate at 50 cycles. In addition, the reverse engineering method (**Figure 2.14-a**) was applied to predict the performance for 5000 randomly generated sets of parameters, and it led to the discovery of 5 news nickel-rich cathodes with superior experimental performance as shown in **Figure 2.14-c**. Similarly, *Kireeva et al* [234] implemented the support vector machines model to predict the initial discharge capacity and of the coulombic efficiency of 98 lithium-rich layered oxide cathodes based on seven different material properties such as the composition, disorder for the site, synthesis data, electronegativity, work function and the Shannon ionic radii. The final optimised models give prediction errors of 18.2 $mAhg^{-1}$ and 4.02% against the hold-out dataset for the predictions of initial discharge capacity and the capacity fading, respectively. It also identified that the manganese content and the Shannon radii of the Li-ion are the two governing variables.

(a)



(b)

Set	Composition	Temperature	Dopant	Washing	Coating	Li	Ni	Co	Mn	Size	Ratio	xrdA	xrdC	Capacity	CRR	Free Li
1	1	700	Bi	1	M	0.990	0.865	0.093	0.043	90.4	1.56	2.876	14.224	212	93.6	646
2	1	700	Bi	1	M	1.006	0.867	0.092	0.042	91.5	1.62	2.876	14.210	215	93.4	621
3	1	700	Bi	1	M	1.003	0.861	0.099	0.042	88.7	1.65	2.874	14.211	218	93.7	560
4	1	700	Bi	1	M	0.961	0.855	0.105	0.041	88.2	1.62	2.876	14.217	213	93.2	1181
5	1	700	Bi	0	N	0.991	0.870	0.091	0.041	91.9	1.68	2.874	14.207	219	93.7	408

(c)

Set	Composition	Temperature	Dopant	Washing	Coating	Li	Ni	Co	Mn	Size	Ratio	xrdA	xrdC	Capacity	CRR	Free Li
1	1	700	Bi	1	M	0.99	0.865	0.092	0.043	90.5	1.56	2.876	14.213	199	92.7	552
2	1	700	Bi	1	M	1.00	0.867	0.091	0.042	91.3	1.62	2.875	14.210	204	93.6	526
3	1	700	Bi	1	M	1.00	0.861	0.097	0.042	88.8	1.65	2.874	14.201	206	93.9	550
4	1	700	Bi	1	M	0.96	0.854	0.104	0.041	88.2	1.62	2.875	14.206	206	95.1	1045
5	1	700	Bi	0	N	0.99	0.869	0.090	0.041	91.9	1.68	2.874	14.207	198	95.6	475

Figure 2.14: The processes of reverse engineering. Reproduced with the permission from [17].

Besides the use of traditional material properties as predictors, there are other examples of generating new predictor variables through molecular stimulation tools. *Choi et al* [235] proposed a 3D quantum artificial neural network model to predict the discharge energy density and capacity fading rate for LIB cathode materials. The novelty of this work is seen from using the distribution of the electrostatic po-

tentials (estimated by DFT) in cathode material supercell as the ML model input variable as opposed to the formula's elemental variables. In the end, the final models demonstrated great prediction power for predicting discharge energy density and capacity fading rate with R^2 scores of 0.905 and 0.979, respectively. *Kauwe et al* [93] used a combination of DFT computed variable (Dipole-moment), elemental variables (average electronegativity, average atomic mass, etc.) and experimental variables (anode material, cathode structure, etc.) to predict the experimental discharge capacity of various cathode materials in LIB. Random forest was selected to be the best performing model with a promising prediction error of 15.7 $mAhg^{-1}$.

The microstructures of electrode materials are influential factors for many essential material properties such as the internal resistance, mechanical strength and the understanding of these correlations are essential for the development of new electrodes with better cyclability. *Takagishi et al* [18] investigated the relationship between the mesoscale structure of the 2000 artificial 3D LIB electrodes and the respective specific resistance in the battery system. First, stimulations of the Li-ion diffusion specific resistance experienced in the active particle of cathode, electrolyte and their interface are performed using a physio-chemical model. In addition, a three hidden layered ANN model was built to predict the resistance based on the porous structure features such as the volume fraction and particle radius of the active materials, the volume ratio of the binder in the cathode, conductivity of the electrolyte and the pressure applied during the compaction process. The ANN final predicted resistances have shown good agreement with the stimulation results with the highest accuracy reported to be R^2=0.99. More recently, *Jiang et al* [18] investigated the correlations of the degree of attachment and detachment of the particle from the carbon/binder domain (CBD) in the $LiNi_{0.8}Co_{0.1}Mn_{0.1}O_2$ (NCM) composite cathode system with the charge and discharge performance. They trained and built a recurrent neural network model to annotate the types of particles, degree of cracking for the rendered CBD microstructure images. **Figure 2.15-a** demonstrates the overall working process of this machine learning approach and **Figure 2.15-b** highlights the results of using the conventional separating algorithm and the recur-

rent network model to identify the particles for 650 microstructure images. The recurrent neural network has higher accuracy in terms of identifying the different phases of the NCM than the conventional method (more uniform in the identified particle colour in **Figure 2.15-b**. In addition, the implementation of this machine learning approach was helpful to characterise the degree of particle detachment for different NCM particles and in the end, the conclusion was drawn that the higher charging rate and the smaller particle can cause greater particle detachment from the CBD matrix.

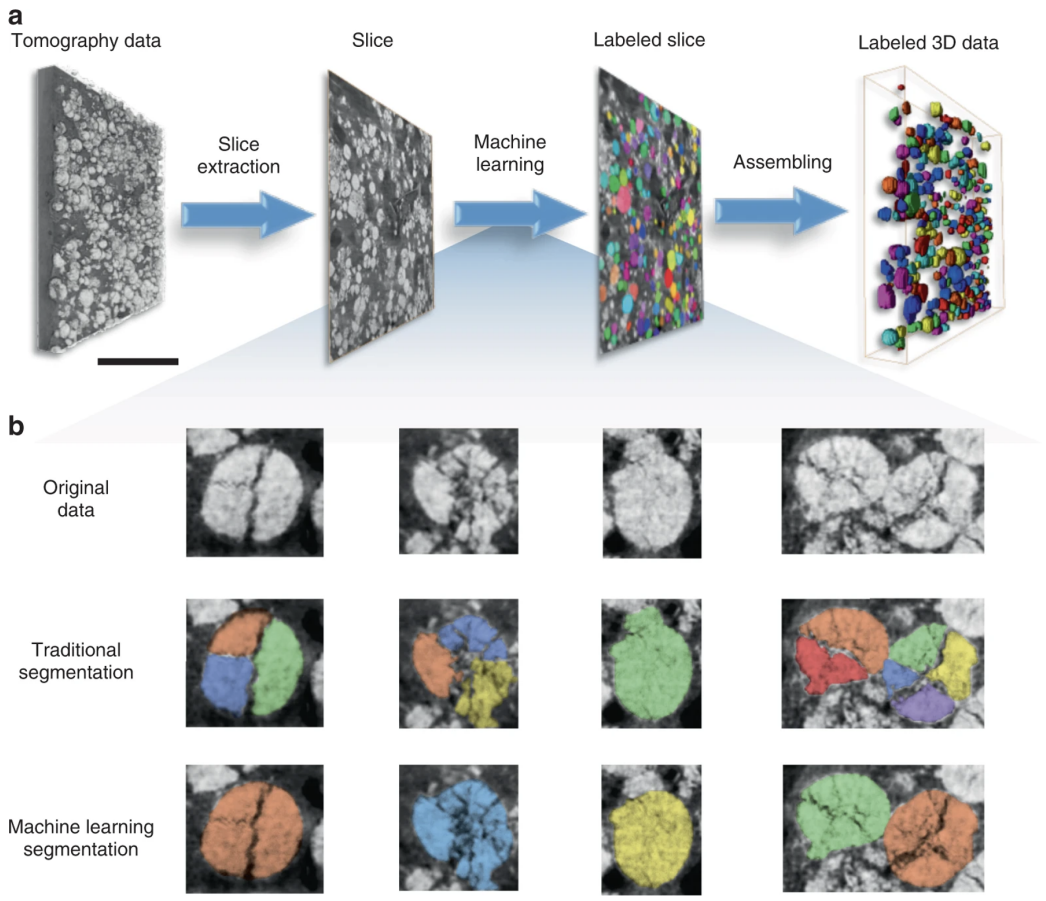


Figure 2.15: (a) The schematic illustration of the workflow for using a recurrent neural network model to perform the segmentation of the electrode materials. (b) In the comparison of the results from the traditional manual segmentation method and the machine learning segmentation method, the colour indicates the different particle labels. Reproduced with permission from [18].

Chapter 3

State of the Art of Machine Learning for Battery Materials Research

3.1 Introduction

Different machine learning algorithms embeds different prediction mechanisms, and this would lead to different prediction performance when they are trained with the same dataset. According to the recent review [236], a wide range of regression machine learning techniques has been implemented in battery materials research such as the ordinary least squares (OLS), decision tree, random forest, artificial neural network and kernel-based algorithms such as the support vector regression and the kernel ridge regression. In this chapter, a detailed discussion will be provided for the implementation example of different machine learning models in battery materials research.

3.2 Performance Comparison between Linear Model and Non-linear Model

Table 3.1 gives a few successful examples of applying the machine learning methods to assist the battery materials research. Various studies can be seen such as to predict the ionic conductivity of the electrolyte materials, the Li-ion diffusivity of materials and investigate the battery materials' formula-performance correlations. It is observed that non-linear algorithms such as support vector machines, neural

networks have been implemented more frequently than the linear algorithms like the least absolute shrinkage and selection operator (LASSO) model and this phenomenon has also been suggested in the recent review [236]. Jalem et al [206, 237] concluded that the optimised artificial neural network model can provide a much better prediction over the Li diffusion barrier of olivine materials ($LiMXO_4$) than the linear partial least square model does. The underlying reason was believed to be that linear models are overly simplified and therefore are not capable of capturing any non-linear correlations between the battery materials structural and performance properties. Other successful application cases of using non-linear models can also be seen from *Ahmad et al* [213] 's work on using kernel ridge regression and the gradient boosting model to predict the elastic constants of the solid electrolyte. The final prediction errors for the shear and bulk moduli of the inorganic materials are shown to be 0.1268 and 0.1013 respectively and twenty new inorganic materials are predicted to have the capability to suppress the Li- dendrite growth on lithium metal anode. These examples demonstrate the great prediction power of the non-linear model in explaining the correlations between the structural properties and the performance of battery materials.

3.3 Data Concerns

The prediction power of the machine learning models is largely governed by the size and the quality of the data. In battery material research, the lack of high-quality data is often reported as one of the biggest constraints in enabling the use of the machine learning (ML) method to conduct research. Simulation data is seen as the most frequently used data source for ML model training as opposed to the experimental data in **Table 3.1**. This is caused by the lack of a high-quality and accessible experimental dataset of the material performance properties in the material science research community. Simulation results could give useful insights into the theoretically calculated material properties, however, requires further experimental work to validate the accuracy of the results. In addition, the collection of experimental data from various publications could involve different degrees of uncertainties such

as the use of unstandardised experimental conditions or misreported information. This issue could be reduced by using various selection rules to guide the collection of publications so that the collected data are more comparable to one the other and could be used to indicate any true correlations of the interested variables.

3.4 Artificial Neural Network Model

Among most machine learning-powered materials science research, artificial neural network (ANN) is shown to be a rather popular choice. *Eslamloueyan et al* [238] built a multi-layer neural network model to predict the thermal conductivity of the electrolyte solutions using the molecular weight and the number of electrons of the solute. Within the study, 660 sets of experimental data were used to train and optimise the ANN model and a prediction error of 7.5×10^{-6} $\mathrm{Wm^{-1}K^{-1}}$ was reached after validating against the hold-out test set. In addition, *Parthiban et al* [239] built an ANN model to predict the charge and discharge capacity of lithium-ion batteries containing CoO anode for up to 50 cycles. In the end, the study yielded a promising 1% for the prediction error after validating against the real experimental results. These successful examples demonstrate the capability of the ANN model in capturing complex non-linear material correlations and provide good estimations of the performance property. On the other hand, the issue of lack of data for training the ANN has been commonly reported by various studies in the battery materials research [232, 240, 241]. ANN model has many neuron weights which requires a large amount of data to optimise the values for. As a way to reduce the requirement of the dataset, recent studies construct one hidden-layer ANN model with much fewer parameters than the multilayer case [238, 241]. On the other hand, this setting could also overly simplify the ANN model architecture and additionally reduce the model's prediction power.

3.5 Tree-based Model

The implementation of tree-based models is also commonly seen in the research, particularly the gradient boosting model and the random forest model. *Nakayama et al* [208] built a gradient boosting machine model and partial least square (PLS)

model using 400 sets of simulated data to predict the migration energy of the Li-Zn-X-O oxide compounds based on their composition details. The final results have indicated that the final prediction errors of the gradient boosting machine model is 0.52eV lower than the PLS and hence hold much higher prediction power. In addition, *Schmidt et al* [242] investigated the prediction power of ridge regression, random forest, gradient boosting trees and the neural network models for the distance of convex hull of cubic perovskite solid systems. The final results confirmed that the gradient boosting tree model hold the best prediction power with its root mean square error estimated to be 121 meV/atom. Other findings were made such that the linear ridge regression model provided the worst prediction over the test dataset and the ANN model was extremely time-consuming to train whilst its final performance was much worse than the tree-based models. In comparison to ANN models, tree-based algorithms such as random forest or gradient boosting machine, offers much fewer hyperparameters to be optimised during the training process and therefore are faster to train. Furthermore, they also seem to work flexibly with different scales of dataset and are shown to provide very reasonable predictions over unseen data.

3.6 Summary

The selection of suitable ML models should be tailored to the research goals and the availability of the dataset. In general, two main goals should be considered during this process, the prediction power and the capability to extract useful knowledge from the model. Linear models such as ridge regression and LASSO are much easier to train and implement and can be easily interpreted through the computed variable coefficients. However, their ability to find the correlations between the battery materials' structure-performance are very limited due to that the overly simplified model structure cannot capture complex underlying electrochemical behaviour. Because of this, many non-linear models such as support vector machines, ridge kernel regression and artificial neural networks receive more popularity in the research as they are known to be more sufficient in capturing non-linear correlations.

Table 3.1: Summary of the machine-learning powered battery research examples with on their research purposes, machine learning methods, achievements and the data source used for model training.

Reference	Research Purpose	ML Methods	Achievements	Data Source
Fujimura et al[243]	Predict the ionic conductivity of the LISICON-type materials	Support vector regressor with gaussian kernels	-Discover that the material Li_4GeO_4 has the highest conductivity -Reached a prediction error of 0.373 for the ionic conductivity	Experiments + Simulations
Jalem et al [237]	Predict the Li-ion diffusion barrier and the cohesive energy of $LiMXO_4$	Artificial neural network	-Reach low prediction errors for both energy properties -Identify 14 new olivine type $LiMXO_4$	Simulation
Duong et al [244]	Understand the effect of additives in the electrolyte on the capacity retention rate at various cycles	Artificial neural network	-Identified the optimal composition of additives and lithium salt that would lead to a high capacity retention rate at 100 cycles -Reached a prediction error of $10.3 mAhg^{-1}$	Experiments
Ishikawa et al [245]	Investigate the correlations between the coordination energy of battery electrolyte solvent and their elemental properties such as the metal ion radius, molecules atomic charge	Least Absolute Shrinkage and selection operator (LASSO), Multiple linear regression	Reached a prediction error of 0.016eV for the coordination energy	Simulations

Among all non-linear models, ANN model is well known for its disadvantages such as the lengthy and costly optimisation (tuning of learning rate, number of layers and neurons) and the low model interpretability (known as the black-box model). These issues then restrict its applicability in many research cases where the available datasets are relatively small. As a comparison, tree-based models such as the random forest and gradient boosting trees are much easier to train as they are constructed using simple decision trees. In addition, the nature of rule-based prediction mechanisms can also give much clearer insights into the correlations between the covariate variable with the response variable and compute the importance of each covariate.

Chapter 4

Methods

4.1 Introduction

To investigate the potentials of using machine learning techniques to predict the discharge performance of the doped cathode materials in lithium-ion batteries, several linear and non-linear regression algorithms are selected to construct the predictive models and compared with their prediction power. This methods chapter is split into two main sections. The first section entails the proposal of the general predictive model used in this thesis and a detailed explanation of the data collection process. Next, the theoretical backgrounds are given for the techniques used in performing data analysis, data engineering, model training and the performance metrics used. In addition, the underlying principles are included for the linear algorithm such as the ordinary linear regression, the penalised regression, Bayesian ridge regression and several non-linear algorithms such as the Artificial neural network, support vector machine, kernel ridge regression, K-nearest neighbour, decision tree, random forest and gradient boosting machine.

4.2 Model Proposal

A predictive model should consist of two major components, the covariate variables and the response variable. The covariate variables are the input variable to the model and also referred to as being predictors. The response variable, on the other hand, is the outcome variable that the model is predicting. Having discussed the current lack of understanding of the structure-property relationship of the cathode in the

previous chapters, we aim to develop a machine learning (ML) model to predict the discharge capacity of the cathodes based on the essential material features properties and the experimental parameters. In the literature, a wide range of covariates have been used to construct the machine learning model to predict different cathode performance properties. These can be summarised into the categories of elemental properties, material structural properties and experimental parameters. **Table 4.1** summarises a selection of potential covariate variables we proposed to implement into the ML models in this thesis. It is also important to note that the selection of covariate might be different in each study because of the difference in the investigated material class and the different times that conducted the research. Furthermore, the initial discharge capacities and the n^{th} cycle discharge capacities are selected as the model response variables, where n indicates a higher charge/discharge cycle number. Most of the material electrochemical testing studies have included the measurement for the initial discharge capacity as it is the initial cycle. However, the maximum performed cycle numbers n for discharge capacity are largely varied across different groups of studies. In this thesis, the scope of collecting the capacities for the higher cycle number "n" is governed by two main conditions. The first is the estimation of which cycle number is the most commonly performed maximum cycle number within each specific material class study. The second is the comparison of the collectable data volume under different cycle numbers. A detailed discussion of the collected dataset in each project is given in their experimental section.

4.3 Model Covariates Selection

The selection of suitable material descriptors has long been stated as a major challenge in the field of machine learning for material science research. In this thesis, the selection of domain covariate variables for the results chapters is decided based on two major criteria; i) the representation of the material features and their relevance to the material electrochemical performance ii) whether such data is widely reported or collectable.

The covariates used to construct the model can be categorised into three classes. These include the curated experimentally reported results such as the material compositions, crystal structure parameters (crystal lattice dimension and volume), cyclic test conditions (charge/discharge current density, operating voltage) and finally, the elemental properties for both the intrinsic material elements and the doping element in the structure.

To illustrate the difference in the structures of material classes, the experimental results from the material characterisation test such as the lattice constant parameters and unit cell volume are collected from the publications. The crystal lattice dimensions and lattice volume reveal the conditions of the material's crystal structure and can control the mechanisms of the Li-ion intercalation/deintercalation during cycling and therefore would influence the discharge capacities at both the initial and the higher cycle. In addition, these material physical properties can be used to represent the differences in the synthesis routine implemented as well as the precursors and are therefore chosen as covariates for the model.

The composition of the elements within material formulas reflect the chemical and physical properties of the materials and are expected to have great influence over the Li-ion intercalation/deintercalation reactions during the cyclic performance. The composition of the manganese and the dopant atoms in spinel cathode materials can influence the valence state of the manganese ions and therefore change the manganese dissolution rate during cycling. In the case of doped NCM layered materials, the compositions of nickel, manganese, cobalt and the dopant in the material formula controls the reaction kinetics, structure stability and the overall capacities of the materials. These compositional ratios should then be investigated for their correlations with the discharging performance at various cycles. In addition, the ratio of the based metal and the dopant elements within the olivine materials formula would indicate the level of the ionic conductivity of the materials and hence could largely influence the discharge performance.

Experimental conditions such as the current density, the minimum and the maximum operating voltage are collected as they account for the number of available

Li-ions for withdrawal in the material and the rate of Li-ion transfer within the battery system. For the carbon-coated olivine cathode material, the amount of carbon coating is selected as the covariate variable because it would influence the ionic conductivity of the material system and hence the discharge capacities. The direct correlations between the amount of carbon coating or the overall thickness and the scale of the improvement in the ionic conductivity are still yet to be confirmed and hence are further investigated in the ML model.

The elemental properties of the atoms in the material formula can indicate the chemical environment within the cathode system and additionally influence the discharging process. In the case of investigation of dopant effects, various elemental properties could be used to describe the unique properties of the atoms. For instance, the electronegativity of the dopant would indicate the characteristics of the bond it has with the surrounding atoms in the structure. This would indicate the stability of the intrinsic material structure and hence alter the longer cycle performance. In addition, other elemental properties such as the dopant ion's ionic size could mitigate the migrations of the Ni^{+} into the Li^{+} layer in the case of doped NCM cathodes. The dopant's electronegativity would influence the type of bond-forming inside of the material structure and hence the overall stability. Likewise, the ionisation energy and the electron affinity would affect the change in the systematic energy during the reactions of Li-ion insertion and the extraction and hence determine the structural stability. The ratios of the active material, conductive and binder during the fabrication of the cathode composites to investigate their influence over the material's electrochemical performance. In particular, the ratio of the active material and the conductive are important as one is directly related to the total amount of active cathode for carrying the discharge performance and the other controls kinetics of the Li-ion transfer within the structure. These materials features are widely available from the publications and can be used as covariate variables in the model.

Netherlevess, the degree of relevance of these covariates to the discharge capacity property are yet still to be investigated. It is important to note that some of these chosen covariates might be suboptimal which are considered less relevant to the

Table 4.1: A summary of the covariate variables for constructing the ML model including elemental properties, material properties and the experiment parameters for the cyclic performance test

Elemental properties	Ref	Material Properties	Experiment Parameters
Molar mass	[249]	Formula composition	Current density
Electronegativity (Pauling Scale)	[249]	Lattice constants "a"	Minimum operating Voltage
Number of electrons	[249]	Lattice constant "c"	Maximum operating Voltage
Number of isotopes	[250]	Crystal volume	
Ionisation energy	[249]		
Ionic radius	[251]		
Atomic radius	[252]		
Electron affinity	[253]		

discharge capacities and might introduce further noises to the model. As a result of this, the model prediction power will be greatly reduced. Some essential properties such as the particle size and the morphological features are known for influencing the discharging performance of the cathode materials [246, 247, 248]. However, such data is difficult to collect due to lack of reporting across literature and the uneven reporting standards (the Scanning electron microscopic figures are often reported in various scales and are difficult to analyse). These new covariates could be involved in the new model and investigated further for their correlations with the model prediction performance.

To investigate the importance of these selected variables, tree-based models such as the random-forest or the gradient boosting tree can be used to analyse the contribution of each variable to the overall prediction. The further details of such a method will be discussed in the following sections.

4.4 Data Collection

In this thesis, we manually curate the data from various open-access publications. The process of data collection involves the initial selection of the journals with criteria such as the type of studies the author had performed, the type characterisation methods performed for the material, the experimental parameters for the electrochemical test and finally, the measured cyclic performance. The initial collection of

the publication is done through keyword research over the research engine such as the web of science and google scholar. A general illustration of the criteria implemented during this data collection process is given below:

1. The publication should investigate the intercalation cathode materials for lithium-ion battery applications

2. The publication should have carried out the essential material characterisation such as the X-ray Diffraction (XRD) to determine the crystal structure of the material. Special cautions should be made on the reported space group, crystal lattice constant, crystal volume.

3. The publication needs to give a clear description of the material synthesis process and fabrications of the cathode electrode such as the ratio of the binder, additives and active materials and indicate whether the doping and/or coating has been applied to the material.

4. The material use of the counter-active anode should be kept at lithium foil and the use of electrolyte should be kept for $LiPF_6$ for standardisation purposes.

5. Experimental parameters such as the current density, operating temperature, the minimum and the maximum operating voltage should be indicated for the cyclic performance test.

6. The number of cyclic performance should also fit with the requirements for each research topic and this depends on the most widely selected cycle number from the study.

For every project carried out in this thesis, the data collection rules are strictly applied to ensure the high accuracy and consistency of the data. Publications that do not fulfil all of the selected criteria are discarded and are not considered for further data collection. For instance, some publications have fulfilled the required number of the charge/discharge cycle for the cyclic performance study, however, are omitted for collection due to the lack of XRD results or the experimental parameters. The

details of these collection rules used in each study are given in their corresponding experimental section.

Most of the reported data such as the crystal lattice constant, crystal volume from the XRD study are reported in the numerical format. On the other hand, the discharge capacities of the investigating cathode under various charge/discharge cycles are reported in the graphical format. A web-based tool named WebPlotDigitizer[254] is used to extract the data from the given cyclic performance plot and the detailed procedures are given in **Figure 4.1**.

The data extraction process using the WebPlotDigitizer tool mainly consists of two steps. The first step is to manually label the axis of the chosen figure as demonstrated in the figure. The data points can then be selected (indicated as the red dots in step 2) and output with their respective x and y information, which corresponds to the discharge capacity and the number of the cycle in this case. In addition, to assess the extent of potential human error exposed during the selection of the data point on the figure, a test was performed to compare the results of the manually collected capacity with the reported capacity in the journal. A deviation of less than 0.03 $mAhg^{-1}$ is obtained of which is trivial to the measured capacity (150 $mAhg^{-1}$) and therefore proving this method being a reasonable estimation for the real discharge performance.

4.5 Feature Analysis and Data Pre-processing

After compiling clean datasets, the next step is to gain some insights into the dataset and processing these data to support the model training.

4.5.1 Pearson Correlation Coefficient Study

Pearson Correlation Coefficient is used to estimate the linear correlation of the two sets of data for two variables and is often treated as the preliminary statistical analysis approach. **Equation 4.1** demonstrates the formula for calculating the Pearson Correlation Coefficient.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.1}$$

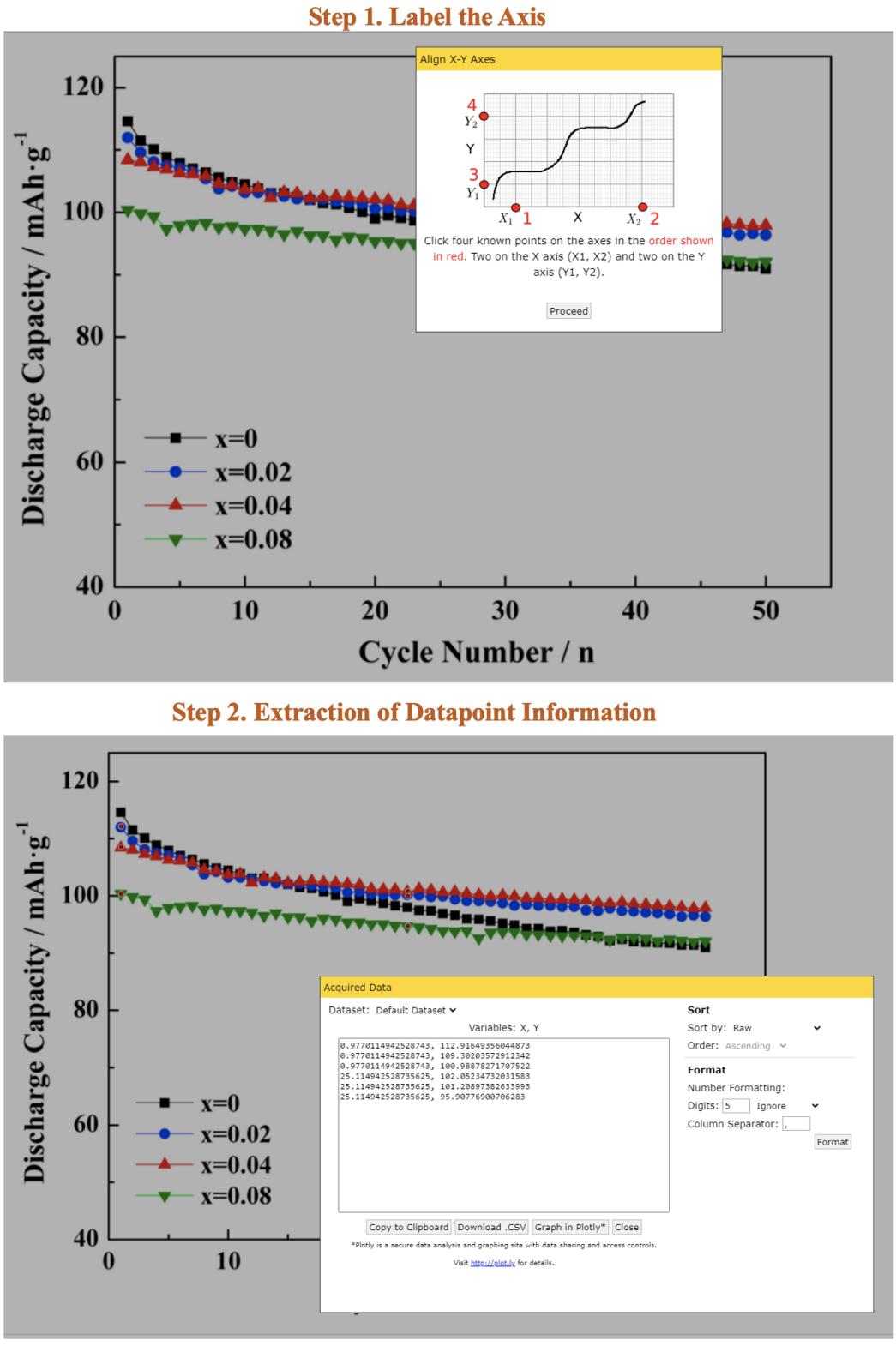


Figure 4.1: A process of using WebPlotDigiztier online tool to extract the data information from the cyclic discharge capacity graph with step one being the labelling of the axis and followed by the second step of selecting the data point of interest. The example discharge capacity figure was taken with permission from [19]

where r_{xy} is the Pearson correlation coefficient, n is the size of the sample, x_i and y_i are the sample points for x and y variable at sample index i, $\bar{x}$ and $\bar{y}$ are the sample mean values for variable x and y.

The values of the Pearson correlation coefficient are always found in the range of -1 to 1, with the value closer to -1 or 1 indicating a more positive or negative linear relationship for the two variables.

4.5.2 Centering and Scaling

Centering-and-scaling is a common data preprocessing technique used to transform the numerical raw data. The data centering is mainly done by subtracting each data point by the mean value of the data group and the new mean value will become 0 in the new data group. Furthermore, the scaling treatment is done by multiplying each data point by a constant to alter the range. Mathematical formula of this is given in **Equation 4.2**

$$x_{\text{standardised}} = \frac{x - \mu}{\sigma} \tag{4.2}$$

Where $x_{\text{standardised}}$ is the standardised data point, x is the original data point in the sample set, μ is the mean value of the sample set, σ is the standard deviation of the sample set. This data treatment is often needed for models such as the support vector regression and K-nearest neighbour because they perform better when all of the variables are on the same scale. This is very useful to transform the variables that have different units into the standardised unit-less data to achieve easy-comparisons. The additional benefit of this treatment is also seen from reducing the computational cost for training the models.

4.6 Model Training and Hyperparameter Tuning

4.6.1 Data Splitting and Model Training

Data splitting is a step to determine how much of the data would be used to train the model and test. In the following projects, a ratio of 8:2 is selected to randomly split the whole dataspace correspond to the model training set and the test set. A model

is trained and constructed based on the training dataset and its prediction power is further examined by predicting against the hold-out test set.

Next, the process of fitting machine learning models to a dataset is equivalent to the purpose of finding the optimal function coefficients or the best splitting rules (for tree-based algorithms) that can minimise a loss function. To begin with, a loss function is a pre-defined error function that measures the model performance by calculating the deviations between the model prediction value and the observed response value. This function is normally used as a guide to the development of models. Building linear or non-linear models requires the estimation of coefficients that minimise the prediction error. In this project, some of the algorithms will use the multiple ordinary least square (OLS) algorithm to estimate the optimal coefficients for each of the covariates that would minimise the loss function. The mathematical formula of OLS is given as **Equation 4.3**.

$$\hat{\beta} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y} \tag{4.3}$$

Where X is a n$\times$p matrix of vectors of p covariates and n numbers of data rows, whose i^{th} row is the vector of covariates for the i^{th} case, $\hat{\beta}$ is the optimal coefficient that minimises the given loss function, T means the transpose matrix and y is the vector of the response variable. The discussions of the loss function used in each ML algorithm are further made in the following sections.

OLS is a non-iterative process that fits the model once when the loss value is minimised and this technique is needed for computing the non-linear models as well as for some penalised models. On the other hand, there are also other iterative optimisation methods such as gradient descent and stochastic gradient descent/ascent. These methods update the weights multiply times based on the loss function and normally require more computing power. The process of these iterative weight-updating algorithms involves the initial inputting of small coefficient values (normally close to 0) into the function, calculating the respective cost from the loss function, computing the derivative of the cost to know in which direction the coefficients should be updated to lower the cost. In the end, the algorithm will

update the coefficient accordingly with a specified learning rate to seek further for the global minima point (lowest error). The learning rate indicates how much the weights will be updated during the optimisation process and should be tuned wisely for each study.

Next, the K-fold cross-validation technique is used to train all the machine learning models. This resampling technique can effectively reduce the likelihood of model overfitting on a given dataset and resulting in models that do not have good prediction power over a new dataset. The technique randomly splits the whole training data space into K small and approximately equal-sized subsets where the K-1 subsets are used for model training and the final subset of data is used for validation. The validation subset is not fixed and should be switched in an alternating manner for every fold iteration. A graphical illustration of this process is given in **Figure 4.2**. The prediction error for each fold training is recorded and additionally, the average value of all folds is calculated to determine the model training performance.

4.6.2 Model Performance Metrics

In this thesis, the regression model performance metrics are used since the response variable, the discharge capacities, have continuous values. This involves the calculation of the mean square error (MSE) or the root mean square error (RMSE) and the coefficient of determination (R^2). The MSE value refers to the square deviations between the observed value and the predicted value and additionally, RMSE is the square root format of the value of MSE. The mathematical formulas of the RMSE error metric are given in **Equation 4.4**.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.4}$$

Where n is the sample size, y_i is the observed variable, $\hat{y}_i$ is the predicted value and X_i is the series of covariate variables. The value of RMSE represents how much does the prediction deviate from the observed values.

During the model training process, MSE values are mostly used to indicate the

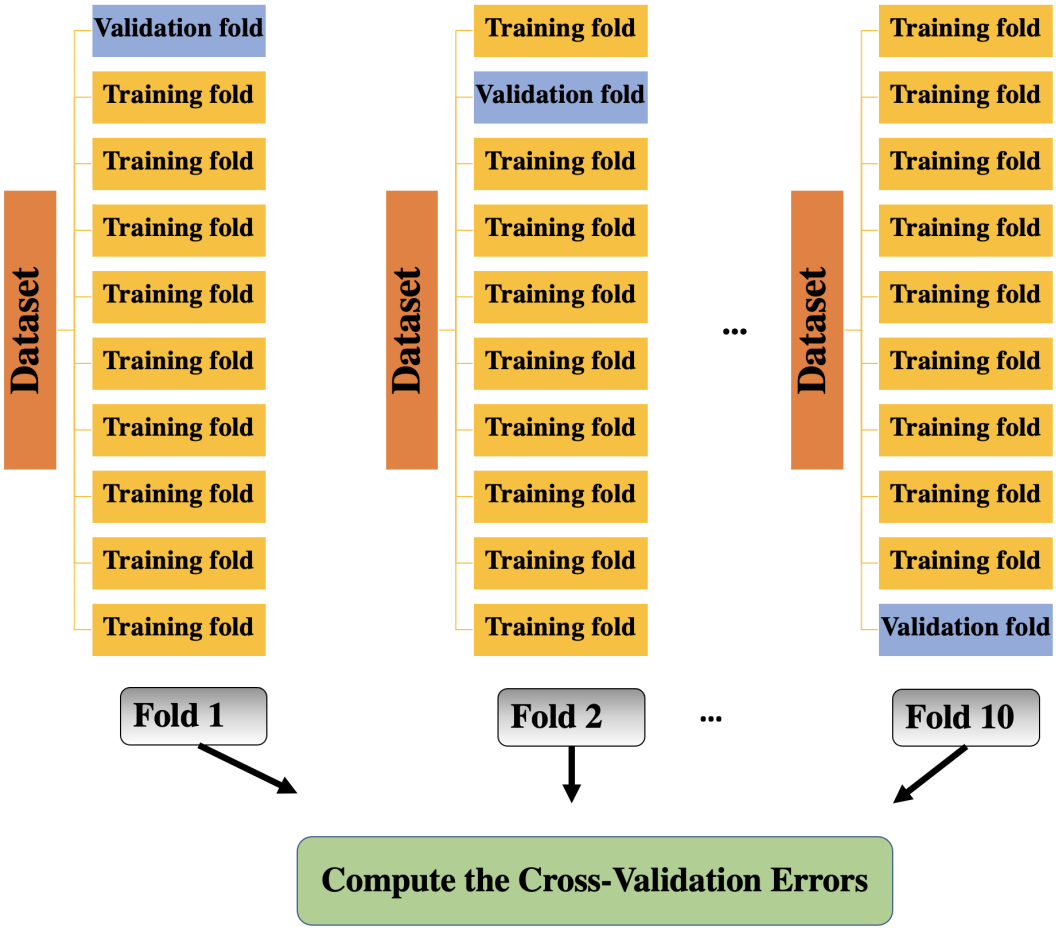


Figure 4.2: A schematic illustration of the ten-fold cross validation for model training

model's prediction performance. This term can be decomposed into three major components as given in **Equation 4.5**.

$$\text{MSE} = \sigma^2 + \text{Bias}^2 + \text{Variance} \tag{4.5}$$

The σ^2 term from **Equation 4.5** represents the irreducible error that naturally occurs from the model and cannot be eliminated. The squared model bias term refers to how close the model prediction is to the real observation. For any given prediction model of $\hat{f}(x)$ estimating $f(x)$, the bias term can be calculated from the equation of $E[\hat{f}(x) - f(x)]$, where $E[.]$ stands for the expected value such as the mean value. The model Variance on the other hand measure how sensitive the model is towards the small change in the training data and can be calculated using

the equation of $E[(\hat{f}(x)^2] - E[\hat{f}(x)]^2$. During the model tuning process, it is very important to control the complexity of the model so that bias and variance errors are both kept at the minimum. Bias and variance errors have the reverse growth pattern with each other, an decrease in the bias error resulting from a more complex model can increase the variance error. The scenarios of an unbalanced trade-off between bias and variance can lead to the results of over-fitting (high variance and low bias) or under-fitting (low variance and high bias). The over-fitting issue occurs when the model is built in a very complex manner that it overly learnt the pattern for the given dataset, captured the data noise and cannot predict well for a new dataset. The under-fitting issue arises when the model is not complex enough and cannot capture the underlying data correlation well.

Next, the coefficient of determination (R^2) is used to determine the proportion of variance of the response variable being accurately described by the selected model and the mathematical representation of it is given in **Equation 4.6**.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.6}$$

where R^2 is the coefficient of determination and its value can be either positive or negative depending on the model performance, y_i is the observed data point of the response variable at "i" indexed, $\hat{y}_i$ is the predicted data point of the response variable at "i" index, $\bar{y}$ is the mean value of the observed values of the response variable.

4.6.3 Model Refinement and Hyperparameter Tuning

The prediction performance of any model is largely controlled by model architecture and the optimisation technique. The hyperparameters in the model are the fundamental features to construct a model and the respective values should be tuned to attain the best possible prediction power for the investigating algorithm. The sets of hyperparameters for investigation are often generated via the grid search or the randomised search techniques. The grid search technique generates a list of hyperparameters incrementally for a set value range. For instance, the investigat-

ing cost factor in a support vector machine model with a target range of 2^{-2} to 2^{-5} would be 2^{-2}, 2^{-3}, 2^{-4} and 2^{-5}. On the other hand, the randomised search technique generates random values from the given value range and are used to test the model performance. These two techniques are often coupled with the cross-validation method to construct a model training plan. The hyperparameter tuning or optimisation is normally carried with the initial proposal of a set of hyperparameters generated from these search techniques. Following by, different predictive models are built based on these architectural hyperparameters and carried out the model training and validating process. The average RMSE value of all ten-fold validations are calculated for every proposed hyperparameter set and conclude the optimal set that gives the lowest RMSE one. Different algorithms will have different sets of hyperparameters that require the users to optimise for and this is largely depending on the computing power available and the targeted model precision. In the following results chapter, the selected hyperparameters and the value range of investigations for each study performed are given in the respective experimental section. The investigating hyperparameter ranges are considered from the existing literature. This includes the artificial neural network [231, 255], support vector machine and kernel ridge regression [93, 232, 256], and tree-based algorithms such as random-forest and gradient boosting machine [255, 257]. Both the grid search technique and the randomised search technique are applied during the hyperparameter optimisation process.

4.6.4 Packages Information

Various machine learning packages are available online for constructing models and implementing optimisation techniques. The common ML packages can be mainly found in two coding languages, R and Python. In the following results chapter, various packages coded in R and Python were used to train, validate the model and test against the hold-out test set. The details of these packages are indicated in the experimental section of each results' chapter.

4.7 Linear Regression Algorithms

4.7.1 General Linear Algorithm

Linear models are commonly used in the early stages of interrogating data. As the name suggests, a linear model is the sum of linear correlations between the input covariate variables X_j, j=1, ..., P and the response variable y_i, i= 1, ..., n, like shown in **Equation 4.7**.

$$y_i = \beta_0 + \sum_{i=1}^{n} \sum_{j=1}^{P} x_{ij}\beta_j + \varepsilon_i \tag{4.7}$$

where β_0 is the intercept, ε_i is the error term that can not be explained by the model and are assumed to be distributed between the mean-zero and the variance ($\varepsilon_i \sim N(0, \sigma^2)$), y_i is the response variable and is normally distributed between $y_i \sim N(X_{ij}\beta_j, \sigma^2)$, β_j is estimated coefficients for the j$^{\text{th}}$ covariate variable and $x_i j$ is the value of the j$^{\text{th}}$ covariate for the i^{th} sample data.

The coefficients in the above general linear model aim to reduce the value of the sum of square error (SSE) and the mathematical formula of this is indicated in **Equation 4.8** .

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.8}$$

Where n is the total number of training points, $\hat{y}_i$ is the model prediction output and y_i is the observed response value. In this thesis, the OLS optimisation algorithm is used to achieve find the optimal coefficients of a linear model that minimise the bias values.

4.7.2 Penalised Regression Algorithms

The parameter coefficients computed through the OLS algorithm are thought to be unbiased as it takes equal consideration for all of the available covariate variables. As a result of this unbiased estimation, the respective variance error in **Equation 4.5** is kept very high as the small fluctuations in the data could lead to a big change to the model structure. On the other, it is commonly known that a small amount of

increase in the bias would produce a significant drop in the variance and hence leading to a much lower overall MSE value from **Equation 4.5** than the ordinary OLS algorithm. One method to increase the bias to the model is to include penalisation terms to the SSE loss function and find the estimates of the coefficients that lower the variance. These models with modified objective functions are known as the penalised regression models.

Penalised regression models include an extra regularisation term targeted to reduce the consideration of less effective variables in the model. There are two commonly used penalised models named after Ridge(L2) [258] and Lasso (Least Absolute Shrinkage and Selection Operator, L1) [259] and their major difference is seen from the add-on regularisation terms, as indicated in **Equation 4.9** and **Equation 4.10**, respectively:

$$\text{Ridge} : \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \tag{4.9}$$

$$\text{LASSO} : \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \tag{4.10}$$

where λ is the weight of the penalty term, β_j^2 and $|\beta_j|$ are the penalty terms for ridge and Lasso regularisation, respectively. The ridge regularisation term $\lambda \sum_{j=1}^{P} \beta_j^2$ controls the contribution of variables to the final prediction by altering their corresponding λ values. For the covariates that are less effective in reducing the objective error function (**Equation 4.9**), the corresponding coefficient will be reduced to closer to 0. Lasso regression is a relatively new and popular alternative to ridge regression, using the absolute value rather than the square penalisation term ($\lambda \sum_{j=1}^{P} \beta_j^2$). A direct consequence of this penalising method is that parameters can be penalised to 0 if they have not shown any improvements in the overall prediction performance. When penalising two closely correlated covariate variables, instead of shrinking both their coefficients toward each other, Lasso regression will selectively pick one and ignore the other. This regularisation method is more advantageous than the previous ridge method in that it reduces the variance of models that contain many unrelated covariates. While the modifications of the regulation terms are compara-

tively smaller from the ridge regression, the practical implications on the other hand are known to be growing very rapidly in various fields.

In conclusion, the penalised regression models are popular to apply for small training datasets as they will provide a more accurate prediction on the test-set by lowering the model variance than the OLS method does.

4.7.3 Bayesian Ridge Regression

The general linear regression technique uses the OLS algorithm to make a single estimate on the best model parameters (β) based on the given training data and this method is susceptible to high variance issues. As an alternative, the Bayesian ridge regression technique assumes that the response variable values (y) are sampled from a probability distribution type and it can perform the predictions on the new observation from this data distribution. An example of a Gaussian distributed response y over the coefficients β^T and covariate variables (X) is given in **Equation 4.11**.

$$y \sim N(\beta^T X, \sigma^2) \tag{4.11}$$

Where σ denotes variance, y is the response variable, N means the distribution. Next, the essential aim of a Bayesian ridge regression algorithm is to maximise the probability of finding the set of coefficient (β) that would best fit with the given response variable like shown in (**Equation 4.12**).

$$\hat{\beta} = \max_{\beta} \frac{P(y|\beta)P(\beta)}{P(y)} \tag{4.12}$$

Where P(y|β) is referred to as the likelihood term, demonstrating the probably of y having generated β, P(β) is normally referred to as being the prior term and this means the probability of observing the settings of β coefficients unconditionally on anything before observing the data, P(y) is the probability of observing y and it has nothing to do with the β optimisation and therefore could be neglected. The numerator of the **Equation 4.12** is normally referred to as being the posterior of the function and should be maximised in this case. Furthermore, the estimation of the suitable β values should therefore involve the estimation of both the likelihood term

and the prior terms indicated in the equation above. In this project, the estimations of both terms are done through the use of uninformative priors which do not hold any assumptions about the data (normally have large standard deviations) and are purely depending on the data.

The estimation of the likelihood term can be made with the assumption that the response variable y is distributed under the Gaussian distribution function and this can be expressed in **Equation 4.13**.

$$P(y|X,\beta,\alpha) = N(y|\beta X,\alpha) \tag{4.13}$$

Where X is the matrix of covariate variable,α is the precision of the noise.

In addition, the prior coefficients term are assumed to be sampled in the Gaussian distribution like shown in **Equation 4.14**.

$$P(\beta|\lambda) = N(\beta|0,\lambda^{-1}I_p) \tag{4.14}$$

Where λ^{-1} is the precision of the weights from the Gaussian distribution, I_p is the identity matrix of p dimensions.

Both the α and the λ parameters are selected to have gamma distribution as the priors and these parameters along with the coefficients are estimated through maximising the logarithmic values of the posterior terms like given in **Equation 4.15** and this is because of the easier computation.

$$\hat{\beta} = \max_{\beta}[\text{logP}(\text{y}|\text{X},\beta,\alpha) + \text{logP}(\beta|\lambda)] \tag{4.15}$$

Furthermore, other Bayesian ridge regression algorithm features are set to be the number of iterations equal to 300 during the estimation of the parameters and the stopping criteria is chosen to be 10^{-3} if the coefficients converged to this value.

In conclusion, the main advantage of applying the Bayesian ridge regression technique is seen from that it can quickly adapt to the given dataset whilst providing a prediction with less variance error than the models that are estimated through the OLS algorithms. On the other hand, since the prediction for the new observations is

based on the inference of the established distribution, this could be very time-wise inefficient.

4.8 Non-linear Regression Algorithms

4.8.1 Artificial Neural Network

Artificial Neural Networks (ANN) is a nature-inspired modelling technique that mimics the structure and the functionality of an animal brain. Following several decades of research, it is noticed that most of the relationships found in nature are formed in a rather complex and nonlinear manner and this prevent the use of conventional prediction approaches such as the human instinct. As an alternative, the characteristics of non-linearity and high flexibility have enabled artificial neural network modelling to predict through these high-dimensional data and effectively show the embedded trends. The construction of artificial neural network [260] involves three important elements, architecture, activation rules and output rule, and the discussion for each is given in the following section.

4.8.1.1 Architecture

The architecture of a typical neural network found in the human brain mainly consists of three important features, neurons, dendrites and axons. At first, a vast amount of signals are received from the dendrite and transferred down to the next neuron unit through the axon bridge. Similarly, an ANN is constructed with several neurons nodes which can be arranged in several ways. An illustration of the structure of a feed-forward neural network with a single hidden layer is given in **Figure 4.3**.

The initial set of nodes with no incoming edge are referred to as the input layer and this is where the information of covariates is fed into ($X_j : j = 1, ...P$). The set of nodes that does not have any outgoing edges is the output layer and this is where the predicted responsive variables are extracted.

The layers of nodes ($Z_m : m = 1, ...M$) that are located in-between these layers are the hidden layers and it contains a different number of hidden neurons that pass forward the information from one layer to the next. The number of the hidden

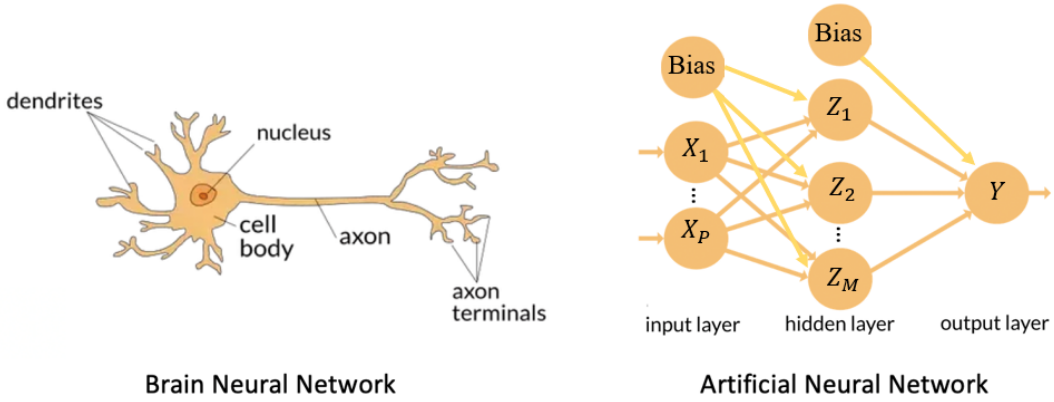


Figure 4.3: Architectures of neural network in human brain (left). Image taken with permission from [20]. The structure of artificial neural network (Right). Image taken with permission from [21].

layers and the number of neurons affect the number of the coefficients that the model need to compute and hence would influence the overall performance. These hyperparameters are optimised carefully in each of the study chapters.

4.8.1.2 Activation Rules and Output Rules

In a feedforward neural network model, the neurons pass the information forwardly from one layer to the next layer. The outputs that are passing down from the preceding layer neurons normally undergo linear aggregation to obtain the intermediate quantity s_υ (**Equation 4.16**). This intermediate quantity output is fed into an activation function first before passing it down to the neuron in the next layer. This transformation process is shown in **Equation 4.17**.

$$s_\upsilon = \sum_{u \to \upsilon} \beta_{u\upsilon}^T x_u + \beta_{0\upsilon} \tag{4.16}$$

$$x_\upsilon = \sigma(s_\upsilon) \tag{4.17}$$

where s_υ is the intermediate quantity, u,υ are the nodes in the neural network and the symbol $u \to \upsilon$ in a feed-forward structure indicates that these nodes are in the successive layers, $\beta_{u\upsilon}$ are the weights for the connection of the preceding node and the successive node, $\beta_{0\upsilon}$ is the bias term for the hidden layer where υ node is

present and this term is used to capture the intercepts in the model and often set to be 1. In addition, the activation function is represented by "$\sigma(\cdot)$".

In the case of a single-layered neural network (left in **Figure 4.3**), the derived features in the first hidden layer is represented as Z_m and can be represented by **Equation 4.18**.

$$Z_m = \sigma(\beta_{0m} + \sum_{j=1}^{P} \beta_{jm}^T X_j) \tag{4.18}$$

The activation function should be chosen to fit the type of prediction task. The sigmoid function, $\sigma(y) = \frac{e^y}{1+e^y}$, is the first activation function used because of its biological interpretation. In addition, the rectifier linear units (ReLU) function with the form of $\sigma(y) = max(0, y)$, has gained great popularity in both the classification and the regression problems because of its simple function and promising performance. In the ReLU activation function, it returns 0 for all the negative inputting y values whilst outputting the same values for all the positive ones. For this thesis, the ReLU function was used as the main activation function during the construction of ANN models.

Furthermore, after the information is processed and transferred to the last hidden layer, a linear output rule is applied instead of the activation function to linearly aggregate all the signals collectively and convert them into a single vector as the model output. The mathematical illustration of this output process for the one-layer neural network example is illustrated in **Equation 4.19**.

$$f(X) = \sum_{m=1}^{M} \gamma_m^T Z_m + \gamma_{mo} \tag{4.19}$$

Where M is the total number of hidden nodes in the previous layer, γ_m is the weights of hidden nodes m and γ_{mo} is the bias term. This linear output is implemented in the following studies.

4.8.1.3 Model Training

The training of the ANN involves the estimation of the optimal neuron weights and bias that could reach the measure of fit criteria. For regression, the sum of square

residues is normally used as the error function to assist these estimations and this is given in **Equation 4.8**.

In the presence of many neurons and bias terms, the weight optimisation process can be very difficult to perform. In the beginning, random values assigned to these weights first and followed by the use of a gradient descent algorithm to search for the optimal values. Due to the composition form of the model, the weights/gradient of the node can be computed through the use of a chain rule for differentiation. This generic approach is commonly referred to as the back-propagation which can sweep forward and backwards over the model and update the weights to each node unit with the specific learning rate. The learning rate is commonly found in the range between 0 and 1 and this value should be tuned carefully to avoid the model from quickly covering the sub-optimal minima instead of the global minima. Nevertheless, using gradient descent as the optimisation algorithm seem to suffer from the slow converging speed particularly for the calculation of the gradients for the large data set. There are other alternative optimisation algorithms available in the literature such as the Mini-Batch Gradient Descent, Momentum, Adagrad and Adaptive Momentum Estimation (Adam) [261]. Among all, the Adam optimiser has demonstrated great convergence speed with promising performance on computing sparse gradients for the noisy dataset. In this thesis, the Adam algorithm is used instead of the gradient descent to train ANN models and the optimisation of the learning rate would be carried out in the logarithmic scale (10^{-3}, 10^{-2}, 10^{-1}).

Next, batch size and epochs are two other parameters to consider. Epoch is the number of times that the dataset will be fed into and pass backwards from the model. Due to the iterative nature of updating the weights in gradient descent, it is normally required to pass through the dataset multiple times to reach the optimal weights. The number of epochs is varied from the different datasets and should be determined case by case. More epochs are generally required for the setting of a smaller learning rate to give smaller updates in the weights. The batch size is the indication of how many small subsets should the whole dataset be split into and used for weights updating. The number of epochs used would be determined by

running a trial test with up to 10000 epochs and the number of the batch is fixed at 10 as this is the most divisible amount by the total sample size.

4.8.2 K-Nearest Neighbour

The K-nearest Neighbour (KNN) algorithm, as the name suggests, predict the new sample based on its K nearest neighbouring samples. In the task of predicting a new regression sample, the model firstly identifies the closest KNN to this sample and outputs the mean of these KNNs as the model response. To identify how close the two measuring points are to each other, a distance metric should be given to guide the search of the nearest neighbours. The Minkowski distance equation and are used as a general formula can be used for this calculation purpose and the mathematical formula of this is given in **Equation 4.20**:

$$(\sum_{j=1}^{P} |x_a - x_b|^q)^{\frac{1}{q}} \tag{4.20}$$

Where x_a and x_b are the two different sample points. The q value should be greater than 0 and the change of this value can give different distance metrics. When q equals 1, the Minkowski equation becomes the Manhattan distance metric and it is normally used for the samples with binary covariates. When q equals 2, the Minkowski equation becomes the commonly seen Euclidean distance of which the distance output is always non-negative for all $q > 0$. In this thesis, the Euclidean distance is used as the distance metric during the training of KNN models. Since the prediction is inherently relying on the distance of the KNN samples, the larger-scale samples would have more impact than the lower scale samples and hence introduce bias to the models. To eliminate this unbiased contribution, covariate variables often required to be centred and scaled before training the KNN model. This data preprocessing method has been discussed and its mathematical formula is given in **Equation 4.2**. Furthermore, the number of neighbours, K, is often treated as the tuning parameter for this model and the optimal value is often selected through the resampling technique such as cross-validation.

In conclusion, KNN is favourable for its easy implementation and good pre-

dictive power. Nevertheless, high computation time is often reported as the major disadvantage because the model requires the loading of the training set and then perform the calculation of the distance between the new sample and the existing training samples to make a prediction.

4.8.3 Support Vector Machine

The theory of support vector machine (SVM) was firstly developed for solving classifications problems by defining the hyperplane that can separate the data into classes. In addition, *Smola and Drucker et al* [262] proposed an SVM regression framework for predicting quantitative response through adapting a few properties of the SVM classifier. To begin with, the linear regression approach uses the SSE error function as a benchmark to estimate the parameters indicated in **Equation 4.7**. On the other hand, an SVM algorithm uses a ε-insensitive loss function to minimise the effects of outliner (observations out of the general trend) during the fitting of the regression function line. The estimation of the coefficient of an SVM regressor requires the minimisation of the objective loss function as shown in **Equation 4.21**.

$$C\sum_{i=1}^{n} L_{\varepsilon}(y_i - \hat{y}_i) + \lambda \sum_{j=1}^{P} \beta_j^2 \tag{4.21}$$

Where C is the cost parameter, β_j^2 term is the ridge penalisation term and the L_{ε} is the ε-insensitive error function and this can be represented in **Equation 4.22**

$$L_{\varepsilon} = \begin{cases} 0 & \text{if } |r| < \varepsilon \\ |r| - \varepsilon & \text{otherwise} \end{cases} \tag{4.22}$$

Where r is the residues and ε is the value that defines the boundary. The underlying principles of these equations are such that the data points with smaller residual (difference between the predicted value and the observed value) than the ε value are ignored (returned as 0) and not used to fit the regression equation. On the other hand, the data points with greater residual values than ε will contribute to the regressor equation fit in a linear scale. A diagrammatic illustration of the ε-insensitive error function is also given in **Figure 4.4**.

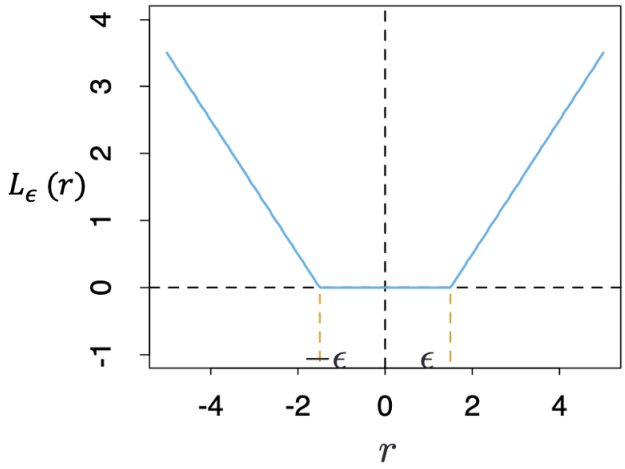


Figure 4.4: The ε-insensitive error function applied by the support vector regressor. Image taken with permission from [22].

In addition, it is important to note that the large residual can be further penalised by the C value and this is treated as a hyperparameter to be tuned for in the model. The SVM regression prediction function is similar to the linear regression prediction function, however, is different for involving some unknown parameters noted as the α_i for each of the datapoints in the training set. The mathematical representation of a SVM regression prediction function is given as **Equation 4.23**:

$$\begin{aligned} \hat{y} &= \beta_0 + \beta_1 u_1 + \ldots + \beta_P u_P \\ &= \beta_0 + \sum_{j=1}^{P} \beta_j u_j \\ &= \beta_0 + \sum_{j=1}^{P} \sum_{i=1}^{n} \alpha_i x_{ij} u_j \\ &= \beta_0 + \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{P} x_{ij} u_j \end{aligned} \tag{4.23}$$

The number of the α_i parameters should be the same as the number of data points in the training set and this could be considered a highly complex over-parameterised model. However, due to the ε-insensitive penalty rules, the perfectly fitted observations points (small prediction residue) are not used to support the com-

putation of the SVM regression hyperplane (their parameters set to be 0). On the other hand, a small subset of data points where the prediction residue is large, have their α_i set not to be 0, are used for the model building and these are referred to as the support vectors for fitting the regression line.

Next, the term $\sum_{j=1}^{P} x_{ij}u_j$ in last part of the **Equation 4.23** is the sum of the dot product of the new sample values and this can be simplified into **Equation 4.24**:

$$f(u) = \beta_0 + \sum_{i-1}^{n} \alpha_i K(x_i, u) \tag{4.24}$$

$K(.)$ is the kernel function used to map the data points from the original data space into a higher dimensional Hilbert space with a much lower computational cost. In this case, the kernel function can be expressed as **Equation 4.25**.

$$K(x_i, u) = \sum_{j=1}^{P} x_{ij}u_i = < \phi(x_i), \phi(u) > \tag{4.25}$$

where K is the kernel function, x_i and u are the data-points from the original data-space, $\phi(.)$ is the mapping function from the original dimension space to a new higher dimensional space, $<,>$ denotes for the dot product of the vectors. There are many forms of kernel functions available for the application of SVM regression. **Equation 4.26** gives three examples of the most popular non-linear kernel functions to create a non-linear SVM regression model and this is referred to as the "kernel" trick.

$$\begin{aligned} \text{d}^{\text{th}}\,\text{degree polynomial} &: K(x_i, u) = (1 + < x_i, u >)^d \\ \text{radial basis function} &: exp(-\gamma||x_i - u||^2) \\ \text{laplacian function} &: exp(-\sigma||x_i - u||) \\ \text{hyperbolic tangent} &: tanh(\kappa_1 < x_i, u > + \kappa_2) \end{aligned} \tag{4.26}$$

Where d, σ and κ_i are the scaling parameters that need to be tuned during the model optimisation stage. In this thesis, the radial basis kernel function is used due to its non-linear feature and the widely acknowledged effectiveness in achieving accurate prediction. Before the SVM model training, centring-and-scaling techniques need to apply to convert all predictor values into the same scale because a large

difference in the predictor's value scale would mislead the estimation of the optimal prediction function.

4.8.4 Kernel Ridge Regression

Kernel ridge regression (KRR) is a similar technique in that it also combines the kernel trick with ridge linear regression (L2 regularisation) to decode the data correlations. The main difference between the support vector regression and the coefficients in a kernel ridge regression function minimise a loss function of the square error loss as opposed to the ε-insensitive loss function indicated in **Equation 4.21**. The mathematical formula for the loss function used in KRR is given as **Equation 4.27**.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \tag{4.27}$$

As the result of the replacement of the ε-insensitive function, KRR eliminates the process of support vector selection from the SVM and uses the whole data-set to do the model training. This seems to provide better training speed when the dataset is small. On the other hand, their prediction performance seems to deteriorate as the size of the data-set increases because it needs to sum up the vectors over the entire dataset than the selected support vectors.

4.9 Tree-based Algorithms

4.9.1 Decision Tree

Decision tree model [263] partitions the whole data space into various subsets and predict for each a response value (mostly a constant) and combine these separate responses to form a complete prediction. The diagrammatic illustration of this partition process is given in **Figure 4.5**.

The entire data space is partitioned into two regions with the rule $X_1 \leq t_1$ applied to the variable X_1, assuming that X_1 is the first optimal variable to split on. Next, a recursive splitting is carried out for the second optimal splitting variable X_2 according to the rule of $X_2 \leq t_2$, and the remaining data space is further split into

region R_2 and R_3. This partition process is also given in the tree evolution format at the right side of **Figure 4.5**, where the top is referred to as the root and the end nodes are referred to as the terminal nodes or the leaves.

During the process of growing a recursive regression tree, the algorithm will have to make decisions on what covariate variable should be split, at what value the split should happen. Suppose a decision tree model is built for a data space that consists of n sets of P predictors ($\chi_i^j : i = 1,....n, j = 1,....P$) and n sets of a single response variable $y_{1,...n}$, the first splitting mechanism will be occur according to the computed decision on the spitting covariate variable v and split point s. This splitting mechanism is driven by the search for the optimal splitting covariate variable v and splitting point s that would reduce the overall residual sum of square, summarised in **Equation 4.28**.

$$min_{v,s}\left[min_{c_1}\sum_{x_i\in R_1(v,s)}(y_i-c_1)^2+min_{c_2}\sum_{x_i\in R_2(v,s)}(y_i-c_2)^2\right] \tag{4.28}$$

Where R_1 and R_2 represent two separate data regions given in **Equation 4.29**.

$$R_1(v,s)=(X|X_v\leq s) \qquad R_2(v,s)=(X|X_v>s) \tag{4.29}$$

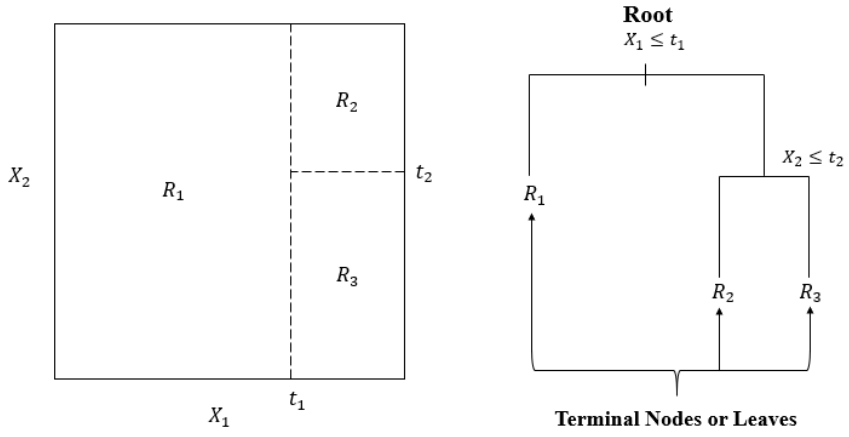


Figure 4.5: Left shows the partition of the data space by recursive binary splitting. Right shows the tree generated from the partition rules. Images taken with permission from [22].

Each region is then represented with the average constant $\hat{c}_n$ of all the response variable from the same region R, given in **Equation 4.30**.

$$\hat{c}_n = \text{average}(y_i | x_i \in R) \tag{4.30}$$

This optimisation will be carried out through all values of every given covariate variable from the dataset at each splitting stage, until stopping rules are instructed by the model builder. For instance, the restriction on the size of the tree or the number of terminal nodes. These parameters can be either predefined by the user or optimised with the cross-validation method. Despite the aforementioned advantages, there are also some shortcomings associated with the embedded building strategy. The rules of splitting are purposely built based on the given dataset, hence any modifications of the training data will directly influence the approximation of the splitting rule at each stage, leading to a big change to the original tree structure. In addition, since the whole dataset is exposed in the training stage, there is a high chance that the model would learn the internal noise and raising the concerns of over-fitting. These issues can be effectively solved through the introduction of randomness to the underlying splitting mechanisms. There are in total two ways of achieving it, one is the implementation of bootstrap aggregation or also known as the bagging method, which is introduced in **subsection 4.9.2**; the other method is known as the gradient boosting, introduced in **subsection 4.9.3**.

4.9.2 Bagged Tree and Random Forest

The bagged tree method or the bootstrap aggregation [264], is an ensemble technique that leverages the power of data aggregation and regression trees to effectively reduce the prediction variance. First, the bootstrap sampling technique is used to randomly draw samples from the original data set with replacement. This means that the selected data point can be drawn again in the future data selection. After generating several bootstraps of samples, the bagged tree model will generate several individual decision trees to the pre-defined maximum depth based on the bootstraps and the final prediction will be taken by averaging every individual pre-

diction result and this is treated as the model output. Bagged trees follow the same splitting pattern as the decision tree where all original covariate variables are exposed to every splitting stage. Nevertheless, this similarity can lead to the creation of very alike (same splitting pattern) and highly inter-correlated trees and increase the overall variance error as a result.

As an alternative approach, random forest [265] introduces the randomness to the construction process of trees by restricting to a subset of all the covariate variables, that are contributed to each tree split. In the end, the outputs of the predictions of all trees are averaged to give the final model output. The detailed growing process of this algorithm is given in **Figure 4.6**.

The major difference between a bagged tree model and a random forest model is seen from their underlying tree construction strategies. In the case of bagged trees, each split is considered across all given covariate variables choices. Random forest on the other hand restricts the number of considered covariate variables at each split stage, represented as the symbol "k" in **Figure 4.6**, and this can be tuned easily with the rule of minimizing the RMSE value from **Equation 4.4**. Since the covariates variables from which the trees are built are randomly selected by the

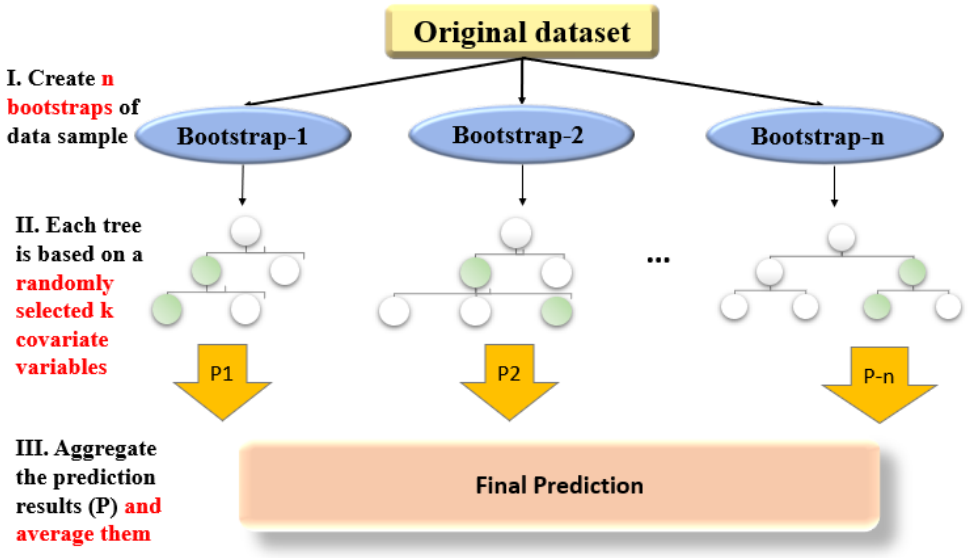


Figure 4.6: A diagrammatic illustration of the growing processes for a random forest model.

algorithm, this largely reduces the interdependence between the trees and reaches a better prediction performance.

4.9.3 Gradient Boosting Machine

Gradient boosting is a modelling technique introduced by *Friedman et al* [266] in 2000, The basic principle is to build additive models on the prediction residuals generated from the weak predictors and in addition to give a better prediction power overall. Weak predictors or learners refer to any models that have a wide selection of tuning parameters and therefore can be made to have weak prediction power. For instance, a decision tree is a classic weak learner where any change in the tree maximum depth would influence its computational time for training, predicting and the final prediction accuracy. In this case, the regression decision tree is thought to be one of the best base boosting learners for numerous reasons. First, the generated trees can be flexibly added together and aggregated into an overall prediction, which fits the gradient boosting the concept of generating additive models. Furthermore, trees can be made into weak predictors by adjusting their parameters such as the tree depth and the number of terminal nodes and hence leaving a big room for adapting the gradient boosting method.

The process of growing a gradient boosting tree model is very much different to that of the random forest model and a graphical illustration of it is given in **Figure 4.7**. Initially, a single leave is generated by taking the average of all of the values of the response from the dataset ($\bar{y}$). The respective residue or gradient can then be computed with the general formula indicated in **Equation 4.31**.

$$r_{mi} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{(m-1)}} \tag{4.31}$$

where r_{mi} is the residue for sample number i at m tree, y_i is the observed response value of sample i, $f(x_i)$ is the predicted value using x_i and $\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}$ is the partial differential of the loss function $L(y_i, f(x_i))$ with respect to the predicted value $f(x_i)$. The commonly used loss function for this gradient boosting model is similar to the SSE and has the form of $\frac{1}{2}(y_i - f(x_i))^2$, where the partial derivative (gradient)

of this function is simply the residue value ($y_i - f(x_i)$). Before the construction of the first tree, the residue can be expressed as the difference between the observed value and the average observed value ($\bar{y}$).

Next, regression trees of a pre-defined depth, D, are built for the computed set of residues ($r_{im} : i = 1,...n, m = 1,...M$) and this produces various terminal leaves labelled as R_{jm}, where j indicates for the leaf number within the range of $j = 1,...J_m$. The final output for each of these terminal nodes, γ_{jm}, can be computed through **Equation 4.32**.

$$\gamma_{im} = arg\,min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma) \tag{4.32}$$

Where the output value γ_{im} for each m terminal leaf is equal to γ value that minimises the summation of the updated residue values ($L(y_i, f_{m-1}(x_i) + \gamma)$) and this can be calculated through the chain rule for differentiation.

In the end, the predicted value of each sample will be updated by adding the predicted value from the previous iteration with the predicted terminal leaf output values produced from the previous step. The mathematical illustration of this is given in **Equation 4.33**.

$$f_m(x) = f_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}) \tag{4.33}$$

where f_{m-1} is the predicted output from the previous three, J_m is the node J at tree m, ν is the shrinkage rate or the learning rate found in the range of 0 to 1 and it is mainly used to control the contribution of each tree on the final prediction. The involvement of ν is commonly referred to as the regularisation strategy of which only add a fraction of the current predicted value to the previously predicted values. This method has shown to be effective in reducing the overfitting issues caused by the gradient optimisation and additionally provide a good prediction power in the long term. This parameter are optimised in the following studies.

Next, interaction depth and the number of iterations are two other main tuning parameters for this algorithm. Interaction depth indicates the depth for how much

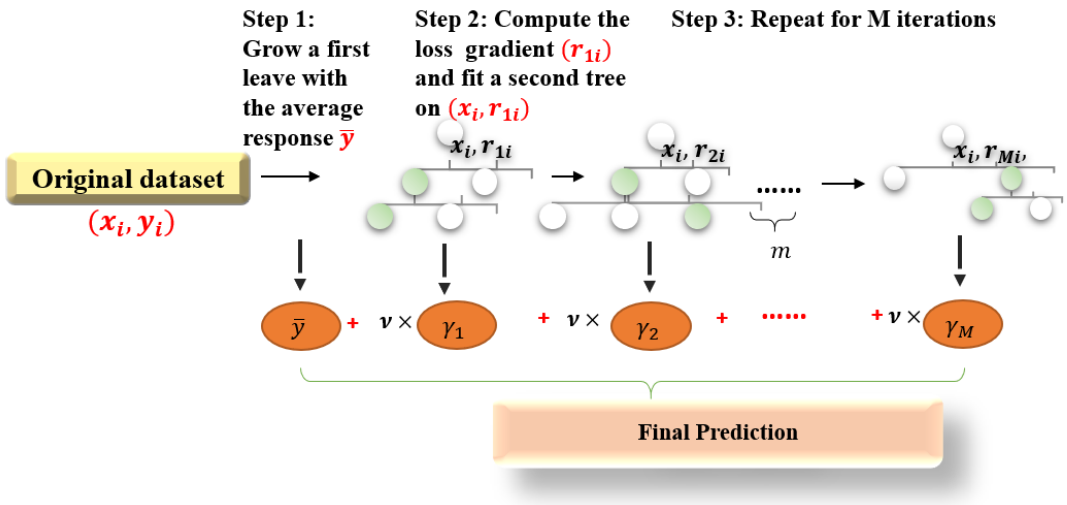


Figure 4.7: A diagrammatic representation of the modelling process of gradient boosting tree. x_i is the covariate variables, y_i is the original responsive variable, r_i^A is the residuals obtained from model A.

the interactions are allowed in-between the previous splits and the subsequent splits, and in practice this value set to be a minimal interaction that is allowed. The number of iterations refers to how many additive trees need to be built to describe the matter of interest. Furthermore, there is also another very similar algorithm named stochastic gradient boosting. The major difference between stochastic gradient boosting and general gradient boosting is seen from that the model is only trained on a fraction of the randomly selected data from the original training dataset. The method itself is borrowed from the concept of random sampling method in the bagging technique and it offers the benefit of lower variance error and lower computational cost. This technique is implemented in some of the following studies and is kept at the recommended value of 0.5 and this means that half of the training set is used to train the gradient boosting model.

Compared to the random forest algorithm, gradient boosting trees have a few distinct differences in the construction process, computational cost and the prediction rules shown in **Table 4.2**.

Random forest model in general is less likely to suffer from the overfitting as gradient boosting model do and this is due to the different splitting and fitting mechanisms. In addition, the random forest method is also less expensive com-

Table 4.2: Feature differences in random forest model and gradient boosting model [30].

	Random Forest	Gradient Boosting Trees
Tree Construction	Independently	One by one (the following tree construction depends on the previous tree)
Tree Depth limit	Maximum Depth	Minimum Interaction Depth
Trees Contribution	Equal	Unequal
Computational Time	Medium	Long

pared to the gradient boosting method as it only has two important hyperparameters to optimise and the trees are growing in batch which is more time-efficient. The gradient boosting model can be applied for a wider range of problems as long as we can determine the gradients such as the Poisson regression. In general, the gradient boosting model statistically outperform the random forest model for most prediction tasks only and if only the hyperparameters are carefully optimised.

4.10 Variable Importance

Variable importance is a measurement of the extent of the contribution of every covariate to the model prediction. This feature is very important to examine the relevance of covariate to the prediction and additionally, to identify the governing factor for the response variable. There are many ways of measuring this importance, one could simply compare the weights/coefficients of each covariate from the constructed penalised linear regression functions in **Equation 4.9** and **Equation 4.10** to gain an insight into the contribution of the variable.

Next, tree-based algorithms can also reveal the variable importance due to the high interpretable nature of the tree structure. Tree-based algorithms such as gradient boosting and random forest typically equip with the estimation approach based on the recording of the node purity improvement. In addition, the random forest also offers the alternative permutation method to measure the strength of the covariates in the prediction.

The first common approach is to record the improvements in the node purity over a certain performance metric for every occurrence of every covariate in the forest. The performance metric for measuring the node purity in the regression model

is often the residual sum of squares errors and has a similar format as indicated in **Equation 4.28**. In addition, the corresponding improvement values for each covariate variable are aggregated across the whole forest and then generating the overall variable importance.

The permutation estimation method is achieved through the out of bag (OOB) technique. This process involves the withdrawal of certain samples from the dataset are taken out from the bootstrap samples. During the growing process of ath tree, these OOB samples are used to grow the tree and the respective prediction error is recorded. In addition, the same tree is constructed again but with the OOB samples randomly permuted with some jth covariate values. The difference between permuted tree performance and non-permuted tree performance is recorded and aggregated across the whole forest to generate a ranking of the variable importance. This process is carried out with one covariate each time for every grown tree to give a fair estimation of the importance of the covariate in terms of their contribution to the overall prediction performance.

4.10.1 Shapley Values and SHAP package

Despite the previous techniques that can give insights into the variable importance, it would be also useful to quantify how the value of each covariate can affect the prediction of each instance, more than just the respective ranking. The contributions of each feature variable for the prediction of one response variable data-point ($\hat{y}$) from a general linear model can be computed through **Equation 4.34**.

$$\phi_j(\hat{y}) = \beta_j x_j - E(\beta_j X_j) \tag{4.34}$$

Where ϕ_j is the estimated contribution from the jth feature to the prediction of $\hat{f}$, β_j is the coefficient of the jth feature, E($\beta_j X_j$) is the mean effect of the jth covariate computed through averaging the contribution of the jth feature among all the predictions of the response variable. The contribution of each feature in the prediction of each datapoint is essentially the difference between the feature effect on a particular instance prediction and the average effect of that feature on the overall

predictions. In addition, the contribution values can either be positive or negative depending on their correlations. This computing method can be used as the basic principle for computing the relative contribution of each feature value based on their relative weights. On the other hand, not all machine learning models are weights-based principles such as tree-based algorithms and this limit the scope of application. To solve it, the concept of the Shapley value from the cooperative game theory can be applied with this method to make it accessible for all model types.

Shapley value is a widely used method in the coalitions game theory where it enables the estimation of how much of the payout should be distributed to each player in the most fairness. This concept can also be applied to estimate the contribution of every covariate variable to the model prediction of every response variable data-point. Each data instance from the covariates are treated as a player in a coalition (feature value combination sets) and the respective Shapley value for each instance can be calculated using the Monte-Carlo sampling method. Given the investigating object to be a value instance from the j^{th} covariate variable, the Shapely value for this instance is calculated through computing the difference in the predictions for with and without the feature value (x^{m}_{+j}) at the m^{th} set of feature combination. This process will be carried for the same covariate until all the possible feature combination sets are reached and the final Shapley value is calculated by taking the mean value of them. A mathematical representation of this is given as

Equation 4.35

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^{M} (\hat{f}(x^m_{+j}) - \hat{f}(x^m_{-j})) \tag{4.35}$$

Where $\hat{\phi}_j$ is the Shapley value for covariate "j", M is the maximum number of feature value combinations, $\hat{f}(x^m_{+j})$ is the for the prediction results of which the single instance x of covariate j is included in the dataset for the training in the m^{th} iteration and $\hat{f}(x^m_{-j})$ means that the prediction for when the single instance x of covariate j is taken out from the predictive model. This process will have to be carried out for every instance of each covariate used in the model to gain an overview of the variable correlations with the response variable. The value of the Shapley indicates

the contribution of this covariate data instance to the prediction of the response data instance. The more positive and more negative Shapley values indicate that that the covariate instance has a bigger contribution to the model prediction. The overall variable importance of the covariates is ranked based on the averaged Shapley values for all data instances of certain covariate like shown in **Equation 4.35**. In addition, the sign of the Shapley value also illustrates the detailed relationship of the investigating covariate with the target variable. A positive Shapley value implies that the covariate instance contributes positively to the overall prediction, in other words, the higher the value of the covariate instance, the higher the value of the response variable instance there would be. On the other hand, a more negative Shapley value of a covariate instance implies a negative correlation with the response variable where the increase of the value of the covariate instance would lead to a decrease of the response variable. The extent of these correlations is again controlled by the magnitude of the computed Shapley values.

In this project, the Shapley Additive exPlanations (SHAP) [267] package is used to compute the Shapley values for the contribution of every covariate value instance to the prediction of discharge capacity.

4.11 Summary

In this chapter, a discussion was made over the data collection rules, data analysis, techniques of data transformation, model training and validating strategy, implemented in this thesis. Different types of linear and non-linear algorithms are selected to compare their potential prediction power to the discharge capacities. The optimal model that gives the best prediction performance will be evaluated with the variable importance technique to explain the contribution of each covariate variable to the prediction. Comparisons of the advantages and disadvantages of the aforementioned ML algorithms that have been discussed are summarised in **Table 4.3** and **Table 4.4**.

Table 4.3: Summary of the advantages and disadvantages of linear and non-linear machine learning algorithms.

	Name	Advantages	Disadvantages	Ref
Linear Algorithms	General Linear Model	-Easy to implement -Good with a small dataset -Highly interpretable	-Does not capture non-linear correlations well -Simple model structure	[268], [236]
	Penalised Regression	-Less prone to overfitting of the training data. -Flexible model complexity (Coefficient shrinkage)	-More computationally expensive than the general linear model (needs to tune lambda hyperparameter)	[236], [269], [190]
	Bayesian Ridge Regression	-Highly adaptable to different datasets -Less prone to overfitting issues	-Time inefficient to do the model inference	[270]
Non-linear Algorithms	Artificial Neural Network	-Robust in explaining complex non-linear correlations -Great with handling large dataset	-Prone to overfitting -High computational cost for training -Black box model and difficult to interpret	[270], [271], [272]
	Support Vector Machine (SVM)	-Less sensitive to outliers -Kernel benefits (a computationally less expensive way to interpret the data in higher dimensions)	-Slow for handling large dataset -Require the rescaling of the dataset	[270], [272], [273]
	Kernel Ridge Regression	-Faster to compute for medium-sized dataset than SVM -Kernel benefits	-Require the rescaling of the dataset -More computationally expensive than SVM as it doesn't have the concept of "supported" vectors	[22], [274]
	K-nearest Neighbour	-Easy to understand and implement -Easy to tune the model parameter (only one hyperpamater presents)	-High computational cost for handling large dataset -Sensitive towards outliners	[269], [270], [205]

Table 4.4: Comparisons of tree-based algorithms including the decision tree, randomforest and gradient boosting machine.

	Name	Advantages	Disadvantages	Refs
Tree-based Algorithms	Decision Tree	-Easy to interpret the model structure -Does not need the dataset to be normalised	-More likely to suffer from overfitting issue -Unstable tree structure and the poor prediction power from the simple model	[190], [270], [272], [273], [22]
	Random Forest	-Flexible with handling different sizes of the training data -Less likely to overfit the dataset -Provide insights into the variable importance -Capable of handling missing data	-A large number of hyperparameters need to be optimised for -Difficult to visualise the whole model structure	[190], [270], [272], [273], [22]
	Gradient Boosting Machine	-Does not require data rescaling -Can ranks the input variable importance	-More computationally expensive to optimise than random forest (more hyperparameters) -Difficult to visualise the whole model internal structure	[190], [22], [275]

Chapter 5

Machine Learning Prediction of the Discharge Capacities of Doped Spinel Cathode ($LiM_xMn_{2-x}O_4$) Materials

5.1 Background

A wide range of attention has been drawn towards the research of $LiMn_2O_4$ (LMO based spinel cathode materials as they are a much cheaper, non-toxic alternative to the widely commercialized lithium cobalt oxide ($LiCoO_2$). In comparison to $LiCoO_2$, spinel materials have a more robust crystal structure that can withstand the structural distortion from the Li-ion intercalation/deintercalation during cycling, and offers a higher Li-ion diffusion rate. In addition, the robust structure enables the deintercalation of the full Li-ion formula unit during charging and thus its practical capacities are closer to the theoretical capacity ($\sim 148 \text{mAhg}^{-1}$). On the other hand, the LMO cathodes are known for their issues in drastic capacity fading during cycling and the limited rate capability which restrict their use in large-scale commercial applications. The underlying reason for these inferior properties is due to the manganese ions in the crystal structure. The first reason is the dissolution reaction of the manganese ions Mn^{3+} from the material surface into various forms of Mn^{4+}(solid) and Mn^{2+}(sol), which reduces the Li-ions site energy and lowers the rate of reversible electrochemical reactions. The second reason is the Jahn-Teller

distortion (JTD) effects initialized from the high spin electrons interactions from the d-orbital electrons of manganese ion (Mn^{3+}), which elongate the crystal structure along the z-axis, reduce its structural stability and lead to a shorter cycle life.

Doping the manganese (III) sites with lower-valence (lower than 3+) dopants **Figure 5.1-a**) seems to be an effective approach to this problem as it increases the average Mn valence in LiMn2O4 to suppresses the JTD effect by reducing the concentration of Mn (III) and eventually decrease the rate of the dissolution reaction. Indeed, promising results have been seen in capacity improvement for lower-valence dopants like Al [276], Cr [277], Fe [278], Gd [279], Ga [280], Mg [153], Nd [281], Ni [146], Ru [282], Sc [283], Zn [284]. On the other hand, recent research seems to oppose this assumption as the use of higher-valence dopants such as Si [285], Sn [286] have also been shown to suppress the intrinsic material's inferior properties and improve their discharge performance. For instance, **Figure 5.1-b** illustrates that the high-valence silicon doped LMO material offers a higher discharge capacity on the 60^{th} cycle than both the lower-valence gallium-doped system the pristine material. This seems contradictory as the introduction of high-valence dopants would be expected to increase the Mn valence in the system, promoting the manganese (III) dissolution reactions and eventually worsen the cyclic performance. This shows that a comprehensive understanding of the principles of doping, and their effects on the spinel cathode discharging capability is still lacking. One root cause is seen from the difficulty in interpreting cross-group experimental results as it requires a deep understanding of the variations of the experimental parameters such as synthesis routes, doping concentration and testing conditions. The present study aims to use robust machine learning techniques to uncover the underlying relationships of these complex material properties and the discharge performance of the doped LMO cathode.

The discharge capacity at a certain cycle is not a quantity that can be easily derived by any existing physical models. In addition, controlling many independent parameters have made the experimental measurement rather difficult to perform. With these stated challenges, this chapter aims to train various machine learning

models with the manually collected experimental data to predict the initial and 20^{th} cycle end discharge capacity for 102 doped spinel systems. These predictions are made based on various reported material properties such as the material molar ratio of the dopant (M) and manganese atom (Mn), crystal lattice constant-a (LC_a), elemental properties such as the material formula molar mass (Mr), dopant's electronegativity (M_EN) and the current density (CD) used for cyclic testing. The diagrammatic illustration of the model design used in this chapter is given in **Figure 5.1-c**. First, seven multivariate models are built for each prediction task and this covers linear algorithms such as the ridge regression (RR), lasso regression (LR), non-linear algorithms like support vector machine (SVM), deep neural network (DNN) and the tree-based algorithms such as the decision-tree (DS), random forest (RF) and the gradient boosting model (GBM). In the end, a variable importance study is performed to gain insights into the governing material properties on the prediction of both discharge capacity from the best performing ML models.

5.2 Methods

5.2.1 Data Collection

To begin with, the study objective of the journal should be focussed on spinel cathodes that are doped with a single dopant. In addition, it is noticed that a considerable number of spinel cathode studies are made on the effectiveness of coating because of their promising results in suppressing manganese dissolution. However, the materials that have received such treatment are neglected for a consistent dataset. Besides the requirement of a single doped, non-coated spinel cathode system, other selection rules are also implemented such as the material should have a space group of Fd-3m and any materials that are either indicated as oxygen-deficient or have used anion doping like fluoride were discarded to maintain high data consistency. In addition, the cyclic performance test must fulfil all listed conditions to qualified for the collection of discharge capacity:

1. used lithium foil as the contrasting anode;
2. used aqueous $LiPF_6$ as the electrolyte;

3. employed a constant current density for charging and discharging and these should be within 150 $mAhg^{-1}$(also called 1C for performing LMO system cyclic test);

4. performed a minimum of 20 charges/discharge cycles;

5. Used standard conditions (T = 25 °C, P = 1atm) to perform the cyclic test.

In addition, 20 cycles are selected as the maximum cycle capacity collection target as it is the most commonly reported cycle across the literature for the studies of doped spinel cathodes. Following these selection rules, the initial 517 collected

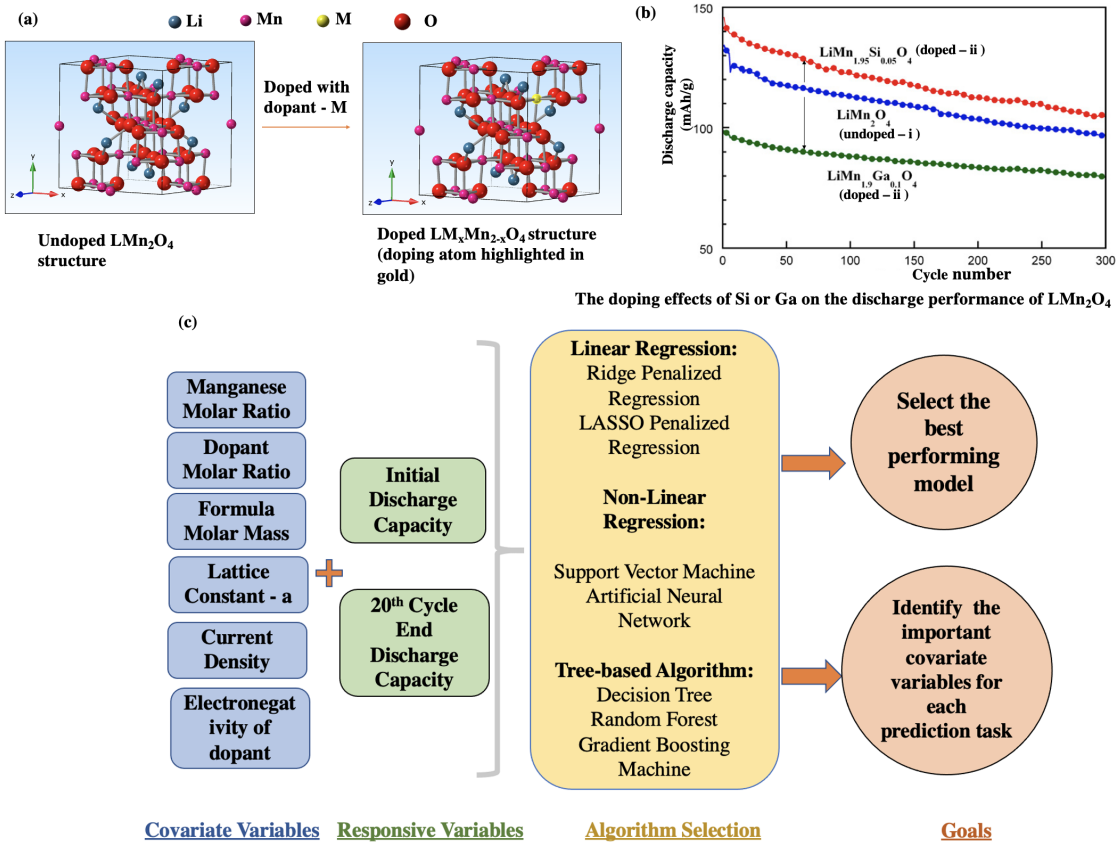


Figure 5.1: a) A schematic illustration of doping the dopant (M) into the crystal structure of $LiMn_2O_4$ (LMO). b) The comparison of the discharge performance of the Silicon doped LMO ($LiMn_{1.95}Si_{0.05}O_4$), Gallium doped LMO ($LiMn_{1.9}Ga_{0.1}O_4$), and the pristine LMO material, arrows indicate the discharge capacity gap of each doped LMO with the pristine LMO. Image reproduced with permission from [23]. Copyright 2020, Journal of Power Sources c) A overall summary of the model design, a selection of the linear, non-linear and tree-based algorithms and the final goals of this project.

publications were reduced to 34 publications which are published in the years between 1998 and 2019. These qualified publications were turned into a dataset of 102 different doped LMO systems with 17 dopant selection variations (B, Co, Ni, Al, Cr, Ce, Sc, Mg, Gd, Zn, Si, Sn, Nd, Ga, Co, In, Ru). For each doped LMO system, the material properties such as the formula ratio of dopant (M), the formula ratio of manganese (Mn), crystal structure lattice constant "a" (LC_a), the material formula molar mass (Mr), the electronegativity value of the dopant atom (M_EN) and the experimental cyclic test current density (CD) were collected. The morphology details of the materials were unable to be collected due to the lack of reporting and the difficulty in interpreting the graphical results.

5.2.2 Model Training and Optimising

All machine learning models in this chapter are constructed using various R (version 3.6.0) libraries, including caret, gbm, randomForest and Keras. To begin with, the variable correlations will be studied through the use of the Pearson Coefficient Correlation technique and it will indicate the extent of linearity between the variables in the dataset. The whole data space was randomly split into the ratio of 4:1 corresponding to the model training set and test set. In addition, the model hyperparameters of all machine learning algorithms were optimised using a ten-fold cross-validation method (on the training data only) and these investigating ranges are indicated in **Table 5.1**. It is important to note that the gradient boosting algorithm implements a stochastic technique of which within each fold of training data, a subsample of 10% of the training data is randomly drawn and used to fit the individual decision tree instead of the whole data space. This method can introduce further randomisation to the model training process and reduce the likelihood of overfitting issues. Further to the model training and optimising, comparisons will be made on all models' prediction power over the test set. The model that equips with the highest prediction power (lowest prediction error) will be examined in the variable importance study to reveal the detailed variable correlation with the response variable. In addition, a randomseed number is selected for reproducible results. **Please note, the data curated for this chap-**

Table 5.1: A summary of the considered hyperparameters for various machine learning models. The random forest model and the gradient boosting models share similar hyperparameters besides the stared variable (*) of which is only available in the gradient boosting algorithm.

Algorithm name	Hyperparameter	Investigating range
Artificial Neural Network	Number of hidden layers	1∼15
	Number of neurons in each hidden layer	1∼15
	Learning Rate	0.001, 0.01, 0.1
Support Vector Machine (Radial Basis Function)	Cost	10^{-4} ∼10^{4}
	Gamma	10^{-3} ∼10^{3}
Random Forest and Gradient Boosting Machine (Stochastic technique)	Number of trees	100∼5000
	The maximum number of features considered at each split	1∼5
	Learning rate*	10^{-4} ∼10^{-1}

ter and the written machine learning codes can be found on the GitHub page: `https://github.com/thepowerligand/LMO-ML`. The detailed references are provided for each set of experimental data collected, in the same GitHub directory.

5.3 Results and Discussion

5.3.1 Preliminary Statistical Analysis

Figure 5.2 represents the matrix results of the Pearson coefficient correlation study performed for each pair of variables in the dataset. The Pearson correlation coefficient (R) values are presented in the range of 0 and 1, with the value closer to either -1 (darker blue) or 1 (darker red) implying a more positive or negative linear relationship between the investigating variable pair. From the results presented, one's intuition should not be misguided by the perfect negative linearity (R = -1) relationship estimated for the M and Mn as it is the result of direct site substitution. In addition, Mr is found to have a relatively stronger linear correlation with IC and with EC than other pairs with coefficients estimated to be R = 0.38 and R = 0.35 respectively. To conclude, there are no strong linear correlations (R $>$ 0.6) being

observed for any of the covariates (M, Mn, LC_a, CD, M_EN) with the two response variables (IC, EC).

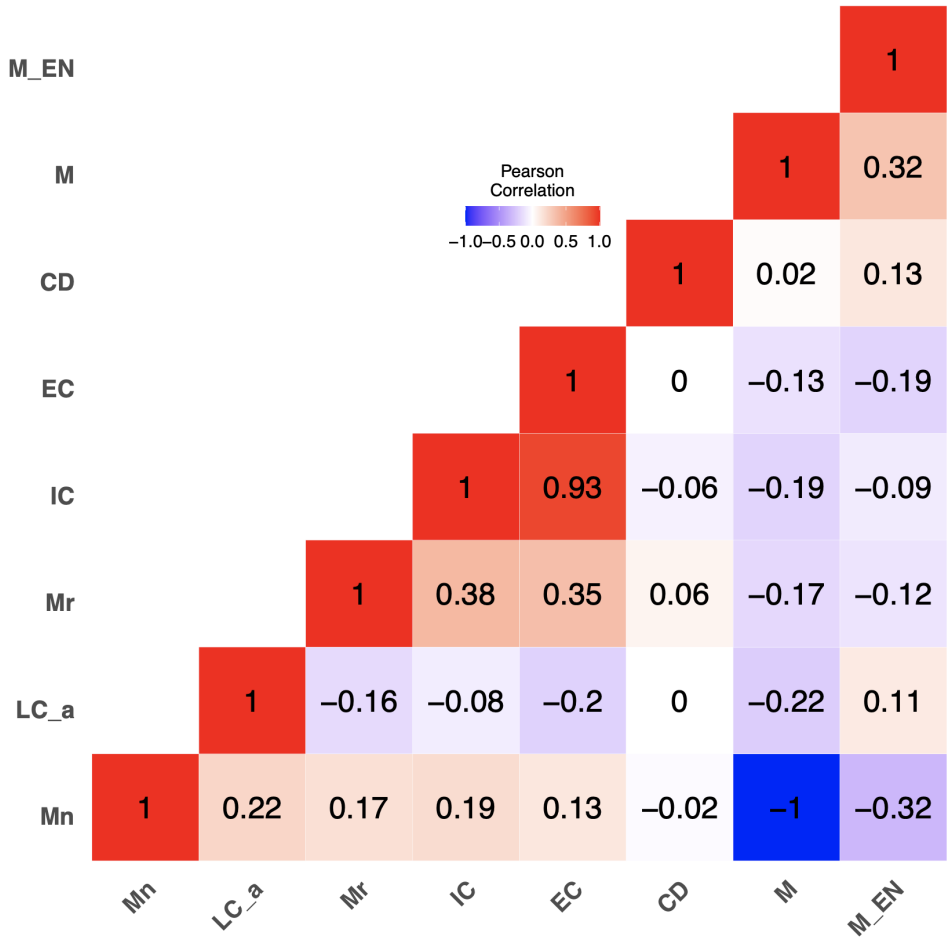


Figure 5.2: The matrix of the Pearson correlation coefficient of every variable pair including the responsive variables, initial discharge capacity (IC), the 20^{th} cycle discharge capacity (EC) of the doped $LiMn_2O_4$ cathode and the covariate variables, the dimension of the unit cells in a single crystal lattice (LC_a), the formula ratio of manganese (Mn), dopant (M), the electronegativity of the selected dopant (M_electron) and the applied current density for the cyclic test (CD).

5.3.2 Penalised Linear Regression

Both the ridge regression and the LASSO regression techniques have been applied to predict the initial and the 20^{th} cycle end discharge capacity of doped-LMO systems. Within each model, the lambda parameter of which controls the penalisation strength of the objective function is optimised with 1000 values set generated with the minimum lambda criteria 10^{-6}. The averaged mean squared errors are estimated

for all ten-fold cross-validation for every lambda value and the optimal value is selected of which the MSE reaches the near minimum. To simplify the result sections, the lambda optimisation results' graphs for each Ridge and LASSO model for the prediction of IC and EC values are given in the Appendix**.** The optimisation graphs for the Ridge model for IC prediction are given in **Figure A.1**, **Figure A.2**, **Figure A.3**, **Figure A.4** and for EC predictions are given as **Figure A.5**, **Figure A.6**, **Figure A.7**, **Figure A.8**. The optimisation graphs for the LASSO model for the prediction of IC are given in **Figure A.9**, **Figure A.10**, **Figure A.11**, **Figure A.12** and for the EC prediction are given in **Figure A.13**, **Figure A.14**, **Figure A.15**, **Figure A.16**.

In general, linear models are sometimes favoured for their clear interpretation of the contributions of each covariate to the prediction of the response variable. **Table 5.2** summarises the weights or the coefficients that are estimated for each of the covariates in the ridge models and LASSO models. First, it is seen that the LASSO models for IC and EC predictions have penalised the weights of some covariate variables to 0, such as the formula ratio of dopant (M), manganese (Mn), the electronegativity value of dopant in LASSO-IC and the dopant ratio and the molar mass in the LASSO-EC model. These results are expected as the LASSO penalisation technique would eliminate the covariates that do not contribute significantly to the overall prediction, by shrinking their weights to 0 in the model. On the other hand, the ridge penalisation models penalise the covariates by shrinking their weights close to 0 to reduce their influence on the overall prediction.

Based on the results by the RR-IC mode and LASSO-IC model, lattice constant "*a*" and the material formula molar mass are shown to have the highest weights in each respective model, implying their important role for the overall prediction. In the case of EC prediction, the RR-EC model seems to assign lattice constant "*a*" (-61.57) with the heaviest weights whereas the LASSO-EC model picked the current density (-70.70) to be the dominating variable with the heaviest weight. In addition, an interesting observation can be such that both ridge models have assigned the most amount of weights to the same covariate (LC_a) and the model lambda values

Table 5.2: Calculated coefficients of the covariate variables by the penalised linear regression model

	RR-IC	LASSO-IC	RR-EC	LASSO-EC
M	-0.23	0.00	-0.10	0.00
Mn	0.04	0.00	1.71	-4.04
M_EN	-0.71	0.00	-6.09	-36.95
Mr	16.78	-45.35	-0.01	0.00
LC_a	-23.69	-1.98	-61.57	-7.70
CD	-0.036	0.00	0.93	-70.70
Lambda Optimal value	111.50	4.54	123.11	4.13
Intercept	100.36	110.66	170.44	104.12

are estimated to be very similar. On the other hand, the LASSO models seem to have selected different important covariates for prediction despite their similarity in the lambda value.

Figure 5.3-a and **Figure 5.4-a** illustrate the predicted initial capacity values computed by the RR-IC model and the LASSO-IC model against the experimental initial capacity values from both training and testing sets. In addition, **Figure 5.3-b** and **Figure 5.4-b** give the predicted 20^{th} cycle end discharge capacity computed by the RR-EC model and the LASSO-EC model, respectively, against the experimental 20^{th} cycle end discharge capacity for both the training and test set. To begin with, all models have shown poor predictions for both the training and test sets as the trend of the plotted points (deep blue and light blue) do not align with the red true line. Furthermore, it is observed that all of the estimated $RMSE_{test}$ values are bigger than the $RMSE_{train}$ which imply a high likelihood of model overfitting to the training set as the training prediction is better than the test prediction. In addition, there have been noticing small changes for all predicted capacities across four results graphs, indicating that the models are not able to predict well with the given covariates. In particular, the coefficient results of many covariates have been shrunk to "0" by the LASSO models as indicated in **Table 5.2**. This implies that only a small fraction of the covariates are accountable for the final model prediction and hence would result in similar predictions.

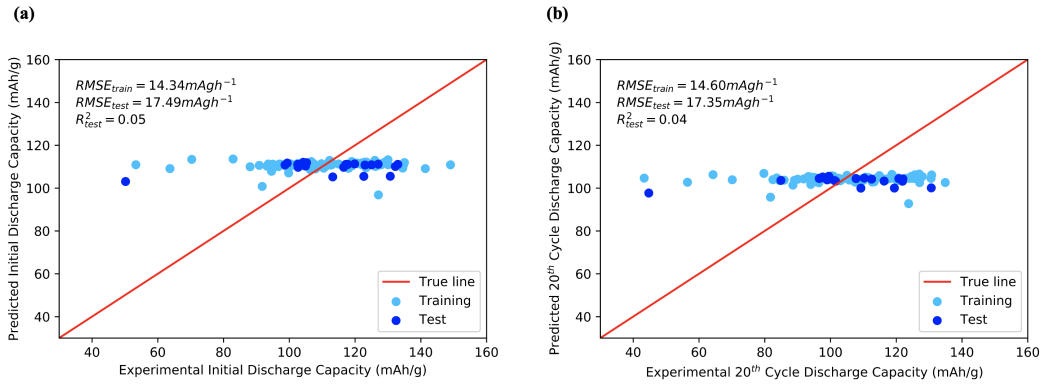


Figure 5.3: Predicted values against the experimental values computed by the ridge regression models for the prediction of (a) initial discharge capacity and (b) 20^{th} cycle discharge capacity of doped $LiMn_2O_4$ cathode.

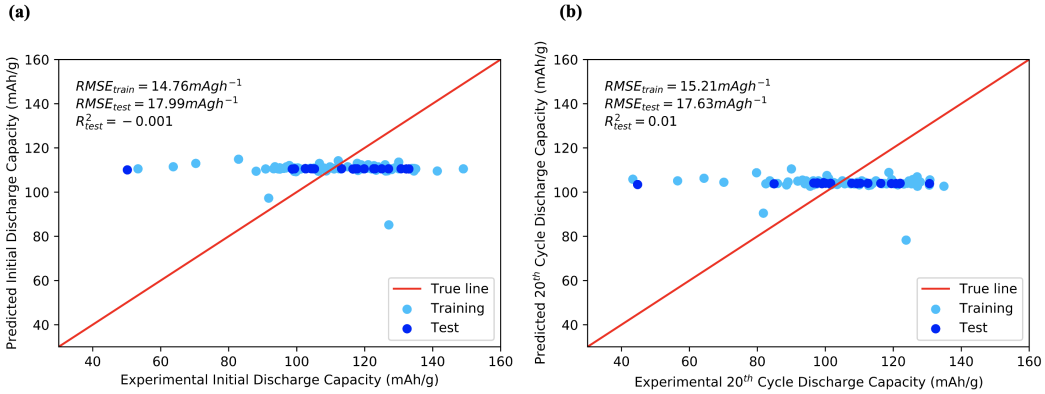


Figure 5.4: Predicted values against the experimental values computed by the LASSO regression models for the prediction of (a) initial discharge capacity and (b) 20^{th} cycle discharge capacity of doped $LiMn_2O_4$ cathode.

5.3.3 Artificial Neural Network

The optimisation of the artificial neural networks models was carried by selecting the suitable number of epochs. The number of epochs indicates the number of times that the whole dataset will be fed to the model and this would influence the model training errors. **Figure 5.5-a** and **Figure 5.5-b** demonstrate the training loss against the number of epochs up to 10000 computed by the ANN-IC model and the ANN-EC model, respectively. From the results graphs, it is seen that the loss values seem to drop as the number of epochs number increases and stabilise after the number of epochs reaches 1000 (indicated by the dotted red line). Hence this is chosen as the optimal number of epochs to train the ANN models.

Next, hyperparameter optimisation is carried out for the number of neurons

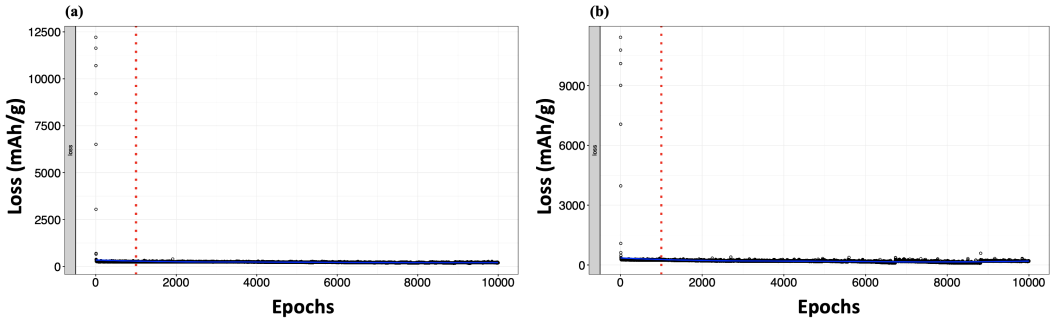


Figure 5.5: The loss values against the number of epochs computed by the artificial neural network models for the prediction of the (a) initial discharge capacity and (b) 20^{th} cycle discharge capacity of doped $LiMn_2O_4$ cathode. The learning rate is set to be 0.001, the number of neurons per layer equals 10 and the number of hidden layers is equal to 10.

per layer, the number of hidden layers and the learning rate for each ANN model. **Figure 5.7** give the colour coded averaged ten-fold cross-validated mean squared error for every combination of investigating hyperparameters within the investigating ranges identified in **Table 5.1**. **Figure 5.7 (a-c)** are the computed error results for the IC prediction and **Figure 5.7 (d-f)** are the computed error results for the EC-prediction, under three different settings of learning rate. From **Figure 5.7 (a-c)**, it is seen that each learning rate group contains a similar number of parameter combinations that give the low mean squared errors ($< 250 \text{mAhg}^{-1}$). A similar trend can also be observed from the EC graphs in **Figure 5.7 (d-f)** where the learning rate of 0.1, 0.01, 0.001 seem to contain a similar amount of low-error hyperparameter combinations. The optimal hyperparameters are selected based on which hyperparameter combination would generate the lowest mean squared errors (MSE) values among all and these are summarised in **Table 5.3**. During the investigation of the optimal hyperparameters sets for the ANN-EC model, the combination of 15×15 (number of neurons and number of layers) under the learning rate of 0.001 has given the lowest MSE. Nevertheless, this hyperparameter set was discarded these would lead the model to be overly parameterised and increases their likelihood of being susceptible to the overfitting issues.

Figure 5.6-a and **Figure 5.6-b** give the results plots of the predicted value against the experimental value for both the training set and the test set computed

Table 5.3: The optimal learning rate, number of hidden neurons per layer and the number of hidden layers used to build the final artificial neural models for the prediction of initial discharge capacity (IC) and the 20th cycle end discharge capacity (EC) of doped $LiMn_2O_4$ cathode.

	LMO-IC	**LMO-EC**
Learning Rate	0.001	0.001
Number of hidden neurons per layer	14	9
Number of hidden layers	5	8
Number of Parameters	888	639

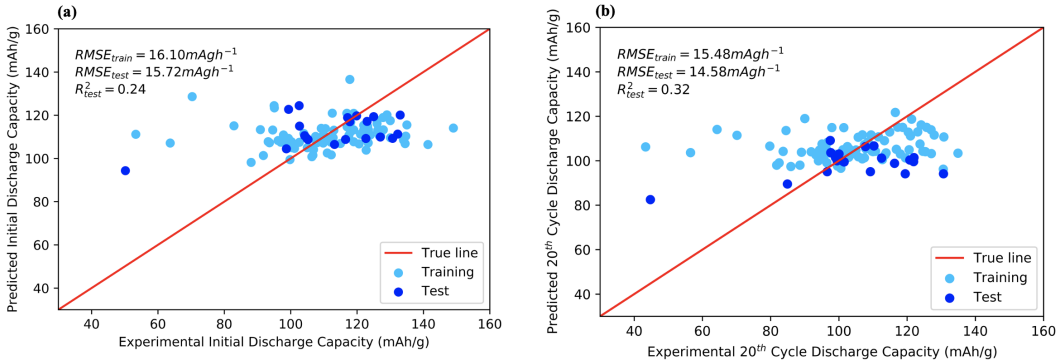


Figure 5.6: Predicted values against the experimental values computed by the artificial neural network models for the prediction of (a) initial discharge capacity and (b) 20th cycle discharge capacity of doped $LiMn_2O_4$ cathode.

by ANN-IC and ANN-EC, respectively. To begin with, RMSE values for the test-set are computed to be 15.72 mAhg^{-1} and 14.58 mAhg^{-1} by the ANN-IC model and the ANN-EC model respectively, which are much lower than the results of penalised regression. Nevertheless, all of the plotted dots for training and testing sets on both results graphs (**Figure 5.6 (a-b)**) are shown to be further from the red true line which indicates that the models cannot generalise well for the whole dataset. These poor prediction performances could be that the model overfits the small sample size of which the overly complex model architectures (shown as the number of parameters from **Table 5.3**) capture more noise from the dataset than any useful signals regarding the variable correlations.

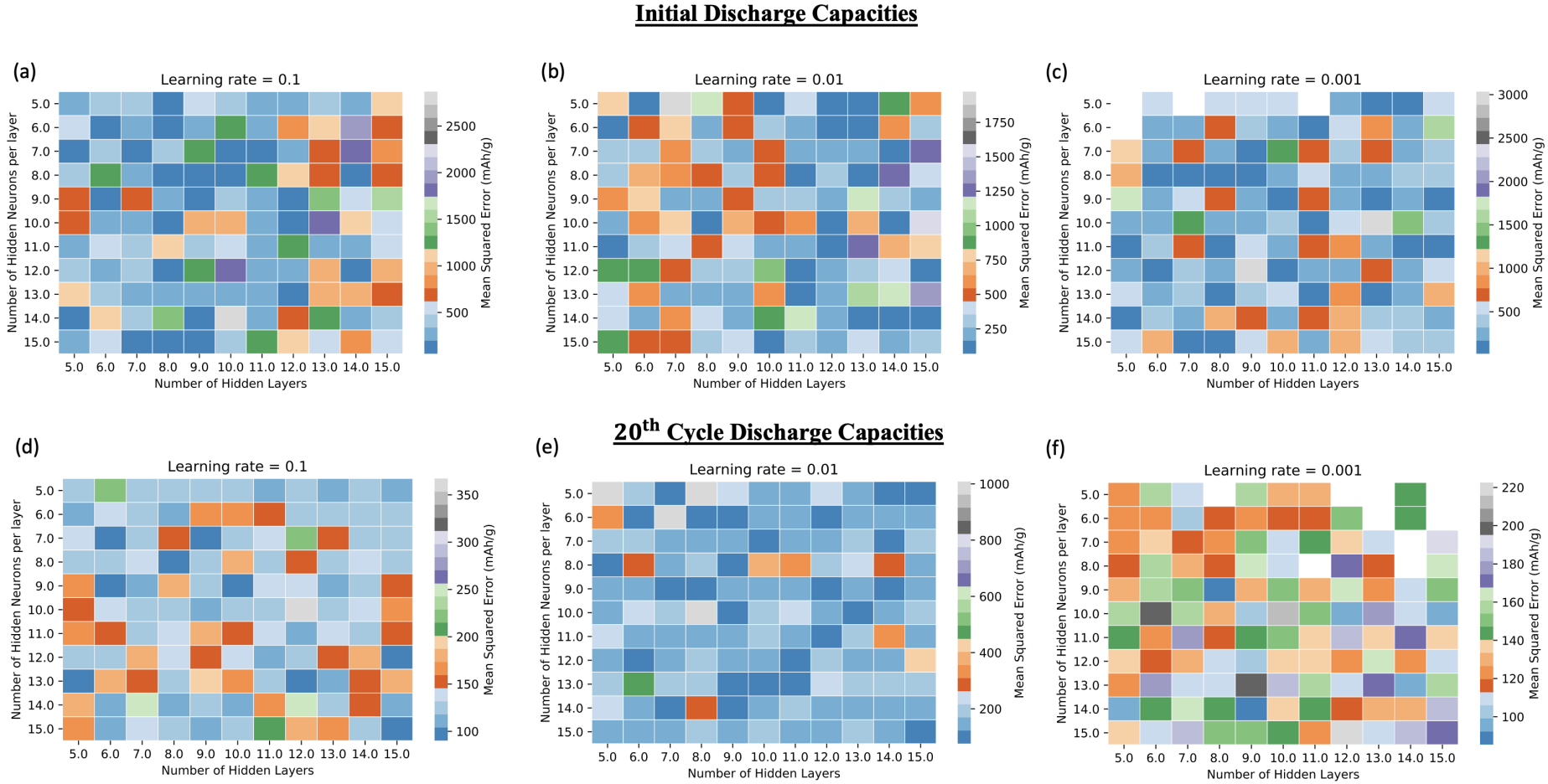


Figure 5.7: The heat-map representation of the hyperparameter optimisation results of artificial neural network for the prediction of (a)-(c) initial discharge capacities and (d)-(f) for the 20th cycle discharge capacities of doped $LiMn_2O_4$ cathode. The blank squares from (c), (d) are the results with significantly large MSE values and are deleted to maintain a similar scale on the colour bar.

5.3.4 Decision Trees

Two decision tree models were constructed with each being optimised for the complexity factor. The complexity factor is used as the minimum improvement criteria that prevent the further growth of the decision tree if the further splits do not meet this benchmark. **Figure 5.8-a** and **Figure 5.8-b** give the mean ten-fold cross-validated error results against the three complexity factors examined for the decision tree models used for the prediction of IC and EC, respectively. A common trend between the two results graphs can be observed that the increase in the complexity factor can increase in the RMSE values. The optimal complexity parameters for the IC and EC decision tree models are selected to be 0.036 and 0.046 accordingly as these produce the lowest error values.

One inherent advantage of using decision tree models is that they have a highly interpretable model structure that reveals the embedded rules of variable splitting at each node. **Figure 5.9** and **Figure 5.10** give the splitting patterns of the decision tree models built upon the optimised complexity parameters, for IC and EC prediction and these models are denoted as DT-IC and DT-EC, respectively. To begin with, DT-IC has a much simpler splitting pattern than DT-EC with only two main covariates variables involved during the decision process, namely, the dopant ratio in the material formula and the material formula molar mass. On the other hand, the DT-EC involves one more splitting variable and the splitting order is given as the formula manganese ratio, material molar mass and the unit cell crystal lattice constant "a". Both models have involved the formula molar mass as the internal splitting variable and this suggests that the Mr is an influential factor in the predictions of both capacities variables.

Figure 5.11-a and **Figure 5.11-b** illustrate the predicted values versus the experimental values estimated by the DT-IC model and the DT-EC model, respectively. The prediction error of the DT-IC model over the test-set validation is estimated to be 15.21 mAhg^{-1} which is an improvement of 0.51 mAhg^{-1} from the results computed by the ANN-IC model (**Figure 5.6-a**). On the other hand, the DT-EC model seems to perform worse over the test-set prediction than the ANN-EC

model as its respective RMSE value is 1.89 $mAhg^{-1}$ higher. In addition, it is observed that similar predictions values have been given to the different experimental in both results graphs. This phenomenon is expected as the decision tree algorithms use the mean value of the group under each node as being the prediction for all the instances within that node group. The R^2 scores for DT-IC and DT-EC model predictions over the test are estimated to be 0.28 and 0.13, which indicate that the models can capture 28% and 13% of the variations of the capacity in the test set. In comparison to the results estimated by both ANN models, both DT models have not shown great improvements in the prediction performance as their R^2 scores are similar.

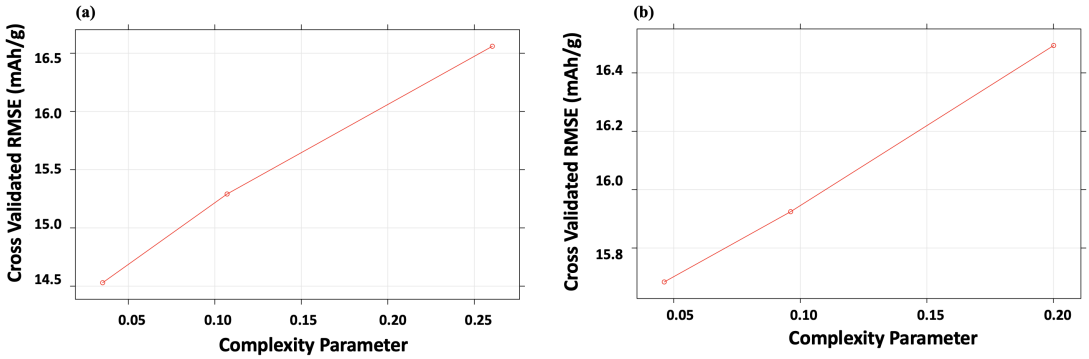


Figure 5.8: The mean values of the root-mean-squared errors from all ten-fold cross-validation estimated by the decision models against the complexity parameter values for (a) initial discharge capacity and (b) 20^{th} cycle discharge capacities of doped $LiMn_2O_4$ cathode, respectively.

5.3.5 Random Forest

The optimisation of the random forest model involves the turning of hyperparameters such as the number of trees and the number of variables considered at each split. The number of trees is selected randomly from the range of 100 to 5000 and the number of covariates considered at each split will be investigated from the range of 1 to 5. For each capacity prediction task, 5000 sets of hyperparameter combinations were generated randomly and were investigated for their prediction performance. **Figure 5.12** and **Figure 5.13** are the 3D plots of the mean ten-fold cross-validated MSE estimated by the RF-IC model, the RF-EC model under these

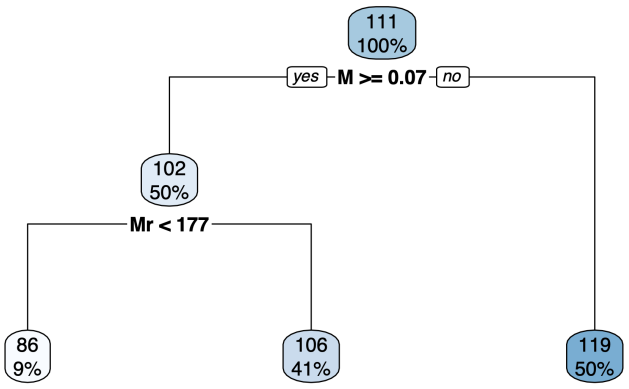


Figure 5.9: A diagrammatic illustration of the decision tree splitting patterns for the prediction of initial discharge capacities of doped $LiMn_2O_4$ cathode with the complex factor fixed at 0.036.

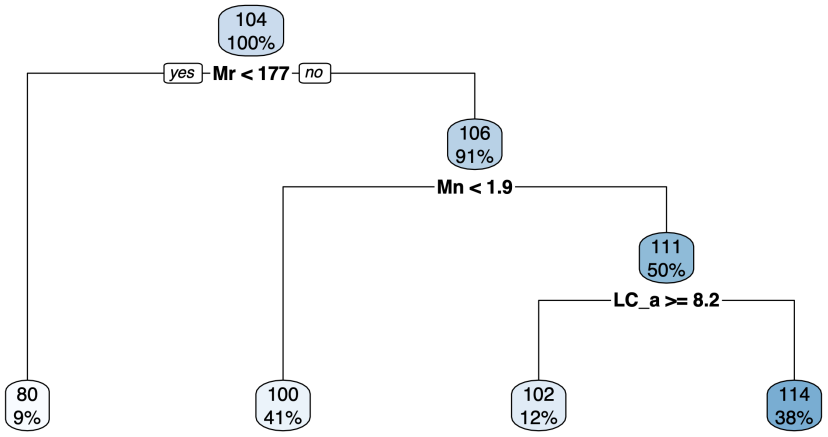


Figure 5.10: A diagrammatic illustration of the decision tree splitting patterns for the prediction of 20^{th} cycle discharge capacities of doped $LiMn_2O_4$ cathode with the complex factor fixed at 0.046.

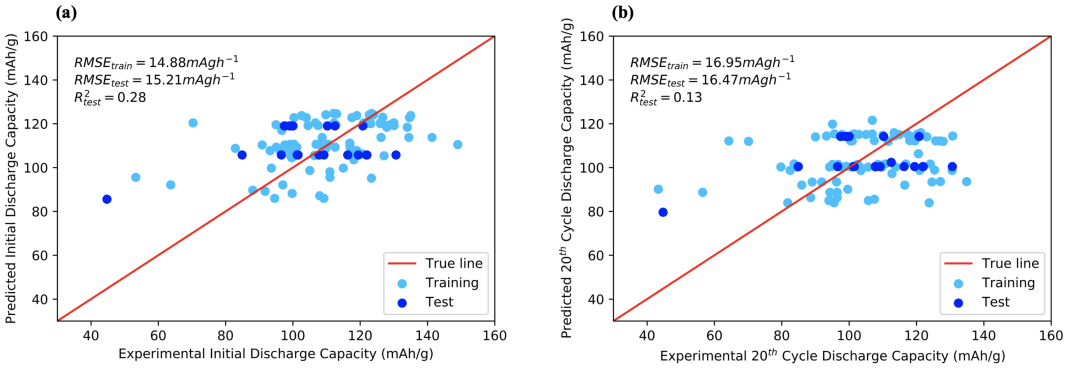


Figure 5.11: Predicted values against the experimental values computed by the decision tree models for the prediction of (a) initial discharge capacity and (b) 20^{th} cycle discharge capacity of doped $LiMn_2O_4$ cathode, respectively.

Table 5.4: The optimal values of the number of trees and the number of covariates considered at each split for the prediction of initial discharge capacity (IC) and the 20^{th} end discharge capacity (EC) of doped $LiMn_2O_4$ cathode and their respective computed mean squared error values.

Hyperparameters	**Initial discharge capacity**	**20^{th} cycle end Discharge capacity**
Number of trees	2510	160
Number of features considered at each split	5	5
Mean squared error (mAhg-1)	230.74	182.46
Number of Parameters	888	639

randomly generated hyperparameter combinations. First, it can be seen that the prediction errors are more evenly distributed (similar colour) across a different number of trees at the same number of splitting covariates as the colour does not change much for the same covariate variable. On the other hand, the MSE values seem to decrease when the number of covariates considered at split increases from one to five across a different number of trees iterations (colour changes from grey, to purple, to green and then to blue) in both results graphs. This implies that the number of covariates variables considered at each split is more influential than the number of trees in the overall model prediction.

Next, two random forest models are constructed with the optimal hyperparameters as listed in **Table 5.4**. **Figure 5.14-a** and **Figure 5.14-b** give the graphs of the

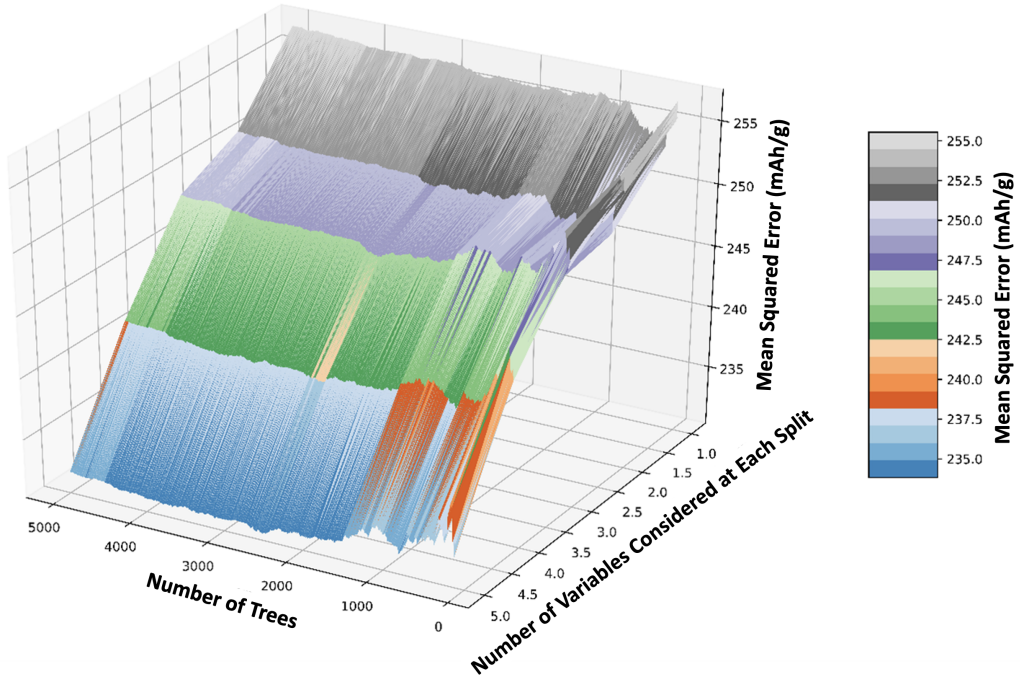


Figure 5.12: The optimisation plot for the computed mean squared error for 5000 combinations of the number of trees, the number of variables considered computed by the random forest models for the prediction of Initial discharge capacity (IC) of doped $LiMn_2O_4$ cathode.

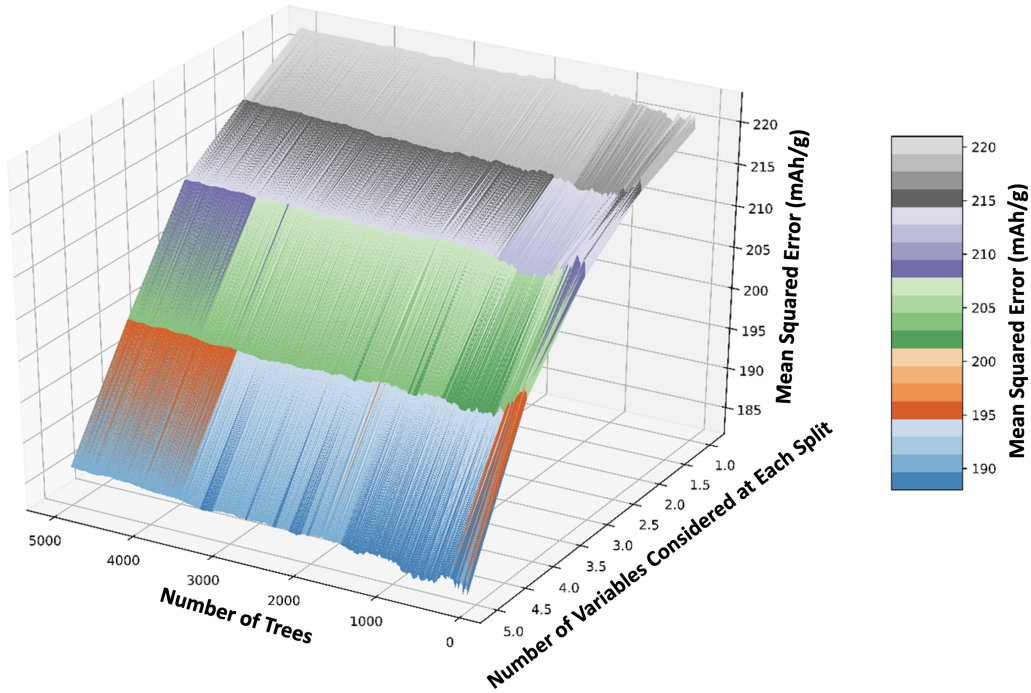


Figure 5.13: The optimisation plot for the computed mean squared error of 5000 combinations of the number of trees, the number of variables considered computed by the random forest models for the prediction of 20^{th} cycle end discharge capacity (EC) of doped $LiMn_2O_4$ cathode.

predicted values against the experimental values estimated by the RF-IC model and the RF-EC model, respectively. Compared to the ANN-IC model, the RF-IC model has more accurate predictions for both the training and testing sets with their RMSE values estimated to be 2.75 $mAhg^{-1}$ and 2.04 $mAhg^{-1}$ lower, respectively. In addition, the R^2 score of the RF-IC model is 0.18 higher than the ANN-IC, indicating more variations in the capacity are being captured from the test set. **Figure 5.14-c** gives the ranking of the importance of the covariates used for prediction from the RF-IC model based on the increase in the mean squared errors value. It is seen that the overall formula mass is estimated to be the most important covariates for the model prediction, followed by the formula ratio of manganese and current density respectively.

In addition to the EC prediction, the RF-EC model has the best predictions so far among all models with their RMSE values estimated to be the lowest among the previously discussed models at 13.32 $mAhg^{-1}$ and 12.93 $mAhg^{-1}$ for the training set and the testing set, respectively. RF-EC model has shown greater prediction power in the test set than the ANN-EC model as the plotted training and testing points are shown to be more aligned with the red true line. In particular, the predictions of the capacities over the range of 120 $mAhg^{-1}$ and 140 $mAhg^{-1}$ are better, as demonstrated from the smaller average deviation from the red true line. **Figure 5.14-d** gives the importance of the variables used in the RF-EC model. The electronegativity of the dopant is the most important variable followed by the current density, which is similar to that of the results from the RF-IC model. In addition, the percentage increase in the mean squared error values for the manganese ratio in the material formula and the materials formula mass is observed to be at a similar level which implies that both variables are equally important to the overall model prediction.

5.3.6 Gradient Boosting Model Regressor

During the optimisation of each gradient boosting machine model, 110 combination sets of the hyperparameters have been examined with the range indicated in **Table 5.1**. The gradient boosting models used for IC and EC prediction is

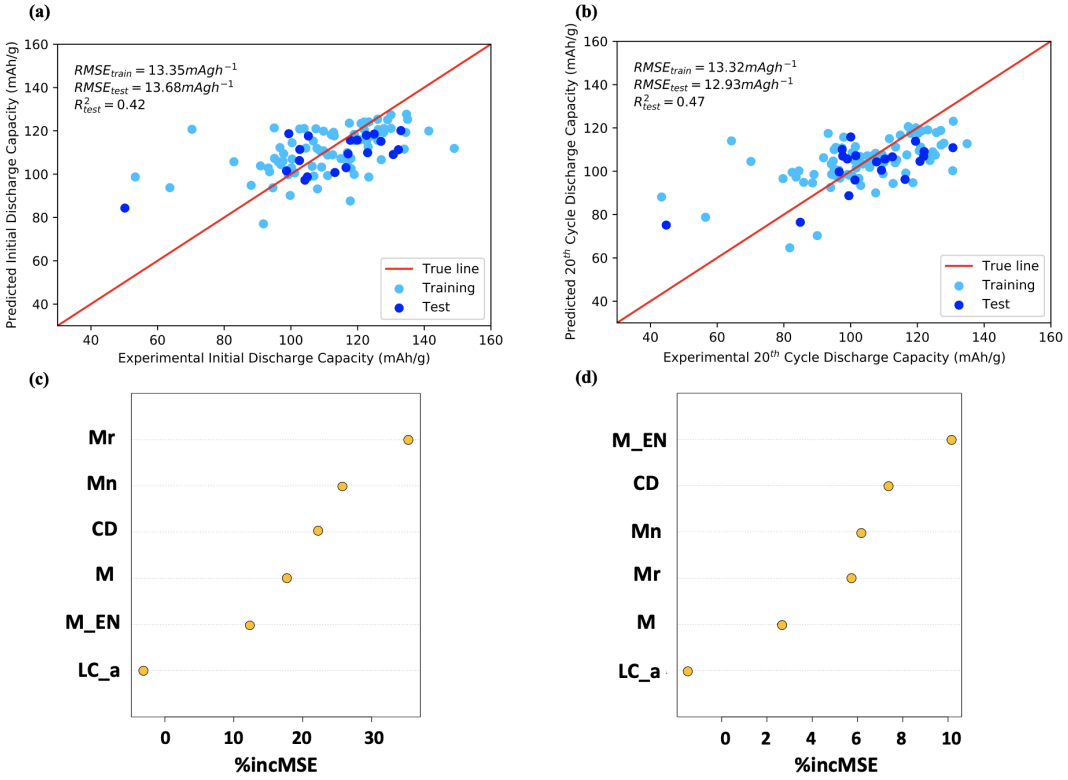


Figure 5.14: The predicted values against the experimental values computed by the random forest models for the prediction of (a) initial discharge capacity (IC) and (b) the 20th end discharge capacity (EC) of doped $LiMn_2O_4$ cathode and their respective ranking of the variable importance in terms of the percentage of increase in the mean squared error values estimated by (c)RF-IC and (d) RF-EC.

named GBM-IC and GBM-EC respectively. The selection of the optimal hyper-parameters was made based on which produces the lowest mean squared errors. To simplify the results chapter, the detailed optimisation results for this gradient boosting machine are available on the Github page (`https://github.com/thepowerligand/LMO-ML`). **Table 5.5** summarises the final optimised hyper-parameters for each gradient boosting model such as the number of iterations, interaction depth, the minimum number of samples remained at each node and the shrinkage rate. Based on the results, several observations can be made such as the optimal number of iterations for the GBM-EC model is much less than the GBM-IC model, whereas its shrinkage rate is nearly twice as much as for the GBM-IC model. This indicates that the contribution of each tree to the overall model prediction is

Table 5.5: The optimal hyperparameters for the gradient boosting models used to predict the initial discharge capacity (IC) and the 20th cycle end discharge capacity (EC) including the number of iterations, interaction depth, the minimum number of samples that remained at the node and the shrinkage rate applied

Parameter	IC	EC
Number of iterations	4082	386
Interaction Depth	10	4
The minimum number of samples remained at the node	5	5
Shrinkage Rate	0.0455	0.09907

heavier in the GBM-EC model than in the GBM-IC model.

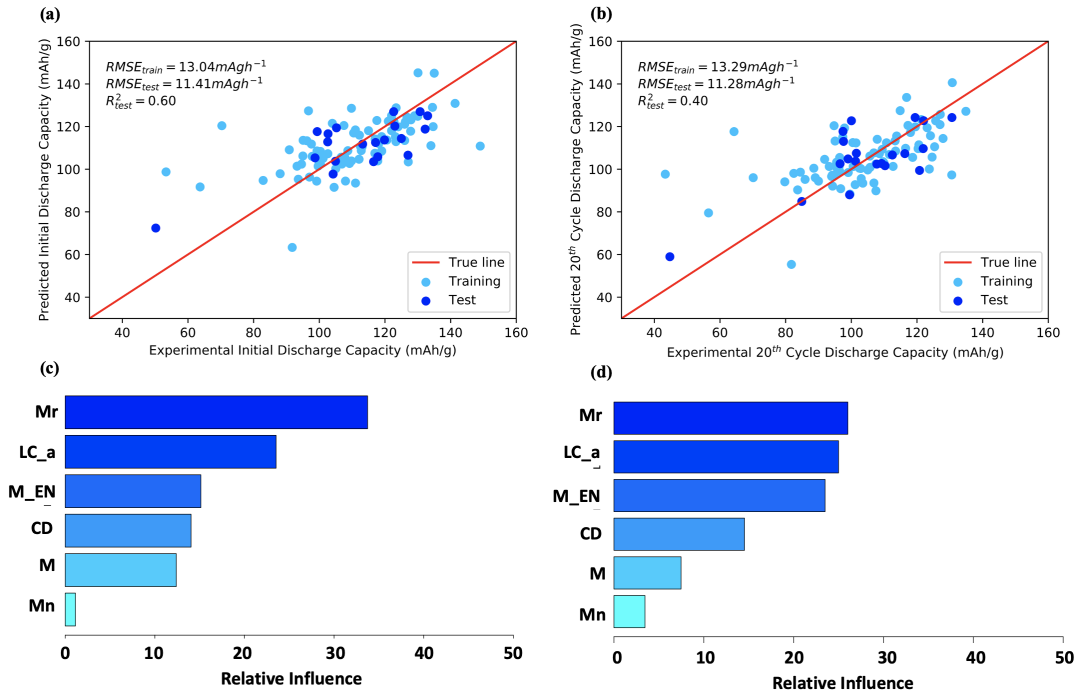


Figure 5.15: The predicted values against the experimental values computed by the GBM-IC model and GBM-EC model for the prediction of (a) initial discharge capacity, (b) 20th cycle discharge capacity of doped $LiMn_2O_4$ cathode and their respective variable importance (c) and (d).

Figure 5.15-a and **Figure 5.15-b** show the predicted values estimated by the GBM-IC model and the GBM-EC model against the experimental values for the prediction of IC and EC, respectively. With the exception of a few outliers, all the plotted points of the training and test set show little deviation from the true line, indicating that the predictions have reached a considerable agreement level with the experimental capacity. In addition, the overall predictive power of the

GBM-IC model is shown greater than the RF-IC model higher as its $RMSE_{train}$ and $RMSE_{test}$values are estimated to be 0.31 $mAhg^{-1}$ and 2.27 $mAhg^{-1}$. This good model prediction performance is also agreed with the R^2 score which is computed to be 0.60, which is calculated to be 0.18 higher than for the RF-IC model. In the case of the GBM-EC model, the overall prediction errors of the model are calculated to be 13.29$mAhg^{-1}$ and 11.28 $mAhg^{-1}$ respectively for the training and test set which are 0.03 $mAhg^{-1}$ and 1.65 $mAhg^{-1}$ lower than the RF-EC mode. In addition, the plotted dots (light blue and deep blue) from **Figure 5.15-b** are aligned well with the red true line, indicating that the overall prediction performance of the GBM-EC model is high across the whole dataset. In comparison to the RF-EC results graph (**Figure 5.14-b**), the GBM-EC model seems to have better prediction power over the lower range capacity at around 40 $mAhg^{-1}$ and the middle range capacity at 83 $mAhg^{-1}$ in the test set as the gaps to the red true line are noticeably smaller. Despite that the R^2 scores of the GBM-EC model are lower than of the RF-EC model, the overall prediction power is seen much higher, and thus the GBM-EC model is selected for further investigation.

Figure 5.15-c and **Figure 5.15-d** give the ranking of the variable importance computed by GBM-IC and GBM-EC respectively. The most important covariates to both models are the material formula molar mass, followed by the lattice constant "*a*", electronegativity of the dopant, current density and the ratios of dopant and the manganese in the material formula. A detailed discussion of these results will be given in **subsection 5.3.9**.

5.3.7 Support Vector Machine

Support vector machine models are optimised for the cost parameter which controls the regularisation strength and the gamma parameter from the radial basis function kernel function. Within the cost function, the value of the epsilon is kept at the default value of 0.1 across all model training. Two trials of hyperparameter optimisation are carried out to refine the investigating ranges indicated in **Table 5.1** and to optimise the cost and gamma values. The optimisation results for each model are summarised in **Table 5.6**. It is seen that the investigated ranges of both Cost and

Table 5.6: Different trials (T-1, T-2) of hyperparameter tuning results for the Support vector machine regressor models for the prediction of initial discharge capacity (IC) and the 20th cycle discharge capacity (EC) of doped $LiMn_2O_4$ cathodes, where Gamma is the parameter inside of the radial basis function kernel function and Cost is the regularisation strength and Min_MSE represents the minimum mean squared error obtained from the investigating range. The optimal values of these parameters are highlighted in red colour.

	IC			EC		
	T-1	T-2	Optimal	T-1	T-2	Optimal
Gamma	$10^{-3} \sim 10^{3}$	$10^{-2} \sim 10^{3}$	$10^{-0.279}$	$10^{-3} \sim 10^{3}$	$10^{-3} \sim 10^{2}$	$10^{2.343}$
Cost	$10^{-4} \sim 10^{4}$	$10^{-3} \sim 10^{3}$	$10^{-0.997}$	$10^{-4} \sim 10^{4}$	$10^{-3} \sim 10^{3}$	$10^{-2.397}$
Min_MSE (mAhg^{-1})	173.65	163.99	-	193.17	190.49	-

Table 5.7: The optimal values of the cost and the gamma for the support vector models for the prediction of initial discharge capacity (IC) and the 20th cycle end discharge capacity (EC) of doped $LiMn_2O_4$ cathodes and their numbers of support vectors as well as the epsilon values from the cost function.

Parameter	**IC**	**EC**
Cost	0.526	220.744
Gamma	0.101	0.00401
Number of support vectors	77	73
Epsilon	0.1	0.1

the Gamma parameter have not shrunk much from the first trial. In addition, it is observed that the minimum mean squared error values are seen to be reduced from the first to the second trial of investigation. For example, the estimated minimal error for the IC support vector machine model has seemed to be reduced from 173.65 mAhg^{-1} in the T-1 trial to 163.99 in the T-2 trial. Similarly, a reduction of 2.68 mAhg^{-1} of the minimum mean squared error is observed after the investigating ranges are refined from value space from T-1 to T-2 for EC support vector machine models. The final optimal cost and gamma values for the SVM-IC and SVM-EC model are given in **Table 5.7**

Figure 5.16-a and **Figure 5.16-b** gives the predicted capacity against the experimental value estimated by the SVM-IC model and the SVM-EC model respectively, for the predictions of IC and EC in the training and testing set. The prediction

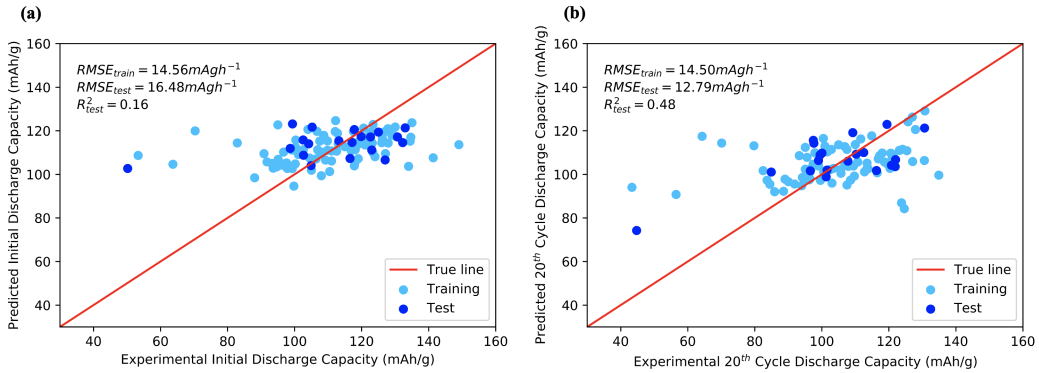


Figure 5.16: Predicted values against the experimental values computed by the support vector machine models for the prediction of (a) initial discharge capacity and (b) 20^{th} cycle discharge capacity of doped $LiMn_2O_4$ cathodes.

performance of SVM-IC is not as good as the GBM-IC model as the plotted points are further from the red true line. In addition, the $RMSE_{train}$(14.56 $mAhg^{-1}$) and $RMSE_{test}$ (16.48 $mAhg^{-1}$) are shown to be larger than the RMSE results from the GBM-IC model, implying that the model has poorer prediction power over the unseen hold-out test set. In addition, this poor prediction performance is reflected in the low R^2 score (0.16) which implies that the model can only capture 16% of the variations in the test set. Furthermore, the prediction of EC in both the training and the testing set by the SVM-EC model seems to be better as all the points shown to be more aligned towards the red line than the results of the SVM-IC model. However, the SVM-EC model seems to generalise more poorly over the training set than GBM-EC as its $RMSE_{train}$ is estimated to be 14.50 $mAhg^{-1}$ with the trend of the light blue dots observed to be around 30 degrees. Similarly, the predictions over the test set are also observed to be poor as all the plotted dots (in deep blue colour) are further from the red line.

5.3.8 Model Performance Evaluations

In this project, fourteen machine learning models are constructed with seven variations of the machine learning algorithms for the prediction of initial discharge capacity and the 20^{th} cycle end discharge capacity of the doped LMO cathodes. **Table 5.8** summarises the prediction performance of each model over both the training and testing set in the metrics of RMSE and R^2 values. To begin with, the ensem-

ble tree-based models such as the GBM and RF have shown much better prediction power over the test-set than linear methods such as the LR and RR. This could be due to the presence of many non-linear correlations between the covariates and the response variables as indicated in **subsection 5.3.2**. Among all models, the GBM model has shown the best performance in generalising for both the initial and 20^{th} cycle end discharge capacities, with the RMSE calculated to be the lowest at 11.41 $mAhg^{-1}$ and 11.28 $mAhg^{-1}$ from the test set prediction. In addition, R^2 scores are estimated to be 0.6 and 0.4 which suggest that the models can capture a sufficient amount of variations of the capacities from the test set. In general, the GBM algorithm is preferred for its robustness in predicting non-linear relationships, offering more flexible hyperparameter tuning options (minimum tree depth, learning rate, number of iterations) and does not require data pre-processing before the model training. Furthermore, the prediction power of DT is shown to be very low and this is due to its simple model structure that is not capable of capturing the complex variable relationships. On the contrary, the ANN model seems to perform better in both prediction tasks than the DT models, however, are not desirable for their issues of over parameterisation as indicated in **Table 5.3**. SVM models have demonstrated reasonable generalisation power over the training and test set, however are not still not comparable with the prediction power of the two GBM models. Both RF models have shown great predictive force over the test-set with their estimated RMSE values computed to be 13.68 $mAhg^{-1}$ and 12.93 $mAhg^{-1}$ for the IC and EC in the test set, respectively. However, the GBM models still showed more superior prediction power for the lower range capacities as discussed previously. At last, the prediction power of all penalised linear methods has shown very poor prediction power both for training and testing as their RMSE values and the R^2 scores are shown to be at the lowest among all. To conclude, GBM models have shown the best prediction on the prediction of both IC and EC and these models will be investigated further in the following section.

Table 5.8: Comparisons of the ten-fold cross-validated mean RMSE values, the RMSE value for testing against holdout test and the R^2 score for the test-set prediction computed by seven different types of machine learning algorithms including gradient boosting model (GBM), random forest (RF), artificial neural network (ANN), support vector machine (SVM), decision tree (DT), ridge regression (RR), LASSO regression (LR), for the prediction of the initial discharge capacity and the 20^{th} cycle end discharge capacity of doped $LiMn_2O_4$ cathodes.

	Initial Discharge Capacity ($mAhg^{-1}$)			**20^{th} 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	13.04	11.41	0.60	13.29	11.28	0.40
RF	13.35	13.68	0.42	13.32	12.93	0.47
ANN	16.10	15.72	0.24	15.48	14.58	0.32
SVM	14.56	16.48	0.16	14.50	12.79	0.48
DT	14.88	15.21	0.28	16.95	16.47	0.13
RR	14.34	17.49	0.05	14.60	17.35	0.04
LR	14.76	17.99	-0.001	15.21	17.63	0.01

5.3.9 Variable Importance Analysis

Most of the non-linear algorithms such as support vector machines and neural networks are known as "black box" algorithms as they do not provide any valuable insights into the internal function used for prediction. On the other hand, tree-based models, including the gradient boosting machine algorithm is more insightful in demonstrating the importance of the covariate variable in terms of their contributions to the overall prediction because their decision tree base learner has a high interpretable rule-guided structure. Since the gradient boosting models have demonstrated the greatest prediction power among all models for the predictions of IC and EC, this section will provide further analysis into the variable ranking estimated by the GBM models and giving an interpretation of the results in the chemical sense.

To begin with, both variable importance graphs computed by the GBM-IC model (**Figure 5.15-c**) and the GBM-EC model (**Figure 5.15-d**) have ranked the material formula molar mass as the most important covariate used in the prediction. For this, Faraday's law stated that the material molar mass and the number of reactive electrons/Li-ions in the electrochemical reactions are inversely proportionally

to the specific capacity of an electrode (the ideal amount of energy available for withdrawing per unit mass). In this regard, the variation in material molar mass would inevitability influence the practical capacities on both the first and the twentieth discharging cycle as they are quantitively related to the specific capacity. **Figure 5.17-a** demonstrates positive linearity trends for the material molar mass with both discharging capacities in the whole dataset and this suggests that the higher molar mass doped LMO formula can lead to higher values for IC and EC.

Next, the crystal lattice constant and the electronegativity of the dopant are the second and the third most important covariates from both gradients boosting models. **Figure 5.17-b** and **Figure 5.17-c** gives the 3D plots of the lattice parameter a and dopant's electronegativity against the colour coded initial capacities and 20^{th} cycle end discharge capacity, respectively. First, variations in the crystal lattice dimension of the doped LMO cathode would also influence the Li-ion site energy which is highly correlated to the discharging performance of the 20^{th} cycle. For this, previous studies have confirmed the important influence of the crystal lattice parameters on the discharging performance of manganese such as the doped spinel($LiMg_xMn_{2-x}O_2$) [287] and the finding was made such as the larger crystal lattice dimension often leads to a poor discharge performance. A clear trend can be observed as the higher IC and EC values (coloured by red) are more likely to occur at smaller lattice parameters ($\sim$8.15 Å). Next, the electronegativity of the dopant can modify the overall structural density of the doped LMO systems as it controls the bonding strength with the surrounding manganese ions and oxygen ions. A small difference in electronegativity can result in the formation of covalent bonds, which in turn makes the structure more stable and less dense than the structure that contains ionic bonds. In addition, the Li-ion site energy changes according to the modified structural density, and this would change the rate of intercalation/deintercalation of the Li-ions from the LMO cathode, resulting in various loading capacities at the first cycle.Based on the plotted 3D graphs, the results seem to suggest that using dopant with a moderate to lower range electronegativity value of ($1.4 \sim 1.8$) can slightly increase both the IC and the EC values. However, this

correlation is not conclusive due to the dispersed distribution observed across all electronegativity values.

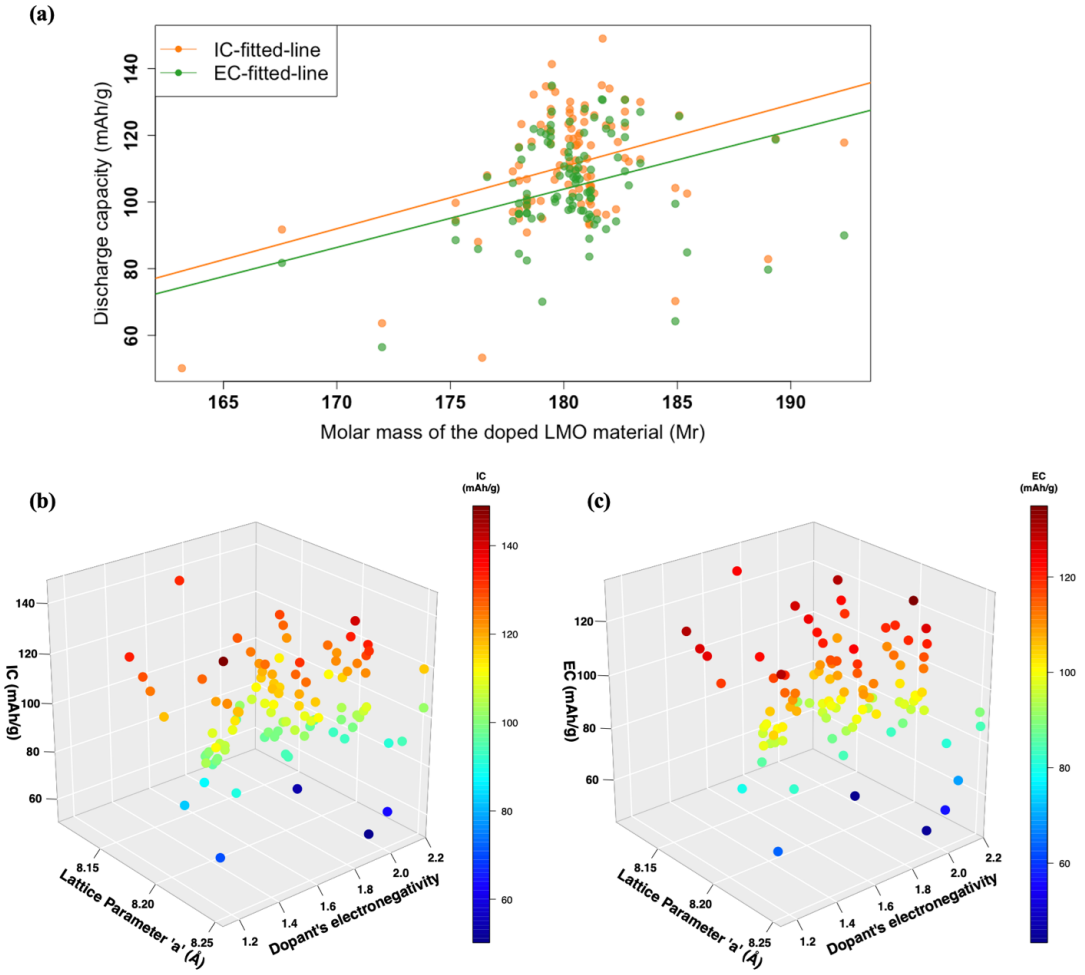


Figure 5.17: The plots of the crystal unit lattice parameter 'a' and the dopant's electronegativity against the values of (a) initial discharge capacities (IC) and (b) 20th cycle end discharge capacities (EC) of doped $LiMn_2O_4$ cathode.

5.4 Summary

In this chapter, seven different machine learning algorithms have been explored for their prediction power in describing the correlations of six different structural and elemental properties of 102 doped LMO systems and the corresponding discharging performance at the first and twentieth cycles. The dataset was manually curated from the literature with data collection rules implemented such as the standardised experimental conditions (i.e. electrolyte, anode, discharging environment temper-

ature). Among all the investigated models, the gradient boosting machine (GBM) models demonstrated the best predicting power for both the first and the twentieth cycle discharge capacities with the root mean square error (RMSE) for each being the lowest at 11.41 $mAhg^{-1}$, 11.28 $mAhg^{-1}$, respectively from predicting against the held-out test set. A variable importance study was conducted using the optimised GBM models and the results suggest a higher molar mass of doped LMOs system can lead to a higher discharge capacity both at the first and at the 20^{th} cycle. In addition, weak negative correlations are captured for the variable pair of the electronegativity of the dopant in LMO with IC and the crystal structure lattice constant-"*a*" with EC, respectively. The results of this study have demonstrated the great potentials in implementing ML algorithms to grasp the complex structure-property correlations of the doped LMO systems which could shorten the testing duration of battery material performance and leading a faster discovery speed of new doped cathode materials for Li-ion batteries with higher storage capacity and longer life cycle.

Chapter 6

Machine Learning Prediction of the Discharge Capacities of doped $LiNi_xCo_yMn_zO_2$ Layered Materials

6.1 Background

Among the various cathode candidates, layered cathodes have received tremendous market success owing to their high practical capacity and the wide operating voltage windows. Quinary oxides (e.g. $LiNi_xCo_yMn_{1-x-y}O_2$) are currently the state-of-art layered cathode materials as they integrate the superior properties of all three fundamental layered materials: $LiCoO_2$ (high kinetics), $LiNiO_2$ (high capacity), $LiMnO_2$ (high safety). **Figure 6.1-a** gives the ternary metal mix-phase diagram with the indication of the stable quinary oxide candidates 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) [24, 288, 289]. Each NCM cathode have different electrochemical performance due to the differences in the composition weighted role of each transition metals during the charge/discharge cycles. To illustrate this process, **Figure 6.1-(b,c,d)** gives the graphical illustrations of the change in the density of states of the transition metals (Ni, Co) in the NCM-333 system during the discharged state, half discharged state and charge state, respectively. During the initial charging period ($1 < x_{Li} <$

2/3) of $Li_xNi_{1/3}Co_{1/3}Mn_{1/3}O_2$, nickel ion acts as the main active component where the reduction reactions take place in the sequence of Ni^{2+}/Ni^{3+} redox couple, to Ni^{3+}/Ni^{4+}. In the further charging process subject to the lithium ratio of ($2/3 < x_{Li} < 0$), the cobalt redox couple of Co^{3+}/Co^{4+} will be responsible for driving the electrochemical reactions until the Fermi level reaches the top of the 2p band of oxygen (**Figure 6.1-d**), which in turn creates the voltage limit for the material. It is important to note that the manganese ions in the formula do not participate in any of the aforementioned redox reactions and are mainly used to stabilize the overall material crystal structure.

The full electrochemical potentials of NCM cathodes are still restricted due to the thermal instability issue during cycling, short cyclic life, and additionally the poor rate capability from the low electronic conductivity [25, 26, 290]. These inadequate performances are caused by the degradation mechanisms found in the crystal structure (cation mixing, oxygen release) and on the material surface (microcrack formation) during cycling. During the Li-ion extraction process, the cation mixing effects occur as Ni^+ ions can spontaneously migrate into vacancy sites in the lithium layer due to the low energy barrier and their similar ionic radius (Ni^{2+}: 0.69 Å and Li^+: 0.76 Å) [291]. This cation mixing would cause anisotropic stress to the local crystal structure, reducing the Li-ion diffusion coefficient and lead to irreversible capacity loss. In addition, the repulsive force between the oxygen and the transition metal layer tends to increase from the repetitive charging of the NCM materials. This effect would lead to the large expansion of the crystal volume along the c-axis accompanied by loss of oxygen, causing undesirable phase transformation from the hexagonal phase into the monoclinic phase and severely degrading the materials' electrochemical performance. As a side effect, the increasing material strain and the changes in the crystal volume tend to facilitate the formation of microcracks on the electrode surface and this severely degrades the material mechanical strength and raising safety concerns for cycling. **Figure 6.2** compares the degree of microcrack formation for various NCM materials and it is found that the nickel rich NCM tends to have more severe microcracking after repetitive charging/discharging

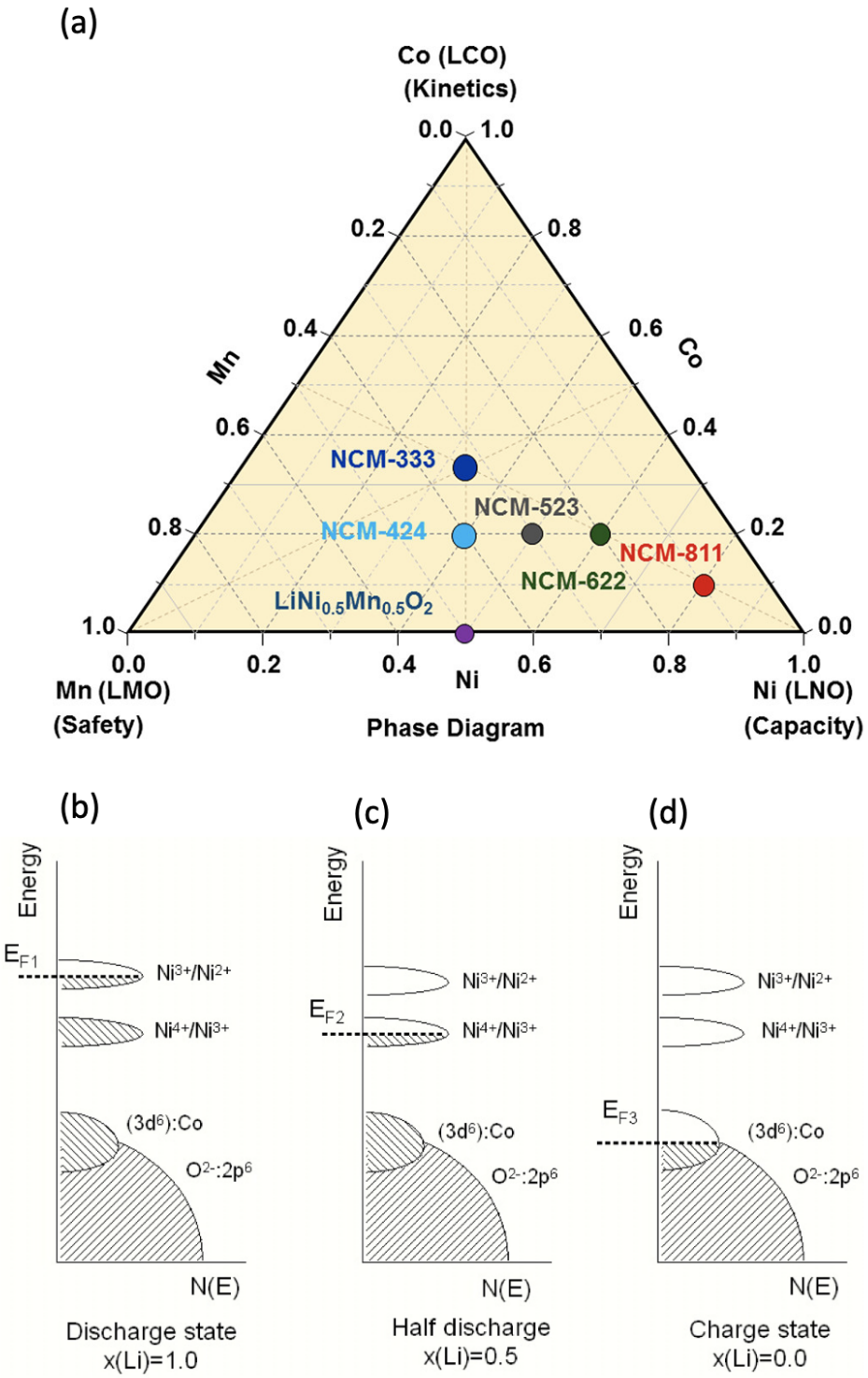


Figure 6.1: (a) The ternary transition metal mix-phase diagrams for the boundaries of lithium nickel oxide (LNO), lithium manganese oxide (LMO) and lithium cobalt oxide (LCO) with the identifications of the NCM materials and the schematic illustration of the energy and density of states for NCM-333 compound at (a) discharged state, (b)half discharged and (c) charged state. Image (a) is reproduced with permission from [24]. Image (b)-(d) are reproduced with permission from [25].

than the low nickel content formula. The increase in the degree of microcracks on the material surface greatly increases the active surface area which promotes undesirable side reactions with the electrode surface and cause great Li-ion losses and, capacity decay as well as the formation of impurity phases (NiO-like as given in **Figure 6.2**).

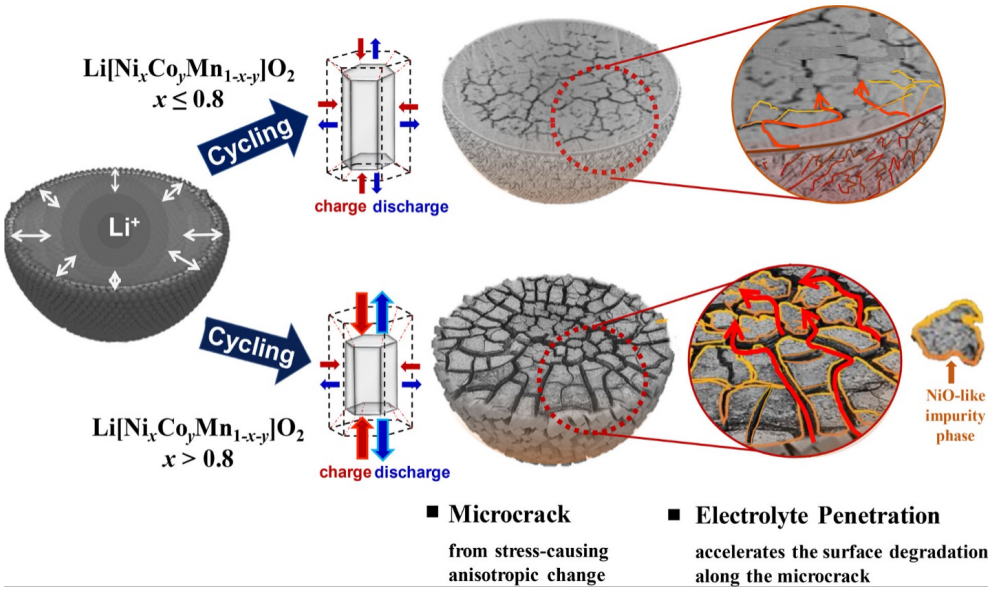


Figure 6.2: Comparisons of the degree of microcrack formation for NCM with low nickel content and high nickel content. Reproduced with permission from [26].

To circumvent the crystal structure-related issues, a wide range of interest is drawn toward applying doping technique to alter the intrinsic properties of the NCM cathodes with the aim of improving their electrochemical performance. From the literature, common doping benefits for the NCM structure are found in the increase in the strength of the oxygen-TM bond and additionally preventing the oxygen release during cycling, mitigating the cationic mixing effect by reducing the migration of the Ni^{+} into the lithium layer and lastly, promoting the Li^{+} transportation and improve both the electronic and the ionic conductivity of the materials. A wide range of doping elements has been investigated in recent years and they have shown great promises in improving the electrochemical performance, such as Al [292], Fe [292], Cu [293], Cr [294, 295], Mg [296], Mo [297], K [298], Pb [299], Ti [300], Si [301], Sn [302]. Nevertheless, obstacles such as the diverse doping principles for different element, different doping sites (Li, Ni, Co, Mn) and the large compositional space

remain as challenges toward finding suitable dopants. The conventional approach to characterise the electrochemical properties of a new doped system is through conducting repetitive experiments which is costly and time-consuming. Another approach based on first-principle computational modelling is also hindered by the expensive computing cost for studying very large super-cell systems. To conquer these shortcomings, this chapter implements various machine learning techniques to investigate the linkages among various doping factors and the experimental cyclic performance of doped NCM cathodes. In this work, 168 distinctive NCM doped systems have been carefully collected covering 20 variations of dopants for all doped NCM -derivate material classes (NCM-333, NCM-523, NCM-622, NCM-811). Firstly, Pearson's correlation coefficient study was performed to investigate the co-linearity of every variable pair. Furthermore, six non-linear algorithms, including gradient boosting machine, random forest, kernel ridge regression, feedforward deep learning, k-nearest neighbours and support vector machine were implemented with the model design given in **Table 6.1** to predict the initial discharge capacities and 50^{th} cycle discharge capacities of the doped NCMs based on twenty covariates (e.g. material characterization results, experimental parameters, elemental properties). By comparing their electrochemical performance against a held-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 discharging performance of the doped NCM systems. These insights should greatly enhance the current understanding of the doping effects and facilitate the design of future experimental work (e.g. independent variable selection, the doping concentrations).

6.2 Methods

6.2.1 Data Collection

During the initial stage of publication filtering, strict rules were applied to ensure that collected data reach a high consistency. These include that the NCM materials

Table 6.1: The proposed ML model including the name 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**, the **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
The volume of the unit cell	CV	**Dopant**'s number of electrons	No_electron_M
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 ionisation energy	E_ion_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	50^{th} cycle end discharge capacity	EC

should be i) single-doped with cation ions, ii) be single-phase iii) have a space group of R-3m; and iv) have no surface coating. Furthermore, the electrochemical testing should also fulfil the following criteria to meet the data collection requirements: i) performed at least 50 cycles of charging/discharging cyclic test, ii) used lithium foil as anode and aqueous $LiPF_6$ as electrolyte iii) applied constant current density for charging and discharging the battery; iv) the cyclic tests are carried out under the atmospheric conditions (i.e. Temperature = 25 $\pm$ 5 ℃, Pressure = 1atm). The cycle of fifty was selected as this is the most commonly performed cycle across different studies on the doped NCM cathodes. A graphical illustration of the publication selection process is given in **Figure 6.4**. After applying these rules, the initial 524 publications were reduced to 59 covering 168 different doped NCM 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, Zr) and these were published in the years between 1998 and 2020.

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 mixture system with solvent and additives of which could lead to different performances if not standardised. The types of electrolyte systems from our collected studies are summarised in **Figure 6.3**. In our dataset, nearly 71% of the investigating electrochemical test were performed from either the mixture of ethylene carbonate/dimethyl carbonate (vol % 1:1) or the ethylene carbonate/dimethyl carbonate/ethyl methyl carbonate (vol% 1:1:1). These systems have similar dielectric constants (**Table B.1**) which should result in similar electrochemical performance. Only seven of the material systems have been tested with the addition of fluoroethylene carbonates and these are used to improve the battery operation safety and hence should not influence the overall data quality by a considerate amount.

6.2.2 Model Outlook and Training

The ML models used in this work were trained using Python programming language and with its relevant ML libraries (Sciki-learn, Pandas). Within the model, 20 co-

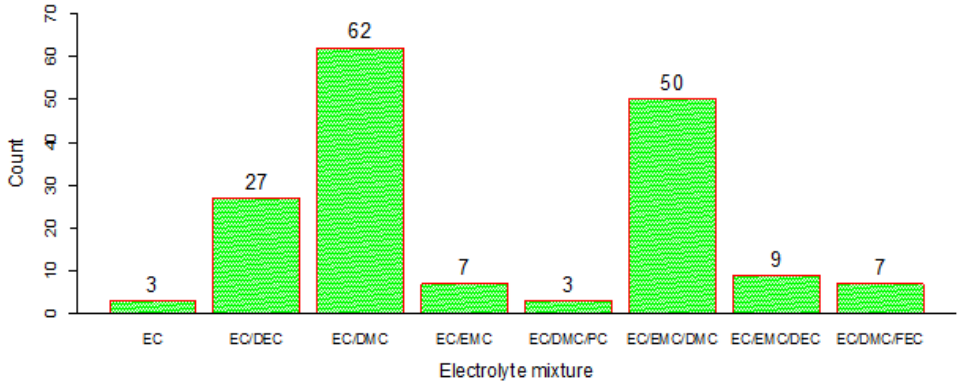


Figure 6.3: The bar plot of the number of the electrolyte mixture (excluding the LiPF6 salt component) used in the studies from the dataset, where EC stands for Ethyl carbonate, DMC stands for Dimethyl carbonates, DEC stands for diethyl carbonates, EMC stands for ethyl methyl carbonates, PC stands for Propylene carbonate, FEC stands for fluoroethylene carbonates.

variate variables are selected to predict the initial and 50^{th} 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, doping element in the material formula (Li, Ni, Mn, Co, M), 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 limit (V_min, V_max). Since this project aims to reveal the important doping features that control the performance of the doped system, seven dopant elemental properties are chosen as covariate variables. These include the dopant's molar mass; the number of electrons; electronegativity; electron affinity; first ionization energy; atomic radius and ionic radius. In this work, six non-linear algorithms were implemented including artificial neural network (ANN), random forest (RF), gradient boosting machine (GBM), support vector machine (SVM), kernel ridge regression (KRR), k-nearest neighbours (KNN).

The whole data space was randomly split into the ratio of 4:1 correspond to the model training set and test set, respectively. Within the training set, the model training process was carried out using the 5-fold cross-validation method to optimise the embedded hyperparameters. **Table 6.2** summarises the hyperparameters considered

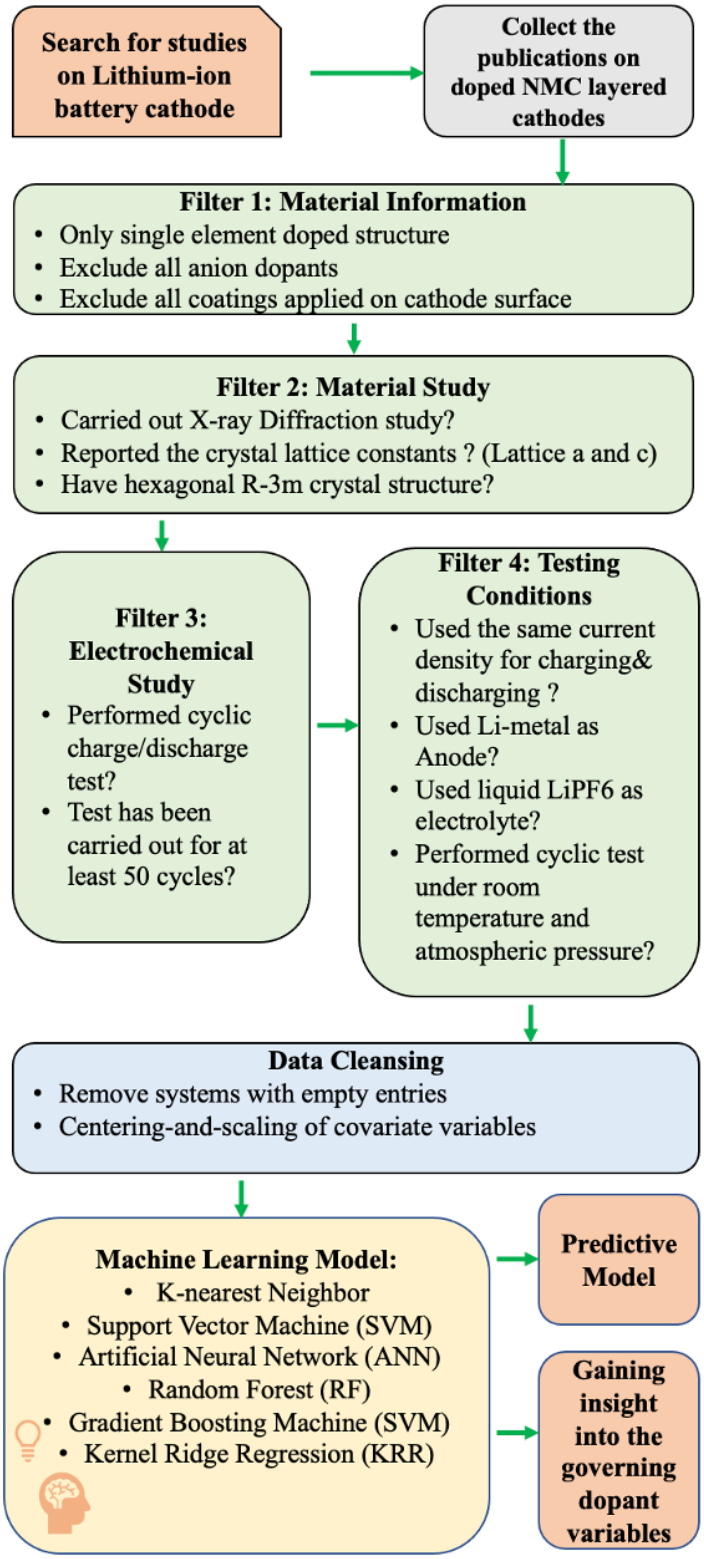


Figure 6.4: An overview of the data collection process with the demonstration of every condition applied in each publication filtering stage for doped nickel-cobalt-manganese oxide layered cathode.

for each of the algorithm and their value range of investigation in this project. In addition, to ensure that the results are reproducible, the same random-seed is assigned to all machine learning models.

Please note, the data curated for this chapter and the written machine learning codes can be found on the GitHub page (`https://github.com/thepowerligand/NCM-ML`). The detailed references are provided for each set of experimental data collected, in the same GitHub directory.

6.3 Results and Discussion

6.3.1 Covariate Selection for the Model

The covariates used to construct the model in this project can be categorised into three classes. These include the curated experimentally reported results such as the material compositions, crystal structure parameters (crystal lattice dimension and volume), cyclic test conditions (charge/discharge current density, operating voltage) and finally, the elemental properties for both the intrinsic material elements and the doping element in the structure. To begin with, the compositions of the lithium, nickel, manganese, cobalt and the dopant in the layered cathode are chosen as they determine the reaction kinetics, structural stability, and overall capacity, respectively. The crystal lattice dimensions and lattice volume reveal the conditions of the material's crystal structure and can control the mechanisms of the Li-ion intercalation/deintercalation during cycling. In addition, these material physical properties can be used to represent the differences in the synthesis routine implemented as well as the precursors. The common synthesis methods for NCM cathode such as the co-precipitation methods, the sol-gel method and the solvothermal method are included in this project. Furthur to the data collection for the cyclic performance results, the corresponding operating voltages and the current density (charging and discharging) are collected as they account for the number of available Li-ions for withdrawal in the material and the rate of Li-ion transfer within the battery system. The cases of using extreme experimental conditions such as extremely high or low current density and of using different current densities for charging and discharging

Table 6.2: A summary of the hyperparameters considered in this project for each of the algorithms, along with their investigating value ranges. Note, the random forest and gradient boosting machine share many common hyperparameters and the stared hyperparameter only applies to the gradient boosting machine model.

Algorithm name	**Hyperparameter**	**Investigating range**
Artificial Neural Network (Feed-forward with Relu activation function and linear output rule)	Number of hidden layers	1∼20
	Number of neurons in each hidden layer	1∼20
	Learning Rate	0.001, 0.01, 0.1
Kernel Ridge Regression	Kernel function	Laplacian, Radial Basis Function (RBF)
	Gamma (RBF)/sigma (Laplacian)	$10^{-4} \sim 10^{4}$
	Lambda	$10^{-4} \sim 10^{4}$
Support Vector Machine (Radial Basis Function)	Cost	$10^{-4} \sim 10^{4}$
	Gamma	$10^{-3} \sim 10^{3}$
K-nearest Neighbour	Number of Neighbours	1 ∼20
Random Forest and Gradient Boosting Machine	Number of trees	100∼5000
	The minimum amount of sample considered for each split	1∼15
	The minimum amount of samples remaining at each leaf	1∼15
	The maximum amount of covariates considered at each split	1∼19
	The maximum tree depth	1∼15
	Learning Rate*	$10^{-4} \sim 10^{-1}$

are discarded to reduce the noise in the dataset. In the end, all the relevant dopant's properties are selected to enhance the prediction power of the ML models. For instance, the dopant ion's ionic size could mitigate the migrations of the Ni^{+} into the Li^{+} layer. The dopant's electronegativity would influence the type of bond-forming inside of the material structure and hence the overall stability. Likewise, the ionisation energy and the electron affinity would affect the change in the systematic energy during the reactions of Li-ion insertion and the extraction and hence determine the structural stability. The importance of these covariate variables in the final model prediction will be examined in the later results section.

6.3.2 Preliminary Statistical Analysis for the Dataset

To gain initial insight into the underlying variable correlations, a Pearson's correlation coefficient study was performed for every pair of variables in the dataset. **Figure 6.5** shows the matrix of correlation values (R) calculated for the twenty covariates and the two response variables. The extent of correlation between every pair is colour-coded with darker orange indicating a strong positive correlation and dark green strong negative correlations.

From the computed R values, it can be seen that there are more strong correlations ($R > 0.75$) being observed between covariates variables than for the covariate variables with either of the two responsive 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 voltage values are often pre-set for the experiments based on 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 the direct TM crystal site substitution. For the model construction, it is

important to find the linkage of covariate variables to the electrochemical properties. First, no strong linear correlations are being observed between covariate variables and response variables which might suggest that the presence of many non-linear correlations in the dataset. 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 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 observation seems to agree with the latest studies showing that the lithium-rich layered cathode ($\sim$200mAhg^{-1}) can provide much higher discharging capacities than the normal NCM-111 compounds ($\sim$165mAhg^{-1}) [303, 304].

6.3.3 Artificial Neural Network

The selection of the number of epochs is made by first running the ANN model up to a high amount (10000) and then choose the epochs number where the five-fold cross-validated prediction errors are reaching a steady pace and are no longer reducing after this point. **Figure 6.6** illustrates the graphical results of the prediction loss against the number of epochs tested in each prediction task. From both graphs, it is seen that the prediction losses seem to stabilise at around 1000 and this was then selected to be the number of training epochs in both models. In addition, the hyperparameter tuning was carried out with various combinations of the neural network hyperparameters including the number of neurons per hidden layer, the number of hidden layers and the learning rate. **Figure 6.8** gives the heatmap results of the prediction performance for different sets of hidden neuron number and hidden layers number under the selected learning rates of 0.1, 0.01 and 0.0001, for the prediction of initial capacity (**a, b, c**) and the prediction of the 50$^{\text{th}}$ cycle discharge capacity (**d, e, f**). First, it is observed that the general mean squared error values across the entire heat maps in **Figure 6.8-c** are shown to be less than 1000mAhg^{-1} (deep blue colour) which are seemingly much lower than the results obtained from **Figure 6.8-a** and **Figure 6.8-b**. This suggests that the lower learning rate (0.001) can reach better overall accuracy than the higher learning rates such as 0.1 and 0.01. Similar observations can be made for the heat map results for the EC model. The lowest

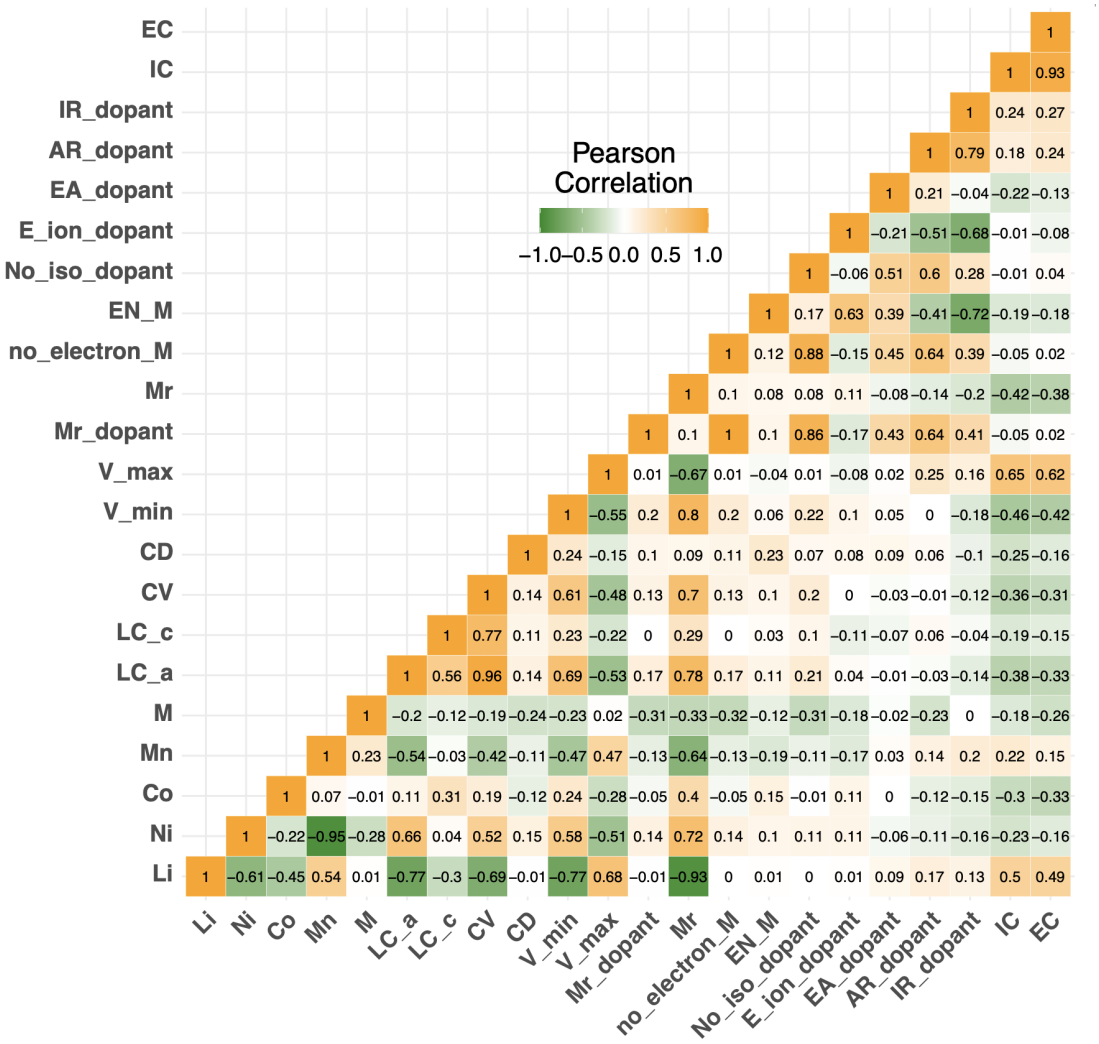


Figure 6.5: Results matrix of Pearson coefficient correlations for every pair of variables in the dataset, including covariate variables: Li, Ni, Co, Mn, M, LC_a, LC_c, CV, V_min, V_max, CD, Mr, Mr_dopant, No_electron_M, EA_dopant, No_iso_dopant, AR_dopant, IR_dopant and with response variables like the initial (IC) and the 50^{th} cycle discharge capacity (EC) of doped nickel-cobalt-manganese oxide layered cathode.

learning rate graph (**Figure 6.8-f**) shows to have the most amount of hyperparameters combinations among the three results graphs with the mean squared error less than 1000 $mAhg^{-1}$. Learning rate indicates the size of the search step for global minima and the results presented indicate that the high learning rate could easily miss capturing the optimal point, resulting in less accurate prediction. In addition, the optimal hyperparameters of ANN models for the IC prediction and the EC prediction are chosen from **Figure 6.8-c** and **Figure 6.8-f**, respectively and **Table 6.3** summarises the hyperparameters used for constructing the ANN models.

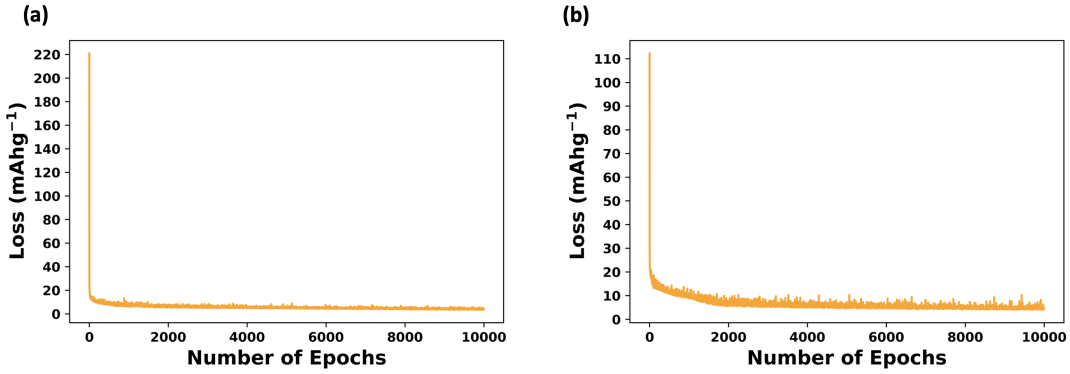


Figure 6.6: The graphs of prediction loss versus the number of epochs computed by (a) the ANN initial discharge model and (ANN-IC) (b) the ANN 50th cycle discharge capacity model (ANN-EC) for doped nickel-cobalt-manganese oxide layered cathode under the hyperparameter setting of learning rate of 0.001, hidden layers of 5 and hidden neurons of 18.

Table 6.3: The optimised hyperparameters for building the ANN models for the prediction of the initial discharge capacity (IC) and the 50th cycle end discharge capacity (EC) of doped nickel-cobalt-manganese oxide layered cathode.

	Learning Rate	**Hidden neurons per layers**	**Number of the Hidden layers**	**Total number of Parameters**
ANN-IC	0.001	18	5	1680
ANN-EC	0.001	20	8	3229

The prediction power of the ANN models with the selected optimal hyperparameters is investigated by using them to predict for the hold-out test set. **Figure 6.7** demonstrate the predicted results of the initial discharge capacities (**Figure 6.7-a**)

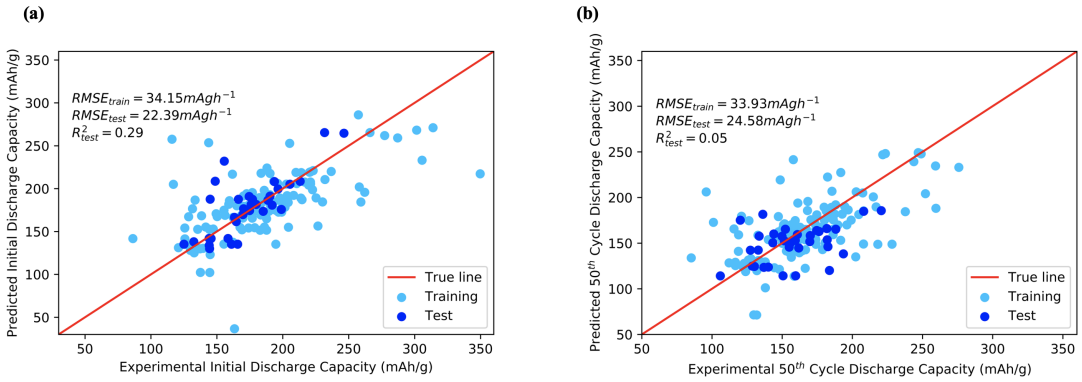


Figure 6.7: The predicted values versus the experimental values for (a) the initial discharge capacity (IC) and (b) 50$^{\text{th}}$ cycle end discharge capacity (EC) of doped nickel-cobalt-manganese oxide layered cathode, computed by the artificial neural network models, ANN-IC and ANN-EC, respectively.

and the 50$^{\text{th}}$ cycle discharge capacity (**Figure 6.7-b**) from the respective ANN models against the experimental observed values in the training set and the test set. It is seen that the ANN model was able to explain more of the variance within the test for initial discharge capacity than for the 50$^{\text{th}}$ cycle discharge as the R-value was shown to be 0.24 higher in the previous case. The estimated test-set RMSE values for both the IC and the EC predictions are seen to be reduced largely from their respective training sets which suggests that the model performed better in the test-set prediction than in the training set. In addition, the ANN-IC model seems to have less power prediction power for the higher range of capacity (at around 250 mAhg^{-1}) in the test set. On the other hand, the ANN-EC have shown better prediction power over a higher range of capacity (200 $\sim$280 mAhg^{-1}) in the test set as the respective points are closer to the true line. Furthermore, it is observed that both models have insufficient prediction powers for lower range capacities ($\sim$150 mAhg^{-1}) as the points are shown the furthest to the true line than the rest of them. These poor prediction powers could be due to the overfitting issues as the overly parametrised ANN models (**Table 6.3**) often require a much larger training dataset (10 times the parameters) to fully capture and learn the underlying relationships between the covariate and the response variable.

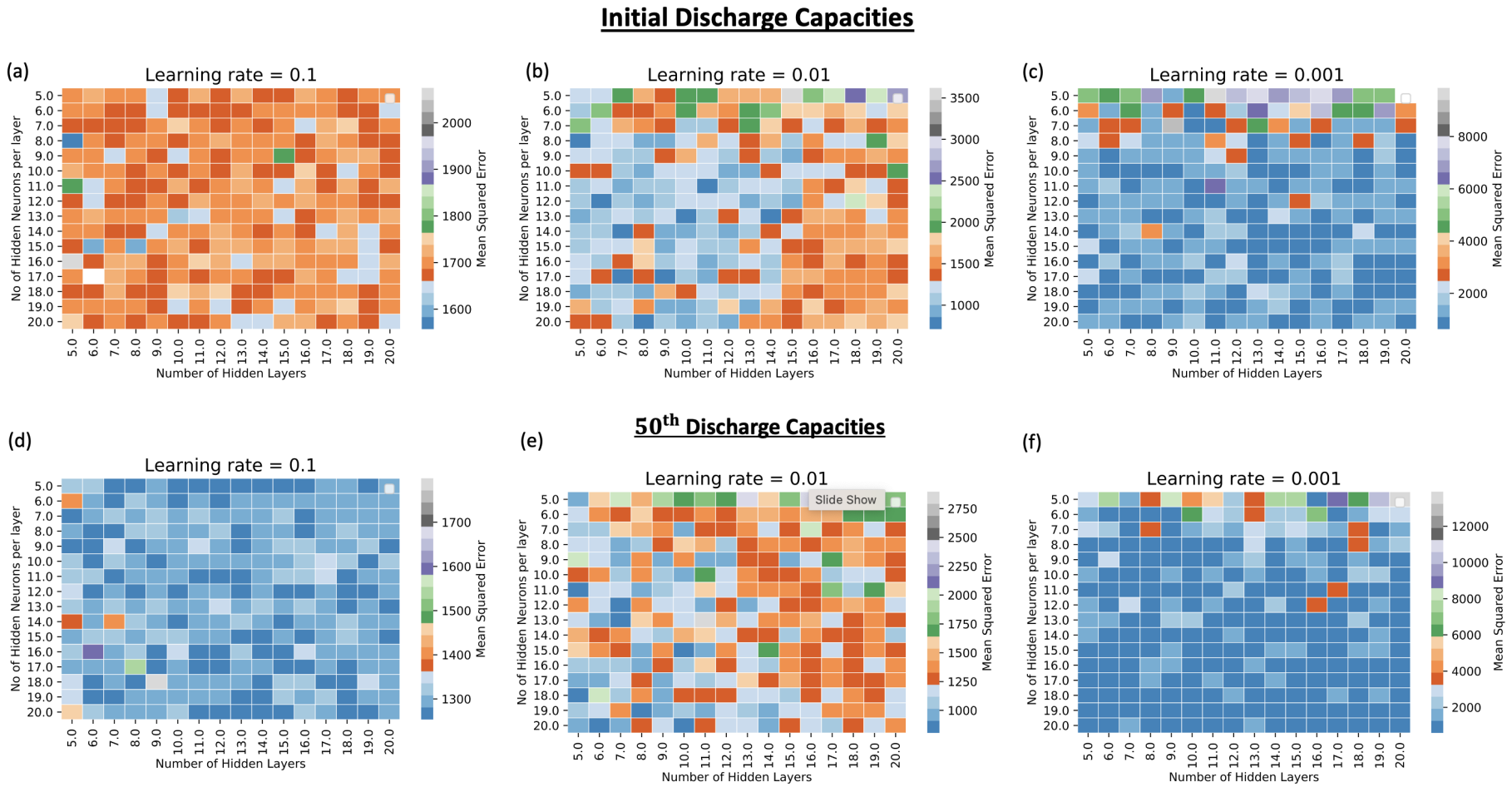


Figure 6.8: The heat-map representation of the hyperparameter optimisation results of artificial neural network for the prediction of (a)-(c) initial discharge capacity and (d)-(f) the 50th cycle discharge capacity of doped nickel-cobalt-manganese oxide layered cathode.

6.3.4 Kernel Ridge Regression

The hyperparameters optimised for the kernel ridge regression models are the choice of the kernel function, the embedded parameter for the selected kernel functions and the lambda values which represent the regularization strength from the I2-ridge regularisation. Laplacian and the radial basis function are two chosen investigated kernel functions in this study as they both have demonstrated superior prediction power in the literature. The additional model optimisations will be carried out for the sigma parameter embedded in the laplacian function and the gamma value embedded in the radial basis function. The equations of each parameter are given in the methods chapter.

Figure 6.9 provides the graphical illustrations on the mean square error calculated using Laplacian kernel function (a, b) or using the Radial basis function (c,d) with various lambda setting from 2 trials runs for the prediction of IC. Similarly, **Figure 6.10** provides the optimisation results for the EC prediction. To begin with, it is seen that the results graphs each respective kernel functions are shown very similar trend for both the IC and EC predictions. Comparing the results graphs of the first trial (T-1) in **Figure 6.9** for the two kernel functions, the Laplacian kernel function seems to generate smaller regions corresponds to low errors (coloured in purple) than the radial basis function. Similar observation can also be made in **Figure 6.10**. In general, the radial basis function kernel function seems to have less predictive power than the Laplacian kernel function as the mean square error values at the minimum point are relatively higher. Under the use of the Laplacian kernel function, the optimal hyperparameters that result in the lowest prediction errors are selected from **Figure 6.9-c** and **Figure 6.10-c**, for the prediction of IC and EC, respectively and are summarised in **Table 6.4**.

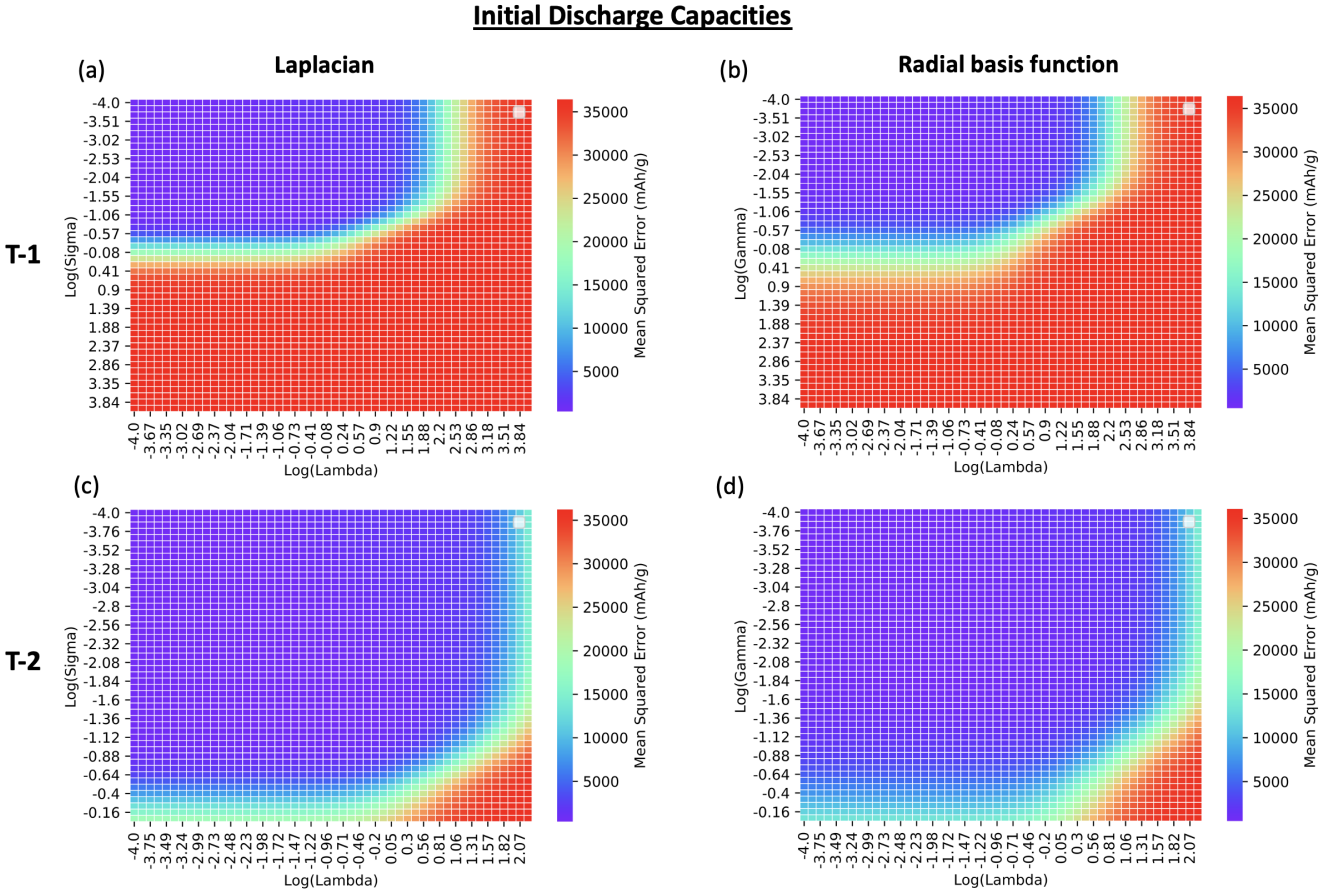


Figure 6.9: Mean squared error ($mAhg^{-1}$) versus the gamma values for different alpha parameters with the kernel functions of Laplacian and radial basis function computed by the kernel ridge regression models for the predictions of the initial discharge capacities of of doped nickel-cobalt-manganese oxide layered cathodes. T implies the trial number of hyperparameter tuning.

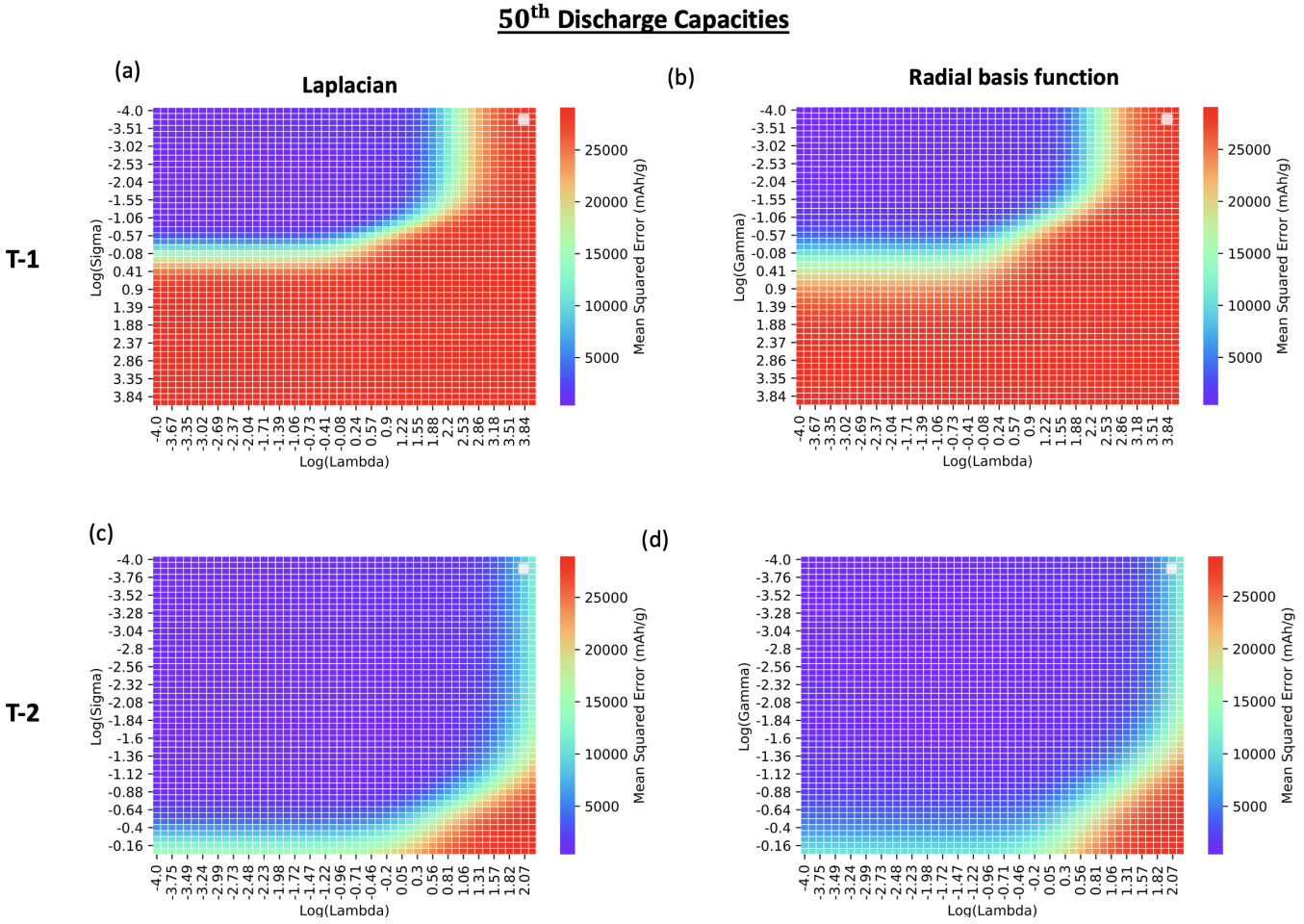


Figure 6.10: Mean squared error ($mAhg^{-1}$) versus the gamma values for different alpha parameters with the kernel functions of Laplacian and radial basis function computed by the kernel ridge regression models for the prediction of the 50^{th} cycle discharge capacities of doped nickel-cobalt-manganese oxide layered cathodes. T implies the trial number of hyperparameter tuning.

Table 6.4: The optimised hyperparameters for the kernel ridge regression models for the prediction of initial discharge capacity and the 50th cycle discharge capacity of doped nickel-cobalt-manganese oxide layered cathodes.

	Kernel	**Sigma**	**Lambda**
IC	Laplacian	0.0120	0.01058
EC	Laplacian	0.00191	0.00791

Next, the prediction power of these optimal KRR models is tested using the hold-out test set. **Figure 6.11-a** and **Figure 6.11-b** demonstrate the predicted discharge capacity against the observed discharge capacity for the prediction of IC and EC, respectively. It is seen that the KRR model used for the initial discharge capacity prediction performs much better than the previous ANN-IC model with its root-mean-squared-error calculated to be 5.11 mAhg^{-1} lower for the test set and the respective R^2 scores to be 0.29 higher. This higher prediction power can be explained as the predicted capacity reaches a higher alignment with the experimental capacity at both the middle range around 150mAhg^{-1} and the higher capacity range around 240 mAhg^{-1} of the test set in **Figure 6.11-a**. For the prediction EC, the KRR-EC model has seemed to outperform the ANN-EC model with better generalisation over both the training and test set with both the RMSE values for training and testing calculated to be lower. In general, both KRR models have shown great prediction power over the two capacity variables.

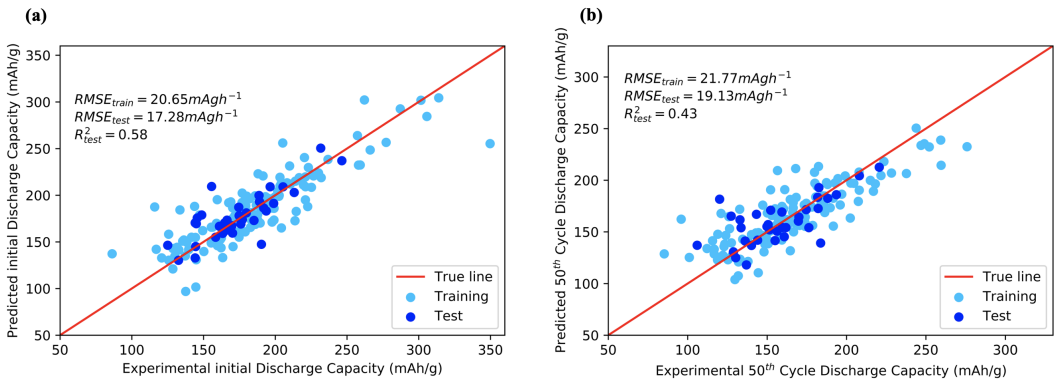


Figure 6.11: The predicted values computed versus the experimental values for (a) the initial discharge capacity (IC) and (b) 50th cycle end discharge capacity (EC) of doped nickel-cobalt-manganese oxide layered cathodes computed by the kernel ridge regression models, KRR-IC and KRR-EC, respectively.

6.3.5 Support Vector Machine

The radial basis function is used as the main kernel function for building support vector machine regression models. The targeted hyperparameters for these models are the cost parameter which controls the strength of the L2 Lasso regularization and the gamma parameter from the kernel function. The optimisation of these hyperparameters was carried in two rounds with each narrowing down the investigating range based on the calculated prediction errors. **Figure 6.12** demonstrates the results graphs of each round of hyperparameter optimisation for the initial discharge capacity (a, b) and the 50^{th} discharge capacity (c, d), respectively. **Figure 6.12-a** and **Figure 6.12-c** correspond to the results from the first round of hyperparameters for the cost and gamma parameters with the investigating range given in **Table 6.2**. It is clear to see that any parameter combinations with gamma value larger than $10^{0.92}$ and costs smaller than $10^{-0.08}$ are undesirable as they are given in grey colour (high prediction error) in both graphs. In addition, the optimal hyperparameter regions represented by the royal blue colour are found to be very similar for both the IC model and the EC model with the gamma and the cost ranges identified to be $10^{-2} \sim 10^{-1}$ and $10^{2} \sim 10^{3}$, respectively. We further analysed these hyperparameter spaces and obtained the results graphs shown as the **Figure 6.12-b** for the initial capacity prediction and **Figure 6.12-d** for the 50^{th} cycle discharge capacity prediction. It is observed that the optimal hyperparameter region is more unified for the IC case and shown as one whole region. On the other hand, the EC model has demonstrated two clusters of optimal hyperparameter regions that contribute to low prediction errors (coloured in deep blue). However, the optimal hyperparameter set from the cluster with lower cost (log(2.0∼2.49)) and smaller gamma (log(-1.76∼1.39)) computes much lower error than the optimal hyperparameter set from the other cluster and hence this was selected as optimal set. In the end, the final selected hyperparameters for building the support vector machine regression models in this study are given in **Table 6.5**.

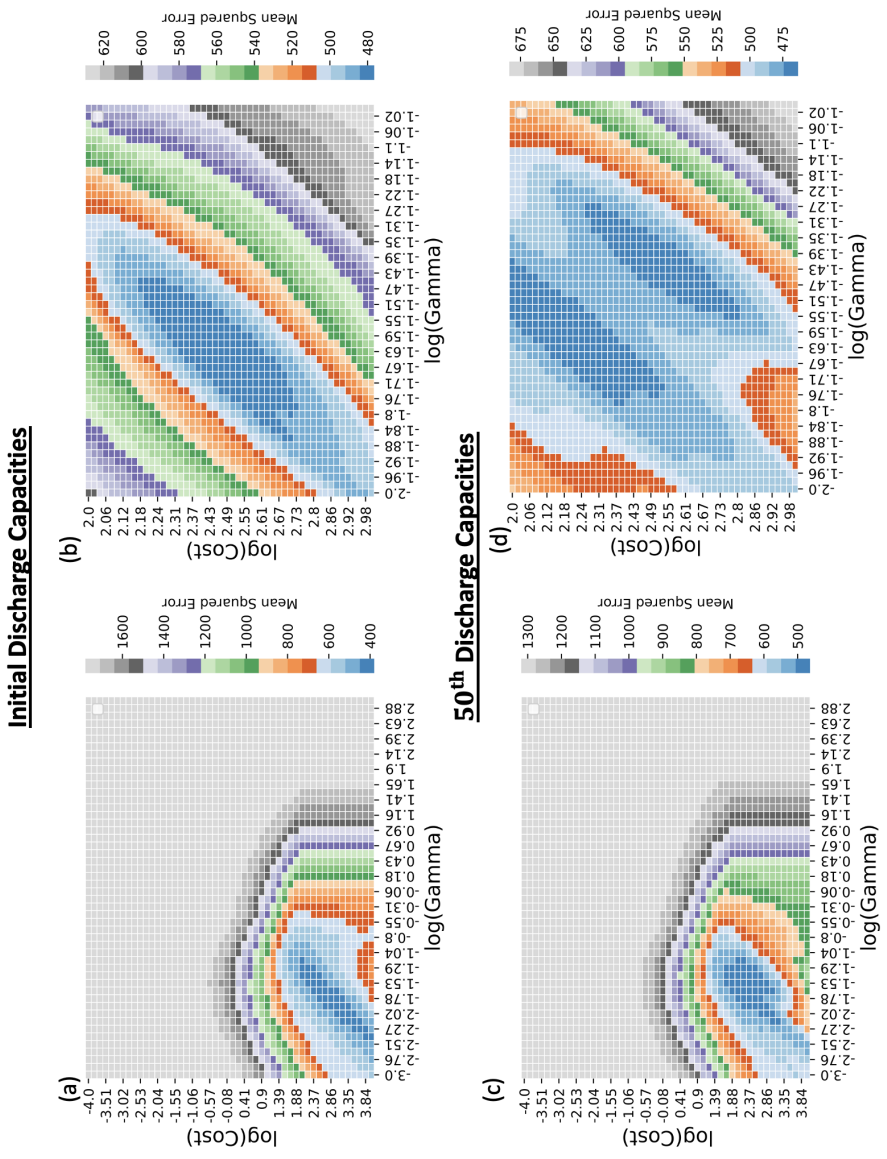


Figure 6.12: The heat-map illustration of the optimisation results for the cost and the gamma parameters in logarithmic scale for the prediction of (a-b) initial discharge capacities and of (c-d) 50^{th} cycle end discharge capacities of doped nickel-cobalt-manganese oxide layered cathodes. (a-b) corresponds to the investigating range noted in Table 2 and (c-d) corresponds to the investigating range of $10^{-2} \sim 10^{-1}$ and $10^{2} \sim 10^{3}$ for gamma and cost, respectively.

Table 6.5: The final optimised hyperparameter for the support vector machine regression model for the prediction of the initial discharge capacity (IC) and the 50th cycle end discharge capacity (EC) of doped nickel-cobalt-manganese oxide layered cathodes.

Optimal hyperparameters and model performance		
	Cost	Gamma
IC	281.177	0.0244
EC	126.486	0.0449

Figure 6.13-a and **Figure 6.13-b** indicate the prediction performance of the SVM models for the initial discharge capacity and the 50th cycle discharge capacity, respectively. Firstly, the SVM regression model for the EC prediction seems to outperform the IC prediction model as its RMSE value and the R^2 value for the test set are calculated to be smaller. In addition, the SVM-EC model also seemed to have better predictive power than the results of the previous KRR-EC model as the plotted points are observed to be closer to the true line. On the other hand, the SVM-IC model seems to perform less accurate prediction for the test set than the KRR-IC model as the RMSE value is 1.32 mAhg^{-1} higher.

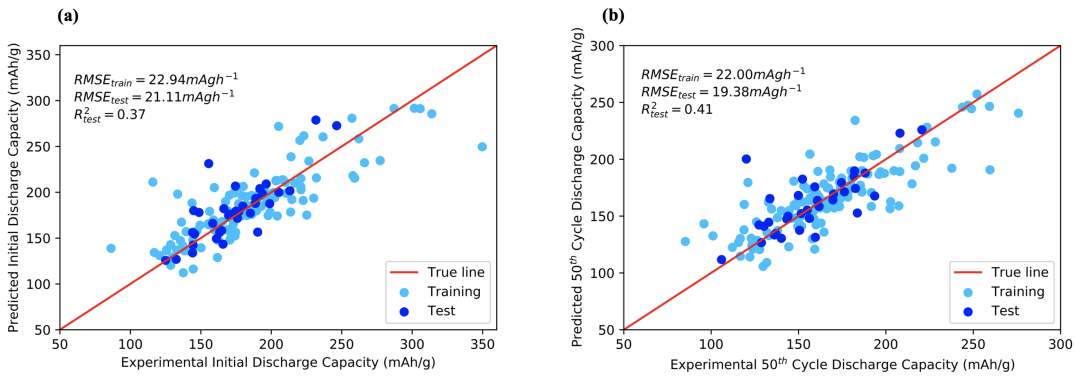


Figure 6.13: The predicted values computed versus the experimental values for (a) the initial discharge capacity (IC) and (b) 50th cycle end discharge capacity (EC) of doped nickel-cobalt-manganese oxide layered cathodes computed by the support vector machine models, SVM-IC and SVM-EC, respectively.

6.3.6 K-Nearest Neighbour

The main hyperparameters to be optimised for K-nearest Neighbour (KNN) is the number of neighbours during the calculations of the Minkowski distances. **Fig-**

ure 6.14 gives the graphical illustrations of the prediction error calculated for the number of neighbours from 1 to 20. To begin with, both models have demonstrated a positive relationship between the prediction error and the number of neighbours with the higher amount leading to higher mean squared errors. In addition, the optimal number of neighbours corresponds to the minimum mean squared error are observed to be 2 and 5 for the prediction of IC and EC, respectively. Furthermore, **Figure 6.15-a** and **Figure 6.15-b** demonstrate the prediction power of these models in the training set and the hold-out test set. It is seen that the prediction for the initial discharge capacities have slightly improved from the previous SVM model as the RMSE error is reduced from 21.11mAhg^{-1} to 18.98mAhg^{-1}. This improvement in the IC prediction accuracy is seen as the result of more accurate predictions over the two test sets of higher capacity at around 250mAhg^{-1} as they appear to be close to the true line. In addition, the performance of the KNN model over the EC prediction is seen worse than the performance of the SVM model with its poor explanations of the capacities in the test-set, particularly at the low-value range at around 100mAhg^{-1} and the high-capacity range at around 225mAhg^{-1}.

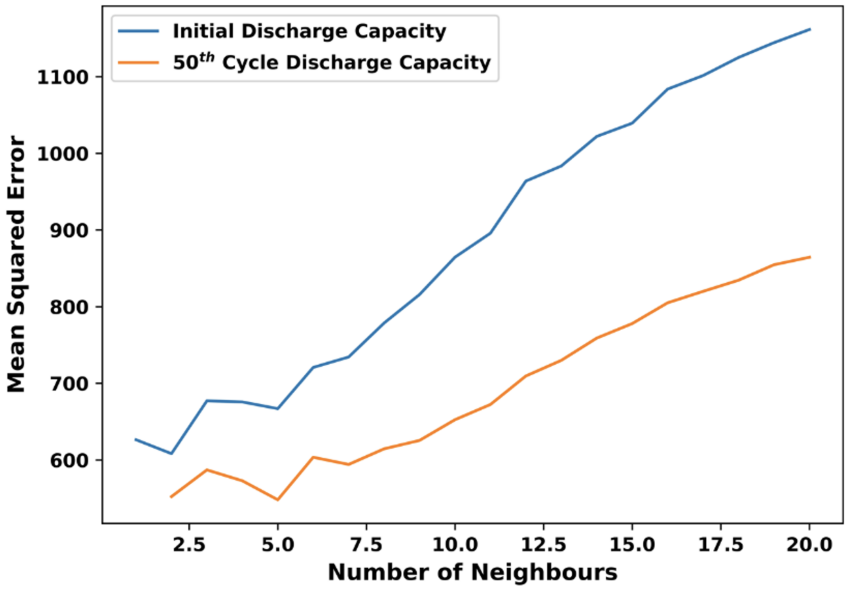


Figure 6.14: Mean Squared Error versus the number of neighbours during the predictions of initial discharge capacity and the 50th cycle discharge capacity of doped nickel-cobalt-manganese oxide layered cathodes.

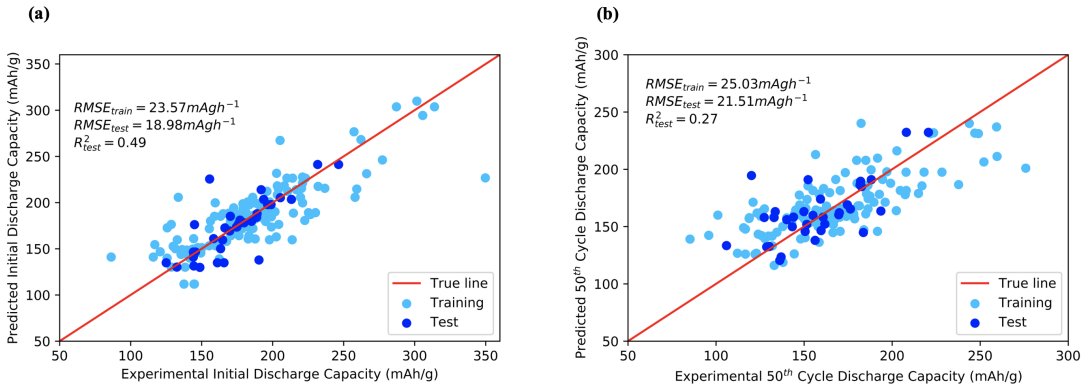


Figure 6.15: The predicted values computed versus the experimental values for (a) the initial discharge capacity (IC) and (b) 50^{th} cycle end discharge capacity (EC) of doped nickel-cobalt-manganese oxide layered cathodes, computed by the 2-nearest neighbour model and 5-nearest neighbour model, respectively.

6.3.7 Random Forest

The optimisation of the random forest model involves the considerations of five different hyperparameters including the number of trees, maximum number of covariates for splitting, maximum tree depth, the minimum amount of samples remaining at each leaf and the minimum amount of samples to be split at each node. The tuning process for these parameters is carried out continuously with each step aimed at refining the investigating value range to search for the optimal set of hyperparameters with lower mean squared prediction errors. The randomised search technique is implemented to automatically select the investigating parameter value from the given range. **Table 6.6** summarises the investigating value range of each random forest hyperparameter at each trial with the identification of the minimum mean squared error against the test set. In addition, the optimal sets of the hyperparameters are chosen after the second trial of hyperparameter tuning and are highlighted in red. During the selection of the optimal hyperparameters, it is noticed that the value of the mean squared error is more sensitive towards the change of the maximum number of covariates feature at each split than the total number of trees in the random forest. This is considered reasonable as different covariates would contribute differently to the splitting of the trees and hence the different prediction powers. A few observations can be made by comparing the optimal forest parameters between the

Table 6.6: Different trials (T-1, T-2) of hyperparameter tuning results for the random forest models for the prediction of initial discharge capacity (IC) and the 50th cycle discharge capacity (EC) of doped nickel-cobalt-manganese oxide layered cathodes, where N_trees indicates the number of trees, min_samples_split and min_samples_leaf imply the minimum amount of samples considered at each split and should remain at each leaf and max_features indicate the maximum number of covariate features should be considered at each split, max_depth denotes for the maximum tree depth in the random forest and Min_MSE represents the minimum mean squared error obtained from the investigating range.

	IC			**EC**		
	T-1	T-2	Optimal	T-1	T-2	Optimal
N_trees	100 ∼ 5000	300 ∼5000	306	100 ∼ 5000	350 ∼ 5000	513
Min_samples_split	1 ∼15	2	2	1 ∼15	2 ∼6	5
Min_sample_leaf	1 ∼15	1	1	1 ∼15	1	1
Max_features	1 ∼20	6 ∼ 11	7	1 ∼20	8, 11, 12	11
Max_depth	1 ∼15	9 ∼15	14	1 ∼15	7 ∼15	7
Min_MSE (mAhg^{-1})	461.41	413.95	-	410.32	406.99	-

IC and the EC models. First, the IC model requires a much smaller number of trees and fewer features at each split to predict than the EC model does. Both models have the same requirement for the minimal amount of samples remained at each leaf and the EC model requires more samples to split than the IC model does.

Next, the prediction power of these random forest models is examined with the hold-out test set. **Figure 6.16-a** and **Figure 6.16-b** illustrate the prediction performance of the random forest models for IC and EC experimental values in the test set. From the results, the random forest model seems to perform better in the prediction of the IC than the EC as the RMSE test value is 2.25 mAhg^{-1} higher in the latter case. Overall, the IC model seems to have a good overall prediction power as all the plotted test points are very close to the true line. On the other hand, the EC model seems to have fewer prediction powers as the test points are further away from the true line than the shown IC model results. In addition, **Figure 6.16-c** and **Figure 6.16-d** reveal the importance of each covariate variable during the model predictions of IC and EC values in the test set. The identifications of the covariate variable associated with each feature number are given in **Table 6.7**. To begin with,

it is seen that the relative importance of the covariate variables in the IC model is more equally distributed than for the EC models. A clear observation can be made as to the feature number 1, 10 and 11 are shown as the most important variables in both prediction models and they are corresponding to the lithium formula ratio, the minimum operating voltage, and the maximum operating voltage in the cyclic test. Furthermore, the molar mass (feature number: 13) and the dopant's electron affinity (feature number: 18) are indicated as the fourth important covariate variable for the IC model and the EC model, respectively.

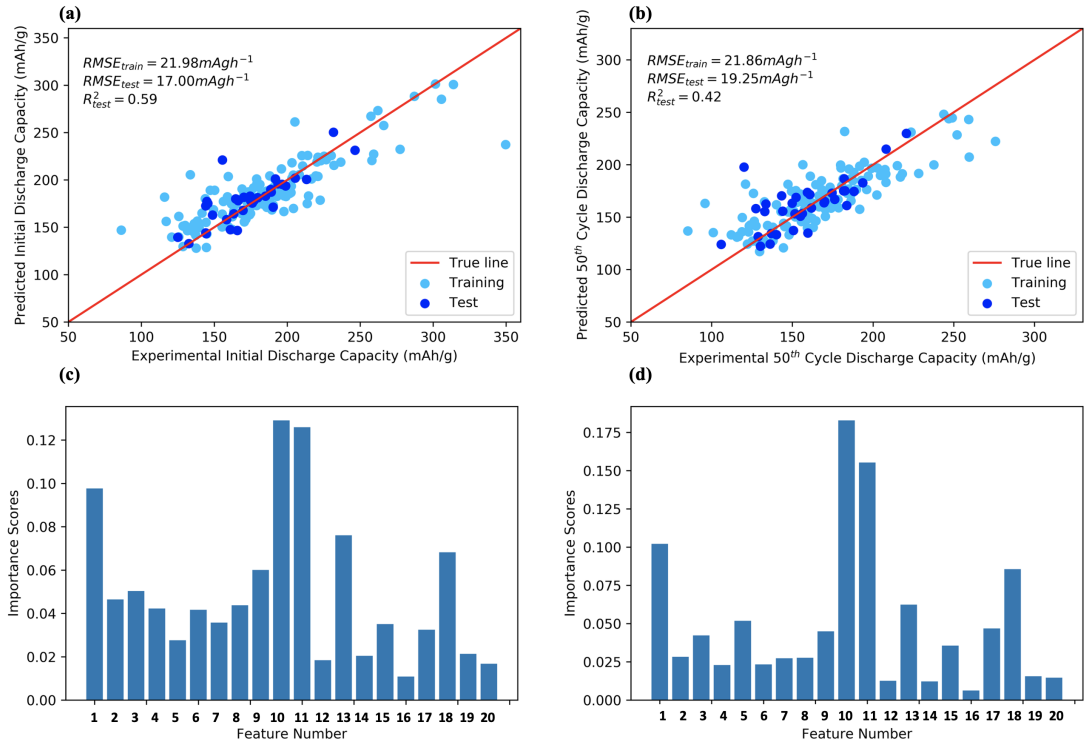


Figure 6.16: The predicted values computed versus the experimental values for (a) the initial discharge capacity (IC) and (b) 50th cycle end discharge capacity (EC) of doped nickel-cobalt-manganese oxide layered cathodes computed by the random forest models and the ranking of the variable importance during the prediction of the hold-out test set for (c) IC and (d) EC, respectively.

6.3.8 Gradient Boosting Machine

The optimisation of the hyperparameters in the gradient boosting machine was carried out in a very similar manner as the previous random forest models. The hyperparameters of the gradient boosting machine include the number of trees, the minimum amount of samples at each split and remained at each leaf, the maximum

Table 6.7: Identification of the covariate variables to the feature number used in the model.

Feature Number	Covariate Variables	Feature Number	Covariate Variables
1	Li formula ratio	11	Maximum operating voltage
2	Ni formula ratio	12	M's molar mass
3	Co formula ratio	13	Material formula's molar mass
4	Mn formula ratio	14	M's number of electrons
5	Dopant (M)'s formula ratio	15	M's electronegativity
6	Lattice constant a	16	M's number of isotopes
7	Lattice constant c	17	M's first ionisation energy
8	Crystal lattice volume	18	M's electron affinity
9	Current Density	19	M's atomic radius
10	Minimum operating voltage	20	M's ionic radius

features considered at each split, maximum depth and the learning rate. **Table 6.8** and **Table 6.9** summarise the trials of investigating ranges for each hyperparameter along with the minimum mean squared errors for IC and EC, respectively. From **Table 6.8**, it is seen that the minimum mean square error value is reduced as the investigating value ranges for each hyperparameter reduce gradually from each trial. On the other hand, despite the fact that the minimum means a squared error that seems to increase from trial 1 to trial 4, the differences are seen to be very small (< 4 mAhg^{-1}). The optimal hyperparameter sets for the IC model is then selected from the T-4 set and the optimal hyperparameter sets for the EC model is selected from the T-1 set and these are coloured in red in each assigned results table. In addition, it is noticed that the variations of learning rate seem to have more control over the model prediction than other parameters like the number of trees and this might be because the learning rate scales the contribution of each tree to the overall prediction.

Figure 6.17-a and **Figure 6.17-b** illustrates the prediction performance of the

Table 6.8: Four trials (T-1, T-2, T-3, T-4) of results of hyperparameter tuning results for the gradient boosting models for the prediction of initial discharge capacity (IC) of doped nickel-cobalt-manganese oxide layered cathodes, where N_trees indicates the number of trees, min_samples_split and min_samples_leaf imply the minimum amount of samples considered at each split and should remain at each leaf and max_features indicate the maximum number of covariate features should be considered at each split, max_depth denotes for the maximum tree depth in the random forest, the learning rate is the scaling factor and Min_MSE represents the minimum mean squared error obtained from the investigating range.

	T-1	**T-2**	**T-3**	**T-4**	**Optimal**
N_trees	100 ∼ 5000	400 ∼4801	1000 ∼ 4801	1000 ∼ 4801	3295
Min_samples_split	1 ∼15	3 ∼12	3 ∼12	4,5,6,9,10	4
Min_sample_leaf	1 ∼15	1 ∼9	2 ∼9	2, 3	2
Max_features	1 ∼20	3 ∼ 9	3 ∼8	5, 6	5
Max_depth	1 ∼15	1 ∼9	1 ∼5	3	3
Learning_rate	10^{-1} ∼ 10^{-4}	10^{-1} ∼ 10^{-3}	10^{-1} ∼ 10^{-3}	10^{-1} ∼ 10^{-3}	0.0850
Min_MSE (mAhg^{-1})	367.81	364.15	357.78	357.76	

Table 6.9: Four trials (T-1, T-2, T-3, T-4) of results of the hyperparameter tuning for the gradient boosting models for the prediction of 50th cycle discharge capacity of doped nickel-cobalt-manganese oxide layered cathodes, where N_trees indicates the number of trees, min_samples_split and min_samples_leaf imply the minimum amount of samples considered at each split and should remain at each leaf and max_features indicate the maximum number of covariate features should be considered at each split, max_depth denotes for the maximum tree depth in the random forest, the learning rate is the scaling factor and Min_MSE represents the minimum mean squared error obtained from the investigating range.

	T-1	**T-2**	**T-3**	**T-4**	**Optimal**
N_trees	100 ∼ 5000	460 ∼5001	460 ∼ 4801	1000 ∼ 4801	4369
Min_samples_split	1 ∼15	2 ∼15	8 ∼15	10 ∼14	2
Min_sample_leaf	1 ∼15	2 ∼15	1,2,3 ,11,14	1, 2	1
Max_features	1 ∼20	7 ∼ 18	7,8,10,11 ,13,15,16	7,8	7
Max_depth	1 ∼15	2 ∼10	2 ∼7	2∼4	3
Learning_rate	10^{-1} ∼ 10^{-4}	10^{-1} ∼ 10^{-3}	10^{-1} ∼ 10^{-3}	10^{-1} ∼ 10^{-3}	0.0495
Min_MSE (mAhg^{-1})	359.07	363.95	363.95	363.54	

gradient boosting models for the IC prediction and EC prediction respectively in the training set and test set. It is seen that both models have performed well in predicting the experimental data in the hold-out test set as many of the test points are shown to be close to the true line. The IC model results in a prediction error at 16.66mAhg^{-1} for the test set with its R^2 scores calculated to be 0.76. In addition, the training set RMSE value (20.26 mAhg^{-1}) is shown to be much higher than the RMSE test value. In the results graph of the EC model (**Figure 6.17-b**), a prediction error of 18.59mAhg^{-1} is achieved for the test set with the R^2 score calculated to be 0.64 and these demonstrate its great prediction power in explaining the test set data. **Figure 6.17- c** and **Figure 6.17- d** provide further insight into the rationale behind the variable importance for the IC and EC models. Like the previous findings from the random forest graphs (**Figure 6.16- (c-d)**), the minimum operating voltage and the maximum operating voltage remain the two most important variables in both gradient boosting models for the IC and EC predictions. In addition, the dopant formula ratio seems to be the third most important variable for the IC prediction whereas the lithium formula ratio remains the third most important variable for the EC prediction.

6.3.9 Model Performance Comparisons

In this project, six different non-linear regression algorithms have been trained, optimised, and validated against a holdout test set for their prediction powers. **Table 6.10** gives the summary of the computed RMSE values for each optimised model during their training stage and for predicting the held-out test set. In addition, the respective R^2 values for the test set prediction are also included to demonstrate 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. Firstly, the ANN models are shown to have the worst performance with the highest test-set RMSE among all. This is because both ANN models are over parameterised (**Table 6.3**) and would require a much larger sample size to estimate these well. Furthermore, the tree-based ensemble methods generally have much lower

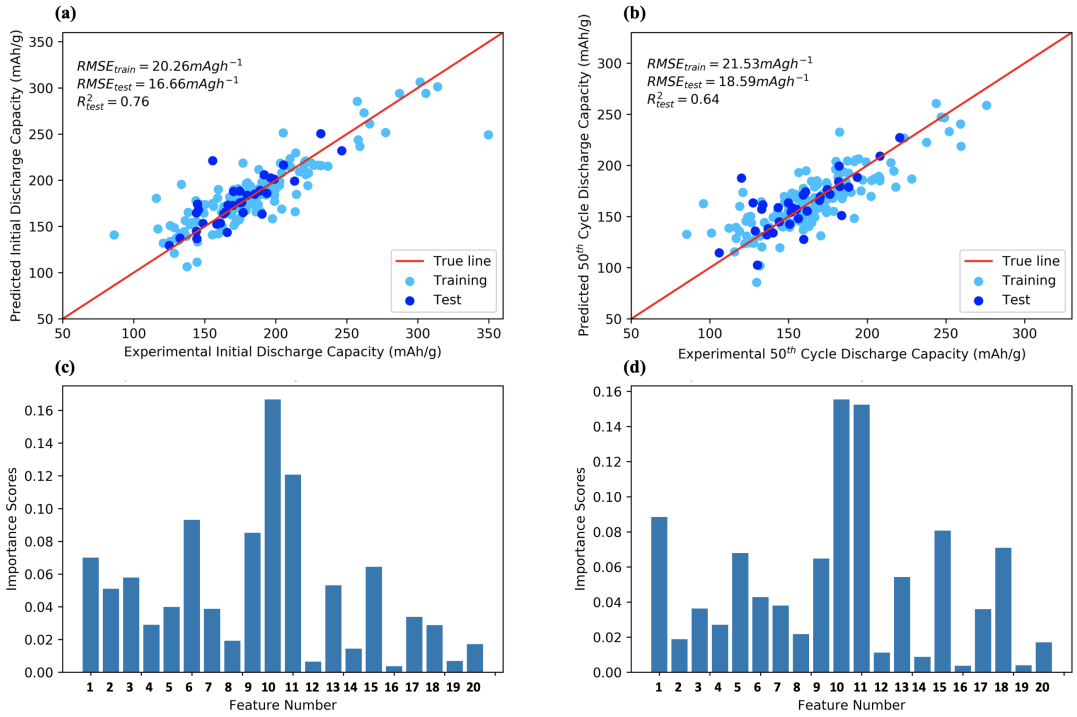


Figure 6.17: The predicted values computed versus the experimental values for (a) the initial discharge capacity (IC) and (b) 50th cycle end discharge capacity (EC) of doped nickel-cobalt-manganese oxide layered cathodes computed by the gradient boosting models and the ranking of the variable importance during the prediction of the hold-out test set for (c) IC and (d) EC, respectively.

test-set RMSE values than other nonlinear models such as SVM and KNN. However, the prediction power of KRR model for the IC prediction is shown to be the third best performing model with the RMSE value computed to be 17.28 mAhg^{-1} and this is very close to the performance of the respective random forest model. In addition, the KRR model has also shown superior prediction power over the EC prediction than the random forest model with its RMSE-test value calculated to be 0.12 mAhg^{-1} higher. Nevetherless, both KRR models are still not comparable with the gradient boosting machine models as they have shown the best prediction performance for both tasks, with their test set RMSE values being the lowest among all, at 16.66 mAhg^{-1} and 18.59 mAhg^{-1}, respectively. In this regard, these optimised GBM models will then be used to carry out the variable correlation study to reveal the relationship of each covariate with the target variables.

Table 6.10: Comparisons of the mean RMSE values during the five-fold cross-validation and for testing against holdout test and the R^2 test score computed by six non-linear models, for the prediction of initial discharge capacity and 50^{th} cycle end discharge capacity of doped nickel-cobalt-manganese oxide layered cathodes.

ML techniques	Initial Discharge Capacity ($mAhg^{-1}$)			50^{th} 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

6.3.10 Variable Correlation Analysis

The covariate variable importance can be estimated through the calculation of the Shapley values from the best performing models' predictions of the hold-out test set. Sharpley values come from the coalitional game theory where each of the covariate variables is treated as the individual "player" and the values estimate the covariate variables' contribution to the final prediction of a response variable instance. It is more desirable than the traditional permutation method for the easier and clearer interpretation of the variable correlation with the response variable. In this project, the treeSHAP (Shapley Additive exPlanations) method proposed by *Lundberg et al* [305], is used to gain insight into the importance of all covariate variables and their feature effect on the prediction. **Figure 6.18-a** and **Figure 6.18-b** show 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 impact of that covariate variable on the model prediction with the more positive or negative value implying a larger overall influence.

To begin with, the minimum cut-off voltage, maximum cut-off voltage and the

current density are ranked within the top 10 important variables. These covariate variables are all the experimental conditions for cycling and thereby are expected to greatly influence the material's discharging performance as a cathode [306]. After excluding these experimental setting variables, one can see that the dopant content ratio and the lithium content ratio are being ranked as the third and the fifth most important features, respectively for the IC test-set prediction from **Figure 6.18-a**.

A negative correlation is identified for the dopant content ratio and the IC as an increase in the corresponding Shapley values lead to the decrease in the IC feature values (shown in the sequence of red to blue). On the contrary, the Shapley values of lithium content ratio are shown to be positively correlated to the IC value, with the colour of the data plot shown to be blue to red. **Figure 6.19-a** shows the 3D plot of the dopant content ratio, Li content ratio correlating to the respective IC values for the entire dataset (train+ test). Firstly, two clusters of data can be identified with one characterized for 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 reduces in the first cluster. Moreover, the latter cluster has much higher average IC values than the first one which implies that a higher Li content is generally more desirable for obtaining a high IC value. This concludes that the layered material formula with 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 that a doped NCM cathode material formula with lower dopant's electronegativity (EN_M), the shorter lattice constant "*a*" (LC_a) and "*c*"(LC_c), smaller the manganese content and the smaller formula molar mass can lead to higher IC values.

Figure 6.18-b shows that the dopant content ratio and its electronegativity value are ranked as the second and the fourth important for the predictions of EC. Electronegativity measures the dopant element's ability to attract electron pairs toward itself. Dopant's EN controls the bonding strength with the surrounding TMs and oxygen atoms and influences structural stability as well as the overall crystal

structure density. During 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 [307]. The involvement of dopant content can greatly improve the structure stability by forming stronger bonds whilst the strategy with doping with a small amount can ensure that no second material phase is formed and also 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 to the EC feature value as the colour of the trend are changing red to blue (left to right). **Figure 6.19-b** displays the 3D intercorrelation of the two dopant-related covariate variables with the respective EC in the whole dataset. A clear trend is observed for EC decreasing with the 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 for when the dopant content ratio is kept low ($x > 0.02$), and this corresponds to the magnesium (1.31) and zirconium (1.33) dopant in the collected dataset. Mg^{2+} ions can occupy the Li^{+} sites in the crystal structure owing to the similar ionic radius and this would prevent the crystal structure from transforming from the layered to the spinel during the cycling. The Zr^{2+} ions can substitute for the TM sites and can be used to alleviate the cation mixing effects between the Li^{+} and the Ni^{2+}. Nevertheless, both doping elements are considered electrochemically inactive during the redox reactions and therefore a small amount is more desirable to preserve the original high capacities of the NCM cathode. Other observations can be made from **Figure 6.18-b** 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 lot of similarities with the previous findings in the IC variable correlations.

To conclude, our results have demonstrated that the materials that both constitute high IC and EC share the common characteristics of high Li content ratio, small dopant ratio and being doped with atoms of low to middle range electronegativity. In addition, it is also encouraged to design a doped NCM material with low formula

molar mass as it improves both discharge capacities.

6.4 Summary

Analysing the past experimental results is a crucial step to better understand the complex correlations of the 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 non-linear machine learning algorithms have been trained and validated with the manually curated 2197 experimental results of 168 doped NCM materials. The models are built on using thirteen material physical properties and seven dopant's elemental properties as covariate variables to predict the initial (IC) and 50^{th} cycle (EC) discharge capacities of each material structure. First, the Pearson coefficient correlation study has indicated that no strong linear correlations are captured for any pairs of covariate variables with either of the two response variables. In addition, gradient boosting models are suggested to hold the best prediction power against the holdout test set for having the lowest root-mean-square error at 16.66 mAhg^{-1}, 18.59 mAhg^{-1} and highest R^2 scores at 0.76, 0.64 during IC and EC prediction respectively. Further insights are gained into the governing material features for each discharging properties. NCM materials with higher lithium content, smaller dopant content and doped with the lower electronegativity value atom seem to bring higher values in both IC and EC. From these promising results, we expect that these machine learning models can be used as a guide tool to estimate the discharging properties of any single doped NCM materials and potentially discover new cathode materials with more advanced electrochemical properties.

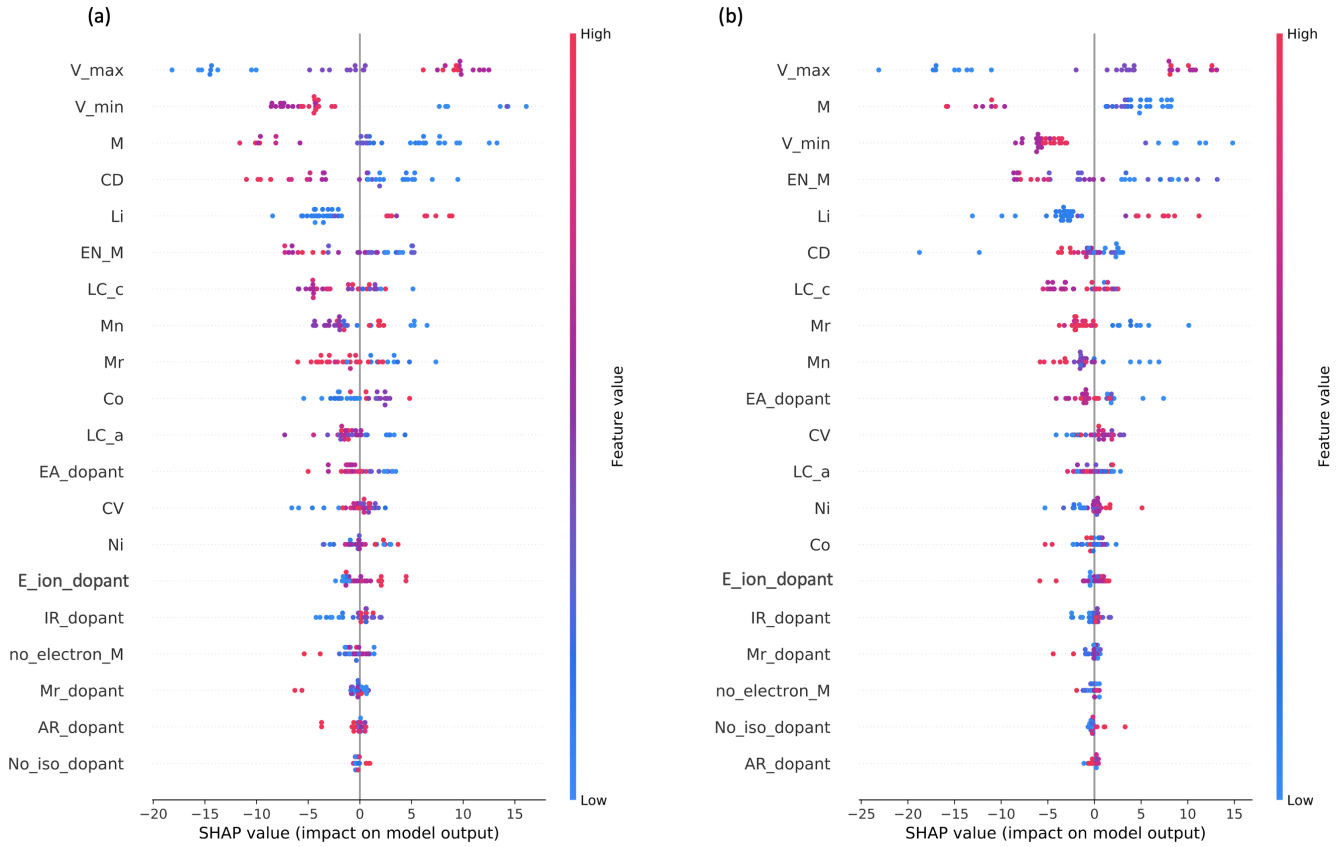


Figure 6.18: The summary plots for the feature contribution of twenty covariate variables in the test-set prediction of (a) initial discharge capacity (IC) of doped nickel-cobalt-manganese oxide layered cathodes, generated based on the GBM-IC model and (b) 50^{th} cycle discharge capacity of doped nickel-cobalt-manganese oxide layered cathodes (EC) of doped nickel-cobalt-manganese oxide layered cathodes, generated based on 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 heat-map showing the values of the respective response feature variable.

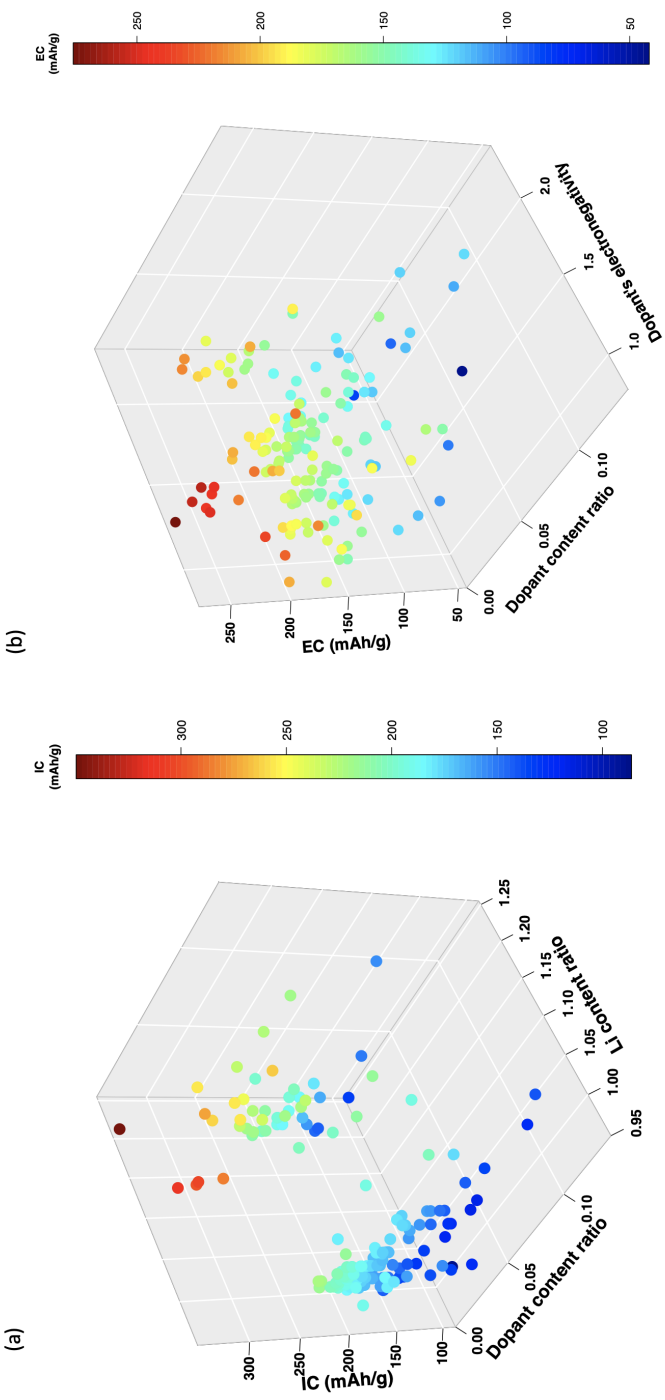


Figure 6.19: The 3D plots give insights into the intercorrelations of (a) initial discharge capacity (IC) of doped nickel-cobalt-manganese oxide layered cathodes with two most important variables (Li content ratio, dopant content ratio) and (b) 50^{th} cycle discharge capacity(EC) of doped nickel-cobalt-manganese oxide layered cathodes with the two most important variables (Dopant content ratio and dopant's electronegativity) in the whole dataset.

Chapter 7

Machine Learning Prediction of the Discharge Capacities of Doped and Carbon-Coated Olivine Cathodes

7.1 Background

Olivine cathode materials such as $LiMnPO_4$ (LMP) and $LiFePO_4$ (LFP) have attracted a wide range of attention due to their steady cyclic performance, good thermal stability, high energy density and low cost. In comparison to other cathode classes, olivine-structured materials offer a unique one-dimensional Li-ion travel path along the [010] tunnel direction as shown in **Figure 7.1-a** and have comparatively greater structural stability during cycling thanks to the strong covalent bonded PO_4^{3-} three-dimensional frameworks. The differences in the redox potentials of various transition metals have led to the different voltage profiles for every olivine compound with a different base metal as shown in **Figure 7.1-b**. LFP tends to generate a flat charge-discharge profile at 3.5V which is much lower than LMP (4.1V), $LiCoPO_4$ (4.8V) and $LiNiPO_4$ (5.1V). In addition, LFP is known to have the same theoretical capacity as the LMP. However, LMP is a more popular option for large scale applications due to its comparable operating voltage with the widely commercialised layered $LiCoO_2$ cathode (4V versus Li^+/Li). The other cobalt-containing olivine compound $LiCoPO_4$ seems to be a more feasible cathode

option than $LiNiPO_4$ as its operating voltage is within the electrochemical window of the conventional organic liquid electrolyte. However, the cost of cobalt is still a big obstacle for the usage of this material in wide applications.

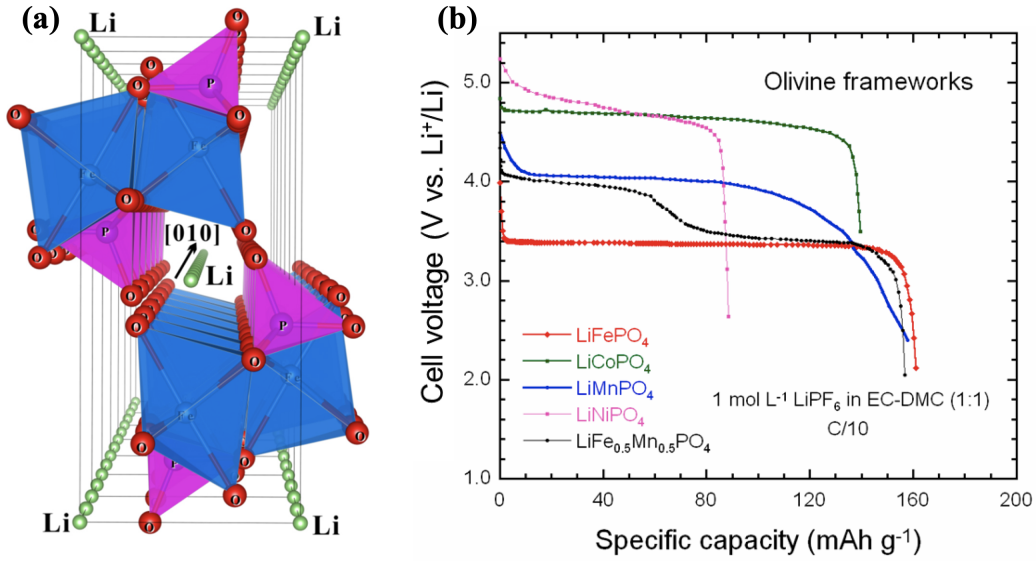


Figure 7.1: The illustration of (a) the crystal structure of $LiFePO_4$ with the indication of the [010] pathway for Li-ion diffusion and (b) the voltage profile of various olivine-structured compounds like $LiFePO_4$, $LiCoPO_4$, $LiMnPO_4$, $LiNiPO_4$, $LiFe_{0.5}Mn_{0.5}PO_4$ against the Li-metal anode and tested under the conditions of the current density of C/10 whilst using 1M $LiPF_6$ as the electrolyte. Image (a) taken with permission from [27] and image (b) taken with permission from [28].

Nevertheless, common drawbacks shared among these olivine cathodes are known to be the low electronic conductivity ($< 10^{-9}$ Scm^{-1}) and low Li-ion diffusion coefficient ($10^{-14} \sim 10^{-16} cm^2 s^{-1}$) and these inferior properties are attributed to the insulation effects of the polyanions groups (PO_4^{3-}) in the material structure. As a result, olivine cathodes often suffer from poor performance such as poor rate capability, low withdrawable capacity, limited cycle life, which restrict their applications in the industry. It is important to note that the low conductivity issue is known to be more severe for the LMP as it is noted that their electronic conductivity ($<10^{-10} Scm^{-1}$) is much less than LFP. LFP tends to form a new $FePO_4$ phrase during the Li-ion extraction process and this new $LiFePO_4/FePO_4$ interface can delay the Li-ion diffusion and worsen the material's rate capability. Similarly, LMP also

suffers various disadvantages such as the large unit cell volumetric expansion during cycling, the low kinetics of Li-ion diffusions due to the high energy barrier of the interface between $LiMnPO_4$ and $MnPO_4$ at charged state (Jahn-teller distortion effects of the Mn^{3+}) and the unstable charged phase where the $MnPO_4$ decomposes to $Li_4P_2O_7$ and O_2 on the electrode surface.

Both the doping treatment and the carbon coating treatment are effective in improving the electronic conductivity of the olivine materials and enhancing their respective cyclic performance. The benefits of carbon coating vary from improving the conduction of the internal particles, avoiding the direct contact of the active material with the electrolyte to prevent the metal dissolution reaction and restricting the unwanted crystal growth. An illustration the applied carbon coating on the surface of the intrinsic olivine cathode is given in **Figure 7.2-(a-d)**. Nevertheless, the extent of these benefits is also controlled by the conditions and the properties of the carbon coating. The carbon structure on a carbon coating is normally measured by the D/G ratio calculating from their respective peaks on the Raman microprobe spectrum and this represents the ratio of the graphitized carbon content to the disordered carbon content on the coating. A similar metric is also seen as the ratio sp^3/sp^2, which indicates the ratio of the number of sp^3 carbon to the number of sp^2 carbon bonds in the coating. Research has indicated that the $C/LiFePO_4$ composite with lower D/G ratio (lower sp^3/sp^2 ratio) seems to have higher conductivity and therefore suggests that the more graphitized carbon coating is more desirable for better discharge performance [308]. In addition, studies have also been performed on replacing the amorphous carbon with other carbon structures such as carbon nanotube [309, 310], carbon nano-fibres [311, 312] and these have seen various scales of improvement in the rate capability and discharge performance. Other carbon coating factors subject to the application for olivine cathode are also studied widely such as the effects of carbon sources [313, 314], carbon coating deposition methods [315, 316] and the carbon surface area and porosity [29, 317, 318, 319, 320, 321].

Next, the doping method with the use of aliovalent ions (Mg^{2+}, Al^{3+}) or prevalent ions (e.g. Ti^{4+}, Nb^{5+}) at either the lithium lattice sites or the transition metal

lattice sites have been previously shown to improve the conductivity of the bulk structure. The underlying reasons are thought to be that doping method improves the kinetics of the phase transformation during the charging/discharging and some studies also suggest that the [010] Li-ion diffusion path could be expanded by the dopant atoms and therefore promoting fast Li-ion transfer during cycling. For instance, *Zhang et al* [29] reported an improvement of 13% on the 50^{th} cycle discharge capacity from doping the pristine $LiFePO_4$ with 5% of the vanadium ions (shown in **Figure 7.2-e**)whilst reducing the carbon coating content from 3.43wt.% to 3.14wt.%. This indicates that the doping methods can effectively enhance the cyclic performance whilst improving the overall energy density by reducing the content of carbon coating in the system. Nonetheless, there are still some remaining challenges to be tackled before the large applications. The first is to determine the optimal amount of carbon coating content to be applied onto the olivine material without reducing for the overall energy density. A heavy amount of carbon coating on the material would directly reduce the overall volumetric energy density, increase the overall Li-ion batteries volume and make them less convenient for transporting. In addition, there is a general lack of understanding in the correlations of the doping treatment with the improved systematic electronic conductivity and the finally enhanced cyclic performance. Some suggest that doping treatment can induce lattice distortion in the olivine structure which weakens the Li-O bond and improve the electronic conductivity from the higher kinetics of Li-ion diffusions. Others question whether the improved electronic conductivity is truly led by the doping effects on the olivine lattice sites or that the doping metals caused the surface conducting percolating [321]. Furthermore, the substitutions of the super-valent metal ions to the M lattice sites of $LiFePO_4$ are modelled to be energetically unfavourable, which makes one wonder whether the improved discharge performance is caused by the carbon coating alone or by the doping treatment [133, 322].

A thorough understanding of the roles of carbon coating and doping in the improvement of cyclic performance is needed for the development of new olivine cathodes with higher electrochemical performance. This work incorporates the power of

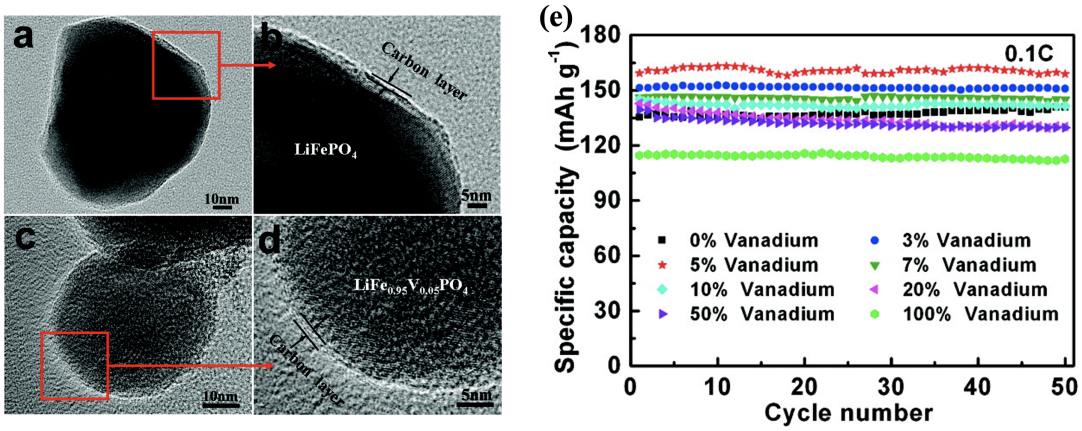


Figure 7.2: The transmission electron microscopy images of (a-b) the undoped, carbon-coated (3.43 wt.%) $LiFePO_4$ compound and (c-d) 5% Vanadium doped, carbon-coated (3.14 wt.%) $LiFePO_4$ and (e) their cyclic performance up to 50 cycles under the current density of 0.1C with the use of Li-metal as the anode and $LiPF_6$ as the electrolyte. Images were taken with permission from [29].

machine learning techniques to unveil the relationships of the aforementioned material properties with their respective discharge capacities. Five different machine learning algorithms were implemented to learn a dataset of 34 doped and carbon-coated olivine materials that includes 24 LFP-based systems and 10 LMP-based systems. 17 different covariates are chosen to predict the material's discharge capacity at the first and the 20^{th} cycle. These cover experimental results such as crystal lattice constants ("a", "b", "c"), the volume of the unit cell (CV), the formula ratio of the based metal element (B) and the dopant element in the material formula (M), material molar mass and the cyclic test parameters such as the charge/discharge current density (CD) and the minimum and maximum operating voltage limit (V_min, V_max). A tabular illustration of the model variables with their abbreviations used in the project is given in **Table 7.1**. Before the construction of the machine learning models, Pearson coefficient correlation analysis is performed to investigate the correlations of the variables in the dataset. Next, machine learning models including Bayesian ridge linear regression, support vector machine, kernel ridge regression, random forest and gradient boosting model are trained to predict the initial and the 20^{th} cycle end discharge capacity with the given covariates. The predictive power of every model will be assessed, and compared for both capacity prediction tasks. The

optimal models will be further used to predict other olivine cathode systems with different transition metals and carbon coating content from the literature. Meanwhile, a variable importance study is conducted to reveal the relationships of the selected covariate variables such as the dopant amount, carbon coating content with the target discharge performance variables.

7.2 Methods

The collection of the dataset for the doped Fe and Mn-based olivine materials was carried out with strict selection rules. The publications should fulfil all following criteria to be considered suitable for the data collection: i) the olivine undoped materials should be either iron-based or manganese-based; ii) the olivine material should be single doped on the transition-metal lattice sites with cationic ions; iii) the material should be coated with carbon and report the amount of carbon coating in the weighted percentage with respect to whole material system; iv) material should have an orthorhombic structure with a space group of Pnmb and reported the lattice parameters; v) the respective experimental section should report the information of the mix ratio of the active materials, conductive and binder; vi) the cyclic test should be carried out up to 20 cycles under the the standardised working conditions (i.e. Temperature = 25 $\pm$ 5 ℃, Pressure = 1atm) using the same current density for charging and discharging; vii) the experimental section should report the minimum and maximum operating voltage used for the cyclic test. In addition, Li metal and $LiPF_6$ should be used as the contrasting electrode and the electrolyte, respectively.

During the investigation of the carbon coating conditions, special attention was paid to check that the coating is coated evenly on the material surface from the reported transmission electron microscopy images. However, the conditions of the carbon coating on the whole of the cathode material are not always available because of the limited reporting and the difficulty in manual interpretation of the particle size and morphology from the reported scanning electron microscopic images. We excluded the olivine materials that had applied other coating materials such as copper, silver, metal oxide, conductive polymer and/or doped with anion dopants

Table 7.1: Overview of the covariate variables and response variables used in the construction of the olivine machine learning models.

Covariate Variables			
Publication Results		Elemental Properties	
Name	Abbreviation	Name	Abbreviation
The ratio of **base-metal (Fe or Mn) and the dopant** in the olivine material formula	B, M	Material formula molar mass	Mr
Crystal lattice constants "a", "b" and "c" and	LC_a, LC_b LC_c	The weighted average value of the first ionisation energy, electronegativity, ionic radius for between the based transition metal and the dopant	Avg_E_ionisation, Avg_En, Avg_IR
Carbon coating content in the system	C_ratio		
The volume of the unit cell	CV		
The ratio of the **active component**, **conductive component** and the **binder component** in the cathode system	Active_r, Conduct_r, Bind_r		
Minimum and maximum operating voltage	V_min, V_max		
The current density for charging and discharging	CD		
Response Variables			
Name	Abbreviation	Name	Abbreviation
Initial discharge capacity	IC	20^{th} cycle end discharge capacity	EC

such as chlorine or fluorine ion for data standardisation purposes. In addition, the 20^{th} cycle as the collection target for the maximum discharge capacity is made based on the most widely reported cycle number from a large number of studies on the olivine cathode systems. After applying these rules, the initial 479 collected olivine cathode journals were reduced to 18 journals, and these were turned into 34 sets of experimental data that covered 24 LFP-based systems and 10 LMP-based systems. In the dataset for the LFP-based doped systems, ten variations of dopants are collected that including Mn, V, Ti, Mo, Cu, Mg, Pt, Co, Gd. On the other hand, the dataset for the LMP-based doped systems covered four types of dopants, namely, Fe, V, Cu, Ce.

The ML models used in this work were trained using Python programming language with its relevant ML libraries (Scikit-learn, Pandas). The models are trained and validated on the mixed dataset of the LFP-based systems and LMP-based systems due to their similarities in structural properties as well as the electrochemical properties. It is important to mention that the values of three elemental properties (electronegativity, first ionisation energy and the ionic radius) are weighted based on the ratio of the base-metal ratio and the dopant metal in the formula and the formula for this is given in **Equation 7.1**.

$$\overline{W}=\frac{n_m \times w_m + n_d \times w_d}{n_m + n_d} \tag{7.1}$$

Where $\overline{W}$ is the weighted mean value of the property, n_m and n_d are the ratios of the base-metal and the dopant in the material formula, w_m and w_d are the elemental properties of the base metal and the doping metal in the material formula, respectively.

Next, five algorithms were implemented including Bayesian ridge regression (BRR), random forest (RF), gradient boosting machine (GBM), support vector machine (SVM), kernel ridge regression (KRR). The whole data space was randomly split into the ratio of 4:1 correspond to the model training set and test set, respectively. Within the training set, the model training process was carried out using the 10-fold cross-validation method to optimise the embedded hyperparameters.

Table 7.2: A summary of the hyperparameters considered in this project for each of the algorithms, along with their investigating value ranges. Note, the random forest and gradient boosting machine share many common hyperparameters and the stared hyperparameters only applies to the gradient boosting machine model.

Algorithm name	**Hyperparameter**	**Investigating range**
Kernel Ridge Regression	Kernel function	Laplacian, Radial Basis Function
	Gamma (RBF)/sigma (Laplacian)	$10^{-4} \sim 10^{4}$
	Lambda	$10^{-4} \sim 10^{4}$
Support Vector Machine (Radial basis function)	Cost	$10^{-4} \sim 10^{4}$
	Gamma	$10^{-3} \sim 10^{3}$
Random Forest and Gradient Boosting Machine	Number of trees	100∼5000
	The maximum number of features considered at each split	1∼17
	The maximum tree depth*	1∼15
	Learning Rate*	$10^{-4} \sim 10^{-1}$

Table 7.2 summarises the hyperparameters considered for each of the algorithms and their value range of investigation in this project. In addition, the Bayesian ridge regression models were trained with the number of iterations set to be 1000, the stopping criteria set to be for when the weights are reaching the values 1×10^{-6} and additionally, an intercept is included during the fitting of the models. The initial alpha value that controls the precision of the noise during convergence is set to be $\frac{1}{\text{variance(y)}}$ and the initial lambda value, that controls the precision of the weights, is set to be 1. In addition, to ensure that the results are reproducible, the same random-seed is assigned to all machine learning models.

Please note, the data curated for this chapter and the machine learning codes can be found on the GitHub page (`https://github.com/thepowerligand/LMP-ML`). The detailed references for each set of experimental data are provided in the same GitHub directory.

7.3 Results and Discussion

7.3.1 Covariate Selection for the Model

The covariates selected for this model are very similar to those of other chapters. On the other hand, a new variable such as carbon coating content is chosen specifically to reflect on the common treatment of the olivine materials to increase their electronic conductivity. The amount of carbon coating is selected as the covariate variable in the model as it is directly related to the overall thickness of the coating applied on the cathode surface. In addition, the ratios of the active material, conductive and binder during the fabrication of the cathode composites to investigate their influence over the material's electrochemical performance. In particular, the ratio of the active material and the conductive are important as one is directly related to the total amount of active cathode for carrying the discharge performance and the other controls kinetics of the Li-ion transfer within the structure. For the elemental properties such as the electronegativity, ionisation energy and the ionic radius, we had taken the weighted mean value of the first ionisation energy, electronegativity, and the ionic radius based on the weights of the base metal and the dopant element, to represent the doping induced changes to the overall internal material properties. For instance, the weighted average of the electronegativity values can indicate the systematic bonding characters of all the metals with both the surrounding oxygen atoms and the phosphorus atoms. In addition, the use of the first ionisation energy of the atoms has been seen in the previous work as the model covariate to predict the lithium diffusion barrier (EA) of the olivine materials [237]. Other experimental variables such as the unit cell dimensions and volume are included to represent the changes in material structural conditions attributed to both the different synthesis routes and the different precursors.

7.3.2 Preliminary Statistical Analysis

Figure 7.3 gives the results matrix of the Pearson correlation coefficients for every pair of the variables included in the dataset. The variable correlations are colour coded and represented by the number from -1 to 1. The closer the numeric value to

-1 or the bluer the square colour is, the more negative linearity of the variable pair is. On the other hand, the closer the numeric value to 1 or the redder the square colour is, the more positive the linearity of the variable pair is. Overall, one can see many clusters of highly positive and negative linearity variable correlations being identified in the map. For instance, the dimension parameters of the unit cells such as the LC_a, LC_b, LC_c and CV are shown to be negative linearly correlated to the average mean values of the ionic radius, electronegativity and the first ionisation energy. This seems to be reasonable as any changes in the ionic radius and the electronegativity would influence both the atomic arrangement and bonding characteristics within the structure and hence lead to different dimensions of the unit cell. Other observations can also be made such that the amount of carbon coating (C_ratio) seems to have a slight positive linear correlation with all lattice parameters, suggesting that any increase in the carbon coating content will slight expand the unit cell of the olivine cathode. This phenomenon could be considered as the strong affinity between the Fe and C which would induce changes in the crystal structures on the cathode surface. Furthermore, there are also many linear correlations observed for the response variables (IC, EC) with the remaining covariates. A clear trend of any expansion in the dimensions of the unit cell of the cathodes seems to deteriorate both capacity variables (all coloured in deep blue). On the other hand, an increase in the weighted mean values of the ionic radius (Avg_IR), the electronegativity (Avg_EN), the first ionisation energy (Avg_E_ionisation) of the base-metal and dopant metal in the material formula, tends to increase both of the capacity response variables. This indicates a close correlation between the properties of the transition metals used in the Li-ion host structure and the capacity of the cathode material.

7.3.3 Bayesian Ridge Linear Regression

The optimal coefficients of the covariates and the final alpha, gamma values in the Bayesian ridge regression (BRR) models for the predictions of IC and EC are given in **Table 7.3**. It is seen that the coefficients of covariates in both models are very similar to each other. For instance, the Ave_E_ionisation and the Avg_IR are observed

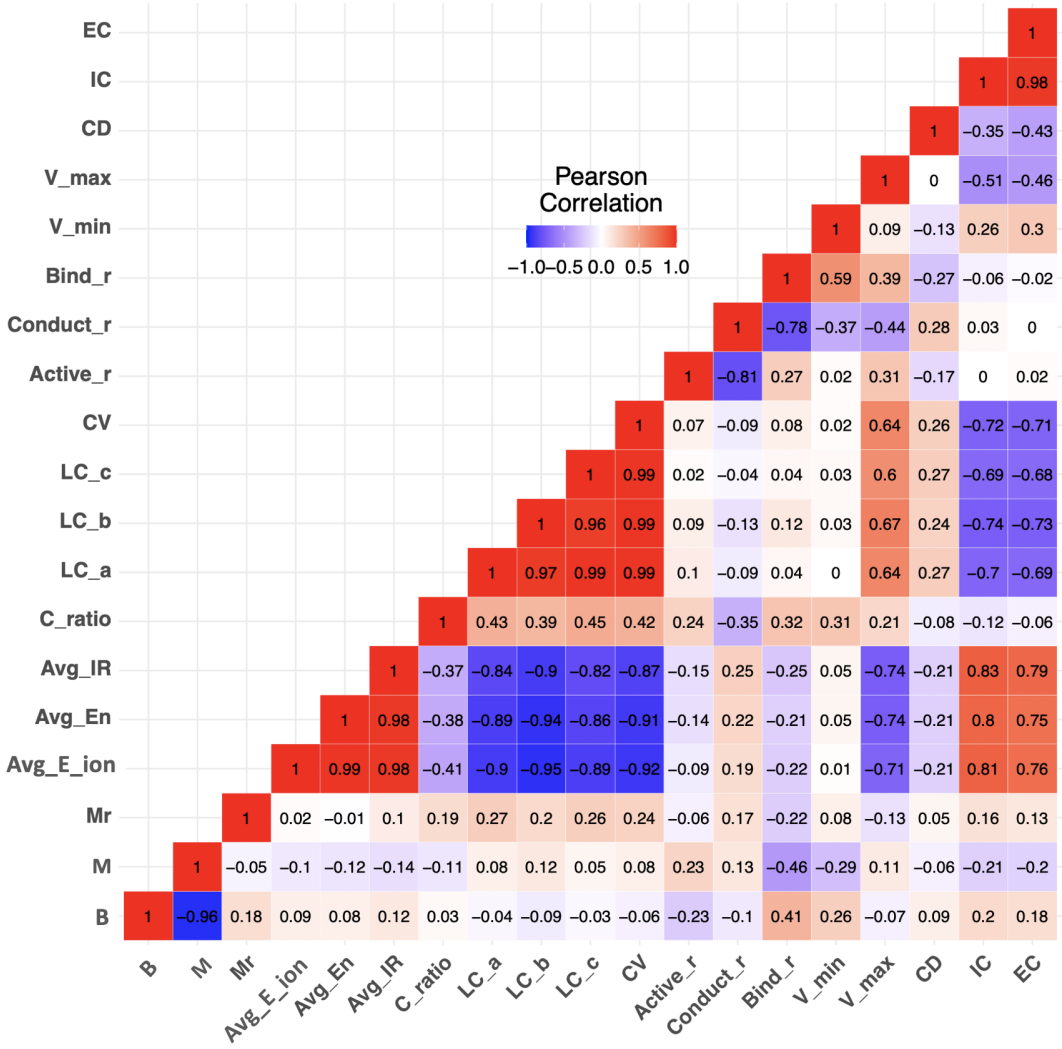


Figure 7.3: Results matrix of Pearson coefficient correlations for every pair of variables in the dataset, including covariate variables: based metal content (M), doping content (D), Average first ionisation energy of between the based metal and dopant (Avg_E_ionisation), average electronegativity of the metal and dopant(Avg_EN), Average ionic radius of the metal and dopant (Avg_IR), Carbon content ratio (C_ratio), Lattice constant-a (LC_a), Lattice constant-b (LC_b), Lattice constant-c (LC_c), crystal lattice volume (CV), the ratio of the active component (Active_r), conductive component (Conduct_r), binder (Bind_r) in the material formula, minimum operating voltage (V_min), maximum operating voltage (V_max), current density (CD), initial discharge capacity (IC), 20th cycle discharge capacity (EC) of doped $LiMPO_4$ olivine cathodes.

to have the first and second highest coefficients for both models. The optimised Alpha values are shown to be very similar to each other with a small deviation of 0.0013. This also applies to the optimised lambda values where the difference is noted to be around -0.88. The detailed discussions of the rationale behind these obtained coefficients of the respective covariates are given in the variable importance section.

Figure 7.4 (a-b) gives the graphical illustrations of the predicted values against the observed values in the case of IC and EC prediction, respectively. The training-set RMSE values in both cases are shown to be much larger than the test set RMSE, with the differences calculated to be 1.82 $mAhg^{-1}$ and 3.69 $mAhg^{-1}$, respectively. These suggest that the BRR models have performed less accurate than in the test set. The reasons behind these phenomena are thought of as the presence of the outliners within the range of 60 $\sim$ 80 $mAhg^{-1}$ and 120$\sim$140$mAhg^{-1}$ in both **Figure 7.4-a** and **Figure 7.4-b**. On the other hand, the prediction power of both models for the test set is observed to be high, with their test-set RMSE values in both models are calculated to be 9.52 $mAhg^{-1}$ and 9.25$mAhg^{-1}$. In addition, the predicted values in the test set are shown to be fairly accurate as the points (coloured in deep blue) are observed to be close to the true line.

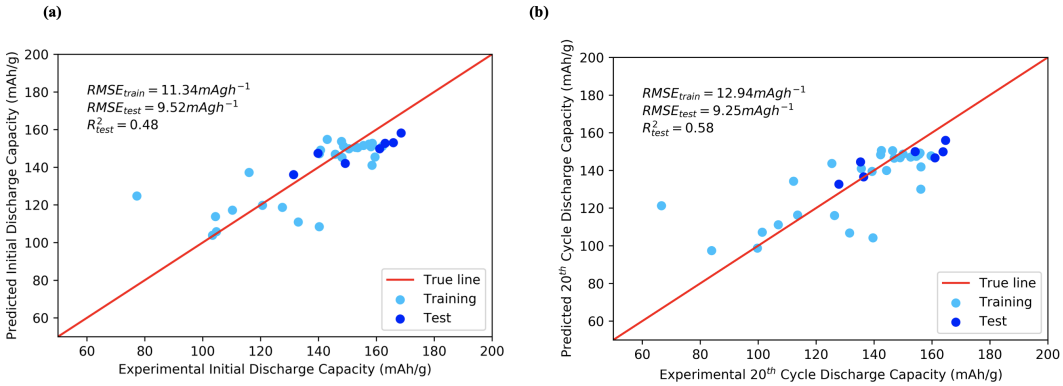


Figure 7.4: The predicted values versus the experimental values for (a) the initial discharge capacity (IC) and (b) 20th cycle end discharge capacity (EC) of doped and carbon-coated $LiMPO_4$ olivine cathodes, computed by the Bayesian Ridge linear regression models.

Table 7.3: Coefficients of the covariate variables in the Bayesian ridge linear regression functions for the prediction of initial discharge capacity (IC) and 20th cycle end discharge capacity (EC) of doped and carbon-coated $LiMPO_4$ olivine cathodes, where M stands for metal.

	IC	EC
B	0.0131	0.0150
M	-0.0146	-0.0168
Mr	0.1232	0.1603
Avg_E_ionisation	0.6573	0.6586
Avg_Electronegativity	0.0040	0.0046
Avg_IR	0.3298	0.3869
Carbon content	0.2661	0.4219
LC_a	-0.0011	-0.0012
LC_b	-0.0009	-0.0014
LC_c	-0.0004	-0.0004
Lattice volume	-0.0957	-0.1257
Active component ratio	0.0026	0.0034
Conductive component ratio	-0.0035	-0.0042
Binder component ratio	0.0008	0.0009
Minimum operating voltage	0.0559	0.0748
Maximum operating voltage	0.0054	-0.0040
Current Density	-0.2689	-0.4340
Optimal Lambda	2.8968	2.0883
Alpha	0.0056	0.0043

7.3.4 Kernel Ridge Regression

The optimisation of the kernel ridge regression models is carried out for selecting the kernel function from either Laplacian or radial basis functions. These two kernel functions have their unique parameter named sigma and gamma, respectively and these will be tuned with the given investigating ranges from **Table 7.2**. In addition, the regularization strength parameter lambda will also be optimised in each kernel function case. **Figure 7.5-(a, c)** and **Figure 7.5-(b, d) provide** the heat map results of the mean squared errors computed under the Laplacian kernel and radial basis function kernel respectively with various internal parameter settings (Sigma for Laplacian and gamma for radial basis function) for the prediction of IC. Similarly, **Figure 7.6-(a, c)** and **Figure 7.6-(b, d)** are the heat map results of the mean squared errors computed under the Laplacian kernel and radial basis function kernel respectively with various parameter settings for the EC prediction. The colour of

each square in the heat map results indicates the value of the mean squared error with more reddish indicate a higher error value and more purple-ish indicating a lower value. First, both the laplacian and the radial basis function have generated similar regions that compute the lower error in the first trial results of hyperparameter tuning for both the IC (**Figure 7.5-(a-b)**) and EC kernel (**Figure 7.6-(a-b)**) ridge regression models. These regions are coloured by purple and can be found in the range of $10^{-0.57} \sim 10^{-4}$ for the sigma/gamma values and $10^{-4} \sim 10^{1.55}$ for the lambda values. We further analysed these low error regions and demonstrated that the parameter combination is shown to generate a similar range of errors as the purple colour are very equally distributed. In addition, the optimal hyperparameters for each model are selected based on which combination would compute the lowest error during training. Across all the trial runs for each model, the laplacian kernel has seemed to generate the lowest errors than the radial basis function. **Table 7.4** demonstrate the optimal combination of hyperparameters (Sigma and the Lambda) for laplacian kernel-based ridge regression models used for the prediction of IC and EC.

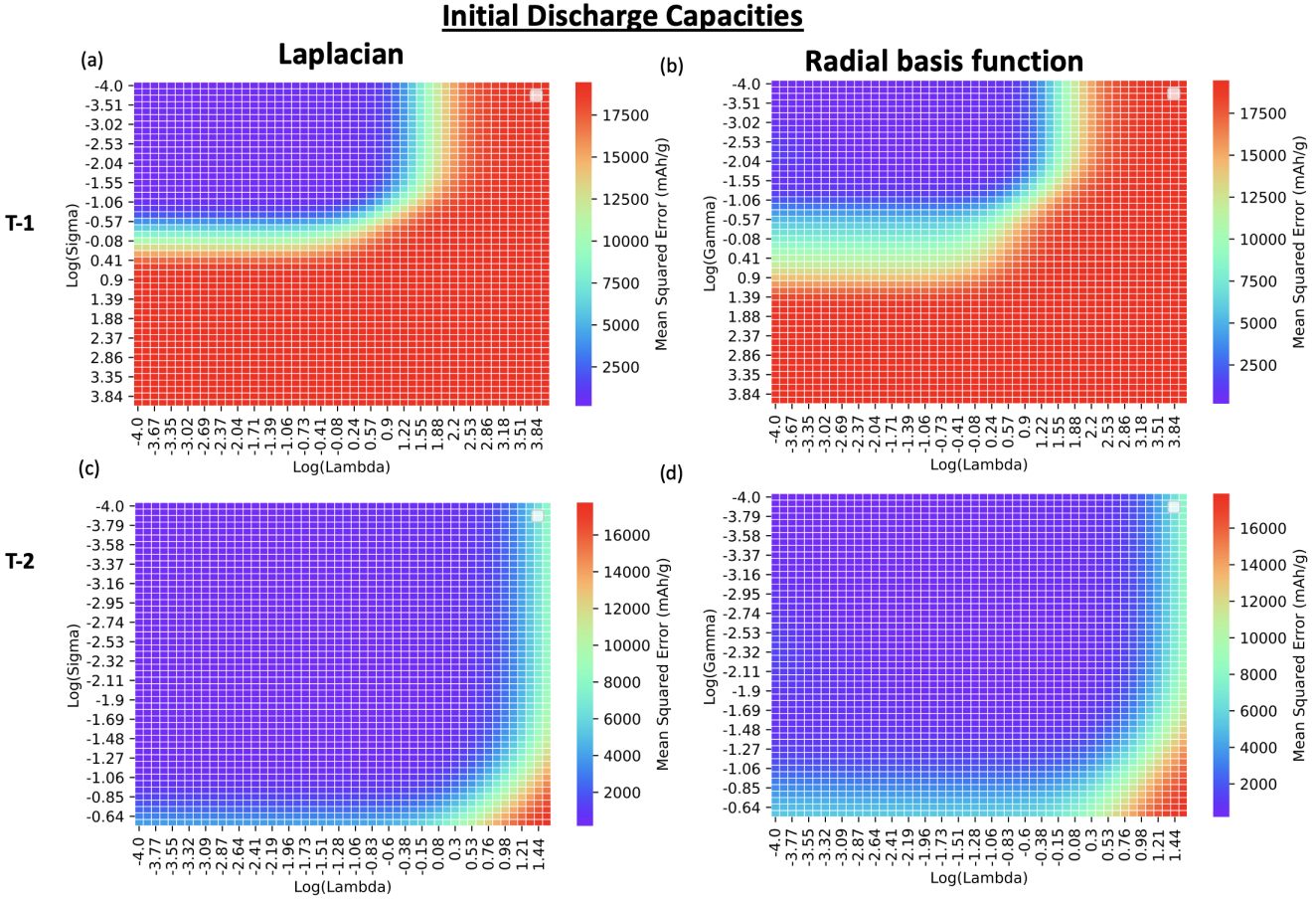


Figure 7.5: Mean squared error ($mAhg^{-1}$) from the prediction of initial discharge capacities of doped and carbon-coated $LiMPO_4$ olivine cathodes computed by Laplacian kernel-based ridge regression model with the parameter settings of the log(sigma) and log(lambda) ((a, c) and by the radial basis function kernel ridge regression model with the parameter settings of computed the log(gamma) and log(lambda) (b, d).

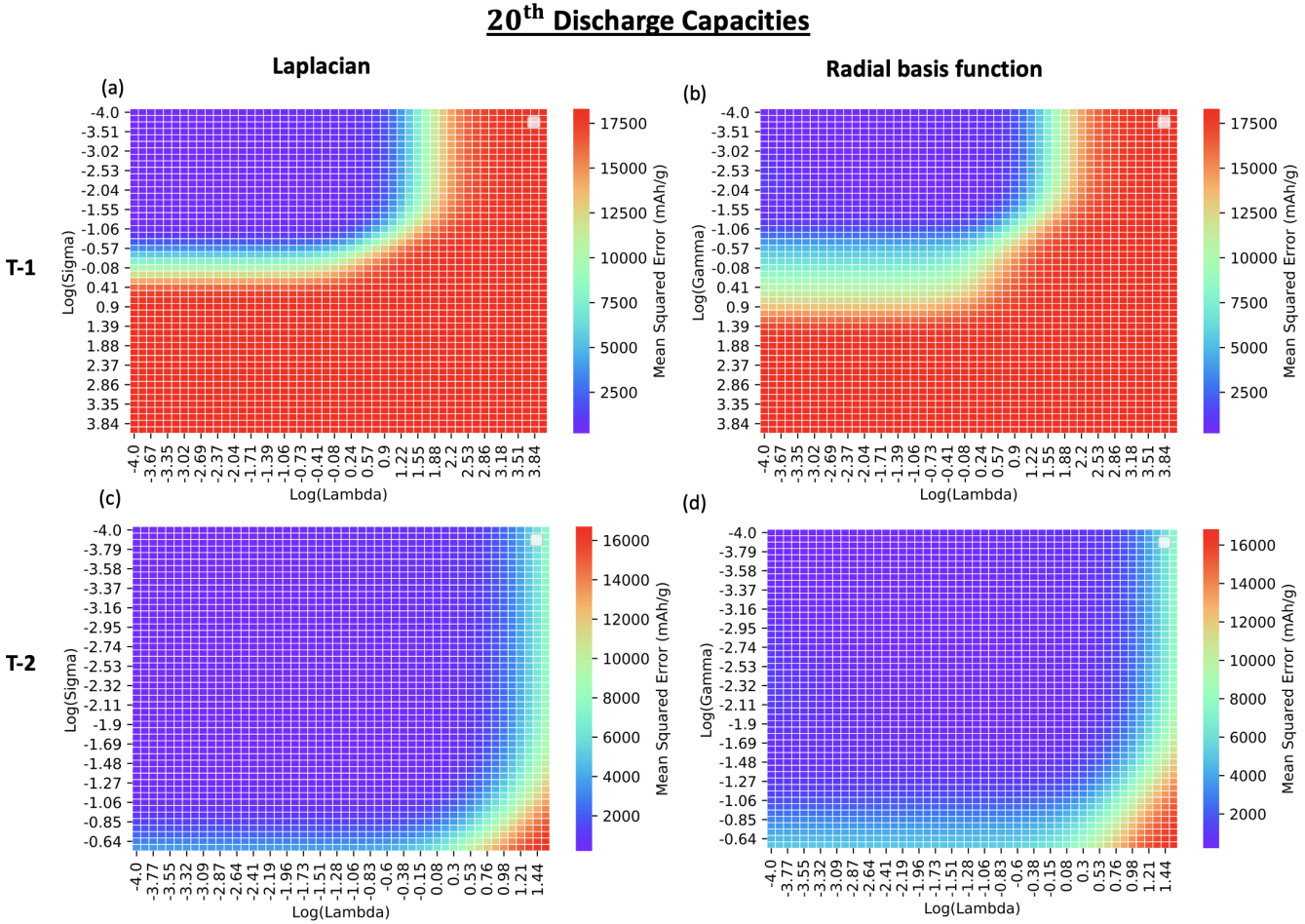


Figure 7.6: Mean squared error ($mAhg^{-1}$) from the prediction of 20^{th} cycle discharge capacities of doped and carbon-coated $LiMPO_4$ olivine cathodes computed by Laplacian kernel-based ridge regression model with the parameter settings of the log(sigma) and log(lambda) ((a, c) and by the radial basis function kernel ridge regression model with the parameter settings of computed the log(gamma) and log(lambda) (b, d).

Table 7.4: The optimal hyperparameters for the kernel ridge regression models for the prediction of initial discharge capacity (IC) and the 20th cycle discharge capacity of doped and carbon-coated $LiMPO_4$ olivine cathodes.

	Kernel	**Sigma**	**Lambda**
IC	Laplacian	0.00912	0.00229
EC	Laplacian	0.0107	0.000130

Figure 7.7-a and **Figure 7.7-b** gives the predicted value versus the observed value for the IC and EC prediction computed by the kernel ridge regression models, respectively. **Figure 7.7-a** indicates that the training error for the KRR-IC is similar to that of the BRR-IC model with a small difference of 0.20 $mAhg^{-1}$. The KRR-IC model seems to have poorer prediction power over the initial capacities under the range of 120 $\sim$140 $mAhg^{-1}$ in the test set than the BRR-IC as the points are shown to be further away from the true line. In addition, the prediction power of the KRR-IC for the test-set is shown to be worse than the BRR-IC as the $RMSE_{test}$ value is 4.51 $mAhg^{-1}$ higher. From **Figure 7.7-b**, it is seen that both the training and testing RMSE values are shown to be larger than those computed by the BRR-EC (**Figure 7.4-b**), and this shows that the KRR-EC has a poorer prediction power than the KRR-EC.

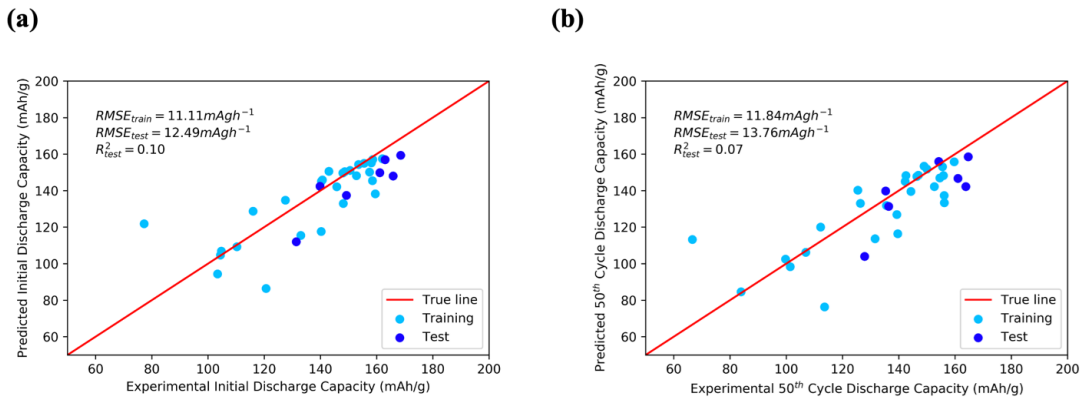


Figure 7.7: The predicted values versus the observed values for (a) the initial discharge capacity (IC) and (b) the 20th cycle end discharge capacity (EC) of doped and carbon-coated $LiMPO_4$ olivine cathodes, computed by the kernel ridge regression models.

7.3.5 Support Vector Machine

The radial basis function is chosen as the major kernel function during the training of the support vector machine and the optimisations are carried out for the hyper-parameters such as the cost and the gamma (embedded in the radial basis function). **Figure 7.8-(a,b)** and **Figure 7.8-(c,d)** give the heatmap illustrations of the mean squared error results during the optimisation of these parameters for IC and EC prediction, respectively. For the initial discharge capacity model, **Figure 7.8-a** illustrates the results of the original investigating range noted in **Table 7.2** and it is seen that the region that can output the lowest prediction errors (coloured in blue), corresponds to the range of $10^{-3} \sim 10^{0}$ for gamma and $10^{1} \sim 10^{3}$ for the cost. Similarly for the EC model, the lowest error prediction errors regions can be identified to be $10^{1} \sim 10^{4}$ for the cost and gamma to be greater than 10^{-1}. Next, further optimisations are carried out with these identified parameter ranges and the results are given in **Figure 7.8-b** and **Figure 7.8-d** for IC and EC, respectively. For the IC prediction, smaller cost values are preferred within the range of $10^{1.25} \sim 10^{1.86}$ and along with the gamma range of $10^{-2.2} \sim 10^{-0.55}$. On the other hand, the EC model can adapt a much wider range of gamma values ($10^{-3.51} \sim 10^{-1.43}$) and cost values ($10^{1.55} \sim 10^{4}$). The final selected hyperparameters for these models are given in **Table 7.5**.

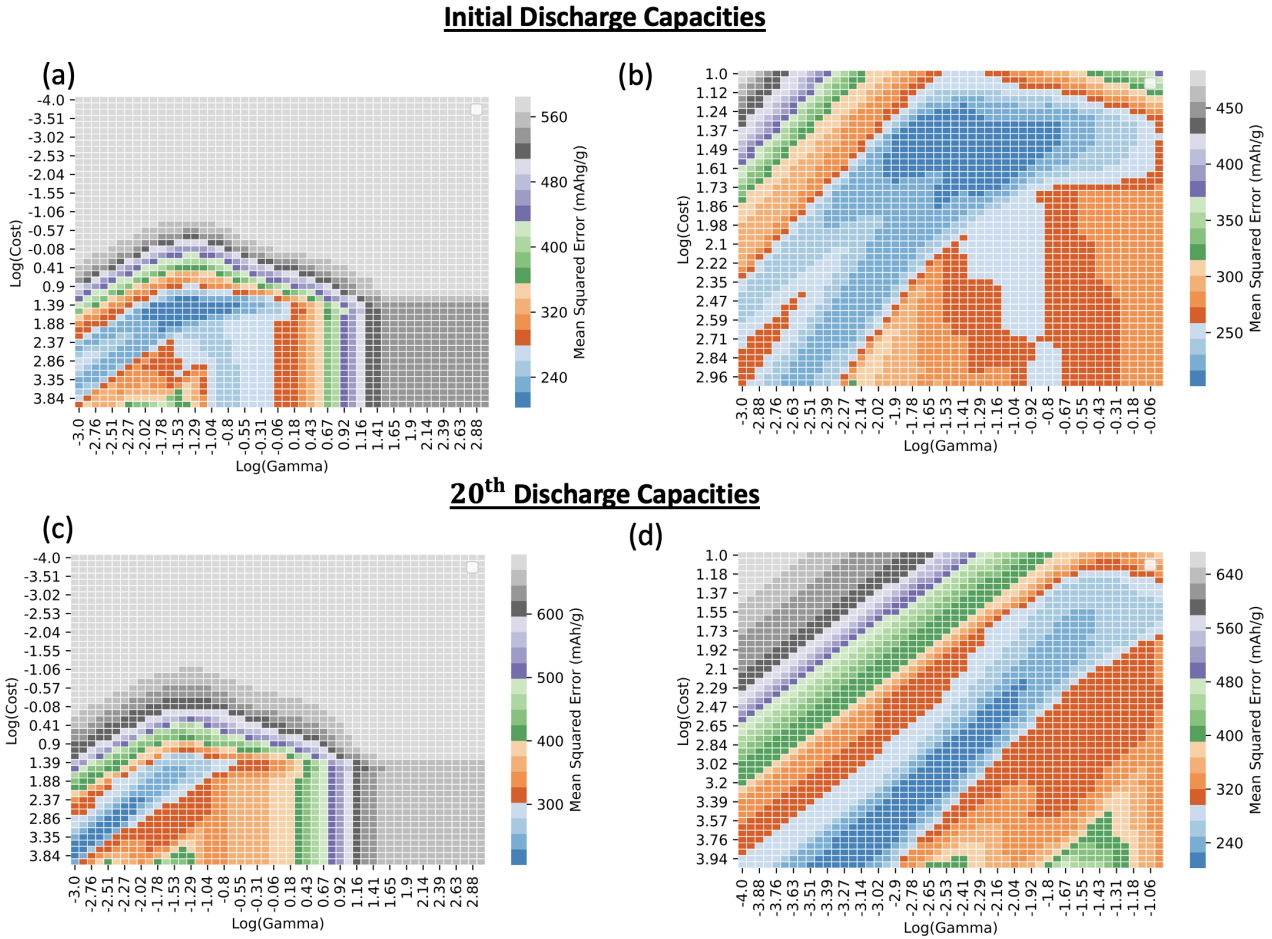


Figure 7.8: The heat-map illustration of the optimisation results for the cost and the gamma parameters in logarithmic scale for the prediction of (a-b) the initial discharge capacities and of (c-d) the 20th cycle end discharge capacities of doped $LiMPO_4$ olivine cathodes. (a-b) corresponds to the investigating range noted in Table 2 and (c-d) corresponds to the investigating gamma range of 10^{-3} $\sim 10^{0}$, 10^{-4} $\sim 10^{-2}$ for IC and EC and cost range of 10^{1} $\sim 10^{3}$, 10^{1} $\sim 10^{4}$ for IC and EC, respectively.

Table 7.5: The final optimised hyperparameter for the support vector machine regression model for the prediction of initial discharge capacity (IC) and of the 20th cycle end discharge capacity (EC) of doped and carbon-coated $LiMPO_4$ olivine cathodes.

	Optimal hyperparameters	
	Cost	Gamma
IC	104.811	0.0910
EC	232.995	0.0595

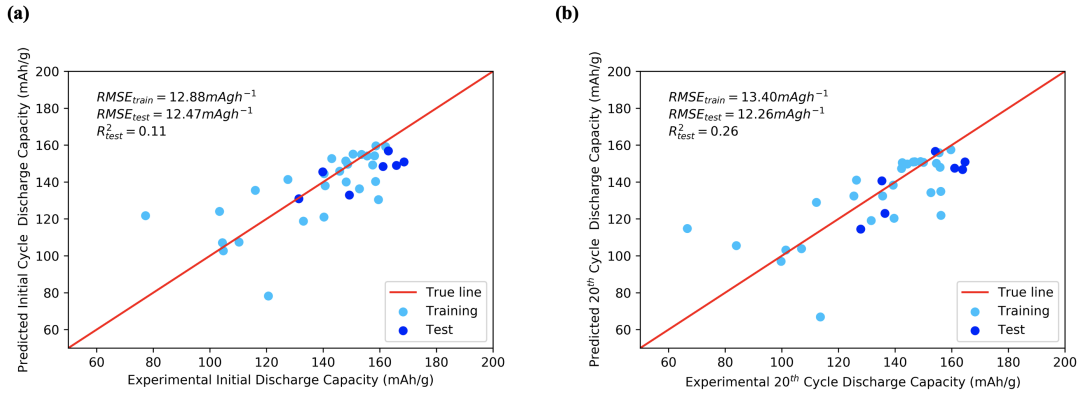


Figure 7.9: The predicted values versus the observed values for (a) the initial discharge capacity (IC) and (b) 20th cycle end discharge capacity (EC) of doped and carbon-coated $LiMPO_4$ olivine cathodes, computed by the support vector regression models.

Next, the prediction power of the two fully optimised support vector machines models, SVM-IC and SVM-EC are assessed through predicting against the hold-out test set. **Figure 7.9** illustrates the predicted values against the observed values for initial and 20th cycle end discharge capacities, respectively. The prediction power of SVM-IC for both the training set and the test set has shown to be worse than the previous KRR-IC models. Both the training and the testing RMSE values from **Figure 7.9-a** are much larger than the error indicated in **Figure 7.7-a**. The R^2 scores for the test-set prediction are also shown to be much smaller in **Figure 7.9-a** which indicate that the support vector machine model does not capture the variations of response variables in the test set well. On the other hand, the SVM-IC seems to be able to predict well for the one initial capacity point between 120$\sim$140 $mAhg^{-1}$ in the test set. This test-set point has commonly been either overly predicted (BRR-

IC) or underpredicted (KRR-IC). Furthermore, the SVM-EC model has shown a much lower RMSE value for the test-set (12.26 $mAhg^{-1}$) than for the training set (13.40 $mAhg^{-1}$). The reason behind this high training error is thought to be the persistent low capacities observed from the range of 60$\sim$85 $mAhg^{-1}$ and the one outliner noticed between 100$\sim$120 $mAhg^{-1}$. The prediction against the test set has not shown any promising improvement from the previous KRR-EC model with the R^2 scores estimated to be 0.26.

7.3.6 Random Forest

The optimisations of the random forest models are carried out for the two major hyperparameters, namely, the number of trees and the maximum number of features/covariates considered at each split. Other tree-based parameters such as the minimum amount of samples for splitting at each leaf and the minimum amount of samples to remain at each leaf are set to be 2 and 1, respectively. **Figure 7.10-(a,b)** illustrate the mean squared error computed for every combination of the hyperparameters values within the ranges identified in **Table 7.2**. **Figure 7.10-a** illustrates the optimisation results of the RF-IC model for the prediction of initial discharge capacity of the doped and carbon-coated olivine systems. In addition, **Figure 7.10-b** gives the optimisation results computed by the RF-EC model for the prediction of the 20^{th} cycle discharge capacities of doped and carbon-coated olivine systems. To begin with, it is seen that both models are more sensitive to the change in the covariates than to change in the number of trees as the colour is more even along the x-axis. In addition, an observation can be made that the RF-IC model seems to suggest a higher number of features than the RF-EC model as the region of low covariate numbers is shown to be yellow colour. The optimal set of these two hyperparameters can be identified through the check of the position of the deep-red troughs. **Table 7.6** summarises the optimised hyperparameters that lead to the lowest errors from **Figure 7.10**. These hyperparameters will be used to train the models and assessed their prediction power through the hold-out test set.

Figure 7.11-a and **Figure 7.11-b** give the predicted values versus the observed value in the test set computed by the optimised RF-IC and RF-EC, respectively.

Figure 7.11-c and **Figure 7.11-d** provide graphical illustrations of the random-forest computed variable importance for every covariate in the prediction for IC and EC, respectively. The identifications of each index covariate are given in **Table 7.7**. To begin with, the overall prediction performance of RF-IC is seen to improve from the previous SVM-IC model with its RMSE values calculated to be 11.77mAhg^{-1} and 10.63 mAhg^{-1}, respectively. The reduction in the training set RMSE error is believed to be caused by the more accurate prediction of the capacities in the lower range (60 $\sim$ 80 mAhg^{-1}) and middle-ranged capacity at around 120 mAhg^{-1}. The prediction over the test-set has shown an improvement comparing to the previous SVM-IC model as its RMSE-test is calculated to be 1.84mAhg^{-1} lower and the R^2 score has increased from 0.11 to 0.35. In addition, it is seen that the three weighted elemental properties, Avg_E_ionisation (Index Number: 3), Avg_Electronegativity (Index Number: 4) and the Avg_IR (index number: 5) are shown as the most important covariates for RF-IC to make good predictions in the test-set. All of the other covariates are shown as much less important as the length of their bars are much shorter as shown in **Figure 7.11-c**. Further to the prediction of the RF-EC model, better prediction power for the middle-ranged capacity at around 120mAhg^{-1} is observed as the respective points in both the training and testing sets are shown closer to the true line than other models' cases (**Figure 7.7-a** and **Figure 7.7-b**). The overall predictions over the test-set are seen fairly accurate as the test-set dots (coloured in deep blue) are closer to the true line. In addition, the variable importance graph (**Figure 7.11-d**) suggest that the three elemental-properties covariates remain as the top three most important covariates for RF-IC to make predictions for the EC values in the test-set. On the other hand, the results graph also indicates that other crystal unit cell parameters are also essential in the prediction of EC. In particular, the variable of the volume of the unit cell (CV) has been ranked nearly as important as the Avg_E_ionisation which is followed by the lattice parameter "c" of the unit cell.

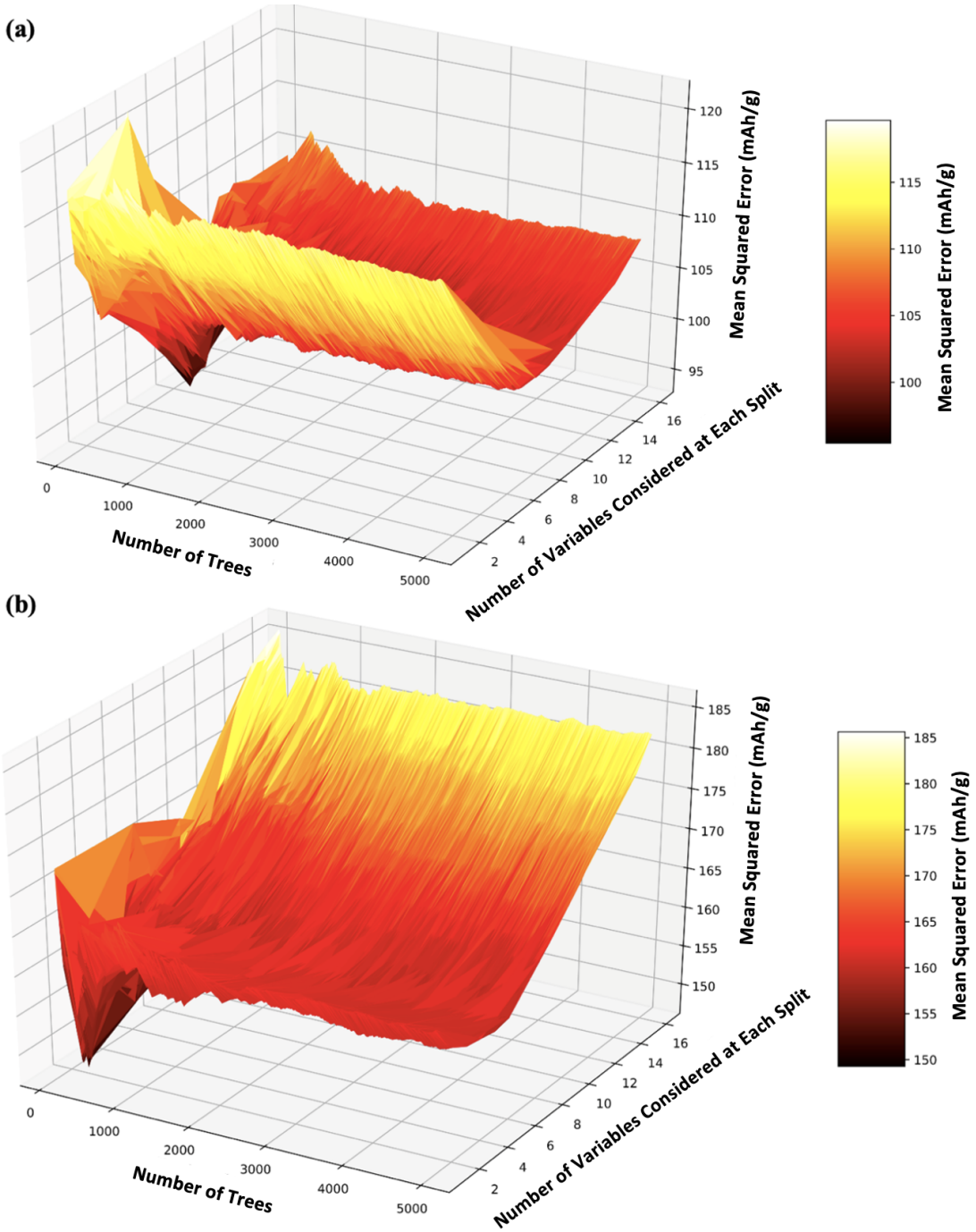


Figure 7.10: Three-dimensional plot of the mean squared error against the number of trees and number of features considered at each split of random forest models for the predictions of (a) Initial discharge capacity (b) 20^{th} cycle end discharge capacity of doped and carbon-coated $LiMPO_4$ olivine cathodes, respectively.

Table 7.6: The optimised hyperparameters for the random forest models including the number of trees and the maximum number of features considered at each split for the prediction of initial discharge capacity (IC) and the 20th cycle discharge capacity (EC) of doped and carbon-coated $LiMPO_4$ olivine cathodes.

	IC	EC
Number of trees	240	156
Maximum number of covariates considered at each split	8	2

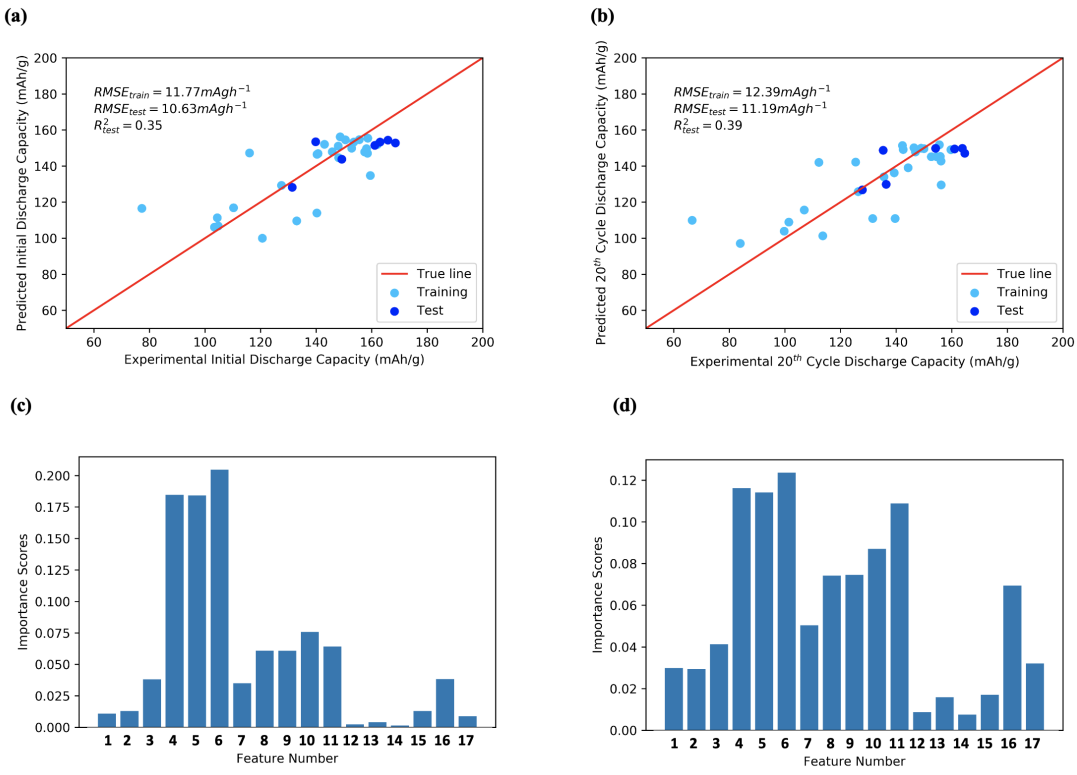


Figure 7.11: The predicted values computed versus the experimental values for (a) the initial discharge capacity (IC) and (b) the 20th cycle end discharge capacity (EC) of doped and carbon-coated $LiMPO_4$ olivine cathodes, computed by the random forest models and the ranking of the variable importance during the prediction of the hold-out test set for (c) IC and (d) EC, respectively.

Table 7.7: Identification of the covariates used in the olivine cathode model.

Feature Number	Covariate Variables	Feature Number	Covariate Variables
0	Base-metal ratio's formula ratio	**9**	The lattice constant "c"
1	Dopant's formula ratio	**10**	Crystal lattice volume
2	The molar mass of the cathode material	**11**	The ratio of the active component
3	Weighted mean of the first ionisation energy between the base-metal ratio and the dopant ratio	**12**	The ratio of the conductive component
4	Weighted mean of the electronegativity values between the base-metal ratio and the dopant ratio	**13**	The ratio of the binder component
5	Weighted mean of the ionic radius between the base-metal ratio and the dopant ratio	**14**	Minimum operating voltage
6	Carbon coating content	**15**	Maximum operating voltage
7	The lattice constant "a"	**16**	Current density
8	The lattice constant "b"		

7.3.7 Gradient Boosting Machine

For the optimisation of the gradient boosting machine models, four major hyperparameters are investigated, namely, the number of trees (N_trees), the maximum number of features considered at each split (Max_features), the maximum interaction depth (Max_depth) and the learning rate for scaling the contribution of every tree in the model. **Table 7.8** and **Table 7.9** summarise the trials of optimisation results including the investigating ranges for each hyperparameter, and the lowest computed mean squared error from those trials of results. The investigating ranges after the first initial trial are selected based on the best performing 10 sets of hyperparameter combinations from the previous trial. In **Table 7.8**, it is seen that the minimum mean squared error is reduced by a slight amount from T-1 to T-2.

Table 7.8: Results of two trials (T-1, T-2) of results of hyperparameter tuning results for the gradient boosting models for the prediction of initial discharge capacity (IC) of doped and carbon-coated $LiMPO_4$ olivine cathodes, where N_trees indicates the number of trees, max_features indicate the maximum number of covariate features should be considered at each split, max_depth denotes for the maximum tree depth in the random forest, the learning rate is the scaling factor and the Min_MSE represents the minimum mean squared error obtained from the investigating range. Variables of min_samples_split and min_samples_leaf are kept at the default setting of 2 and 1 respectively.

	T-1	**T-2**	**Optimal**
N_trees	$100 \sim 5000$	$500 \sim 4500$	1041
Max_features	$1 \sim 16$	5,7,11,12,13,15,16	5
Max_depth	$1 \sim 20$	$2 \sim 5$	3
Learning_rate	$10^{-1} \sim 10^{-4}$	$10^{-1} \sim 10^{-3}$	0.0153
Min_MSE ($mAhg^{-1}$)	80.119	79.453	

Table 7.9: Results of two trials (T-1, T-2) of results of hyperparameter tuning results for the gradient boosting models for the prediction of 20^{th} cycle discharge capacity (EC) of doped and carbon-coated $LiMPO_4$ olivine cathodes, where N_trees indicates the number of trees, max_features indicate the maximum number of covariate features should be considered at each split, max_depth denotes for the maximum tree depth in the random forest, the learning rate is the scaling factor and the Min_MSE represents the minimum mean squared error obtained from the investigating range. Variables of min_samples_split and min_samples_leaf are kept at the default setting of 2 and 1 respectively.

	T-1	**T-2**	**Optimal**
N_trees	$100 \sim 5000$	$500 \sim 4500$	2432
Max_features	$1 \sim 16$	$1 \sim 6$	1
Max_depth	$1 \sim 20$	$2 \sim 20$	2
Learning_rate	$10^{-1} \sim 10^{-4}$	$10^{-1} \sim 10^{-3}$	0.0105
Min_MSE ($mAhg^{-1}$)	131.333	131.400	

It is important to mention that the GBM-IC model is more sensitive towards the change in the maximum number of interactions and the learning rate than towards the N_trees during the training process. The optimal hyperparameters for the GBM-IC model are selected from the T-2 trial and are highlighted in red. In **Table 7.9**, the minimum mean squared error results from T-2 has not seemed to change much compared to the result of the previous T-1 trial. The optimised hyperparameters for the GBM-EC are highlighted in red.

Next, the prediction powers of both GBM-IC and GBM-EC models will be

assessed through their prediction against the hold-out test set. **Figure 7.12-a** and **Figure 7.12-b** give the graphs of the predicted value against the observed values for IC and EC, computed by the GBM-IC and GBM-EC, respectively. First, the prediction power of the GBM-EC model for the test-set is much higher than the GBM-IC model as the test-set RMSE is calculated to be much lower. The prediction performance of the GBM-IC for both the training and testing set is shown to be very poor as the data points in **Figure 7.12-a** are far from the true line. On the other hand, the GBM-EC model seems to hold good prediction power over the test set as its RMSE value is calculated to be 9.19mAhg^{-1} whilst the R^2 score is computed to be 0.59. There are a few outliners identified at the lower capacity range (60$\sim$80 mAhg^{-1}) and middle capacity range (100$\sim$120 mAhg^{-1}), which increases the value of the training RMSE. The importance of the covariate in GBM-IC and GBM-EC are ranked and given as **Figure 7.12-c** and **Figure 7.12-d**, respectively. In **Figure 7.12-c**, it is seen that the three elemental properties are ranked as the top three important variables, and this agrees with the conclusion from **Figure 7.11-c.** In addition, **Figure 7.12-d** seems to suggest that Avg_Electronegativity (index number: 4) and Avg_IR (index number: 5) are the two most important covariates which agree with the previous findings from **Figure 7.11-d**. On the other hand, the lattice constant "a" (LC_a, index number: 7) seems to be the third most important variable for the prediction of EC instead of the Avg_E_ionisation (index number: 3). In general, one can conclude that the predictions of the IC from both RF-IC and GBM-IC model are mainly made from the three elemental properties, whereas the EC predictions require much broader covariate variables as the relative size of the bars are shown much spread out among different covariates in **Figure 7.11-d** and **Figure 7.12-d**.

7.3.8 Model Performance Comparisons

In this project, five different regression algorithms are used to train and build the models for the datasets Mn and Fe based materials. **Table 7.10** summarises the RMSE values computed for the training and testing set of every optimised model, along with the R^2 scores for the test set predictions. It is seen that the BRR mod-

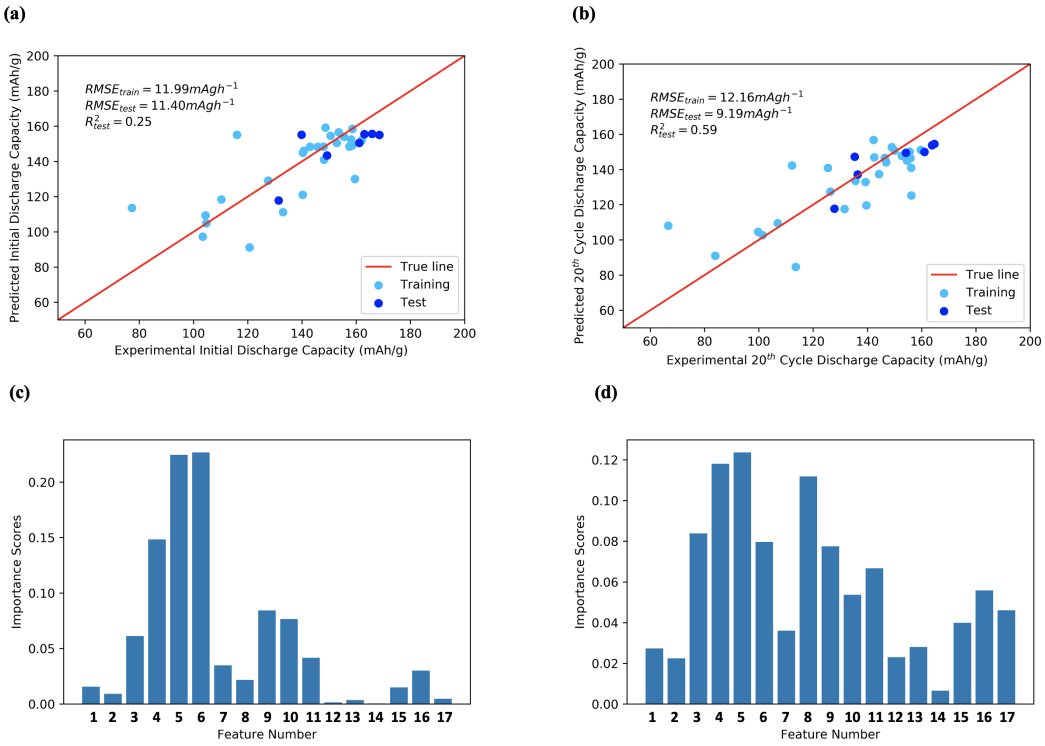


Figure 7.12: The predicted values computed versus the experimental values for (a) the initial discharge capacity (IC) and (b) the 20th cycle end discharge capacity (EC) of doped and carbon-coated $LiMPO_4$ olivine cathodes, computed by the gradient boosting models and the ranking of the variable importance during the prediction of the hold-out test set for (c) IC and (d) EC, respectively.

els seem to have the highest prediction power for IC among all other models with their test-set RMSE values calculated to be the lowest (9.52mAhg^{-1}) and the R^2 score (0.48) calculated to be highest among all models. In the case of EC prediction, the GBM seems to have slightly better prediction performance than the BRR model as the RMSE values for both the training and testing set are shown to be 0.78mAhg^{-1} and 0.06mAhg^{-1} lower, respectively. Similarly, the R^2 score for the test-set prediction is slightly better for the GBM model than for the BRR model. BRR algorithm is known for its advantages of efficient adaption to the dataset as well as the implementation of the regularization technique to reduce the likelihood of model overfitting to any dataset. On the other hand, the GBM algorithm is preferred for its robust power in explaining the non-linear relationships of the variables in the dataset.

Table 7.10: Comparisons of the mean RMSE values during the 10-fold cross-validation and for testing against holdout test and the R^2 test score computed by five ML models, for the prediction of the initial discharge capacity and the 20^{th} cycle end discharge capacity of doped and carbon-coated $LiMPO_4$ olivine cathodes.

ML Techniques	Initial Discharge Capacity ($mAhg^{-1}$)			20^{th} 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**
BRR	11.34	9.52	0.48	12.94	9.25	0.58
GBM	11.99	11.40	0.25	12.16	9.19	0.59
RF	11.77	10.63	0.35	12.39	11.19	0.39
SVM	12.88	12.47	0.11	13.40	12.26	0.26
KRR	11.11	12.49	0.10	11.84	13.76	0.07

7.3.9 Variable Importance Analysis

Next, investigations of correlations of every covariate to the capacity prediction can be made through estimating their contribution towards the model prediction against the test-set. In this project, further analysis will be carried for the BRR-IC model and the GBM-EC model as they were tested to be the best-performing ones among all. BRR model is a linear model that outputs the coefficients, or the weights of every covariate involved in the prediction. By comparing the weight value for each covariate where the values have been standarised, we can estimate the correlations of the covariates for the target IC. Furthermore, Shapley values can be computed using the SHAP package to estimate the prediction contributions of each covariate to the GBM-EC's predictions on the test set. The more positive or more negative the respective Shapley value, the more contribution the instance of the covariate has on the prediction of that EC instance.

Table 7.3 summarises the computed coefficients of every covariate used for the prediction of IC. To begin with, it is seen that the Avg_E_ionisation is ranked as the most important variable with its coefficient estimated to be highest at 0.6573. This value suggests that the higher initial capacity can be achieved through an increase in the weighted average ionisation energy between the base metal and the

dopant. For the olivine systems, the conditions of the material crystallinity are commonly governed by how well-defined the peaks of the M-O (metal and oxygen bond) from their octahedron networks show on their Fourier transform infrared spectrum (FT-IR) [323]. The higher the crystallinity of the olivine materials, the better the electrochemical performance it would be as there are no other material impurities present. Recent research has confirmed that the large difference of the ionisation energy between the inserted dopant and the base-metal material can induce peak splitting effects which produce many well-defined MO_6 bands in the FT-IR at around 650cm^{-1} and consequently reflecting on higher material crystallinity [324, 325]. Our results have suggested that a higher weighted mean ionisation energy between the base-metal and the dopant metal would lead to a higher IC. In addition, the Avg_IR and carbon_content seem to be ranked as the second and the third most important variables for the BRR-IC model. The increase in the Avg_IR in the olivine system could lead to a larger unit cell volume and additionally expand the [010] pathway to promote Li-ion transfer. Interestingly, the coefficient results of the unit cell volume and the lattice parameters had also suggested the reverse trend where the reductions in the unit cell dimensions and size would lead to a larger IC. However, the contributions of these unit cell parameters to the overall predictions are shown to be small as their coefficients are computed to be two-fold smaller than Avg_IR. Furthermore, an increase in the amount of carbon coating content is suggested to lead to a higher IC as the respective coefficient is calculated to be 0.266. However, it is also suggested that a thick carbon coating would also greatly inhibit the Li-ion from penetrating the cathode and lead to lower capacity performance. Other small correlations can be observed that the higher base-metal content and smaller dopant content is preferable for reaching high IC. This is considered reasonable since the high content of dopant would increase the likelihood of blocking the [010] diffusion path and constrain the kinetics of Li-ion diffusion within the olivine system. In conclusion, the covariates that relate to the doping treatment (Avg_IR, Avg_E_ionisation) have much higher coefficient values than the variable of carbon coating content, which might suggest their greater influence on the initial

discharge performance of the doped olivine system.

Figure 7.13 gives a graphical illustration of the Shapley values computed for every covariate during the prediction of each EC instance in the test-set using GBM-EC. The order of the covariates on the y-axis shows the ranking of the importance of the covariate for the predictions with the higher indicating greater importance towards the model prediction. To begin with, the Avg_electronegativity is ranked as the top covariate by the GBM-EC. The values of electronegativity can determine the characteristics of the bond formed with the other element. From the results, it is observed that the increase in the metal elements' (include the base metal and the dopant metal) average electronegativity values seem to improve the respective EC as the colour is changing points are changing from blue to red. An increase in the weighted metals' elements electronegativity in the olivine system can lead to the formation of more covalent bonds with the surrounding P and O atoms. The covalent bonding characters will make the M-P bonds become longer to push further the PO_4 tetrahedra groups. This will eventually reduce the [010] diffusion path for Li-ions to travel through, promoting the efficient transfer of the Li-ions and hence leading to a higher discharge capacity [326]. Similarly, an observation of an increase in the Avg_IR result to a larger EC can also be explained as the increase in the ionic radius would increase the bond length with surround P atoms from the reduced charge density and hence reducing the Li-ion diffusion path. Higher carbon coating content is shown to be more desirable as it increases the EC values. The underlying reasons could be thought as the reduction in the metal dissolution reactions with the electrolyte, which occurs during the frequent charging cycle, and this would both destabilise the cathode's Li-ion hosting structure and reduce the kinetics of Li-ion transfer from the surficial crystal growth. This finding seems to agree with the previous correlation of IC results from **Table 7.3**. In addition, the contractions of both the lattice dimension 'a' (LC_a) and "c" (LC_c) are shown to lead to higher EC values. Previous research has reported that the shortening of the dimensions along "a" and "c" axes would narrow the [010] diffusion channel in the olivine systems and hence boost up their electrochemical performance [326]. In ad-

dition, the correlations of the material formula molar mass (Mr) with the 20th cycle discharge capacity (EC) cannot be confirmed due to the unevenly distributed data trend; some datapoints with lower formula molar mass seems to provide generate a higher EC than the ones with higher formula molar mass. Other useful observations can be made such as that the higher EC can be achieved with smaller dopant content and higher base-metal content. For the mixing ratios of binder, active component and the conductive, the only clear correlation is observed for the binder content where the less respective binder content is suggested for reaching a higher EC value. In conclusion, the Shapley values computed for the elemental properties related variables such as Avg_EN, Avg_IR are much larger than the carbon coating content variable which potentially implies that the dopant plays a bigger role in the prediction of 20th cycle discharge capacity.

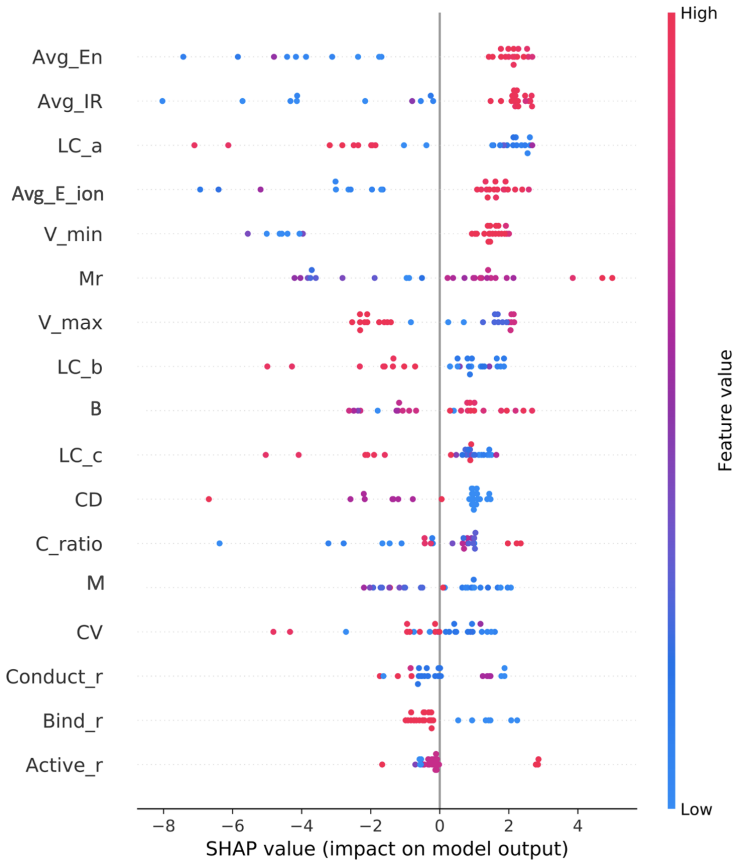


Figure 7.13: The summary plot of the Shapley values computed from the gradient boosting model for the prediction of the 20th cycle end discharge capacity of the carbon-coated, doped cobalt olivine systems in the test set.

7.4 Summary

Understanding the relative importance of the doping treatment and the carbon coating is vital for the development of better olivine materials. In this project, we trained and optimised ten machine learning models from five underlying machine learning algorithms including Bayesian ridge regression, support vector machine, kernel ridge regression, random forest and the gradient boosting machine to predict for the initial and the 20^{th} cycle end discharge capacity of carbon-coated, doped, for the iron-based and the manganese-based olivine systems. Among all models, the Bayesian ridge regression model has shown greater prediction power for the initial discharge capacity with its root-mean-square-error computed to be lowest at 9.54 mAhg^{-1} in the hold-out test-set. In addition, the gradient boosting machine model has demonstrated good prediction power for the 20^{th} cycle discharge capacity at 9.25 mAhg^{-1}. The variable importance analysis reveals that the doping treatment has played a greater role in performing these predictions than the carbon coating content. In the end, the variable correlations suggest higher initial and 20^{th} cycle discharge are more attainable for the olivine systems should have higher base-metal ratio, lower dopant ratio, smaller crystal dimensions and are doped with dopant atoms that have higher electronegativity, ionic radius and ionisation energy.

Chapter 8

Overall Discussion, Conclusions and Future Work

8.1 General Discussion and Conclusions

The market for lithium-ion batteries (LIB) is undergoing a big transformation as the renewable energy industry demands devices with cheaper prices, higher capacity, higher charge/discharge rate and longer lifetimes. As such, finding new cathode material has become the key as it is known to be the constraining factor for the LIB performance. The conventional materials discovery method is very inefficient and costly due to the slow processes in the trial-error experimenting and battery performance testing. Leveraging the high volume of experimental data from over 7000 publications made each year on LIB materials, this thesis mainly investigates the power of using various machine learning techniques for the prediction of the discharge performance based on a collection of variables such as the systematic materials properties and the experimental conditions. Three promising classes of doped cathode materials are investigated in this thesis, namely, layered materials, spinel cathode and carbon-coated olivine cathodes. In addition, the underlying complex variable correlations have been analysed to identify the governing material variable for both the initial and the higher cycle of discharge capacities. In this section, an overall discussion over the three results chapter will be given from the aspects of data quality, quantity as well as the model outlook. To aid with the discussion, a

Table 8.1: A comparison of the size of the dataset, model outlook and the best performing ML model for each of the aforementioned results chapters.

Name	**Doped LMO cathodes**		**Doped NCM cathodes**		**Doped and Carbon-coated Olivine cathodes**	
Dataset size	102		168		34	
Number of Covariates	6		20		17	
Response Variable	IC	20^{th} cycle discharge capacity	IC	50^{th} cycle discharge capacity	IC	20^{th} cycle discharge capacity
Best Performing Model	GBM	GBM	GBM	GBM	BRR	GBM

comparison of these aspects is given in **Table 8.1**.

The prediction power of any machine learning model is closely related to the quality and the quantity of the data used during the training process. To ensure the high quality of data being collected, several selection rules have been implemented in the three projects reported. The final curated size of the data used for **Chapter 5**, **Chapter 6** and **Chapter 7** are indicated in **Table 8.1** as 102, 168 and 34, respectively. A clear observation can be seen that the dataset for the NCM project is much bigger than both the doped spinel project and the doped olivine project. The reasons are thought as doped NCM cathodes have more available transition metal doping sites than the other two cathode systems, which implies that there is more possible scope to study the dopant effects on the intrinsic properties. In terms of electrochemical performance, the layered materials receive more attention among the three cathode systems as it has the highest capacities and widest operating voltage range. On the other hand, the collation of feasible data for doped olivine cathode systems has been very difficult because most publications do not report key information such as the carbon coating content on the surface of the cathode materials nor the mixing ratio of binder/active/conductive components in cathode systems. In addition, the research for the olivine system is seen as less popular because it involves extra consideration of carbon precursors, coating process and doping strategy, which are more complex than the other two projects. During the collection

of datasets for all three projects, it is important to note that a lot of publications have used different current densities for discharging than charging and these were neglected from the data collection process to ensure a more standardised dataset. However, this has largely reduced the number of publications qualifying for the collection criteria.

In Chapter 5, 102 doped spinel cathode ($LiM_xMn_{2-x}O_4$) materials with 17 dopant variations were investigated using seven machine learning techniques for the correlations of six material-related properties and their discharge performance. Gradient boosting machine models have shown the best prediction power for predicting the initial and the 20^{th} cycle discharge capacity with test-set error computed to be 11.41mAhg^{-1} and 11.28 mAhg^{-1}, respectively. Among all covariate variables, material formula molar mass is estimated to be the most important covariate by both gradients boosting models and the additional variable importance study reveals that doped spinel cathode with higher material formula molar mass, shorter crystal lattice constant "*a*" and doping with smaller electronegativity value dopant can lead to higher discharge capacities in both cycles.

In Chapter 6, 168 doped nickel-cobalt-manganese (NCM) materials with twenty dopant variations were investigated using six non-linear machine learning algorithms for correlation of twenty covariate variables (including material structural properties, experimental conditions and dopant elemental properties) with their discharge performance. Among all investigated models, gradient boosting models have shown the best prediction power over the initial and 50^{th} cycle discharge capacities, with their test-set error estimated to be 16.66 mAhg^{-1} and 18.59 mAhg^{-1}, respectively. In addition, artificial neural networks models are shown to have the worst prediction power among all models. The variable correlation study reveals that Greater initial and 50^{th} cycle discharge capacities can be achieved by increasing the Li content in the doped NCM formula, reducing the overall material molar mass, increasing the nickel content and doping the system with a small amount of low electronegativity dopant.

In Chapter 7, a mixture of 34 carbon-coated and doped olivine materials con-

sists of 24 $LiFePO_4$ systems and 10 $LiMnPO_4$ systems are investigated using five different machine learning algorithms for the correlations of 17 material related covariate variables (carbon coating content, crystal structure lattice constant.. etc) and the respective discharge performance. Among all models, the Bayesian ridge regression model and the gradient boosting model have shown the best prediction power for the initial and the 20^{th} cycle discharge capacity, respectively, with each error calculated to be 9.52 $mAhg^{-1}$ and 9.19 $mAhg^{-1}$. These optimal models have been used to predict the capacity of carbon-coated and doped cobalt-based systems and resulted in capacity error percentages of 6% and 6.1% (on average) than the reported values. Furthermore, higher values of both capacities can be achieved by altering the material system to have higher base-metal content, lower dopant content, smaller unit cell lattice dimension and are additionally doped with dopant atoms of higher electronegativity, higher ionisation energy and a larger ionic radius.

In conclusion, all of the results present in this thesis have demonstrated the benefits of implementing machine learning techniques into the LIB research; i) helps to construct highly predictive models to given fast predictions over the material properties; ii) illustrate the highly complex correlations of the doping-induced changes in material systematic properties and their respective discharging performance; iii) help to propose the potential direction of new cathode properties that could lead to better discharging performance. This includes the selection of dopant, material formula stoichiometries and crystal structural properties for all three investigated cathode classes. Non-linear models such as gradient boosting machines models have shown superior prediction performance for the discharge performance of two of the material classes and this potentially suggests the nature of non-linear material property correlations. This thesis can be seen as a milestone in achieving the predictions of discharge performance and is useful in shortening the testing stage as well as facilitating the discovery of new highly functional cathode materials for Li-ion battery applications.

8.2 Machine Learning Model Choices

All of the investigated projects share a common goal of investigating the correlations between the material's structural properties and the short and long cycle discharge capacities of the cathode material. Since all three cathode materials have similar electrochemical reactions and underlying discharging mechanisms, it is assumed that the underlying variable correlations between the materials' structural properties and the performance properties should be similar. In addition, this could then lead to the assumption that the knowledge of the prediction power of certain ML algorithms prediction from one project could be used to guide the selection of the algorithms for the next project.

Chapter 5 investigated the prediction power of both linear and non-linear algorithms as to not include any initial bias in the algorithm selection. This includes the investigation of ridge regression, LASSO regression, ANN, gradient boosting model, etc. After comparing the prediction errors of each model, it was concluded that the non-linear models such as support vector machine and gradient boosting model have much higher prediction power than linear models such as the ridge regression models. From this, an assumption was made that the correlations between the materials' properties and the discharge capacities are in a non-linear manner for the NCM materials investigated in **Chapter 6**. In addition, the ANN model has demonstrated much poorer prediction performance than other tree-based models and this might be due to the limited dataset used for training. After gaining these insights, only six different non-linear algorithms were investigated in **Chapter 6** and all of the linear algorithms investigated in **Chapter 5** were excluded. ANN algorithm was again investigated as the size of the dataset implemented in **Chapter 6** is much bigger than in **Chapter 5**. The tree-based model, gradient boosting machine demonstrated the highest prediction power among all models and this agrees with the findings from **Chapter 5**. In addition, ANN has again shown the worst prediction power among all non-linear models which then suggest that the current dataset is still very limited to unleash its full prediction power. This insight then suggests that the ANN model should not be further investigated in the next project.

In **Chapter 7**, the investigated ML models are very similar to the ones investigated in **Chapter 6** besides the inclusion of a new linear Bayesian ridge regression model. Bayesian ridge regression algorithm is known for being less prone to overfitting issues as well as highly adaptable to various scales of the dataset. This algorithm was only learnt in the later stage of the research project and hence was only applied in the last results chapter. As discussed in the previous project, Artificial neural network has received high popularity in machine learning power battery material research for their ability to capture non-linear correlations. However, its great prediction power was not validated from the results of all three chapters. In addition, tree-based models like random forest and gradient boosting machine have consistently demonstrated great prediction power over the structure-property-performance correlations of three different classes of cathode materials.

8.3 Model Performance Validations

From the optimised ML models, various chemical insights can be gained through interpreting the model structure and the output from the best performing model. These variable correlations could be used to guide the of design new battery materials with better discharging performance. However, attention should also be made to investigate the truthfulness of these obtained correlations and this can be achieved through conducting experiments.

Min et al [17] first generated 50,000 sets of experimental input parameters within the sensible range and used the best performing machine learning model to reversely predict their discharging performance. A few sets of experimental parameters that were predicted to lead to outstanding discharging performance were validated experimentally and the final experimental results have shown a high level of agreement. In the end, 15 new high-performing layered cathode materials were discovered.

Following this example above, various experiments could be designed using the correlations identified in the previous results chapter and the final experimental results could be used to validate them. The first step of this validation process

involves the generation of different sets of covariates values using the correlations that have been identified. For instance, the new doped spinel cathode formula could have a higher formula ratio of dopant with heavier molar mass such as molybdenum or technetium as this was indicated to lead to higher initial and 20th cycle end discharge capacity. After selecting a range of dopants and their respective ratios for the new testing formulas, one could use DFT simulation modelling tools to access their structural stability through estimating the relative free energies. In addition, the results of these simulations will also give the optimal structural properties which could then be used to be fed back into the ML model for discharge capacity prediction. After estimating the structural free energy and the respective discharge capacities, further experiments could be carried out for candidates that have the lowest free energy (most stable) and are predicted to possess high initial and higher cycle discharge capacities. This method would greatly fasten the process of discovering new materials with better discharging capacities.

To investigate the real prediction power of the optimal machine learning models from each project, a list of real experimental data have been collected and used to compare with the predictions made by the optimal models in each project. **Table 8.2**, **Table 8.3** and **Table 8.4** gives the prediction errors computed using the optimal machine learning for the predictions of initial and higher cycle discharge capacities of doped spinel cathode, doped NCM cathode and doped olivine cathode, respectively.

In **Table 8.2**, it is seen that the gradient boosting models have performed reasonably well for most of the discharge capacities of the cathode materials with the averaging deviation percentage estimated to be 8.5%. On the other hand, it is noticed that a high error percentage (17.5%) is reached for EC prediction for $LiBa_{0.01}Mn_{1.99}\ O_4$ materials. The respective experimental EC value is very similar to $LiBa_{0.02}Mn_{1.98}\ O_4$, however, the latter case has demonstrated a much lower prediction error percentage error (8.5%). Both of these materials formulas have molar mass and this implies that the molar mass difference is not the key contributor to these prediction deviations. The major difference between these two materials is

Table 8.2: The predicted initial (IC) and the 20th cycle discharge capacities (EC) of the doped spinel cathode materials, calculated by the gradient boosting models and the respective experimental values.

Material formulas	$LiTi_{0.03}Mn_{1.97}O_4$		$LiBa_{0.01}Mn_{1.99}O_4$		$LiBa_{0.02}Mn_{1.98}O_4$		$LiBa_{0.05}Mn_{1.95}O_4$	
Lattice constant (Å)	8.233		8.242		8.236		8.238	
Capacities	IC	EC	IC	EC	IC	EC	IC	EC
Experimental Capacities ($mAhg^{-1}$)	128.34	123.66	115.482	111.12	115.345	112.902	89.867	89.474
Predicted Capacities ($mAhg^{-1}$)	119.2	112.5	105.1	91.7	111.4	103.3	98.7	96.5
Deviations (%)	7.1	9.0	9.0	17.5	3.4	8.5	9.8	7.9
Reference	*Zhang et al* [327]				*Sahan et al* [328]			

seen as $LiBa_{0.01}Mn_{1.99}O_4$'s lattice constant is longer than $LiBa_{0.02}Mn_{1.98}$ O_4 and this could be the cause for the prediction difference as materials' lattice constant was previously ranked as a governing covariate by the optimal gradient boosting models.

In **Table 8.3**, it is seen that the deviation percentage between the predicted capacities and the experimental capacities for the IC and EC are very different across all three materials. $Li_{1.21}Ni_{0.125}Co_{0.125}$ $Mn_{0.525}In_{0.015}O_2$ is shown to have the highest average error percentage, whereas $Li_{1.21}Ni_{0.125}Co_{0.125}$ $Mn_{0.53}In_{0.02}O_2$ has the lowest average error. A trend is observed that as the capacity values increase, the prediction power of the gradient boosting machine model reduces. For instance, $Li_{1.21}Ni_{0.125}Co_{0.125}Mn_{0.525}In_{0.015}O_2$ has the highest IC and EC among the three materials and its deviation percentages are much higher than the case of $Li_{1.21}Ni_{0.125}Co_{0.125}Mn_{0.53}In_{0.02}O_2$ where both experimental IC and EC are shown to be the lowest. This suggests that the optimal GBM model can predict well for materials with lower capacities (< 250 $mAhg^{-1}$), however, cannot predict well for materials with higher discharge capacities. A possible cause of this lack of prediction power could be the lack of data for training the machine learning models,

Table 8.3: The predicted initial (IC) and the 50th cycle discharge capacities (EC) of the doped NCM layered materials, calculated by the gradient boosting models and the respective experimental values.

Material formulas	$Li_{1.21}Ni_{0.125}Co_{0.125}Mn_{0.53}In_{0.01}O_2$		$Li_{1.21}Ni_{0.125}Co_{0.125}Mn_{0.525}In_{0.015}O_2$		$Li_{1.21}Ni_{0.125}Co_{0.125}Mn_{0.53}In_{0.02}O_2$	
Capacities	IC	EC	IC	EC	IC	EC
Experimental Capacities ($mAhg^{-1}$)	249.47	240.96	296.72	293.99	208.847	170.9131
Predicted Capacities ($mAhg^{-1}$)	223.546	183.145	228.262	180.187	219.729	179.807
Deviations (%)	10.4	24.0	23.1	38.7	5.2	5.2
Reference	*Etefagh et al* [329]					

Table 8.4: The predicted initial (IC) and the 20th cycle discharge capacities (EC) of the carbon-coated and doped olivine materials, calculated by the gradient boosting model and bayesian ridge regression model respectively and the experimental results.

Material Formula	**$LiNi_{0.2}Co_{0.8}PO_4$**		**$LiCo_{0.95}Mn_{0.05}PO_4$**	
Carbon Content (wt.%)	4.08		2.97	
Capacity	IC	EC	IC	EC
Experimental (mAh/g)	141	129	141.23	135
Predicted (mAh/g)	149.03	137.32	150.76	142.41
Deviations (%)	5.6	6.4	6.7	5.5
Reference	*Ornek et al* [330]		*Ornek et al* [331]	

which leads to the model not being able to capture these correlations precisely.

Table 8.4 gives the prediction of the IC and EC values for two doped and carbon-coated olivine systems, namely, $LiNi_{0.2}Co_{0.8}PO_4$ and $LiCo_{0.95}Mn_{0.05}PO_4$. Small prediction deviations are observed among the predictions for IC and EC, with the average deviation calculated to be 6.15% and 5.95% respectively for the two systems. This indicates that our models have sufficiently captured the underlying correlations between the material properties and their discharge performance and can additionally predict the capacities of other metal-based olivine systems. In this case, it was the cobalt and nickel doped olivine systems.

Overall, promising predictions of the discharge capacities have been achieved by some ML models for a few experimental reported material cases. However,

many high prediction deviation percentages still can be observed as 38.7% and 24%. These high errors suggest that the ML models are still yet to provide highly accurate predictions over the experimental results and still have a big room for improvement. One direct method to improve their prediction power is to increase the size of the dataset for training. Further discussion over the future work can be found in the later section.

8.4 Limitations of the Study

The potentials of using machine learning techniques for predicting the discharge capacities of cathode materials and for uncovering the complex correlations of battery materials have been revealed in this thesis. However, there are still some challenges that remain to be solved. This section reveals two aspects of the limitations and challenges that are faced in each study, namely, the material data and the ML model optimisation for achieving higher prediction power.

The lack of agreed standardisation of experimental conditions, both on material synthesis and the electrochemical test has been a major challenge when collecting data. In **Chapter 5** and **Chapter 6**, it has been observed that battery cell fabrication methods such as the mixing ratio for anode and cathode components are varied across different research teams (e.g. binder, conductive additives, and active material). The ratio of binder and additive can affect the mechanical stability and electrical conductivity of the cathode material and therefore affects the durability of the battery and their long-term discharge capacities [80]. However, such information is often not reported, hindering the selection of this covariate as a part of the ML model from the first two projects. The effects of these mixing ratios are only investigated in the last olivine system because the respective dataset is much smaller and hence, it is easier to collate full mixing information for all cathode systems. Nevertheless, other key experimental information have also been missing such as the volume of the electrolyte used and the measurement of the surface area of cathode and anode. 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 [27].

Next, the microstructure of the cathode material such as particle sizes and their distribution, as well as their surface morphologies are known to affect the electrical conductivity and the rate of electrochemical reactions of the cathode [276, 332]. The collection of such data is found to be challenging as either the measurements are not performed in some publications or the results are not published in a numerical format, and thus would require further interpretation. The publications that have reported SEM images of the surface have used various scales (e.g. D10, D50, D90) which adds up the complexity of data collecting. A similar issue has also been reported by *Kauwe et al* [93] during data collection work for their studies of Li-ion battery materials. In addition, such information is particularly important when it comes to analysing the properties of the carbon coating on the surface of doped olivine material and their effects on the respective electrochemical performance. For instance, several studies have demonstrated the effects of the properties of carbon coating such as carbon surface distribution, surface area, morphology and porosity, on the rate of electrons transfer from the olivine cathode [133, 333, 334, 335, 336]. These properties were not widely available from the publications as previously discussed and therefore remain a limitation for the studies.

Despite various findings that have been made previously such as identifying the best performing ML model for each cathode system, there are still many uncertainties relating to the covariates that remain. These could both affect the validity of the obtained variable correlations. To begin with, the experimental data collection involves a lot of uncertainties such as misreading or misreporting the results from the equipment. These errors will lead to the collection of invalid data which would only add noise to the ML model and not guide the useful research. The lack of reporting standards is also another concern when it comes to covariate uncertainty. A common phenomenon is observed in that different numbers of significant figures are used to report the same covariate variable across different publications. One example of this can be seen from that certain publications report lattice constants in three significant figures and some in five significant figures and this could add

another layer of uncertainty to the model.

In addition, the variance of the covariates used for model training could also play a key part in the ML model prediction accuracy. Materials molar mass was ranked as being the top governing variable in **Chapter 5**. However, it is noticed that this variable also has the highest variance than other materials properties such as the electronegativity and this could mislead the gradient boosting model in using this variable more during the tree splitting process and hence would potentially reduce the model prediction force.This issue could be potentially resolved by collating more data to improve the variance of each covariate. Furthermore, multicollinearity of the covariates is also another issue that could affect the model prediction power. This occurs when the degree of linearity correlation of two independent variables is very high and would make the model variable coefficient more sensitive to small changes and additionally reduce the overall model prediction power. This issue could be mostly found in the linear regression algorithms where estimations of the variable coefficients are required. In **Chapter 5**, a high linear correlation was identified between the ratio of manganese and the dopant in the formula and this could be one of the reasons for the observed poor prediction power of the linear models. In summary, there are many uncertainties relating to the covariates that would lead to the reduction of model prediction power. The most effective way to access the real prediction power of the ML model is to utilise them to predict against the real experimental results and compare the differences in the results. In this thesis, the prediction power of the best performing ML model under each project has been investigated by validating with the real experimental results and this is indicated in the previous validation section.

The model training process and the selection of hyperparameters for the machine learning models can largely affect their prediction power. The artificial neural network models investigated in **Chapters 5** and **Chapter 6** have demonstrated poor prediction power over targeted discharge capacity and as discussed, this was potentially caused by the overfitting issue, both from the small dataset and the over-parameterisation. In addition, the general lack of available high-quality datasets is

also seen as the biggest challenge across all three research projects.

8.5 Material Correlation Insights

Through interpreting the covariate contribution in the best performing machine learning models, insights can be gained on the governing variable and their correlations with the response variables. This section entails the discussion over the chemical intuitions behind the observed correlations under each results chapter.

In **Chapter 5**, the molar mass of the doped spinel cathode material was determined to be the governing variable for the discharge capacities of the first and the twentieth cycle. The importance of the cathode material molar mass has been widely observed from Faraday's law where it stated that the material molar mass and the number of reactive Li-ions are inversely proportional to the specific capacities of an electrode and therefore would have a significant impact on it. A positive correlation is observed between the material molar mass the discharge performance which implies that higher discharge capacities could be achieved with lower material formula mass. This, on the other hand, contradicts the relationship mentioned previously where the lower material molar would be considered to be better in terms of leading to higher discharging performance. After interpreting the variable importance results from both the NCM and the olivine ML models, a common finding has made that the initial and higher cycle discharge capacities of both material systems are shown inversely proportionally to the material molar mass and this agrees with the previously stated Faraday's law.

Crystal lattice dimensions of the material control the Li-ion site energy which in turn affect the lithium-ion diffusion rate and hence could greatly affect the discharging performance of the materials. During the investigation of the correlations between the crystal dimension lattice of the cubic doped spinel cathode, the inverse correlations have been identified where the smaller the crystal dimension lattice, the higher the discharging performance at both the first and the twentieth cycle. The underlying reason could be that a shorter crystal lattice dimension would greatly reduce the diffusion pass for the Li-ions and hence enhance the transfer rate dur-

ing the charging/discharging which leads to higher discharge capacities at various cycles. Similar observations can also be made from other material variable studies such as the shorter the lattice dimensions "a" and "c" of doped NCM cathode would lead to higher the first and the 50^{th} cycle capacities and the shorter lattice dimensions of "a", "b" and "c" of carbon-coated and doped olivine system, the higher the initial and the 20^{th} cycle discharge capacities would be. The chemical intuition behind this could be because the smaller lattice parameter means a shorter diffusion path for the Li-ions to travel and this increases the rate of electrochemical reactions which in turn delivers more energy during the discharging performance.

Electronegativity measures the dopant element's ability to attract electron pairs toward itself. Dopant's EN controls the bonding strength with the surrounding TMs and oxygen atoms and influences structural stability as well as the overall crystal structure density. During a long cyclic charging and discharging process, the overall crystal structure often becomes unstable which then triggers significant lattice collapses and leads to severe capacity fading [307]. The involvement of dopant content can greatly improve the structure stability by forming stronger bonds whilst the strategy with doping with a small amount can ensure that no second material phase is formed and also the whole crystal structure is not modified significantly to disturb the Li-ion intercalation/deintercalation mechanisms. In the first results chapter, it is found that higher discharge capacities of the first and the twentieth cycle can be achieved using dopant of a moderate to lower range electronegativity value ($1.4 \sim 1.8$). The average electronegativity of this value is close to 1.6 which is 0.1 higher than the manganese elements. This suggests that the dopant will establish bonds with more covalent characters with surrounding oxygen atoms than manganese atoms do. Dopants such as vanadium or chromium could be useful for improving the discharging performance of spinel cathodes according to this correlation. In **Chapter 6**, negative correlations between the dopant's electronegativity value and the discharge performance are obtained, indicating that more ionic bonds within the material structure are better in terms of promoting greater discharge capacities. Similarly, the last results chapter has indicated that having a greater av-

erage electronegativity of the based metal and the dopant atom can lead to higher discharge capacities at both the initial and the twentieth cycle. This is because an increase in the weighted metals' elements electronegativity in the olivine system can lead to the formation of more covalent bonds with the surrounding P and O atoms. The covalent bonding characters will make the M-P bonds become longer to push further the PO_4 tetrahedra groups. This will eventually reduce the [010] diffusion path for Li-ions to travel through, promoting the efficient transfer of the Li-ions and hence leading to a higher discharge capacity [326].

The material composition determines the chemical environment within the cathode material which influence the structural physical and electrochemical properties directly and hence the respective discharging performance. The ratio of dopant in the material formula is one of the commonly discussed issues in material research as an increased amount will significantly affect the material intrinsic structure and lead to unwanted change to the discharge performance. In the case of doping electrochemical inactive dopants such as titanium would greatly reduce the rate of the electrochemical reactions during charge/discharge performance. During the investigation of the dopant ratio on the electrochemical performance of the NCM materials, it is found that a lower amount of dopant is preferred as it would lead to an increase in the discharge capacities of the initial and the 50^{th} cycle. This could be understood as such that a larger amount of dopant would greatly alter the intrinsic properties of the NCM cathodes and hence would restrict the rate of Li-ion transfer during the charge/discharge process. Similarly, for the carbon-coated olivine cathode materials, a lower dopant ratio can lead to an increase in the initial and the 20^{th} cycle discharge capacities. This is thought to be reasonable since a high content of dopant would increase the likelihood of blocking the [010] diffusion path and constrain the kinetics of Li-ion diffusion within the olivine system. Overall, over-doping is undesirable as it would greatly reduce the discharge capacities of the cathode materials at both the initial and the higher cycle in general. The underlying reason could be that over-doping would largely replace the electrochemically active atoms within the structure of the material and hence lower the rate of

electrochemical reactions during the charging/discharging process.

In **Chapter 7**, Carbon coating content was introduced as new covariate as the material system requires the application of it to boost the ionic conductivities. On the other hand, the selection of the optimal amount of carbon coating content for reaching the best discharge performance is very difficult. Thick carbon coating would also greatly inhibit the Li-ion from penetrating the cathode and lead to the lower capacity performance of the whole cathode material. In addition, a smaller amount of carbon coating would not enhance the Li-ion conductivity enough to boost up the overall discharge performance. The results of the variable importance study revealed that a higher carbon coating content would lead to higher discharge capacities at the first and the twentieth cycle. This could be because higher carbon coating content would lead to a great boost in the ionic conductivity of the materials. However, this still requires further investigations on the upper limit carbon coating amount for reaching the best discharging capacity before it gets deteriorated.

8.6 Future Work

This thesis lays a solid foundation in using machine learning techniques to investigate the doping effects on the discharge capacities of three classes of cathode material for lithium-ion batteries. This section summarises some valuable prospects that could be further investigated as additional future work of this thesis:

- In response to the potential overfitting issue caused by small datasets and a large number of covariates, dimension reduction techniques can be applied before training the ML model. For instance, the principal component analysis is a useful method that selects the principal component based on covariates' relevance to the final response variable [337]. The use of such a technique will greatly eliminate irrelevant covariates, reduce computational cost whilst maintaining useful information and most importantly, reducing the overfitting effects. Another similar dimensionality reduction approach is linear discriminate analysis (LCA), which has also demonstrated great help in reducing dimensionalities of variable space whilst improving the prediction power of

ML models [338]. These techniques can be applied to all of the datasets used in the aforementioned results chapters, particularly for **chapter 7**, where there are only 34 sets of data with the presence of 17 covariates.

- The prediction power of the artificial neural network has shown to be very poor in **Chapter 5** and **Chapter 6** and this is partially due to the models being overly parameterised. As a method to improve it, the L2 regularisation method introduced in **subsection 4.7.2** can be used to regulate the weights of hidden neurons to improve the model prediction power. This technique works by shrinking the weights of neurons that do not contribute much towards the prediction to avoid the addition of noise to the neural network model [339, 340]. This would not only greatly improve the learning speed of the neural network model, but also helps to reduce the likelihood of the model being overfitted to the dataset. However, implementation of such a technique requires optimisation of the lambda value (controls the strength of regularisation), which involves a higher computational cost. As a potential approach to reduce this long computing time, an early-stopping technique can be used in parallel with the regularisation method. The early-stopping technique involves reducing the length of the training process and has shown improvements in model prediction power [341, 342, 343].

- The Bayesian ridge regression algorithm has demonstrated great prediction power for the initial and the twentieth cycle of the discharge capacity for the carbon-coated and doped olivine cathode. This technique could also be investigated for its prediction power over the dataset of other different cathode systems.

- New covariates could be introduced towards the future model construction to boost up prediction power. The valence number of manganese ions in doped spinel cathode material can be calculated and used as an additional descriptor, as it is an indication of the manganese dissolution reaction in the spinel system. In addition, it has been reported during the synthesis of NCM materi-

als, some synthesis parameters such as temperature used for sintering or calcinating could influence both the materials morphologies, surface chemistry between electrode and electrolyte (oxygen loss) and hence, the respective discharge capacities [344, 345, 346, 347]. However, this correlation has not yet been investigated quantitatively in the literature, which leaves room for future exploitation. Similarly, despite the effects of material properties such as particle size and size distribution being discussed to be influential for the materials' electrochemical performance. there are still difficulties remaining such as identification of the most suitable particle size to reach the best cyclic performance. One could potentially implement the power of a convolutional neural network to investigate the linkage between the Scanning Electron Microscopy (SEM) images of the cathode material surface and their respective discharge performance. After identifying the best particle size and particle distribution for reaching higher performance, one could use 3D printing techniques to print out the cathode material with the desirable microstructure that leads to high discharge capacities.

- The lack of agreed experimental reporting standards is one of the major causes for the relatively small datasets used in this thesis. This problem could be potentially solved by integrating this machine learning project with some experimental work on the side. In such a way, experimentalists could consistently control the experimental conditions such as to standardise the chemical usage and cyclic performance conditions across all experiments. In this regard, a large amount of high-quality data can be produced and made available for later ML training. Further utilisation of such detailed experimental data could be to find the linkage between synthesis parameters and properties of the final cathode materials. These models could be applied consecutively to the models built from this thesis to identify the desirable synthesis parameters that would lead to great discharge capacities, and in general help to streamline the whole process of material discovery.

- The application of reverse engineering has shown great promise in identifying

new cathode materials with desired properties. This process involves generating a new dataset of covariates within the chemically intuitive value ranges and fed back into the ML models to predict their discharge capacities. After identifying the new cathode structure with desirable dopant and the relative amount, one can perform density-functional-theory calculations to assess the stability of the structure and select the most stable one to perform the additional experimental synthesis. This approach will help to save a lot of time and money from conducting trial-and-error experiments that may not lead to the discovery of new materials of the desired property.

Appendix A

List of Supplementary Figures

This section includes the optimisation graphs generated during the computation of the ridge regression and LASSO regression models for the prediction of the initial and the 20^{th} cycle discharge capacities of the doped spinel cathode materials($LiM_xMn_{2-x}O_4$). The software used for generating these graphs are from a R package named ”glmnet”.

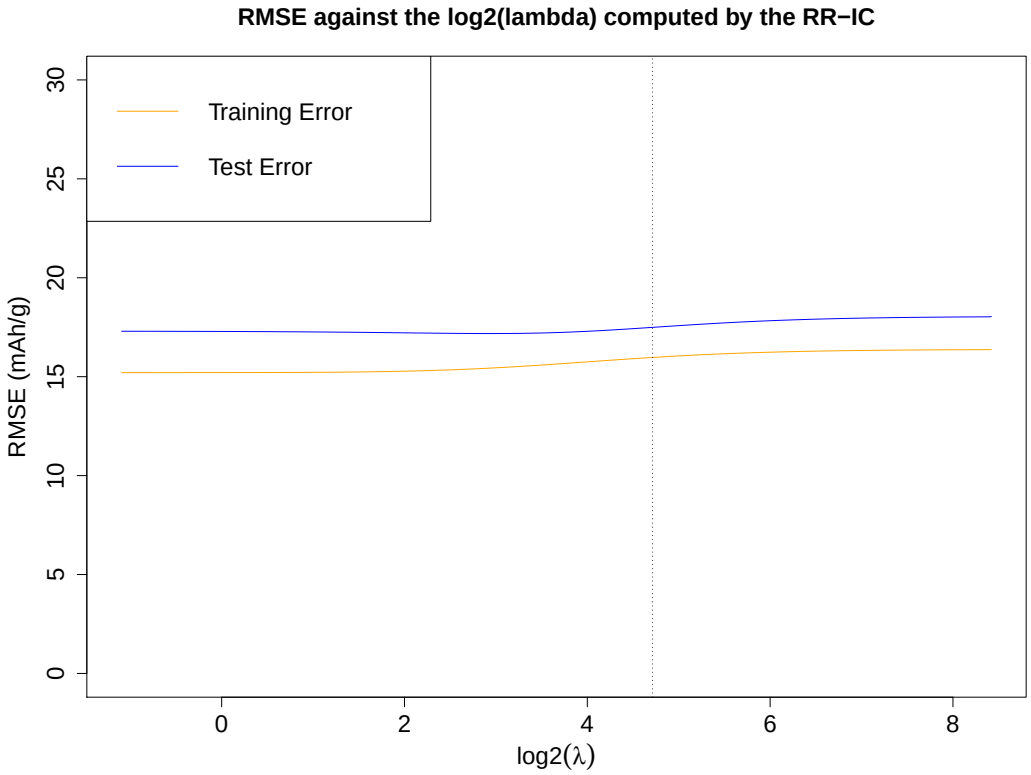


Figure A.1: The training and testing errors versus the variations of lambda values computed by the Ridge model for the prediction of initial discharge capacity (IC) of doped $LiMn_2O_4$ cathodes. The dotted line indicates the optimal lambda value used in the model.

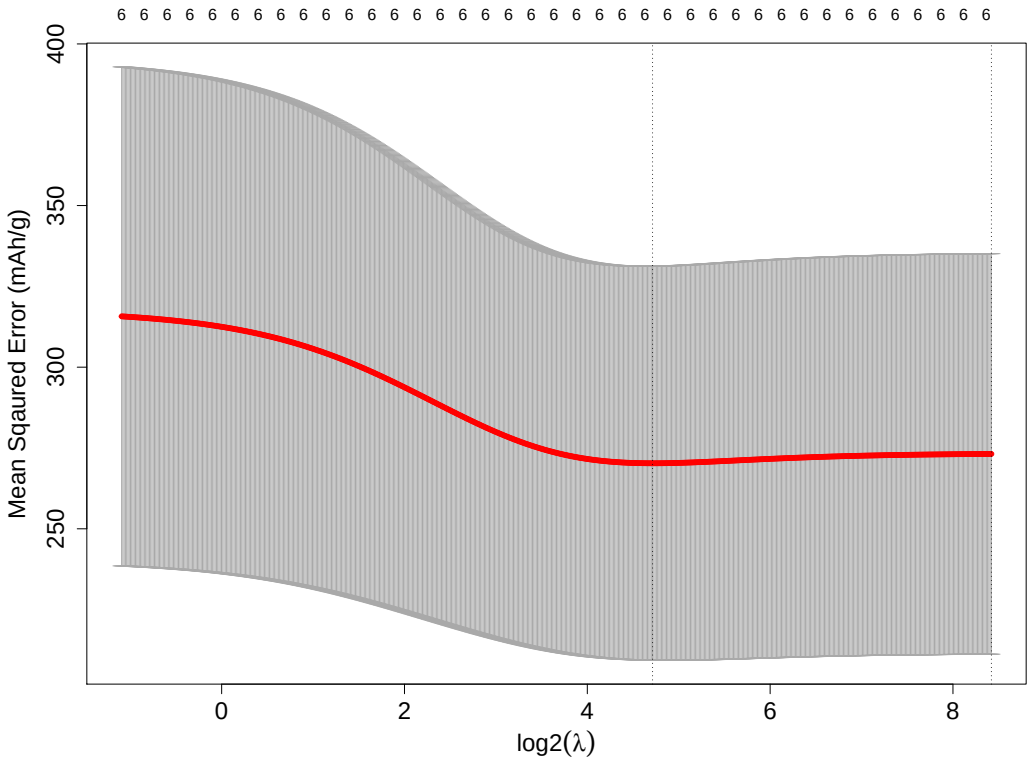


Figure A.2: The computed mean squared error with the change in the lambda values (in logarithmic scale with the base of 2) computed by the ridge regression model for the prediction of initial discharge capacity (IC) of doped $LiMn_2O_4$ cathodes. The red indicates the averaged validation error across the ten-folds and the dotted line indicate the optimal lambda values.

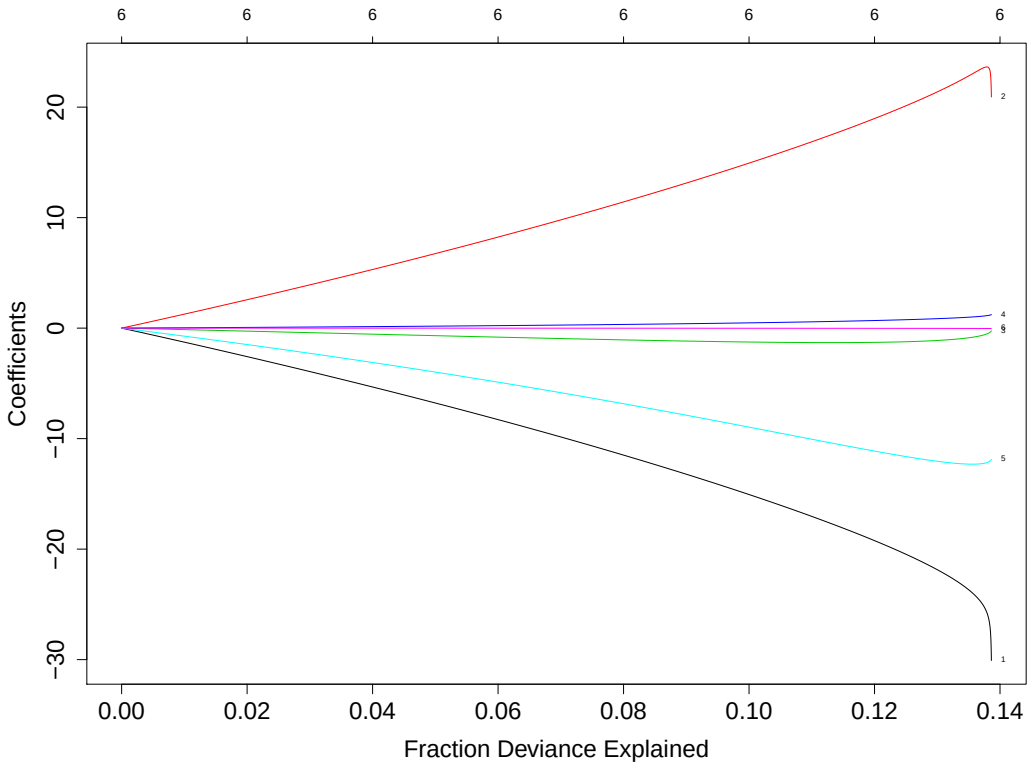


Figure A.3: The changes in the coefficients of covariates variable (labelled by numbers) against the fraction of the deviance explained by the ridge regression model for the prediction of initial discharge capacity (IC) of doped $LiMn_2O_4$ cathodes. The model complexity increases from left to right.

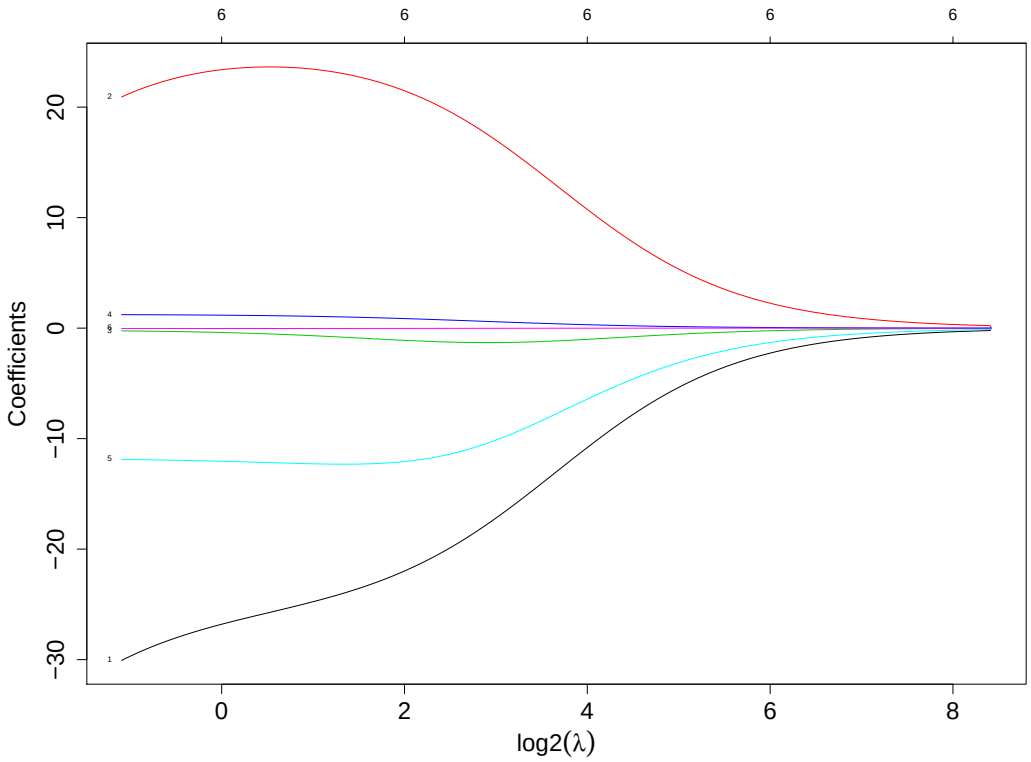


Figure A.4: The changes in the coefficients of covariates variable (labelled by numbers) against the change in the lambda values (in logarithmic scale with the base of 2) computed by the ridge regression models for the prediction of initial discharge capacity (IC) of doped $LiMn_2O_4$ cathodes.

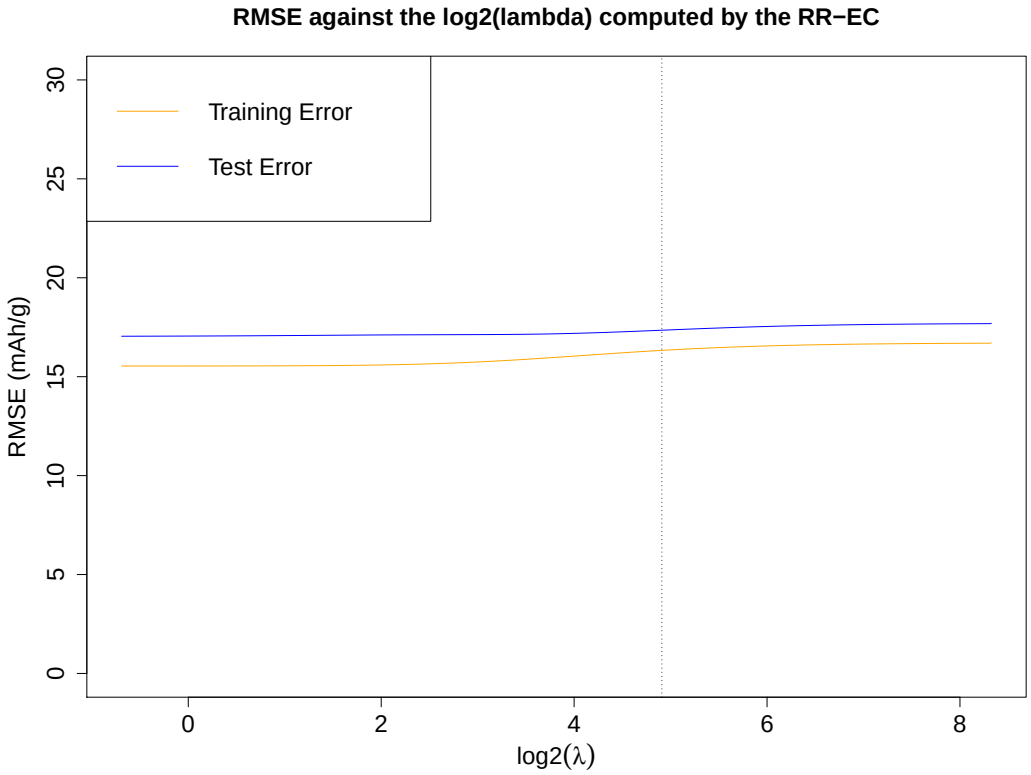


Figure A.5: The training and testing errors versus the variations of lambda values computed by the Ridge model for the prediction of 20^{th} cycle end discharge capacity (EC) of doped $LiMn_2O_4$ cathodes. The dotted line indicates the optimal lambda value used in the model.

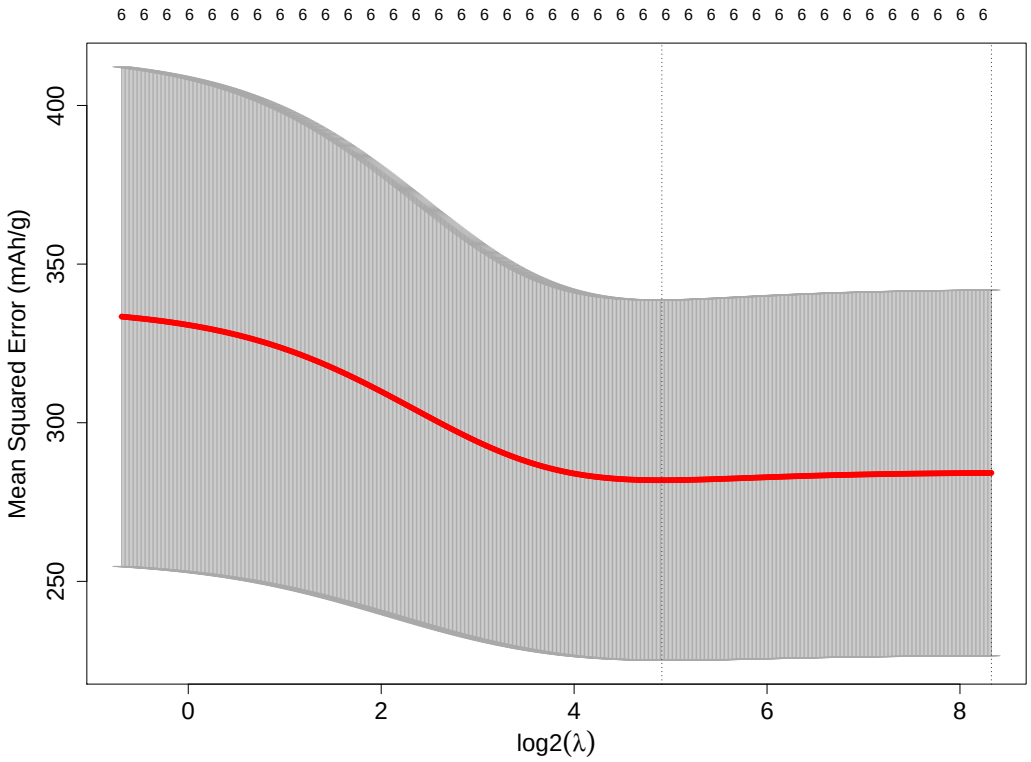


Figure A.6: The computed mean squared error with the change in the lambda values (in logarithmic scale with the base of 2) computed by the ridge regression model for the prediction of 20^{th} cycle end discharge capacity (EC) of doped $LiMn_2O_4$ cathodes. The red indicates the averaged validation error across the ten-folds and the dotted line indicate the optimal lambda values.

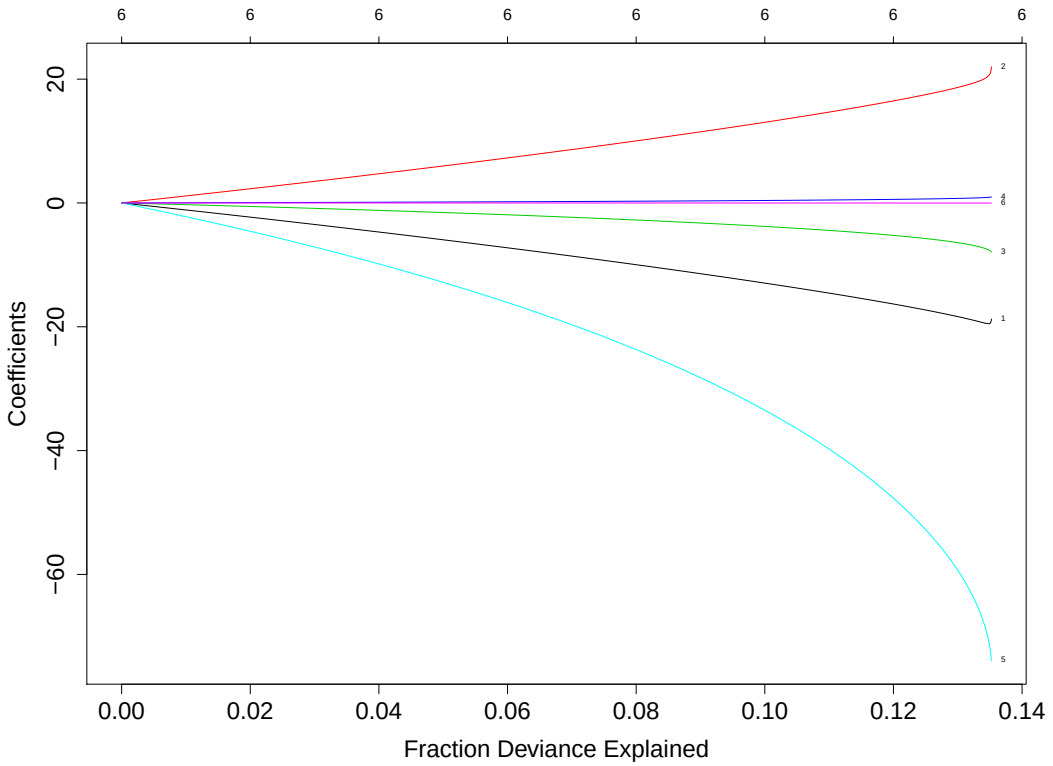


Figure A.7: The changes in the coefficients of covariates variable (labelled by numbers) against the fraction of the deviance explained by the ridge regression model for the prediction of 20^{th} cycle end discharge capacity (EC) of doped $LiMn_2O_4$ cathodes. The model complexity increases from left to right.

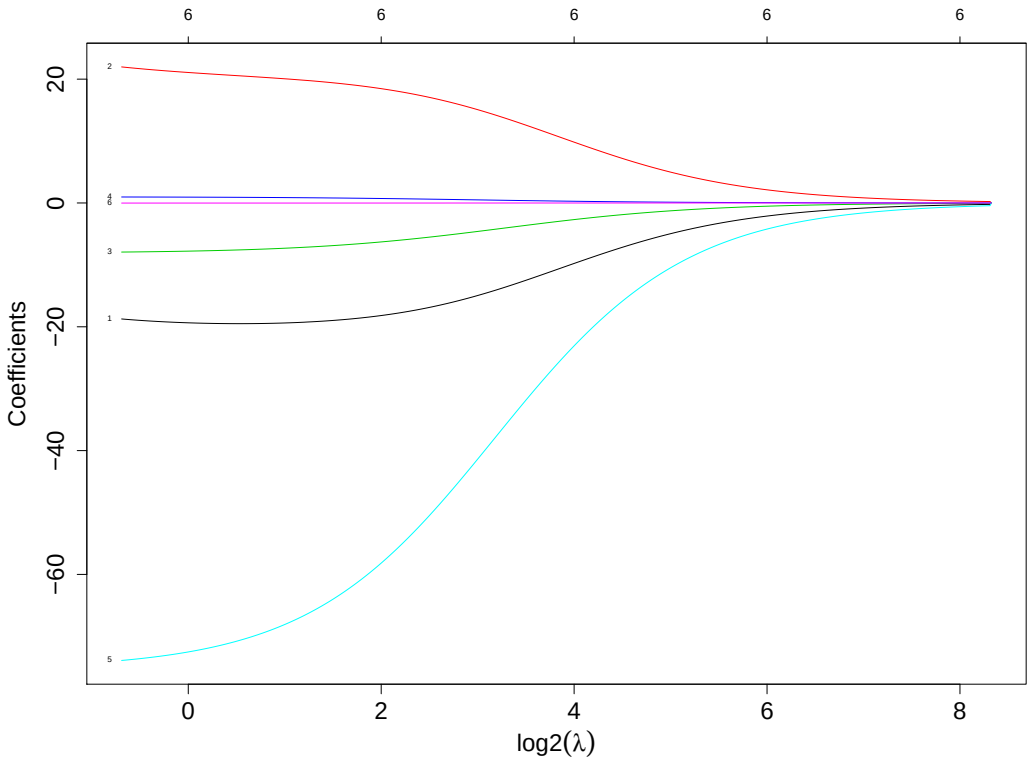


Figure A.8: The changes in the coefficients of covariates variable (labelled by numbers) against the change in the lambda values (in logarithmic scale with the base of 2) computed by the ridge regression models for the prediction of 20^{th} cycle end discharge capacity (EC) of doped $LiMn_2O_4$ cathodes.

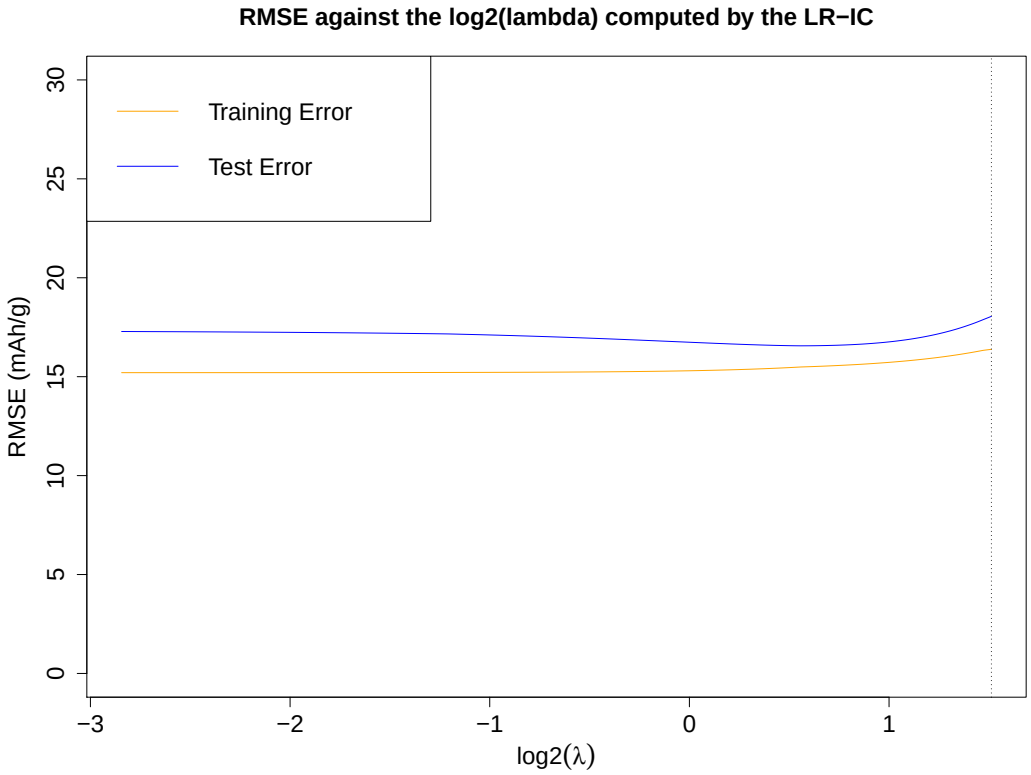


Figure A.9: The training and testing errors versus the variations of lambda values computed by the LASSO model for the prediction of initial discharge capacity (IC) of doped $LiMn_2O_4$ cathodes. The dotted line indicates the optimal lambda value used in the model.

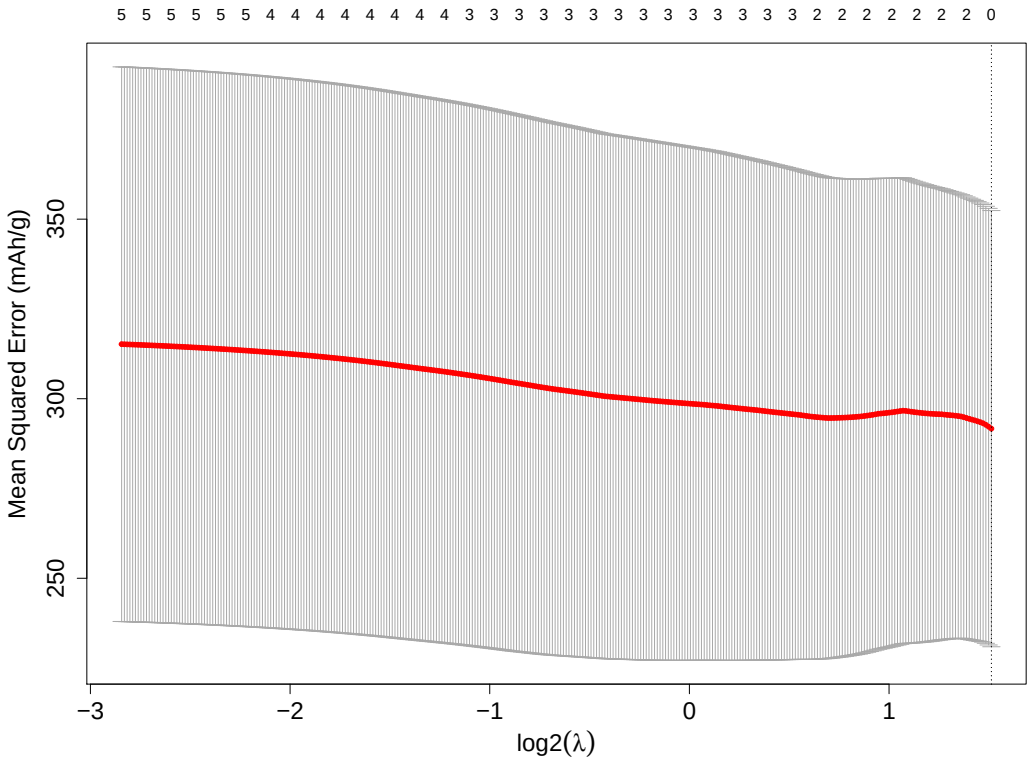


Figure A.10: The computed mean squared error with the change in the lambda values (in logarithmic scale with the base of 2) computed by the LASSO regression model for the prediction of initial discharge capacity (IC) of doped $LiMn_2O_4$ cathodes. The red indicates the averaged validation error across the ten-folds and the dotted line indicate the optimal lambda values.

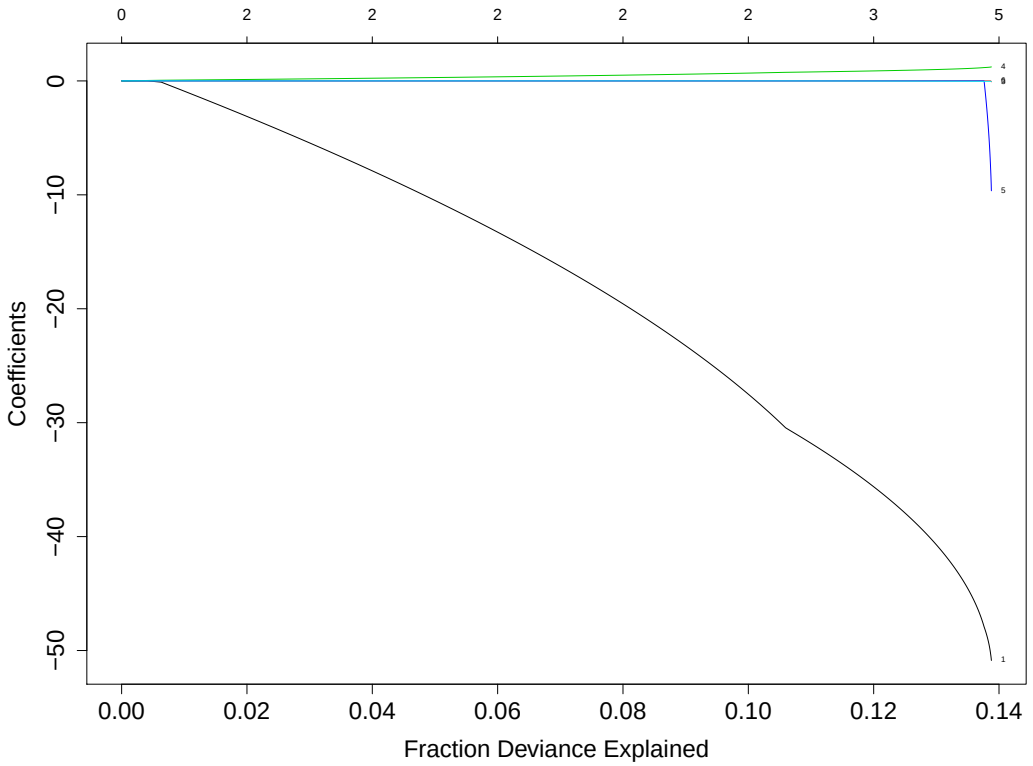


Figure A.11: The changes in the coefficients of covariates variable (labelled by numbers) against the fraction of the deviance explained by the LASSO regression model for for the prediction of initial discharge capacity (IC) of doped $LiMn_2O_4$ cathodes. The model complexity increases from left to right.

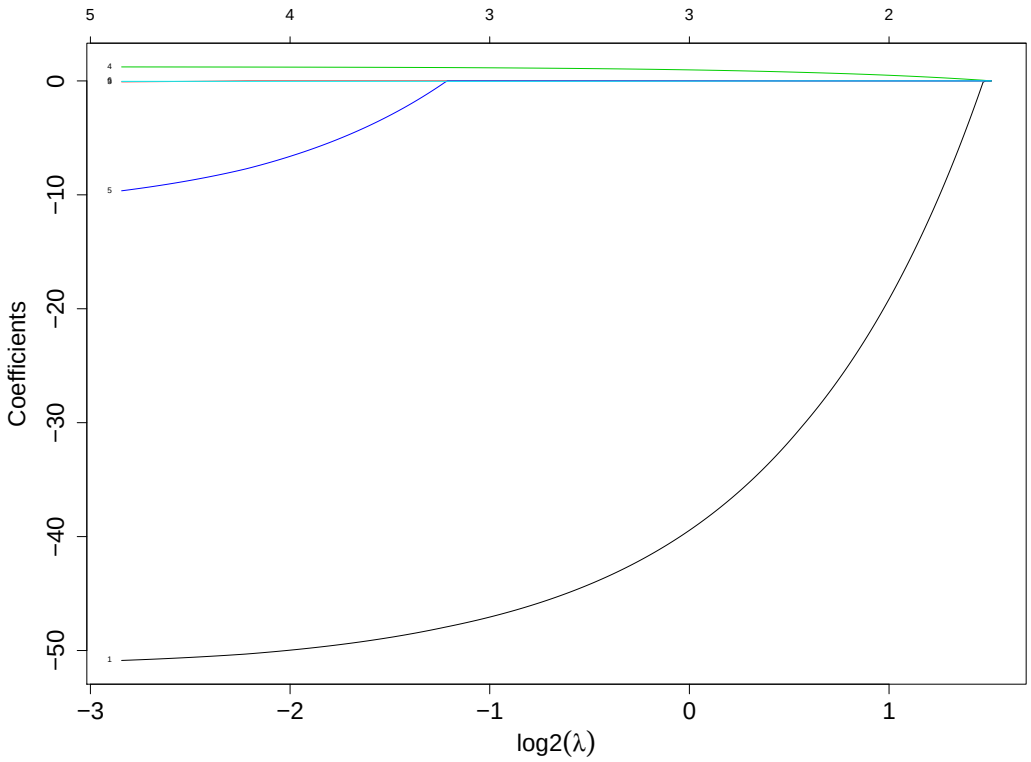


Figure A.12: The changes in the coefficients of covariates variable (labelled by numbers) against the change in the lambda values (in logarithmic scale with the base of 2) computed by the LASSO regression models for for the prediction of initial discharge capacity (IC) of doped $LiMn_2O_4$ cathodes.

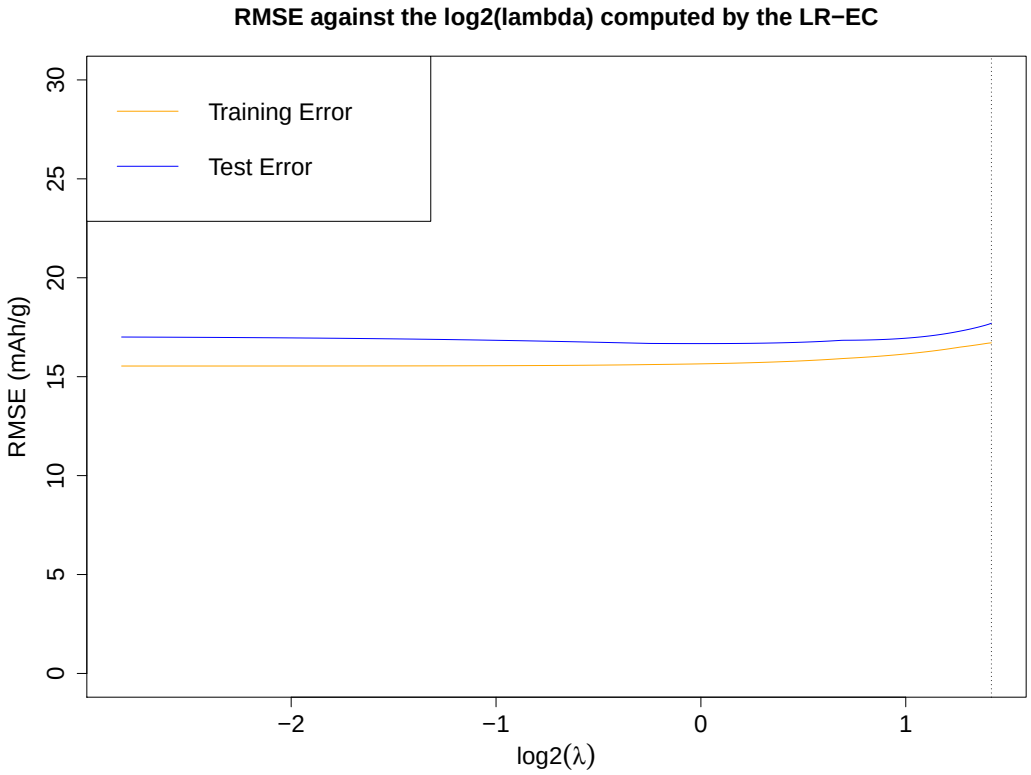


Figure A.13: The training and testing errors versus the variations of lambda values computed by the LASSO model for the prediction of 20^{th} cycle end discharge capacity (EC) doped $LiMn_2O_4$ cathodes. The dotted line indicates the optimal lambda value used in the model.

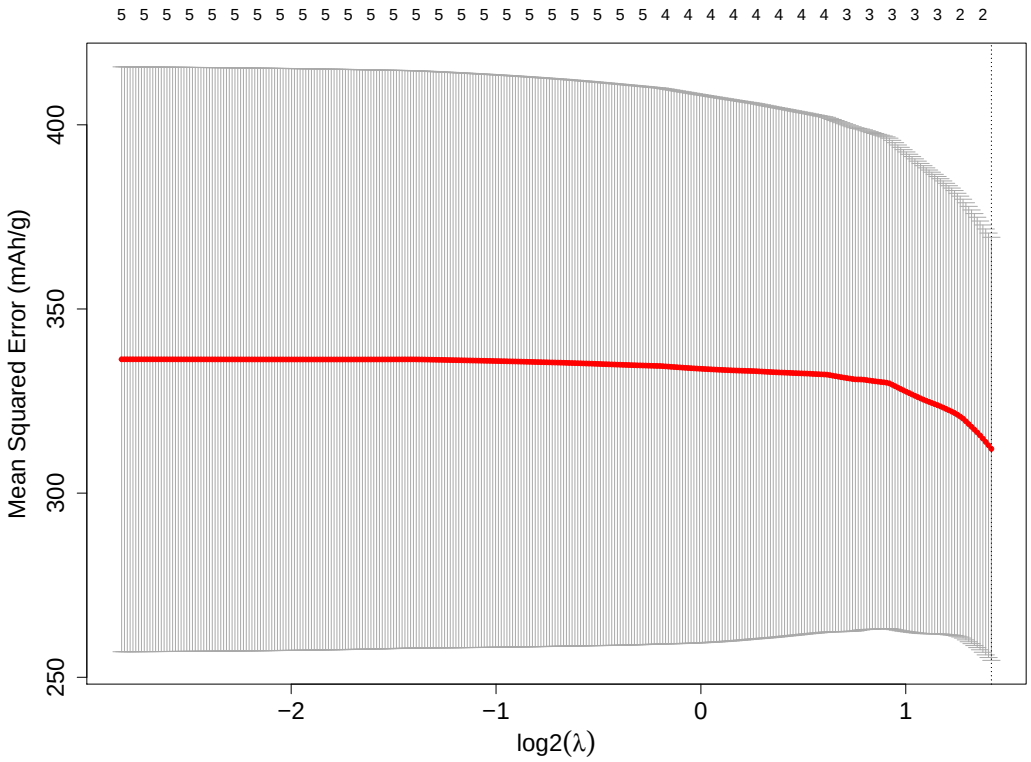


Figure A.14: The computed mean squared error with the change in the lambda values (in logarithmic scale with the base of 2) computed by the LASSO regression model for the prediction of 20^{th} cycle end discharge capacity (EC) doped $LiMn_2O_4$ cathodes. The red indicates the averaged validation error across the ten-folds and the dotted line indicate the optimal lambda values.

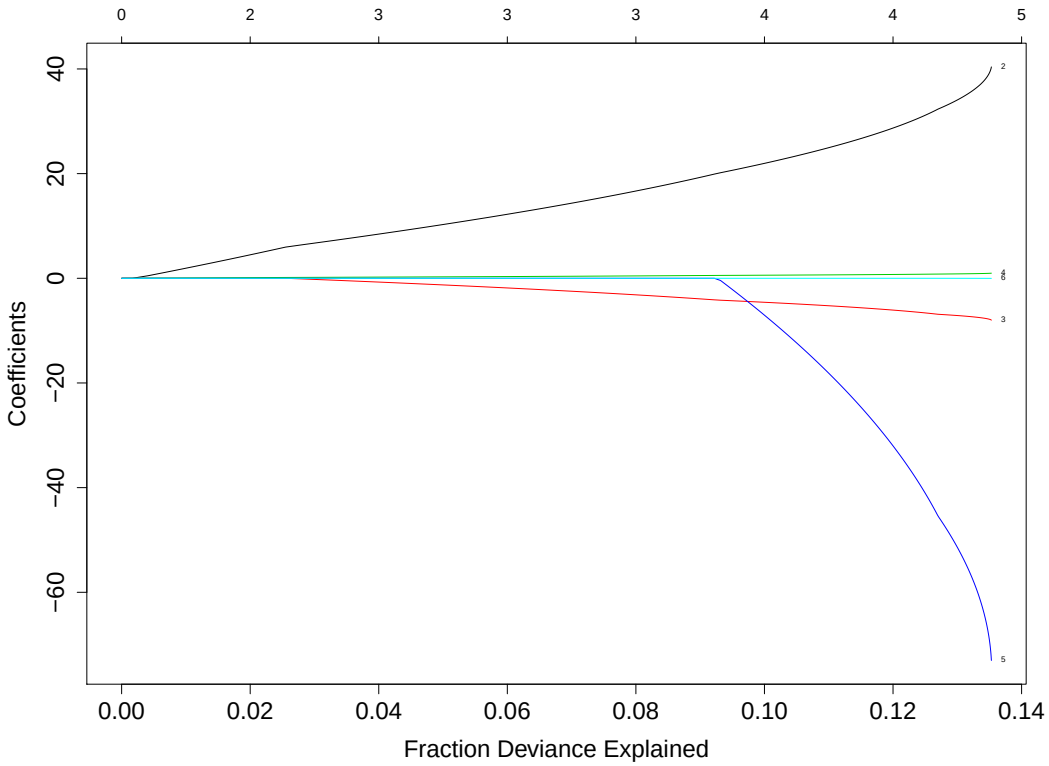


Figure A.15: The changes in the coefficients of covariates variable (labelled by numbers) against the fraction of the deviance explained by the LASSO regression model for the prediction of 20^{th} cycle end discharge capacity (EC) doped $LiMn_2O_4$ cathodes. The model complexity increases from left to right.

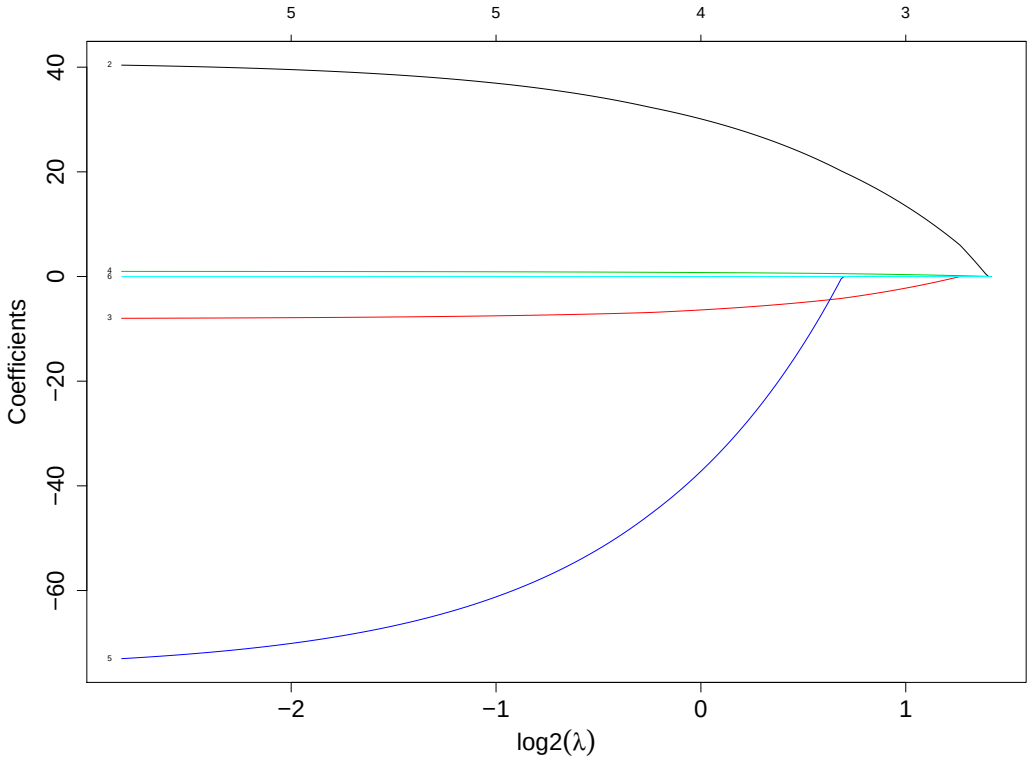


Figure A.16: The changes in the coefficients of covariates variable (labelled by numbers) against the change in the lambda values (in logarithmic scale with the base of 2) computed by the LASSO regression models for the prediction of 20th cycle end discharge capacity (EC) doped $LiMn_2O_4$ cathodes.

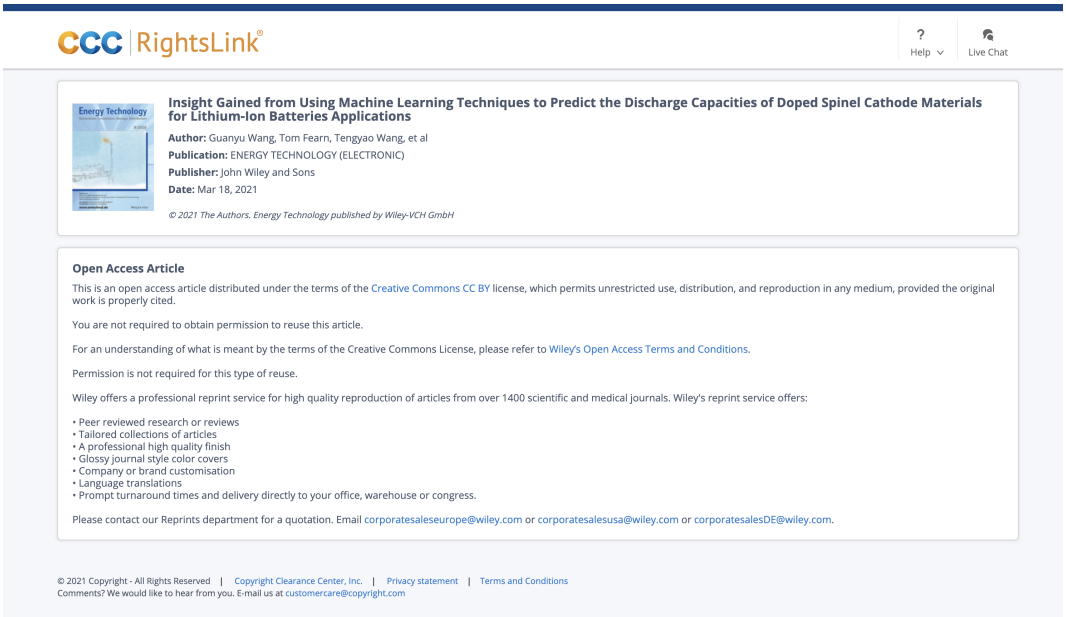


Figure A.17: Capture of the pen license agreement on the use of materials for constructing the **Chapter 3** in this thesis, with the Publisher, John Wiley and Sons.

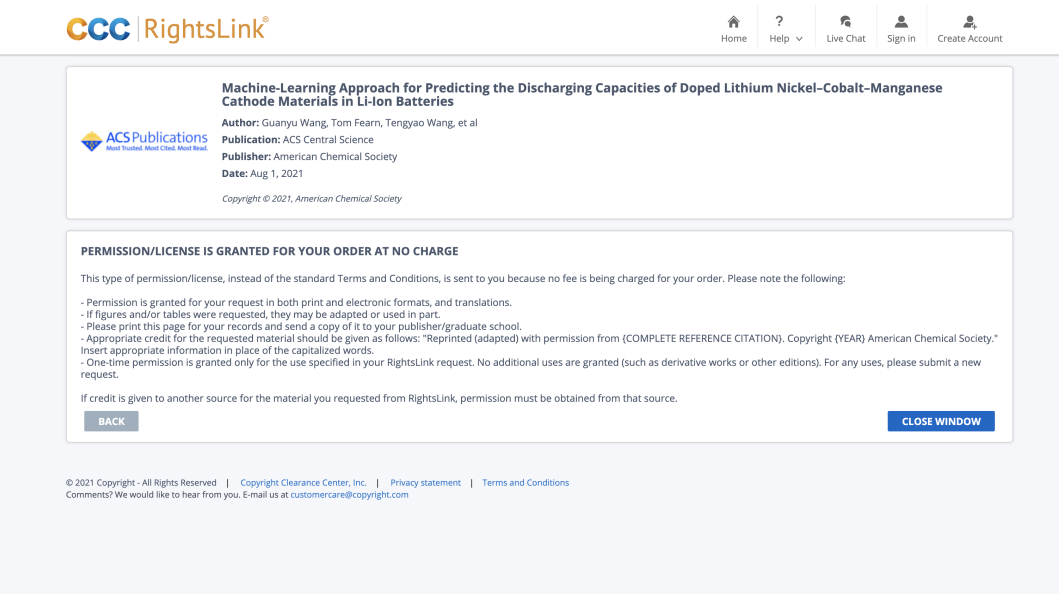


Figure A.18: Capture of the pen license agreement on the use of materials for constructing the **Chapter 4** in this thesis, with the Publisher, American Chemical Science.

Appendix B

List of Supplementary Tables

Table B.1: The dielectric constants of various compounds are used in the electrolyte mixture for lithium-ion batteries applications. Adapted with permission from [31].

Compound	**Dielectric Constant (25°)**
Ethyl Carbonate (EC)	95.3
Dimethyl Carbonates (DMC)	3.12
Diethyl Carbonates (DEC)	2.82
Propylene carbonate (PC)	65.5
Ethyl Methyl Cabornates (EMC)	2.4

Bibliography

[1] Energy Transitions: Global and National Perspectives (Second expanded and updated edition) – Vaclav Smil.

[2] Ray and Douglas. Lazard's Levelized Cost of Energy Analysis—Version 13.0. Technical report, 2020.

[3] Isidor Buchmann and Cadex Electronics Inc. *Batteries in a portable world : a handbook on rechargeable batteries for non-engineers*. Cadex Electronics, 2001.

[4] Shaik Amjad, S. Neelakrishnan, and R. Rudramoorthy. Review of design considerations and technological challenges for successful development and deployment of plug-in hybrid electric vehicles, 4 2010.

[5] Chaofeng Liu, Zachary G. Neale, and Guozhong Cao. Understanding electrochemical potentials of cathode materials in rechargeable batteries, 3 2016.

[6] John B. Goodenough and Youngsik Kim. Challenges for rechargeable Li batteries, 2 2010.

[7] Xianxia. Yuan, Hansan. Liu, and Jiujun. Zhang. *Lithium-ion batteries : advanced materials and technologies*. CRC Press, 2012.

[8] Christian Julien, Alain Mauger, Karim Zaghib, and Henri Groult. Comparative Issues of Cathode Materials for Li-Ion Batteries. *Inorganics*, 2(1):132–154, 3 2014.

[9] Brent C. Melot and J. M. Tarascon. Design and preparation of materials for advanced electrochemical storage. *Accounts of Chemical Research*, 46(5):1226–1238, 5 2013.

[10] Chaofeng Liu, Zachary G. Neale, and Guozhong Cao. Understanding electrochemical potentials of cathode materials in rechargeable batteries, 3 2016.

[11] Hyun Deog Yoo, Elena Markevich, Gregory Salitra, Daniel Sharon, and Doron Aurbach. On the challenge of developing advanced technologies for electrochemical energy storage and conversion, 4 2014.

[12] Thierry Drezen, Nam-Hee Hee Kwon, Paul Bowen, Ivo Teerlinck, Motoshi Isono, and Ivan Exnar. Effect of particle size on LiMnPO4 cathodes. *Journal of Power Sources*, 174(2):949–953, 12 2007.

[13] Yue Liu, Tianlu Zhao, Wangwei Ju, Siqi Shi, Siqi Shi, and Siqi Shi. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177, 9 2017.

[14] Mlitz Kimberly. • Global Big Data market size 2011-2027 — Statista.

[15] Wenwen Li, Yasunobu Ando, Emi Minamitani, and Satoshi Watanabe. Study of Li atom diffusion in amorphous Li3PO4 with neural network potential. *Journal of Chemical Physics*, 147(21):214106, 12 2017.

[16] Tanmay Sarkar, Alind Sharma, Abhik Kumar Das, Dipti Deodhare, and Mridula D. Bharadwaj. A neural network based approach to predict high voltage li-ion battery cathode materials. In *Proceedings of the IEEE International Caracas Conference on Devices, Circuits and Systems, ICCDCS*. Institute of Electrical and Electronics Engineers Inc., 10 2014.

[17] Kyoungmin Min, Byungjin Choi, Kwangjin Park, and Eunseog Cho. Machine learning assisted optimization of electrochemical properties for Ni-rich cathode materials. *Scientific Reports*, 8(1):1–7, 12 2018.

[18] Zhisen Jiang, Jizhou Li, Yang Yang, Linqin Mu, Chenxi Wei, Xiqian Yu, Piero Pianetta, Kejie Zhao, Peter Cloetens, Feng Lin, and Yijin Liu. Machine-learning-revealed statistics of the particle-carbon/binder detachment in lithium-ion battery cathodes. *Nature Communications*, 11(1):1–9, 12 2020.

[19] Cancan Peng, Jijun Huang, Yujiao Guo, Qiling Li, Hongli Bai, Yonghui He, Changwei Su, and Junming Guo. Electrochemical performance of spinel LiAlxMn2-xO4 prepared rapidly by glucose-assisted solid-state combustion synthesis. *Vacuum*, 120(PA):121–126, 7 2015.

[20] Michael A. Arbib. The handbook of brain theory and neural networks. page 1118, 1995.

[21] Perceptrons - the most basic form of a neural network · Applied Go.

[22] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning. 2009.

[23] A. Iturrondobeitia, A. Goñi, V. Palomares, I. Gil De Muro, L. Lezama, and T. Rojo. Effect of doping LiMn 2O 4 spinel with a tetravalent species such as Si(IV) versus with a trivalent species such as Ga(III). Electrochemical, magnetic and ESR study. *Journal of Power Sources*, 216:482–488, 10 2012.

[24] Arup Chakraborty, Sooraj Kunnikuruvan, Sandeep Kumar, Boris Markovsky, Doron Aurbach, Mudit Dixit, and Dan Thomas Major. Layered Cathode Materials for Lithium-Ion Batteries: Review of Computational Studies on LiNi1- x- yCoxMnyO2 and LiNi1- x- yCoxAlyO2. *Chemistry of Materials*, 32(3):915–952, 2 2020.

[25] Christian Julien, Alain Mauger, Karim Zaghib, and Henri Groult. Optimization of layered cathode materials for lithium-ion batteries, 2016.

[26] Qiqi Tao, Liguang Wang, Caihong Shi, Jun Li, Guang Chen, Zheng Xue,

Jichang Wang, Shun Wang, and Huile Jin. Understanding the Ni-rich layered structure materials for high-energy density lithium-ion batteries, 3 2021.

[27] Zhangxian Chen, Weixin Zhang, and Zeheng Yang. A review on cathode materials for advanced lithium ion batteries: Microstructure designs and performance regulations, 2020.

[28] Alain Mauger and Christian Julien. Olivine Positive Electrodes for Li-Ion Batteries: Status and Perspectives. *Batteries*, 4(3):39, 8 2018.

[29] Lu Lu Zhang, Gan Liang, Alexander Ignatov, Mark C. Croft, Xiao Qin Xiong, I. Ming Hung, Yun Hui Huang, Xian Luo Hu, Wu Xing Zhang, and Yun Long Peng. Effect of vanadium incorporation on electrochemical performance of LiFePO4 for lithium-ion batteries. *Journal of Physical Chemistry C*, 115(27):13520–13527, 7 2011.

[30] Thomas G. Dietterich. Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 8 2000.

[31] David S. Hall, Julian Self, and J. R. Dahn. Dielectric Constants for Quantum Chemistry and Li-Ion Batteries: Solvent Blends of Ethylene Carbonate and Ethyl Methyl Carbonate. *Journal of Physical Chemistry C*, 119(39):22322–22330, 9 2015.

[32] Siavash Asiaban, Nezmin Kayedpour, Arash E. Samani, Dimitar Bozalakov, Jeroen D. M. De Kooning, Guillaume Crevecoeur, and Lieven Vandevelde. Wind and Solar Intermittency and the Associated Integration Challenges: A Comprehensive Review Including the Status in the Belgian Power System. *Energies 2021, Vol. 14, Page 2630*, 14(9):2630, 5 2021.

[33] Vijay Devabhaktuni, Mansoor Alam, Soma Shekara Sreenadh Reddy Depuru, Robert C. Green, Douglas Nims, and Craig Near. Solar energy: Trends and enabling technologies. *Renewable and Sustainable Energy Reviews*, 19:555–564, 3 2013.

[34] S. Mekhilef, R. Saidur, and A. Safari. A review on solar energy use in industries. *Renewable and Sustainable Energy Reviews*, 15(4):1777–1790, 5 2011.

[35] Manisha Sawant, Sameer Thakare, A. Prabhakara Rao, Andrés E. Feijóo-Lorenzo, and Neeraj Dhanraj Bokde. A Review on State-of-the-Art Reviews in Wind-Turbine- and Wind-Farm-Related Topics. *Energies 2021, Vol. 14, Page 2041*, 14(8):2041, 4 2021.

[36] Radian Belu. Assessment and Analysis of Offshore Wind Energy Potential. *Entropy and Exergy in Renewable Energy [Working Title]*, 12 2020.

[37] Qi Li, Yang Liu, Shaohua Guo, and Haoshen Zhou. Solar energy storage in the rechargeable batteries. *Nano Today*, 16:46–60, 10 2017.

[38] Tesla battery in South Australia expanded by 50 per cent, energy minister lauds benefits - ABC News.

[39] ALEXANDER MAURO. The Role of the Voltaic Pile in the Galvani-Volta Controversy Concerning Animal vs. Metallic Electricity. *Journal of the History of Medicine and Allied Sciences*, XXIV(2):140–150, 4 1969.

[40] P. Kurzweil. Gaston Planté and his invention of the lead-acid battery-The genesis of the first practical rechargeable battery. *Journal of Power Sources*, 195(14):4424–4434, 7 2010.

[41] Anthony Green. Characteristics of the nickel-cadmium battery for energy storage. *Power Engineering Journal*, 13(3):117–121, 1999.

[42] Uwe Koehler, Franz J. Kruger, Joerg Kuempers, Matthias Maul, Eberhard Niggemann, and Herbert H. Schoenfelder. High performance nickel-metal hydride and lithium-ion batteries. *Proceedings of the Intersociety Energy Conversion Engineering Conference*, 1:93–98, 1997.

[43] M. Stanley Whittingham. Intercalation chemistry and energy storage. *Journal of Solid State Chemistry*, 29(3):303–310, 9 1979.

[44] K. Mizushima, P. C. Jones, P. J. Wiseman, and J. B. Goodenough. LixCoO2 (0¡x1): A new cathode material for batteries of high energy density. *Solid State Ionics*, 3-4(C):171–174, 8 1981.

[45] Qi Liu, Xin Su, Dan Lei, Yan Qin, Jianguo Wen, Fangmin Guo, Yimin A. Wu, Yangchun Rong, Ronghui Kou, Xianghui Xiao, Frederic Aguesse, Javier Bareño, Yang Ren, Wenquan Lu, and Yangxing Li. Approaching the capacity limit of lithium cobalt oxide in lithium ion batteries via lanthanum and aluminium doping. *Nature Energy 2018 3:11*, 3(11):936–943, 6 2018.

[46] Da Deng. Li-ion batteries: basics, progress, and challenges. *Energy Science & Engineering*, 3(5):385–418, 9 2015.

[47] Kazunori Ozawa. Lithium-ion rechargeable batteries with LiCoO2 and carbon electrodes: the LiCoO2/C system. *Solid State Ionics*, 69(3-4):212–221, 8 1994.

[48] Ana Irina Stan, Maciej Swierczynski, Daniel Ioan Stroe, Remus Teodorescu, and Soren Juhl Andreasen. Lithium ion battery chemistries from renewable energy storage to automotive and back-up power applications - An overview. In *2014 International Conference on Optimization of Electrical and Electronic Equipment, OPTIM 2014*, pages 713–720. IEEE Computer Society, 2014.

[49] Qinglin Zhang, Jie Pan, Peng Lu, Zhongyi Liu, Mark W. Verbrugge, Brian W. Sheldon, Yang Tse Cheng, Yue Qi, and Xingcheng Xiao. Synergetic Effects of Inorganic Components in Solid Electrolyte Interphase on High Cycle Efficiency of Lithium Ion Batteries. *Nano Letters*, 16(3):2011–2016, 3 2016.

[50] Kai Chao Pu, Xin Zhang, Xiao Lei Qu, Jian Jiang Hu, Hai Wen Li, Ming Xia Gao, Hong Ge Pan, and Yong Feng Liu. Recently developed strategies to restrain dendrite growth of Li metal anodes for rechargeable batteries, 6 2020.

[51] He Liu, Xin Bing Cheng, Jia Qi Huang, Hong Yuan, Yang Lu, Chong Yan, Gao Long Zhu, Rui Xu, Chen Zi Zhao, Li Peng Hou, Chuanxin He, Ste-

fan Kaskel, and Qiang Zhang. Controlling Dendrite Growth in Solid-State Electrolytes, 3 2020.

[52] Sasithorn Klongkan and Jantrawan Pumchusak. Effects of Nano Alumina and Plasticizers on Morphology, Ionic Conductivity, Thermal and Mechanical Properties of PEO-LiCF3SO3 Solid Polymer Electrolyte. *Electrochimica Acta*, 161:171–176, 4 2015.

[53] Wangyu Li, Ying Pang, Jingyuan Liu, Guanghui Liu, Yonggang Wang, and Yongyao Xia. A PEO-based gel polymer electrolyte for lithium ion batteries. *RSC Advances*, 7(38):23494–23501, 4 2017.

[54] Snehashis Choudhury, Sanjuna Stalin, Duylinh Vu, Alexander Warren, Yue Deng, Prayag Biswal, and Lynden A. Archer. Solid-state polymer electrolytes for high-performance lithium metal batteries. *Nature Communications*, 10(1):1–8, 12 2019.

[55] Liping Yue, Jun Ma, Jianjun Zhang, Jingwen Zhao, Shanmu Dong, Zhihong Liu, Guanglei Cui, and Liquan Chen. All solid-state polymer electrolytes for high-performance lithium ion batteries, 10 2016.

[56] Ava Hosseinioun, Pinchas Nürnberg, Monika Schönhoff, Diddo Diddens, and Elie Paillard. Improved lithium ion dynamics in crosslinked PMMA gel polymer electrolyte. *RSC Advances*, 9(47):27574–27582, 8 2019.

[57] M. Z. Kufian, M. F. Aziz, M. F. Shukur, A. S. Rahim, N. E. Ariffin, N. E.A. Shuhaimi, S. R. Majid, R. Yahya, and A. K. Arof. PMMA–LiBOB gel electrolyte for application in lithium ion batteries. *Solid State Ionics*, 208:36–42, 2 2012.

[58] Hong Ryun Jung and Wan Jin Lee. Electrochemical characteristics of electrospun poly(methyl methacrylate)/polyvinyl chloride as gel polymer electrolytes for lithium ion battery. *Electrochimica Acta*, 58(1):674–680, 12 2011.

[59] Yan Qiu Shen, Fang Lei Zeng, Xin Yu Zhou, An bang Wang, Wei kun Wang, Ning Yi Yuan, and Jian Ning Ding. A novel permselective organo-polysulfides/PVDF gel polymer electrolyte enables stable lithium anode for lithium–sulfur batteries. *Journal of Energy Chemistry*, 48:267–276, 9 2020.

[60] W Liu, X K Zhang, F Wu, and Y Xiang. A study on PVDF-HFP gel polymer electrolyte for lithium-ion batteries. *IOP Conference Series: Materials Science and Engineering*, 213(1):012036, 6 2017.

[61] Anil Arya and A. L. Sharma. Polymer electrolytes for lithium ion batteries: a critical study. *Ionics 2017 23:3*, 23(3):497–540, 1 2017.

[62] Yan Jie Wang and Dukjoon Kim. Crystallinity, morphology, mechanical properties and conductivity study of in situ formed PVdF/LiClO4/TiO2 nanocomposite polymer electrolytes. *Electrochimica Acta*, 52(9):3181–3189, 2 2007.

[63] Shyamal K. Das, Soumit S. Mandal, and Aninda J. Bhattacharyya. Ionic conductivity, mechanical strength and Li-ion battery performance of mono-functional and bi-functional ("Janus") "soggy sand" electrolytes. *Energy & Environmental Science*, 4(4):1391–1399, 3 2011.

[64] Young Woo Lee, Won Kyung Shin, and Dong Won Kim. Cycling performance of lithium-ion polymer batteries assembled using in-situ chemical cross-linking without a free radical initiator. *Solid State Ionics*, 255:6–12, 2 2014.

[65] Xiang Guan, Qingping Wu, Xiaowan Zhang, Xuhong Guo, Chilin Li, and Jun Xu. In-situ crosslinked single ion gel polymer electrolyte with superior performances for lithium metal batteries. *Chemical Engineering Journal*, 382:122935, 2 2020.

[66] Candice F. J. Francis, Ilias L. Kyratzis, and Adam S. Best. Lithium-Ion Battery Separators for Ionic-Liquid Electrolytes: A Review. *Advanced Materials*, 32(18):1904205, 5 2020.

[67] Zhifang Liu, Yingjun Jiang, Qiaomei Hu, Songtao Guo, Le Yu, Qi Li, Qing Liu, and Xianluo Hu. Safer Lithium-Ion Batteries from the Separator Aspect: Development and Future Perspectives. *ENERGY & ENVIRONMENTAL MATERIALS*, page eem2.12129, 10 2020.

[68] Xiaoyu Jiang, Xiaoming Zhu, Xinping Ai, Hanxi Yang, and Yuliang Cao. Novel Ceramic-Grafted Separator with Highly Thermal Stability for Safe Lithium-Ion Batteries. *ACS Applied Materials and Interfaces*, 9(31):25970–25975, 8 2017.

[69] Wonjun Na, Ki Hwan Koh, Albert S. Lee, Sangho Cho, Byoeri Ok, Suk Won Hwang, Jin Hong Lee, and Chong Min Koo. Binder-less chemical grafting of SiO2 nanoparticles onto polyethylene separators for lithium-ion batteries. *Journal of Membrane Science*, 573:621–627, 3 2019.

[70] Xiaosong Huang. Performance evaluation of a non-woven lithium ion battery separator prepared through a paper-making process. *Journal of Power Sources*, 256:96–101, 6 2014.

[71] Wang Yi, Zhan Huaiyu, Hu Jian, Liang Yun, and Zeng Shushu. Wet-laid nonwoven fabric for separator of lithium-ion battery. *Journal of Power Sources*, 189(1):616–619, 4 2009.

[72] Hun Lee, Meltem Yanilmaz, Ozan Toprakci, Kun Fu, and Xiangwu Zhang. A review of recent developments in membrane separators for rechargeable lithium-ion batteries, 12 2014.

[73] Zhongliang Hu, Jingying Li, Xiaojing Zhang, and Yirong Zhu. Strategies to Improve the Performance of Li Metal Anode for Rechargeable Batteries. *Frontiers in Chemistry*, 8:409, 5 2020.

[74] Chih Kai Yang. Erratum: A metallic graphene layer adsorbed with lithium (Applied Physics Letters (2009) 94 (163115) DOI: 10.1063/1.3194145), 7 2009.

[75] Guoxiu Wang, Xiaoping Shen, Jane Yao, and Jinsoo Park. Graphene nanosheets for enhanced lithium storage in lithium ion batteries. *Carbon*, 47(8):2049–2053, 7 2009.

[76] Eun Joo Yoo, Jedeok Kim, Eiji Hosono, Hao Shen Zhou, Tetsuichi Kudo, and Itaru Honma. Large reversible Li storage of graphene nanosheet families for use in rechargeable lithium ion batteries. *Nano Letters*, 8(8):2277–2282, 8 2008.

[77] Hailiang Wang, Li Feng Cui, Yuan Yang, Hernan Sanchez Casalongue, Joshua Tucker Robinson, Yongye Liang, Yi Cui, and Hongjie Dai. Mn3O4-graphene hybrid as a high-capacity anode material for lithium ion batteries. *Journal of the American Chemical Society*, 132(40):13978–13980, 10 2010.

[78] M. N. Obrovac and V. L. Chevrier. Alloy negative electrodes for Li-ion batteries, 12 2014.

[79] Aashutosh N. Mistry, Kandler Smith, and Partha P. Mukherjee. Secondary-Phase Stochastics in Lithium-Ion Battery Electrodes. *ACS Applied Materials and Interfaces*, 10(7):6317–6326, 2 2018.

[80] Simon Hein, Timo Danner, Daniel Westhoff, Benedikt Prifling, Rares Scurtu, Lea Kremer, Alice Hoffmann, André Hilger, Markus Osenberg, Ingo Manke, Margret Wohlfahrt-Mehrens, Volker Schmidt, and Arnulf Latz. Influence of Conductive Additives and Binder on the Impedance of Lithium-Ion Battery Electrodes: Effect of Morphology. *Journal of The Electrochemical Society*, 167(1):013546, 1 2020.

[81] Honghe Zheng, Ruizhi Yang, Gao Liu, Xiangyun Song, and Vincent S. Battaglia. Cooperation between active material, polymeric binder and conductive carbon additive in lithium ion battery cathode. *Journal of Physical Chemistry C*, 116(7):4875–4882, 2 2012.

[82] Ozan Toprakci, Hatice A.K. Toprakci, Ying Li, Liwen Ji, Leigang Xue, Hun Lee, Shu Zhang, and Xiangwu Zhang. Synthesis and characterization of

xLi2MnO3·(1 - X)LiMn1/3Ni1/3Co1/3O2 composite cathode materials for rechargeable lithium-ion batteries. *Journal of Power Sources*, 241:522–528, 11 2013.

[83] Jiexi Wang, Zhixing Wang, Xinhai Li, Huajun Guo, Xianwen Wu, Xiaoping Zhang, and Wei Xiao. XLi3V2(PO4)3·LiVPO 4F/C composite cathode materials for lithium ion batteries. *Electrochimica Acta*, 87:224–229, 1 2013.

[84] Takahiro Kozawa, Toshiya Harata, and Makio Naito. Fabrication of an LiMn ¡sub¿2¡/sub¿ O ¡sub¿4¡/sub¿ @LiMnPO ¡sub¿4¡/sub¿ composite cathode for improved cycling performance at high temperatures. *Journal of Asian Ceramic Societies*, 8(2):309–317, 4 2020.

[85] Zhengrui Xu, Libin Gao, Yijing Liu, and Le Li. Review—Recent Developments in the Doped LiFePO 4 Cathode Materials for Power Lithium Ion Batteries . *Journal of The Electrochemical Society*, 163(13):A2600–A2610, 9 2016.

[86] Wuwei Yan, Shunyi Yang, Youyuan Huang, Yong Yang, and Guohui Yuan. A review on doping/coating of nickel-rich cathode materials for lithium-ion batteries, 4 2020.

[87] Miklos Lengyel, Kuan Yu Shen, Deanna M. Lanigan, Jonathan M. Martin, Xiaofeng Zhang, and Richard L. Axelbaum. Trace level doping of lithium-rich cathode materials. *Journal of Materials Chemistry A*, 4(9):3538–3545, 2 2016.

[88] Tahir Sattar, Seung Hwan Lee, Bong Soo Jin, and Hyun Soo Kim. Influence of Mo addition on the structural and electrochemical performance of Ni-rich cathode material for lithium-ion batteries. *Scientific Reports*, 10(1):1–10, 12 2020.

[89] Ilkyu Hwang, Chul Wee Lee, Jae Chang Kim, and Songhun Yoon. Particle size effect of Ni-rich cathode materials on lithium ion battery performance. *Materials Research Bulletin*, 47(1):73–78, 1 2012.

[90] Naoaki Yabuuchi, Kei Kubota, Yoshinori Aoki, and Shinichi Komaba. Understanding Particle-Size-Dependent Electrochemical Properties of Li2MnO3-Based Positive Electrode Materials for Rechargeable Lithium Batteries. *Journal of Physical Chemistry C*, 120(2):875–885, 1 2016.

[91] Zonghai Chen, Yan Qin, Khalil Amine, and Y. K. Sun. Role of surface coating on cathode materials for lithium-ion batteries. *Journal of Materials Chemistry*, 20(36):7606–7612, 9 2010.

[92] Peiyuan Guan, Lu Zhou, Zhenlu Yu, Yuandong Sun, Yunjian Liu, Feixiang Wu, Yifeng Jiang, and Dewei Chu. Recent progress of surface coating on cathode materials for high-performance lithium-ion batteries, 4 2020.

[93] Steven Kauwe, Trevor Rhone, and Taylor Sparks. Data-Driven Studies of Li-Ion-Battery Materials. *Crystals*, 9(1):54, 1 2019.

[94] Guanyu Wang, Tom Fearn, Tengyao Wang, and Kwang-Leong Choy. Insight Gained from Using Machine Learning Techniques to Predict the Discharge Capacities of Doped Spinel Cathode Materials for Lithium-Ion Batteries Applications. *Energy Technology*, page 2100053, 3 2021.

[95] Guanyu Wang, Tom Fearn, Tengyao Wang, Kwang-leong Choy, and G Y Wang. Machine Learning Approach for Predicting the Discharging Capacities of Doped lithium Nickel-Cobalt-Manganese Cathode Materials in Li-ion Batteries. *ACS Central Science*, pages 1–19, 8 2021.

[96] John B. Goodenough and Kyu Sung Park. The Li-ion rechargeable battery: A perspective, 1 2013.

[97] E. Plichta, M. Salomon, S. Slane, M. Uchiyama, D. Chua, W. B. Ebner, and H. W. Lin. A rechargeable Li/LixCoO2 Cell. *Journal of Power Sources*, 21(1):25–31, 8 1987.

[98] N. Furukawa. Development and commercialization of nickel-metal hydride secondary batteries. *Journal of Power Sources*, 51(1-2):45–59, 1994.

[99] Atsushi Ito, Decheng Li, Yuichi Sato, Masazumi Arao, Manabu Watanabe, Masaharu Hatano, Hideaki Horie, and Yasuhiko Ohsawa. Cyclic deterioration and its improvement for Li-rich layered cathode material Li[Ni0.17Li0.2Co0.07Mn0.56]O2. *Journal of Power Sources*, 195(2):567–573, 1 2010.

[100] Na Yeon Kim, Taeeun Yim, Jun Ho Song, Ji Sang Yu, and Zonghoon Lee. Microstructural study on degradation mechanism of layered LiNi0.6Co0.2Mn0.2O2 cathode materials by analytical transmission electron microscopy. *Journal of Power Sources*, 307:641–648, 3 2016.

[101] Shijian Zheng, Rong Huang, Yoshinari Makimura, Yoshio Ukyo, Craig A.J. Fisher, Tsukasa Hirayama, and Yuichi Ikuhara. Microstructural changes in LiNi0.8Co0.15Al 0.05O2 positive electrode material during the first cycle. *Journal of the Electrochemical Society*, 158(4), 2011.

[102] M. Broussely, F. Perton, P. Biensan, J. M. Bodet, J. Labat, A. Lecerf, C. Delmas, A. Rougier, and J. P. Pérès. LixNiO2, a promising cathode for rechargeable lithium batteries. *Journal of Power Sources*, 54(1):109–114, 1995.

[103] Tae Joon Kim, Dongyeon Son, Jaephil Cho, and Byungwoo Park. Enhancement of the electrochemical properties of o-LiMnO2 cathodes at elevated temperature by lithium and fluorine additions. *Journal of Power Sources*, 154(1):268–272, 3 2006.

[104] Shuang Liu, L. Xiong, and C. He. Long cycle life lithium ion battery with lithium nickel cobalt manganese oxide (NCM) cathode. *Journal of Power Sources*, 261:285–291, 9 2014.

[105] Chung Hsin Lu and Lee Wei-Cheng. Reaction mechanism and kinetics analysis of lithium nickel oxide during solid-state reaction. *Journal of Materials Chemistry*, 10(6):1403–1407, 1 2000.

[106] Xiaoqing Wang, Hiroyoshi Nakamura, and Masaki Yoshio. Capacity fad-

ing mechanism for oxygen defect spinel as a 4 V cathode material in Li-ion batteries. *Journal of Power Sources*, 110(1):19–26, 7 2002.

[107] Atsuo Yamada, Masahiro Tanaka, Koichi Tanaka, and Koji Sekai. Jahn-Teller instability in spinel Li-Mn-O. *Journal of Power Sources*, 81-82:73–78, 9 1999.

[108] Jina Cao, Gaoshao Cao, Hongming Yu, Jian Xie, and Xinbing Zhao. Synthesis and electrochemical performance of YF3-coated LiMn2O4 cathode materials for Li-ion batteries. *Rare Metals*, 30(1):39–43, 2 2011.

[109] Lifen Xiao, Yanqiang Zhao, Yanyan Yang, Yuliang Cao, Xinping Ai, and Hanxi Yang. Enhanced electrochemical stability of Al-doped LiMn2O4 synthesized by a polymer-pyrolysis method. *Electrochimica Acta*, 54(2):545–550, 12 2008.

[110] Hongyuan Zhao, Xiuzhi Bai, Jing Wang, Dongdong Li, Bo Li, Yashuang Wang, Li Dong, Binbin Liu, and Sridhar Komarneni. Enhanced Cycling Stability through Erbium Doping of LiMn2O4 Cathode Material Synthesized by Sol-Gel Technique. *Materials*, 11(9):1558, 8 2018.

[111] M. Michalska, D. A. Ziółkowska, J. B. Jasiński, P. H. Lee, P. Ławniczak, B. Andrzejewski, A. Ostrowski, W. Bednarski, S. H. Wu, and J. Y. Lin. Improved electrochemical performance of LiMn2O4 cathode material by Ce doping. *Electrochimica Acta*, 276:37–46, 6 2018.

[112] Subramani Bhuvaneswari, U. V. Varadaraju, R. Gopalan, and Raju Prakash. Structural stability and superior electrochemical performance of Sc-doped LiMn 2 O 4 spinel as cathode for lithium ion batteries. *Electrochimica Acta*, 301:342–351, 4 2019.

[113] D. Liu, W. Zhu, J. Trottier, C. Gagnon, F. Barray, A. Guerfi, A. Mauger, H. Groult, C. M. Julien, J. B. Goodenough, and K. Zaghib. Spinel materials for high-voltage cathodes in Li-ion batteries, 11 2014.

[114] Gemeng Liang, Vanessa K. Peterson, Khay Wai See, Zaiping Guo, and Wei Kong Pang. Developing high-voltage spinel LiNi0.5Mn1.5O4cathodes for high-energy-density lithium-ion batteries: Current achievements and future prospects, 8 2020.

[115] Weikang Li, Yoon Gyo Cho, Weiliang Yao, Yixuan Li, Ashley Cronk, Ryosuke Shimizu, Marshall A. Schroeder, Yanbao Fu, Feng Zou, Vince Battaglia, Arumugam Manthiram, Minghao Zhang, and Ying Shirley Meng. Enabling high areal capacity for Co-free high voltage spinel materials in next-generation Li-ion batteries. *Journal of Power Sources*, 473:228579, 10 2020.

[116] Shiyou Li, Konglei Zhu, Dongni Zhao, Qiuping Zhao, and Ningshuang Zhang. Porous LiMn2O4 with Al2O3 coating as high-performance positive materials. *Ionics*, 25(5):1991–1998, 5 2019.

[117] Dongsheng Guan, Judith A. Jeevarajan, and Ying Wang. Enhanced cycleability of LiMn2O4 cathodes by atomic layer deposition of nanosized-thin Al2O3 coatings. *Nanoscale*, 3(4):1465–1469, 4 2011.

[118] Hyung Wook Ha, Nan Ji Yun, and Keon Kim. Improvement of electrochemical stability of LiMn2O4 by CeO2 coating for lithium-ion batteries. *Electrochimica Acta*, 52(9):3236–3241, 2 2007.

[119] Chunbo Qing, Ying Bai, Jueming Yang, and Weifeng Zhang. Enhanced cycling stability of LiMn2O4 cathode by amorphous FePO4 coating. *Electrochimica Acta*, 56(19):6612–6618, 7 2011.

[120] Artur Tron, Yeong Don Park, and Junyoung Mun. AlF3-coated LiMn2O4 as cathode material for aqueous rechargeable lithium battery with improved cycling stability. *Journal of Power Sources*, 325:360–364, 9 2016.

[121] Quanqi Chen, Yaobin Wang, Tingting Zhang, Wumei Yin, Jianwen Yang, and Xianyou Wang. Electrochemical performance of LaF 3-coated LiMn 2O

4 cathode materials for lithium ion batteries. *Electrochimica Acta*, 83:65–72, 11 2012.

[122] Qing Zhu, Shuai Zheng, Xuwu Lu, Yi Wan, Quanqi Chen, Jianwen Yang, Ling Zhi Zhang, and Zhouguang Lu. Improved cycle performance of LiMn2O4 cathode material for aqueous rechargeable lithium battery by LaF3 coating. *Journal of Alloys and Compounds*, 654:384–391, 1 2016.

[123] Shu Juan Bao, Chang Ming Li, Hu Lin Li, and John H.T. Luong. Morphology and electrochemistry of LiMn2O4 optimized by using different Mn-sources. *Journal of Power Sources*, 164(2):885–889, 2 2007.

[124] Hua Li Zhu, Zhao Yong Chen, Shan Ji, and Vladimir Linkov. Influence of different morphologies on electrochemical performance of spinel LiMn2O4. *Solid State Ionics*, 179(27-32):1788–1793, 9 2008.

[125] Donglei Guo, Zhaorong Chang, Hongwei Tang, Bao Li, Xinhong Xu, Xiao Zi Yuan, and Haijiang Wang. Electrochemical performance of solid sphere spinel LiMn2O 4 with high tap density synthesized by porous spherical Mn 3O4. *Electrochimica Acta*, 123:254–259, 3 2014.

[126] Zhumabay Bakenov and Izumi Taniguchi. LiMnPO 4 olivine as a cathode for lithium batteries. *Open Materials Science Journal*, 5:222–227, 2011.

[127] Wei Jun Zhang. Structure and performance of LiFePO4 cathode materials: A review, 3 2011.

[128] Zhumabay Bakenov and Izumi Taniguchi. Electrochemical performance of nanocomposite LiMnPO4/C cathode materials for lithium batteries. *Electrochemistry Communications*, 12(1):75–78, 1 2010.

[129] R. Sharabi, E. Markevich, V. Borgel, G. Salitra, D. Aurbach, G. Semrau, M. A. Schmidt, N. Schall, and C. Stinner. Significantly improved cycling performance of LiCoPO4 cathodes. *Electrochemistry Communications*, 13(8):800–802, 8 2011.

[130] Joanna L Dodd. Phase Composition and Dynamical Studies of Lithium Iron Phosphate. Technical report, 2007.

[131] Xiaoyu Zhang, Martijn Van Hulzen, Deepak P. Singh, Alex Brownrigg, Jonathan P. Wright, Niels H. Van Dijk, and Marnix Wagemaker. Direct view on the phase evolution in individual LiFePO 4 nanoparticles during Li-ion battery cycling. *Nature Communications*, 6(1):1–7, 9 2015.

[132] Mingyuan Zhu, Linfeng Cheng, Yang Liu, Wenxian Li, Pengfei Hu, Hongming Jin, Yemin Hu, and Ying Li. LiFePO4/(C+Cu) composite with excellent cycling stability as lithium ion battery cathodes synthesized via a modified carbothermal reduction method. *Ceramics International*, 44(11):12106–12111, 8 2018.

[133] Jiajun Wang and Xueliang Sun. Understanding and recent development of carbon coating on LiFePO 4 cathode materials for lithium-ion batteries, 1 2012.

[134] K. S. Park, J. T. Son, H. T. Chung, S. J. Kim, C. H. Lee, K. T. Kang, and H. G. Kim. Surface modification by silver coating for improving electrochemical properties of LiFePO4. *Solid State Communications*, 129(5):311–314, 2 2004.

[135] Fiona C. Strobridge, Hao Liu, Michal Leskes, Olaf J. Borkiewicz, Kamila M. Wiaderek, Peter J. Chupas, Karena W. Chapman, and Clare P. Grey. Unraveling the Complex Delithiation Mechanisms of Olivine-Type Cathode Materials, LiFexCo1-xPO4. *Chemistry of Materials*, 28(11):3676–3690, 6 2016.

[136] D. Yu Gryzlov, S. A. Novikova, T. L. Kulova, A. M. Skundin, and A. B. Yaroslavtsev. The Effect of Particle Size on the Processes of Charging and Discharging of the LiFe0.97Ni0.03PO4/C/Ag Cathode Material. *Russian Journal of Electrochemistry*, 54(5):442–450, 5 2018.

[137] Min Zhang, Nuria Garcia-Araez, and Andrew L. Hector. Understanding and

development of olivine LiCoPO4 cathode materials for lithium-ion batteries, 7 2018.

[138] J. Yao, S. Bewlay, K. Konstantionv, V. A. Drozd, R. S. Liu, X. L. Wang, H. K. Liu, and G. X. Wang. Characterisation of olivine-type LiMnxFe1-xPO4 cathode materials. *Journal of Alloys and Compounds*, 425(1-2):362–366, 11 2006.

[139] Claus Daniel, Debasish Mohanty, Jianlin Li, and David L Wood. Cathode materials review ARTICLES YOU MAY BE INTERESTED IN Cathodes-Technological review AIP Conference Material review of Li ion battery separators AIP Conference Fundamental principals of battery design: Porous electrodes AIP Conference. 1597:14, 2014.

[140] K. Vijaya Babu, L. Seeta Devi, V. Veeraiah, and K. Anand. Structural and dielectric studies of LiNiPO 4 and LiNi 0.5 Co 0.5 PO 4 cathode materials for lithium-ion batteries. *Journal of Asian Ceramic Societies*, 4(3):269–276, 9 2016.

[141] Tao Deng, Xiulin Fan, Longsheng Cao, Dong Su, Xiao-Qing Yang, Chunsheng Wang, Ji Chen, Singyuk Hou, Xiao Ji, Long Chen, Shuang Li, Xiuquan Zhou, and Enyuan Hu. Designing In-Situ-Formed Interphases Enables Highly Reversible Cobalt-Free LiNiO 2 Cathode for Li-ion and Li-metal Batteries. *Joule*, 3, 2019.

[142] Hongwei Yan, Xuejie Huang, and Liquan Chen. Microwave synthesis of LiMn2O4 cathode material. *Journal of Power Sources*, 81-82:647–650, 1999.

[143] Jinyun He, Weimin Wang, Zhengguang Zou, Fei Long, and Zhengyi Fu. Solvothermal synthesis and electrochemical performance of Ag-doped V6O13 as cathode material for lithium-ion battery. *Ionics*, 20(8):1063–1070, 8 2014.

[144] Arumugam Manthiram, Katharine Chemelewski, and Eun Sung Lee. A perspective on the high-voltage LiMn1.5Ni0.5O 4 spinel cathode for lithium-ion batteries. In *Energy and Environmental Science*, volume 7, pages 1339–1350. Royal Society of Chemistry, 3 2014.

[145] Yuki Orikasa, Takehiro Maeda, Yukinori Koyama, Haruno Murayama, Katsutoshi Fukuda, Hajime Tanida, Hajime Arai, Eiichiro Matsubara, Yoshiharu Uchimoto, and Zempachi Ogumi. Transient phase change in two phase reaction between LiFePO4 and FePO4 under battery operation. *Chemistry of Materials*, 25(7):1032–1039, 4 2013.

[146] Yuzhen Duan, Junming Guo, Mingwu Xiang, Jinyu Zhu, Changwei Su, Hongli Bai, Xiaofang Liu, Wei Bai, and Rui Wang. Single crystalline polyhedral LiNi x Mn 2-x O 4 as high-performance cathodes for ultralong cycling lithium-ion batteries. *Solid State Ionics*, 326(July):100–109, 2018.

[147] Jong H. Lee, Jin K. Hong, Dong H. Jang, Y. K. Sun, and Seung M. Oh. Degradation mechanisms in doped spinels of LiM0.05Mn1.95O4 (M = Li, B, Al, Co, and Ni) for Li secondary batteries. *Journal of Power Sources*, 89(1):7–14, 7 2000.

[148] Y. L. Ding, J. Xie, G. S. Cao, T. J. Zhu, H. M. Yu, and X. B. Zhao. Enhanced elevated-temperature performance of Al-doped single-crystalline LiMn2O4 nanotubes as cathodes for lithium ion batteries. *Journal of Physical Chemistry C*, 115(19):9821–9825, 5 2011.

[149] Burçak Ebin, Göran Lindbergh, and Sebahattin Gürmen. Preparation and electrochemical properties of nanocrystalline LiBxMn2-xO4 cathode particles for Li-ion batteries by ultrasonic spray pyrolysis method. *Journal of Alloys and Compounds*, 620:399–406, 1 2015.

[150] Gui Ming Song, Yu Jin Wang, and Yu Zhou. Synthesis and electrochemical performance of LiCrxMn 2-xO4 powders by mechanical activation and rotary heating. *Journal of Power Sources*, 128(2):270–277, 4 2004.

[151] D. Zhang, B. N. Popov, and R. E. White. Electrochemical investigation of CrO2.65 doped LiMn2O4 as a cathode material for lithium-ion batteries. *Journal of Power Sources*, 76(1):81–90, 11 1998.

[152] D. Arumugam and G. Paruthimal Kalaignan. Synthesis and electrochemical characterizations of nano size Ce doped LiMn2O4 cathode materials for rechargeable lithium batteries. *Journal of Electroanalytical Chemistry*, 648(1):54–59, 9 2010.

[153] Jijun Huang, Fangli Yang, Yujiao Guo, Cancan Peng, Hongli Bai, Jinhui Peng, and Junming Guo. LiMgxMn2-xO4 (x0.10) cathode materials with high rate performance prepared by molten-salt combustion at low temperature. *Ceramics International*, 41(8):9662–9667, 9 2015.

[154] Xiaoqing Wang, Osamu Tanaike, Masaya Kodama, and Hiroaki Hatori. High rate capability of the Mg-doped Li-Mn-O spinel prepared via coprecipitated precursor. *Journal of Power Sources*, 168(1 SPEC. ISS.):282–287, 5 2007.

[155] Monika Bakierska, Michał Świętosławski, Krystian Chudzik, Marcelina Lis, and Marcin Molenda. Enhancing the lithium ion diffusivity in LiMn2O4ySy cathode materials through potassium doping. *Solid State Ionics*, 317:190–193, 4 2018.

[156] Lilong Xiong, Youlong Xu, Cheng Zhang, Zhengwei Zhang, and Jiebin Li. Electrochemical properties of tetravalent Ti-doped spinel LiMn 2O 4. *Journal of Solid State Electrochemistry*, 15(6):1263–1269, 6 2011.

[157] Thangaian Kesavan, Chenrayan Senthil, and Manickam Sasidharan. Solvothermally synthesized Ti-rich LiMnTiO4 as cathode material for high Li storage. *Journal of Materials Science*, 53(6):4406–4416, 3 2018.

[158] Zhi Yuan Tang, Xing He Lu, and Na Zhang. The anion-cation multiple doping effect of spinel cathode materials on electrochemical speciality. *Acta Physico - Chimica Sinica*, 21(8):934–938, 8 2005.

[159] Azhar Iqbal, Yousaf Iqbal, Lin Chang, Safeer Ahmed, Zhiyong Tang, and Yan Gao. Enhanced electrochemical performance of La- and Zn-co-doped LiMn 2O4 spinel as the cathode material for lithium-ion batteries. *Journal of Nanoparticle Research*, 14(10):1–14, 10 2012.

[160] Yue Yu, Junming Guo, Mingwu Xiang, Changwei Su, Xiaofang Liu, Hongli Bai, Wei Bai, and Kaijiao Duan. Enhancing the durable performance of LiMn2O4 at high-rate and elevated temperature by nickel-magnesium dual doping. *Scientific Reports*, 9(1), 12 2019.

[161] Zengrong Yang, Yujie Wang, Xianchun Chen, Hao Wu, and Yun Zhang. Mg ¡sup¿2+¡/sup¿ and Ti ¡sup¿4+¡/sup¿ Co–Doped Spinel LiMn ¡sub¿2¡/sub¿ O ¡sub¿4¡/sub¿ as Lithium-Ion Battery Cathode. *ChemistrySelect*, 4(33):9583–9589, 9 2019.

[162] Thierry Drezen, Nam Hee Kwon, Paul Bowen, Ivo Teerlinck, Motoshi Isono, and Ivan Exnar. Effect of particle size on LiMnPO4 cathodes. *Journal of Power Sources*, 174(2):949–953, 12 2007.

[163] Yonggao Xia, Masaki Yoshio, and Hideyuki Noguchi. Improved electrochemical performance of LiFePO4 by increasing its specific surface area. *Electrochimica Acta*, 52(1):240–245, 10 2006.

[164] D. Arumugam and G. Paruthimal Kalaignan. Synthesis and electrochemical characterizations of nano size Ce doped LiMn2O4 cathode materials for rechargeable lithium batteries. *Journal of Electroanalytical Chemistry*, 648(1):54–59, 9 2010.

[165] Wei Liu, Weiyang Li, Denys Zhuo, Guangyuan Zheng, Zhenda Lu, Kai Liu, and Yi Cui. Core–Shell Nanoparticle Coating as an Interfacial Layer for Dendrite-Free Lithium Metal Anodes. *ACS Central Science*, 3(2):135–140, 2 2017.

[166] Weigang Wang, Guorong Hu, Zhongdong Peng, Ke Du, Yanbing Cao, and Jianguo Duan. Nano-sized over-lithiated oxide by a mechano-chemical

activation-assisted microwave technique as cathode material for lithium ion batteries and its electrochemical performance. *Ceramics International*, 44(2):1425–1431, 2 2018.

[167] Shubiao Xia, Jian Jun Liu, Fushao Li, Feixiang Cheng, Xue Li, Chengke Sun, and Hong Guo. Structure and morphology evolution in solid-phase synthesis lithium ion battery LiNi0.80Co0.15Al0.05O2 cathode materials with different micro-nano sizes of raw materials. *Ceramics International*, 44(8):9294–9302, 6 2018.

[168] Yingchao Chen, Kai Xie, Yi Pan, and Chunman Zheng. Nano-sized LiMn 2 O 4 spinel cathode materials exhibiting high rate discharge capability for lithium-ion batteries. *Journal of Power Sources*, 196(15):6493–6497, 8 2011.

[169] Renjie Chen, Taolin Zhao, Xiaoxiao Zhang, Li Li, and Feng Wu. Advanced cathode materials for lithium-ion batteries using nanoarchitectonics, 11 2016.

[170] Zuguang Yang, Wei Xiang, Zhenguo Wu, Fengrong He, Jun Zhang, Yao Xiao, Benhe Zhong, and Xiaodong Guo. Effect of niobium doping on the structure and electrochemical performance of LiNi 0.5 Co 0.2 Mn 0.3 O 2 cathode materials for lithium ion batteries. *Ceramics International*, 43(4):3866–3872, 3 2017.

[171] Qiao Ying Huang, Zhi Wu, Jing Su, Yun Fei Long, Xiao Yan Lv, and Yan Xuan Wen. Synthesis and electrochemical performance of Ti-Fe co-doped LiMnPO4/C as cathode material for lithium-ion batteries. *Ceramics International*, 42(9):11348–11354, 7 2016.

[172] Kaipeng Wu, Shan Yin, Shen Wang, Jinliang Zhu, and Weitang Yao. Construction of submicron-sized LiFe0.4Mn0.6PO4/C enwrapped into graphene framework for advanced Li-storage. *Carbon*, 169:55–64, 11 2020.

[173] Yao Fu, Hao Jiang, Yanjie Hu, Ling Zhang, and Chunzhong Li. Hierarchical porous Li4Mn5O12 nano/micro structure as superior cathode materials for Li-ion batteries. *Journal of Power Sources*, 261:306–310, 9 2014.

[174] Zhihong Lei, Jiulin Wang, Jun Yang, Yanna Nuli, and Zifeng Ma. Nano-/Microhierarchical-Structured LiMn 0.85 Fe 0.15 PO 4 Cathode Material for Advanced Lithium Ion Battery. *ACS Applied Materials and Interfaces*, 10(50):43552–43560, 12 2018.

[175] Alexandru Horatiu Marincaş, Firuţa Goga, Sorin Aurel Dorneanu, and Petru Ilea. Review on synthesis methods to obtain LiMn2O4-based cathode materials for Li-ion batteries, 3 2020.

[176] Annelise Kopp Alves, Carlos P. Bergmann, and Felipe Amorim Berutti. Hydrothermal Synthesis. pages 61–76. 2013.

[177] Arup Chakraborty, Mudit Dixit, Doron Aurbach, and Dan T. Major. Predicting accurate cathode properties of layered oxide materials using the SCAN meta-GGA density functional. *npj Computational Materials*, 4(1):60, 12 2018.

[178] Mahesh Datt Bhatt and Colm O'Dwyer. Recent progress in theoretical and computational investigations of Li-ion battery materials and electrolytes. *Physical Chemistry Chemical Physics*, 17(7):4799–4844, 2 2015.

[179] Guo Xu Zhang, Anthony M. Reilly, Alexandre Tkatchenko, and Matthias Scheffler. Performance of various density-functional approximations for cohesive properties of 64 bulk solids. *New Journal of Physics*, 20(6):063020, 6 2018.

[180] Tai Yu Pan, Ngoc Thanh Thuy Tran, Yin Chia Chang, and Wen Dung Hsu. First-principles study on the initial reactions at LiNi1/3Co1/3Mn1/3O2 cathode/electrolyte interface in lithium-ion batteries. *Applied Surface Science*, 507:144842, 3 2020.

[181] Juan J. De Pablo, Barbara Jones, Cora Lind Kovacs, Vidvuds Ozolins, and Arthur P. Ramirez. The Materials Genome Initiative, the interplay of experiment, theory and computation. *Current Opinion in Solid State and Materials Science*, 18(2):99–117, 2014.

[182] Elizabeth Gibney. Google AI algorithm masters ancient game of Go, 1 2016.

[183] John Nawara. Machine Learning: Face Recognition Technology Evidence in Criminal Trials. *University of Louisville Law Review*, 49, 2010.

[184] Karol J. Piczak. Environmental sound classification with convolutional neural networks. In *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, volume 2015-Novem. IEEE Computer Society, 11 2015.

[185] Jack Stilgoe. Machine learning, social learning and the governance of self-driving cars. *Social studies of science*, 48(1):25–56, 2 2018.

[186] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 7 2018.

[187] Yun Liu, Oladapo Christopher Esan, Zhefei Pan, and Liang An. Machine learning for advanced energy materials. *Energy and AI*, 3:100049, 3 2021.

[188] Teng Zhou, Zhen Song, and Kai Sundmacher. Big Data Creates New Opportunities for Materials Research: A Review on Methods and Applications of Machine Learning for Materials Design, 2019.

[189] Chi Chen, Yunxing Zuo, Weike Ye, Xiangguo Li, Zhi Deng, and Shyue Ping Ong. A Critical Review of Machine Learning of Energy Materials. *Advanced Energy Materials*, 10(8):1–36, 2020.

[190] Jonathan Schmidt, Mário R.G. Marques, Silvana Botti, and Miguel A.L. Marques. Recent advances and applications of machine learning in solid-state materials science, 12 2019.

[191] Jose F. Rodrigues, Larisa Florea, Maria C. F. de Oliveira, Dermot Diamond, and Osvaldo N. Oliveira. Big data and machine learning for materials science. *Discover Materials 2021 1:1*, 1(1):1–27, 4 2021.

[192] Geun Ho Gu, Juhwan Noh, Inkyung Kim, and Yousung Jung. Machine learning for renewable energy materials, 2019.

[193] Yun Zhang and Xiaojie Xu. Machine learning lattice constants for cubic perovskite A22+BBO6 compounds. *CrystEngComm*, 22(38):6385–6397, 10 2020.

[194] Yun Zhang and Xiaojie Xu. Machine learning lattice constants for spinel compounds. *Chemical Physics Letters*, 760:137993, 12 2020.

[195] Logan Williams, Arpan Mukherjee, and Krishna Rajan. Deep Learning Based Prediction of Perovskite Lattice Parameters from Hirshfeld Surface Fingerprints. *Journal of Physical Chemistry Letters*, 11(17):7462–7468, 9 2020.

[196] Yuxin Li, Wenhui Yang, Rongzhi Dong, and Jianjun Hu. Mlatticeabc: Generic Lattice Constant Prediction of Crystal Materials Using Machine Learning. *ACS Omega*, 6(17):11585–11594, 5 2021.

[197] Dong Yeon Kim, Taeseung Yoon, Youn Jeong Jang, Jin Ho Lee, Younghoon Na, Byeong Jun Lee, Jae Sung Lee, and Kwang S. Kim. Band Gap Narrowing of Zinc Orthogermanate by Dimensional and Defect Modification. *Journal of Physical Chemistry C*, 123(23):14573–14581, 6 2019.

[198] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *The Journal of Physical Chemistry Letters*, 9(7):1668–1673, 4 2018.

[199] Vladislav Gladkikh, Dong Yeon Kim, Amir Hajibabaei, Atanu Jana, Chang Woo Myung, and Kwang S. Kim. Machine Learning for Predicting the Band Gaps of ABX3 Perovskites from Elemental Properties. *Journal of Physical Chemistry C*, 124(16):8905–8918, 4 2020.

[200] Vishwesh Venkatraman. The utility of composition-based machine learn-

ing models for band gap prediction. *Computational Materials Science*, 197:110637, 9 2021.

[201] Kenji Homma, Yu Liu, Masato Sumita, Ryo Tamura, Naoki Fushimi, Junichi Iwata, Koji Tsuda, and Chioko Kaneta. Optimization of a Heterogeneous Ternary Li3PO4–Li3BO3–Li2SO4 Mixture for Li-Ion Conductivity by Machine Learning. *The Journal of Physical Chemistry C*, 124(24):12865–12870, 6 2020.

[202] Austin D. Sendek, Ekin D. Cubuk, Evan R. Antoniuk, Gowoon Cheon, Yi Cui, and Evan J. Reed. Machine Learning-Assisted Discovery of Solid Li-Ion Conducting Materials. *Chemistry of Materials*, 31(2):342–352, 1 2019.

[203] Pikee Priya and N. R. Aluru. Accelerated design and discovery of perovskites with high conductivity for energy applications through machine learning. *npj Computational Materials 2021 7:1*, 7(1):1–12, 6 2021.

[204] Yijie Xu, Yun Zong, and Kedar Hippalgaonkar. Machine learning-assisted cross-domain prediction of ionic conductivity in sodium and lithium-based superionic conductors using facile descriptors. *Journal of Physics Communications*, 4(5):055015, 5 2020.

[205] Hongshuai Wang, Yujin Ji, and Youyong Li. Simulation and design of energy materials accelerated by machine learning. *WIREs Computational Molecular Science*, 10(1), 1 2020.

[206] Randy Jalem, Takahiro Aoyama, Masanobu Nakayama, and Masayuki Nogami. Multivariate method-assisted ab initio study of olivine-type LiMXO 4 (main group M2+-X5+ and M3+-X 4+) compositions as potential solid electrolytes. *Chemistry of Materials*, 24(7):1357–1364, 4 2012.

[207] Randy Jalem, Kenta Kanamori, Ichiro Takeuchi, Masanobu Nakayama, Hisatsugu Yamasaki, and Toshiya Saito. Bayesian-Driven First-Principles Calculations for Accelerating Exploration of Fast Ion Conductors for Rechargeable Battery Application. *Scientific Reports*, 8(1):1–10, 12 2018.

[208] Masanobu Nakayama, Kenta Kanamori, Koki Nakano, Randy Jalem, Ichiro Takeuchi, and Hisatsugu Yamasaki. Data-Driven Materials Exploration for Li-Ion Conductive Ceramics by Exhaustive and Informatics-Aided Computations. *The Chemical Record*, 19(4):771–778, 4 2019.

[209] Dipendra Jha, Logan Ward, Arindam Paul, Wei keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Scientific Reports*, 8(1):1–13, 12 2018.

[210] Tian Xie and Jeffrey C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120(14):145301, 4 2018.

[211] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials*, 31(9):3564–3572, 5 2019.

[212] Maarten de Jong, Wei Chen, Thomas Angsten, Anubhav Jain, Randy Notestine, Anthony Gamst, Marcel Sluiter, Chaitanya Krishna Ande, Sybrand van der Zwaag, Jose J Plata, Cormac Toher, Stefano Curtarolo, Gerbrand Ceder, Kristin A. Persson, and Mark Asta. Charting the complete elastic properties of inorganic crystalline compounds. *Scientific Data 2015 2:1*, 2(1):1–13, 3 2015.

[213] Zeeshan Ahmad, Tian Xie, Chinmay Maheshwari, Jeffrey C. Grossman, and Venkatasubramanian Viswanathan. Machine Learning Enabled Computational Screening of Inorganic Solid Electrolytes for Suppression of Dendrite Formation in Lithium Metal Anodes. *ACS Central Science*, 4(8):996–1006, 8 2018.

[214] Danny E. P. Vanpoucke, Onno S. J. van Knippenberg, Ko Hermans, Katrien V. Bernaerts, and Siamak Mehrkanoon. Small data materials design

with machine learning: When the average model knows best. *Journal of Applied Physics*, 128(5):054901, 8 2020.

[215] David H. Wolpert. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7):1341–1390, 10 1996.

[216] Fang Ren, Logan Ward, Travis Williams, Kevin J. Laws, Christopher Wolverton, Jason Hattrick-Simpers, and Apurva Mehta. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Science Advances*, 4(4), 2018.

[217] Maximilian Amsler, Logan Ward, Vinay I. Hegde, Maarten G. Goesten, Xia Yi, and Chris Wolverton. Ternary mixed-anion semiconductors with tunable band gaps from machine-learning and crystal structure prediction. *Physical Review Materials*, 3(3), 3 2019.

[218] Ruoqian Liu, Abhishek Kumar, Zhengzhang Chen, Ankit Agrawal, Veera Sundararaghavan, and Alok Choudhary. A predictive machine learning approach for microstructure optimization and materials design. *Scientific Reports*, 5(1):1–12, 6 2015.

[219] Geoffroy Hautier, Christopher C. Fischer, Anubhav Jain, Tim Mueller, and Gerbrand Ceder. Finding natures missing ternary oxide compounds using machine learning and density functional theory. *Chemistry of Materials*, 22(12):3762–3767, 6 2010.

[220] Mariette Hellenbrandt. The inorganic crystal structure database (ICSD) - Present and future. In *Crystallography Reviews*, volume 10, pages 17–22. Mariette Hellenbrandt, 1 2004.

[221] Harry E. Pence and Antony Williams. Chemspider: An online chemical information resource, 11 2010.

[222] Antanas Vaitkus, Andrius Merkys, and Saulius Gražulis. Validation of

the Crystallography Open Database using the Crystallographic Information Framework. *Journal of Applied Crystallography*, 54(2):54, 4 2021.

[223] Jordan O'Mara, Bryce Meredig, and Kyle Michel. Materials Data Infrastructure: A Case Study of the Citrination Platform to Examine Data Import, Storage, and Access. *JOM*, 68(8):2031–2034, 8 2016.

[224] Anubhav Jain, Joseph Montoya, Shyam Dwaraknath, Nils E. R. Zimmermann, John Dagdelen, Matthew Horton, Patrick Huck, Donny Winston, Shreyas Cholia, Shyue Ping Ong, and Kristin Persson. The Materials Project: Accelerating Materials Design Through Theory-Driven Data and Tools. In *Handbook of Materials Modeling*, pages 1–34. Springer International Publishing, 2018.

[225] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM*, 65(11):1501–1509, 11 2013.

[226] Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H. Taylor, Lance J. Nelson, Gus L.W. Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, Natalio Mingo, and Ohad Levy. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science*, 58:227–235, 6 2012.

[227] Paul Raccuglia, Katherine C. Elbert, Philip D.F. Adler, Casey Falk, Malia B. Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A. Friedler, Joshua Schrier, and Alexander J. Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73–76, 5 2016.

[228] Yihang Li, Bin Xiao, Yuchao Tang, Fu Liu, Xiaomeng Wang, Feinan Yan, and Yi Liu. Center-Environment Feature Model for Machine Learning Study

of Spinel Oxides Based on First-Principles Computations. *Journal of Physical Chemistry C*, 124(52):28458–28468, 12 2020.

[229] Roman A. Eremin, Pavel N. Zolotarev, Olga Yu Ivanshina, and Ivan A. Bobrikov. Li(Ni,Co,Al)O2 Cathode Delithiation: A Combination of Topological Analysis, Density Functional Theory, Neutron Diffraction, and Machine Learning Techniques. *Journal of Physical Chemistry C*, 121(51):28293–28305, 12 2017.

[230] M. Attarian Shandiz and R. Gauvin. Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries. *Computational Materials Science*, 117:270–278, 5 2016.

[231] Omar Allam, Byung Woo Cho, Ki Chul Kim, and Seung Soon Jang. Application of DFT-based machine learning for developing molecular electrode materials in Li-ion batteries. *RSC Advances*, 8(69):39414–39420, 11 2018.

[232] Rajendra P. Joshi, Jesse Eickholt, Liling Li, Marco Fornari, Veronica Barone, and Juan E. Peralta. Machine Learning the Voltage of Electrode Materials in Metal-Ion Batteries. *ACS Applied Materials and Interfaces*, 11(20):18494–18503, 2019.

[233] Gregory Houchins and Venkatasubramanian Viswanathan. An accurate machine-learning calculator for optimization of Li-ion battery cathodes. *Journal of Chemical Physics*, 153(5):054124, 8 2020.

[234] Natalia Kireeva and Vladislav S. Pervov. Materials Informatics Screening of Li-Rich Layered Oxide Cathode Materials with Enhanced Characteristics Using Synthesis Data. *Batteries & Supercaps*, 3(5):427–438, 5 2020.

[235] Cleber F.N. Marchiori, Daniel Brandell, and C. Moyses Araujo. Predicting Structure and Electrochemistry of Dilithium Thiophene-2,5-Dicarboxylate Electrodes by Density Functional Theory and Evolutionary Algorithms. *Journal of Physical Chemistry C*, 123(8):4691–4700, 2 2019.

[236] Yue Liu, Biru Guo, Xinxin Zou, Yajie Li, and Siqi Shi. Machine learning assisted materials design and discovery for rechargeable batteries, 10 2020.

[237] Randy Jalem, Masanobu Nakayama, and Toshihiro Kasuga. An efficient rule-based screening approach for discovering fast lithium ion conductors using density functional theory and artificial neural networks. *Journal of Materials Chemistry A*, 2(3):720–734, 1 2014.

[238] Reza Eslamloueyan, Mohammad H. Khademi, and Saeed Mazinani. Using a multilayer perceptron network for thermal conductivity prediction of aqueous electrolyte solutions. *Industrial and Engineering Chemistry Research*, 50(7):4050–4056, 4 2011.

[239] Thirumalai Parthiban, R. Ravi, and N. Kalaiselvi. Exploration of artificial neural network [ANN] to predict the electrochemical characteristics of lithium-ion cells. *Electrochimica Acta*, 53(4):1877–1882, 12 2007.

[240] Randy Jalem, Masanobu Nakayama, and Toshihiro Kasuga. An efficient rule-based screening approach for discovering fast lithium ion conductors using density functional theory and artificial neural networks. *Journal of Materials Chemistry A*, 2(3):720–734, 12 2013.

[241] Ali Zeinolabedini Hezave, Sona Raeissi, and Mostafa Lashkarbolooki. Estimation of Thermal Conductivity of Ionic Liquids Using a Perceptron Neural Network. *Industrial and Engineering Chemistry Research*, 51(29):9886–9893, 7 2012.

[242] Jonathan Schmidt, Jingming Shi, Pedro Borlido, Liming Chen, Silvana Botti, and Miguel A.L. Marques. Predicting the Thermodynamic Stability of Solids Combining Density Functional Theory and Machine Learning. *Chemistry of Materials*, 29(12):5090–5103, 6 2017.

[243] Koji Fujimura, Atsuto Seko, Yukinori Koyama, Akihide Kuwabara, Ippei Kishida, Kazuki Shitara, Craig A. J. Fisher, Hiroki Moriwake, and Isao

Tanaka. Accelerated Materials Design of Lithium Superionic Conductors Based on First-Principles Calculations and Machine Learning Algorithms. *Advanced Energy Materials*, 3(8):980–985, 8 2013.

[244] Minh Van Duong, Man Van Tran, Akhil Garg, Hoang Van Nguyen, Tuyen Thi Kim Huynh, and My Loan Phung Le. Machine learning approach in exploring the electrolyte additives effect on cycling performance of LiNi0.5Mn1.5O4 cathode and graphite anode-based lithium-ion cell. *International Journal of Energy Research*, 45(3):4133–4144, 3 2021.

[245] Atsushi Ishikawa, Keitaro Sodeyama, Yasuhiko Igarashi, Tomofumi Nakayama, Yoshitaka Tateyama, and Masato Okada. Machine Learning Prediction of Coordination Energies for Alkali Group Elements in Battery Electrolyte Solvents. *Physical Chemistry Chemical Physics*, 2019.

[246] Amalia Christina Wagner, Nicole Bohn, Holger Geßwein, Matthias Neumann, Markus Osenberg, André Hilger, Ingo Manke, Volker Schmidt, and Joachim R. Binder. Hierarchical Structuring of NMC111-Cathode Materials in Lithium-Ion Batteries: An In-Depth Study on the Influence of Primary and Secondary Particle Sizes on Electrochemical Performance. *ACS Applied Energy Materials*, 3(12):12565–12574, 12 2020.

[247] Marianna Hietaniemi, Tao Hu, Juho Välikangas, Janne Niittykoski, and Ulla Lassi. Effect of precursor particle size and morphology on lithiation of Ni0.6Mn0.2Co0.2(OH)2. *Journal of Applied Electrochemistry*, 51(11):1545–1557, 11 2021.

[248] Marcus Muller, Luca Schneider, Nicole Bohn, Joachim R. Binder, and Werner Bauer. Effect of Nanostructured and Open-Porous Particle Morphology on Electrode Processing and Electrochemical Performance of Li-Ion Batteries. *ACS Applied Energy Materials*, 4(2):1993–2003, 2 2021.

[249] David R Lide, Mary Eleanor Lide David Alston Lide, and Jr Grace Eileen

Lide David Austell Whitcomb Kate Elizabeth Whitcomb. CRC Handbook of Chemistry and Physics, 84th Edition, 2003-2004. 2003.

[250] R. D. Vocke. Atomic Weights of the Elements 1997. *Pure and Applied Chemistry*, 71(8):1593–1607, 1999.

[251] Martin Rahm, Roald Hoffmann, and N. W. Ashcroft. Atomic and Ionic Radii of Elements 1–96. *Chemistry – A European Journal*, 22(41):14625–14632, 10 2016.

[252] F Albert Cotton, Geoffrey Wilkinson, Carlos a. Murillo, and Manfred Bochmann. Advanced Inorganic Chemistry, 6th Edition. *Wiley-Interscience: New York*, pages 1–1376, 1999.

[253] † Jonathan C. Rienstra-Kiracofe, Gregory S. Tschumper, , Henry F. Schaefer III*, Sreela Nandi and, and G. Barney Ellison*. Atomic and Molecular Electron Affinities: Photoelectron Experiments and Theoretical Computations. *Chemical Reviews*, 102(1):231–282, 1 2002.

[254] WebPlotDigitizer - Extract data from plots, images, and maps.

[255] Chunhao Yang, Wuning Ma, Jianlin Zhong, and Zhendong Zhang. Comparative Study of Machine Learning Approaches for Predicting Creep Behavior of Polyurethane Elastomer. *Polymers 2021, Vol. 13, Page 1768*, 13(11):1768, 5 2021.

[256] Nooshin Saadatkhah, Seyedfoad Aghamiri, Mohammad Reza Talaie, and Gregory S. Patience. Flame-assisted spray pyrolysis to size-controlled LiyAlxMn2-xO4: A supervised machine learning approach. *CrystEngComm*, 20(46):7590–7601, 11 2018.

[257] Fleur Legrain, Jesús Carrete, Ambroise Van Roekeghem, Georg K.H. Madsen, and Natalio Mingo. Materials Screening for the Discovery of New Half-Heuslers: Machine Learning versus ab Initio Methods. 122(2):625–632, 1 2018.

[258] Erika Cule and Maria De Iorio. Ridge Regression in Prediction Problems: Automatic Choice of the Ridge Parameter. *Genetic Epidemiology*, 37(7):704–714, 11 2013.

[259] Sungwan Bang, Hyungjun Cho, and Myoungshic Jhun. Adaptive lasso penalised censored composite quantile regression. *International Journal of Data Mining and Bioinformatics*, 15(1):22–46, 4 2016.

[260] Jacek M Zurada and St Paul New York Los Angeles San Francisco. Introduction to Art if icial Neural Sy s. Technical report, 1992.

[261] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 12 2014.

[262] Harris Drucker·, Chris J C Burges, Linda Kaufman, Alex Smola··, and Vladimir Vapoik. Support Vector Regression Machines.

[263] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 3 1986.

[264] D. A. Freedman. Bootstrapping Regression Models. *The Annals of Statistics*, 9(6):1218–1228, 11 1981.

[265] Andy Liaw and Matthew Wiener. Classification and Regression by RandomForest. Technical report, 2002.

[266] Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378, 2 2002.

[267] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence 2020 2:1*, 2(1):56–67, 1 2020.

[268] Dastan Maulud and Adnan M. Abdulazeez. A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, 1(4):140–147, 12 2020.

[269] Max Kuhn and Kjell Johnson. Applied predictive modeling. *Applied Predictive Modeling*, pages 1–600, 1 2013.

[270] Amanpreet Singh, Narina Thakur, and Aakanksha Sharma. A review of supervised machine learning algorithms, 2016.

[271] Jack V. Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225–1231, 11 1996.

[272] Jiazhen Cai, Xuan Chu, Kun Xu, Hongbo Li, and Jing Wei. Machine learning-driven new material discovery, 8 2020.

[273] K. Uma Pavan Kumar, Ongole Gandhi, M. Venkata Reddy, and S. V. N. Srinivasu. Usage of KNN, Decision Tree and Random Forest Algorithms in Machine Learning and Performance Analysis with a Comparative Measure. pages 473–479, 2021.

[274] Arun Mannodi-Kanakkithodi, Ghanshyam Pilania, and Rampi Ramprasad. Critical assessment of regression-based machine learning methods for polymer dielectrics. *Computational Materials Science*, 125:123–135, 12 2016.

[275] Min Hsuan Lee. Insights from Machine Learning Techniques for Predicting the Efficiency of Fullerene Derivatives-Based Ternary Organic Solar Cells at Ternary Blend Design. *Advanced Energy Materials*, 9(26):1–9, 2019.

[276] Won Hee Ryu, Ji Yong Eom, Ri Zhu Yin, Dong Wook Han, Won Keun Kim, and Hyuk Sang Kwon. Synergistic effects of various morphologies and Al doping of spinel LiMn2O4 nanostructures on the electrochemical performance of lithium-rechargeable batteries. *Journal of Materials Chemistry*, 21(39):15337–15342, 9 2011.

[277] Z. D. Peng, Q. L. Jiang, K. Du, W. G. Wang, G. R. Hu, and Y. X. Liu. Effect of Cr-sources on performance of Li1.05Cr0.04Mn1.96O4 cathode materials prepared by slurry spray drying method. *Journal of Alloys and Compounds*, 493(1-2):640–644, 3 2010.

[278] Haiqiang Liu, Ruiyuan Tian, Yi Jiang, Xinghua Tan, Jiankun Chen, Lina Zhang, Yanjun Guo, Hanfu Wang, Lianfeng Sun, and Weiguo Chu. On the drastically improved performance of Fe - Doped LiMn2O4 nanoparticles prepared by a facile solution - Gelation route. *Electrochimica Acta*, 180:138–146, 10 2015.

[279] Pura Ram, A. Gören, S. Ferdov, M. M. Silva, Ganpat Choudhary, Rahul Singhal, C. M. Costa, Rakesh K. Sharma, and S. Lanceros-Méndez. Synthesis and improved electrochemical performance of LiMn 2 – x Gd x O 4 based cathodes. *Solid State Ionics*, 300:18–25, 2017.

[280] Su Cheol Han, Satendra Pal Singh, Yun-hwa Hwang, Eun Gyoung Bae, Byung Kyu Park, Kee-Sun Sohn, and Myoungho Pyo. Gadolinium-Doped LiMn 2 O 4 Cathodes in Li Ion Batteries: Understanding the Stabilized Structure and Enhanced Electrochemical Kinetics . *Journal of The Electrochemical Society*, 159(11):A1867–A1873, 2012.

[281] Rahul Singhal, Suprem R. Das, Maharaj S. Tomar, Osbert Ovideo, Santander Nieto, Ricardo E. Melgarejo, and Ram S. Katiyar. Synthesis and characterization of Nd doped LiMn2O4 cathode for Li-ion rechargeable batteries. *Journal of Power Sources*, 164(2):857–861, 2 2007.

[282] M. V. Reddy, S. Sundar Manoharan, Jimmy John, Brajendra Singh, G. V. Subba Rao, and B. V. R. Chowdari. Synthesis, Characterization, and Electrochemical Cycling Behavior of the Ru-Doped Spinel, Li[Mn[sub 2x]Ru[sub x]]O[sub 4] (x=0, 0.1, and 0.25). *Journal of The Electrochemical Society*, 156(8):A652, 2009.

[283] Yanting Xie, Yanbin Xu, Lan Yan, Zhenyin Yang, and Rudong Yang. Synthesis and electrochemical properties of Sc-doped Li 1.05ScxMn2-xO4 spinel as cathodic material for rechargeable Li-battery. *Solid State Ionics*, 176(35-36):2563–2569, 11 2005.

[284] Prakash Chand, Vivek Bansal, Sukriti, and Vishal Singh. Investigations of spinel LiZnxMn2xO4 (x0.03) cathode materials for a lithium ion battery application. *Materials Science and Engineering B: Solid-State Materials for Advanced Technology*, 238-239:93–99, 12 2018.

[285] Hongyuan Zhao, Xingquan Liu, Cai Cheng, Qiang Li, Zheng Zhang, Yue Wu, Bing Chen, and Weiqiang Xiong. Synthesis and electrochemical characterizations of spinel LiMn1.94MO4 (M = Mn0.06, Mg0.06, Si0.06, (Mg0.03Si0.03)) compounds as cathode materials for lithium-ion batteries. *Journal of Power Sources*, 282:118–128, 5 2015.

[286] Shaohua Guo, Shichao Zhang, Xiangming He, Weihua Pu, Changyin Jiang, and Chunrong Wan. Synthesis and Characterization of Sn-Doped LiMn[sub 2]O[sub 4] Cathode Materials for Rechargeable Li-Ion Batteries. *Journal of The Electrochemical Society*, 155(10):A760, 2008.

[287] Mingwu Xiang, Chang Wei Su, Lili Feng, Minglong Yuan, and Junming Guo. Rapid synthesis of high-cycling performance LiMgxMn 2-xO4 (x 0.20) cathode materials by a low-temperature solid-state combustion method. *Electrochimica Acta*, 125:524–529, 4 2014.

[288] Jing Xu, Feng Lin, Marca M. Doeff, and Wei Tong. A review of Ni-based layered oxides for rechargeable Li-ion batteries. *Journal of Materials Chemistry A*, 5(3):874–901, 1 2017.

[289] Florian Schipper, Evan M. Erickson, Christoph Erk, Ji-Yong Shin, Frederick Francois Chesneau, and Doron Aurbach. Review—Recent Advances and Remaining Challenges for Lithium Ion Battery Cathodes. *Journal of The Electrochemical Society*, 164(1):A6220–A6228, 12 2017.

[290] Zhengcheng Ye, Lang Qiu, Wen Yang, Zhenguo Wu, Yuxia Liu, Gongke Wang, Yang Song, Benhe Zhong, and Xiaodong Guo. Nickel-Rich Layered Cathode Materials for Lithium-Ion Batteries. *Chemistry – A European Journal*, 27(13):4249–4269, 3 2021.

[291] Yongseon Kim, Doyu Kim, and Shinhoo Kang. Experimental and first-principles thermodynamic study of the formation and effects of vacancies in layered lithium nickel cobalt oxides. *Chemistry of Materials*, 23(24):5388–5397, 12 2011.

[292] Kyung K. Lee, Won S. Yoon, Kwang B. Kim, Ki Y. Lee, and Seung T. Hong. Characterization of LiNi0.85Co0.10M0.05O2 (M = Al, Fe) as a cathode material for lithium secondary batteries. In *Journal of Power Sources*, volume 97-98, pages 308–312, 7 2001.

[293] Qina Sa, Joseph A. Heelan, Yuan Lu, Diran Apelian, and Yan Wang. Copper Impurity Effects on LiNi1/3Mn1/3Co1/3O2 Cathode Material. *ACS Applied Materials and Interfaces*, 7(37):20585–20590, 9 2015.

[294] Umair Nisar, Ruhul Amin, Abdul Shakoor, Rachid Essehli, Siham Al-Qaradawi, Ramazan Kahraman, and Ilias Belharouak. Synthesis and electrochemical characterization of Cr-doped lithium-rich Li1.2Ni0.16Mn0.56Co0.08-xCrxO2 cathodes. *Emergent Materials*, 1(3-4):155–164, 12 2018.

[295] Yucheng Sun, Yonggao Xia, and Hideyuki Noguchi. The improved physical and electrochemical performance of LiNi0.35Co0.3-xCrxMn0.35 O2 cathode materials by the Cr doping for lithium ion batteries. *Journal of Power Sources*, 159(2):1377–1382, 9 2006.

[296] Yanling Jin, Youlong Xu, Fang Ren, and Penggang Ren. 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*, 336:87–94, 8 2019.

[297] Ortal Breuer, Arup Chakraborty, Jing Liu, Tatyana Kravchuk, Larisa Burstein, Judith Grinblat, Yaron Kauffman, Alexandr Gladkih, Prasant Nayak, Merav Tsubery, Anatoly I. Frenkel, Michael Talianker, Dan T. Major, Boris Markovsky, and Doron Aurbach. Understanding the Role of Minor Molybdenum Doping in LiNi ¡sub¿0.5¡/sub¿ Co ¡sub¿0.2¡/sub¿ Mn ¡sub¿0.3¡/sub¿ O ¡sub¿2¡/sub¿ Electrodes: from Structural and Surface Analyses and Theoretical Modeling to Practical Electrochemical Cells. *ACS Applied Materials & Interfaces*, 10(35):29608–29621, 9 2018.

[298] Zuguang Yang, Xiaodong Guo, Wei Xiang, Weibo Hua, Jun Zhang, Fengrong He, Kai Wang, Yao Xiao, and Benhe Zhong. K-doped layered LiNi0.5Co0.2Mn0.3O2 cathode material: Towards the superior rate capability and cycling performance. *Journal of Alloys and Compounds*, 699:358–365, 3 2017.

[299] Xueqian Zhang, Yali Xiong, Mengfei Dong, and Zhiguo Hou. Pb-Doped Lithium-Rich Cathode Material for High Energy Density Lithium-Ion Full Batteries. *Journal of The Electrochemical Society*, 166(13):A2960–A2965, 8 2019.

[300] Isaac M. Markus, Feng Lin, Kinson C. Kam, Mark Asta, and Marca M. Doeff. Computational and experimental investigation of Ti substitution in Li1(NixMnxCo1-2x-yTiy)O2 for lithium ion batteries. *Journal of Physical Chemistry Letters*, 5(21):3649–3655, 11 2014.

[301] Seong Hwan Na, Hyun Soo Kim, and Seong In Moon. The effect of Si doping on the electrochemical characteristics of LiNi xMnyCO(1-x-y)O2. *Solid State Ionics*, 176(3-4):313–317, 1 2005.

[302] Qi Qi Qiao, Lei Qin, Guo Ran Li, Yong Long Wang, and Xue Ping Gao. Sn-stabilized Li-rich layered Li(Li0.17Ni0.25Mn0.58)O2 oxide as a cathode for advanced lithium-ion batteries. *Journal of Materials Chemistry A*, 3(34):17627–17634, 7 2015.

[303] Michael F. Toney. Li gradients for Li-rich cathodes, 12 2019.

[304] Lingqun Xu, Zhenhe Sun, Yu Zhu, Yu Han, Manman Wu, Yanfeng Ma, Yi Huang, Hongtao Zhang, and Yongsheng Chen. A Li-rich layered-spinel cathode material for high capacity and high rate lithium-ion batteries fabricated via a gas-solid reaction. *Science China Materials*, 63(12):2435–2442, 12 2020.

[305] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv*, 2 2018.

[306] Yao Wu, Peter Keil, Simon F. Schuster, and Andreas Jossen. 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 . *Journal of The Electrochemical Society*, 164(7):A1438–A1445, 5 2017.

[307] Wangda Li, Hooman Yaghoobnejad Asl, Qiang Xie, and Arumugam Manthiram. Collapse of LiNi1- x- yCoxMnyO2 Lattice at Deep Charge Irrespective of Nickel Content in Lithium-Ion Batteries. *Journal of the American Chemical Society*, 141(13):5097–5101, 4 2019.

[308] Marca M. Doeff, James D. Wilcox, Robert Kostecki, and Grace Lau. Optimization of carbon coatings on LiFePO4. *Journal of Power Sources*, 163(1 SPEC. ISS.):180–184, 12 2006.

[309] Yunjian Liu, Xinhai Li, Huajun Guo, Zhixing Wang, Wenjie Peng, Yong Yang, and Rufu Liang. Effect of carbon nanotube on the electrochemical performance of C-LiFePO4/graphite battery. In *Journal of Power Sources*, volume 184, pages 522–526. Elsevier B.V., 10 2008.

[310] R. Muruganantham, M. Sivakumar, and R. Subadevi. Enhanced rate performance of multiwalled carbon nanotube encrusted olivine type composite cathode material using polyol technique. *Journal of Power Sources*, 300:496–506, 12 2015.

[311] Yan L. Cheah, Song Y. Choy, Tingji Toh, Subodh Mhaisalkar, and Madhavi Srinivasan. Olivine-carbon nanofibrous cathodes for lithium ion batteries. In *Materials Research Society Symposium Proceedings*, volume 1266, pages 19–24. Materials Research Society, 2 2010.

[312] L. Dimesso, C. Spanheimer, W. Jaegermann, Y. Zhang, and A. L. Yarin. LiFePO 4 - 3D carbon nanofiber composites as cathode materials for Li-ions batteries. *Journal of Applied Physics*, 111(6), 3 2012.

[313] Zhao yong Chen, Hua li Zhu, Shan Ji, Rushanah Fakir, and Vladimir Linkov. Influence of carbon sources on electrochemical performances of LiFePO4/C composites. *Solid State Ionics*, 179(27-32):1810–1815, 9 2008.

[314] Yaping Wang, Huanhuan Li, Min Chen, Xiaoliang Yang, and Deli Jiang. Synthesis and electrochemical performance of LiFePO4/C cathode materials from Fe2O3 for high-power lithium-ion batteries. *Ionics*, 23(2):377–384, 2 2017.

[315] Yulong Liu, Jian Liu, Jiajun Wang, Mohammad Norouzi Banis, Biwei Xiao, Andrew Lushington, Wei Xiao, Ruying Li, Tsun Kong Sham, Guoxian Liang, and Xueliang Sun. Formation of size-dependent and conductive phase on lithium iron phosphate during carbon coating. *Nature Communications*, 9(1):1–8, 12 2018.

[316] Bing Zhao, Yong Jiang, Haijiao Zhang, Haihua Tao, Mingyang Zhong, and Zheng Jiao. Morphology and electrical properties of carbon coated LiFePO4 cathode materials. *Journal of Power Sources*, 189(1):462–466, 4 2009.

[317] Xiaoning Tian, Weiheng Chen, Zhongqing Jiang, and Zhong Jie Jiang. Porous carbon-coated LiFePO4 nanocrystals prepared by in situ plasma-assisted pyrolysis as superior cathode materials for lithium ion batteries. *Ionics*, 26(6):2715–2726, 6 2020.

[318] Dung V. Trinh, Mai T. T. Nguyen, Hue T. M. Dang, Dung T. Dang, Hang T. T. Le, Huynh T. N. Le, Hoang V. Tran, and Chinh D. Huynh. Hydrother-

mally synthesized nanostructured LiMnxFe1xPO4 (x=0–0.3) cathode materials with enhanced properties for lithium-ion batteries. *Scientific Reports*, 11(1):12280, 12 2021.

[319] L. N. Wang, X. C. Zhan, Z. G. Zhang, and K. L. Zhang. A soft chemistry synthesis routine for LiFePO4-C using a novel carbon source. *Journal of Alloys and Compounds*, 456(1-2):461–465, 5 2008.

[320] Nan Zhou, Yanyi Liu, Jiangang Li, Evan Uchaker, Suqin Liu, Kelong Huang, and Guozhong Cao. Synthesis and characterization of high power LiFePO 4/C nano-plate thin films. *Journal of Power Sources*, 213:100–105, 9 2012.

[321] P. Subramanya Herle, B. Ellis, N. Coombs, and L. F. Nazar. Nano-network electronic conduction in iron and nickel olivine phosphates. *Nature Materials*, 3(3):147–152, 2 2004.

[322] Craig A.J. Fisher, Veluz M.Hart Prieto, and M. Saiful Islam. Lithium battery materials LiMPO4 (M = Mn, Fe, Co, and Ni): Insights into defect association, transport mechanisms, and doping behavior. *Chemistry of Materials*, 20(18):5907–5915, 9 2008.

[323] C. M. Julien, A. Mauger, K. Zaghib, R. Veillette, and H. Groult. Structural and electronic properties of the LiNiPO 4 orthophosphate. *Ionics*, 18(7):625–633, 7 2012.

[324] A. Ait Salah, P. Jozwiak, J. Garbarczyk, K. Benkhouja, K. Zaghib, F. Gendron, and C. M. Julien. Local structure and redox energies of lithium phosphates with olivine- and Nasicon-like structures. *Journal of Power Sources*, 140(2):370–375, 2 2005.

[325] Gioele Pagot, Marco Bandiera, Keti Vezzù, Andrea Migliori, Renzo Bertoncello, Enrico Negro, Vittorio Morandi, and Vito Di Noto. High valence transition metal-doped olivine cathodes for superior energy and fast cycling lithium batteries. *Journal of Materials Chemistry A*, 8(48):25727–25738, 12 2020.

[326] Oriele Palumbo, Jessica Manzi, Daniele Meggiolaro, Francesco M. Vitucci, Francesco Trequattrini, Mariangela Curcio, Annalisa Paolone, and Sergio Brutti. Effect of transitional metals (Mn and Ni) substitution in LiCoPO4 olivines. *Molecules*, 25(3), 1 2020.

[327] Yaqing Zhang, Hongyan Xie, Huixin Jin, Xiaohui Li, Qiang Zhang, Yezhu Li, Kai Feng Li, Fenglan Luo, Wenlei Li, and Chenzhe Li. Enhancing the Electrochemical Properties of Ti-Doped LiMn2O4 Spinel Cathode Materials Using a One-Step Hydrothermal Method. *ACS Omega*, 6(33):21304–21315, 8 2021.

[328] H Sahan, H Göktepe, and S Patat. Cycling behaviour of barium doped LiMn 2 O 4 cathode materials for Li ion secondary batteries. *Materials Science-Poland*, 28(4), 2010.

[329] Reihane Etefagh, · Seyed, Mohammad Rozati, · Ahmad, Ahmadi Daryakenari, · Fatemeh Poursalehi, · Majid, Karami Keshmarzi, and * Seyed. Enhanced Li-storage performance of In-doped Li1.21[Mn0.54Ni0.125Co0.125]O2 as Li- and Mn-rich cathode materials for lithium-ion batteries. *Journal of Applied Electrochemistry 2021*, 1:1–15, 1 2022.

[330] Ahmet Örnek, Emrah Bulut, and Mustafa Can. Influence of gradual cobalt substitution on lithium nickel phosphate nano-scale composites for high voltage applications. *Materials Characterization*, 106:152–162, 6 2015.

[331] Ahmet Örnek, Mustafa Can, and Ali Yeşildağ. Improving the cycle stability of LiCoPO4 nanocomposites as 4.8 V cathode: Stepwise or synchronous surface coating and Mn substitution. *Materials Characterization*, 116:76–83, 6 2016.

[332] Jing Mao, Kehua Dai, Minjie Xuan, Guosheng Shao, Ruimin Qiao, Wanli Yang, Vincent S. Battaglia, and Gao Liu. Effect of Chromium and Niobium Doping on the Morphology and Electrochemical Performance of High-

Voltage Spinel LiNi0.5Mn1.5O4 Cathode Material. *ACS Applied Materials and Interfaces*, 8(14):9116–9124, 4 2016.

[333] Wang Y, Wang Y, Hosono E, Wang K, and Zhou H. The design of a LiFePO4/carbon nanocomposite with a core-shell structure and its synthesis by an in situ polymerization restriction method. *Angewandte Chemie (International ed. in English)*, 47(39):7461–7465, 9 2008.

[334] Mu Rong Yang, Wei Hsin Ke, and She Huang Wu. Preparation of LiFePO4 powders by co-precipitation. *Journal of Power Sources*, 146(1-2):539–543, 8 2005.

[335] Weijia Zhou, Wen He, Zhengmao Li, Hongshi Zhao, and Shunpu Yan. Biosynthesis and electrochemical characteristics of LiFePO4/C by microwave processing. *Journal of Solid State Electrochemistry 2008 13:12*, 13(12):1819–1823, 12 2008.

[336] Xuedong Yan, Guiling Yang, Jing Liu, Yucui Ge, Haiming Xie, Xiumei Pan, and Rongshun Wang. An effective and simple way to synthesize LiFePO4/C composite. *Electrochimica Acta*, 54(24):5770–5774, 10 2009.

[337] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 8 1987.

[338] Alan Julian Izenman. Linear Discriminant Analysis. pages 237–280, 2013.

[339] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of Neural Networks using DropConnect, 5 2013.

[340] Ismoilov Nusrat and Sung-Bong Jang. A Comparison of Regularization Techniques in Deep Neural Networks. *Symmetry 2018, Vol. 10, Page 648*, 10(11):648, 11 2018.

[341] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks, 6 2020.

[342] Xing Xing Wu and Jin Guo Liu. A new early stopping algorithm for improving neural network generalization. *2009 2nd International Conference on Intelligent Computing Technology and Automation, ICICTA 2009*, 1:15–18, 2009.

[343] David Duvenaud, Dougal Maclaurin, and Ryan Adams. Early Stopping as Nonparametric Variational Inference, 5 2016.

[344] Xinhui Zhang, Dong Luo, Guangshe Li, Jing Zheng, Chuang Yu, Xiangfeng Guan, Chaochao Fu, Xiangdong Huang, and Liping Li. Self-adjusted oxygen-partial-pressure approach to the improved electrochemical performance of electrode Li[Li0.14Mn0.47Ni0.25Co0.14]O2 for lithium-ion batteries. *Journal of Materials Chemistry A*, 1(34):9721–9729, 8 2013.

[345] Jatinkumar Rana, Richard Kloepsch, Jie Li, Tobias Scherb, Gerhard Schumacher, Martin Winter, and John Banhart. On the structural integrity and electrochemical activity of a 0.5Li2MnO3·0.5LiCoO2 cathode material for lithium-ion batteries. *Journal of Materials Chemistry A*, 2(24):9099–9110, 5 2014.

[346] Haijun Yu, Yeong-Gi So, Akihide Kuwabara, Eita Tochigi, Naoya Shibata, Tetsuichi Kudo, Haoshen Zhou, and Yuichi Ikuhara. Crystalline Grain Interior Configuration Affects Lithium Migration Kinetics in Li-Rich Layered Oxide. *Nano Letters*, 16(5):2907–2915, 5 2016.

[347] Alpesh Khushalchand Shukla, Quentin M. Ramasse, Colin Ophus, Hugues Duncan, Fredrik Hage, and Guoying Chen. Unravelling structural ambiguities in lithium- and manganese-rich transition metal oxides. *Nature Communications 2015 6:1*, 6(1):1–9, 10 2015.