Machine Learning Driven High-Throughput Discovery of Battery Electrodes and Electrolytes

Ph.D. Thesis

by

SOUVIK MANNA

DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF TECHNOLOGY INDORE
OCTOBER 2025

Machine Learning Driven High-Throughput Discovery of Battery Electrodes and Electrolytes

A THESIS

Submitted in partial fulfilment of the

requirements for the award of the degree

of

DOCTOR OF PHILOSOPHY

by

SOUVIK MANNA

DEPARTMENT OF CHEMISTRY

INDIAN INSTITUTE OF TECHNOLOGY INDORE

OCTOBER 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Machine Learning Driven High-Throughput Discovery of Battery Electrodes and Electrolytes** in the partial fulfilment of the requirements for the award of the degree of **DOCTOR OF PHILOSOPHY** and submitted in the **DEPARTMENT OF CHEMISTRY, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from December 2020 to October 2025 under the supervision of **Prof. BISWARUP PATHAK**, Professor, Department of Chemistry, IIT Indore.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Souvik Manna 06/10/2025

Signature of the student with date
(SOUVIK MANNA)

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.

06/10/2025

Signature of Thesis Supervisor with date
(Prof. BISWARUP PATHAK)

SOUVIK MANNA has successfully given his Ph.D. Oral Examination held on 30/01/2026

30/01/2026

Signature of Thesis Supervisor with date
(Prof. BISWARUP PATHAK)

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Prof. Biswarup Pathak, for his unwavering support, mentorship, and guidance throughout my Ph.D. His encouragement to develop and implement my own research ideas has been instrumental in shaping me into an independent researcher. I am especially thankful for his patience during the early years of my Ph.D., when I struggled to publish, and for always motivating me to maintain a healthy work-life balance. His optimism, even during moments of academic frustration, helped me persist and grow. He has been not just a mentor but a source of steady inspiration over the last five years.

I'm also grateful to my PSPC members, Dr. Deepak Kumar Roy and Prof. Nirmala Menon, for their insightful feedback and constructive criticism. I thank all the faculty, staff, and technicians of the Department of Chemistry, IIT Indore, for their continuous support during coursework, departmental activities, and administrative procedures.

My sincere thanks to my collaborators Prof. Rahul Banerjee (IISER Kolkata), Prof. T. Pradeep (IIT Madras), and Prof. Tushar Kanti Mukherjee (IIT Indore) for the opportunity to explore new tools and work on interdisciplinary projects. I acknowledge IIT Indore and the Chemistry Department for providing lab and computational facilities. I'm also thankful to CSIR and PMRF India for fellowships and grants, and to ANRF for travel support.

I am deeply indebted to all past and present CMDG lab members. A heartfelt thanks to Dr. Akhil S. Nair and Dr. Shyama C. Mandal for their constant support and companionship. Dr. Akhil S. Nair's deep knowledge and approachable nature helped me develop confidence in my technical skills. He never hesitated to explain concepts multiple times until I truly understood, and for that, I remain grateful. I'd like to thank Dr. Diptendu Roy for introducing me to machine learning, and Dr. Sneha Mittal for involving me in DNA sequencing research. I am also thankful to Dr. Milan

Kumar Jena and Dr. Nishchal Bharadwaj for their friendly nature towards me. I want to sincerely thank Dr. Amitabha Das, Surya Sekhar Manna and Dr. Eti Mahal for their warmth, guidance, and unwavering support throughout my journey. They were more than mentors, treated me like family, offering professional advice and emotional strength when I needed it most.

I'll forever be grateful to Dr. Sandeep Das, whose mentorship shaped not just my research but my personal growth - always guiding me, encouraging me, and helping me navigate the early turbulence of Ph.D. life. He believed in me at times when I didn't believe in myself and always showed up when it mattered most. His presence gave me a sense of stability and reassurance that went beyond academics.

Special thanks to Ms. Poulami Paul for her persistent friendship and emotional support. Thanks to Ms. Priyanka Ghosh for her caring presence. To my childhood friends, Sontu and Dipta - thank you. Sontu has been my rock since school, and Dipta's help in machine learning proved invaluable in my early research years.

I am also thankful to Dr. Argha Chakraborty, Dr. Sayantan Sarkar, Dr. Bittu Mondal, and Dr. Harish Sahu for their crucial support during tough times.

My academic journey would be incomplete without acknowledging Serampore College and Presidency University for shaping my B.Sc. and M.Sc. years. I thank Prof. Dibyendu Mallick for introducing me to computational chemistry, and Prof. Pulak Kumar Ghosh for inspiring me to pursue research. I thank Serampore College and SC sir, DC sir, and SG ma'am for their foundational support.

Finally, I am forever grateful to my parents, Mr. Sambhu Manna and Mrs. Suvra Manna, and my sister, Sumana Manna, for their unconditional love and support throughout this journey.

SOUVIK MANNA

Dedicated to

Science

and

Humanities

SYNOPSIS

1. Introduction

As global energy demand rises, the shift from fossil fuels to renewables like solar, wind, and hydro has become essential.*[1]* However, the intermittent nature of these sources poses challenges for consistent energy supply.*[2]* Rechargeable batteries play a vital role in addressing this gap, enabling energy storage and controlled release.*[3]* Lithium-ion batteries (LIBs) currently dominate due to their high energy density and long cycle life, but concerns over lithium scarcity, safety, cost, and environmental impact are driving the search for alternative battery chemistries.*[4, 5]*

Potassium-ion (K-ion), sodium-ion (Na-ion), magnesium-ion (Mg-ion), calcium-ion (Ca-ion), and aluminum-ion (Al-ion) batteries have attracted significant attention as potential successors to LIBs, owing to the earth-abundance and favourable electrochemical properties of these elements.*[6]* Nevertheless, the successful deployment of such systems hinges on the rational design of battery components such as electrodes, electrolytes, and interfaces that are both chemically compatible and electrochemically stable. In this regard, a fundamental scientific bottleneck persists: the discovery and optimization of materials for diverse battery chemistries remain heavily reliant on resource-intensive trial-and-error experimental protocols and high-cost quantum mechanical simulations.*[7]*

To address this bottleneck, machine learning (ML) has emerged as a transformative tool in materials science, capable of accelerating the screening and design of functional materials through data-driven predictive modeling.*[8]* However, the application of ML in battery research is still in its early stages, particularly in relation to interpretability, generalization across chemistries, and integration with first-principles approaches. Key research gaps include (i) the lack of scalable, accurate models for predicting battery-relevant properties across diverse material families, (ii) the

underutilization of ML in solvent screening and interfacial analysis, and (iii) the absence of unified frameworks capable of navigating the vast compositional and structural spaces encountered in energy storage systems. This thesis aims to bridge these gaps through the development and application of supervised, unsupervised, and physics-informed machine learning approaches, specifically tailored for materials discovery in metal-ion and Al-S batteries. The work spans a range of key battery challenges like from predicting the specific capacity of bulk electrode materials and simulating voltage profiles of two-dimensional (2D) systems, to modeling metal–solvent interactions and mapping electrochemical stability windows of electrolytes. Additionally, the role of surface chemistry in stabilizing sulfur intermediates in aluminum–sulfur (Al–S) batteries is examined via ML-guided screening of MXene materials.

Each chapter in this thesis demonstrates how targeted ML methodologies e.g., kernel ridge regression, graph neural networks, gradient boosting, and interpretable frameworks like Shapash can drastically reduce computational overhead while maintaining high predictive accuracy.*[9, 10]* Moreover, by integrating domain knowledge with algorithmic modeling, this work moves beyond black-box predictions toward interpretable, scalable tools that guide both theoretical and experimental efforts.

In summary, this thesis presents a unified, data-centric strategy to tackle the multi-dimensional challenges of battery material discovery. It leverages the synergy between quantum chemistry and machine learning to create efficient pipelines for screening, predicting, and understanding battery materials. The insights and models developed here are expected to contribute to the rational design of metal-ion and Al-S batteries with improved performance, cost-effectiveness, and environmental compatibility, ultimately supporting the global pursuit of sustainable energy storage technologies.

2. Objectives

The core objective of the thesis is machine learning (ML) assisted high throughput screening of electrode and electrolyte and physical insight of ML predicted result for metal-ion and Al-S batteries. Specifically, the objectives of the thesis are as follows:

i) To propose optimize ML model for the prediction of specific capacity of unknown electrode materials from Materials Project database for metal-ion batteries and importance physical attributes affecting the specific capacity.
ii) Development of automated pipeline with the help of universal ML potential for the prediction of stepwise voltage profile of 2D electrode materials.
iii) Investigation of solvent-electrolyte interaction in voltage determination and interpretable ML model to interpret ML predicted results.
iv) Supervised learning-based screening of solvent-electrolyte based on the electrochemical window (ECW) and unsupervised learning-based clustering of large solvent-electrolyte space.
v) Identification of suitable anchoring cathode materials for aluminium-sulphur battery from a large space of 2D MXene materials with the help of machine learning guided screening.

3. Summary of the research work

The contents of each chapter included in the thesis are discussed as follows:

3.1. Introduction (Chapter 1)

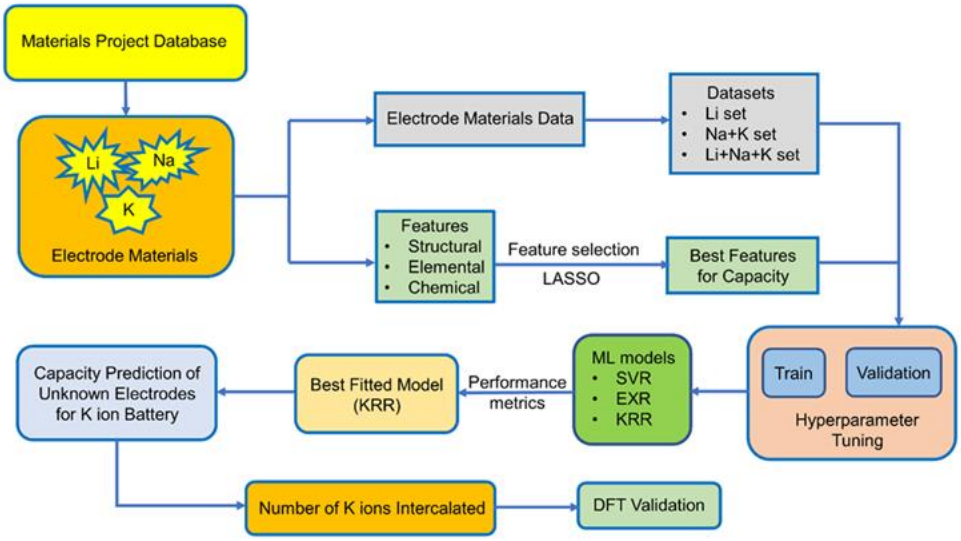
In this chapter, we have briefly discussed the developing field of metal-ion batteries (MIBs) and metal-sulphur batteries giving more emphasis on its working mechanism. An elaborate discussion is presented about the various part of MIBs such as cathode, anode, solvent-electrolyte and their role in improving electrochemical performance of MIBs. A brief discussion on the application of ML methodology to solve the battery related problem has also been provided. In addition, we have also reviewed the recent advancements using DFT and ML in the field of batteries.

Our thesis work involves the density functional theory (DFT), and predictive machine learning (ML) model and ML potential based model to investigate for the high throughput screening of electrode and electrolyte screening for the applicability in meta-ion and Al-S batteries. Therefore, this chapter also includes a brief discussion of these concepts and their importance in theoretical understanding of batteries. This chapter also covers the computational techniques which are used to explain the results of the computation.

3.2. Specific Capacity Prediction of Cathode Materials for Li/Na/K ion Batteries (Chapter 2)

In this chapter, a machine learning (ML) based methodology is developed to predict the specific capacity of electrode materials for potassium-ion (K-ion) batteries **(Scheme 1)**. Various ML models were evaluated using structural features and compositional descriptors derived from elemental properties. Among the tested models, Kernel Ridge Regression exhibited the highest predictive accuracy. The analysis utilized data from the Materials Project database, incorporating Li-, Na-, and K-ion battery materials with negative formation energies, ensuring thermodynamic stability. Feature engineering was performed to extract meaningful patterns, and statistical techniques such as box plots, joint plots, and correlation heatmaps were used to understand data distribution and feature interdependence. The trained model was then applied to predict the specific capacities of unexplored K-ion electrode materials. From these predictions, the number of K ions that could be intercalated per formula unit was estimated. Selected predictions were validated through density functional theory (DFT) calculations to confirm the structural stability and intercalation feasibility of the electrode compounds. The findings demonstrate that ML models can effectively accelerate the high-throughput screening of electrode materials, offering a computationally efficient

alternative to exhaustive DFT-based methods for identifying high-capacity candidates for next-generation K-ion battery technologies.

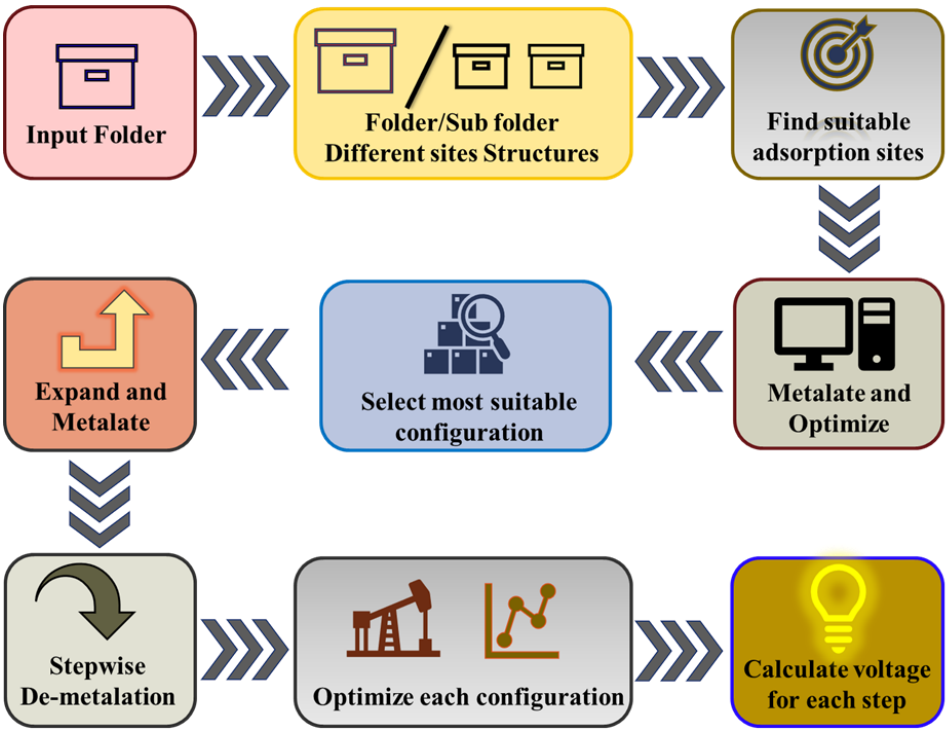


Scheme 1: Illustration of the systematic steps followed in the present work.

3.3. Automated Pipeline for High throughput Screening of Electrode Materials (Chapter 3)

In this chapter, a fully automated high-throughput computational framework is introduced to identify viable two-dimensional (2D) MT_2-type electrode materials for Li, Na, and K-ion battery applications **(Scheme 2)**. The workflow employs CHGNet (Crystal Hamiltonian Graph Neural Network), a graph neural network-based machine learning potential pretrained and fine-tuned to achieve near-density functional theory (DFT) accuracy in predicting total energies and structural relaxations. Starting solely from the unit cell structure, the pipeline performs a sequence of tasks including the identification of favorable adsorption sites, iterative simulation of ion intercalation and deintercalation, voltage calculation, and monitoring of structural evolution during each stage. The methodology allows the accurate prediction of voltage profiles and energetic stability for a diverse range of materials, capturing both geometric and energetic changes upon ion insertion. The role of metal and terminal group present in

the 2D MT_2 type material has explained on voltage determination. Finally, out of 289 MT_2 candidates screened, the pipeline identifies 46, 39, and 16 potential electrode materials for Li^+, Na^+, and K^+ ion batteries, respectively based on the stepwise voltage change and voltage difference of consecutive intercalation or deintercalation steps. The inclusion of CHGNet enables direct insights into chemical reactivity and structural responses, going beyond traditional ML-based screening techniques. This approach significantly reduces the computational overhead associated with conventional first-principles methods and provides a robust, chemistry-aware framework for accelerating the discovery of 2D electrode materials, thereby supporting more targeted experimental validation efforts.



Scheme 2: Automated pipeline for the determination of voltage profile of electrode materials.

3.4. Role of Metal-solvent interaction on voltage in Metal Ion Battery (Chapter 4)

This chapter presents a machine learning (ML) based framework for investigating the role of metal-solvent interaction energies in determining the anodic half-cell voltage in metal-ion batteries (MIBs). A dataset comprising 1584 metal-solvent systems including six commonly used metals (Li, Na, Mg, Al, K, Ca) and 66 different solvents was used to train and evaluate various ML algorithms. Among them, Gradient Boosting Regression (GBR) was identified as the most accurate model, yielding a root mean square error of 0.489 eV and a mean absolute error of 0.326 eV for interaction energy prediction. The study systematically evaluates how increasing solvent coordination around the metal center influences the interaction energy and, in turn, the resulting voltage. Anodic half-cell voltages were calculated using the GBR-predicted interaction energies for each metal-solvent combination, considering a graphite anode as the reference. An inverse correlation between interaction energy and voltage was observed, where weaker metal-solvent interactions generally led to lower anodic voltages **(Figure 1)**. Based on this analysis, five optimal solvents were identified for each metal, offering the most favorable voltage characteristics. In addition to predictive modeling, interpretability of the ML predictions was ensured using the Shapash library, enabling both global and local feature importance analysis. This interpretable approach supports reliable solvent screening for metal-ion batteries and demonstrates the potential of ML in guiding experimental efforts by reducing the reliance on computationally expensive methods such as density functional theory (DFT).

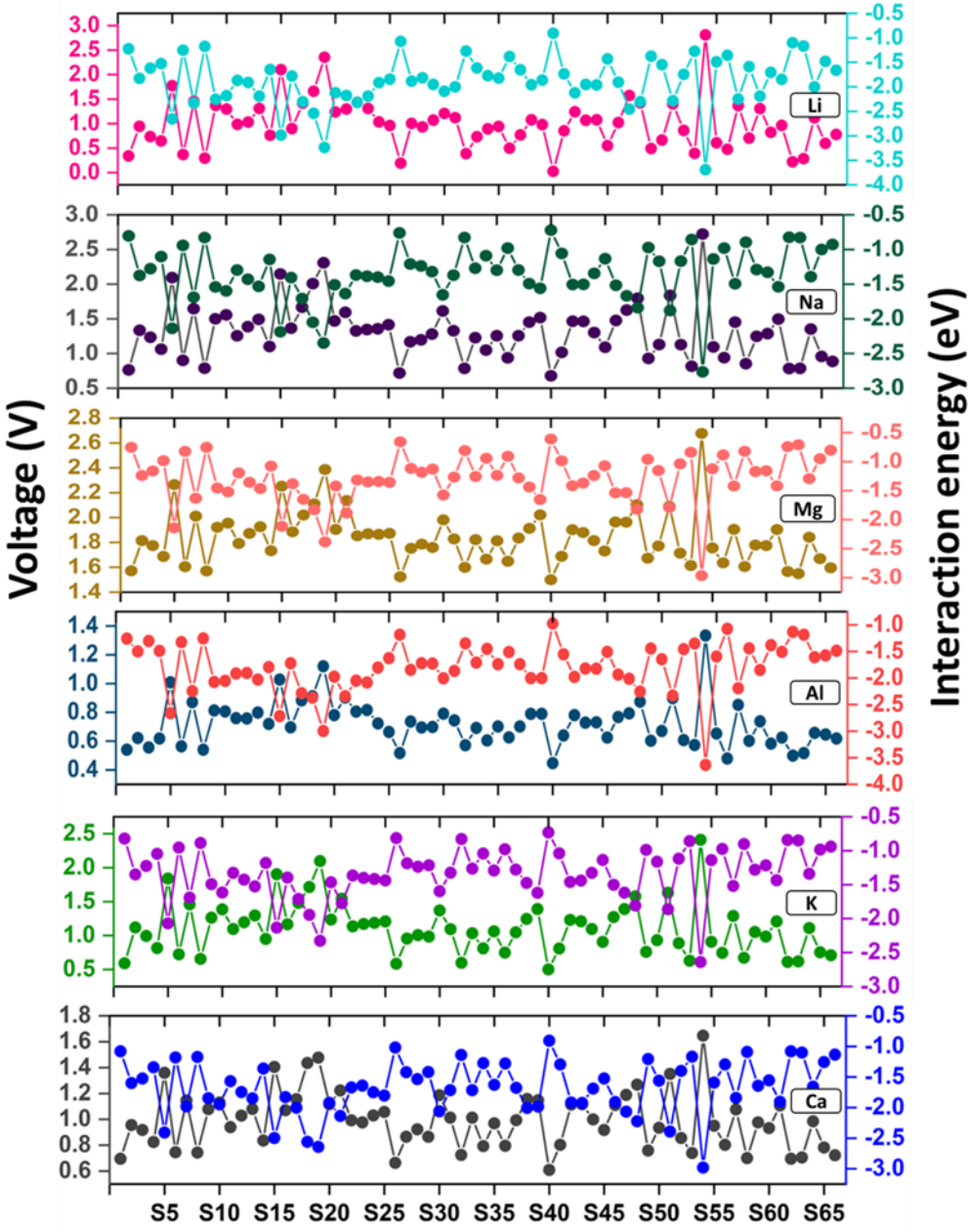


Figure 1: Average voltage and interaction energy plot for the combination of all considered six metals and 66 solvents.

3.5. Screening and Clustering of High ECW Solvent Electrolytes for Battery Applications (Chapter 5)

In this chapter, an integrated machine learning (ML) approach is developed to systematically identify and analyze solvent electrolytes suitable for high-voltage rechargeable metal-ion batteries. The study combines supervised

learning techniques with unsupervised clustering to predict the electrochemical windows (ECWs) comprising oxidation and reduction potentials of 4882 solvent molecules. The optimized Extreme Gradient Boosting Regression (XGBR) model, selected through cross-validation, feature selection (ANOVA-F value and Select-K-Best), and hyperparameter tuning, demonstrated high predictive accuracy, surpassing the performance of previous DFT-based methods. To further refine the vast solvent space, K-means clustering was applied, resulting in 11 distinct clusters characterized by a range of ECW distributions and functional group frequencies **(Figure 2)**. These clusters were further categorized into three boundary classes representing solvents with low, moderate, and high ECWs. Notably, solvents containing oxygen and fluorine functional groups predominantly contributed to high ECW clusters, highlighting the chemical trends underlying optimal electrochemical behaviour. The combined supervised-unsupervised approach provides a computationally efficient method for screening and organizing large-scale solvent datasets, facilitating the targeted selection of candidates for experimental validation. This framework significantly reduces the time and resources typically required for solvent discovery, enabling more focused development of electrolyte systems compatible with advanced electrode materials in battery technologies.

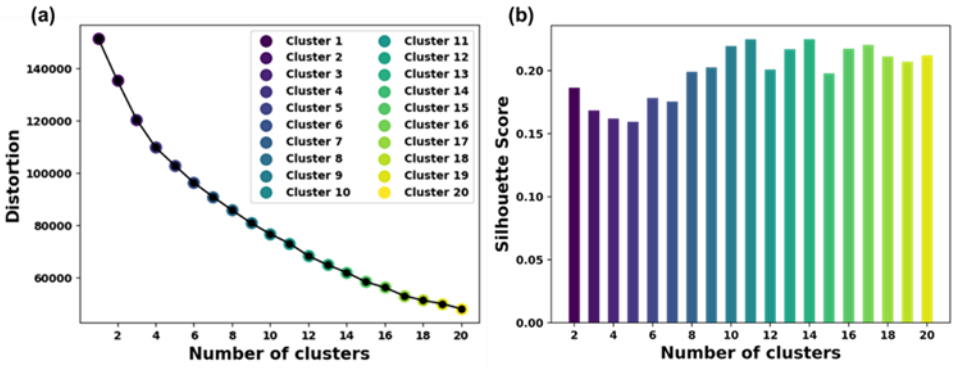


Figure 2: Optimization of number of clusters where (a) elbow curve, and (b) bar plot of Silhoutte score. We have considered the maximum number of clusters as 20.

3.6. Screening of MXene with Superior Anchoring Effect in Al–S Batteries (Chapter 6)

In this chapter, a combined density functional theory (DFT) and machine learning (ML) approach is presented for identifying MXene materials capable of effectively anchoring polysulfide intermediates in aluminum–sulfur (Al–S) batteries. The dissolution of intermediate species into the electrolyte remains a critical challenge for Al–S battery development. A high-throughput screening of $M_1M_2XT_2$ type MXenes was conducted to predict their adsorption strength toward various Al–S polysulfide species. A set of boosting tree-based ML models was developed and evaluated, with the Extreme Gradient Boosting Regression (XGBR) model demonstrating the highest accuracy. Feature selection using ANOVA F-value analysis improved model performance by prioritizing the most relevant descriptors. Based on the ML predictions, 42 MXene candidates were identified as having optimal anchoring behavior, characterized by adsorption energies that balance the need for sufficient binding without leading to irreversible interactions **(Figure 3)**. Analysis revealed that the nature of surface terminal groups plays a pivotal role, with MXenes functionalized by oxygen and fluorine groups exhibiting superior anchoring effects across multiple intermediate species. The developed ML framework offers a computationally efficient approach to down-select materials from a vast candidate space, enabling accelerated design and experimental validation of advanced sulfur host materials for Al–S batteries.

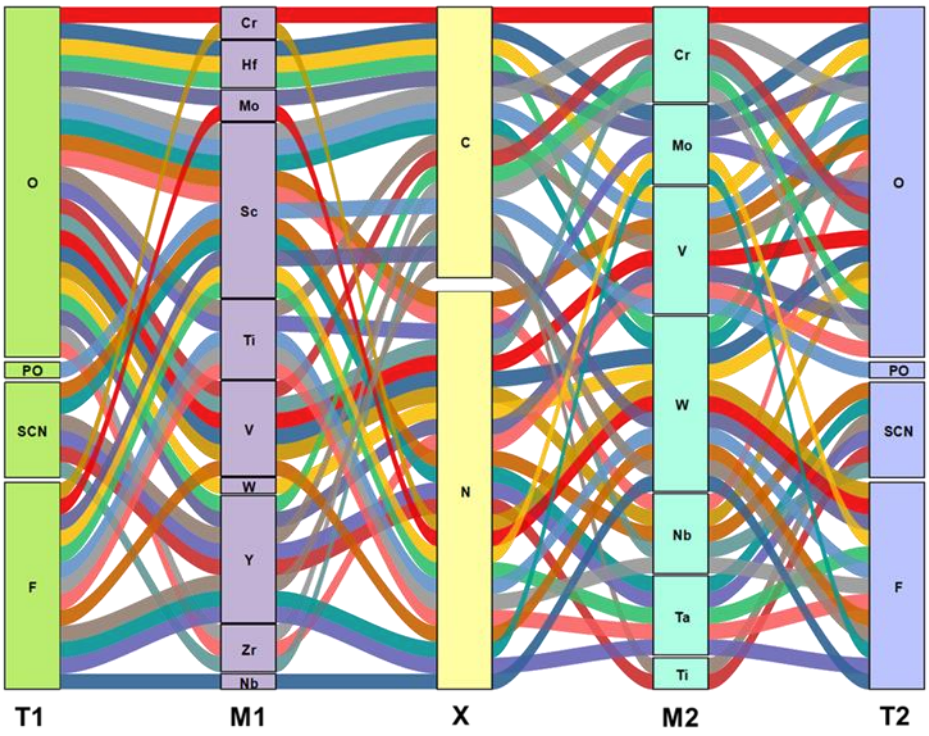


Figure 3: Alluvial plot showing the combinations of T1, M1, X, M2, and T2 leading to MXenes with optimum adsorption for all of the polysulfides.

4. Conclusions

This thesis presents a comprehensive exploration of machine learning (ML)-driven methodologies to accelerate the discovery and optimization of materials for metal-ion and Al-S battery applications. By leveraging both supervised and unsupervised learning strategies, we have tackled critical challenges in predicting electrode properties, solvent compatibility, and interfacial behavior, thereby advancing the rational design of next-generation battery components. The key outcomes from Chapters 2 to 6 are summarized below, with an emphasis on their integration within the broader context of energy storage and data-driven materials design:

1. Beginning with the prediction of specific capacity, we demonstrated how ML models, particularly Kernel Ridge Regression, can effectively map compositional and structural features to performance metrics. This model enabled rapid pre-screening of electrode materials, significantly

reducing the need for exhaustive density functional theory (DFT) calculations and setting the stage for high-throughput discovery workflows.

2. To expand the search for efficient battery electrodes, a fully automated pipeline integrating graph neural network (CHGNet) was constructed to evaluate the voltage profiles of two-dimensional (2D) MT_2-type materials for Li, Na, and K-ion batteries. This pipeline enabled high-throughput analysis of ion intercalation behavior, ultimately identifying 46, 39, and 16 promising candidates for Li^+, Na^+, and K^+ batteries, respectively, by starting from only the unit cell structure.

3. Extending the ML paradigm to the electrolyte domain, we investigated how metal-solvent interaction energies influence the anodic half-cell voltage. Using gradient boosting regression, we uncovered an inverse relationship between interaction strength and voltage and identified optimal solvent systems for six different metal ions. The interpretability layer, enabled by Shapash, provided mechanistic insights into how molecular features affect electrochemical performance.

4. To address the vast chemical design space of electrolytes, an integrated supervised-unsupervised learning framework was developed to map the electrochemical windows (ECWs) of nearly 5000 solvents. The optimized XGBR model predicted redox potentials with high accuracy, while K-means clustering efficiently segmented the solvent landscape into functional groups based on ECW characteristics. This dual approach enabled targeted solvent selection aligned with the voltage requirements of high-energy battery systems.

5. Finally, the ML-DFT hybrid strategy was extended to identify suitable sulfur host materials for aluminum–sulfur (Al–S) batteries. A total of 42 $M_1M_2XT_2$-type MXenes were identified with optimal anchoring properties, balancing intermediate adsorption to suppress the polysulfide shuttle effect. The dominant role of terminal functional

groups—especially oxygen and fluorine—in governing adsorption behaviour underscores the predictive power of ML in tuning material interfaces.

References

1. World Energy Outlook 2023 – Analysis – IEA. (n.d.) World Energy Outlook 2023 – Analysis – IEA. (accessed 2025-07-25)
2. Creutzig F., Agoston P., Goldschmidt J. C., Luderer G., Nemet G., Pietzcker R. C. (2017), The underestimated potential of solar energy to mitigate climate change, Nat. Energy, 2 (9), 1–9 (DOI: 10.1038/nenergy.2017.140)
3. Nykvist B., Nilsson M. (2015), Rapidly falling costs of battery packs for electric vehicles, Nat. Clim. Chang., 5 (4), 329–332 (DOI: 10.1038/nclimate2564)
4. Goodenough J. B., Park K. S. (2013), The Li-ion rechargeable battery: a perspective, J. Am. Chem. Soc., 135 (4), 1167–1176 (DOI: 10.1021/ja3091438)
5. Tarascon J. M., Armand M. (2001), Issues and challenges facing rechargeable lithium batteries, Nature, 414 (6861), 359–367 (DOI: 10.1038/35104644)
6. Kim H., Kim J. C., Bianchini M., Seo D. H., Rodriguez-Garcia J., Ceder G. (2018), Recent progress and perspective in electrode materials for K-ion batteries, Adv. Energy Mater., 8 (9), 1702384 (DOI: 10.1002/aenm.201702384)
7. Jain A., Ong S. P., Hautier G., Chen W., Richards W. D., Dacek S., Cholia S., Gunter D., Skinner D., Ceder G., Persson K. A. (2013), Commentary: The Materials Project: a materials genome approach to accelerating materials innovation, APL Mater., 1 (1), 011002 (DOI: 10.1063/1.4812323)

8. Butler K. T., Davies D. W., Cartwright H., Isayev O., Walsh A. (2018), Machine learning for molecular and materials science, Nature, 559 (7715), 547–555 (DOI: 10.1038/s41586-018-0337-2)
9. Jordan M. I., Mitchell T. M. (2015), Machine learning: trends, perspectives, and prospects, Science, 349 (6245), 255–260 (DOI: 10.1126/science.aaa8415)
10. Deng B., Zhong P., Jun K. J., Riebesell J., Han K., Bartel C. J., Ceder G. (2023), CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling, Nat. Mach. Intell., 5 (9), 1031–1041 (DOI: 10.1038/s42256-023-00716-3)

LIST OF PUBLICATIONS

1) Manna S., Roy D., **Das S.**, Pathak B. (2022), Capacity prediction of K-ion batteries: a machine learning based approach for high throughput screening of electrode materials, Mater. Adv., 3, 7833-7845. (DOI: 10.1039/D2MA00746K) (Impact Factor: **5.5)**
2) Manna S., Manna S. S., **Das S.**, Pathak. B. (2023) Metal-Solvent Interaction Contribution on Voltage for Metal Ion Battery: An Interpretable Machine Learning Approach, Electrochim. Acta, 467, 143148. (DOI: 10.1016/j.electacta.2023.143148) (Impact Factor: 6.6)
3) Manna S., Das A., Das S., Pathak B. (2024) Machine learning assisted screening of MXene with superior anchoring effect in Al–S batteries, ACS Mater. Lett., 6 (2), 572–582 (Impact Factor: 8.7)
4) Manna S., Manna S. S., Pathak B. (2024) Integrated Supervised and Unsupervised Machine Learning Approach to Map the Electrochemical Windows Over 4500 Solvents for Battery Applications, ACS Appl. Mater. Interfaces, 16, 42138–42152 (Impact Factor: 8.2)
5) Manna S., Paul P., Manna S. S., Das S., Pathak B. (2025) Utilizing machine learning to advance battery materials design: challenges and prospects, Chem. Mater., 37, 1759–1787 (Impact Factor: 6.97)
6) Singh S., Datta S., Manna S., Pathak B., Mukherjee T. K. (2025) Unraveling the hidden pathway of catalyst-free direct photochemical conversion of sulfides to sulfoxides: a universal pathway under UVA radiation, J. Phys. Chem. Lett., 16, 6106–6115 (Impact Factor: 4.6)
7) Das A., Roy D., Manna S., Pathak B. (2024) Harnessing the potential of machine learning to optimize the activity of Cu-based dual atom catalysts for CO_2 reduction reaction, ACS Mater. Lett., 6, 5316–5324 (Impact Factor: 8.7)
8) Karak S., Singh H., Biswas A., Paul S., Manna S., Nishiyama Y., Pathak B., Banerjee A., Banerjee R. (2024) Lithiophilic dibenzamide linkages

to impart lithium storage capacity in porous polybenzamides, J. Am. Chem. Soc., 146, 20183–20192 (Impact Factor: 15.6)

9) Paul P., Das S., Manna S., Manna S. S., Pathak B. (2024) Integration of Density Functional Theory and Machine Learning for Electrolyte Optimization in High-Voltage Dual-Ion Battery Design, ACS Appl. Mater. Interfaces, 16, 33, 43591–43601 (Impact Factor: 8.2)
10) Das S., Manna S., Pathak B. (2023) Unlocking the potential of dual-ion batteries: identifying polycyclic aromatic hydrocarbon cathodes and intercalating salt combinations through machine learning, ACS Appl. Mater. Interfaces, 15, 54520–54529 (Impact Factor: 8.2)
11) Manna S. S., Manna S., Pathak B. (2023) Molecular dynamics-machine learning approaches for the accurate predictions of electrochemical windows of ionic liquids electrolytes for dual-ion batteries, J. Mater. Chem. A, 11, 21702–21712 (Impact Factor: 9.5)
12) Mittal S., Manna S., Jena M. K., Pathak B. (2023) Artificial intelligence aided recognition and classification of DNA nucleotides using MoS_2 nanochannel, Digit. Discov., 2, 1589–1600 (Impact Factor: 5.66)
13) Mittal S., Manna S., Jena M. K. (2023) Decoding both DNA and methylated DNA using a MXene-based nanochannel device: supervised machine learning assisted exploration, ACS Mater. Lett., 5, 1570–1580 (Impact Factor: 8.7)
14) Roy D., Das A., Manna S., Pathak B. (2023) A route map of machine learning approaches in heterogeneous CO_2 reduction reaction, J. Phys. Chem. C, 127, 871–881 (Impact Factor: 3.2)
15) Mittal S., Manna S., Pathak B. (2022) Machine learning prediction of transmission function for protein sequencing with graphene nanoslit, ACS Appl. Mater. Interfaces, 14, 51645–51655 (Impact Factor: 8.2)

Table of Contents

List of Figures

Chapter 3

Chapter 5

List of Tables

Chapter 2

Acronyms

LIB	Lithium Ion Battery
NIB	Sodium Ion Battery
KIB	Potassium Ion Battery
MIB	Metal Ion Battery
MLIP	Machine Learning Interatomic Potential
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
DFT	Density functional theory
ML	Machine Learning
LDA	Local density approximation
GGA	Generalized Gradient Approximation
PBE	Perdew-Burke-Ernzerhof
PAW	Projector augmented wave
VASP	Vienna Ab initio Simulation Package
PDOS	Projected Density of States
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
SVR	Support Vector Regression
KRR	Kernel Ridge Regression
GBR	Gradient Boosting Regression
XGBR	eXtreme Gradient Boosting Regression
DTR	DecisionTreeRegressor
EXTR	ExtraTreesRegressor
GNN	Graph Neural Network
CHGNet	Crystal Hamiltonian Graph neural Network

Chapter 1

Introduction

1.1. Global Energy Challenge and the Need for Sustainable Storage

With the rapid rise in population, industrial growth, and technological expansion, global energy consumption has reached an unprecedented scale. Traditional fossil fuel sources, such as coal, oil, and natural gas, continue to meet the bulk of energy demand.*[1]* However, these sources are finite, environmentally damaging, and geopolitically complex. The transition to renewable energy technologies, such as solar, wind, hydro, and geothermal, is seen as an essential shift toward sustainable energy.*[2]* Yet, despite their promise, renewable energy sources suffer from spatial limitations and temporal intermittency. Solar and wind energy, for example, depend on time-of-day and weather conditions. Consequently, energy storage systems capable of storing surplus energy and releasing it on demand are essential for stabilizing the energy supply.*[3]*

Electrochemical energy storage devices, especially rechargeable batteries, have emerged as a vital technological solution in this context.*[4]* Batteries offer scalable, modular, and flexible energy storage systems that can be integrated across residential, industrial, and grid-level applications. These batteries function by converting electrical energy into chemical energy during charging and reversing the process during discharge. Rechargeable batteries are widely used across a range of applications, from personal electronics to large-scale grid storage, contributing significantly to modern convenience and technological advancement. Among the various battery technologies, lithium-ion batteries (LIBs) have gained prominence due to their high energy density, low weight, and strong electrochemical performance. LIBs operate through the reversible movement of lithium ions between electrodes, a mechanism often referred to as a "rocking chair" system **(Figure 1.1)**.*[5]* Despite their widespread use, LIBs face limitations in terms of resource availability and long-term sustainability.*[6]* This has led to growing interest in alternative chemistries, such as other metal-ion batteries (MIBs), which may offer more abundant materials and complementary performance characteristics.*[7]*

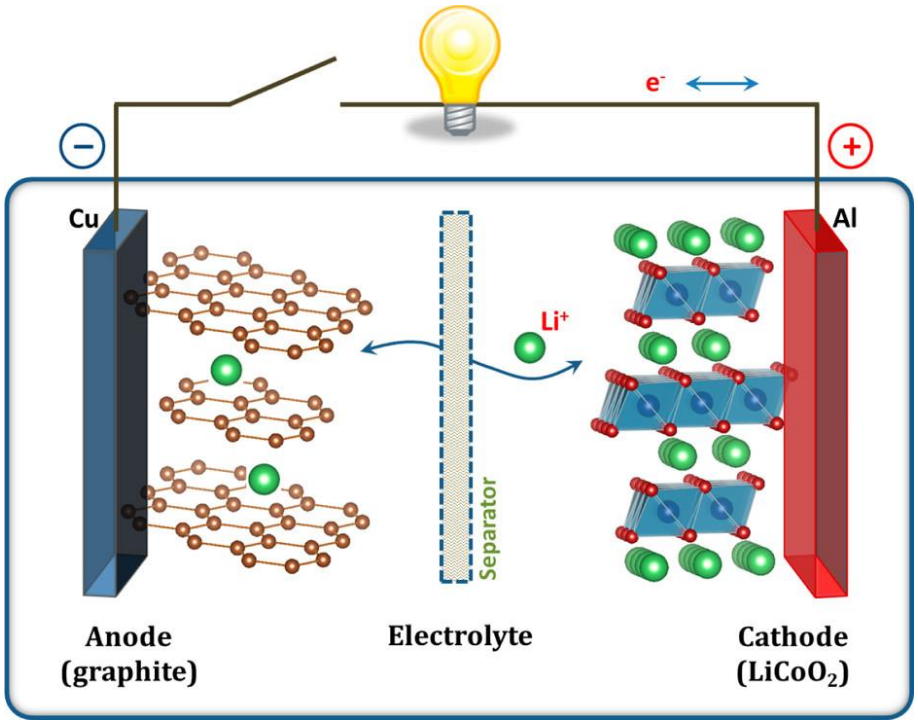


Figure 1.1: Schematic representation of a lithium-ion battery. Figure reprinted with permission from Ref. 4. Copyrights 2013, American Chemical Society.

1.2. From Lithium-Ion Batteries to Beyond

Lithium-ion batteries (LIBs) have revolutionized the energy storage market over the last three decades.*[4]* Their high energy density, long cycle life, and commercial availability have made them the dominant battery technology in smartphones, laptops, electric vehicles, and even grid storage.*[6]* LIBs operate on a "rocking chair" mechanism, where lithium ions shuttle between the cathode and anode during charge-discharge cycles.*[8]*

However, LIBs are not without shortcomings such as:*[9]*

• Limited abundance of lithium, leading to supply risks and rising costs*[10]*

• High environmental footprint associated with mining and disposal*[11]*

• Safety hazards such as dendrite formation, thermal runaway, and flammability*[12]*

• Scalability limitations for long-term grid-level deployment*[13]*

These challenges have sparked interest in post-Li battery chemistries, such as sodium-ion (Na-ion), potassium-ion (K-ion), magnesium-ion (Mg-ion), calcium-ion (Ca-ion), and aluminum-sulfur (Al–S) batteries. These systems offer the advantages of elemental abundance, low toxicity, and potentially higher capacities.*[14-18]*

1.2.1. Historical Background of Computational Materials Discovery in Batteries

Historically, the discovery of novel battery materials has followed a sequence of experimental and theoretical innovations. Initially, empirical studies guided the identification of new battery materials, particularly during the development of classical battery systems such as lead-acid, nickel-metal hydride, and LIBs.*[4]* With advancements in computational chemistry, Density Functional Theory (DFT) emerged as a dominant approach in the early 2000s, allowing researchers to simulate the structural and electronic behavior of candidate materials from first principles.*[19]* DFT provided insights into voltage profiles, diffusion barriers, electronic structure, and reaction mechanisms, thus reducing the cost and time of experimental exploration.

However, the DFT method, while highly accurate, suffers from several inherent limitations. First, it is computationally expensive, especially for systems with large unit cells, defects, or complex interfaces. Second, DFT simulations are typically limited to ground-state properties and small time/length scales, making them unsuitable for exploring kinetic phenomena such as long-term cycling behavior or electrolyte degradation. Third, exploring an entire chemical space comprising millions of possible materials through DFT is practically infeasible.

These limitations motivated the materials science community to seek alternatives that can complement or even replace DFT in certain applications. As a result, machine learning (ML) began gaining traction around 2015, driven largely by initiatives like the Materials Project, the Open Quantum Materials Database (OQMD), and the Novel Materials Discovery (NOMAD) project, which provided open-access databases of DFT-calculated properties.*[20]* Researchers began developing ML models trained on these databases to predict target properties such as formation energy, specific capacity, bandgap, and intercalation voltage.*[21]* Initially, these models relied on simple linear regressions or tree-based algorithms using handcrafted features (e.g., elemental statistics, coordination environments).*[22]*

More recently, there has been a transition to graph-based neural networks (GNNs) that operate directly on crystal structures, bypassing the need for manual feature engineering.*[23]* These models can approximate DFT-level accuracy with dramatically reduced computational requirements. The emergence of interpretable ML frameworks like SHAP and Shapash has also added transparency, making it possible to understand how specific atomic or molecular features influence predicted properties.*[24]* Moreover, generative models like variational autoencoders (VAEs) and diffusion models are beginning to play a transformative role by enabling the inverse design of novel materials that satisfy multiple design constraints.*[25-27]*

In essence, the historical trajectory from empirical discovery → DFT → data-driven ML → generative AI marks a profound paradigm shift in battery materials research.*[28]* This thesis positions itself at the frontier of this evolution by developing scalable, interpretable, and generative ML models to solve longstanding bottlenecks in battery design and electrolyte optimization.

1.2.2. Specific Capacity

Specific capacity is a core performance metric for evaluating the suitability of a material as a battery electrode.*[7]* It quantifies the amount of electric charge, in milliampere-hours (mAh), that can be stored and released per gram of active material. A high specific capacity directly translates to a longer operating time for battery-powered devices, making it crucial for both small-scale and large-scale energy storage applications.*[29]*

The theoretical specific capacity C, measured in mAh/g, is mathematically expressed as,

$$C = \frac{x \,.\, n \,.\, F}{M_f}$$

where x and n are the charge of the metal ion and number of metal-ion adsorbed in the formula unit, respectively. F is the Faraday constant, and M_f is the molecular mass of formula unit.

Traditional approaches to determining this value based on experimental electrochemical measurements or quantum mechanical simulations such as DFT are computationally and resource intensive, particularly when applied across a wide materials search space.

In recent years, the demand for low-cost, sustainable alternatives to lithium-ion batteries has prompted growing interest in potassium-ion (K-ion) batteries, due to the earth-abundance and low extraction cost of potassium.*[30]* However, the successful development of K-ion batteries depends heavily on identifying electrode materials that not only offer high specific capacity but also exhibit structural and chemical stability under repeated charge–discharge cycles. While DFT can simulate K^+ intercalation into candidate materials and thereby estimate capacity, the time and computational power required to do this across thousands of materials make the process prohibitively slow. In this context, machine learning (ML) provides an attractive alternative.*[31]* By learning the underlying structure–

property relationships from previously computed or experimentally measured data, ML models can predict specific capacities of new materials without requiring new quantum-level calculations.*[32]*

This chapter introduces a supervised ML-based framework aimed at predicting specific capacity for a large class of electrode materials, particularly with applications in K-ion batteries.*[33]* A dataset of electrode materials containing lithium, sodium, and potassium ions was curated from the Materials Project database, with the inclusion criteria focused on materials with negative formation energies to ensure thermodynamic viability.*[34]* Specific capacity labels were obtained either directly from the database or computed based on the number of intercalated ions per formula unit. A range of compositional descriptors such as average electronegativity, atomic radius, valence electron count, and atomic mass were extracted for each compound using Matminer and Pymatgen libraries.*[35]* These features are directly relevant to charge transport, ion intercalation, and redox activity, which collectively influence a material's capacity. Multiple regression models were trained and compared, including Kernel Ridge Regression (KRR), ExtraTrees Regression, and Support Vector Regression (SVR). Among them, KRR demonstrated superior performance in terms of R^2 score and Mean Absolute Percentage Error (MAPE), achieving both high accuracy and good generalization on the validation set.

1.2.3. 2D Materials and Stepwise Voltage Profile

The design of high-performance electrode materials for rechargeable batteries hinges on a careful balance between structural stability, ion mobility, and redox energetics. Among the central parameters that define battery efficiency, voltage is critical as it directly controls the energy output of a cell and strongly affects both power density and safety. Voltage in metal-ion batteries arises from the difference in electrochemical potentials

between the anode and cathode, and for a cathode host material undergoing ion insertion, the voltage is typically defined as:

$$V = -\frac{\Delta G}{nF}$$

where ΔG is the Gibbs free energy change per mole of reaction, n is the number of electrons transferred, and F is Faraday's constant. Furthermore, for stepwise ion intercalation (which often leads to multi-plateau voltage profiles), the problem becomes not only computationally demanding but also sensitive to structural relaxation accuracy and atomic-scale ion coordination environments.*[36]*

In recent years, two-dimensional (2D) materials have emerged as a promising class of electrode materials due to their high surface area, tunable interlayer spacing, and favorable ion diffusion characteristics.*[37]* The MT_2 (M = transition metal, T = chalcogen) family of layered compounds, in particular, has drawn increasing attention for metal-ion batteries including Li, Na, and K systems. These materials offer natural interlayer galleries for ion storage, good structural flexibility, and electrochemical resilience. Previous reports have demonstrated high reversible capacities and moderate-to-high operating voltages for MoS_2, TiS_2, and their analogues, making the broader MT_2 family a fertile ground for discovering next-generation cathode materials. However, given the combinatorial explosion of possible M–T combinations, heteroatom doping, and surface functionalization, exhaustive DFT calculations for voltage profiling are prohibitively expensive.

To address this challenge, the community has started turning to machine learning (ML) potentials, particularly those based on graph neural networks (GNNs), to predict atomic energies and forces with near-DFT accuracy but at a fraction of the computational cost. These models leverage local atomic environments and long-range bonding information to learn potential energy surfaces from large datasets of DFT-calculated structures.*[38]* This makes

them especially well-suited for predicting stepwise ion intercalation, capturing the voltage change across multiple compositional states in a single framework.

This chapter presents a physics-informed ML approach integrated with an automated pipeline for screening 2D battery electrode materials. The automated pipeline enables interpretable, scalable, and accurate voltage profiling of 2D electrode materials, advancing both the methodology and the materials space. It bridges the gap between fast but coarse empirical screening and slow but precise DFT studies, offering a third way that is fast, precise, and extensible.

1.3. Electrolyte Optimization

In electrochemical energy storage systems, the electrolyte plays a central role that extends far beyond simply conducting ions between electrodes. It governs interfacial stability, redox compatibility, solvation dynamics, and ultimately impacts key battery metrics like voltage, energy density, and cycle life. While considerable efforts have gone into optimizing electrode materials in metal-ion batteries (MIBs), a growing body of work has emphasized that electrolyte–metal interactions can significantly influence battery voltage and electrochemical reversibility. Despite this, the metal–solvent interaction energy remains an underexplored parameter in voltage design, especially across a diverse range of metals such as lithium (Li), sodium (Na), magnesium (Mg), potassium (K), calcium (Ca), and aluminium (Al).

Most prior computational work, rooted in DFT or ab initio molecular dynamics (AIMD), has focused either on the electrochemical window or decomposition pathways of solvents, leaving a blind spot in understanding how specific metal–solvent interaction strengths affect the thermodynamics of ion intercalation and stripping, which are directly tied to cell voltage. Moreover, while the concept of solvation energy has been studied in the context of diffusion and SEI formation, its systematic relation to anodic

voltage—particularly across multiple metals and solvent chemistries—has not been rigorously evaluated.

1.3.1. Metal-Solvent Interaction

In metal-ion batteries (MIBs), the role of the electrolyte extends well beyond being an inert medium for ion transport. The nature of the solvent molecules and their interaction with metal ions significantly influences ion solvation, desolvation kinetics, interfacial energetics, and ultimately, the overall electrochemical performance of the cell.*[39-41]* One of the most critical but historically underexamined aspects of electrolyte design is the metal–solvent interaction energy, which quantifies the thermodynamic strength of the solvation shell formed around the active metal ion.

Traditionally, studies have focused on solvation energies in the context of lithium-ion batteries, especially in relation to SEI formation, solvent decomposition pathways, and ion transport.*[42]* Seminal works using ab initio molecular dynamics and DFT have shown that solvents with high dielectric constants can stabilize Li^+ more effectively, leading to better ionic mobility and lower overpotentials. However, these analyses have generally remained limited to small data sets, often covering just a few solvents and a single metal ion—typically Li^+.*[43]* Moreover, these studies rarely go beyond solvation to explicitly connect the strength of metal–solvent interactions with the battery voltage, a parameter central to energy density and power output.

In this context, the thesis work represents a critical advancement by introducing a large-scale machine learning (ML) framework to understand and quantify the connection between metal–solvent interaction energy and anodic voltage. Specifically, evaluation of a total of 1584 systems, spanning six metals (Li, Na, K, Mg, Ca, Al) and 66 diverse solvents, with variations in dielectric constant, donor number, molecular geometry, and electronic structure.*[44]*Using these inputs, the ML pipeline centered around a Gradient Boosting Regressor (GBR) achieved a remarkably accurate

prediction of metal–solvent interaction energies. The regression model captured complex, non-linear relationships between solvent features (e.g., HOMO–LUMO gap, dipole moment, molecular volume, polarizability) and the thermodynamic stability of the solvated metal complexes.

A key insight from your study is the inverse relationship between metal–solvent interaction energy and anodic half-cell voltage, modeled using the graphite reference electrode. In electrochemical thermodynamics, this is consistent with the notion that weakly coordinated metal ions are more readily desolvated, thus requiring less energy (lower voltage) for stripping and intercalation.

1.3.2. Electro Chemical Window

The widely accepted concept, first introduced by Goodenough and Kim, proposes that the formation of the solid electrolyte interphase (SEI) occurs when the electrode redox potentials fall outside the electrochemical stability window of the electrolyte.*[45, 46]* This stability window is typically described in terms of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) energy levels of the solvent molecules. As depicted in Figure 1.2, When an electron's energy surpasses the LUMO level, the solvent is prone to reduction; conversely, electron energies below the HOMO level can result in solvent oxidation.

However, defining electrolyte stability solely through the HOMO-LUMO framework is an oversimplification. Real electrolytes are complex mixtures containing solvents, salts, and various additives that interact with electrode surfaces in ways that are highly dependent on molecular configuration and conformation.*[47-50]* Additionally, the HOMO and LUMO energy levels are typically derived from isolated molecules using approximate electronic structure methods. These values may not accurately reflect the behavior of molecules undergoing redox processes.

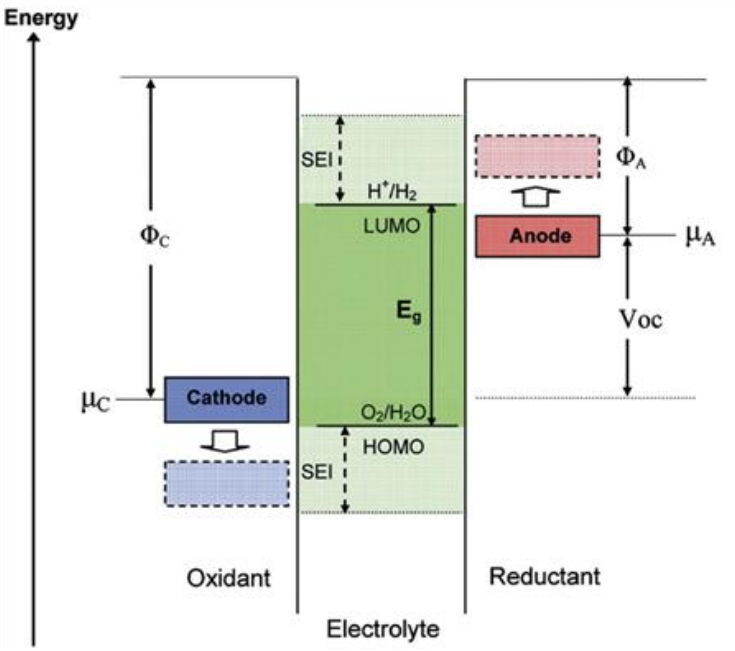


Figure 1.2: Energy diagram of electrolyte, anode and cathode in battery. Φ_A and Φ_C are the anode and cathode work functions. E_g is the electrochemical stability window of electrolyte. μ_A and μ_C are redox potential of anode and cathode, respectively. Figure reprinted with permission from Ref. 51. Copyrights 2018, Springer Nature.

In reality, redox potentials are more accurately determined by the difference in Gibbs free energy between the initial and final states of a redox reaction.*[51]* As such, relying solely on orbital energy levels can lead to inaccurate conclusions. A more reliable definition of electrochemical stability considers the potential at which electrolyte reduction occurs at the negative end, and the potential at which solvent oxidation takes place on the positive end of the electrochemical window.*[49]*

1.4. Metal-Sulphur Battery

Metal–sulphur batteries have emerged as a promising next-generation energy storage technology owing to their exceptionally high theoretical energy densities, cost-effectiveness, and abundant material resources.*[52-*

54] Among these, lithium–sulphur (Li–S) batteries have been the most extensively studied, demonstrating a theoretical energy density of ~2600 Wh/kg, significantly outperforming traditional lithium-ion batteries. However, the practical deployment of sulphur-based batteries faces several bottlenecks such as the dissolution and shuttle of polysulfide intermediates, poor conductivity of sulphur and its discharge products, and irreversible side reactions at the electrode–electrolyte interface.*[55]*

Expanding beyond Li–S systems, aluminium–sulphur (Al–S) batteries have recently garnered interest due to the natural abundance and low cost of aluminium, which also provides a high volumetric capacity (~8046 mAh/cm³), trivalent redox nature, and excellent environmental safety.*[56-58]* In a typical Al–S battery, aluminium serves as the anode, and elemental sulphur is used as the cathode. During discharge, sulphur is progressively reduced to polysulfide species, which can dissolve in the electrolyte and migrate to the aluminium anode resulting in capacity fading, low coulombic efficiency, and poor cycle life. Addressing this issue requires cathode host materials that can physically or chemically immobilize polysulfide intermediates and improve the electrochemical reversibility of the system.

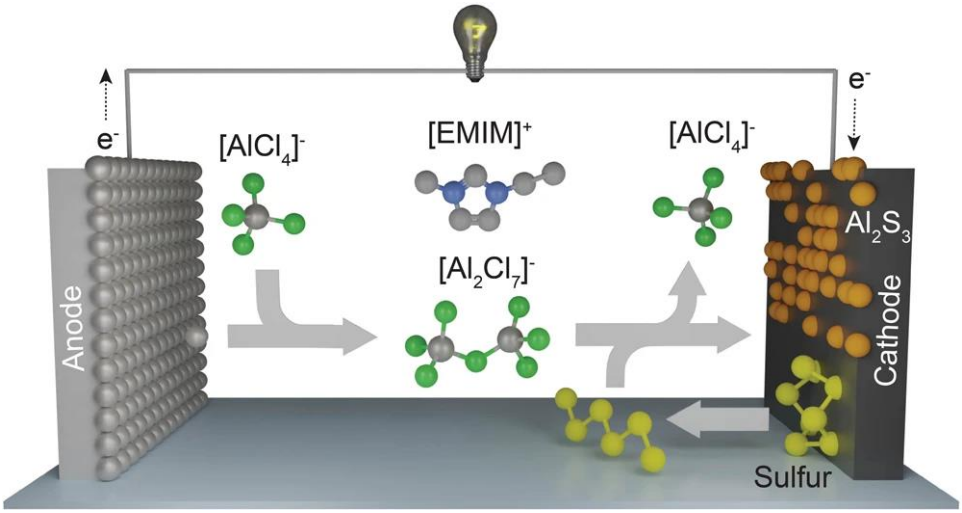


Figure 1.3: Working mechanism of aluminium-sulphur battery. Figure reprinted with permission from Ref. 56. Copyrights 2018, Springer Nature.

Over the past decade, researchers have explored various materials as cathode hosts, including carbon nanostructures, metal–organic frameworks (MOFs), and conductive polymers. However, many of these fail to provide adequate anchoring, leading to significant loss of active material. In this context, two-dimensional (2D) materials, especially MXenes, have emerged as one of the most promising classes of anchoring substrates due to their tunable surface chemistry, high conductivity, and rich chemistry derived from surface terminations.*[52, 59-60]*

1.4.1. 2D MXene Materials and Anchoring Effect

MXenes are a large family of 2D transition metal carbides and nitrides with a general formula $M_{n+1}X_nT_x$, where M is an early transition metal (e.g., Ti, Mo, V), X is carbon and/or nitrogen, and T_x represents surface terminations such as –O, –F, or –OH. Discovered in the early 2010s, MXenes have rapidly grown into a diverse family with hundreds of theoretically predicted and experimentally realized members.*[61, 62]* Their metallic conductivity, high mechanical strength, and versatile chemistry make them excellent candidates for applications in batteries, supercapacitors, and electrocatalysis.

The anchoring effect in the context of Al–S batteries refers to the ability of a host material to adsorb and immobilize polysulfide species, thereby preventing their dissolution into the electrolyte. This anchoring can be physical (via van der Waals interaction or pore trapping) or chemical (via bond formation or charge transfer).*[63, 64]* MXenes, owing to their high surface area and functional terminal groups, are especially effective in chemically anchoring intermediate species. For example, oxygen or fluorine terminated MXenes can form strong interactions with polysulfides, stabilizing the discharge products and suppressing the shuttle effect. Recent studies have demonstrated the ability of $Ti_3C_2T_x$ and Mo_2CT_x MXenes to adsorb Li_2S or Na_2S, and this behavior is now being extended to Al–S systems.

Despite these advances, exploring the vast MXene chemical space experimentally or even through DFT is computationally prohibitive. Many unexplored combinations of transition metals (M1, M2), carbonitrides (X), and terminal groups (T_x) may possess superior anchoring properties but remain undiscovered. Therefore, a high-throughput screening strategy assisted by machine learning is necessary to accelerate this discovery process.

A combined DFT and machine learning pipeline was implemented to screen $M1M2XT_2$-type MXenes for their anchoring capabilities in Al–S batteries.*[65]* Using adsorption energy as a target variable for anchoring strength, and features derived from atomic and structural descriptors, an XGBoost regression model was trained and validated. With growing interest in non-lithium batteries, this work stands at the intersection of computational chemistry, materials science, and data-driven design, offering a robust framework to identify and engineer next-generation cathode host materials.

1.5. Theory

This section outlines the foundational theoretical concepts and computational strategies utilized in this study.

1.5.1. Schrödinger Equation

To explore the electronic structure of materials and molecular systems, we rely on the time-independent form of the Schrödinger equation. This fundamental equation is given as:

$$\mathrm{H}\Psi(\mathrm{r},\mathrm{R}) = \mathrm{E}\Psi(\mathrm{r},\mathrm{R}) \qquad (1.1)$$

In this expression, H denotes the Hamiltonian operator, which encapsulates the total energy contributions of the system. The wavefunction $\Psi(r, R)$ encodes comprehensive information about the quantum state, including both electronic and nuclear components. The eigenvalue E represents the total energy associated with that state.

The Hamiltonian operator can be expressed using equation 1.2 as follows:

$$H = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 - \sum_I \frac{\hbar^2}{2M_I}\nabla_I^2 + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r_i}-\mathbf{r_j}|} + \frac{1}{2}\sum_{I\neq J}\frac{Z_I Z_J e^2}{|\mathbf{R_I}-\mathbf{R_J}|} - \sum_{i,I}\frac{Z_I e^2}{|\mathbf{r_i}-\mathbf{R_I}|}$$

(1.2)

In this formulation, the first term represents the kinetic energy of the electrons, where m_e is the mass of an electron, $\hbar$ is the reduced Planck's constant, and ∇_i^2 is the Laplacian operator acting on the position coordinates of electron i. The second term denotes the kinetic energy of the nuclei, with M_I being the mass of the I^{th} nucleus and ∇_I^2 the Laplacian operator acting on nuclear coordinates.

The third term accounts for the electron–electron repulsion, where e is the elementary charge, and $|r_i - r_j|$ is the distance between electrons i and j. The factor of $\frac{1}{2}$ ensures that each electron pair is counted only once. The fourth term represents nucleus–nucleus repulsion, where Z_I and Z_J denote the atomic numbers of nuclei I and J, respectively, and $|R_I - R_J|$ is the internuclear distance. Similarly, the pre-factor of $\frac{1}{2}$ avoids double-counting interactions between distinct nuclei. The final term corresponds to the electron–nucleus attraction, which arises due to the Coulombic attraction between negatively charged electrons and positively charged nuclei. This interaction is attractive in nature, and the negative sign reflects this in the total energy expression.

Altogether, Equation (1.2) provides a complete quantum mechanical description of the total energy of a system composed of interacting electrons and nuclei. However, solving this equation exactly for systems with more than a few particles is intractable, necessitating the use of approximations and computational methods such as the Born–Oppenheimer approximation and density functional theory (DFT).*[66]*

1.5.1.1. Born-Oppenheimer (BO) Approximation

The Born-Oppenheimer (BO) approximation simplifies the many-body Schrödinger equation by separating nuclear and electronic motion *[66]*. Since nuclei are approximately 1836 times more massive than electrons, their movement is much slower and can be treated as static during electronic calculations. This allows the omission of the nuclear kinetic energy term from the Hamiltonian, leading to a reduced form that focuses solely on the electronic structure. Thus, the Hamiltonian operator is as follows,

$$H = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r_i}-\mathbf{r_j}|} + \frac{1}{2}\sum_{I\neq J}\frac{Z_I Z_J e^2}{|\mathbf{R_I}-\mathbf{R_J}|} - \sum_{i,I}\frac{Z_I e^2}{|\mathbf{r_i}-\mathbf{R_I}|} \quad (1.3)$$

In systems containing only a single nucleus, the nucleus-nucleus repulsion term becomes irrelevant. The Hamiltonian then reduces to:

$$H = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r_i}-\mathbf{r_j}|} - \sum_{i,I}\frac{Z_I e^2}{|\mathbf{r_i}-\mathbf{R_I}|} \quad (1.4)$$

Despite the simplifications from the Born-Oppenheimer approximation, solving the Schrödinger equation remains computationally expensive. Methods like Hartree-Fock and, more efficiently, Density Functional Theory (DFT) are used to make these problems tractable. The next section focuses on DFT and its role in materials modelling.

1.5.2. Density Functional Theory (DFT)

Density Functional Theory (DFT) is a quantum mechanical method used to investigate the electronic structure of many-body systems, with electron density as its central variable. Unlike wavefunction-based approaches, which scale poorly with system size, DFT simplifies the problem by relying solely on electron density.

The foundations of DFT trace back to the Thomas-Fermi model, which described a gas of non-interacting electrons. However, modern DFT was formalized through the Hohenberg-Kohn theorems, which establish that all ground-state properties of a system are uniquely determined by its electron

density. These theorems provide the theoretical basis for expressing the energy of a system as a functional of the density, enabling practical and efficient electronic structure calculations.*[67-68]* The next sections provide an overview of the theoretical framework and key components of DFT.

1.5.2.1. The Hohenberg-Kohn Theorems

Theorem 1: The first Hohenberg-Kohn theorem states that the ground-state properties of a many-electron system are uniquely determined by its electron density $\rho(\mathbf{r})$ given an external potential $V_{ext}(\mathbf{r})$. In other words, there exists a one-to-one correspondence between the ground-state electron density $\rho_0(\mathbf{r})$ and the external potential, and thus the total ground-state energy is a unique functional of $\rho(\mathbf{r})$. This foundational result implies that all ground-state observables can, in principle, be derived from the electron density alone.

Theorem 2: The second Hohenberg-Kohn theorem introduces a universal energy functional $E[\rho(r)]$ which depends on the electron density $\rho(r)$ and includes the effects of the external potential $V_{ext}(r)$. It is expressed as:

$$E[\rho(r)] = E_{HK}[\rho(r)] + \int V_{ext}(r)\rho(r)dr \quad (1.5)$$

Here, $E_{HK}[\rho(\mathbf{r})]$ is the universal functional, encompassing both the kinetic energy and the electron-electron interactions. This theorem confirms that the ground-state energy of a many-electron system is minimized by the correct ground-state electron density. However, the particular form of the E_{HK} is unknown, necessitating approximations in practical DFT calculations.

1.5.2.2. Kohn-Sham Equations

Solving the full many-electron Schrödinger equation is computationally infeasible for most real-world systems. The Kohn-Sham (KS) approach, introduced by Walter Kohn and Lu Jeu Sham, provides a practical workaround by replacing the complex, interacting system of electrons with

a simplified one: a system of non-interacting electrons moving in an effective potential that still reproduces the correct ground-state electron density [94].

In this framework, the total energy of a system is written as a functional of the electron density $\rho(r)$:

$$E[\rho(r)] = T_0[\rho(r)] + \frac{1}{2}\iint \frac{\rho(r)\rho(r)'drdr'}{|r-r'|} + \int V_{ext}(r)\rho(r)dr + E_{xc}[\rho(r)dr] + E_{II} \quad (1.6)$$

where, the $T_0[\rho(r)]$, $\frac{1}{2}\iint \frac{\rho(r)\rho(r)'drdr'}{|r-r'|}$, $\int V_{ext}(r)\rho(r)dr$, and $E_{xc}[\rho(r)dr]$ desccribe the kinetic energy of the simple single non-interacting electron, the electron-electron Coulombic interaction, the potential energy of the valence and the core electrons and the exchange-correlation interaction respectively. The E_{II} term indicates the nuclei-nuclei interactions. However, the above Kohn-Sham equation can be further reduced as using equation 1.7,

$$\left[-\frac{1}{2}\nabla^2 + V_{eff}(r)\right]\Psi_i(r) = E_i\Psi_i(r) \quad (1.7)$$

Here, $\Psi_i(r)$ describe the Kohn-Sham orbitals, and V_{eff} is the summation of external potential (V_{ext}), Coulomb interaction ($V_{Hartree}$), and exchange correlation (V_{xc}).

$$V_{eff} = V_{Hartree} + V_{ext} + V_{xc} \quad (1.8)$$

In the Kohn-Sham formalism, the effective potential experienced by a non-interacting particle consists of three main components: the classical Coulomb force, the external field, and exchange-correlation interactions. However, accurately defining the exchange-correlation potential remains a significant challenge. To address this, a variety of approximations have been proposed and are now routinely applied in computational studies of both molecular and solid-state systems to estimate solutions to equations (1.20) and (1.21). *[69]*

1.5.2.3. Exchange-Correlation Functional

The exchange-correlation functional in the Kohn-Sham framework, whose exact form is not known, is typically divided into two distinct terms: one representing exchange effects and the other accounting for correlation. This decomposition is represented in equation (1.9).

$$\mathrm{E_{xc}}\big(\rho(\mathbf{r})\big) = \mathrm{E_x}\big(\rho(\mathbf{r})\big) + \mathrm{E_c}\big(\rho(\mathbf{r})\big) \tag{1.9}$$

In equation 1.9, $\mathrm{E_{xc}}\big(\rho(\mathbf{r})\big)$ describe exchange-correlation functional, whereas $\mathrm{E_x}\big(\rho(\mathbf{r})\big)$ and $\mathrm{E_c}\big(\rho(\mathbf{r})\big)$ describe the exchange and correlation component of the structure. The exchange-correlation functional $E_{xc}(n(r))$ is often estimated using local approximations, which are outlined in the following sections.

1.5.2.4. Local Density Approximation (LDA)

The Local Density Approximation (LDA) is commonly employed to estimate the exchange-correlation functional $\mathrm{E_{xc}}\big(\rho(\mathbf{r})\big)$.*[70-71]* This method assumes a uniformly distributed electron gas, serving as the basis for the approximation, and is defined as follows:

$$\mathrm{E_{xc}^{LDA}} = \int \mathrm{d^3r}\, \rho(\mathbf{r})\, \varepsilon_{\mathrm{xc}}^{\mathrm{hom}}((\mathbf{r})) \tag{1.10}$$

In equation 1.10, $\varepsilon_{\mathrm{xc}}^{\mathrm{hom}}(\rho(\mathbf{r}))$ represent the exchange-correlation energy per particle with the electron density $\rho(\mathbf{r})$ in the homogeneous gas. This approximation proves effective for estimating the ground-state properties of solids where electron density changes gradually. However, it significantly underperforms when calculating values such as cohesive energy, formation energy, bond dissociation energy, and adsorption energies, often deviating from experimental results. Moreover, it fails to accurately capture the band gaps of semiconductors and insulators. These drawbacks underline the necessity for improved approaches and more refined models to better describe these aspects of solid-state systems.*[72]*

1.5.2.5. Generalized Gradient Approximation (GGA)

To overcome the shortcomings of the Local Density Approximation (LDA), the Generalized Gradient Approximation (GGA) was developed. This method refines the estimation of the exchange-correlation functional by incorporating not just the electron density, but also its gradient. By accounting for the spatial variation in electron distribution, GGA offers improved accuracy over LDA. The functional form used in GGA is presented in equation (1.11).

$$E_{xc}^{GGA} = \int d^3r\, \rho(\mathbf{r})\, \varepsilon_{xc}^{GGA}\big(\rho(\mathbf{r}), \nabla\rho(\mathbf{r})\big) \tag{1.11}$$

where, $\varepsilon_{xc}^{GGA}\big(\rho(\mathbf{r}), \nabla\rho(\mathbf{r})\big)$ represents the exchange-correlation energy per electron, incorporating both the local density and its gradient. Among the GGA formulations, the Perdew-Burke-Ernzerhof (PBE) functional is particularly popular, especially for systems with rapidly changing electron densities. GGA-PBE has demonstrated reliable performance in computing key properties such as total energy, cohesive and formation energies, adsorption characteristics, and structural parameters like lattice constants. The exchange energy component within the GGA-PBE framework can be expressed as:

$$E_x^{PBE} = \int d^3r\, \rho(\mathbf{r})\, \varepsilon_x^{PBE}\big(\rho(\mathbf{r}), s(\mathbf{r})\big) \tag{1.12}$$

In equation (1.12), the exchange energy within the PBE formulation is expressed as a product of two terms: the enhancement factor F_x^{PBE} and the LDA exchange energy. This relationship is further detailed in equation (1.13).

$$\varepsilon_x^{PBE}\big(\rho(\mathbf{r}), s(\mathbf{r})\big) = \varepsilon_x^{LDA}\Big(\rho(\mathbf{r})) * F_x^{PBE}\,(s(\mathbf{r})\Big) \tag{1.13}$$

In addition to PBE, several other GGA-based functionals such as PW91 by Perdew and Wang, PBEsol, and revised PBE—have been developed to compute exchange-correlation energies with improved accuracy for specific material classes.*[73–75]*

1.5.2.6. Projector Augmented Wave (PAW) Method

In an atom, core and valence electrons exhibit markedly different behaviors due to their spatial proximity to the nucleus. Core electrons, being closer to the nucleus, have wavefunctions that oscillate rapidly, while valence electrons display much smoother wavefunctions. To accurately model these distinctions, different basis sets are applied. Valence electrons are typically described using plane wave basis sets, which are efficient and widely adopted in solid-state calculations. However, the complex, highly localized nature of core electron wavefunctions makes plane waves unsuitable for their description. Instead, the projector augmented wave (PAW) method is employed. This technique uses a combination of plane waves for valence electrons and a partial wave expansion for the core region, allowing for both computational efficiency and accurate electronic structure representation.*[76–79]*

Using a linear transformation operator (T) the all-electron wavefunction (Ψ_n) is mapped onto a corresponding pseudo-wavefunction ($\widetilde{\Psi}_n$). This pseudo-wavefunction is constructed as a functional transformation of the original wavefunction Ψ_n can be written as:

$$|\Psi_n\rangle = T|\widetilde{\Psi}_n\rangle \tag{1.14}$$

where T is the transformation operator and $\widetilde{\Psi}_n$ and Ψ_n can be represented as linear combination of partial waves for each augmentation regions.

$$|\Psi_n\rangle = \sum_i c_i |\phi_i\rangle \tag{1.15}$$

$$|\widetilde{\Psi}_n\rangle = \sum_i c_i |\widetilde{\phi}_i\rangle \tag{1.16}$$

Therefore, the transformation operator T can be expressed as equation 1.17,

$$T = 1 + \sum_i (|\phi_n\rangle - |\widetilde{\phi}_n\rangle) \langle\tilde{p}_i| \tag{1.17}$$

where, $\langle\tilde{p}_i|$ denotes the projection function used in the transformation process. The pseudopotential method effectively addresses the difficulties

posed by the complex, oscillatory behaviour of core electron wavefunctions by converting them into smoother, more computationally manageable forms. This approach is especially useful for exploring the electronic properties of solid-state systems. The PAW method builds on this concept and has been further enhanced by combining it with techniques like ultra-soft pseudopotentials and linear augmented plane wave (LAPW) features. In this thesis, electronic structure calculations were performed using the PAW method implemented within the Vienna Ab-initio Simulation Package (VASP). This combination offers a robust and efficient framework for accurately modelling the electronic behaviour of materials.*[80]*

1.5.3. Dispersion in Density Functional Theory

The approaches discussed so far often fail to capture long-range dispersion interactions, which arise from distance-dependent forces between electron densities. Notably, Coulombic and exchange interactions are influenced by the transition density between interacting fragments, and their accurate treatment is essential for describing dispersion forces. To address this, specialized methods have been developed, such as dispersion-corrected DFT (DFT-D), van der Waals (vdW) functionals, and empirical force field approaches. These techniques explicitly account for long-range interactions, offering a more complete and reliable description of dispersion effects in both molecular and solid-state systems.

$$E_{Disp}^{(2)} = \sum_{ia} \sum_{jb} \frac{(ia|jb)[(ia|jb)-(ja|ib)]}{\varepsilon_a + \epsilon_b - \varepsilon_i - \varepsilon_j} \quad (1.18)$$

In this context, particle-hole excitations between orbitals $i \rightarrow a$ and $j \rightarrow b$, localized on fragments A and B respectively, contribute to dispersion interactions. However, standard DFT lacks access to orbital energy differences ε, making it insufficient for accurately capturing long-range dispersion effects.*[81]* A widely adopted solution to this problem is Grimme's DFT-Dn approach, an empirical correction method that

introduces dispersion terms into the DFT framework. This method provides a general expression for dispersion energy, formulated as follows *[82]*:

$$E_{Disp}^{DFT-D} = -\sum_{AB}\sum_{n=6,8,10....} S_n \frac{C_n^{AB}}{R_{AB}^n} f_{damp}(R_{AB}) \quad (1.19)$$

In this expression, R_{AB} denotes the distance between fragments A and B, while C_n^{AB} represents the dispersion coefficient specific to the interacting pair. The term S_n, acts as a scaling factor to adjust repulsive interactions*[83]* and the damping function $f_{damp}(R_{AB})$ is introduced to mitigate the double-counting of correlation effects at intermediate distances.*[84]*

In this work, Grimme's DFT-D3 method was utilized for the majority of DFT-based simulations. This approach incorporates dispersion effects by accounting for three-body interactions among atomic triplets, allowing for a more realistic treatment of long-range forces. While the dispersion correction does not modify intrinsic electronic properties like the wavefunction, it does influence the computed atomic forces. As a result, geometric optimizations may differ from those obtained without dispersion corrections. Incorporating these effects enhances the accuracy of intermolecular interaction modeling, ultimately leading to more reliable predictions of structural and energetic properties.

1.6. Machine Learning Methods

Machine learning (ML) is broadly categorized into supervised and unsupervised learning, distinguished by the presence or absence of labeled output data.*[85]* In supervised learning, the algorithm is trained on input data paired with known output labels to learn a mapping from features to outcomes.*[86]* Unsupervised learning, by contrast, deals with unlabeled data and tries to uncover underlying patterns or groupings without explicit guidance.*[87]* In the context of materials design (e.g., battery materials), both paradigms are useful: supervised methods can learn relationships between material features and properties, while unsupervised methods can reveal natural clusters or reduce data complexity.*[88]*

1.6.1. Supervised Learning

Supervised learning algorithms utilize labeled datasets (feature vectors with target values) to train models that can predict outputs for new, unseen inputs. The two major types of supervised learning tasks are classification and regression, depending on the nature of the output. We have discussed each step of general ML workflow in the next section and different regression models as the thesis mainly oriented on the regression problem.

1.6.1.1. Regression

In regression tasks, the goal is to predict a continuous numerical value based on input features. A simple example is predicting a battery material's voltage and volume change from its descriptors.*[89, 90]* The most important part of any ML related problem is data which either needs to be extracted from the experimental report or some database or generated through DFT.

1.6.1.2. Classification

Classification, on the other hand, involves assigning inputs to discrete categories or classes. Instead of predicting a numeric value, the model answers questions like "Is this material stable or unstable?" or "Is this electrolyte flammable or non-flammable?" Here, the target variable is categorical, and the model learns to distinguish between classes based on patterns in the feature space.*[91]*

Applications of classification in materials research include identifying crystal structures, predicting phase stability, and screening materials for specific functional groups or electronic behavior.*[92]* Classification methods such as logistic regression, k-nearest neighbors (KNN), support vector machines (SVM), and neural networks are especially powerful when the dataset captures rich, discriminative features.

Both regression and classification are supervised learning methods, meaning they rely on known outcomes during training. The choice between

them depends entirely on the nature of the prediction task, regression is used when outputs are continuous, and classification is used when they are categorical. In this work, regression serves as the primary modeling strategy, enabling precise numerical predictions of materials-related properties.

1.6.2. Feature Representations and Selection

How we represent materials or input data as features is crucial for ML model performance. In general, feature representation involves encoding the raw data (whether crystalline structures, compositions, or experimental conditions) into numerical descriptors that the model can ingest. In battery materials design, a variety of feature types can be used:

- **Compositional features**: descriptors derived from chemical formulas (e.g., fractions of certain elements, average atomic number, electronegativities, ionic radii, etc.). Tools like matminer provide many composition-based featurizers.*[93]*
- **Structural features**: for crystalline materials, features can include symmetry functions, radial distribution function statistics, formation energy per atom from DFT, volume per atom, bandgap, etc., possibly obtained from databases like Materials Project.*[94]*
- **Nanostructure or 2D-specific features**: for 2D materials, one may include layer thickness, surface area, functional groups on surfaces, etc. In the aNANt database context, features might be derived from the computed electronic band structure or formation energies of monolayers.
- **Engineered descriptors**: domain knowledge can guide creation of specific features, such as a known performance metric (e.g., the ratio of certain elemental properties that correlates with capacity or diffusion barriers).

Proper feature scaling (normalizing or standardizing features) is often performed so that features with larger numeric ranges do not dominate those with smaller ranges. Categorical inputs (if any) might be one-hot encoded. In summary, preparing high-quality, informative features that capture the underlying physics or chemistry of the problem greatly assists the ML model in learning relevant patterns.

In many cases, we have an abundance of candidate features, and not all are useful. Thus, feature selection or dimensionality reduction steps are taken to reduce the feature space to the most informative variables. One approach we employed is SelectKBest with an ANOVA F-value (Analysis of Variance F-test) to rank features by their correlation with the target variable.*[95]* The ANOVA F-test can be used for regression or classification to score each feature individually: it essentially measures how much variance in the target is explained by differences in that feature, relative to within-group variance. Features with higher F-scores are more strongly related to the target. Using this univariate feature selection, we can pick the top *K* features that have the highest F-values (or p-values below a threshold). This is a fast way to eliminate less relevant descriptors. For example, if we have 100 initial descriptors for each material, ANOVA F-value ranking might tell us that only, say, 20 of them have statistically significant correlation with battery capacity; we might then retain those 20 for model training to reduce noise and overfitting risk. It's important to compute the F-test on training data in each fold of cross-validation to avoid bias (when using it in a pipeline), since peaking at the full data can overestimate feature importance.

Another form of feature selection arises inherently from certain models: as noted, Lasso regression drives some coefficients to zero, effectively performing feature selection as part of model fitting.*[96]* Decision tree-based models also perform an implicit feature selection by splitting on informative features; we can extract feature importance scores from random

forests or gradient boosted trees to see which inputs are most influential.[97] In one part of our work, we complemented ANOVA selection with SHAP values (described below) to interpret feature importance after model training – ensuring that the chosen features not only correlate with the target but indeed have a consistent impact on model predictions.*[98]*

Dimensionality reduction techniques can also be applied either as preprocessing or integrated in modeling. Principal Component Analysis (PCA) can project features to principal components that capture the majority of variance, allowing one to reduce dimensionality while retaining most information.*[99]* Partial Least Squares (PLS) (as introduced earlier) is another supervised dimensionality reduction, finding latent features that explain both predictors and response well.*[100]* Reducing dimensionality is especially valuable when dealing with high-throughput datasets from materials databases (which might have dozens or hundreds of descriptors per entry); it simplifies the model and often improves generalization.

1.6.3. Train and Test Data

A fundamental practice in machine learning is to split data into separate training and testing sets.*[101]* The model is fitted (trained) on the training set and then evaluated on the independent test set to assess its performance on unseen data. This simulates how the model would perform in real-world predictions and guards against overfitting.*[86]* In a typical workflow, one might use ~70-80% of the data for training and hold out ~20-30% for testing (or use other splits as appropriate). If N is the total number of samples, and we choose a train–test split fraction of p (for training), then the split sizes are $N_{\text{train}} = pN$ and $N_{\text{test}} = (1-p)N$ (rounded to integers). For example, an 80/20 split on a dataset of size N means $N_{\text{train}} = and\ N_{\text{test}} = 0.2N$. This separation mimics prospective prediction on unseen data and helps detect overfitting. Often the data is shuffled randomly before splitting to ensure each subset is representative.

In addition, one may set aside a validation set (especially if also tuning hyperparameters) or use cross-validation.*[102]* The key principle is that no information from the test set should be used in model training. Only after finalizing the model, we evaluate it on the test set to obtain an unbiased estimate of its true performance.*[103]*

Training error alone is not a reliable indicator of model performance – a model can overfit the training data (achieving very low error) yet fail to generalize to new data. By evaluating on the test set (which the model never saw during training), we obtain an unbiased estimate of predictive accuracy.*[104]*

1.6.4. Cross-Validation and Averaging

Cross-validation (CV) provides a more robust way to estimate model performance (and perform model selection) by averaging results across multiple train–test splits.*[102]* In K-fold cross-validation, the dataset is partitioned into K roughly equal folds. The model is trained on $K-1$ folds and evaluated on the remaining 1 fold; this process is repeated K times, each time using a different fold as the validation set. The K resulting validation scores (e.g. accuracies, MSEs, etc.) are then averaged to yield the cross-validation score:

$$\text{CV-}K\text{ Score} = \frac{1}{K}\sum_{k=1}^{K} \text{Score}_{\text{valid}}^{(k)}$$

which is an estimate of the model's generalization performance. By averaging over folds, CV reduces the variance associated with a single train–test split. Common cases are $K = 5$ or $K = 10$ (5-fold or 10-fold CV), as well as leave-one-out CV (LOOCV, where $K = N$).*[86]* Cross-validation is often used for hyperparameter tuning, one chooses the hyperparameters that maximize the average validation score. Note that while CV uses all data for training (across folds), one still should have an

independent test set for final evaluation once the model and hyperparameters are set, to avoid bias from having used all data in training.

1.6.4.1. K-Fold Cross-Validation

In K-Fold Cross-Validation, the dataset is split into k equal subsets, or "folds." The model is trained on k−1 folds and validated on the remaining one. This process is repeated k times, with each fold serving as the test set exactly once. The final performance metric is then averaged across all iterations, yielding a robust measure of the model's generalization ability.

K-Fold CV strikes a practical balance between computational cost and evaluation reliability. It maximizes data usage, critical when samples are limited and provides a richer understanding of model behavior compared to a one-time train-test split. By mitigating variance in the evaluation process, it also reduces the risk of overfitting or underfitting to any particular data subset.

1.6.4.2. Repeated K-Fold Cross-Validation

Repeated K-Fold CV builds upon standard K-Fold by adding an element of randomization.*[105]* The dataset is repeatedly split into new K-Fold partitions n times, with random reshuffling occurring in each repetition. This strategy introduces diversity across folds and reduces evaluation variance, especially when datasets are small or have subtle patterns.

Repeated K-Fold is particularly valuable when a single round of K-Fold might be skewed by chance data splits. By averaging results across many random splits, the evaluation becomes more statistically reliable and less sensitive to outlier partitions.

1.6.4.3. Leave-One-Out Cross-Validation (LOOCV)

In Leave-One-Out Cross-Validation (LOOCV), every individual data point takes a turn as the test set while the remaining n-1 samples are used for training. This process is repeated n times, once for each data point.*[106]*

LOOCV is exhaustive and nearly unbiased, making it ideal for very small datasets where every sample is precious. However, it is computationally expensive for large datasets, as it requires training the model n separate times. Despite this, it remains a gold-standard benchmark when a detailed and low-bias estimate is required.

1.6.5. Evaluation Metrics: RMSE, MAE, R^2

Evaluating model performance quantitatively is vital, both to compare different models and to judge whether a model is sufficiently accurate for a given application. For regression models (which are prevalent in material property prediction), common error metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). Each metric provides a slightly different perspective on the model's predictive errors:

- **Mean Absolute Error (MAE):** The average of absolute residuals. For predictions $\hat{y}_i$ on true values y_i,

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

MAE is a linear error measure – each error contributes proportionally to its magnitude. It is more robust to outliers than RMSE and is measured in the same units as the target.*[107]*

- **Root Mean Square Error (RMSE):** The square root of the average of squared residuals.*[108]* It is essentially the standard deviation of the prediction errors. The formula is:

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

which is the square root of mean squared error (MSE). RMSE penalizes large errors more strongly than MAE (due to squaring). It is widely used

and directly interpretable as "typical magnitude of error." Lower RMSE indicates better fit.

- **Coefficient of Determination (R^2):** Also known as R^2 or *R-squared*, this is the fraction of variance in y explained by the model.*[109]* One definition is:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

where $SS_{\text{res}} = \sum_i (y_i - \hat{y}i)^2$ *is the residual sum of squares and* $SS\text{tot} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares around the mean of y. Thus $R^2 = 1$ indicates a perfect fit (zero residual error), and $R^2 = 0$ indicates the model predicts no better than the mean of y. In simple linear regression, R^2 is the square of the Pearson correlation between y_i and $\hat{y}_i$. In multiple regression, it generalizes to the proportion of variance explained by all predictors.

R^2 can be negative if the model is worse than the baseline mean predictor (when not using an intercept or in some nonlinear fits). Often the adjusted R^2 is reported, which penalizes inclusion of additional features to compensate for R^2 always increasing (or staying the same) when more predictors are added.

In summary, MAE and RMSE give absolute error scales (RMSE emphasizing larger errors), while R^2 is unitless and tells how well the model captures variance in the data.

1.6.6. Machine Learning Algorithms

Having covered the general concepts of ML, we now discuss specific algorithms applied in this thesis. We employed a range of algorithms, from simple linear models to complex ensemble techniques, each chosen based on the nature of the task (regression) and the size/complexity of available data. Below we introduce each major algorithm category, explain how it works, and note any particular considerations or advantages relevant to materials design problems.

1.6.6.1. Linear Regression

Linear Regression is the simplest and most interpretable regression model.*[110]* It assumes a linear relationship between features and target: $\hat{y} = w_0 + \sum_j w_j x_j$. The model parameters w_j are fitted by minimizing the sum of squared residuals between predictions and true values (Ordinary Least Squares solution). Despite its simplicity, linear regression can be surprisingly effective for well-behaved data and offers clear insights into feature influence (via the weights). However, plain linear regression can overfit if there are many features or multicollinearity. We mitigated this by using regularized variants like ridge (L2 penalty) or lasso (L1 penalty) when needed, as well as by feature selection as described earlier.

For example, we applied linear regression to a dataset of material formation energies with features like elemental fractions and electronegativities. The linear model provided a baseline with an R^2 of around 0.5. By examining the learned coefficients, we could confirm known chemical trends (e.g., a large positive weight on a feature representing an element's atomic radius might indicate that increasing that radius tends to raise the formation energy). This interpretability is a strength of linear models – each coefficient directly indicates the direction and magnitude of that feature's effect on the prediction, assuming other features held fixed.

Linear regression models a continuous target as a linear combination of features, $y \approx X\beta$ + noise

The optimal coefficients β are obtained by minimizing the sum of squared residuals $S(\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2$

Setting the gradient to zero yields the normal equations $X^T X, \hat{\beta} = X^T y$

The closed-form solution (for full column rank X) is:

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

which is the ordinary least squares (OLS) estimator. Here X is the $n \times p$ design matrix (with a column of ones for the intercept), and y is the $n \times 1$ target vector. The predicted outcomes are $\hat{y} = X\hat{\beta}$. This solution minimizes the mean squared error on the training data and provides the foundation for many extensions.

1.6.6.2. Ridge Regression

Ridge regression adds an L_2 penalty to the linear regression loss to prevent overfitting.*[110]* The ridge objective function is:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 + \alpha \sum_{j=1}^{p} \beta_j^2$$

where $\hat{y}i = x_i^T$ *is the predicted value, and* $\alpha > 0$ *is the regularization strength. The second term* $\alpha \sum_j \beta_j^2$ *penalizes large coefficients. The normal equations become* $(X^T X + \alpha I)\widehat{\beta_{\text{ridge}}} = X^T y$, so the ridge solution can be written as:

$$\widehat{\beta_{ridge}} = (X^T X + \alpha I)^{-1} X^T y$$

analogous to OLS but with a shrinkage term. Ridge regression shrinks coefficients towards zero (but never exactly zero) to reduce variance at the cost of some bias. Ridge tends to shrink weights towards zero (but not exactly zero), especially useful in cases of many correlated features. It helped in our work to avoid overestimation of effects and to handle multicollinearity in descriptors (which is common if, say, two features are both measuring atomic size in different ways).*[111]*

1.6.6.3. Lasso Regression

not only prevents overfitting but also sets some weights exactly to zero, thus performing feature selection. When we had high-dimensional descriptor sets, lasso was valuable to pinpoint which features actually mattered.*[110]* A combination approach known as Elastic Net (not explicitly listed but

worth noting) blends L1 and L2 penalties and was also considered to get a balance of ridge and lasso benefits.

Lasso regression adds an L_1 penalty to the OLS loss, promoting sparsity in the coefficients. The lasso optimization problem is:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y_i})^2 + \alpha \sum_{j=1}^{p} |\beta_j|$$

with $\alpha > 0$ controlling the strength of regularization. Unlike ridge, the L_1term can force some $\widehat{\beta_j}$ exactly to zero, performing feature selection.*[96]* There is no closed-form solution for lasso; instead, techniques like coordinate descent are used to solve it. The subgradient optimality conditions yield that each coefficient's solution is a soft-thresholded version of the OLS estimate. In summary, lasso yields a sparse coefficient vector by minimizing squared error with an absolute penalty.

1.6.6.4. Kernel Ridge Regression

Kernel ridge regression (KRR) extends ridge regression to non-linear functions by using kernel functions. Kernel Ridge Regression is an extension of ridge regression that can model nonlinear relationships by using the kernel trick. Essentially, KRR performs ridge regression in a high-dimensional feature space implicitly mapped by a kernel function.*[112]* One commonly used kernel is the Gaussian RBF kernel, $K(x_i, x_j) = \exp(-\gamma|x_i - x_j|^2)$, which allows the model to fit nonlinear patterns in the original input space. The dual form of ridge regression uses kernel functions, and the solution involves kernel matrix inversions (hence it's more computationally expensive, scaling with $O(N^3)$ for N training points, but works well for moderate N).

In the dual form, the model is $f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$, where $K(\cdot,\cdot)$ is a positive-definite kernel and α_i are coefficients. The training objective in

dual form is $\min_{\alpha}|y - K\alpha|_2^2 + \lambda\,\alpha^T K$, leading to the normal equations $(K + \lambda I)\hat{\alpha} = y$. The closed-form dual solution is:

$$\hat{\alpha} = (K + \lambda I)^{-1} y$$

where K is the $n \times n$ kernel matrix with $K_{ij} = K(x_i, x_j)$. Prediction for a new input x is $\hat{y}(x) = \sum_{i=1}^{n} \hat{\alpha}_i\, , K(x_i, x)$. Kernel ridge regression thus finds a function in the reproducing kernel Hilbert space that fits the data with an L_2-regularization on function norm, enabling non-linear regression via the kernel trick.

We used KRR in scenarios where we suspected a nonlinear relationship between features and target that simpler linear models couldn't capture. For example, predicting the bandgap of materials from compositional attributes might not be strictly linear (due to threshold effects, etc.), and KRR with an RBF kernel could flexibly fit such data. KRR has two hyperparameters to tune: the regularization strength (analogous to ridge's alpha) and the kernel length-scale (like γ\gamma in the RBF kernel). We tuned these via cross-validation. KRR often gave us a boost in accuracy over linear regression at the cost of model interpretability and computational efficiency. It serves as a nice intermediate between linear models and more complex machine learning models – it's a nonparametric model that can achieve high performance on smaller datasets, which is often the regime in materials informatics.

1.6.6.5. Decision Tree Regressor

A Decision Tree Regressor is a non-parametric model that learns piecewise constant predictions by recursively partitioning the feature space. Each split in a regression tree is based on a feature and a cutoff value, chosen to minimize the target variance in the resulting two subsets (equivalently, to maximize the reduction in sum of squared errors). The tree grows until some stopping criteria (like minimum samples per leaf or maximum depth) are met, or until the leaves are pure (perfectly fit the training data). Decision

trees can capture nonlinear relationships and feature interactions in an intuitive if-else rules manner.*[110]*

For regression trees, splits are chosen to maximize the reduction in target variance (equivalently, minimize mean squared error) after the split. If a node T with sample set S is split into two subsets S_L and S_R, the variance reduction achieved is:

$$\Delta \mathrm{V} = \mathrm{Var(S)} - \left(\frac{|\mathrm{S_L}|}{|\mathrm{S}|} \mathrm{Var(S_L)} + \frac{|\mathrm{S_R}|}{|\mathrm{S}|} \mathrm{Var(S_R)} \right)$$

where $Var(S)$ denotes the variance of target values in set S. Expanding $Var(S)$ in terms of sums of squared deviations, this is equivalent to comparing the total within-node sum of squared errors before and after splitting. A split is chosen to maximize ΔV, i.e. to achieve the largest drop in weighted variance. In practice, this means picking the feature and threshold that yield the most homogeneous (lowest variance) child nodes. If $\overline{y_S}$ is the mean of S, note that

$$\mathrm{Var}(S) = \frac{1}{|S|} \sum_{i \in S} (y_i - \overline{y_S})^2$$

Thus, the criterion aligns with minimizing MSE in children. (For classification trees, analogous impurity measures like Gini index or entropy are used.)

We utilized decision tree regression for its interpretability and ability to handle feature interactions. A decision tree on a materials dataset might produce rules like: “if formation energy < X and atomic weight > Y then predicted capacity = Z”. Such rules can sometimes correspond to human-understandable thresholds. However, single decision trees are prone to overfitting – they can create very complex branching structures that fit noise. We pruned trees or set maximum depth to avoid this. In our experiments, a standalone decision tree was usually not the best-performing model (often it underfit if heavily pruned, or overfit if grown too deep). But

it laid the groundwork for understanding the data structure and served as the base learner for ensemble methods like Random Forests and Gradient Boosting.

1.6.6.6. Random Forest Averaging

A Random Forest is an ensemble of decision trees, introduced by Breiman (2001), which reduces overfitting by averaging many trees trained with randomness. In a random forest regressor, each tree is trained on a bootstrap sample of the data (bagging) and at each split, the algorithm considers a random subset of features rather than all features. This randomness injects diversity into the trees such that their predictions are not perfectly correlated. The forest's prediction is the average of predictions from all individual trees. By the law of large numbers, this averaging tends to cancel out the individual trees' errors (especially the high-variance errors) and yield a more robust predictor.*[110]*

A random forest is an ensemble of B decision trees trained on random subsets of data and/or features. For regression tasks, the forest prediction is the average of the individual tree predictions. If $f^{(b)}(x)$ is the prediction of the b-th tree for input x, the random forest output is:

$$\widehat{y_{\mathrm{RF}}}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{(b)}(x)$$

Each tree in the forest is grown (typically to full depth) on a bootstrap sample of the training data with random feature selection at splits, which injects diversity. By averaging many de-correlated trees, the random forest reduces variance while remaining approximately unbiased. For classification, the analogous operation is majority voting among trees (or averaging predicted class probabilities). Random forests thus improve generalization by combining multiple high-variance learners into a low-variance ensemble.*[97]*

RFR often provided excellent accuracy out-of-the-box with relatively few hyperparameters to tune (mainly the number of trees and perhaps max features per split). Second, it handles nonlinear relationships and interactions naturally, which is useful for complex material datasets where properties emerge from combinations of features. Third, random forests give useful measures of feature importance (based on how much each feature split reduces error on average) that we used as a guide alongside SHAP values. For instance, if a random forest consistently uses "atomic radius" in top splits across many trees, it will show a high importance for that feature, suggesting it's a key driver in the property prediction.

1.6.6.7. Gradient Boosting Regression

Gradient Boosting Regression refers to ensemble models where trees are added sequentially (instead of in parallel like a forest), each new tree correcting the errors of the current ensemble. The algorithm (proposed by Friedman) works by fitting a small decision tree to the residuals of the model's current predictions, thereby boosting performance step by step. Essentially, gradient boosting performs a stage-wise optimization of a loss function (typically squared error for regression) by taking gradient steps in function space. Each tree is usually shallow (often called a "weak learner"), and hundreds of such trees can be combined.*[113]*

Gradient boosting builds an additive model in a forward stagewise fashion, where each new learner $h_m(x)$ is fit to the pseudo-residuals (negative gradients of the loss) of the current model. Starting from an initial prediction $F_0(x)$(often F_0 is a constant like the mean of y), at each iteration $m = 1, \dots, M$ the model is updated as:

$$F_m(x) = F_{m-1}(x) + \nu h_m(x)$$

where $h_m(x)$ is the new weak learner (e.g. a regression tree) trained on the residuals from stage $m-1$, and $0 < \nu \leq 1$ is a learning rate (shrinkage factor). In practice, $h_m(x)$ is fit to approximate the negative gradient of the

loss: $h_m(x) \approx -\frac{\partial}{\partial F_{m-1}} \sum_i L\big(y_i, F_{m-1}(x_i)\big)$, so that adding ν, h_m in that direction most reduces the loss. Optionally, a line search may be done to find an optimal multiplier γ_m: $F_m(x) = F_{m-1}(x) + \nu, \gamma_m h_m(x)$, where $\gamma_m = \arg\min_{\gamma} \sum_i L\big(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\big)$. Gradient boosting thus iteratively refines the model by greedily adding predictors that correct the remaining errors (gradients), yielding a powerful ensemble.

One should note that while gradient boosting often gives top accuracy, it can be sensitive to hyperparameters and noise. We found it crucial to do proper cross-validation when using boosting, as it can otherwise overfit the training set if, say, too many trees are used. Regularization options like subsampling of data and columns per tree, and minimum leaf sizes were also leveraged to keep the model general.

1.6.6.8. XGBoost

XGBoost (Extreme Gradient Boosting) uses a specific form of gradient boosting with a regularized objective to build decision-tree ensembles. Technically, XGBoost includes additional regularization terms in the objective, support for parallel tree construction, and other engineering improvements that make it faster and often more accurate than vanilla gradient boosting. When we refer to XGBR, we mean using XGBoost's regressor for our problems.

At iteration t, let $f_t(x)$ be the new tree (with T leaves) to add. XGBoost's objective can be written as:

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t)$$

where $l(y, \hat{y})$ is the training loss (e.g. squared error or logistic loss) and $\Omega(f_t)$ is the regularization term for the tree (penalizing model complexity). For a tree model $f(x)$ with leaf weights $w_1, \ldots, w_T$, a common regularizer

is $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ with $\gamma, \lambda \geq 0$) which penalizes number of leaves and magnitude of weights. Using a second-order Taylor expansion of l around the current prediction, the approximate gain from adding a tree can be derived. If I_j is the set of data indices in leaf j: define $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$ as the sum of first and second derivatives (gradients g_i and Hessians h_i) of the loss for that leaf. The optimal weight for leaf j is:

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

and the corresponding second-order *approximate* reduction in objective from that leaf is $\frac{1}{2}\frac{G_j^2}{H_j+\lambda}$. The overall tree's score (objective decrease) is $\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda} - \gamma T$. A split is made if it yields a positive gain larger than a threshold. In summary, XGBoost's tree-building algorithm uses these formulas to greedily split nodes (maximizing gain) and set leaf weights, resulting in an efficiently optimized boosted tree model.*[114]*

XGBoost also provided model interpretability. One could extract feature importances (much like random forest) and even use SHAP (Shapley Additive Explanations) specifically geared for tree models (TreeSHAP) to interpret how features affect predictions. We discuss SHAP in the next section, but it's worth noting here that tools like SHAP have made interpreting complex models like XGBoost more feasible. Thus, we were able to extracting insights, like which features increase or decrease the predicted target and by how much on average along with XGBoost's accuracy.

1.6.6.9. Partial Least Squares (PLS) Regression

Partial Least Squares regression finds a set of latent factors by simultaneously considering predictors X and response Y. These latent

factors (or components) are linear combinations of original features, chosen to explain as much covariance as possible between X and Y. Then a regression is performed in this latent space. Essentially, PLS projects the data into a lower-dimensional subspace (like PCA but supervised) and then does linear regression in that space.*[100]*

PLS is particularly valuable in situations with many features that are highly collinear and relatively few data points – a scenario not uncommon in materials informatics where one might compute dozens of descriptors for only tens or hundreds of materials. In such cases, ordinary linear regression would be ill-posed or overfit, whereas PLS can yield a stable solution by using a limited number of components.

Partial Least Squares regression finds a set of latent variables (components) that both capture variance in the predictors X and have high correlation with the responses Y. PLS decomposes the data as:

$$X = TP^T + E, \qquad Y = UQ^T + F$$

where T and U are matrices of h latent scores, P and Q are loadings, and E, F are residuals. Each PLS component is obtained by finding weight vectors w (for X) and c (for Y) that maximize the covariance between the transformed variables $t = Xw$ and $u = Yc$. Equivalently, the first PLS weight $w^{(1)}$ is the top singular vector of the cross-covariance X^TY, so that $t^{(1)} = Xw^{(1)}$ and $u^{(1)} = Yc^{(1)}$ have maximal covariance. Subsequent components are computed on deflated matrices (after removing the variance explained by earlier components). In effect, PLS finds an optimal low-dimensional representation of X (scores T) that is most predictive of Y. A concise objective for PLS (one component case) can be written as minimizing the reconstruction error $\sum_i (y_i - x_i^T w, c)^2$ over w, c (with normalization constraints), which leads to the same solution as the covariance maximization. PLS is especially useful in high-dimensional settings where X has many variables collinear with each other.

1.6.7. Unsupervised Learning

Unsupervised learning finds hidden structure in unlabeled data.*[115]* It is particularly useful for exploring material datasets where class labels may not be pre-defined. A primary unsupervised task is clustering, where the aim is to group materials or data points such that similar items fall into the same cluster.*[116]*

K-Means Clustering is one of the simplest and most popular clustering algorithms that we have discussed in the next section.*[117]*

1.6.7.1. K-Means Clustering

K-means clustering partitions n observations into K clusters by minimizing the within-cluster sum of squared distances. The objective function for *K*-means is:

$$\arg\min_{\mu_1,\ldots,\mu_K} \sum_{i=1}^{K} \sum_{x_j \in C_i} |x_j - \mu_i|^2$$

where C_i is the set of points assigned to cluster i and μ_i is the centroid (mean) of cluster i. In words, *K*-means seeks cluster centers μ_i that minimize the sum of squared Euclidean distances from each data point to its nearest center. This is a non-convex problem but the standard *K*-means algorithm (Lloyd's algorithm) finds a local minimum by alternating between assignment (assign each point to its closest μ_i) and update (recompute each μ_i as the mean of points in C_i).*[118]* The objective $E(K)$ is also called the within-cluster sum of squares (WCSS). Because the cost decreases with increasing K (clusters can always be split to reduce error), one must use methods like the elbow plot or cross-validation to choose a suitable number of clusters.*[119]*

1.6.7.1.1. Elbow Method for Optimal *K*

The elbow method is a heuristic for choosing the number of clusters in *K*-means by looking at how the clustering cost (SSE or WCSS) decreases as

K increases.*[120]* One computes the total within-cluster sum of squared errors (SSE) for different K values:

$$\mathrm{SSE}(K) = \sum_{i=1}^{K} \sum_{x \in C_i} |\, x - \mu_i|^2$$

which is simply the *K*-means objective value for K clusters. This SSE will always decrease (or stay the same) as K increases, since adding more clusters can only reduce within-cluster variance. The elbow method involves plotting $\mathrm{SSE}(K)$ against K and looking for a "bend" or elbow: the value K beyond which the marginal gain (SSE reduction) diminishes significantly. The elbow point is taken as the optimal number of clusters, as it balances model complexity with explained variance. Essentially, before the elbow each additional cluster provides substantial improvement, while after the elbow the improvements level off (indicating possible overfitting if more clusters are added).*[121]* This method is subjective, but when a clear elbow exists it provides a reasonable choice for K.

1.6.7.1.2. Silhouette Score

The silhouette score measures clustering quality by assessing how similar an object is to others in its cluster (cohesion) compared to objects in other clusters (separation). For each data point i, define:

- $a(i)$ the average distance from i to all other points in the same cluster. This measures how tightly i is coupled with its cluster mates (lower $a(i)$ means i is well inside its cluster).
- $b(i)$ the minimum over all other clusters C of the average distance from i to points in cluster C. This is the distance from i to its nearest neighboring cluster (smaller $b(i)$ means a closer alternative cluster).

The silhouette score for point i is then defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

which always lies in the range $-1 \leq s(i) \leq +1$. A score $s(i)$ close to $+1$ means i's cluster assignment is appropriate (it is much closer to its own cluster than to any other), whereas $s(i)$ near 0 indicates the point lies on the boundary between clusters. Negative values mean i is closer on average to a different cluster than its own, suggesting potential mis-clustering. The overall silhouette score for the clustering is the average $s(i)$ over all points. This metric provides a way to evaluate clustering validity and also to choose an optimal K (by maximizing average silhouette over different K).

The formula above is equivalent to

$s(i) = 1 - \frac{a(i)}{b(i)}$ if $a(i) < b(i)$,

$s(i) = 0$ if $a(i) = b(i)$, and

$s(i) = \frac{b(i)}{a(i)} - 1$ if $a(i) > b(i)$.

It thus intuitively reflects the relative difference between within-cluster and nearest-other-cluster distances. Another important unsupervised technique is dimensionality reduction, which aims to simplify high-dimensional data while preserving essential structure. Techniques like Principal Component Analysis (PCA) and t-SNE/UMAP can be used to visualize complex composition–structure–property datasets in 2D or 3D, revealing patterns such as clustering of similar materials.*[121-123]* While PCA wasn't explicitly listed in the outline, it's worth noting that PLS (mentioned above) can be seen as a supervised dimensionality reduction, and unsupervised PCA could likewise be applied to reduce feature space before clustering or modeling.

1.6.8. ANOVA F-Test Formula

In one-way Analysis of Variance (ANOVA), we compare k group means to see if at least one differs significantly. The ANOVA F-statistic is the ratio of between-group variance to within-group variance.*[124, 125]* Computed from sums of squares, the formula is:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{\frac{SS_{\text{between}}}{k-1}}{\frac{SS_{\text{within}}}{N-k}}$$

where SS_{between} is the between-groups sum of squares, SS_{within} the within-groups (residual) sum of squares, and N the total sample size. Here MS denotes mean square (sum of squares divided by degrees of freedom). Expanded, if $\bar{y}_j$ is the mean of group j and $\bar{y}$ the overall mean, then

- $SS_{\text{between}} = \sum_{j=1}^{k} n_j \left(\bar{y}_j - \bar{y}\right)^2$ with degrees of freedom (df) $= k - 1$
- $SS_{\text{within}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(y_{ij} - \bar{y}_j\right)^2$ with df $= N - k$

F follows an $F \sim F(k-1, N-k)$ distribution under the null hypothesis that all group means are equal. A large F (significantly greater than 1) indicates that between-group variability dominates within-group variability, providing evidence against H_0. In essence, ANOVA partitions the total variance ($SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}}$) and F quantifies whether the explained variance (between) is large relative to unexplained variance (within).

1.6.9. Hyperparameter Tuning

Hyperparameter tuning is a critical phase in machine learning (ML) model development, directly influencing a model's performance, learning efficiency, and ability to generalize.*[126, 127]* Unlike model parameters which are learned from the data during training hyperparameters are predefined settings that govern the behavior of the learning algorithm itself. Examples include the learning rate in gradient-based models, the maximum depth of a decision tree, or the number of hidden layers in a neural network.

The primary objective of hyperparameter tuning is to identify the combination of settings that yields the best model performance. Well-tuned hyperparameters can dramatically improve both accuracy and robustness,

whereas poorly chosen ones may lead to underfitting (failing to learn enough) or overfitting (memorizing the training data too well). Two widely used methods for tuning hyperparameters are GridSearchCV and RandomSearchCV, both of which are available through the popular scikit-learn Python library.

1.6.9.1. Grid Search and Hyperparameter Tuning

Grid search is an exhaustive strategy for hyperparameter tuning where one specifies a discrete set of values for each hyperparameter and trains/evaluates the model for every combination.*[128]* The goal is to identify the hyperparameter tuple that yields the best performance on a validation set or via cross-validation.

for a chosen evaluation metric on the validation folds. For example, one might search over a grid of parameters like $\{\text{max\_depth} \in \{3,5,7\}, \text{learning\_rate} \in \{0.1,0.01\}\}$ for a gradient boosting model; this entails training models for each combination (e.g. 6 models) and picking the best. In practice, grid search is often combined with cross-validation (GridSearchCV in scikit-learn) to robustly estimate performance of each setting. The result is the set of hyperparameters that optimizes the cross-validation metric. Other strategies like random search or Bayesian optimization can also be used to explore the hyperparameter space more efficiently, but the concept remains evaluating the model under different hyperparameter values and choosing the optimum. Once the best hyperparameters are found, a final model is typically retrained on the full training set with those settings and then evaluated on the test set (which was not used during tuning) to report the generalization performance.

1.6.9.2. RandomSearchCV

RandomSearchCV, in contrast, samples combinations of hyperparameters at random from specified distributions. This strategy allows it to explore the

search space more broadly and often more efficiently, especially when only a subset of the parameters heavily influences performance.

RandomSearchCV is especially well-suited to high-dimensional or loosely constrained search spaces, where exhaustive search is infeasible.*[126]* While it doesn't guarantee finding the global optimum, it frequently identifies high-performing configurations with significantly fewer iterations. This makes it a practical choice for time-sensitive or computationally limited experiments.

Together, GridSearchCV and RandomSearchCV provide powerful tools for model optimization, each with its own strengths. In practice, the choice between them often depends on the complexity of the model, available resources, and time constraints. Regardless of method, hyperparameter tuning remains a cornerstone of building reliable, high-performing ML models—especially in domains like materials science, where predictive accuracy can directly impact experimental direction and discovery.

1.6.10. SHAP (Shapley Additive exPlanations)

SHAP values provide an interpretation for individual predictions by attributing the prediction difference (from the dataset baseline) to features, using concepts from cooperative game theory.*[98, 129]* The SHAP value for feature i is essentially the Shapley value of a "feature contribution game," which fairly distributes the prediction among features. The formula for the Shapley value of player (feature) i is:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\ (n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

where $N = 1, \ldots, n$ is the full set of features and the sum is over all subsets S not containing feature i. Here $v(S)$ represents the *value* (in SHAP context, usually the model output expectation) obtained by a coalition of features S being present (and the others absent). Intuitively, we consider all possible orders in which features can enter, and for each such ordering, compute

feature i's marginal contribution $v(S \cup i) - v(S)$; the Shapley value ϕ_i is the average of these contributions over all orderings. The factorial terms $\frac{|S|!(n-|S|-1)!}{n!}$ indeed correspond to the proportion of orderings where S is the set of features that come before i.

In an ML setting, $v(S)$ is often taken as the expected model prediction when only features in S are known (features not in S are marginalized out). Then ϕ_i attributes how much feature i contributes to the prediction compared to the baseline (average prediction over all data). By construction, SHAP values satisfy desirable properties: they sum to the difference between the actual prediction and the baseline, and they are fair in the sense that if two features contribute equally in any context, they get equal credit (Symmetry), and features that contribute nothing get zero ϕ (Dummy). Computing SHAP values exactly requires summing over 2^{n-1} subsets, which is generally intractable, but approximations or model-specific algorithms (for trees, linear models, etc.) can compute them efficiently. SHAP values are a popular choice for explaining complex models because of their clear game-theoretic interpretation and consistency with human notions of feature importance.

1.6.11. ML Potential

In atomistic modeling of materials (common in battery materials design), machine learning potentials often assume the total potential energy of a structure can be decomposed as a sum of contributions from individual atoms or local environments. Formally, if a structure contains M atoms, the predicted total energy is expressed as

$$E_{\text{total}} = \sum_{i=1}^{M} E_i$$

where E_i is the energy contribution associated with atom i (as a function of its local chemical environment). This assumption – used in High-Dimensional Neural Network potentials (HDNNPs) and other interatomic

ML potentials – allows the model to be extensive (energy scales with system size) and leverages locality. For example, Behler–Parrinello neural networks compute a set of descriptors for each atom's environment and then use a neural network to predict an energy E_i for that atom; the sum yields the total energy.*[130]* By training on reference calculations (e.g. DFT energies), the model learns to estimate atomic contributions such that their sum matches the total target energy.*[131]* For example, CHGNet (Crystal Hamiltonian Graph neural Network) an universal ML potential model is a graph-based deep learning model designed to predict energies, forces, magnetic moment and stresses from atomic configurations.*[132]* CHGNet represents crystal structures as graphs, incorporating atom, bond, and angle embeddings, and uses message passing through graph convolutional layers. Beyond accuracy, CHGNet is specifically optimized for large-scale simulations, making it suitable for exploring long-time dynamics and thermodynamic stability in materials. By unifying multiple physical quantities into a single model, CHGNet enables consistent predictions across diverse material classes, reducing the need for system-specific retraining. Moreover, its integration into high-throughput workflows allows accelerated screening of novel functional materials with near-DFT accuracy but significantly lower computational cost.

This decomposition has advantages such as the energy prediction automatically scales to larger systems (adding an atom just adds one more term), and the model can be trained on small structures and transferred to bigger ones.*[133]* It also provides some interpretability, as atomic energies E_i can indicate which sites or environments are energetically favorable. Note that long-range interactions (e.g. electrostatics) often violate the strict locality assumption, but they can be incorporated via specialized terms or by extending the descriptor range. Overall, the additive decomposition is a widely-used principle in ML interatomic potentials for breaking a complex quantum-mechanical energy into manageable local pieces.

1.6.12. Graph Neural Network

Graph Neural Networks (GNNs) have emerged as powerful models for predicting properties of molecules and materials (such as total energy) by operating on a graph representation of the structure.*[134]* In a typical materials GNN, each atom is a graph node with an initial feature (e.g. atom type, possibly local environment descriptors), and edges represent interatomic bonds or neighbor relations. Through iterative message-passing updates, the GNN computes node embeddings h_i that encode the local chemistry around atom i. To predict a total energy from the graph, a common approach is to use a readout function that aggregates node contributions. For example, the network might assign an energy to each atom and sum them (similar to the ML potential above):

$$\widehat{E_{\text{total}}} = \sum_{i=1}^{M} f_\theta(h_i)$$

where f_θ is a learnable function (often a small neural network) that maps the node embedding for atom i to a scalar atomic energy. The sum over nodes ensures the output is extensive (scales with system size) and permutation-invariant to atom indexing, both desirable properties for energy models. Alternatively, some GNNs perform a global pooling: $\hat{E} = g_\theta(\{h_i\}_{i=1}^{M})$, where g_θ could be a sum or more complex symmetric function.

The GNN is trained on known energies (and optionally forces) of example structures, adjusting weights so that the predicted energy matches the reference.*[135]* Notably, by using the graph structure, GNNs can capture both local and non-local interactions: message passing allows information to propagate and update atomic representations based on neighbors (and neighbors-of-neighbors, etc., over multiple layers). For instance, a GNN-based interatomic potential can naturally learn bond formations, angle strain, and other many-body effects. Once trained, the GNN can generalize to new structures, predicting their energy. Modern GNN architectures (like SchNet, MEGNet, DimeNet, GemNet, etc.) have achieved high accuracy in

reproducing quantum-calculated energies for molecules and crystals.*[136, 137]* In summary, a graph neural network predicts energy by learning an embedding for each atom and then combining atomic contributions (often via summation) to yield the total energy. This combines the physics-inspired idea of atomic energy additivity with the expressive power of neural networks that can capture complex dependencies in the graph of atoms.

1.7. References

(1) World Energy Outlook 2023 – Analysis - IEA (accessed 2025-07-25), https://www.iea.org/reports/world-energy-outlook-2023

(2) Lund H., Østergaard P.A., Connolly D., Ridjan I., Mathiesen B.V., Hvelplund F., Thellufsen J.Z., Sorknæs P. (2016), Energy storage and smart energy systems, Int. J. Sustain. Energy Plan. Manag., 11, 3–14 (DOI: 10.5278/ijsepm.2016.11.2)

(3) Rechargeable lithium batteries: from fundamentals to applications (accessed 2025-07-25), https://books.google.co.in/books?hl=en&lr=&id=eUSdBAAAQBAJ&oi=fnd&pg=PP1&dq=Rechargeable+lithium-ion+batteries:+From+fundamental+to+applications&ots=A4AxEAG4tp&sig=0WT0ru-akoONHQq28FKD_foA4SY

(4) Goodenough J.B., Park K.S. (2013), The Li-ion rechargeable battery: a perspective, J. Am. Chem. Soc., 135 (4), 1167–1176 (DOI: 10.1021/ja3091438)

(5) Nykvist B., Nilsson M. (2015), Rapidly falling costs of battery packs for electric vehicles, Nat. Clim. Chang., 5 (4), 329–332 (DOI: 10.1038/nclimate2564)

(6) Tarascon J.M., Armand M. (2001), Issues and challenges facing rechargeable lithium batteries, Nature, 414 (6861), 359–367 (DOI: 10.1038/35104644)

(7) Manthiram A. (2020), A reflection on lithium-ion battery cathode chemistry, Nat. Commun., 11, 1550 (DOI: 10.1038/s41467-020-15355-0)

(8) Dunn B., Kamath H., Tarascon J.M. (2011), Electrical energy storage for the grid: a battery of choices, Science (1979), 334 (6058), 928–935 (DOI: 10.1126/science.1212741)

(9) Xu B., Qian D., Wang Z., Meng Y.S. (2012), Recent progress in cathode materials research for advanced lithium ion batteries, Mater. Sci. Eng. R Rep., 73 (5–6), 51–65 (DOI: 10.1016/j.mser.2012.05.003)

(10) Marom R., Amalraj S.F., Leifer N., Jacob D., Aurbach D. (2011), A review of advanced and practical lithium battery materials, J. Mater. Chem., 21 (27), 9938–9954 (DOI: 10.1039/c0jm04225k)

(11) Gaines L. (2014), The future of automotive lithium-ion battery recycling: charting a sustainable course, Sustain. Mater. Technol., 1–2, 2–7 (DOI: 10.1016/j.susmat.2014.10.001)

(12) Zhang S.S. (2007), A review on the separators of liquid electrolyte Li-ion batteries, J. Power Sources, 164 (1), 351–364 (DOI: 10.1016/j.jpowsour.2006.10.065)

(13) Zubi G., Dufo-López R., Carvalho M., Pasaoglu G. (2018), The lithium-ion battery: state of the art and future perspectives, Renew. Sustain. Energy Rev., 89, 292–308 (DOI: 10.1016/j.rser.2018.03.002)

(14) Slater M.D., Kim D., Lee E., Johnson C.S. (2013), Sodium-ion batteries, Adv. Funct. Mater., 23 (8), 947–958 (DOI: 10.1002/adfm.201200691)

(15) Deng J., Luo W.B., Chou S.L., Liu H.K., Dou S.X. (2018), Sodium-ion batteries: from academic research to practical commercialization, Adv. Energy Mater., 8 (4), 1701428 (DOI: 10.1002/aenm.201701428)

(16) Muldoon J., Bucur C.B., Gregory T. (2014), Quest for nonaqueous multivalent secondary batteries: magnesium and beyond, Chem. Rev., 114 (23), 11683–11720 (DOI: 10.1021/cr500049y)

(17) Ponrouch A., Frontera C., Bardé F., Palacín M.R. (2016), Towards a calcium-based rechargeable battery, Nat. Mater., 15 (2), 169–172 (DOI: 10.1038/nmat4462)

(18) Lin M.C., Gong M., Lu B., Wu Y., Wang D.Y., Guan M., Angell M., Chen C., Yang J., Hwang B.J., Dai H. (2015), An ultrafast rechargeable aluminium-ion battery, Nature, 520 (7547), 325–328 (DOI: 10.1038/nature14340)

(19) Hafner J. (2008), Ab-initio simulations of materials using VASP: density-functional theory and beyond, J. Comput. Chem., 29 (13), 2044–2078 (DOI: 10.1002/jcc.21057)

(20) Jain A., Ong S.P., Hautier G., Chen W., Richards W.D., Dacek S., Cholia S., Gunter D., Skinner D., Ceder G., Persson K.A. (2013), Commentary: the Materials Project: a materials genome approach to accelerating materials innovation, APL Mater., 1 (1), 011002 (DOI: 10.1063/1.4812323)

(21) Ward L., Agrawal A., Choudhary A., Wolverton C. (2016), A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater., 2 (1), 1–7 (DOI: 10.1038/npjcompumats.2016.28)

(22) Choudhary K., Decost B., Tavazza F. (2018), Machine learning with force-field-inspired descriptors for materials: fast screening and mapping energy landscape, Phys. Rev. Mater., 2 (8), 083801 (DOI: 10.1103/physrevmaterials.2.083801)

(23) Xie T., Grossman J.C. (2018), Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett., 120, 145301 (DOI: 10.1103/physrevlett.120.145301)

(24) Lundberg S.M., Allen P.G., Lee S.I. (2017), A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst., 30

(25) Sanchez-Lengeling B., Aspuru-Guzik A. (2018), Inverse molecular design using machine learning: generative models for matter engineering, Science (1979), 361 (6400), 360–365 (DOI: 10.1126/science.aat2663)

(26) Dan Y., Zhao Y., Li X., Li S., Hu M., Hu J. (2020), Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials, npj Comput. Mater., 6 (1), 1–7 (DOI: 10.1038/s41524-020-00352-0)

(27) Chen C.T., Gu G.X. (2020), Generative deep neural networks for inverse materials design using backpropagation and active learning, Adv. Sci., 7 (5) (DOI: 10.1002/advs.201902607)

(28) Butler K.T., Davies D.W., Cartwright H., Isayev O., Walsh A. (2018), Machine learning for molecular and materials science, Nature, 559 (7715), 547–555 (DOI: 10.1038/s41586-018-0337-2)

(29) Tarascon J.M., Armand M. (2010), Issues and challenges facing rechargeable lithium batteries, Mater. Sustain. Energy, 171–179 (DOI: 10.1142/9789814317665_0024)

(30) Kim H., Kim J.C., Bianchini M., Seo D.H., Rodriguez-Garcia J., Ceder G. (2018), Recent progress and perspective in electrode materials for K-ion batteries, Adv. Energy Mater., 8 (9), 1702384 (DOI: 10.1002/aenm.201702384)

(31) Reiser P., Neubert M., Eberhard A., Torresi L., Zhou C., Shao C., Metni H., van Hoesel C., Schopmans H., Sommer T., Friederich P. (2022),

Graph neural networks for materials science and chemistry, Commun. Mater., 3 (1), 1–18 (DOI: 10.1038/s43246-022-00315-6)

(32) Si K., Sun Z., Song H., Jiang X., Wang X. (2025), Machine learning-assisted design and prediction of materials for batteries based on alkali metals, Phys. Chem. Chem. Phys., 27 (11), 5423–5442 (DOI: 10.1039/d4cp04214j)

(33) Manna S., Roy D., Das S., Pathak B. (2022), Capacity prediction of K-ion batteries: a machine learning based approach for high throughput screening of electrode materials, Mater. Adv., 3 (21), 7833–7845 (DOI: 10.1039/d2ma00746k)

(34) Ong S.P., Richards W.D., Jain A., Hautier G., Kocher M., Cholia S., Gunter D., Chevrier V.L., Persson K.A., Ceder G. (2013), Python materials genomics (pymatgen): a robust, open-source python library for materials analysis, Comput. Mater. Sci., 68, 314–319 (DOI: 10.1016/j.commatsci.2012.10.028)

(35) Ward L., Dunn A., Faghaninia A., Zimmermann N.E.R., Bajaj S., Wang Q., Montoya J., Chen J., Bystrom K., Dylla M., Chard K., Asta M., Persson K.A., Snyder G.J., Foster I., Jain A. (2018), Matminer: an open source toolkit for materials data mining, Comput. Mater. Sci., 152, 60–69 (DOI: 10.1016/j.commatsci.2018.05.018)

(36) Bhauriyal P., Mahata A., Pathak B. (2018), Graphene-like carbon-nitride monolayer: a potential anode material for Na- and K-ion batteries, J. Phys. Chem. C, 122 (5), 2481–2489 (DOI: 10.1021/acs.jpcc.7b09433)

(37) Xue Y., Xu T., Wang C., Fu L. (2024), Recent advances of two-dimensional materials-based heterostructures for rechargeable batteries, iScience, 27 (8), 110392 (DOI: 10.1016/j.isci.2024.110392)

(38) Deng B., Zhong P., Jun K.J., Riebesell J., Han K., Bartel C.J., Ceder G. (2023), CHGNet as a pretrained universal neural network potential for

charge-informed atomistic modelling, Nat. Mach. Intell., 5 (9), 1031–1041 (DOI: 10.1038/s42256-023-00716-3)

(39) Yamada Y., Wang J., Ko S., Watanabe E., Yamada A. (2019), Advances and issues in developing salt-concentrated battery electrolytes, Nat. Energy, 4 (4), 269–280 (DOI: 10.1038/s41560-019-0336-z)

(40) Peljo P., Villevieille C., Girault H.H. (2025), The redox aspects of lithium-ion batteries, Energy Environ. Sci., 18 (4), 1658–1672 (DOI: 10.1039/d4ee04560b)

(41) Hu Y.S., Pan H. (2022), Solvation structures in electrolyte and the interfacial chemistry for Na-ion batteries, ACS Energy Lett., 7 (12), 4501–4503 (DOI: 10.1021/acsenergylett.2c02529)

(42) Xu K. (2014), Electrolytes and interphases in Li-ion batteries and beyond, Chem. Rev., 114 (23), 11503–11618 (DOI: 10.1021/cr500003w)

(43) Cheng L., Assary R.S., Qu X., Jain A., Ong S.P., Rajput N.N., Persson K., Curtiss L.A. (2015), Accelerating electrolyte discovery for energy storage with high-throughput screening, J. Phys. Chem. Lett., 6 (2), 283–291 (DOI: 10.1021/jz502319n)

(44) Manna S., Manna S.S., Das S., Pathak B. (2023), Metal-solvent interaction contribution on voltage for metal ion battery: an interpretable machine learning approach, Electrochim. Acta, 467, 143148 (DOI: 10.1016/j.electacta.2023.143148)

(45) Kumar N., Siegel D.J. (2016), Interface-induced renormalization of electrolyte energy levels in magnesium batteries, J. Phys. Chem. Lett., 7 (5), 874–881 (DOI: 10.1021/acs.jpclett.6b00091)

(46) Goodenough J.B. (2013), Electrochemical energy storage in a sustainable modern society, Energy Environ. Sci., 7 (1), 14–18 (DOI: 10.1039/c3ee42613k)

(47) Borodin O. (2019), Challenges with prediction of battery electrolyte electrochemical stability window and guiding the electrode–electrolyte stabilization, Curr. Opin. Electrochem., 13, 86–93 (DOI: 10.1016/j.coelec.2018.10.015)

(48) Hausbrand R. (2020), Electronic energy levels at Li-ion cathode-liquid electrolyte interfaces: concepts, experimental insights, and perspectives, J. Chem. Phys., 152 (18) (DOI: 10.1063/1.5143106)

(49) Peljo P., Girault H.H. (2018), Electrochemical potential window of battery electrolytes: the HOMO–LUMO misconception, Energy Environ. Sci., 11 (9), 2306–2309 (DOI: 10.1039/c8ee01286e)

(50) Leung K. (2013), Electronic structure modeling of electrochemical reactions at electrode/electrolyte interfaces in lithium ion batteries, J. Phys. Chem. C, 117 (4), 1539–1547 (DOI: 10.1021/jp308929a)

(51) Wang A., Kadam S., Li H., Shi S., Qi Y. (2018), Review on modeling of the anode solid electrolyte interphase (SEI) for lithium-ion batteries, npj Comput. Mater., 4 (1), 1–26 (DOI: 10.1038/s41524-018-0064-0)

(52) Guo Y., Jin H., Qi Z., Hu Z., Ji H., Wan L.J. (2019), Carbonized-MOF as a sulfur host for aluminum–sulfur batteries with enhanced capacity and cycling life, Adv. Funct. Mater., 29 (7), 1807676 (DOI: 10.1002/adfm.201807676)

(53) Guo Y., Hu Z., Wang J., Peng Z., Zhu J., Ji H., Wan L.J. (2020), Aluminium-sulfur battery with improved electrochemical performance by cobalt-containing electrocatalyst, Angew. Chem. Int. Ed., 59, 22963–22967 (DOI: 10.1002/anie.202008481)

(54) Ng S.F., Lau M.Y.L., Ong W.J. (2021), Lithium–sulfur battery cathode design: tailoring metal-based nanostructures for robust polysulfide adsorption and catalytic conversion, Adv. Mater., 33 (50), 2008654 (DOI: 10.1002/adma.202008654)

(55) Zhang Q., Wang Y., Seh Z.W., Fu Z., Zhang R., Cui Y. (2015), Understanding the anchoring effect of two-dimensional layered materials for lithium-sulfur batteries, Nano Lett., 15 (6), 3780–3786 (DOI: 10.1021/acs.nanolett.5b00367)

(56) Klimpel M., Kovalenko M.V., Kravchyk K.V. (2022), Advances and challenges of aluminum–sulfur batteries, Commun. Chem., 5 (1), 1–7 (DOI: 10.1038/s42004-022-00693-5)

(57) Guo Y., Jin H., Qi Z., Hu Z., Ji H., Wan L.J. (2019), Carbonized-MOF as a sulfur host for aluminum–sulfur batteries with enhanced capacity and cycling life, Adv. Funct. Mater., 29 (7), 1807676 (DOI: 10.1002/adfm.201807676)

(58) Yu X., Manthiram A. (2017), Electrochemical energy storage with a reversible nonaqueous room-temperature aluminum–sulfur chemistry, Adv. Energy Mater., 7 (18) (DOI: 10.1002/aenm.201700561)

(59) Seh Z.W., Fredrickson K.D., Anasori B., Kibsgaard J., Strickler A.L., Lukatskaya M.R., Gogotsi Y., Jaramillo T.F., Vojvodic A. (2016), Two-dimensional molybdenum carbide (MXene) as an efficient electrocatalyst for hydrogen evolution, ACS Energy Lett., 1 (3), 589–594 (DOI: 10.1021/acsenergylett.6b00247)

(60) Naguib M., Halim J., Lu J., Cook K.M., Hultman L., Gogotsi Y., Barsoum M.W. (2013), New two-dimensional niobium and vanadium carbides as promising materials for Li-ion batteries, J. Am. Chem. Soc., 135 (43), 15966–15969 (DOI: 10.1021/ja405735d)

(61) Khazaei M., Ranjbar A., Arai M., Yunoki S. (2016), Topological insulators in the ordered double transition metals M_2MC_2 MXenes (M = Mo, W; M = Ti, Zr, Hf), Phys. Rev. B, 94, 125152 (DOI: 10.1103/PhysRevB.94.125152)

(62) Anand R., Ram B., Umer M., Zafari M., Umer S., Lee G., Kim K.S. (2022), Doped MXene combinations as highly efficient bifunctional and multifunctional catalysts for water splitting and metal–air batteries, J. Mater. Chem. A, 10, 23771–23784 (DOI: 10.1039/d2ta06297f)

(63) Bhauriyal P., Pathak B. (2020), Superior anchoring effect of a Cu-benzenehexathial MOF as an aluminium–sulfur battery cathode host, Mater. Adv., 1, 3572–3580 (DOI: 10.1039/d0ma00546k)

(64) Sim E.S., Yi G.S., Je M., Lee Y., Chung Y.C. (2017), Understanding the anchoring behavior of titanium carbide-based MXenes depending on the functional group in Li–S batteries: a density functional theory study, J. Power Sources, 342, 64–69 (DOI: 10.1016/j.jpowsour.2016.12.042)

(65) Manna S., Das A., Das S., Pathak B. (2024), Machine learning assisted screening of MXene with superior anchoring effect in Al–S batteries, ACS Mater. Lett., 6 (2), 572–582 (DOI: 10.1021/acsmaterialslett.3c01043)

(66) Born M., Oppenheimer J. (1927), Zur quantentheorie der molekeln, J. Ann. Physik, 84, 457 (DOI: 10.1002/andp.19273892002)

(67) Hohenberg P., Kohn W. (1964), Inhomogeneous electron gas, Phys. Rev. B, 136, B864-B871 (DOI: 10.1103/PhysRev.136.B864)

(68) Kohn W., Sham L. J. (1965), Self-consistent equations including exchange and correlation effects, Phys. Rev., 140, A1133-A1138 (DOI: 10.1103/PhysRev.140.A1133)

(69) Xia B. Y., Wu H. B., Wang X., Lou X. W. (2013), Index facets and enhanced electrocatalytic properties, Angew. Chem. Int. Ed., 52, 12337-12340 (DOI: 10.1002/anie.201307518)

(70) Martin, R. M. (2012), Electronic structure: basic theory and practical methods, Cambridge University press, ISBN: 9780511805769 (DOI: 10.1017/CBO9780511805769)

(71) Ceperley D. M., Alder B. (1980), Ground State of the Electron Gas by a Stochastic Method, Phys. Rev. Lett., 45, 566-569 (DOI: 10.1103/PhysRevLett.45.566)

(72) van de Walle A., Ceder G. (1999), Correcting over binding in local-density-approximation calculations, Phys. Rev. B, 59, 14992-15001 (DOI: 10.1103/PhysRevB.59.14992)

(73) Perdew J. P., Burke K., Ernzerhof M. (1996), Generalized gradient approximation made simple, Phys. Rev. Lett., 77, 3865-3868 (DOI: 10.1103/PhysRevLett.77.3865)

(74) Perdew J. P., Wang Y. (1992), Accurate and simple analytic representation of the electron-gas correlation energy, Phys. Rev. B, 45, 13244-13249 (DOI: 10.1103/PhysRevB.45.13244)

(75) Perdew J. P., Ruzsinszky A., Csonka G. I., Vydrov O. A., Scuseria G. E., Constantin L. A., Zhou X., Burke K. (2008), Restoring the density-gradient expansion for exchange in solids and surfaces, Phys. Rev. Lett., 100, 136406 (DOI: 10.1103/PhysRevLett.100.136406)

(76) Vanderbilt D. (1990), Soft self-consistent pseudopotentials in a generalized eigenvalue formalism, Phys. Rev. B, 41, 7892-7895 (DOI: 10.1103/PhysRevB.41.7892)

(77) Blochl P. E. (1994), Projector augmented-wave method, Phys. Rev. B, 50, 17953-17979 (DOI: 10.1103/PhysRevB.50.17953)

(78) Andersen, O. K. (1975), Linear methods in band theory, Phys. Rev. B., 12, 3060-3083 (DOI: 10.1103/PhysRevB.12.3060)

(79) Hamann D. R., Schlüter M., Chiang C. (1979), Norm-conserving pseudopotentials, Phys. Rev. Lett., 43, 1494-1497 (DOI:10.1103/PhysRevLett.43.1494)

(80) Mills G., Jónsson H. (1994), Quantum and thermal effects in H2 dissociative adsorption: Evaluation of free energy barriers in

multidimensional quantum systems, Phys. Rev. Lett. 72, 1124 (DOI:10.1103/PhysRevLett.72.1124)

(81) Grimme S., Antony J., Ehrlich S., Krieg H. (2010), A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, J. Chem. Phys. 132, 154104 (DOI: https://doi.org/10.1063/1.3382344)

(82) Grimme S. (2011), Density functional theory with London dispersion corrections, WIREs Comput. Mol. Sci., 1, 211-228 (DOI: 10.1002/wcms.30)

(83) Grimme S. (2004), Accurate description of van der Waals complexes by density functional theory including empirical corrections, J. Comput. Chem., 25, 1463-1473 (DOI:10.1002/jcc.20078)

(84) Chai J. D., Head-Gordon M. (2008), Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections, Phys. Chem. Chem. Phys., 10, 6615-6620 (DOI: 10.1039/B810189B)

(85) Jordan M.I., Mitchell T.M. (2015), Machine learning: trends, perspectives, and prospects, Science, 349 (6245), 255–260 (DOI: 10.1126/science.aaa8415)

(86) Hastie T., Tibshirani R., Friedman J. (2009), The elements of statistical learning, Springer (DOI: 10.1007/978-0-387-84858-7)

(87) Morgan D., Jacobs R. (2020), Opportunities and challenges for machine learning in materials science, Annu. Rev. Mater. Res., 50, 71–103 (DOI: 10.1146/annurev-matsci-070218-010015)

(88) Butler K.T., Davies D.W., Cartwright H., Isayev O., Walsh A. (2018), Machine learning for molecular and materials science, Nature, 559 (7715), 547–555 (DOI: 10.1038/s41586-018-0337-2)

(89) Joshi R.P., Eickholt J., Li L., Fornari M., Barone V., Peralta J.E. (2019), Machine learning the voltage of electrode materials in metal-ion

batteries, ACS Appl. Mater. Interfaces, 11 (20), 18494–18503 (DOI: 10.1021/acsami.9b04933)

(90) Moses I.A., Joshi R.P., Ozdemir B., Kumar N., Eickholt J., Barone V. (2021), Machine learning screening of metal-ion battery electrode materials, ACS Appl. Mater. Interfaces, 13 (45), 53355–53362 (DOI: 10.1021/acsami.1c04627)

(91) Jeschke S., Johansson P. (2021), Supervised machine learning-based classification of Li–S battery electrolytes, Batter. Supercaps, 4 (7), 1156–1162 (DOI: 10.1002/batt.202100031)

(92) Wang J., Wang X., Feng S., Miao Z. (2024), Studying the thermodynamic phase stability of organic–inorganic hybrid perovskites using machine learning, Molecules, 29 (13), 2974 (DOI: 10.3390/molecules29132974)

(93) Ward L., Dunn A., Faghaninia A., Zimmermann N.E.R., Bajaj S., Wang Q., Montoya J., Chen J., Bystrom K., Dylla M., Chard K., Asta M., Persson K.A., Snyder G.J., Foster I., Jain A. (2018), Matminer: an open source toolkit for materials data mining, Comput. Mater. Sci., 152, 60–69 (DOI: 10.1016/j.commatsci.2018.05.018)

(94) Jain A., Ong S.P., Hautier G., Chen W., Richards W.D., Dacek S., Cholia S., Gunter D., Skinner D., Ceder G., Persson K.A. (2013), Commentary: the Materials Project: a materials genome approach to accelerating materials innovation, APL Mater., 1 (1) (DOI: 10.1063/1.4812323)

(95) CrossRef (2000), Listing of deleted DOIs, CrossRef, 1 (DOI: 10.1162/153244303322753616)

(96) Tibshirani R. (1996), Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Series B Stat. Methodol., 58 (1), 267–288 (DOI: 10.1111/j.2517-6161.1996.tb02080.x)

(97) Breiman L. (2001), Random forests, Mach. Learn., 45 (1), 5–32 (DOI: 10.1023/a:1010933404324)

(98) Lundberg S.M., Lee S.I. (2017), A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst., 2017-December, 4766–4775

(99) Jollife I.T., Cadima J. (2016), Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci., 374 (2065) (DOI: 10.1098/rsta.2015.0202)

(100) Wold S., Sjöström M., Eriksson L. (2001), PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst., 58 (2), 109–130 (DOI: 10.1016/S0169-7439(01)00155-1)

(101) Goodfellow I., Bengio Y., Courville A. (2016), Deep learning, MIT Press (accessed 2025-07-24)

(102) Efron B. (1983), Estimating the error rate of a prediction rule: improvement on cross-validation, J. Am. Stat. Assoc., 78 (382), 316–331 (DOI: 10.1080/01621459.1983.10477973)

(103) Bishop C.M. (2006), Pattern recognition and machine learning, Springer (accessed 2025-07-24)

(104) Zhang Y., Yang Y. (2015), Cross-validation for selecting a model selection procedure, J. Econom., 187 (1), 95–112 (DOI: 10.1016/j.jeconom.2015.02.006)

(105) Arlot S., Celisse A. (2009), A survey of cross-validation procedures for model selection, Stat. Surv., 4, 40–79 (DOI: 10.1214/09-SS054)

(106) Stone M. (1974), Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. Series B Stat. Methodol., 36 (2), 111–133 (DOI: 10.1111/j.2517-6161.1974.tb00994.x)

(107) Willmott C.J., Matsuura K. (2005), Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Clim. Res., 30, 79–82 (accessed 2025-07-24)

(108) Chai T., Draxler R.R. (2014), Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 7 (3), 1247–1250 (DOI: 10.5194/gmd-7-1247-2014)

(109) Seber G.A.F., Lee A.J. (2012), Linear regression analysis, Wiley (DOI: 10.1002/9780471722199)

(110) James G., Witten D., Hastie T., Tibshirani R. (2021), An introduction to statistical learning, Springer (DOI: 10.1007/978-1-0716-1418-1)

(111) Hoerl A.E., Kennard R.W. (1970), Ridge regression: biased estimation for nonorthogonal problems, Technometrics, 12 (1), 55–67 (DOI: 10.2307/1267351)

(112) McDonald G.C. (2009), Ridge regression, Wiley Interdiscip. Rev. Comput. Stat., 1 (1), 93–100 (DOI: 10.1002/wics.14)

(113) Friedman J.H. (2001), Greedy function approximation: a gradient boosting machine, Ann. Stat., 29 (5), 1189–1232 (DOI: 10.1214/aos/1013203451)

(114) Chen T., Guestrin C. (2016), XGBoost: a scalable tree boosting system, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 13–17 August 2016, 785–794 (DOI: 10.1145/2939672.2939785)

(115) Murphy K.P. (1991), Machine learning: a probabilistic perspective, MIT Press, 73–78, 216–244

(116) Ikotun A.M., Ezugwu A.E., Abualigah L., Abuhaija B., Heming J. (2023), K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data, Inf. Sci., 622, 178–210 (DOI: 10.1016/j.ins.2022.11.139)

(117) Lloyd S.P. (1982), Least squares quantization in PCM, IEEE Trans. Inf. Theory, 28 (2), 129–137 (DOI: 10.1109/tit.1982.1056489)

(118) MacQueen J. (1967), Some methods for classification and analysis of multivariate observations, Proc. Fifth Berkeley Symp. Math. Stat. Probab., 5 (1), 281–298

(119) Ketchen D.J., Shook C.L. (1996), The application of cluster analysis in strategic management research: an analysis and critique, Strateg. Manag. J., 17 (6), 441–458 (DOI: 10.1002/(sici)1097-0266(199606)17:6<441::aid-smj819>3.0.co;2-g)

(120) Thorndike R.L. (1953), Who belongs in the family?, Psychometrika, 18 (4), 267–276 (DOI: 10.1007/bf02289263)

(121) Rousseeuw P.J. (1987), Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math., 20 (C), 53–65 (DOI: 10.1016/0377-0427(87)90125-7)

(122) Maaten L. van der, Hinton G. (2008), Visualizing data using T-SNE, J. Mach. Learn. Res., 9 (86), 2579–2605

(123) McInnes L., Healy J., Saul N., Großberger L. (2018), UMAP: uniform manifold approximation and projection, J. Open Source Softw., 3 (29), 861 (DOI: 10.21105/joss.00861)

(124) Eichner D., Cariello R., Lee J., Minyard D., Realuyo A., Bellocchio A. (2025), Advancing the design space of a tactical, air-launched balloon, Proc. AIAA Aviat. Forum Expo., AIAA

(125) Snedecor G.W., Cochran W.G. (1989), Statistical methods, 8th ed., Iowa State Univ. Press, Ames (accessed 2025-07-24)

(126) Bergstra J., Bengio Y. (2012), Random search for hyper-parameter optimization, J. Mach. Learn. Res., 13 (10), 281–305

(127) Pedregosa F., Michel V., Grisel O., Blondel M., Prettenhofer P., Weiss R., Vanderplas J., Cournapeau D., Varoquaux G., Gramfort A., Thirion B., Dubourg V., Passos A., Brucher M., Perrot M., Duchesnay É. (2011), Scikit-learn: machine learning in Python, J. Mach. Learn. Res., 12 (85), 2825–2830

(128) Hsu C.-W., Chang C.-C., Lin C.-J. (2009), A practical guide to support vector classification, Dept. of Computer Science, National Taiwan Univ.

(129) Kuhn H.W., Tucker A.W. (1951), Nonlinear programming, in: Neyman J. (Ed.), Proc. Second Berkeley Symp. Math. Stat. Probab., 481–492

(130) Behler J., Parrinello M. (2007), Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett., 98 (14), 146401 (DOI: 10.1103/physrevlett.98.146401)

(131) Chmiela S., Tkatchenko A., Sauceda H.E., Poltavsky I., Schütt K.T., Müller K.R. (2017), Machine learning of accurate energy-conserving molecular force fields, Sci. Adv., 3 (5), eaao1495 (DOI: 10.1126/sciadv.1603015)

(132) Deng B., Zhong P., Jun K., Riebesell J., Han K., J. Bartel C., and Ceder G. (2023), CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling, Nat. Mach. Intell., 5, 1031–1041 (DOI: 10.5281/zenodo.8173515)

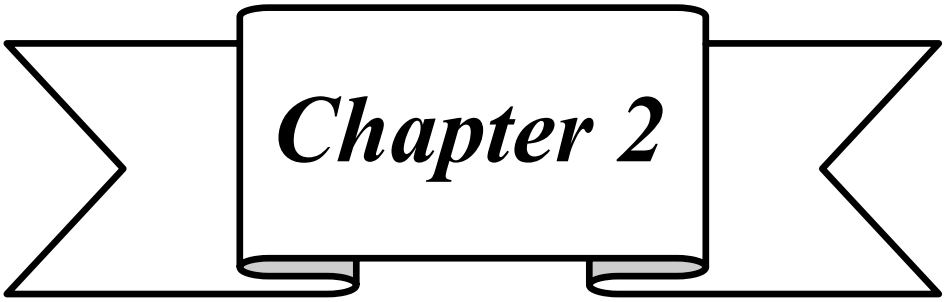
(133) Zhang L., Han J., Wang H., Car R., E W. (2018), Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics, Phys. Rev. Lett., 120 (14), 143001 (DOI: 10.1103/physrevlett.120.143001)

(134) Chen C., Ye W., Zuo Y., Zheng C., Ong S.P. (2019), Graph networks as a universal machine learning framework for molecules and crystals, Chem. Mater., 31 (9), 3564–3572 (DOI: 10.1021/acs.chemmater.9b01294)

(135) Fung V., Zhang J., Juarez E., Sumpter B.G. (2021), Benchmarking graph neural networks for materials chemistry, npj Comput. Mater., 7 (1), 84 (DOI: 10.1038/s41524-021-00554-0)

(136) Schütt K.T., Sauceda H.E., Kindermans P.J., Tkatchenko A., Müller K.R. (2018), SchNet - a deep learning architecture for molecules and materials, J. Chem. Phys., 148 (24), 241722 (DOI: 10.1063/1.5019779)

(137) Gasteiger J., Groß J., Günnemann S. (2020), Directional message passing for molecular graphs, Proc. 8th Int. Conf. Learn. Represent., ICLR



Specific Capacity Prediction of Cathode Materials for Li/Na/K ion Batteries

2.1. Introduction

With the increase in energy demands, harnessing energy from renewable sources has become increasingly important for sustainable development. However, renewable energy sources are intermittent in nature as they depend on factors like weather, location, efficiency, and available infrastructure. Thus, efficient large-scale energy storage systems, are required to store, transfer, and utilize the energy produced from renewable energy sources.*[1]* Rechargeable metal-ion batteries are used extensively to store energy in the form of chemical energy which can be converted back to electrical energy whenever required. Among all the metal-ion battery, Li-ion batteries (LIBs) are leading the energy storage devices market, especially in portable devices such as smartphones and laptops.*[2,3]* Li metal-ion batteries has even opened extraordinary possibilities in automotive sector and electric vehicles market recently.*[4,5]* The long cycle life, high efficiencies and high energy densities are the main reason behind the success of LIBs.*[2,6]* However, for large scale energy storage, LIBs have certain shortcomings such as its relatively low energy density, and safety issue owing its high reactivity in air.*[4,6–11]* Very low abundance of Li sources is also a major concern which ultimately contributes to the high price of these batteries.*[6,12,13]* These issues demand for cheap, efficient and sustainable alternatives of LIBs.

Potassium (K) is one of the metal ions which could replace lithium in energy storage devices. K is more abundant compared to Li sources and hence reduces the production cost.*[14]* K-ion batteries have a similar rocking chair mechanism like LIBs. K^+ having large atomic radius (1.38 Å) has a small Stokes radius in various organic electrolytes which results in higher ionic conductivity.*[15]* The standard potential of K/K^+ in nonaqueous medium specially in the most common solvent propylene carbonate is -2.88 V, which is more negative compared to Li and Na.*[16]* Organic electrode materials are also used as cathode material for K ion batteries.*[17]* K ions having low de-solvation energy possesses faster diffusion over the

electrode/electrolyte interface.*[18]* Readily available and cheaper electrolyte solutions and salt of K ion batteries is also a reason for their low price compared to LIBs. For example, KPF_6 is much cheaper than the similar analogous of Li and Na.*[19]* Currently extensive investigation is going on the layered transition metal oxides cathode materials having larger interlayer distance and diffusion path for K ion battery.*[20]* Adopting K ion battery technology can also lead to production of cheaper Co-free batteries, often based on transition metals such as Fe, Mn, and V.*[21]* Though the electrode materials for LIBs have been extensively explored, but the same is not true for K-ion batteries. However, seeking suitable electrode materials for K ion battery is experimentally challenging and even theoretically, requires high computational facilities. Majority of the electrode materials used for LIBs are still unexplored for K-ion battery due to the difficulties in experimental and computational screening of large number of electrode materials with high accuracy.*[22–24]* Therefore, machine learning (ML) could be an advanced tool which can save both time and cost, and at the same time screen many electrodes with minimum computational cost. Among different factors affecting the success of ML, data takes the central position. Every ML model depends on the amount and quality of data needed for training. Taking advantage of different computational material database like Materials Project, OQMD, AFLowLib, ESP, CMR, NOMAD, applications of ML for determining battery properties are increasing day by day. Besides computational data, ICSD and COD provides data from published literature, while NASA battery datasets contains experimental battery data and so on.*[25–32]* Although use of DFT based data is not standard for every context, still it delivers sensible insights which ultimately helps in the guidance of experimental research.*[33,34]* ML combined with data from various databases can be used for predicting any specific property of interest for a particular battery material.*[23,29,35–39]* Application of ML in the field of material science can be found in the prediction of microscopic properties like band structure, formation energy, density of

states, etc. which play an essential role in research areas like solar cells, batteries, and catalysis.*[40–55]* Kernel ridge regression (KRR) and support vector regression (SVR) has been used by Seko et al. for the prediction of thermal conductivity and cohesive energy of binary and tertiary compounds.*[56,57]* ML techniques have also been used for the prediction of different properties for their applicability as materials in photovoltaic cells and glass alloys.*[58–60]* Application of ML on battery systems was first carried out by Salkind et al., predicting state of charge and state of health and from then onwards investigation on the application of ML in battery monitoring has continued.*[61]* Siqui Shi and coworkers have predicted the activation energy in cubic Li-argyrodites with hierarchically encoding crystal structure based descriptors.*[62]* Application of ML has also been reported for the determination of interphase stability of Li doped $Li_7La_3Zr_2O_{12}$.*[63]* Sendek and coworkers have used the logistic regression for screening of 12000 Li containing solids as solid-state electrolytes for LIBs by rapid screening.*[52]* ML has been also applied for the identification of chemical factors and descriptors affecting reaction kinetics of Li battery.*[64,65]* Meredig et al. built ML model to estimate thermodynamic stability and proposed around 4500 stable novel materials.*[66]* Multilayer automated feature selection method has been reported to incorporate expert knowledge.*[67]* ML has also been used for an alternative as well as faster method than DFT, prediction of thermal, electronic and mechanical properties.*[29,68–70]* However, certain challenges exist in applying ML to materials research such as contradictions between high dimension and small sample data, conflict and compromise between complexity and accuracy of machine learning models, and inconsistency and collaboration between learning results and domain expert knowledge.*[71,72]* Method development and guidelines for different ML-based publications highlighting supervised learning and its interpretability has been elaborated recently by Rodrigues and co-workers.*[73]*

Capacity is one of the important metrics for the measurements of battery performance. The longevity of a battery mainly depends on cycle life of a battery and the former directly related to the capacity of a battery. The capacity can be calculated from the number of ions intercalated in electrode materials and in order to do so quantum mechanically, we need to perform time consuming DFT calculations for each individual electrode material. However, we can utilize the different advance ML model as a tool to speed up the screening of electrode materials based on capacity as target variable. Very few studies have been carried out on the experimental capacity prediction on the basis of cycle life via ML for a particular electrode material.*[74,75]* In a recent report, ML has been used for the prediction of voltage for large number of electrode materials for metal ion battery.*[76]* They have considered both low and high metal ion concentration. However, we want to calculate specific capacity of non-intercalated systems by learning from known electrode materials without the help of high ion concentration. In this study we have utilized the Li, Na, and K ion battery data for the training of ML models in order to predict the capacity of those electrode materials for the K ion battery. To the best of our knowledge, this is the first work regarding prediction of theoretical capacity on the basis of structure of electrode material for metal ion batteries. Here we have directly predicted the capacity of a non-intercalated electrode material without knowing the number of K ion getting intercalated i.e., without doing any DFT calculation. The capacity of different electrode materials varies rapidly, and the range of minimum capacity and maximum capacity is very high. Keeping that in mind, we have only considered the monovalent ions and not bivalent and trivalent ions for intercalation. We have also not considered the lower alkali metal ions since the radius of those ions will increase as we go down the group. Among the metal ion batteries, LIBs have been explored extensively, however, experimentally or by DFT calculation testing all those LIBs electrode materials for K ion batteries is a lengthy process. Therefore, after considering the Li, Na and K ion battery

data as training set, we have replaced the Li and Na by K for an approximate estimation of capacity with the help of different ML models. Here, we have used Support Vector Machine (SVM), ExtraTrees Regression (EXR) and Kernel Ridge Regression (KRR) to fit training dataset. With addition to our particular interest i.e., capacity, we further used the predicted capacity for the calculation of number of K ion that could be intercalated in the electrode materials for LIBs and sodium ion batteries.

2.2. Data Preprocessing

2.2.1. Data and Features

In Materials Project database, there are many instances of same electrode with different number of intercalated ion and capacity. However, we have considered only the non-intercalated system for a particular ion intercalation with maximum capacity as the target variable, so that machine can learn about the capacity by maximum intercalation of specific ions for a fixed electrode material.

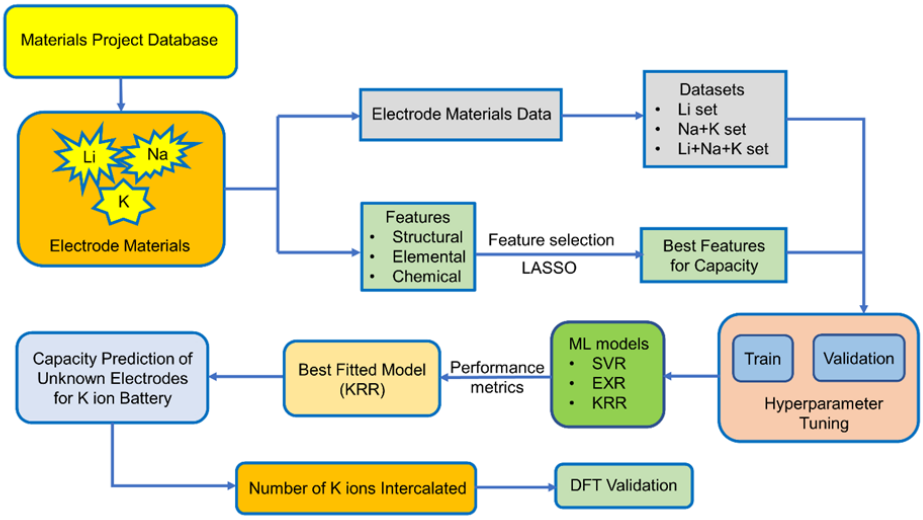


Figure 2.1: Illustration of the systematic steps followed for the prediction of specific capacity using machine learning models.

The performance of different ML model was assessed by mean absolute percentage error (MAPE). DFT calculation for few unknown electrode

materials has been performed to validate the machine learning model. We have provided a schematic diagram **(Figure 2.1)**, which shows the steps followed in our work. The training data for these metal ion batteries has been retrieved from the Materials Project database.*[30,77]* Overall, 2118 data have been considered in the training set, among which 69.53% of Li, 22.41% of Na and 8.06% of K ion battery data. We have excluded the repeating formula unit cells, as we are considering non-intercalated electrode materials for learning about the capacity. From the dataset, the ML model may overconcern with LIBs electrode materials, and ignore some knowledge about those for Na/K ion since the contribution in the overall data from LIBs is very high compared to other two metal ion batteries. The overall known dataset is divided in training set and validation set. The training set has been used to train the ML model whereas the validation set was used to validate the performance of our ML models. The validation set is composed of 20% of total data and rest of the data is used for training. The training set remains unique for all the ML model used and same is true for the validation set. The amount of Li, Na and K ion data used for training has been shown in **Figure 2.2**. To describe the electrode materials, we have generated 196 unique elemental features depending on chemical formula of individual electrode materials using choice-based feature vectorization.*[78]* Along with these features, other structural parameters such as lattice parameter (a, b, c), lattice angle (α, β, γ), volume of void, have also been considered, so that these features can represent each electrode materials uniquely. The volume of void (V_{void}) can be calculated as,

$$V_{void} = V_{crystal} - (\sum_{i=1}^{n} x_i \frac{4}{3}\pi r_i^3)$$

Where, $V_{crystal}$ is the volume of the lattice, i stands for the constituent elements in the lattice, xi stands for the number of atoms of the constituent element i, r_i stands for atomic radius of the constituent elements. In order to

specify the intercalated ion, we have also included some elemental properties like ionic radius, ionization energy, heat of atomization, among others. After the generation of features, scaling has been performed on each descriptor except on target variable using StandardScaler module of python package to bring down the all the features in the same scale to avoid the biasness of our data set based on the magnitude of each descriptor of electrode materials by the machine.

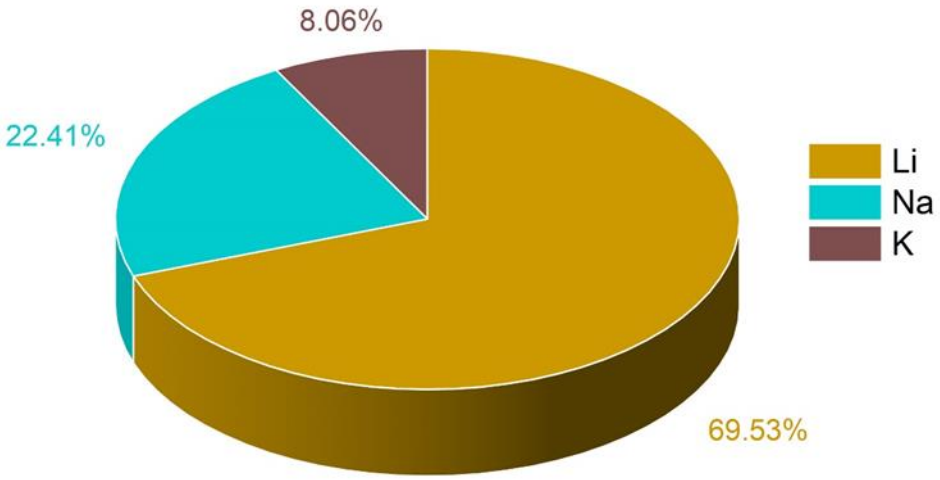


Figure 2.2: Distribution of different metal ions battery data used in machine learning model for training.

2.2.2. Feature Elimination

Further, to check the importance of considered features towards the desired target variable, we have performed Lasso Regression. Using Lasso regression, we have calculated the feature importance of each individual features and based on the magnitude of the feature importance we have screened the features. We have only considered the features having feature importance greater than zero and eliminated the rest during the fitting of ML models. The mathematical expression of LASSO Regression is given as,

$$\sum_{i=1}^{n}(y_i - y_i')^2 + \lambda m$$

where, y is the actual value and y' is the value of best fitted line. The value of y_i and y_i' varies from i =1 to n, where n is the number of observations. Using Lasso regression, we have calculated the slope (m) value for each descriptor. λ is a constant and considering its value equal to one, the slope for each descriptor has been established. The features having m value equal to zero were considered as irrelevant and those features were dropped. As a result, the number of considered features has shrunk from 199 to 71 i.e., 36% of features remain after Lasso regression. The list of selected features based on feature importance can be found in the Chapter-2 of Github repository (https://github.com/Souvik-ml/Thesis_data). It has been observed that lattice parameters (a, b, c), S orbital contribution, average number of valence electrons, volume of void, Allred Rochow electronegativity etc., contributed more towards the target variable specific capacity. Variance in Allred Rochow electronegativity and average number of valence electrons are found to be among the most important features.

2.3. Results and Discussion

2.3.1. Feature Correlation

For finding out the correlation among the features the heat map is generated as shown in **Figure 2.3** using the correlation function from the seaborn library. From the correlation values of different features, most of the features are found to be independent of each other. Some features are positively correlated while some show negative correlation. For example, Variance in Allred Rochow electronegativity is negatively correlated whereas average valence electrons are found to be positively correlated with target variable (capacity). Hence, the choice of these features will be able to represent each electrode materials uniquely. Though few elemental features are found to be dependent on each other, those features are not dropped so as to represent the intercalating ions.

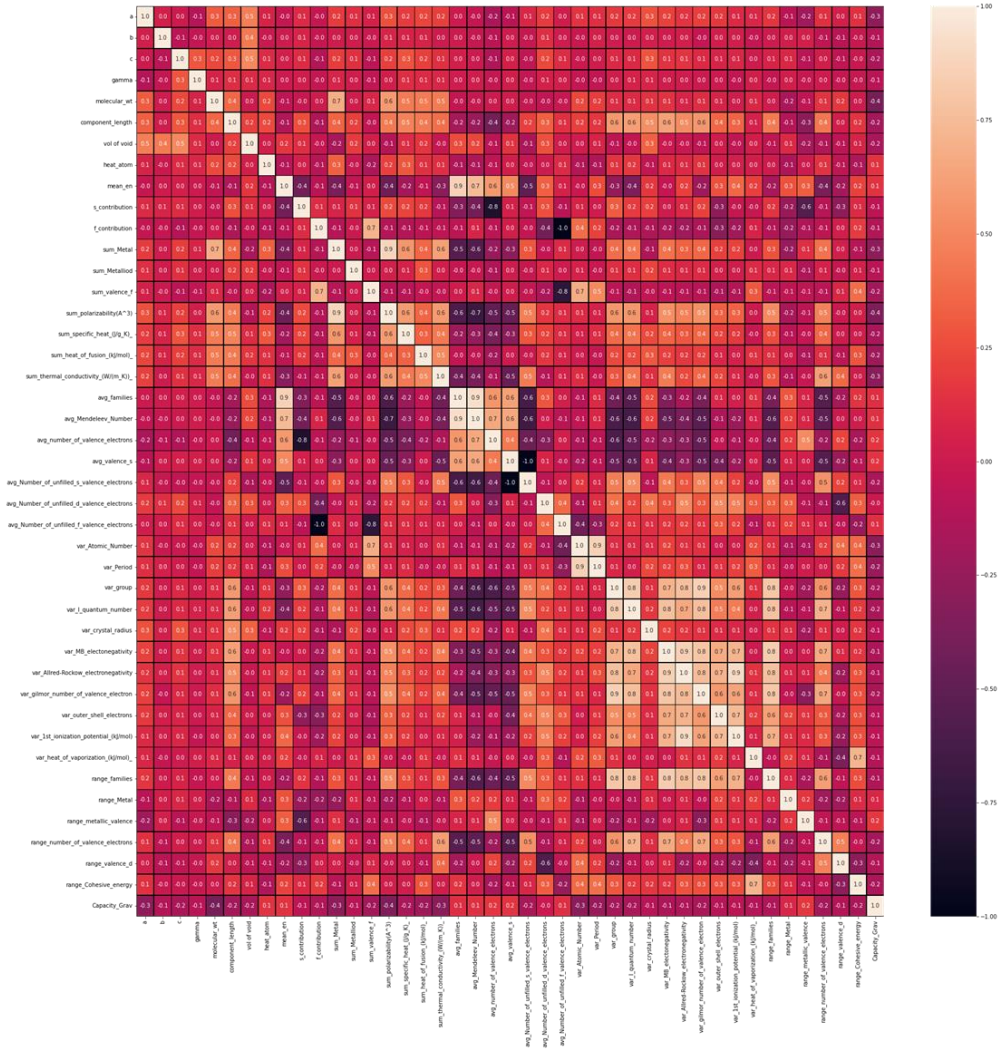


Figure 2.3: Heatmap showing the correlation among the considered features.

The elemental descriptors considered for the generation of input features are provided in the chapter-2 of Github repository (https://github.com/Souvik-ml/Thesis_data). Understanding the nature of dependence of features is highly important and, in this regard, joint plot helps us to find out the density of features. From the **Figure 2.4**, it is observed that molecular weight and polarizability follow almost same trend whereas Pauling electronegativity values are diverse at higher magnitude.

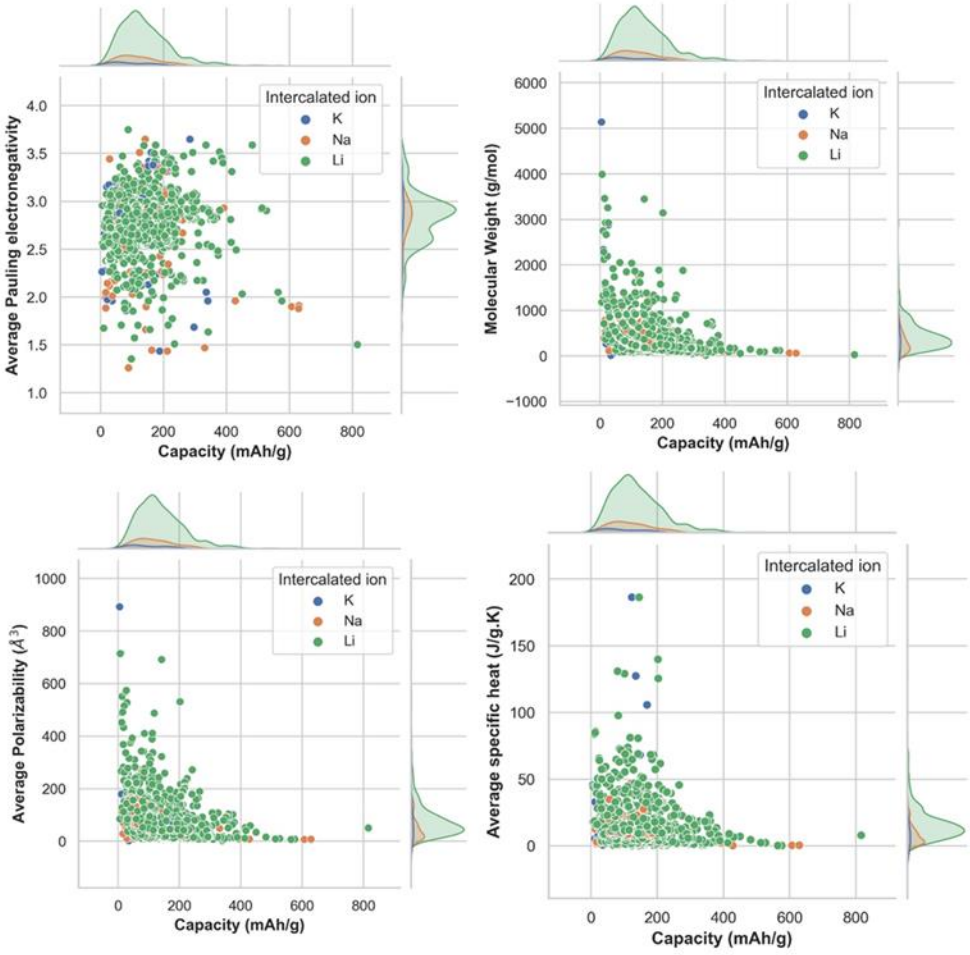


Figure 2.4: Joint plots for the density and distribution of capacity with respect to molecular properties, (a) average Pauling electronegativity, (b) molecular weight, (c) average polarizability, and (d) average specific heat, of constituent elements in the electrode material formula unit.

There is an indirect correlation between molecular weight and polarizability as electron density increases with the increase in atomic mass.*[79]* Therefore, it is very likely to observe the similar trend between these two parameters. Any trend in change of capacity with respect to average Pauling electronegativity could not be identified. The high range of electronegativity for most of the electrode materials may be the cause for this. To understand how the capacity of electrode materials change with the change in the electronic properties of those materials, average s, d, f valence electrons have been calculated by taking the average of the valence

electrons of the constituent atoms of electrode materials which is then plotted as the contribution of valence electrons with change in specific capacity. From the **Figure 2.5**, it is evident that the average capacity values fall in the range of higher s electron contribution, whereas in case of d orbital valence electron contribution, the capacity values are on the lower side. Thus, large number of electrode materials have more s orbital valence electrons and fewer d orbital valence electrons.

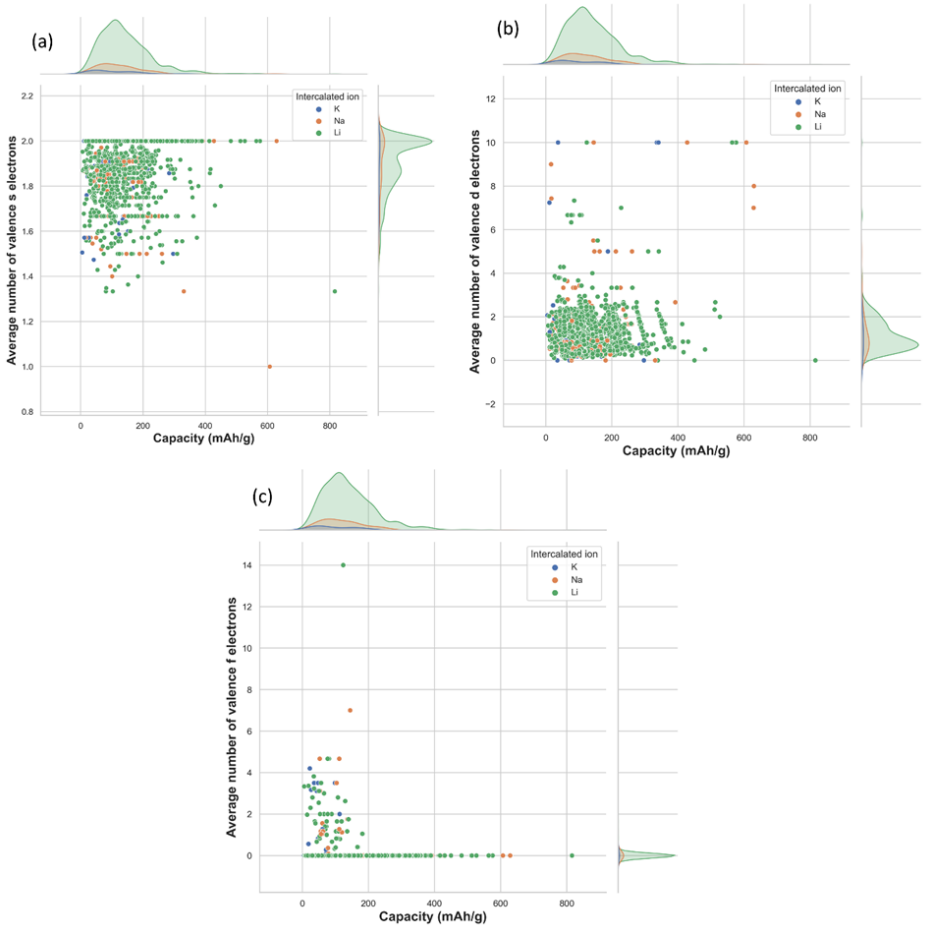


Figure 2.5: Joint plots for the distribution plot across electronic properties. Change of capacity with (a) s valence electrons, (b) d valence electrons, and (c) f valence electrons.

However, the capacity range varies from low magnitude to high magnitude of f-orbital valence electron. The reason behind the observed phenomenon

may be that most of the electrode materials in our data consists of transition metals with valence d orbital electrons and filled s orbital electrons. In the **Figure 2.6**, the distribution of capacity with the different lattice parameters (a, b, c) and lattice angle γ, of electrode materials has been presented. From the plot, the distribution range of a and c are found to be wide whereas the distribution range of the lattice parameter b is found to be very limited. The capacity is found to vary considerably for similar values of lattice parameter b. Similarly, with change in gamma the capacity is found to change without any uniform trend. Though the capacity values distribution is dispersed with respect to the lattice parameter, still the consideration of lattice parameters as descriptors is important to include the domain knowledge and structural properties of the various electrode materials. For instance, there are multiple unit cell structures for the same electrode material compound in the Materials Project database and hence, lattice parameters as descriptors help in distinguishing them.

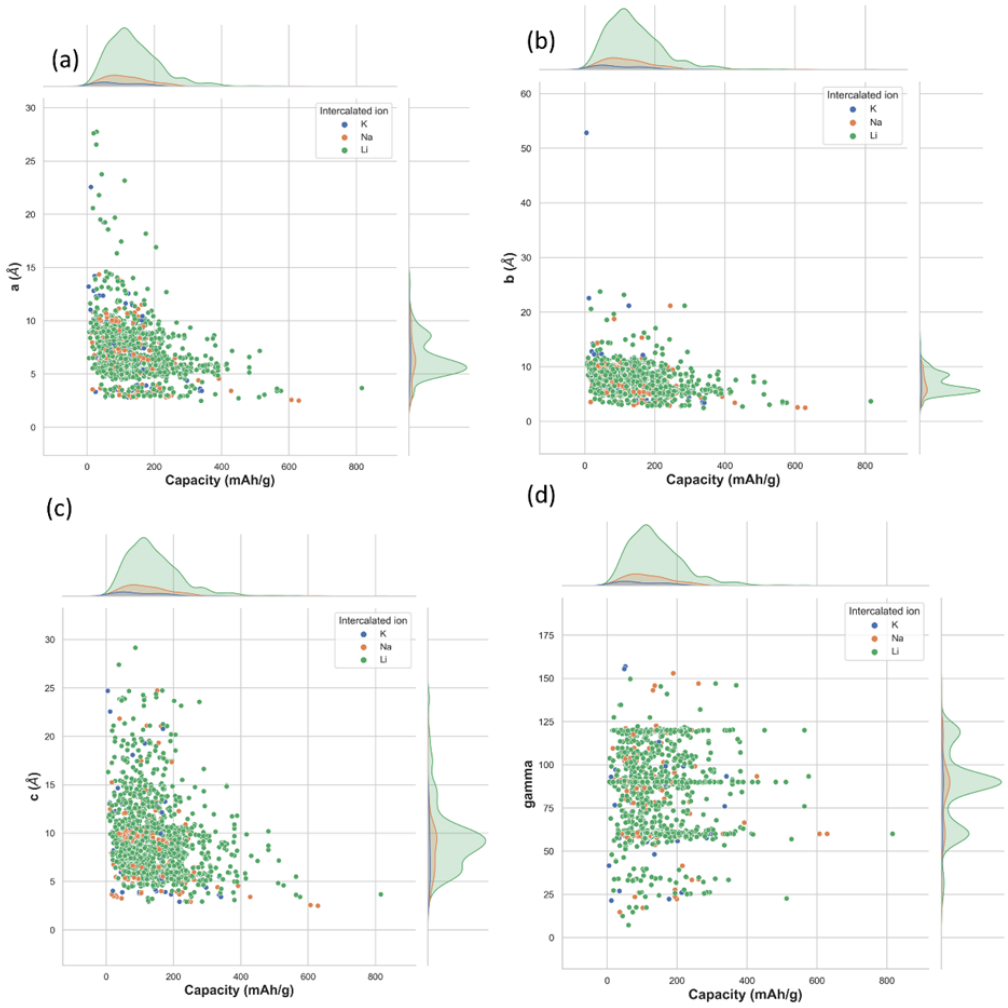


Figure 2.6: Distribution of capacity with respect to different lattice parameters of electrode materials. Change in capacity with lattice parameter (a) a, (b) b, (c) c, (d) γ.

The box plot between gravimetric capacity and the ionic radius of intercalated ions has been presented in **Figure 2.7**. The mid-line in the box plot is the median, lower line outside the box is the minimum range and the upper line outside the box is the maximum range of our property of interest. The average capacity for Li ion battery is found to be higher followed by Na and K ion batteries. With increase in ionic radius, the number of intercalated ions within an electrode material is expected to decrease and so is the specific capacity of electrode material.

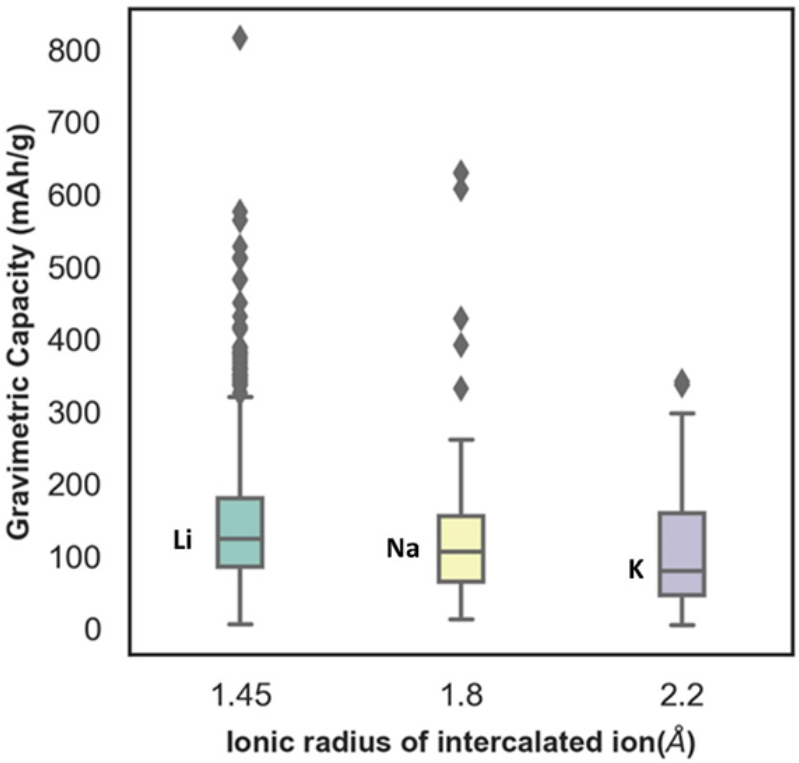


Figure 2.7: Distribution of specific capacity across the ionic radius of Li, Na and K where the Li, Na and K having ionic radius 1.45 Å, 1.8Å and 2.2Å respectively are represented by the first, second, and third box of the boxplot.

Since the target variable capacity varies rapidly with a slight change in the electrode material, so in order to understand the distribution of the electrode materials across the capacity we have plotted the range of % electrodes across per 100 mAh/g intervals of capacity as represented in **Figure 2.8**.

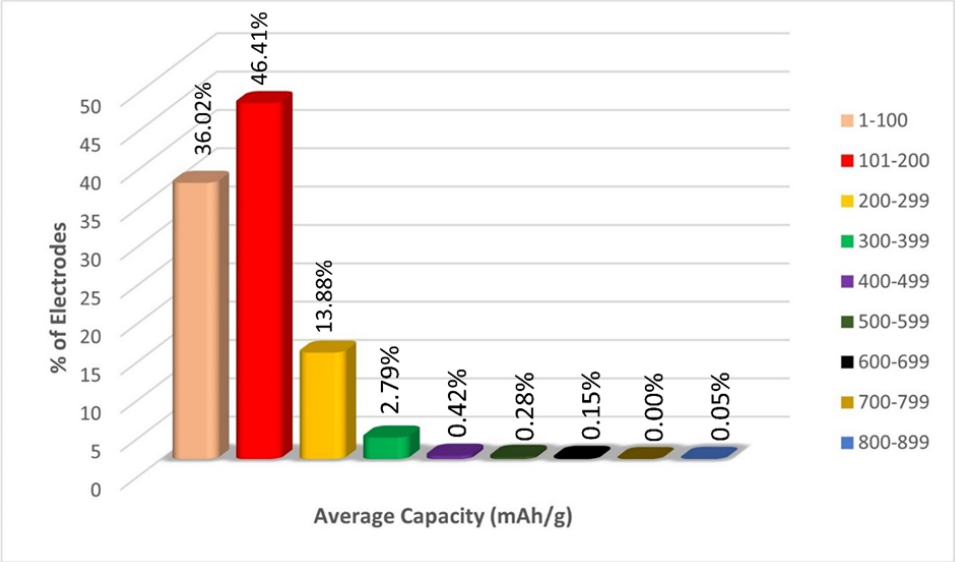


Figure 2.8: Distribution of capacity range across different electrode materials.

From **Figure 2.8**, we can observe more than 44% electrodes lie in the capacity range 101 to 200 mAh/g, around 35% electrodes lie in the capacity range 1 to 100 mAh/g and 15% electrodes have capacity 200 to 299 mAh/g. % of electrodes having capacity greater than 299 mAh/g is very less compared to the first three group of capacity ranges. Therefore, the sampling of target variable is highly heterogeneous which might cause a misinterpretation in the nature of data by machine which may lead to overestimation of capacity data in the range of 1 to 299. To avoid this overestimation, we fit the ML models in three different data set Li, Na+K, and Li+Na+K which has been discussed later.

2.3.2. Machine Learning

The analysis of our data set begins with fixing the target variable as capacity. The overall data set has been split into two set, training set composed of 80% of data and validation set composed of 20% data. Here we have compared three different machine learning algorithms namely, Support Vector Machine (SVM), ExtraTress Regression (EXR) and Kernel Ridge Regression (KRR). After analyzing the complexity of the dataset, we

have chosen these non-linear ML algorithms to fit the data. Different types of non-linear kernels present in each of these two ML algorithms like Radial basis function (rbf), polynomial, Laplacian have been used for the training of machines. Since our data set is medium in size, KRR comes very handy as it is an advanced ML algorithm compared to SVR having an additional parameter namely kernel trick. The KRR fitting is much faster compared to the fitting of the SVR. Therefore, first we have fitted our dataset with an SVR algorithm to see the performance. Further we have checked using KRR and our results show that the KRR fitted well compared to SVR. Ensemble-based ML algorithm namely ExtraTrees Regressor which takes decisions from the combination of a large number of decision trees has also been applied on the training data set. This algorithm has been chosen as it divides the whole data into a further small dataset and each model predicts some different values and the result is basically the average result of each model. These models have been found to be applied on capacity prediction for a particular cycle life of electrode materials.*[80]*

2.3.3. Hyperparameter Tuning

Since we are predicting continuous value via ML, therefore it belongs to a regression problem and hence we have used Support Vector Regression (SVR) which is a sub part of SVM. SVR contains two important parameters, C (penalty term) and gamma which should be optimized before fitting our dataset in SVR model. We have also tested our training data considering different kernel function within SVR like linear function, radial basis function (RBF) and polynomial function to select the most optimized kernel through the assessment of loss function as mean absolute percentage error (MAPE). As discussed earlier, the large contribution of LIBs in overall dataset might result in mimicking of Li data, the data set has been divided in three sets Li, Na+K, and Li+Na+K dataset. The training set is divided into 10 folds so that for each cross-validation test, 9 folds are used for the training whereas the remaining 1-fold is used for the assessment of the model performance in terms of MAPE as loss function. The standard

deviation for each 10-fold cross validation set has also been calculated. By the cross validation test we have tried to sample our data in such a way so that machine does not overfit certain data which could lead to a good training score but a very bad test score. The details regarding testing of different kernels for SVR are shown in **Table 2.1**. Among different kernels, RBF kernel function is found to fit well with less error as we have assessed our SVR model performance by checking the cross-validation score (CV_i).

Table 2.1: 10-fold cross validation (CVi) score, standard deviation (SD), Mean absolute percentage error (MAPE) on full data set (Li+Na+K) having different kernel of Support vector regression (SVR).

CV_i	**Linear**	**RBF**	**Polynomial**
CV_1	0.62	0.46	0.60
CV_2	0.36	0.36	0.39
CV_3	0.37	0.32	0.64
CV_4	0.49	0.33	0.55
CV_5	0.46	0.35	0.46
CV_6	0.38	0.40	0.44
CV_7	0.32	0.23	0.44
CV_8	0.29	0.19	0.31
CV_9	0.29	0.20	0.43
CV_{10}	0.36	0.35	0.53
SD	**0.10**	**0.08**	**0.10**
Mean MAPE	**0.39**	**0.32**	**0.48**
$MAPE_V$	**0.31**	**0.24**	**0.40**

The cross validation has been also performed for all three different dataset and shown in **Table 2.2.** The Mean MAPE shows the error on training set whereas MAPEV shows the error on validation set. For all the three dataset the similar trend has been observed with respect to training error and validation error. Though lesser error is expected for Na+K data set as it

contains lowest number of data points however, the less error for Li+Na+K data set compared to other two dataset evidences a better sampling of the data. The standard deviation for the dataset containing Li, Na and K data is lower compared to dataset having Na and K data which indicates that in overall dataset the deviation mainly arises from the Na+K data and not from the Li data. However, this data set can play an important role in the prediction of the capacity for the K-ion battery as Na is the closer element of K in the alkali metal group compared to Li.

Table 2.2: 10-fold cross validation (CV_i), standard deviation (SD), Mean absolute percentage error (MAPE) on three different dataset (Li+Na+K, Na+K, Li) having RBF kernel of Support vector regression (SVR).

CV_i	**Li (C=100, gamma=0.05)**	**Na+K (C=75, gamma = 0.01)**	**Li+Na+K (C=100, gamma=0.05)**
CV_1	0.44	0.54	0.46
CV_2	0.36	0.50	0.36
CV_3	0.35	0.25	0.32
CV_4	0.31	0.38	0.33
CV_5	0.40	0.24	0.35
CV_6	0.32	0.18	0.40
CV_7	0.25	0.12	0.23
CV_8	0.21	0.19	0.19
CV_9	0.19	0.18	0.20
CV_{10}	0.34	0.23	0.35
SD	**0.08**	**0.14**	**0.08**
Mean MAPE	**0.32**	**0.28**	**0.32**
$MAPE_V$	**0.26**	**0.28**	**0.23**

The hyperparameter tuning on C and gamma for SVR ML model has been illustrated in **Figure 2.9(a, b)**. The best parameters for Li+Na+K and Li data are found to be same whereas for Na+K data is different **(Table 2.2)**. After finding out the best hyperparameters for three different data set we have fitted SVR model on the training set and then validated the model in terms of MAPE utilizing the validation set.

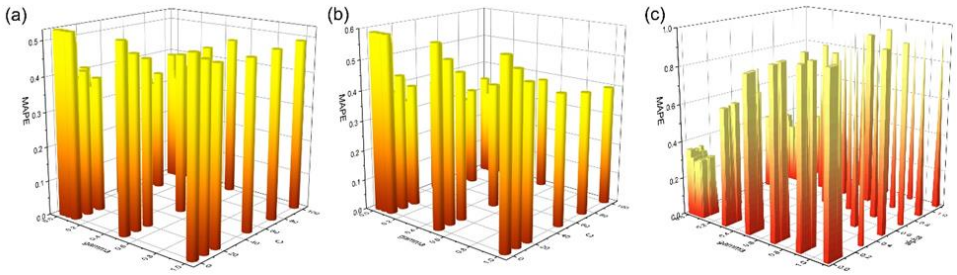


Figure 2.9: (a) Tuning of C and gamma parameter for Li+Na+K data set for SVR ML model. (b) Tuning of C and gamma parameter for Na+K data for SVR ML model. (c) Tuning of alpha and gamma parameter for Li+Na+K data set for KRR ML model.

2.3.4. Parity Plot

The comparison between DFT calculated capacity and SVR Predicted capacity has been shown in **Figure 2.10.** We have plotted the DFT calculated capacity vs ML predicted capacity using the best hyperparameters of SVR for all three different datasets.

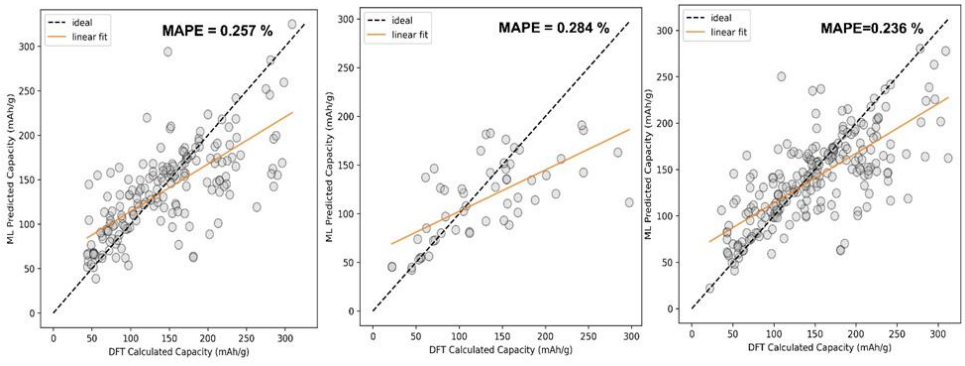


Figure 2.10: Comparison between ML predicted capacity and DFT calculated capacity after fitting SVR ML model using. (a) RBF kernel,

C=100, gamma=0.05 hyperparameters on Li dataset. (b) RBF kernel, C=75, gamma=0.01 hyperparameters on Na+K dataset (c) RBF kernel, C=100, gamma= 0.05 hyperparameters on Li+Na+K dataset.

Similarly, we have fitted our dataset in a tree-based ML model, ExtraTrees Regression (EXR). The cross-validation score using EXR algorithm has been presented in **Table 2.3**. The number of trees and other parameters are optimized before fitting the EXR ML model. However, the optimized parameters remain same for all three different datasets. We have found the same error trend as the SVR model. The Li+Na+K data has given less error on validation set compared to other two data sets. As we have compared the performance of SVR model in three different sets here also we have plotted the same plot using EXR ML model after fitting **(Figure 2.11)**.

Table 2.3: Cross validation score (CV_i), standard deviation (SD), Mean MAPE on training set and MAPE on validation set using EXR ML model.

CV_i	**Li**	**Na+K**	**Li+Na+K**
CV_1	0.43	0.60	0.47
CV_2	0.37	0.31	0.33
CV_3	0.32	0.16	0.26
CV_4	0.36	0.24	0.32
CV_5	0.37	0.17	0.34
CV_6	0.31	0.16	0.39
CV_7	0.21	0.14	0.23
CV_8	0.20	0.16	0.19
CV_9	0.23	0.21	0.25
CV_{10}	0.38	0.25	0.33
SD	**0.08**	**0.14**	**0.08**
Mean MAPE	**0.32**	**0.24**	**0.31**
$MAPE_V$	**0.26**	**0.28**	**0.24**

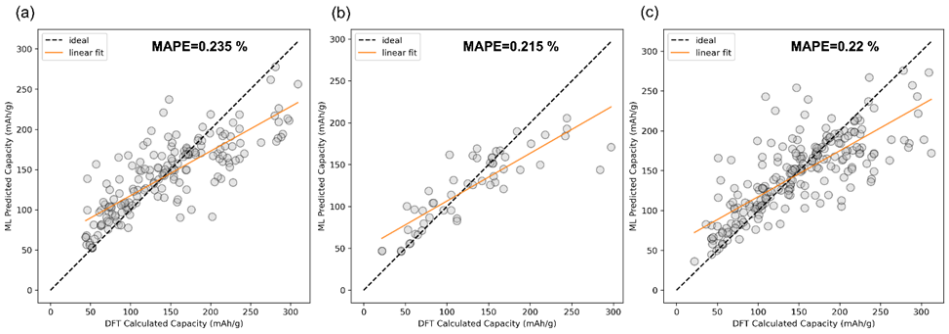


Figure 2.11: Comparison between ML predicted capacity and DFT calculated capacity for EXR ML model having number of trees=800, min_samples_leaf=3, min_samples_split=2 hyperparameters on (a) Li dataset. (b) Na+K dataset. (c) Li+Na+K dataset.

Table 2.4: MAPE distribution of capacity, standard deviation (SD), Mean MAPE on training set and MAPE on validation set (MAPEV) for 10 folds of training (CV_i) in KRR ML model trained with Na+K, Li, Li+Na+K data.

CV_i	**Li**	**Na+K**	**Li+Na+K**
CV_1	0.42	0.50	0.40
CV_2	0.33	0.19	0.30
CV_3	0.31	0.16	0.22
CV_4	0.33	0.19	0.32
CV_5	0.36	0.12	0.34
CV_6	0.32	0.14	0.40
CV_7	0.24	0.19	0.25
CV_8	0.20	0.19	0.18
CV_9	0.22	0.13	0.23
CV_{10}	0.38	0.24	0.33
SD	**0.07**	**0.11**	**0.07**
Mean MAPE	**0.31**	**0.21**	**0.30**
MAPEv	**0.24**	**0.15**	**0.21**

Furthermore, KRR has been used for the fitting of the data where we have again checked 10-fold cross validation result after choosing the optimized hyperparameters. The result of 10-fold cross validation test is shown in **Table 2.4**. Among all these three ML algorithms, KRR has fitted the Na+K data well compared to the others having MAPEV of 0.153%. Gamma and alpha are two important parameters for KRR algorithm. The optimization of these parameters is shown in **Figure 2.9(c)**.

The comparison between DFT calculated capacity and KRR predicted capacity for three different datasets has been displayed in **Figure 2.12**. Though the trend in training error and validation error in KRR is slightly different from the SVR and EXR ML model, the overall performance of KRR ML model is better than the rest of two as KRR is able to mimic the nature of Na+K data better which is more important than to mimic Li ion data considering our goal of predicting the capacity for K ion battery electrode materials. Therefore, from overall analysis on different dataset it is evident that KRR performs better than other considered models.

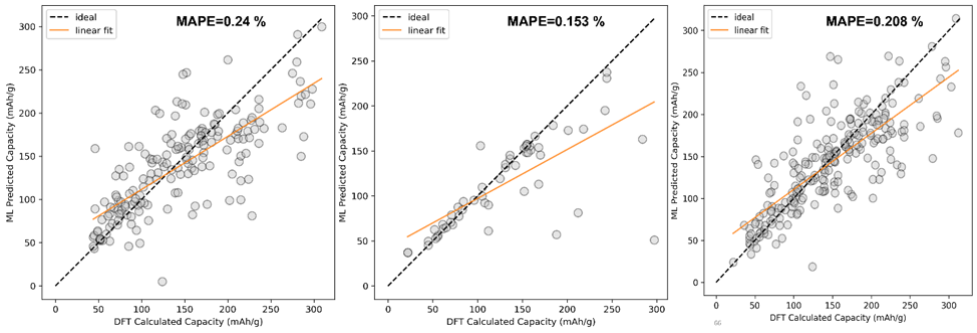


Figure 2.12: Comparison between ML predicted capacity and DFT calculated capacity after fitting KRR ML model (kernel=Laplacian, alpha=0.024239, gamma=0.047051, degree=2 hyperparameters) on (a) Li dataset, (b) Na+K dataset. (c) Li+Na+K dataset.

We have also fitted random forest regression (RFR) 470 number of optimized trees with a average cross validation MAPE of 0.39 **(Figure 2.13, Table 2.5)**. Similarly, decision tree regression model has been also fitted

with optimized hyperparameters to predict capacity having average MAPE of 0.38 **(Table 2.6)**.

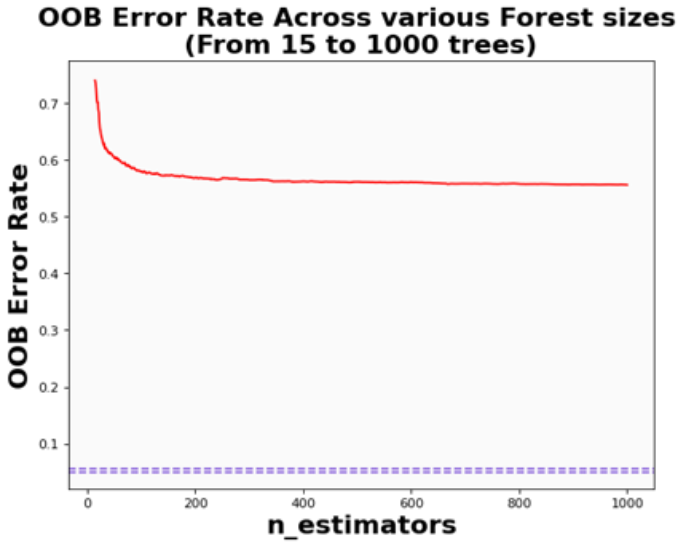


Figure 2.13: Estimation of optimized number of trees for Random Forest ML model.

Table 2.5: Cross validation score for Random Forest Regression (Number of Trees = 470).

Number of CV fold	**MAPE**
1	0.56
2	0.36
3	0.48
4	0.38
5	0.38
6	0.30
7	0.23
8	0.45
9	0.31
10	0.47
Mean MAPE	**0.39**

Table 2.6: Optimized hyperparameters and mean absolute percentage error (MAPE) for Decision Trees Regression.

Max_depth	Min_samples_leaf	Min_samples_split	splitter	MAPE
17	3	2	Random	0.38

2.4. DFT Validation

To validate our calculated capacity for various electrode materials, we have considered five structurally different sample electrode materials (Mn_4NiO_8, FeO_2, $Fe(CoO_3)_2$, V_5O_{12} and $CoPO_4$) and checked their maximum specific capacity by carrying out first principles calculations using the projector augmented wave (PAW) method as implemented in the Vienna Ab initio Simulation Package. [81–86] The selected materials are taken in such a way that they belong to different crystal lattice structures and stoichiometry of constituent elements. Moreover, the generalized gradient approximation of Perdew–Burke–Ernzerhof (GGA-PBE) has been considered as the exchange correlation potentials and the energy cutoff is set to 470 eV. Furthermore, the dispersion energy corrections have been considered by incorporating DFT-D3 method of Grimme. [87,88] All of the structures are relaxed until the Hellmann–Feynman force criteria of < 0.01 eV/Å and the total energy convergence criteria of 10^{-4} eV is reached. The structures of the system have been taken from Materials Project database. Using the value of specific capacity from ML results, we obtained the number of intercalating ions using the equation,

$$C = \frac{ZxF}{M_F}$$

where z represents the charge on intercalating ions (1 in case of K), x represents the number of intercalating ions and F is the Faraday constant (26.8 Ah mol^{-1}). M_f represents the molecular weight of the formula unit of the electrode material. Using the ML predicted data from KRR model we have found the number of intercalated K ions and rounded off to the nearest whole number for DFT validation as presented in **Table 2.7**. The rounding

off is carried out to decrease the computational cost as modelling fractional ion intercalation will result in heavier DFT calculations. The DFT optimized fully intercalated systems with maximum capacity are represented in **Figure 2.14**. This proves that the number of K intercalation obtained from ML predicted data also agrees with the DFT optimized structures.

Table 2.7: Details regarding the number of intercalated K ions predicted by ML and the corresponding values chosen for DFT validation.

Electrode materials	No. of intercalating K ions predicted by ML	No. of intercalating K ions considered for DFT
Mn_4NiO_8	3.1	3
FeO_2	0.6	1
$Fe(CoO_3)_2$	2.5	2
VFeO4	1.1	1
$CoPO_4$	0.9	1

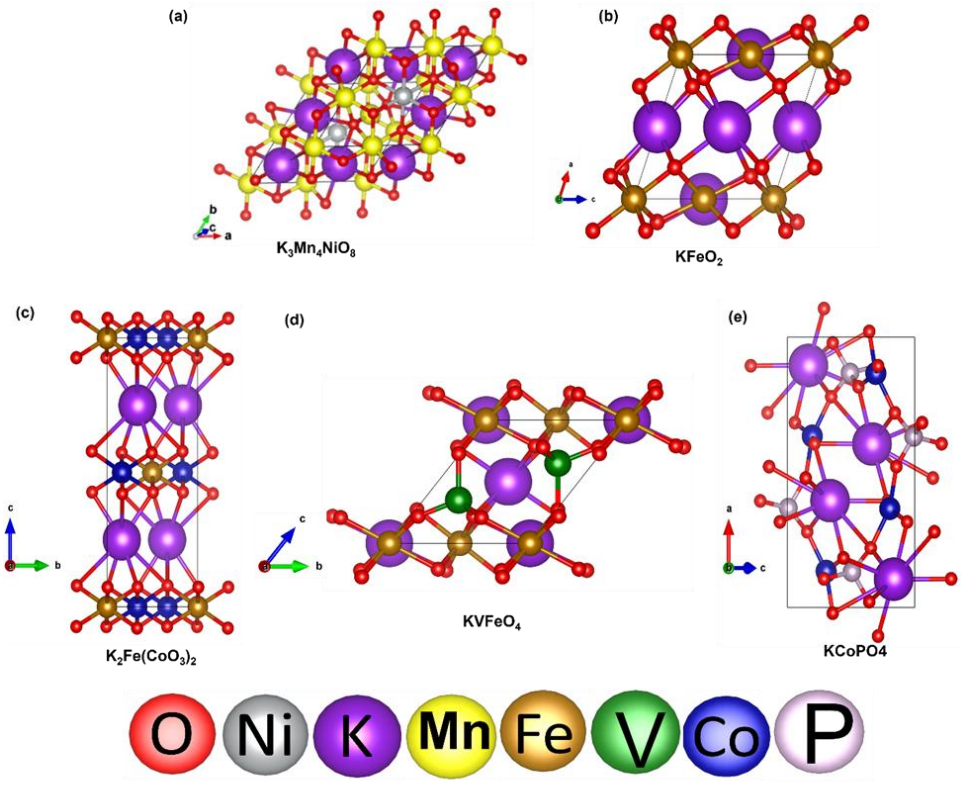


Figure 2.14: DFT optimized structures of K intercalated electrode materials (a) Mn_4NiO_8, (b) FeO_2, (c) $Fe(CoO_3)_2$, (d) $VFeO_4$, and (e) $CoPO_4$.

The gradual intercalation of K has also been shown for a sample electrode material Mn_4NiO_8, where the negative binding energy of K insertion shows the favourability of intercalation in the considered electrode material **(Figure 2.15 and Table 2.8)**.

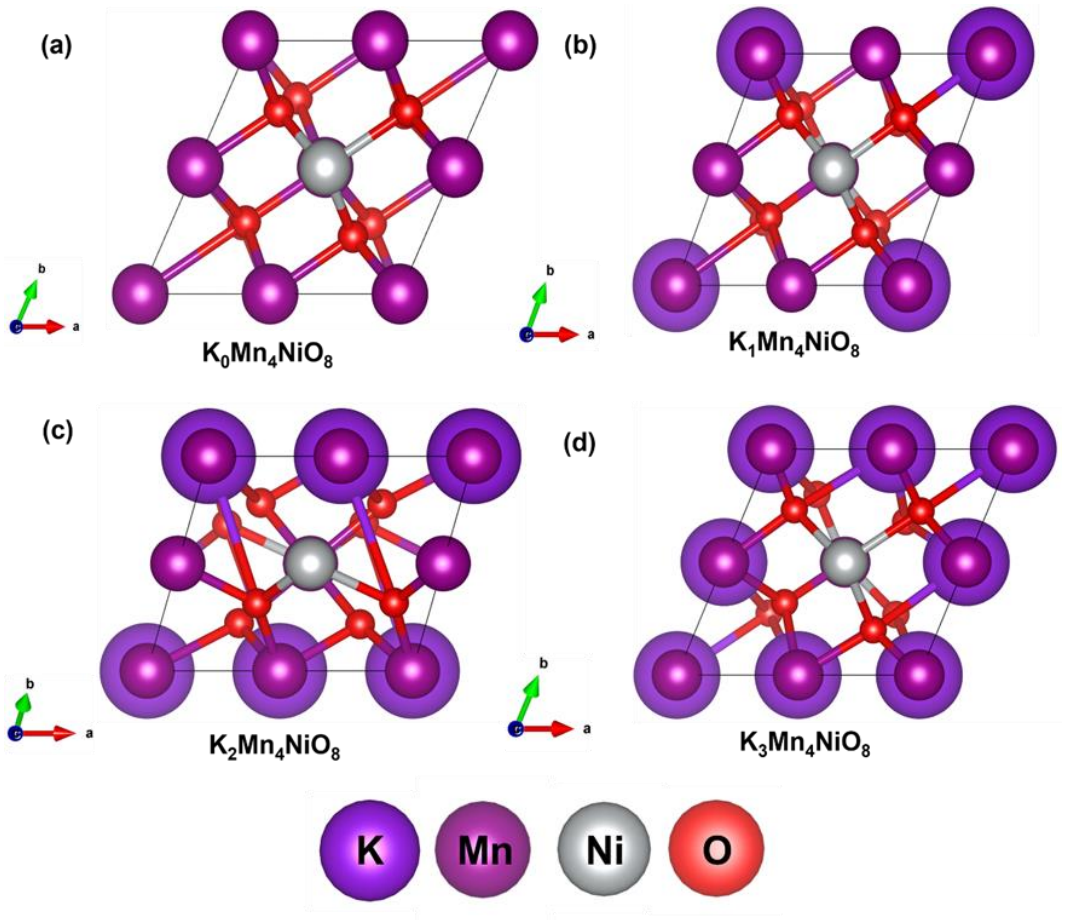


Figure 2.15: Gradual insertion of K ions in electrode material, Mn_4NiO_8 (a) $K_0Mn_4NiO_8$, (b) $K_1Mn_4NiO_8$, (c) $K_2Mn_4NiO_8$, (d) $K_3Mn_4NiO_8$.

Table 2.8: The calculated binding energy per K insertion for different concentration of K in Mn_4NiO_8.

Electrode with K insertion	Binding Energy/K insertion (eV)
$K_1Mn_4NiO_8$	-1.83
$K_2Mn_4NiO_8$	-3.15
$K_3Mn_4NiO_8$	-1.97

Furthermore, we have intercalated the fourth K-ion into the Mn_4NiO_8 in two possible ways to check if further intercalation is possible. The huge distortion in optimized structures of $K_4Mn_4NiO_8$ **(Figure 2.16)** as well as the abrupt increase in RMSD value **(Figure 2.17)** of the intercalated system with respect to non-intercalated system shows that intercalation of fourth K-ion is not suitable. This further validates that the ML predicted capacity values correspond to the maximum intercalation of K-ions in the electrode materials.

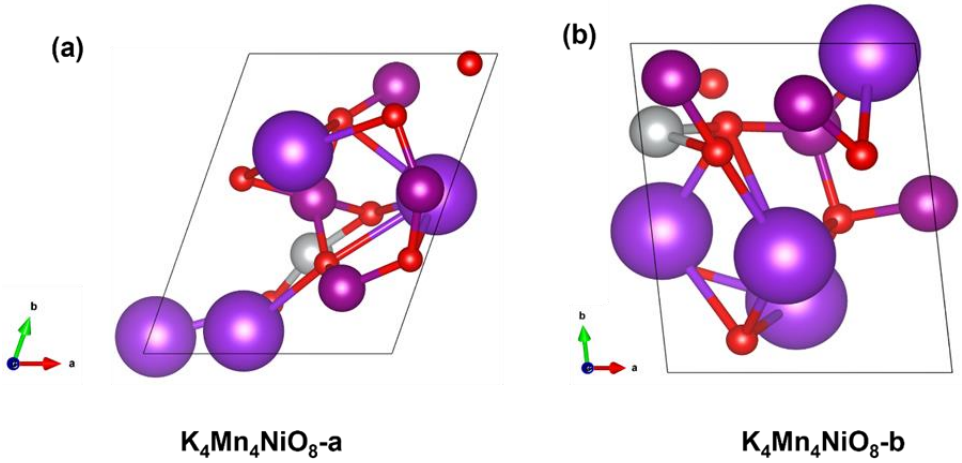


Figure 2.16: DFT optimized structures of Mn_4NiO_8 upon intercalation by four K ions. Here (a) and (b) represent two possibilities.

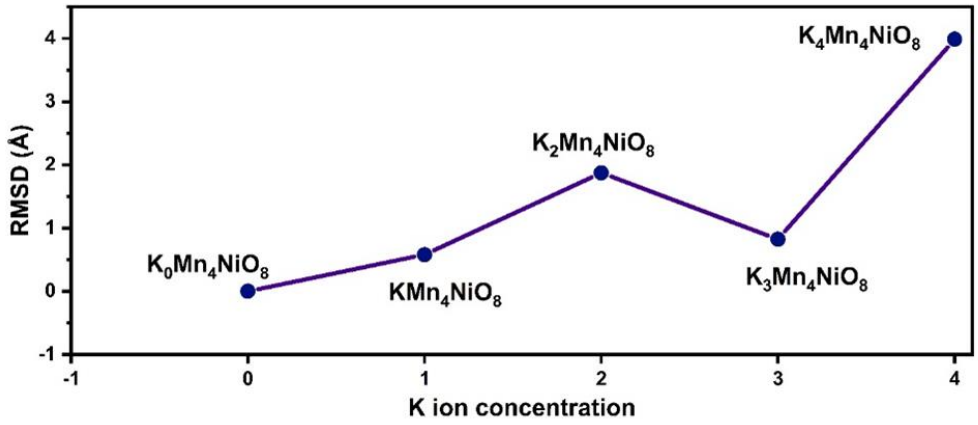


Figure 2.17: Root mean square displacement (RMSD) of Mn_4NiO_8 structure upon intercalation of K-ions with respect to the unintercalated structure.

2.5. Conclusion

In this work we have predicted specific capacity of prospective K-ion battery electrode materials based on the structural properties (e.g., lattice parameter, lattice angle, void space, etc.) and choice based feature vectorization generated from elemental properties (e.g., atomic number, electronegativity, ionic radius, valence electrons, etc.). We have considered Li, Na and K-ion electrode materials and their available battery data from Materials Project database. The electrode materials extracted from materials project database can be considered as stable as their formation energies are negative. Suitable features have been considered and developed to train the various machine learning algorithms. The available data has been divided into training set and validation set. The training set has been fitted using various ML algorithms like Support Vector Regression, ExtraTrees Regression and Kernel Ridge Regression to learn the nature of the data and features. Some statistical methods of data analysis like box plot and joint plot to understand the distribution of features, heatmap for the correlation metrics have been utilized. We have evaluated the performance of considered machine learning models by compairing the mean absolute percentage error between training set and validation set in each case. Further, adopting Kernel Ridge Regression we have predicted the capacity of unknown electrode materials for K-ion battery (Chapter-2 of Gihub repository, https://github.com/Souvik-ml/Thesis_data). Using the value of specific capacity, the number of intercalated K ions in the formula unit of the non-intercalated electrode material compounds have been calculated. DFT calculations have been performed for sample electrode materials to verify that our ML model can give similar results. Thus, implementing ML approach is much more faster compared to the computationally demanding quantum mechanecial methods for quick screening of electrode materials which will help to guide the experiments for devoloping electrode materials for metal ion batteries.

2.6. References

1. Yang Z., Zhang J., Kintner-Meyer M. C. W., Lu X., Choi D., Lemmon J. P., Liu J. (2011), Electrochemical energy storage for green grid, Chem. Rev., 111, 3577–3613 (DOI: 10.1021/cr100290z)
2. Nitta N., Wu F., Lee J. T., Yushin G. (2015), Li-ion battery materials: present and future, Mater. Today, 18, 252–264 (DOI: 10.1016/j.mattod.2015.10.011)
3. Winter M., Barnett B., Xu K. (2018), Before Li ion batteries, Chem. Rev., 118, 11433–11456 (DOI: 10.1021/acs.chemrev.8b00237)
4. Dunn B., Kamath H., Tarascon J. M. (2011), Electrical energy storage for the grid: a battery of choices, Science, 334, 928–935 (DOI: 10.1126/science.1212741)
5. Cheng F., Liang J., Tao Z., Chen J., Cheng F. Y., Liang J., Tao Z. L., Chen J. (2011), Functional materials for rechargeable batteries, Adv. Mater., 23, 1695–1715 (DOI: 10.1002/adma.201003649)
6. Tarascon J. M. (2010), Is lithium the new gold?, Nat. Chem., 2, 510–510 (DOI: 10.1038/nchem.713)
7. Joshi R. P., Ozdemir B., Barone V., Peralta J. E. (2015), Hexagonal BC_3: a robust electrode material for Li, Na, and K ion batteries, J. Phys. Chem. Lett., 6, 2728–2732 (DOI: 10.1021/acs.jpclett.5b01100)
8. Bhauriyal P., Mahata A., Pathak B. (2017), Hexagonal BC_3 electrode for a high-voltage Al-ion battery, J. Phys. Chem. C, 121, 9748–9756 (DOI: 10.1021/acs.jpcc.7b02847)
9. Posada J. O. G., Rennie A. J. R., Villar S. P., Martins V. L., Marinaccio J., Barnes A., Glover C. F., Worsley D. A., Hall P. J. (2017), Aqueous batteries as grid scale energy storage solutions, Renew. Sust. Energ. Rev., 68, 1174–1182 (DOI: 10.1016/j.rser.2016.03.038)
10. Liu K., Liu Y., Lin D., Pei A., Cui Y. (2018), Materials for lithium-ion battery safety, Sci. Adv., 4(6), eaas9820 (DOI: 10.1126/sciadv.aas9820)

11. Tarascon J. M., Armand M. (2010), Issues and challenges facing rechargeable lithium batteries, Mater. Sustain. Energy, 171–179 (DOI: 10.1038/35104644)
12. Nithya C., Gopukumar S. (2015), Sodium ion batteries: a newer electrochemical storage, Wiley Interdiscip. Rev.: Energy Environ., 4, 253–278 (DOI: 10.1002/wene.145)
13. Larcher D., Tarascon J. M. (2014), Towards greener and more sustainable batteries for electrical energy storage, Nat. Chem., 7, 19–29 (DOI: 10.1038/nchem.2085)
14. Scrosati B., Garche J. (2010), Lithium batteries: status, prospects and future, J. Power Sources, 195, 2419–2430 (DOI: 10.1016/j.jpowsour.2009.11.048)
15. Zhang W., Liu Y., Guo Z. (2019), Approaching high-performance potassium-ion batteries via advanced design strategies and engineering, Sci. Adv., 5(5), eaav7412 (DOI: 10.1126/sciadv.aav7412)
16. Eftekhari A., Jian Z., Ji X. (2017), Potassium secondary batteries, ACS Appl. Mater. Interfaces, 9, 4404–4419 (DOI: 10.1021/acsami.7b00552)
17. Zhang W., Huang W., Zhang Q. (2021), Organic materials as electrodes in potassium-ion batteries, Chem. Eur. J., 27, 6131–6144 (DOI: 10.1002/chem.202004813)
18. Rajagopalan R., Tang Y., Ji X., Jia C., Wang H. (2020), Advancements and challenges in potassium ion batteries: a comprehensive review, Adv. Funct. Mater., 30, 1909486 (DOI: 10.1002/adfm.201909486)
19. Pramudita J. C., Sehrawat D., Goonetilleke D., Sharma N. (2017), An initial review of the status of electrode materials for potassium-ion batteries, Adv. Energy Mater., 7, 1602911 (DOI: 10.1002/aenm.201602911)
20. Li W., Bi Z., Zhang W., Wang J., Rajagopalan R., Wang Q., Zhang D., Li Z., Wang H., Wang B. (2021), Advanced cathodes for potassium-ion

batteries with layered transition metal oxides: a review, J. Mater. Chem. A, 9, 8221–8247 (DOI: 10.1039/d1ta01720a)

21. Kim H., Ji H., Wang J., Ceder G. (2019), Next-generation cathode materials for non-aqueous potassium-ion batteries, Trends Chem., 1, 682–692 (DOI: 10.1016/j.trechm.2019.09.006)
22. Kirkpatrick P., Ellis C. (2004), Chemical space, Nature, 432, 823– (DOI: 10.1038/432823a)
23. von Lilienfeld O. A. (2018), Quantum machine learning in chemical compound space, Angew. Chem. Int. Ed., 57, 4164–4169 (DOI: 10.1002/anie.201707858)
24. Mullard A. (2017), The drug-maker's guide to the galaxy, Nature, 549, 445–447 (DOI: 10.1038/549445a)
25. Draxl C., Scheffler M. (2018), NOMAD: the FAIR concept for big data-driven materials science, MRS Bull., 43, 676–682 (DOI: 10.1557/mrs.2018.157)
26. Saal J. E., Kirklin S., Aykol M., Meredig B., Wolverton C. (2013), Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD), JOM, 65, 1501–1509 (DOI: 10.1007/s11837-013-0755-4)
27. Kirklin S., Saal J. E., Meredig B., Thompson A., Doak J. W., Aykol M., Rühl S., Wolverton C. (2015), The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, npj Comput. Mater., 1, 15010 (DOI: 10.1038/npjcompumats.2015.10)
28. Curtarolo S., Setyawan W., Hart G. L. W., Jahnatek M., Chepulskii R. v., Taylor R. H., Wang S., Xue J., Yang K., Levy O., Mehl M. J., Stokes H. T., Demchenko D. O., Morgan D. (2012), AFLOW: an automatic framework for high-throughput materials discovery, Comput. Mater. Sci., 58, 218–226 (DOI: 10.1016/j.commatsci.2012.02.005)
29. Gossett E., Toher C., Oses C., Isayev O., Legrain F., Rose F., Zurek E., Carrete J., Mingo N., Tropsha A., Curtarolo S. (2017), AFLOW-ML: a RESTful API for machine-learning predictions of materials

properties, Comput. Mater. Sci., 152, 134–145 (DOI: 10.1016/j.commatsci.2018.01.015)

30. Jain A., Ong S. P., Hautier G., Chen W., Richards W. D., Dacek S., Cholia S., Gunter D., Skinner D., Ceder G., Persson K. A. (2013), Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater., 1, 011002 (DOI: 10.1063/1.4812323)
31. Ong S. P., Cholia A., Jain A., Brafman M., Gunter D., Ceder G., Persson K. A. (2015), The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles, Comput. Mater. Sci., 97, 209–215 (DOI: 10.1016/j.commatsci.2014.10.037)
32. Ling C. (2022), A review of the recent progress in battery informatics, npj Comput. Mater., 8, 1–22 (DOI: 10.1038/s41524-022-00843-1)
33. Joshi R. P., Trepte K., Withanage K. P. K., Sharkas K., Yamamoto Y., Basurto L., Zope R. R., Baruah T., Jackson K. A., Peralta J. E. (2018), Fermi-Löwdin orbital self-interaction correction to magnetic exchange couplings, J. Chem. Phys., 149, 164101 (DOI: 10.1063/1.5029232)
34. Kaloni T. P., Joshi R. P., Adhikari N. P., Schwingenschlögl U. (2014), Band gap tunning in BN-doped graphene systems with high carrier mobility, Appl. Phys. Lett., 104, 073116 (DOI: 10.1063/1.4865960)
35. Ramprasad R., Batra R., Pilania G., Mannodi-Kanakkithodi A., Kim C. (2017), Machine learning in materials informatics: recent applications and prospects, npj Comput. Mater., 3, 1–13 (DOI: 10.1038/npjcompumats.2017.8)
36. Bassman L., Rajak P., Kalia R. K., Nakano A., Sha F., Sun J., Singh D. J., Aykol M., Huck P., Persson K., Vashishta P. (2018), Active learning for accelerated design of layered materials, npj Comput. Mater., 4, 1–9 (DOI: 10.1038/s41524-018-0065-0)

37. Zhang Y., Ling C. (2018), A strategy to apply machine learning to small datasets in materials science, npj Comput. Mater., 4, 1–8 (DOI: 10.1038/s41524-018-0128-8)
38. Butler K. T., Davies D. W., Cartwright H., Isayev O., Walsh A. (2018), Machine learning for molecular and materials science, Nature, 559, 547–555 (DOI: 10.1038/s41586-018-0337-2)
39. Schleder G. R., Padilha A. C. M., Acosta C. M., Costa M., Fazzio A. (2019), From DFT to machine learning: recent approaches to materials science–a review, J. Phys.: Mater., 2, 032001 (DOI: 10.1088/2515-7639/ab1425)
40. Dong Y., Wu C., Zhang C., Liu Y., Cheng J., Lin J. (2019), Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride, npj Comput. Mater., 5, 1–8 (DOI: 10.1038/s41524-019-0179-0)
41. Zhuo Y., Mansouri Tehrani A., Brgoch J. (2018), Predicting the band gaps of inorganic solids by machine learning, J. Phys. Chem. Lett., 9, 1668–1673 (DOI: 10.1021/acs.jpclett.8b00527)
42. Kolb B., Lentz L. C., Kolpak A. M. (2017), Discovering charge density functionals and structure-property relationships with PROPhet: A general framework for coupling machine learning and first-principles methods, Sci. Rep., 7, 1–9 (DOI: 10.1038/s41598-017-08843-w)
43. Pilania G., Wang C., Jiang X., Rajasekaran S., Ramprasad R. (2013), Accelerating materials property predictions using machine learning, Sci. Rep., 3, 1–6 (DOI: 10.1038/srep02828)
44. Pilania G., Mannodi-Kanakkithodi A., Uberuaga B. P., Ramprasad R., Gubernatis J. E., Lookman T. (2016), Machine learning bandgaps of double perovskites, Sci. Rep., 6, 1–10 (DOI: 10.1038/srep34256)
45. Takahashi K., Takahashi L., Miyazato I., Tanaka Y. (2018), Searching for hidden perovskite materials for photovoltaic systems by combining data science and first principle calculations, ACS Photonics, 5, 771–775 (DOI: 10.1021/acsphotonics.7b01560)

46. Sodeyama K., Igarashi Y., Nakayama T., Tateyama Y., Okada M. (2018), Liquid electrolyte informatics using an exhaustive search with linear regression, Phys. Chem. Chem. Phys., 20, 22585–22591 (DOI: 10.1039/c8cp02939k)
47. Okamoto Y., Kubo Y. (2018), Ab initio calculations of the redox potentials of additives for lithium-ion batteries and their prediction through machine learning, ACS Omega, 3, 7868–7874 (DOI: 10.1021/acsomega.8b01553)
48. Jalem R., Nakayama M., Kasuga T. (2013), An efficient rule-based screening approach for discovering fast lithium ion conductors using density functional theory and artificial neural networks, J. Mater. Chem. A, 2, 720–734 (DOI: 10.1039/c3ta13487b)
49. Fujimura K., Seko A., Koyama Y., Kuwabara A., Kishida I., Shitara K., Fisher C. A. J., Moriwake H., Tanaka I. (2013), Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms, Adv. Energy Mater., 3, 980–985 (DOI: 10.1002/aenm.201200671)
50. Kireeva N., Pervov V. S. (2017), Materials space of solid-state electrolytes: unraveling chemical composition–structure–ionic conductivity relationships in garnet-type metal oxides using cheminformatics virtual screening approaches, Phys. Chem. Chem. Phys., 19, 20904–20918 (DOI: 10.1039/c7cp04025a)
51. Cubuk E. D., Sendek A. D., Reed E. J. (2019), Screening billions of candidates for solid lithium-ion conductors: a transfer learning approach for small data, J. Chem. Phys., 150, 214701 (DOI: 10.1063/1.5078680)
52. Sendek A. D., Yang Q., Cubuk E. D., Duerloo K. A. N., Cui Y., Reed E. J. (2017), Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials, Energy Environ. Sci., 10, 306–320 (DOI: 10.1039/c6ee02619a)

53. Schütt K. T., Glawe H., Brockherde F., Sanna A., Müller K. R., Gross E. K. U. (2014), How to represent crystal structures for machine learning: Towards fast prediction of electronic properties, Phys. Rev. B. Condens. Matter., 89, 205118 (DOI: 10.1103/physrevb.89.205118)
54. Roy D., Mandal S. C., Pathak B. (2021), Machine learning-driven high-throughput screening of alloy-based catalysts for selective CO_2 hydrogenation to methanol, ACS Appl. Mater. Interfaces, 13, 56151–56163 (DOI: 10.1021/acsami.1c16145)
55. Roy D., Mandal S. C., Pathak B. (2022), Machine learning assisted exploration of high entropy alloy-based catalysts for selective CO_2 reduction to methanol, J. Phys. Chem. Lett., 13, 5991–6002 (DOI: 10.1021/acs.jpclett.2c01729)
56. Seko A., Hayashi H., Nakayama K., Takahashi A., Tanaka I. (2017), Representation of compounds for machine-learning prediction of physical properties, Phys. Rev. B, 95, 144110 (DOI: 10.1103/physrevb.95.144110)
57. Seko A., Maekawa T., Tsuda K., Tanaka I. (2014), Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids, Phys. Rev. B. Condens. Matter., 89, 054303 (DOI: 10.1103/physrevb.89.054303)
58. Ward L., Agrawal A., Choudhary A., Wolverton C. (2016), A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater., 2, 1–7 (DOI: 10.1038/npjcompumats.2016.28)
59. Faber F. A., Lindmaa A., von Lilienfeld O. A., Armiento R. (2016), Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals, Phys. Rev. Lett., 117, 135502 (DOI: 10.1103/PhysRevLett.117.135502)
60. Deml A. M., O'Hayre R., Wolverton C., Stevanović V. (2016), Predicting density functional theory total energies and enthalpies of

formation of metal-nonmetal compounds by linear regression, Phys. Rev. B, 93, 085142 (DOI: 10.1103/PhysRevB.93.085142)

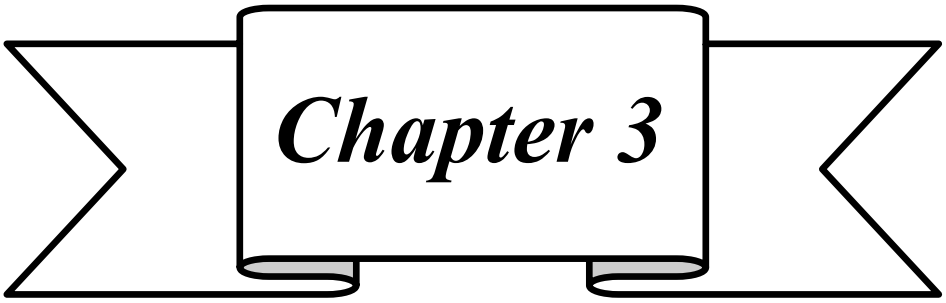
61. Shi S., Gao J., Liu Y., Zhao Y., Wu Q., Ju W., Ouyang C., Xiao R. (2015), Multi-scale computation methods: Their applications in lithium-ion battery research and development, Chin. Phys. B, 25, 018212 (DOI: 10.1088/1674-1056/25/1/018212)
62. Zhao Q., Avdeev M., Chen L., Shi S. (2021), Machine learning prediction of activation energy in cubic Li-argyrodites with hierarchically encoding crystal structure-based (HECS) descriptors, Sci. Bull., 66, 1401–1408 (DOI: 10.1016/j.scib.2021.07.009)
63. Liu B., Yang J., Yang H., Ye C., Mao Y., Wang J., Shi S., Yang J., Zhang W. (2019), Rationalizing the interphase stability of Li|doped-$Li_7La_3Zr_2O_{12}$ via automated reaction screening and machine learning, J. Mater. Chem. A, 7, 19961–19969 (DOI: 10.1039/c9ta07357k)
64. Wang A., Zou Z., Wang D., Liu Y., Li Y., Wu J., Avdeev M., Shi S. (2021), Identifying chemical factors affecting reaction kinetics in Li-air battery via ab initio calculations and machine learning, Energy Storage Mater., 35, 595–601 (DOI: 10.1016/j.ensm.2020.07.026)
65. Zhao Q., Zhang L., He B., Ye A., Avdeev M., Chen L., Shi S. (2021), Identifying descriptors for Li^+ conduction in cubic Li-argyrodites via hierarchically encoding crystal structure and inferring causality, Energy Storage Mater., 40, 386–393 (DOI: 10.1016/j.ensm.2021.02.016)
66. Meredig B., Agrawal A., Kirklin S., Saal J. E., Doak J. W., Thompson A., Zhang K., Choudhary A., Wolverton C. (2014), Combinatorial screening for new materials in unconstrained composition space with machine learning, Phys. Rev. B. Condens. Matter., 89, 094104 (DOI: 10.1103/PhysRevB.89.094104)
67. Liu Y., Wu J. M., Avdeev M., Shi S. Q. (2020), Multi-layer feature selection incorporating weighted score-based expert knowledge toward

modeling materials with targeted properties, Adv. Theory Simul., 3, 1900215 (DOI: 10.1002/adts.201900215)

68. Brockherde F., Vogt L., Li L., Tuckerman M. E., Burke K., Müller K. R. (2017), Bypassing the Kohn-Sham equations with machine learning, Nat. Commun., 8, 1–10 (DOI: 10.1038/ncomms8721)
69. Mills K., Spanner M., Tamblyn I. (2017), Deep learning and the Schrödinger equation, Phys. Rev. A, 96, 042113 (DOI: 10.1103/PhysRevA.96.042113)
70. Li L., Snyder J. C., Pelaschier I. M., Huang J., Niranjan U. N., Duncan P., Rupp M., Müller K. R., Burke K. (2016), Understanding machine-learned density functionals, Int. J. Quantum Chem., 116, 819–833 (DOI: 10.1002/qua.25171)
71. Liu Y., Guo B., Zou X., Li Y., Shi S. (2020), Machine learning assisted materials design and discovery for rechargeable batteries, Energy Storage Mater., 31, 434–450 (DOI: 10.1016/j.ensm.2020.05.020)
72. Liu Y., Zhao T., Ju W., Shi S. (2017), Materials discovery and design using machine learning, J. Materiomics, 3, 159–177 (DOI: 10.1016/j.jmat.2017.03.002)
73. Bender A., Schneider N., Segler M., Walters W. P., Engkvist O., Rodrigues T. (2022), Evaluation guidelines for machine learning tools in the chemical sciences, Nat. Rev. Chem., 6, 428–442 (DOI: 10.1038/s41570-022-00391-5)
74. Wu J., Zhang C., Chen Z. (2016), An online method for lithium-ion battery remaining useful life estimation using importance sampling and neural networks, Appl. Energy, 173, 134–140 (DOI: 10.1016/j.apenergy.2016.03.038)
75. Cheng D., Sha W., Wang L., Tang S., Ma A., Chen Y., Wang H., Lou P., Lu S., Cao Y. C. (2021), Solid-state lithium battery cycle life prediction using machine learning, Appl. Sci., 11, 4671 (DOI: 10.3390/app11094671)

76. Joshi R. P., Eickholt J., Li L., Fornari M., Barone V., Peralta J. E. (2019), Machine learning the voltage of electrode materials in metal-ion batteries, ACS Appl. Mater. Interfaces, 11, 18494–18503 (DOI: 10.1021/acsami.9b05103)
77. Zhou F., Cococcioni M., Marianetti C. A., Morgan D., Ceder G. (2004), First-principles prediction of redox potentials in transition-metal compounds with LDA + U, Phys. Rev. B. Condens. Matter., 70, 1–8 (DOI: 10.1103/PhysRevB.70.235121)
78. Wang A. Y. T., Murdock R. J., Kauwe S. K., Oliynyk A. O., Gurlo A., Brgoch J., Persson K. A., Sparks T. D. (2020), Machine learning for materials scientists: An introductory guide toward best practices, Chem. Mater., 32, 4954–4965 (DOI: 10.1021/acs.chemmater.0c01420)
79. Atkins P., Atkins P. W., de Paula J. (2014), Chapter 16, Atkins' Physical Chemistry, Oxford University Press, 59
80. Wang G., Fearn T., Wang T., 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., 7, 1551–1560 (DOI: 10.1021/acscentsci.1c01447)
81. Blöchl P. E. (1994), Projector augmented-wave method, Phys. Rev. B, 50, 17953 (DOI: 10.1103/PhysRevB.50.17953)
82. Kresse G., Joubert D. (1999), From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B, 59, 1758 (DOI: 10.1103/PhysRevB.59.1758)
83. Kresse G., Hafner J. (1994), Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium, Phys. Rev. B, 49, 14251 (DOI: 10.1103/PhysRevB.49.14251)
84. Kresse G., Hafner J. (1993), Ab initio molecular dynamics for liquid metals, Phys. Rev. B, 47, 558 (DOI: 10.1103/PhysRevB.47.558)

85. Kresse G., Furthmüller J. (1996), Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci., 6, 15–50 (DOI: 10.1016/0927-0256(96)00008-0)
86. Kresse G., Furthmüller J. (1996), Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B, 54, 11169 (DOI: 10.1103/PhysRevB.54.11169)
87. Perdew J. P., Burke K., Ernzerhof M. (1996), Generalized gradient approximation made simple, Phys. Rev. Lett., 77, 3865 (DOI: 10.1103/PhysRevLett.77.3865)
88. Grimme S., Antony J., Ehrlich S., Krieg H. (2010), A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu, J. Chem. Phys., 132, 154104 (DOI: 10.1063/1.3382344)



Chapter 3

Automated Pipeline for High throughput Screening of Electrode Materials

3.1. Introduction

The rapid advancement of energy storage technologies has driven extensive research into high-performance rechargeable batteries. Among the key components of these batteries, anode plays a crucial role in determining overall performance, including energy density, cycle life, and rate capability. Traditional graphite anodes, while widely used in lithium-ion (Li ion) batteries, exhibit limitations in terms of capacity and compatibility with larger alkali metal ions such as sodium (Na) and potassium (K). This has motivated the search for novel anode materials with superior electrochemical properties.

Two-dimensional (2D) materials have emerged as promising candidates for next-generation anodes due to their unique structural, electronic, and mechanical properties.*[1–3]* In particular, 2D MT2-type materials (M: transition metal and T: terminal functional group), have garnered significant interest for metal-ion batteries.*[4–7]* These materials exhibit high surface area, tunable electronic properties, and favorable ion diffusion pathways, making them suitable for accommodating alkali metal ions. While Li ion batteries dominate the current energy storage market, exploring alternative metal-ion batteries such as Na and K ion batteries along with Li ion batteries are critical for sustainable energy solutions. The growing demand for lithium, coupled with its uneven geographical distribution and high cost, poses challenges for large-scale deployment. In contrast, sodium and potassium are more abundant and cost-effective, making them attractive alternatives for next-generation batteries. Additionally, Na-ion and K-ion batteries offer promising electrochemical performance and potential advantages in large-scale energy storage applications. Investigating 2D materials for Li-ion, Na-ion and K-ion batteries can lead to new breakthroughs in energy storage technology.

Despite these promising characteristics, the discovery and optimization of new 2D electrode materials for battery applications remain challenging. The

conventional approach, based on experimental and density functional theory (DFT) calculations, is highly time consuming and computationally intensive, particularly when evaluating a large number of candidate materials. Most of the DFT or experimentally reported battery electrode materials are concerned about one or a few materials at a time and hence exploration of a large material database for battery applications needs some innovative tools which can quickly filter the potential electrode candidates. The need to determine optimal adsorption sites, simulate fully intercalated structures followed by stepwise de-metalation, and compute voltage profiles at multiple stages further exacerbates the computational burden, making high-throughput screening via DFT unviable.

To overcome these limitations, machine learning (ML) models and machine learning-based interatomic potentials have emerged as a powerful alternative for accelerating computational materials discovery. Establishment of structure-property relationships through various ML models based on suitable features has accelerated the material discovery processes specially for rechargeable batteries.*[8–13]* Recently various ML potential models are coming very handy to quickly filter out potential candidates for energy storage devices, Crystal Hamiltonian Graph Neural Network (CHGNet) developed by Bowen et el. is one of them which can identify suitable energy storage material in no time compared to DFT with DFT level of accuracy.*[14–22]*

In this work we adopted a fully automated workflow that identifies stable adsorption sites, metal-ion adsorption, and predicts voltage profiles for Li-, Na-, and K-ion storage with the help of a universal machine learning potential. The proposed automated pipeline helps to achieve significant computational efficiency, enabling rapid exploration of a vast chemical space while maintaining accuracy comparable to DFT calculations. An optimum selection criterion has been set to identify potential 2D electrode materials for Li-, Na-, and K-ion batteries. The automated pipeline not only

accelerate the material discovery process but also offer valuable insights into the electrochemical behavior of 2D MT2 type systems, guiding future experimental and theoretical studies in the field of rechargeable batteries. The adopted automated pipeline for determination of voltage profile has been represented in **Figure 3.1**.

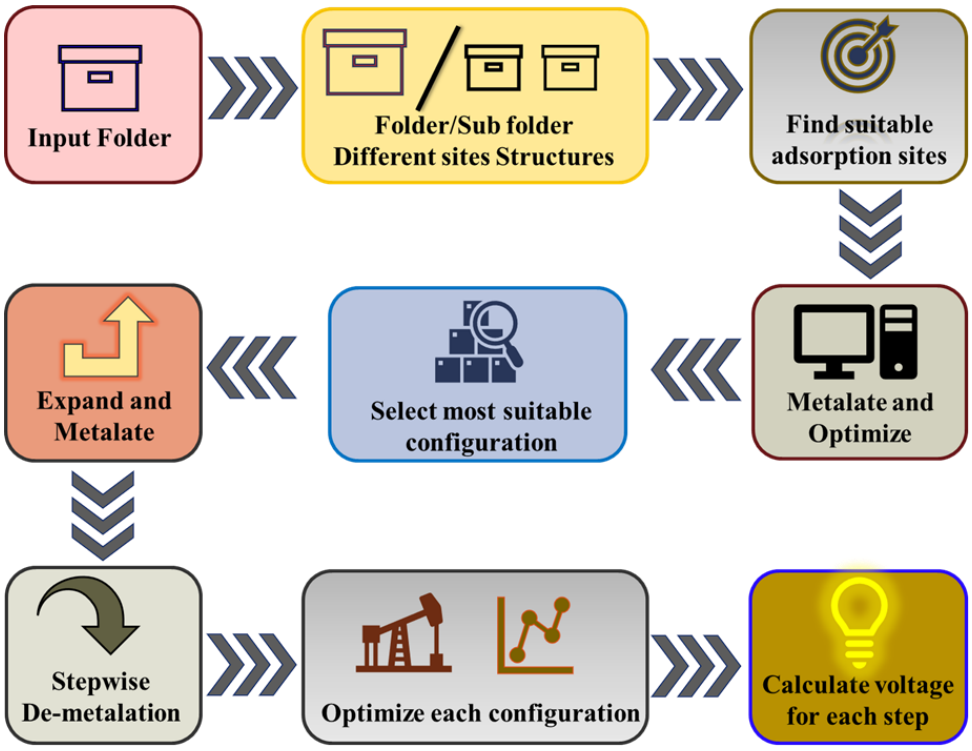


Figure 3.1: Automated pipeline for the determination of voltage profile of electrode materials.

3.2. Computational Details

The geometry optimizations were performed using the Vienna Ab Initio Simulation Package (VASP) with the projector augmented wave (PAW) method.*[23–28]* The Perdew–Burke–Ernzerhof (PBE) form of the generalized gradient approximation (GGA) was employed to describe the exchange–correlation interactions, and the plane wave basis was truncated at an energy cutoff of 470 eV.*[29]* We have incorporated van der Waals interactions via Grimme's DFT-D3 dispersion correction.*[30]* Structural relaxations were carried out until the Hellmann–Feynman forces on each

atom were less than 10^{-2} eV/Å and the total energy converged to within 10^{-5} eV. A Γ-centered k-point grid of $2 \times 2 \times 1$ was used for Brillouin zone sampling, and a vacuum spacing of approximately 10 Å was introduced along the z-direction to avoid spurious interactions between periodic images. The DFT generated data has been utilized for the training of pre-trained (Crystal Hamiltonian Graph neural Network) CHGNet model which is a graph-based deep learning model designed to predict energies, forces, and stresses from atomic configurations. CHGNet represents crystal structures as graphs, incorporating atom, bond, and angle embeddings, and uses message passing through graph convolutional layers. For each of the 289 materials, four different adsorption sites were evaluated for Li^+, Na^+, and K^+ ions. After identifying the most stable configuration, five sequential de-metalation steps were performed to simulate ion removal during battery discharge resulted in a total of 7,803 calculations.

3.3. Materials

The considered MT2 type based 2D materials are extracted from the aNANt database (anant.mrc.iisc.ac.in/apps) which are mainly composed of one metal layer (M) sandwich between two terminal (T) layers. For our study, we have considered MT2 types of materials having same terminal groups in both ends.

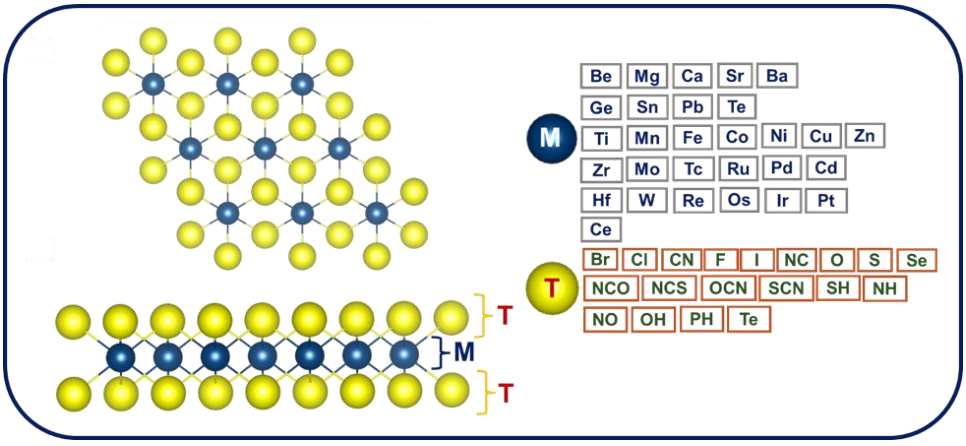


Figure 3.2: General representation of considered 2D materials (top and side views) with various metal atoms and terminal atoms/groups.

The metal and terminal groups considered in our study are given in **Figure 3.2**. The total number of 2D materials having the same terminal group considered for our study is 289. The combination of 29 different metals (M) and 19 terminal groups (T) creates a highly diverse and chemically rich dataset. This complexity poses a significant challenge for traditional machine learning (ML) models, which often struggle to capture the nuanced chemistry of ion adsorption based on limited DFT calculated data and optimized structures. Conventional ML approaches rely heavily on extensive data preprocessing and handcrafted feature engineering, and even then, their reliability hinges on data quantity and quality. A critical issue arises when researchers discard structures that become unstable after ion adsorption, labeling them as outliers. While this may simplify the dataset, it introduces bias because when the same model is applied to unknown materials, it may mistakenly predict unstable configurations as stable ones, especially since re-optimizing thousands of candidates via DFT is not practically feasible. To overcome these limitations, we leverage the Crystal Hamiltonian Graph Neural Network (CHGNet), developed by Bowen et al.*[14]* Unlike conventional models, CHGNet is a universal ML potential trained on the Materials Project database. It requires only an input structure and can rapidly predict key physical quantities such as energy and force offering DFT level accuracy with near-zero computational cost. Crucially, it also captures structural distortions upon ion intercalation, making it ideal for high-throughput screening.

In the next section, we detail the conventional DFT based method for voltage calculation and show how CHGNet enables full automation of this workflow.

3.4. Results and Discussion

3.4.1. Training of CHGNet

At first, we have generated a DFT dataset of total energy, energy per atom, atomic forces, and stress tensors of Li^+, Na^+, and K^+ intercalated MT2 based

electrode materials. We have considered dataset of every fifth ionic relaxation step to balance the temporal resolution with data compactness. For each configurations, we extracted the crystal structure coordinates, total energy, energy per atom, atomic forces, and stress tensors. This has been done to efficiently capture relevant structural as well energetic and other related informations to train the CHGNet model. While CHGNet demonstrates remarkable accuracy in predicting material stability on the Materials Project (MP) dataset, its pretraining was conducted exclusively on 3D bulk structures. In our case, however, the electrode materials of interest are 2D, and the CHGNet model must be fine-tuned to accurately capture their properties. In particular, interatomic interactions, electronic structure, and adsorption energetics in low-dimensional systems differ fundamentally from bulk behavior, necessitating a retraining or fine-tuning strategy to adapt the model to the 2D domain.

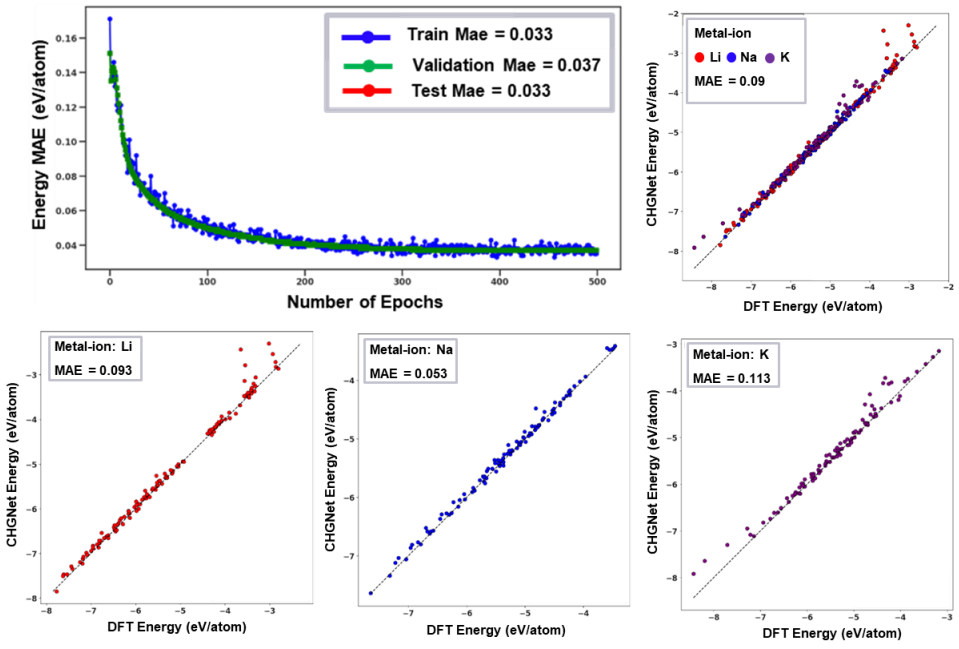


Figure 3.3: Learning curve of the training of CHGNet, and Parity plot comparing the CHGNet predicted energies with the DFT energies for metal-ion intercalated structures.

The learning curve **(Figure 3.3)** shows a steady decline in both training and validation MAE, eventually converging to 0.038 eV/atom on the test set. The small and consistent gap between training and validation error across 500 epochs indicates that the model avoids overfitting while learning effectively from the data. This performance suggests strong generalization to new chemical environments within the given compositional space. To further assess the model's performance, we compared CHGNet-predicted energies against DFT-calculated energy values for individual ion types. As shown in **Figure 3.3**, CHGNet achieves a MAE of 0.093, 0.056, and 0.113 eV/atom for Li-, Na-, and K-ion systems respectively with an overall MAE of 0.09 eV/atom. While the predictions generally follow the ideal line, slight deviations appear at higher energy values for Li-ion and at lower energy values for K-ion. Nevertheless, CHGNet is able to track these subtle distortions reasonably well, suggesting that it captures the overall energy trends while remaining sensitive to outlier structures.

3.4.2. Adsorption Sites

Because several adsorption sites can exist, it is important to identify the most stable site to construct the metal-ion–adsorbed structures. We used a tool in the pymatgen package (AdsorbateSiteFinder) to automatically generate possible adsorption sites. Four possible sites were identified: top, bridge, hollow-1, and hollow-2, as shown in Figure 3. All these structures were relaxed, and the lowest-energy configuration for each metal-ion was selected for voltage calculations. The most stable Li, Na, and K adsorption sites for each material are listed in **Table 3.1** of chapter-3, Github repository (https://github.com/Souvik-ml/Thesis_data). Starting from the most stable site, we expanded the supercell, fully loaded with metal ions, and then sequentially removed the metal-ions as shown in **Figure 3.4.**

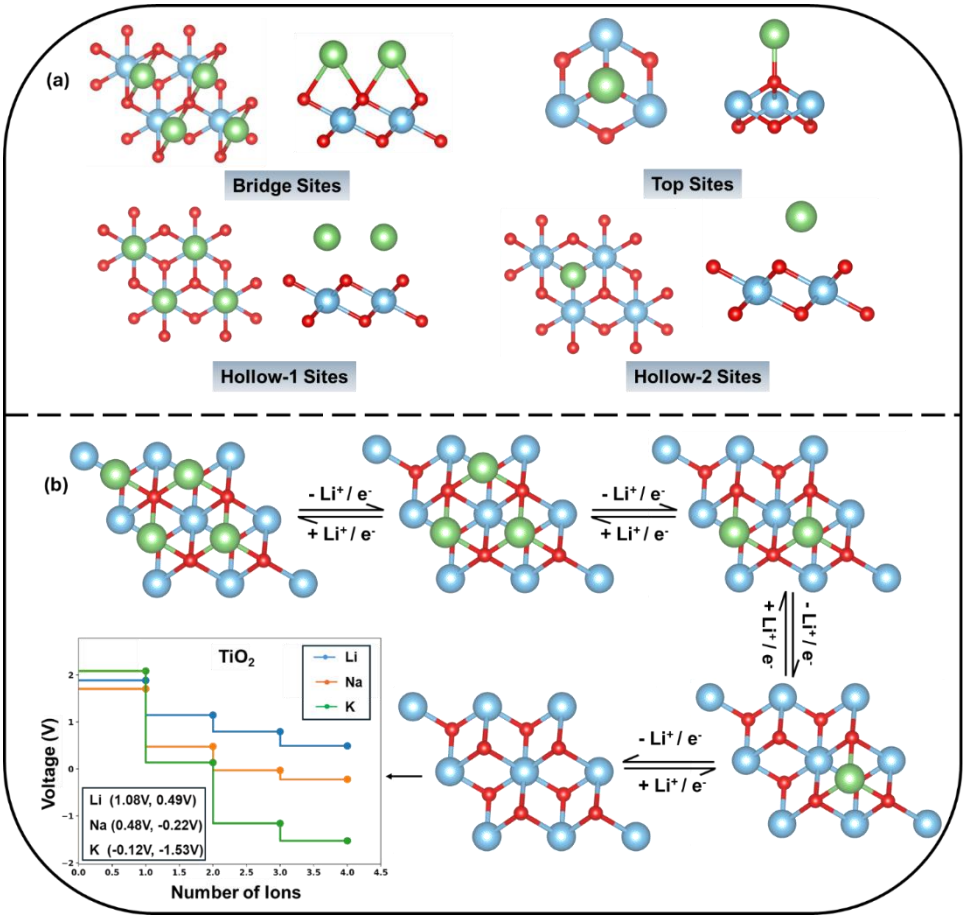


Figure 3.4. (a) Top and side views of different adsorption sites for Li^+ adsorption on TiO_2, and (b) schematic diagram of the de-lithiation steps from the lithiated system for voltage calclations.

3.4.3. Voltage Calculation

For voltage calculations, a fully lithiated structure was first constructed, followed by stepwise de-lithiation (from four to zero) to compute the stepwise voltage. Here, we present the voltage equation considering Li as the metal-ion. For Na- and K-ion batteries, Li is replaced by Na or K. The half-cell reaction is expressed as follows:

$$MT_2 + nLi^+ + ne^- \rightleftharpoons MT_2Li_n$$

Where MT_2 and MT_2Li_n are the non-lithiated and lithiated structures respectively, and n is the number of adsorb Li in each step.

The voltage (E_{cell}) for each step is expressed in terms of Gibbs free energy change (ΔG) of the above reaction, where ΔG is approximated by the change in internal energy (ΔE).

$$E_{cell} = -\frac{\Delta G}{nF} \approx -\frac{\Delta E}{nF}$$

Where ΔE is written as,

$$\Delta E = E_{MT_2Li_n} - (E_{MT_2} + n\,.\,E_{Li})$$

Thus, the general voltage equation can be written as follows,

$$E_{cell} = -\frac{E_{MT_2Li_n} - (E_{MT_2} + n\,.\,E_{Li})}{nF}$$

Where $E_{MT_2Li_n}$ and E_{MT_2} are the total energies of MT_2 system with Li and without Li respectively, E_{Li} is the energy per atom for the bulk metal and F is the Faraday constant.

To assess the practical applicability of CHGNet in predicting voltages for MT_2 type 2D electrode materials, we benchmarked its performance against both DFT computed values and conventional ML models predicted values. As shown in **Figure 3.5a**, CHGNet achieves a strong correlation with DFT calculated voltages across all three ion types (Li^+, Na^+, and K^+), with an overall MAE of 0.253 V

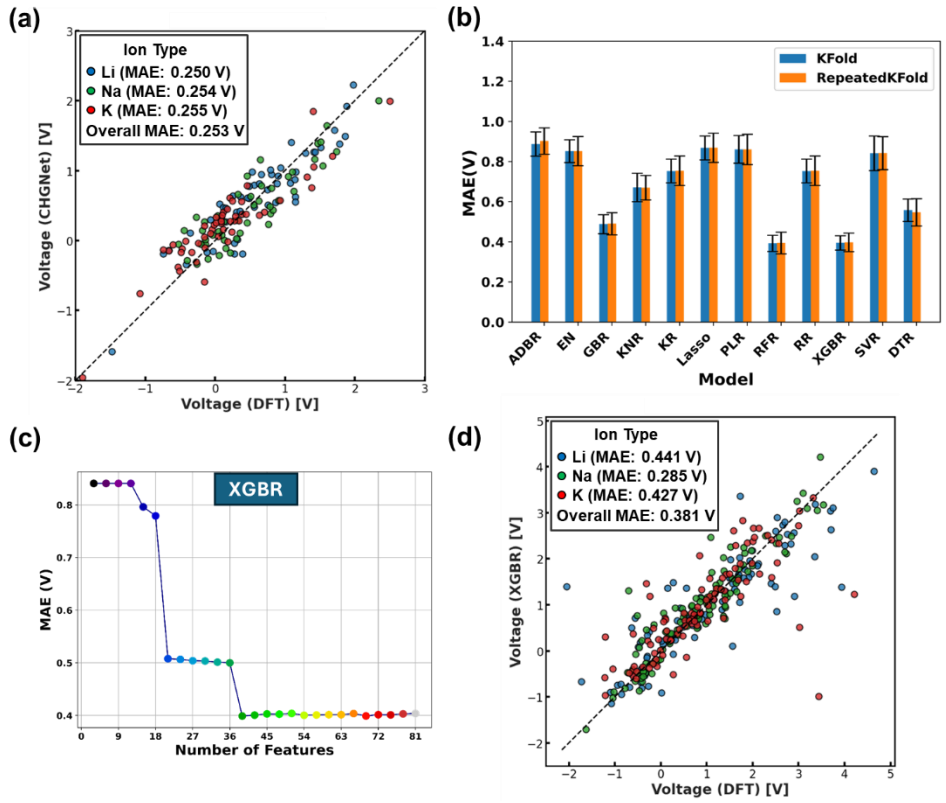


Figure 3.5: (a) Parity plot comparing DFT-calculated voltage with CHGNet-predicted voltage, (b) Bar plot of mean absolute errors across different machine learning models, (c) Optimization of feature selection for the XGBR model using the Select-K-Best method and (d) Parity plot comparing DFT-calculated voltages with XGBR-predicted voltages.

The individual MAEs for Li (0.250 V), Na (0.254 V), and K (0.255 V) are very close, suggesting that CHGNet maintains robust performance across different ionic sizes. This level of agreement is notable, given that voltages are highly sensitive to subtle changes in local bonding environments and structural relaxation, particularly in 2D materials. Further, we evaluated a conventional supervised eXtreme Gradient Boosting Regression (XGBR) model's performance, a widely used tree-based model in materials informatics for the prediction of voltage. We employed both K-Fold and Repeated K-Fold Cross Validation **(Figure 3.5b)** with elemental feature set **(Table 3.2)** for different ML models.

Table 3.2: Abbreviation of initially considered features along with description for machine learning model. Here, "M", "T", stand for metal and terminal group in MT_2 respectively and "I" stands for metal ion (Li/Na/K).

Abbreviation	Description
VE_M, VE_T	Valence electron of M, and T
VS_M, VS_T	Valence s electron of M, and T
VP_M, VP_T	Valence p electron of M, and T
VD_M, VD_T	Valence d electron of M, and T
VF_M, VF_T	Valence f electron of M, and T
UVS_M, UVS_T	Unfilled s valence electrons of M, and T
UVP_M, UVP_T	Unfilled p valence electrons of M, and T
UVD_M, UVD_T	Unfilled d valence electrons of M, and T
UVF_M, UVF_T	Unfilled f valence electrons of M, and T
OSE_M, OSE_T	Outer shell electrons of M, and T
IE_M, IE_T, IE_I	First ionization energy of M, T, and I
PO_M, PO_T, PO_I	Polarizability of M, T, and I
IR_M, IR_T, IR_I	Ionic radius of M, T, and I
CR_M, CR_T, CR_I	Covalent radius of M, T, and I
PE_M, PE_T	Pauling electronegativity of M, and T
MB_M, MB_T	MB electronegativity of M, and T
ME_M, ME_T, ME_I	Mulliken electronegativity of M, T, and I
MV_M, MV_T	Metallic valence of M, and T
Number_of_ion	Number of adsorbed metal-ions
AN_M, AN_T, AN_I	Atomic number of M, T, and I
AW_M, AW_T, AW_I	Atomic weight of M, T, and I
P_M, P_T	Period number of M, and T
G_M, G_T	Group numbers of M, and T

MN_M, MN_T	Mendeleev number of M, and T
MP_M, MP_T, MP_I	Melting point of M, T, and I
BP_M, BP_T, BP_I	Boiling point of M, T, and I
D_M, D_T, D_I	Density of M, T, and I
SH_M, SH_T, SH_I	Specific heat of M, T, and I
HF_M, HF_T, HF_I	Heat of fusion of M, T, and I
HV_M, HV_T, HV_I	Heat of vaporization of M, T, and I
TC_M, TC_T, TC_I	Thermal conductivity of M, T, and I
HA_M, HA_T, HA_I	Heat atomization of M, T, and I
CE_M, CE_T, CE_I	Cohesive energy of M, T, and I

We started with twelve different ML algorithms covering both the linear and non-linear ML models namely AdaBoostRegressor (ADBR), ElasticNet (EN), GradientBoostingRegressor (GBR), KneighborsRegressor (KNR), KernelRidge (KR), Lasso, PLSRegression (PLR), RandomForestRegressor (RFR), Ridge (RR), XGBRegressor (XGBR), SVR, and DecisionTreeRegressor (DTR). Two different cross-validation methods – KfoldCV (K=10), and RepeatedKFoldCV (K=10, Repetation=5) along with mean-absolute error (MAE) as scoring metrics have been considered to check the stability and generalizability of the considered ML models. Among all the models, XGBR has been found to show lowest MAE from both CV methods.

Furthermore, to reduce dimensionality, we applied Select-K-Best feature ranking using statistical correlation between input features and the voltage target as shown in **Figure 3.5c**. This yielded a sharp drop in error once the top 39 features were retained, beyond which model performance plateaued suggesting redundancy or noise in lower-ranked features. The list of selected features for XGBR model has been tabulated in **Table 3.3**. With

the optimized feature set, hyperparameter tuning of XGBR was conducted using RandomSearchCV to minimize the MAE on the test set. As depicted in **Figure 3.5d**, the resulting model achieved an overall MAE of 0.381 V, with highest deviations observed for Li (0.441 V) and K (0.427 V), and the lowest for Na (0.285 V). These values clearly fall short of the accuracy attained by CHGNet. Thus, we have utilizing the automated pipeline we have determined the voltage of all the considered 2D electrode materials. The voltage profile for all the materials for Li-, Na-, and K-ions batteries are provided in the chapter-3 of the Github repository (https://github.com/Souvik-ml/Thesis_data).

Table 3.3: Selected features based on the Select-K-Best method during the application of XGBR algorithm.

Select-K-Best and XGBR selected features (39) | 0.39 V
Number_of_ion, AN_M, G_M, MN_M, IR_M, CR_M, PE_M, MB_M, ME_M, MV_M, VE_M, VS_M, VD_M, UVS_M, UVD_M, OSE_M, IE_M, PO_M, MP_M, BP_M, D_M, HF_M, HV_M, HA_M, CE_M, G_T, MV_T, VE_T, VP_T, VD_T, UVS_T, MP_T, BP_T, SH_T, HF_T, HV_T, TC_T, CE_T, D_I

3.4.4. Voltage Data Analysis

To evaluate the effect of terminal groups on metal-ion adsorption behavior, we have analyzed the voltage distributions for Li-, Na-, and K-ions across a diverse set of MT2-based 2D materials (**Figure 3.6**). For most of the voltage profile the path of voltage towards downside as reported previously.*[31]* The voltage axis scales differ across the subplots to accurately reflect the range of voltage values predicted for each class of terminal groups depending on the interaction with different metal-ions.

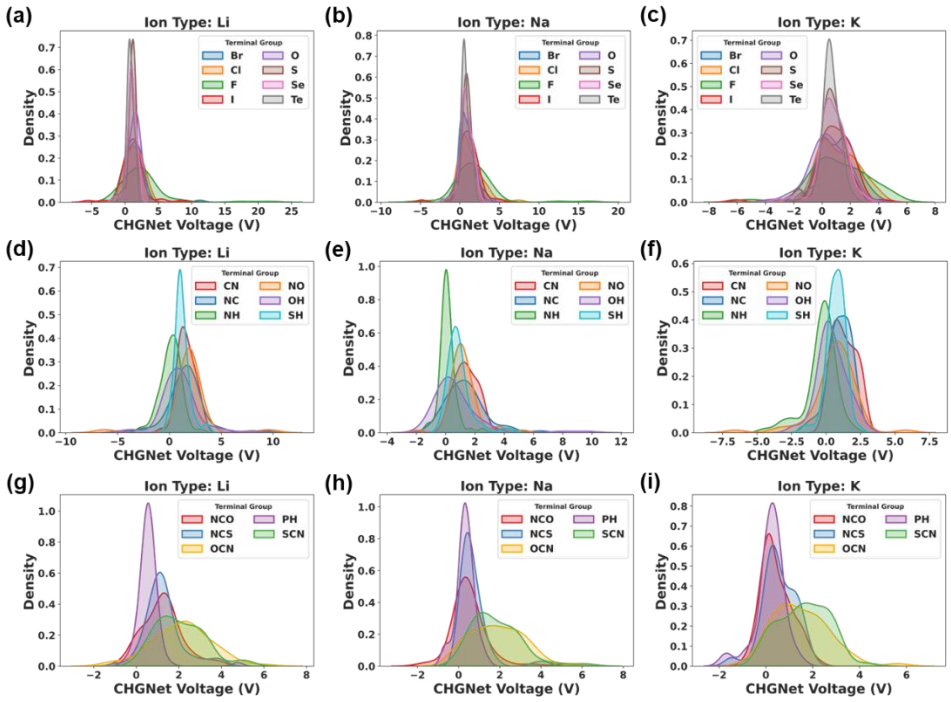


Figure 3.6: Voltage density distribution for Li-, Na-, and K-ion batteries across monoatomic (a–c), diatomic (d–f), and polyatomic (g–i) terminal groups.

In order to systematically interpret the influence of large number of terminal groups on voltage behavior, the terminal moieties have been categorized into three distinct classes based on their structural complexity such as monoatomic (O, F, Cl, Te, Br, I), diatomic (CN, NC, OH, NO, SH), and polyatomic groups (NCO, OCN, PH, SCN). Monoatomic terminal groups consist of a single atom bonded to the metal atom of the MT_2 material whereas diatomic terminal groups are composed of two atoms, and polyatomic terminal groups contain three atoms.

For monoatomic terminal groups, voltage values for Li^+ span from approximately –5 V to 10 V, though the majority of the data is tightly clustered between 0 and 3 V, with oxygen-, sulfur-, selenium- and tellurium-terminated systems showing sharp peaks around 0 to 3.5 V. These predictions align well with previous experimental studies. For instance, Te-terminated and S-terminated MT_2-type 2D anode materials have

demonstrated voltage ranges of 0.1–2.5 V and 0.005–3.0 V, respectively, in Li- and Na-ion batteries.*[32-34]* Similarly, Se-terminated materials such as WSe_2 and $MoSe_2$ have shown experimental voltage ranges consistent with our predicted values.*[35-36]* In contrast, halogen atoms, the voltage distributions become broader, indicating weaker interactions. A similar trend is observed for Na^+ and K^+, but with wider and higher distributions. For Na^+, voltages mostly fall between -5 V and 10 V, with O, S and Te again producing the most defined peaks. K^+ voltages range from -6 to 6 V, with a broad density centered between 0 and 4 V, showing increased variability in adsorption behavior across terminal atoms.

In the case of diatomic terminal groups, Li^+ voltages range from –8 to 10 V, with most systems falling between 0 and 4 V. CN-terminated structures display sharper peaks and slightly higher voltages than NC, which can be attributed to the more electronegative nitrogen atom being exposed more for interaction with the Li-cation, resulting in stronger ion-dipole interactions. OH groups show broader distributions, likely due to configurational flexibility or hydrogen bonding effects. NO-terminated systems exhibit the widest and most flattened voltage profiles, extending beyond 5 V making them least suitable terminal group for anode materials. Similar behaviors are seen for Na^+ and K^+ with these terminations, though the voltage distributions shift slightly higher for both ions. Na^+ voltages range from about –2 to 7 V, with a peak density between -1 and 3.0 V. K^+ voltages span roughly -7.5 to 7.5 V, with a flattened but high-density region between 0 and 2.5 V. The increased ionic radius of K^+ likely reduces its sensitivity to directional bonding, explaining the smoother, broader KDE curves across all groups.

For polyatomic terminal groups, the voltage distributions show distinct behavior depending on the intercalated ion. For Li^+, voltages range from –2 to 6 V, with most values concentrated between 0 and 2.5 V. Among these, PH-terminated systems exhibit the sharpest and most pronounced peak

wheras NCO and OCN produce broader voltage distributions. For Na^+, voltages span a similar range, with PH and NCS showing relatively well-defined peaks, while OCN and SCN remain more diffuse. For K^+, all polyatomic groups lead to broad voltage profiles extending up to 5 V. comparatively broad peaks for K^+ suggests weaker adsorption behavior, possibly due to the larger ionic size and lower charge density of K^+, which reduces its interaction strength with surface functional groups.

To gain deeper insight into how the metal layer influences voltage, we analyzed voltage distributions with respect to the metal's position in the periodic table. In particular, we categorized the MT_2 materials based on whether the metal originates from the s-, p-, d-, or f-block elements. The results are presented in two separate figures. Figure 3.6 displays voltage distributions for metals from the s-, p-, and f-blocks, whereas Figure 3.7 provides the corresponding analysis for the 3d-, 4d-, and 5d-transition metals.

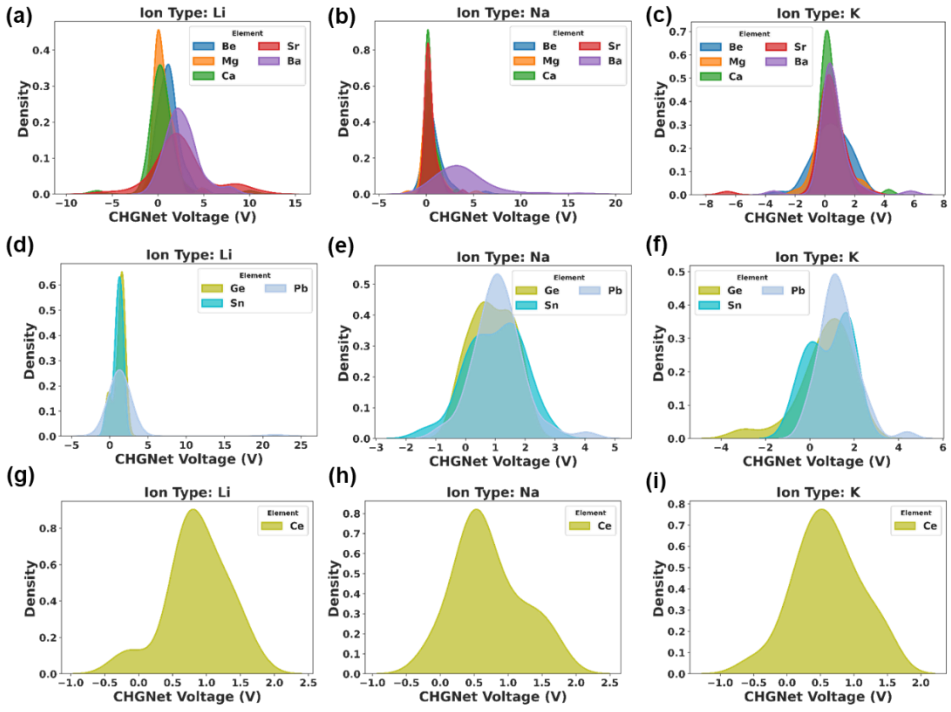


Figure 3.7: Voltage density distribution of Li-, Na-, and K-ion batteries with metal layers consisting of s-, p-, and f-block elements.

Figure 3.7 reveals that the voltage trends depend on both ionic size and electronic flexibility. For Li^+, lighter metals such as Be, Mg, and Ca yield sharp, narrow peaks centered around 0–2.5 V, indicating strong and stable Li-ion adsorption. As the metal becomes heavier, moving toward Sr and Ba, the voltage profiles broaden significantly and shift to higher values (>10 V). This trend suggests a decrease in adsorption strength and increased variability in Li-ion interaction as atomic size increases. A similar but more pronounced trend is observed for Na^+. However, Ba result in significantly higher and more dispersed voltages exhibiting long tails extending up to 15–20 V making it less favorable for Na-ion battery. For K^+, the trend is less distinct with all s-block metals, including Ba, show voltage distributions between -4 and 4 V. This uniformity suggests that K-ion adsorption is less dependent on the metal in the MT_2 structure, likely due to the larger size and lower charge density of K^+, which weakens its interaction with the host lattice. Overall, for Li^+ and Na^+ systems, lighter s-block metals promote more favorable and consistent ion adsorption whereas K^+ adsorption exhibits weaker sensitivity to metal identity, resulting in more uniform voltage behavior across the s-block series.

The voltage distributions **(Figure 3.7)** for across Ge-, Sn-, and Pb-based MT_2 materials reval that moving from Li^+ to Na^+ to K^+, the calculated voltages become more spread out and less sharply distributed. For Li^+, all three p-block (Ge, Sn, Pb) metal based electrodes perform well, with Ge and Sn giving particularly clean and stable voltage predictions. For Na^+, Sn and Ge still behave reasonably, but Pb starts to show signs of instability with a broader voltage range. However, for K^+, Pb becomes the most stable electrode, while Ge starts to misbehave, showing negative voltages and irregular peaks. The calculated result indicate Ge and Sn based electrode materials are better suited for Li- and Na-ion batteries, offering more optimum adsorption while Pb may be less ideal for smaller ions but becomes more favorable for K-ion systems.

We have also examined the influence of only f-block element for the Ce based material through voltage distributions plot which shows relatively stable and symmetric profiles across all three ion types. For both Li^+ and Na^+, voltage ranges mostly from -1.0 to 2.5 V with peak centered around ~0.75-1.0 V. For K^+, the voltage range remains consistent, centered around 0.5 V. Across all cases, Ce-based structures demonstrate moderate and uniform ion adsorption behavior.

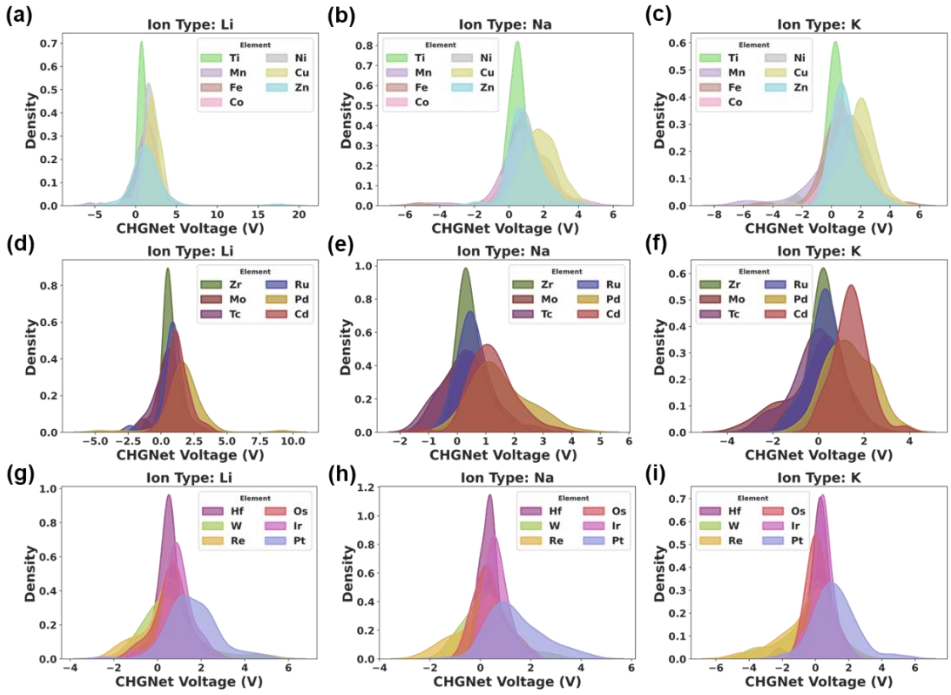


Figure 3.8: Voltage density distribution of Li-, Na-, and K-ion batteries with metal layers consisting of 3d-, 4d-, and 5d-elements.

After examining the s-, p-, and f-block elements we next explored the full range of transition metals across the 3d-, 4d-, and 5d-blocks to understand how increasing d-orbital occupancy and atomic mass influence metal-ion adsorption behavior.

As shown in **Figure 3.8**, 3d transition metal-based electrode materials exhibit wide-ranging voltage distributions for alkali-ion intercalation. Ti-based electrode materials mostly display sharp, narrow voltage peaks (~0.5–1.5 V) across Li^+, Na^+, and K^+ systems, indicating stable and consistent

adsorption behavior. In contrast, Mn-, Fe-, and Co-based electrode materials show broader distributions with negative tails, reflecting structural variability and potential instability, especially upon K^+ insertion. Zn exhibits the widest spread (–5 V to >10 V for Li^+), highlighting poor predictability and weak interaction. For Na^+, voltage profiles generally shift rightward and broaden, though Ti containing electrode materials remains relatively stable. Under K^+ insertion, most 3d metals demonstrate reduced stability, with Mn, Fe, and Co showing distributions spanning –8 V to 5 V. In the 4d series, Li^+ intercalation mostly yields narrow distributions for Zr and Ru (centered at ~1.0–1.5 V), signifying robust adsorption. Mo and Tc present broader peaks (~1.5–3.0 V), whereas Pd and Cd show dispersed profiles with high-voltage tails. Na^+ and K^+ intercalation further broaden these distributions. Among the 5d metal-based electrode materials, Hf consistently shows sharp, symmetric peaks near 0 to1 V across all ion types, suggesting strong ion-host interactions. W, Os, and Ir also exhibit favorable profiles with moderate sharpness and centering. Pt demonstrates broad, high-voltage distributions, especially with Na^+ and K^+, indicating weaker binding and possible cathodic applicability. Overall, early transition metals (Ti, Zr, Mo, Hf, W) exhibit sharp, well-centered voltage distributions and strong, predictable ion adsorption, while late transition metals (Zn, Cd, Pt) display broader, right-shifted profiles with reduced stability and weaker interactions.

3.5. Potential Electrode Materials

To identify promising 2D MT_2 type anode materials for alkali metal-ion batteries, we implemented a screening strategy that accounts for both electrochemical performance and structural stability. Since these materials are intended for anode applications, a low insertion voltage is desirable, as the overall battery voltage is determined by the potential difference between the cathode and anode, with a lower anode voltage resulting in a higher full-cell voltage. Stepwise voltage calculations during ion insertion typically show a gradual decrease with increasing ion concentration, consistent with

expected behavior in layered systems, as demonstrated in the voltage profile of $Mn(OH)_2$ during Li^+ insertion **(Figure 3.9)**.

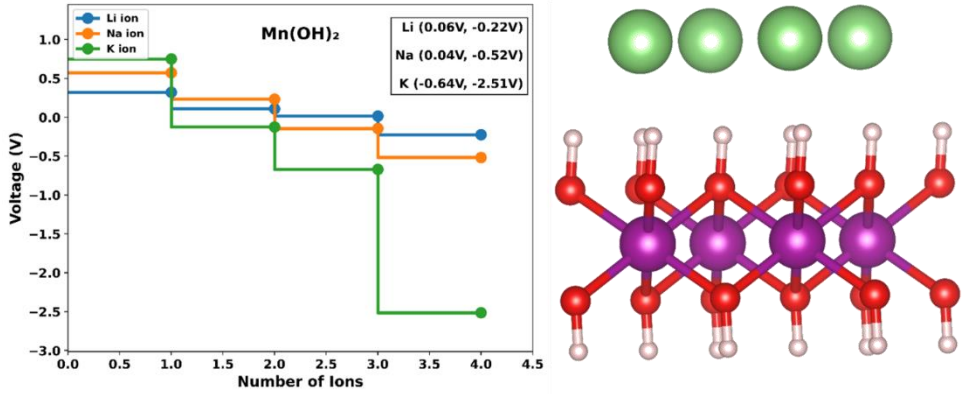


Figure 3.9: Voltage profile of $Mn(OH)_2$ and it's stable lithiated structure. The green, red, pink, and purple color sphere represent the Li, O, H, and Mn atom respectively.

This smooth voltage decline reflects stable ion accommodation within the host structure, and the voltage difference between any two consecutive steps remains below 0.26 V, indicating good electrochemical and structural stability.

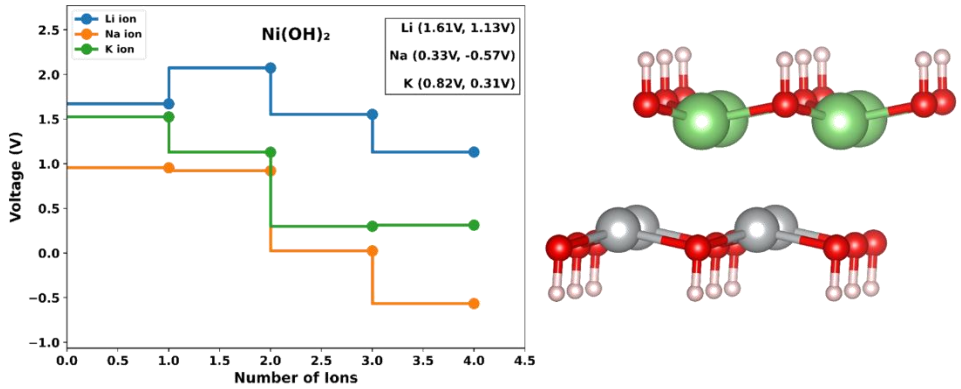


Figure 3.10: Voltage profile of $Ni(OH)_2$ and it's unstable lithiated structure. The green, red, pink, and grey color sphere represent the Li, O, H, and Mn atom respectively. In the inset the first value is the average voltage whereas the second value indicates the open circuit voltage.

However, not all materials follow this ideal trend. In several candidates, non-monotonic voltage behavior is observed, where insertion steps alternate between increasing and decreasing voltages, an effect clearly seen in the irregular voltage profile of $Ni(OH)_2$ during Li^+ insertion **(Figure 3.10)**. In such cases, the voltage difference for at least one step exceeds the 0.26 V threshold, correlating with structural distortion upon lithiation. To ensure both electrochemical and structural consistency, we imposed a screening criterion that any material exhibiting a voltage difference greater than 0.26 V between two consecutive insertion steps was considered at risk for reduced cell performance and structural distortion and was therefore excluded from further consideration.*[37, 38]*

Table 3.4: Potential electrode materials for Li-, Na-, and K-ion batteries.

Metal-ion	Potential anode materials
Li	$CaBr_2$, $MgBr_2$, $PtBr_2$, $CaCl_2$, $MgCl_2$, CaI_2, MgI_2, PtI_2, $Mg(NCO)_2$, $Pt(NCO)_2$, $Be(NCS)_2$, $Ge(NH)_2$, $Cu(NO)_2$, $Pd(NO)_2$, $Pt(NO)_2$, $Cd(OH)_2$, $Mn(OH)_2$, GeO_2, $Ce(Ph)_2$, $Hf(Ph)_2$, $Pd(Ph)_2$, $Pt(Ph)_2$, $Re(Ph)_2$, $Ti(Ph)_2$, $Zr(Ph)_2$, $IrSe_2$, $PdSe_2$, $PtSe_2$, $TiSe_2$, $ZrSe_2$, $Cd(SH)_2$, $Pb(SH)_2$, $Pt(SH)_2$, IrS_2, PdS_2, PtS_2, RuS_2, TiS_2, $GeTe_2$, $HfTe_2$, $IrTe_2$, $PdTe_2$, $PtTe_2$, $RuTe_2$, $TiTe_2$, $ZrTe_2$
Na	$BeBr_2$, $CaBr_2$, $MgBr_2$, $SrBr_2$, $CaCl_2$, $MgCl_2$, $SrCl_2$, CaF_2, MgF_2, SrF_2, BeI_2, CaI_2, MgI_2, SrI_2, $Ca(NCO)_2$, $Cd(NCO)_2$, $Mg(NCO)_2$, $Sr(NCO)_2$, $Ge(NH)_2$, $Pt(NH)_2$, $Sn(NH)_2$, $Ca(OH)_2$, $Mg(OH)_2$, $Sr(OH)_2$, $Ce(Ph)_2$, $Hf(Ph)_2$, $Ir(Ph)_2$, $Pd(Ph)_2$, $Pt(Ph)_2$, $Pb(SH)_2$, $Sr(SH)_2$, $HfTe_2$, $IrTe_2$, $PdTe_2$, $PtTe_2$, $RuTe_2$, $SnTe_2$, $TiTe_2$, $ZrTe_2$
K	$CdBr_2$, $SrBr_2$, $BaCl_2$, $CaCl_2$, $SrCl_2$, BaF_2, CaI_2, CdI_2, SrI_2, ZnI_2, $Sr(OH)_2$, $Ce(Ph)_2$, $SnSe_2$, $Ba(SH)_2$, $Pt(SH)_2$, $Sr(SH)_2$

However, certain considered materials display negative average voltages, indicating not suitable for anode material. Thus, based on the screening criteria, we identified 46 potential anode materials for Li-ion batteries, 39 for Na-ion, and 16 for K-ion systems. The complete list of selected materials is presented in **Tables 3.4**. This framework ensures that only structurally robust and electrochemically feasible materials are advanced for further development in metal-ion battery applications.

3.6. Conclusion

In this study, we present a fully automated pipeline for discovering promising two-dimensional (2D) anode materials for Li-, Na-, and K-ion batteries. The framework leverages the Crystal Hamiltonian Graph Neural Network (CHGNet) to achieve density functional theory (DFT)-level accuracy while significantly reducing computational costs. The pipeline autonomously performs key tasks, including identifying favorable adsorption sites, simulating stepwise ion insertion and extraction, calculating voltages, and generating voltage profiles. It also monitors structural changes during metal-ion adsorption, providing valuable insights into the chemical behavior of these materials.

Applying the proposed screening criteria, we identified 46 out of 289 materials as suitable anode candidates for Li-ion batteries, 39 for Na-ion, and 16 for K-ion systems. The framework requires only the unit cell structure as input and outputs corresponding voltage profiles, making it highly efficient and user-friendly. Furthermore, this approach can be readily extended to other 2D material databases, offering a practical and scalable solution for high-throughput screening. We believe this framework will support experimental efforts by narrowing down viable candidates, reducing trial-and-error, and deepening understanding of the underlying electrochemical mechanisms.

3.7. References

1. Wan Z., Chen X., Kang Y., Zhou Z., Jiang X., Xiang Z., Xu D., Luo X. (2024), Computational screening of 2D anode materials with robust thermal and electrical properties for lithium-ion batteries, J. Energy Storage, 75, 109577 (DOI: 10.1016/j.est.2023.109577)
2. Wang D., Zhang F., Wang J., Shi X., Gong P., Liu H., Wu M., Wei Y., Lian R. (2024), High-throughput screening of stable layered anode materials A_2TMO_3Cl for chloride-ion batteries, J. Mater. Chem. A, 12 (14), 8302–8310 (DOI: 10.1039/d3ta08094c)
3. Manna S., Das A., Pathak B. (2024), Machine learning assisted screening of MXene with superior anchoring effect in Al–S batteries, ACS Mater. Lett., 6 (2), 572–582 (DOI: 10.1021/acsmaterialslett.3c01043)
4. Jing Y., Zhou Z., Cabrera C. R., Chen Z. (2013), Metallic VS_2 monolayer: a promising 2D anode material for lithium ion batteries, J. Phys. Chem. C, 117 (48), 25409–25413 (DOI: 10.1021/jp410969u)
5. Fang W., Zhao H., Xie Y., Fang J., Xu J., Chen Z. (2015), Facile hydrothermal synthesis of VS_2/graphene nanocomposites with superior high-rate capability as lithium-ion battery cathodes, ACS Appl. Mater. Interfaces, 7 (23), 13044–13052 (DOI: 10.1021/acsami.5b03124)
6. Zhang X., Yu Z., Wang S. S., Guan S., Yang H. Y., Yao Y., Yang S. A. (2016), Theoretical prediction of MoN_2 monolayer as a high capacity electrode material for metal ion batteries, J. Mater. Chem. A, 4 (39), 15224–15231 (DOI: 10.1039/c6ta07065e)
7. Wan M., Zhao S., Zhang Z., Zhou N. (2022), Two-dimensional BeB_2 and MgB_2 as high capacity Dirac anodes for Li-ion batteries: a DFT study, J. Phys. Chem. C, 126 (23), 9642–9651 (DOI: 10.1021/acs.jpcc.2c02563)
8. Manna S., Roy D., Das S., Pathak B. (2022), Capacity prediction of K-ion batteries: a machine learning based approach for high throughput

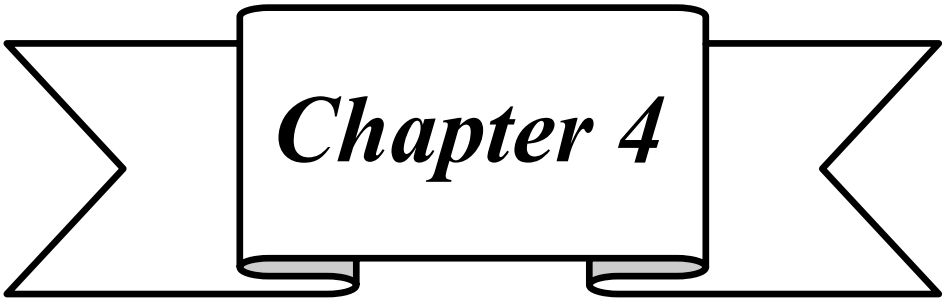
screening of electrode materials, Mater. Adv., 3 (21), 7833–7845 (DOI: 10.1039/d2ma00746k)

9. Manna S., Manna S. S., Pathak B. (2024), Integrated supervised and unsupervised machine learning approach to map the electrochemical windows over 4500 solvents for battery applications, ACS Appl. Mater. Interfaces, 16 (32), 42138–42152 (DOI: 10.1021/acsami.4c06243)
10. Manna S., Manna S. S., Das S., Pathak B. (2023), Metal-solvent interaction contribution on voltage for metal ion battery: an interpretable machine learning approach, Electrochim. Acta, 467, 143148 (DOI: 10.1016/j.electacta.2023.143148)
11. Liu Y., Zou X., Ma S., Avdeev M., Shi S. (2022), Feature selection method reducing correlations among features by embedding domain knowledge, Acta Mater., 238, 118195 (DOI: 10.1016/j.actamat.2022.118195)
12. Liu Y., Ge X., Yang Z., Sun S., Liu D., Avdeev M., Shi S. (2022), An automatic descriptors recognizer customized for materials science literature, J. Power Sources, 545, 231946 (DOI: 10.1016/j.jpowsour.2022.231946)
13. Liu Y., Wang S., Yang Z., Avdeev M., Shi S. (2023), Auto-MatRegressor: liberating machine learning alchemists, Sci. Bull. (Beijing), 68 (12), 1259–1270 (DOI: 10.1016/j.scib.2023.05.017)
14. Deng B., Zhong P., Jun K. J., Riebesell J., Han K., Bartel C. J., Ceder G. (2023), CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling, Nat. Mach. Intell., 5 (9), 1031–1041 (DOI: 10.1038/s42256-023-00716-3)
15. Artrith N., Morawietz T., Behler J. (2011), High-dimensional neural-network potentials for multicomponent systems: applications to zinc oxide, Phys. Rev. B, 83 (15), 153101 (DOI: 10.1103/physrevb.83.153101)

16. Zhang L., Wang H., Car R., Weinan E. (2021), Phase diagram of a deep potential water model, Phys. Rev. Lett., 126 (23), 236001 (DOI: 10.1103/physrevlett.126.236001)
17. Batzner S., Musaelian A., Sun L., Geiger M., Mailoa J. P., Kornbluth M., Molinari N., Smidt T. E., Kozinsky B. (2021), E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nat. Commun., 13 (1), 1–11 (DOI: 10.1038/s41467-022-29939-5)
18. Takamoto S., Izumi S., Li J. (2022), TeaNet: universal neural network interatomic potential inspired by iterative electronic relaxations, Comput. Mater. Sci., 207, 111280 (DOI: 10.1016/j.commatsci.2022.111280)
19. Chen C., Ong S. P. (2022), A universal graph deep learning interatomic potential for the periodic table, Nat. Comput. Sci., 2 (11), 718–728 (DOI: 10.1038/s43588-022-00349-3)
20. Choudhary K., De Cost B., Major L., Butler K., Thiyagalingam J., Tavazza F. (2023), Unified graph neural network force-field for the periodic table: solid state applications, Digit. Discov., 2 (2), 346–355 (DOI: 10.1039/d2dd00096b)
21. Ko T. W., Finkler J. A., Goedecker S., Behler J. (2021), A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer, Nat. Commun., 12 (1), 1–11 (DOI: 10.1038/s41467-020-20427-2)
22. Zubatyuk R., Smith J. S., Nebgen B. T., Tretiak S., Isayev O. (2021), Teaching a neural network to attach and detach electrons from molecules, Nat. Commun., 12 (1), 1–11 (DOI: 10.1038/s41467-021-24904-0)
23. Blöchl P. E. (1994), Projector augmented-wave method, Phys. Rev. B, 50 (24), 17953 (DOI: 10.1103/physrevb.50.17953)

24. Kresse G., Joubert D. (1999), From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B, 59 (3), 1758 (DOI: 10.1103/physrevb.59.1758)
25. Kresse G., Hafner J. (1993), Ab initio molecular dynamics for liquid metals, Phys. Rev. B, 47 (1), 558 (DOI: 10.1103/physrevb.47.558)
26. Kresse G., Hafner J. (1994), Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium, Phys. Rev. B, 49 (20), 14251 (DOI: 10.1103/physrevb.49.14251)
27. Kresse G., Furthmüller J. (1996), Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci., 6 (1), 15–50 (DOI: 10.1016/0927-0256(96)00008-0)
28. Kresse G., Furthmüller J. (1996), Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B, 54 (16), 11169 (DOI: 10.1103/physrevb.54.11169)
29. Perdew J. P., Burke K., Ernzerhof M. (1996), Generalized gradient approximation made simple, Phys. Rev. Lett., 77 (18), 3865 (DOI: 10.1103/physrevlett.77.3865)
30. Grimme S., Antony J., Ehrlich S., Krieg H. (2010), A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu, J. Chem. Phys., 132 (15), 154104 (DOI: 10.1063/1.3382344)
31. Bhauriyal P., Mahata A., Pathak B. (2018), Graphene-like carbon-nitride monolayer: a potential anode material for Na- and K-ion batteries, J. Phys. Chem. C, 122 (5), 2481–2489 (DOI: 10.1021/acs.jpcc.7b09433)
32. Panda M. R., Gangwar R., Muthuraj D., Sau S., Pandey D., Banerjee A., Chakrabarti A., Sagdeo A., Weyland M., Majumder M., Bao Q., Mitra S. (2020), High Performance Lithium-Ion Batteries Using Layered 2H-$MoTe_2$ as Anode, Small,16, 2002669 (DOI: 10.1002/smll.202002669)

33. Zhu C., Mu X., van Aken P. A., Yu Y., Maier J. (2014), Single-Layered Ultrasmall Nanoplates of MoS_2 Embedded in Carbon Nanofibers with Excellent Electrochemical Performance for Lithium and Sodium Storage, Angew. Chem. Int. Ed., 53, 2152–2156 (DOI: 10.1002/anie.201308354)
34. Sen U. K., Johari P., Basu S., Nayak C., Mitra S. (2014), An experimental and computational study to understand the lithium storage mechanism in molybdenum disulfide, Nanoscale, 6, 10243 (DOI: 10.1039/c4nr02480j)
35. Luo Z., Zhou J., Wang L., Fang G., Pan A., Liang S. (2016), Two-dimensional hybrid nanosheets of few layered $MoSe_2$ on reduced graphene oxide as anodes for long-cycle-life lithium-ion batteries, J. Mater. Chem. A, 4, 15302 (DOI: 10.1039/c6ta04390a)
36. Yang W., Wang J., Si C., Peng Z., Zhang Z. (2017), Tungsten diselenide nanoplates as advanced lithium/sodium ion electrode materials with different storage mechanisms, Nano Research, 10 (8), 2584–2598 (DOI: 10.1007/s12274-017-1460-3)
37. Yu S., Kim S.-O., Kim H.-S., Choi W. (2019), Computational screening of anode materials for potassium-ion batteries, Int. J. Energy Res., 43, 7646–7654 (DOI: 10.1002/er.4771)



Chapter 4

Role of Metal-solvent interaction on voltage in Metal Ion Battery

4.1. Introduction

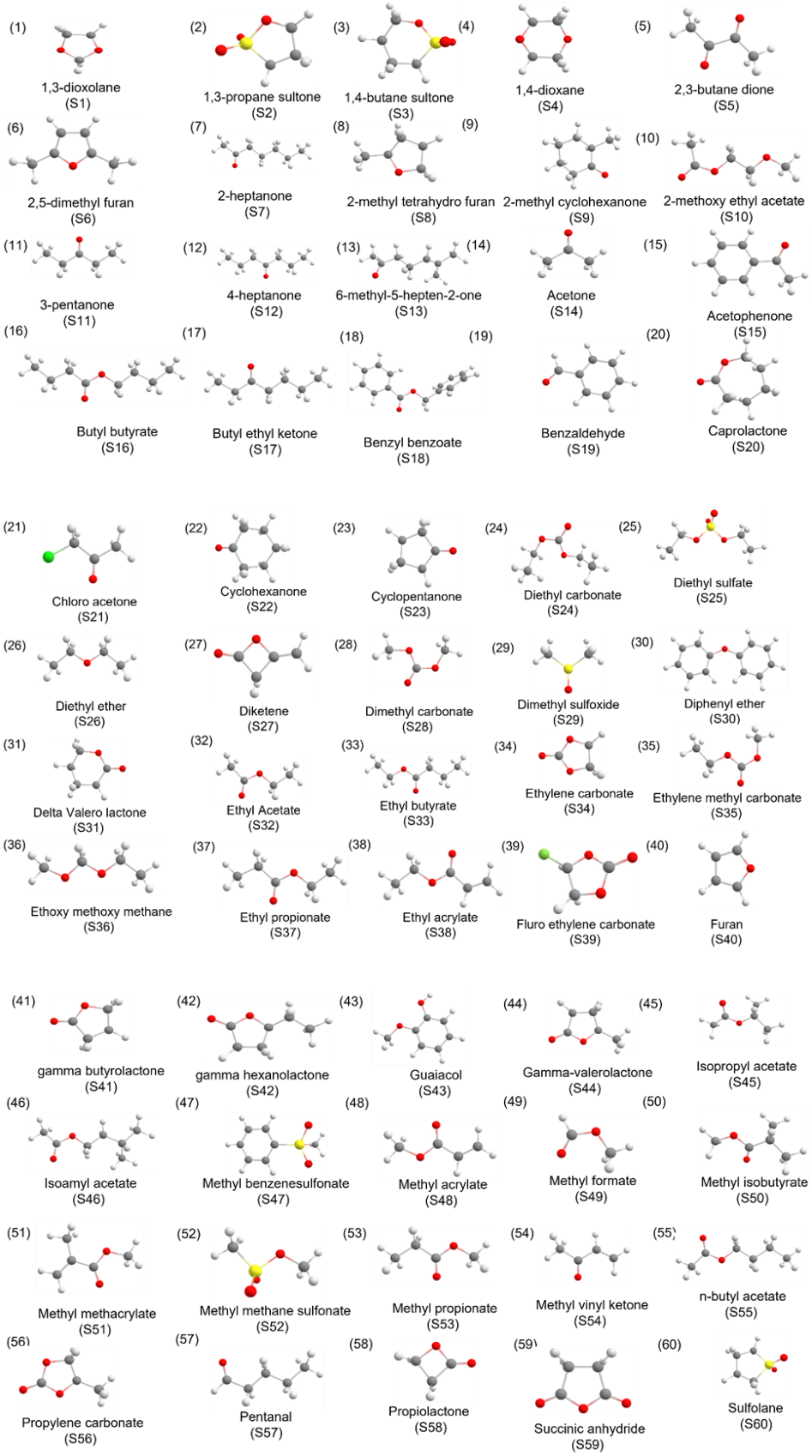
Among the metal ion batteries (MIBs), lithium-ion batteries (LIBs) have shaken up the energy storage and telecommunications sectors in terms of voltage, lifetime, and weight.*[1,2]* Extraordinary opportunities for green technologies have been developed based on LIB technology.*[1,3,4]* Despite currently dominating the market of energy storage devices, some major issues e.g., low abundance of Li raw material, high price, safety concerns, and high energy requirements demand for cheaper, sustainable, and well-performing alternate MIBs.*[5–8]* Several battery mechanisms for monovalent (Na and K) and multivalent (Mg, Ca, Al) metals have been also reported.*[5,6,9–12]* As batteries are multifaceted electrochemical ensemble composed of various components such as cathode, electrolyte, anode, separator, current collectors, etc., designing MIBs through traditional experiment and DFT-based simulation approach needs large research resources combined with sophisticated domain knowledge for the improvement of trial-and-error approach. In current years, machine learning (ML) techniques have appeared as the fourth paradigm of materials research in parallel to DFT based computational materials science.*[13–16]* ML based tools have been thriving in materials characterization, hastening atomic simulations, experimental design, and the detection of numerous functional candidates with an exceptional rate.*[17–31]* High accuracy ML models have been reported based on the combination of elemental and geometrical descriptors generated through column matrix and SOAP which are invariant with respect to translation, rotation and permutation.*[27,32–35]* Integrating ML into experimental and computational techniques has the potential to achieve success in various aspects of battery research. Voltage, capacity, and energy density being among the most important electrochemical parameters of batteries, the research community is focused on improving these parameters by investigating and selecting from a large number of electrode materials. Since the number of possible electrodes is in the order of thousands, applications of various ML techniques have been

reported for the screening of electrode materials based on voltage as the target variable.*[36–41]* Machine learning potential has been implemented by Kim et. al., to resolve the capacity fading issues due to irreversible structural instability of electrode materials.*[42]* However, there are hardly any reports regarding the role of metal-solvent interaction in the voltage determination for MIBs.

It is a general understanding that the performance of MIBs depend on the considered electrode materials as the working ions are contributed by them rather than the electrolyte. The electrolyte is only expected to maintain stability and act as a medium for the movement of ions during the working of the battery. However, we envisage that as the metal ion needs to overcome the metal-solvent interaction energy to intercalate in the electrode material, the metal-solvent interaction energy can play a very important role in battery performance in terms of voltage. There are reports regarding stability of electrolytes concerning electrode materials. It has been reported that organic solvents (dimethyl carbonate (DMC), ethylene carbonate (EC), ethyl methyl carbonate (EMC), diethyl carbonate (DEC), etc.) are well compatible with the graphite anode material for Li-ion batteries where the organic solvents get stabilized through the formation of stable solid electrolyte interface at the anode.*[43]* However, various combination of metal-solvent chemical space is very high to be explorable through experiments or conventional DFT calculations for various MIBs. Hence, attention can be paid towards data-driven ML techniques for the exploration of metal-solvent chemical space. To explore high dimensional chemical space, ML techniques has evolved as a handy tool that can carry out screening process in a very short time with minimum resources. Though metal-solvent interaction energy has been reported through the ML approach in a previous study, in that case, the interaction energy between one metal and one solvent was determined.*[44]* There are hardly any reports on the effect of metal-solvent interaction energy on voltage determination. From a practical point of view, the metal ion can be solvated by more than

one number of solvents, and thus the interaction energy will vary with the number of coordinated solvent molecules. The varying interaction energy will surely affect the half-cell voltage depending on the ease of ions getting desolvated from the solvent and intercalating or adsorbing on the electrode surface.

Here, the research work is devoted to finding out the contribution of metal-solvent interaction in battery performance through a data-driven ML approach and correlating the interaction energy with the voltage of the battery. The role of fundamental parameters affecting the metal-solvent interaction i.e., the interpretability of the utilized ML model based on the different input parameters has also been explored in our study. We have considered six metal ions (Li, Na, Mg, Al, K, and Ca) and 66 commonly used battery solvents **(Figure 4.1)** for this work. To capture a more realistic picture of metal-solvent interaction, all the possible combination has been considered for the interaction of a metal with a solvent by varying the number of coordinated solvents from 1 to 4. Therefore, for each metal, there are $66 \times 4 = 264$ combinations possible which will lead to 264 different voltage values for a particular electrode material for a single MIB. Thus, we have opted to utilize various ML models for the prediction of interaction energy for six metal systems (Li, Na, Mg, Al, K, and Ca) thereby resulting in $264 \times 6 = 1584$ voltage values. The novelty of the work lies in proposing the correlation among interaction energy and voltage for the first time through ML techniques as well as interpreting the feature importance towards the predicted values. We believe our result can be very handy for the experimental researchers to further investigate the proposed metal-solvent combination for the improvement of battery performance.


(1)
1,3-dioxolane
(S1)
(2)
1,3-propane sultone
(S2)
(3)
(4)
1,4-butane sultone
(S3)
1,4-dioxane
(S4)
(5)
2,3-butane dione
(S5)
(6)
2,5-dimethyl furan
(S6)
(7)
2-heptanone
(S7)
(8)
(9)
2-methyl tetrahydro furan
(S8)
2-methyl cyclohexanone
(S9)
(10)
2-methoxy ethyl acetate
(S10)
(11)
3-pentanone
(S11)
(12)
4-heptanone
(S12)
(13)
(14)
6-methyl-5-hepten-2-one
(S13)
Acetone
(S14)
(15)
Acetophenone
(S15)
(16)
Butyl butyrate
(S16)
(17)
Butyl ethyl ketone
(S17)
(18)
(19)
Benzyl benzoate
(S18)
Benzaldehyde
(S19)
(20)
Caprolactone
(S20)
(21)
Chloro acetone
(S21)
(22)
Cyclohexanone
(S22)
(23)
Cyclopentanone
(S23)
(24)
Diethyl carbonate
(S24)
(25)
Diethyl sulfate
(S25)
(26)
Diethyl ether
(S26)
(27)
Diketene
(S27)
(28)
Dimethyl carbonate
(S28)
(29)
Dimethyl sulfoxide
(S29)
(30)
Diphenyl ether
(S30)
(31)
Delta Valero lactone
(S31)
(32)
Ethyl Acetate
(S32)
(33)
Ethyl butyrate
(S33)
(34)
Ethylene carbonate
(S34)
(35)
Ethylene methyl carbonate
(S35)
(36)
Ethoxy methoxy methane
(S36)
(37)
Ethyl propionate
(S37)
(38)
Ethyl acrylate
(S38)
(39)
Fluro ethylene carbonate
(S39)
(40)
Furan
(S40)
(41)
gamma butyrolactone
(S41)
(42)
gamma hexanolactone
(S42)
(43)
Guaiacol
(S43)
(44)
Gamma-valerolactone
(S44)
(45)
Isopropyl acetate
(S45)
(46)
Isoamyl acetate
(S46)
(47)
Methyl benzenesulfonate
(S47)
(48)
Methyl acrylate
(S48)
(49)
Methyl formate
(S49)
(50)
Methyl isobutyrate
(S50)
(51)
Methyl methacrylate
(S51)
(52)
Methyl methane sulfonate
(S52)
(53)
Methyl propionate
(S53)
(54)
Methyl vinyl ketone
(S54)
(55)
n-butyl acetate
(S55)
(56)
Propylene carbonate
(S56)
(57)
Pentanal
(S57)
(58)
Propiolactone
(S58)
(59)
Succinic anhydride
(S59)
(60)
Sulfolane
(S60)

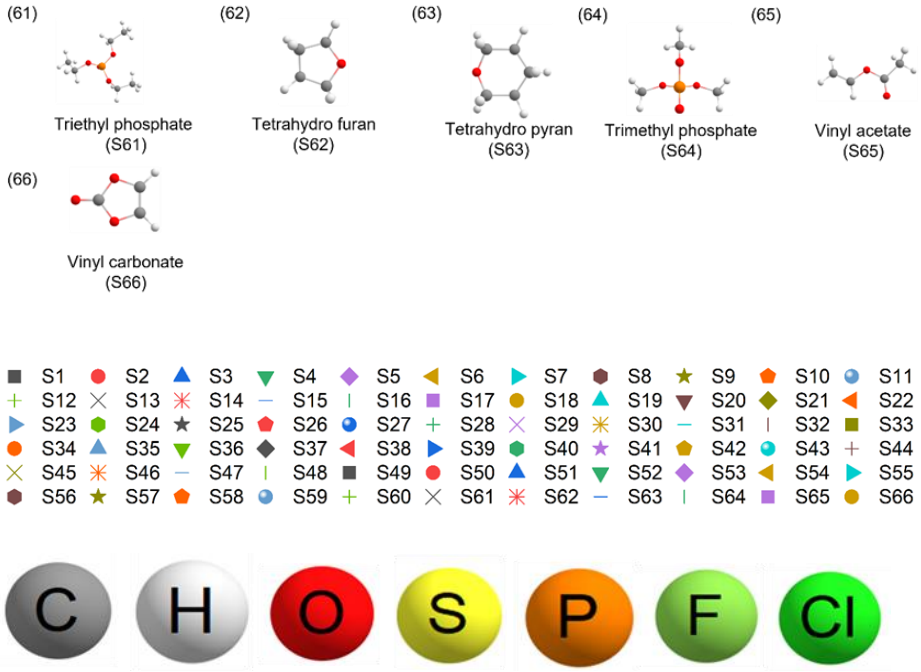


Figure 4.1: All the 66 optimized solvent structures considered for our study.

4.2. Methods

4.2.1. Computational Details

Due to unavailability of suitable database for metal-solvent interaction energies we have calculated the interaction energy of a few systems through DFT in order to use the data for training and testing of various ML models. The interaction energy must depend on the number of solvents (n) coordinating a particular metal ion. Hence, we have varied the number of solvents from 1 to 4, for a particular metal ion to calculate the metal-solvent interaction energy **(Figure 4.2)**.

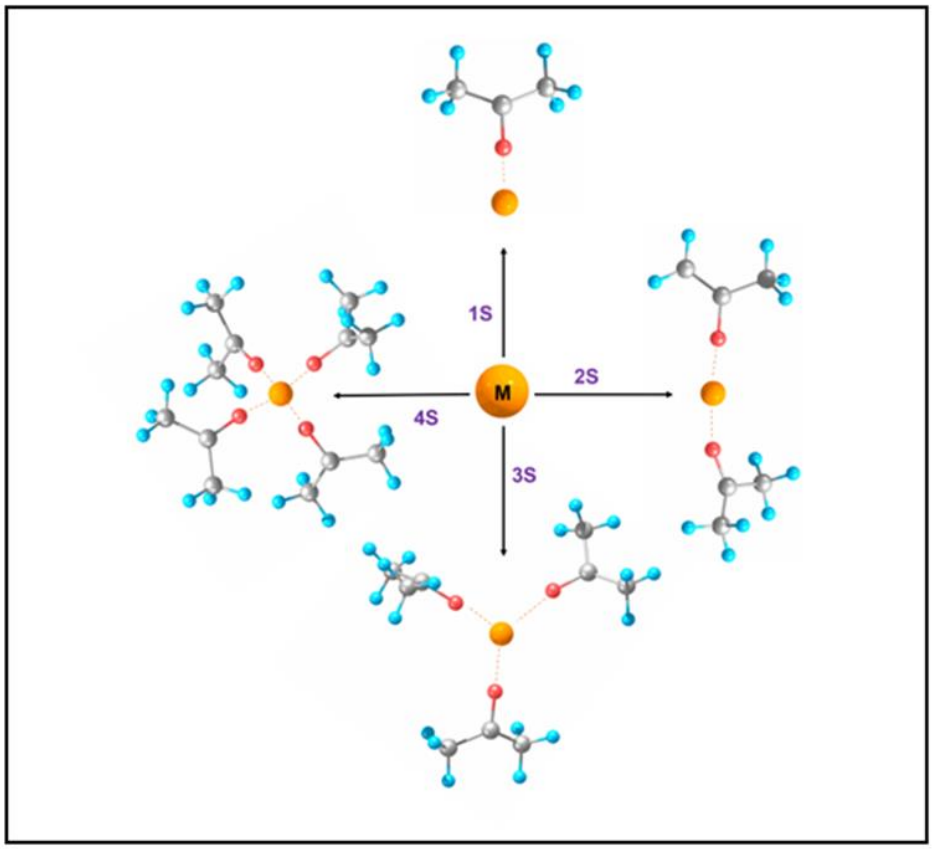


Figure 4.2: Schematic diagram of interaction energy model, where M and S stands for metal and solvent, respectively. Here acetone is considered as the sample solvent. Orange, red, grey and cyan represent metal, oxygen, carbon and hydrogen atoms, respectively.

Considering n number of solvent (S) interacting with metal (M), the interaction reaction can be shown as **equation R1**, where n varies from 1 to 4 and M are Li, Na, Mg, Al, K, and Ca.

$$M + nS \leftrightarrow MS_n \qquad (R1)$$

The interaction energy (E_{int}) of the R1 can be given as,

$$E_{int} = E_{MS_n} - E_M - n \times E_S \qquad (1)$$

Where, E_{MS_n}, E_M, and E_S are the total energies of metal-solvent system, single metal atom and solvent molecule, respectively. Thus, some combinations of M-S_n complexes **(Figure 4.2)** have been modelled and

optimized through DFT to obtain the necessary parameters for ML. The interaction energies for those combinations have also been calculated using **equation 1**. All the DFT calculations have been carried out with the Gaussian 09 package using Becke's three-parameter hybrid exchange functional and Lee–Yang–Parr's correlation functional (B3LYP).*[45]* The Pople diffuse basis set 6-31++G(d,p) was considered for all elements.*[46–48]* Non-covalent interactions have been considered utilizing Grimme's DFT-D3 potential.*[49]*

4.2.2. Machine Learning

First, we describe the data pre-processing which includes the selected suitable input features, the global correlation matrix followed by the application of various ML models. The performance of utilized ML models has been assessed by comparing the performance metrics, RMSE (root mean squared error) and MAE (mean absolute error). The final optimized ML model has been used to predict the E_{int} for the rest of the metal-solvent combination. Further, we have determined the voltage from the predicted E_{int} for all combinations of a half-cell by fixing anode material for comparison which led us to find out the effect of metal-solvent interaction in voltage determination for MIBs. Later we have discussed about application of model dependent interpretable ML technique to interpret the ML results. We have analysed our results and calculate the global and local feature importance of the considered features using shapash package namely shapash analysis.

4.3. Results and Discussion

4.3.1. Data pre-processing

The known database (generated through DFT) consists of 225 data points including six metal ions that interact with different solvents. The percentage of six metals considered interacting with various solvents utilized for the training and testing of ML models has been shown in **Figure 4.3**. We have

included all the metal-solvent data in some extent in the training set so that the ML predicted result must not be biased for a particular metal.

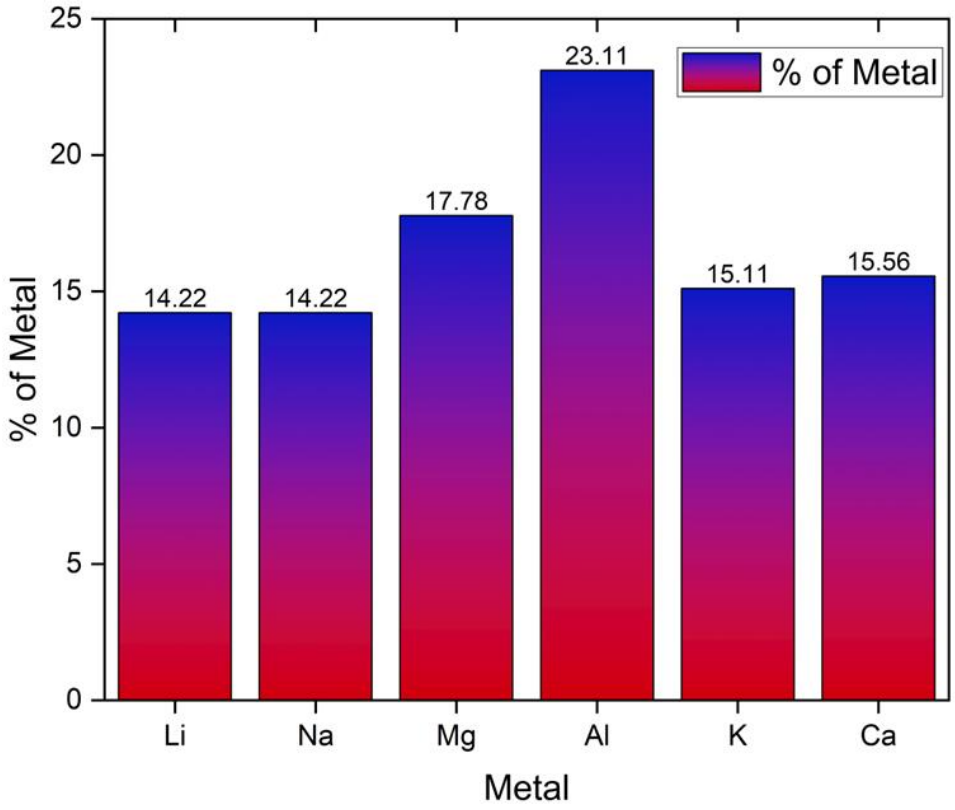


Figure 4.3: Percentage of metals in the DFT calculated dataset of 225 metal-solvent combinations.

ML studies on well sampled small data set has already reported.*[50–53]* Though the numbers of data for training is less, we have tried to generate the data in such a way that it should be well sampled. Thus, we have incorporated the homogeneity in the known data set so that the small data can be used farther to train various ML models. Considering the interaction energy as the target variable, suitable features have been selected to define the feature space needed for ML models. The features (charge on electronegative atom, dipole moment, HOMO and LUMO energy among others) have been selected based on the domain knowledge. To predict interaction energy, we have considered only those features which are important for the same. For example, we have considered formula weight

instead of molecular weight as the number of a particular solvent surrounding a metal could be an important factor for determining metal-solvent interaction along with number of C, H, O, S, F, Cl, P atom and natural charge on the electronegative atom. All the selected features have been tabulated in **Table 4.1**. For effective ML applications, features must be invariant against translation, rotation and permutation in order to have high-accuracy prediction models.*[54]* Our selected features satisfy this criterion of invariance. The features are selected in such a way that they can represent each system appropriately. Here, the total energy refers to the electronic energy of the solvent molecules. The total energy of solvent has been considered as input vector to incorporate the electronic properties of the different solvents.

Table 4.1: The considered features for the preparation of feature space. Elemental properties (1,2,3), and physical properties (19,20,21,22) have been taken from the literature, and rest of the features' value determined through DFT calculations.

	Features
1,2,3	Ionic radius (1), electronegativity (2), atomic number (3) of metal
4	Number of solvent (n)
5	Total energy of solvent (E_s)
6	Total energy of Metal (E_M)
7,8,9,10,11,12,13	Number of C atom (7), H atom (8), O atom (9), S atom (10), F atom (11), Cl atom (12), P atom (13)
14	Formula weight of solvent (FW)
15	Dipole moment of solvent (TD)
16	Natural charge on electronegative atom of solvent
17, 18	LUMO (17) and HOMO (18) of solvent
19,20, 21, 22	Boiling point (19), melting point (20), flashing point (21), density (22) of solvent

Pearson's correlation matrix has been plotted **(Figure 4.4)** to analyse the correlation among the considered features. The matrix delivers the correlation coefficient among every pair of the input features as well as with the output (target variable). The ticks in **Figure 4.4** correspond to the considered features in **Table 4.1**. It should be noted that tick number 23 in the correlation plot is the target variable which has been included in the plot to find out the correlation among features and the target variable. Tick number 12 i.e., number of Cl atom in the solvent has been dropped from the heat map since there is no such solvent in the train set where Cl atom is present. However, in the overall database, there are Cl atom in some solvent, and hence we have considered the number of Cl as feature.

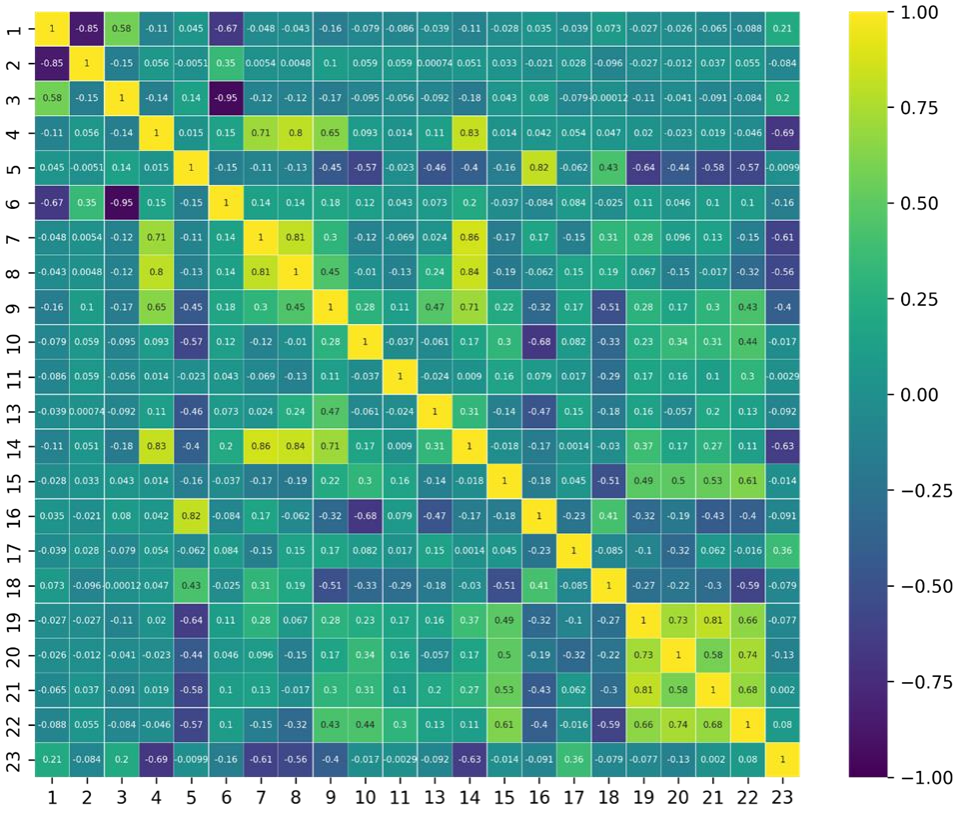


Figure 4.4: Pearson's correlation matrix regarding the correlation among the input features (1-22, **Table 4.1**) as well as with the target variable interaction energy (23).

From **Figure 4.4**, a strong negative correlation (-0.85) between feature 1 (ionic radius) and feature 2 (electronegativity) has been observed. Similarly, there is a strong negative correlation (-0.95) between feature 3 (atomic number of metal) and feature 6 (energy of metal). A moderately strong positive correlation between feature 7 (number of C atom) and 14 (formula weight) as well as feature 8 (number of H atom) and 14 (formula weight) is observed. This is because, with the increase in number of C and H atoms, formula weight also increases, which leads to a linear relationship among these features. The target variable is found to have a strong negative correlation with features 4 (number of solvent) and 7 (number of C atom) which is obvious since the number of solvents varying affects the interaction energy inversely. It has been found that feature 17 (LUMO) varies proportionally with the target variable whereas feature 18 (HOMO) varies inversely, which indicates that wider the HOMO-LUMO gap higher can be the interaction energy. Since we have not found any correlation coefficient greater than 0.95, this shows that the selections of features are appropriate, and all the features have been further processed to be considered for training, testing, and prediction.

4.3.2. ML methods and interaction energy

For our supervised ML models, the known data set has been split into a train set and test set in the 80:20 ratio.*[55–58]* The train set and the test set remain same throughout the work so that the results are reproducible. Nine ML algorithms namely, Ridge Regression (RR), Kernel Ridge Regression (KRR), Random Forest Regression (RFR), Extra Trees Regression (EXR), XGBoost Regression (XGBR), Gradient Boosting Regression (GBR), AdaBoost Regression (ADBR), LightGBM and Neural Network (NN) have been applied to train the machine and further, the optimized trained model has been utilized for the prediction of interaction energy of test set. The performance of each ML model has been evaluated by comparing three performance metrics, namely root mean square error (RMSE), mean

absolute error (MAE), and R^2 defined as equation (i), (ii), and (iii) respectively,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i}^{N}(y_i - y_p)^2} \qquad (i)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i}^{N}(y_i - y_p)^2 \qquad (ii)$$

$$R^2 = 1 - \frac{\sum_{i}^{N}(y_i - y_p)^2}{\sum_{i}^{N}(y_i - y_a)^2} \qquad (iii)$$

Here, y_i, y_p and y_a indicate DFT calculated, ML predicted and average interaction energies, respectively. N is the total number of data point in the test set. To find out the best fitted line from each ML models and to increase the prediction accuracy, all the hyperparameters corresponding to each model has been tuned utilizing the RandomizedSearchCV as executed in scikit-learn library.*[59]* The calculated MAE and RMSE values along with the optimized hyperparameters for each regression ML models have been tabulated in the **Table 4.2**.

Table 4.2: ML models utilized for the prediction of interaction energy along with the optimized hyperparameters, RMSE and MAE.

ML Models	**Optimized hyperparameters**	**RMSE (eV)**	**MAE (eV)**
RR	Alpha=0.56, random_state=42	0.680	0.462
KRR	alpha=0.1, gamma=0.001, kernel=Laplacian, coef0=2, degree=1	0.650	0.462
RFR	max_depth=30, min_samples_leaf=5, min_samples_split=8,	0.544	0.354

	n_estimators=200, random_state=42		
EXR	max_depth=20, max_features=19, max_samples_split=4, min_samples_leaf=3, max_leaf_nodes=25, random_state=42	0.517	0.380
XGBR	learning_rate=0.005, n_estimators=1500, random_state=42	0.517	0. 354
GBR	learning_rate=0.01, max_depth=4, max_features=10, n_estimators=500, min_samples_split=3, random_state=42	0.489	0.326
ADBR	learning_rate=0.02, loss='exponential', n_estimators=200, random_state=42	0.571	0.408
LightGBM	learning_rate=0.05, max_depth=5, n_estimators=200	0.517	0.367
NN	Optimizer=Adam, learning_rate= 0.001, Loss_function= mean absolute error, Hidden layer = 16, nodes = 160 (hidden layers)	0.734	0.490

Figure 4.5a shows the error bar on applying different ML models to fit the train set and prediction of test set. From **Figure 4.5a** and **Table 4.2**, it is evident that machine has been trained well with tree-based algorithms as there is a sharp change observed in the RMSE and MAE from KRR to RFR which indicates tree-based algorithms are the better choice for the prediction of the target variable.

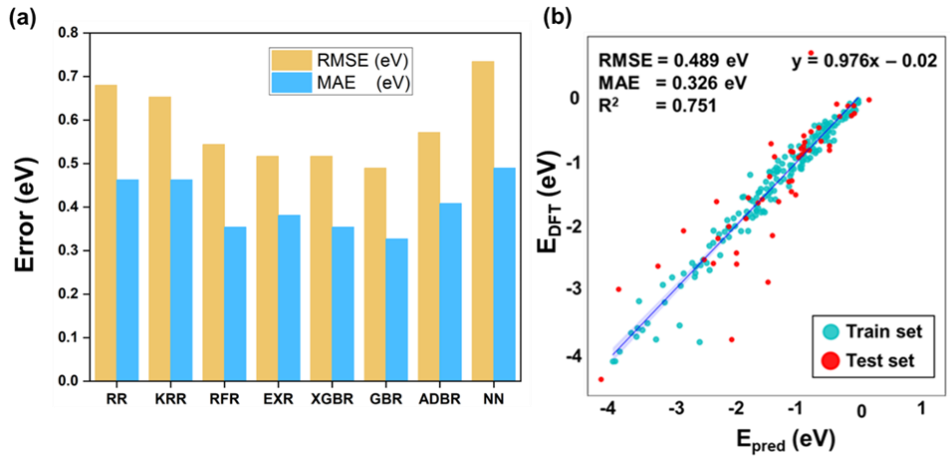


Figure 4.5: (a) Error bar plot of all the utilized ML models, and (b) scatter plot of DFT calculated interaction energy vs predicted interaction energy in GBR ML model.

Among the considered ML models, GBR model has been found to be best fitted model for the prediction of target variable. From the scatter plot **(Figure 4.5b)**, it is evident that the test data points are close to train data points but not exactly overlapping with the train data points. Some points are also far from train data points confirming that there is no overfitting in the result obtained from the ML models. The performance of EXR and XGBR are almost same as they yield same RMSE (0.517 eV) error. However, GBR model has been found to be the most effective and best fitted model with the least RMSE and MAE of 0.489 eV and 0.326 eV, respectively for the prediction of interaction energy. The loss function used in GBR is mean squared error which has been optimized with low learning

rate 0.01 **(Table 4.2)**. The R^2 value for GBR is maximum (75.1%) which indicates high accuracy for the test set **(Figure 4.5b)**.

Further, to check the stability and to ensure there is no overfitting of the GBR model, two methods, namely, K-fold cross validation and leave one out cross validation (LOOCV) have been performed.*[60]* For the K-fold cross validation, number of folds for the training set has been varied from 2 to 10. For each fold we have calculated **(Table 4.3)** the average mean absolute error ($\overline{\text{MAE}}$). The standard deviation of the K-fold cross validation is 0.011 eV. The mean absolute error of K-fold cross validation with respect to GBR model has been found to be 0.025 eV. From the standard deviation value, we can say the change in $\overline{\text{MAE}}$ is less irrespective of different K-fold CV. Hence, it confirms that our considered GBR model is transferable i.e., not overfitted and is suitable to predict the interaction energy of unknown system. Further, LOOCV method has been applied on test set and the $\overline{\text{MAE}}$ has been evaluated to check the model stability on each label explicitly. Unlike K-fold CV, in LOOCV, one sample remains in the test set and rest of the sample remains in the train set. The LOOCV helps to measure the error explicitly for each sample as in this case ML is predicting interaction energy for each label separately. After the MAE calculation of each label, average MAE on the overall train set has been determined for the comparison of MAE with the K-fold cross validation and optimized GBR model. The average MAE for LOOCV method has been found to be 0.344 eV which indicates that the error range is stable and not varying irrespective of different cross-validation method. This further shows that the consider model is not overfitted. Hence, based on the least RMSE, MAE and high R^2, GBR model has been selected for the prediction of interaction energy for all unknown system having same optimized hyperparameters. To explore the trend of interaction energy between metal ions and solvents, interaction energy with respect to number of solvents has been plotted **(Figure 4.6)**.

Table 4.3: Comparison of K-fold cross-validation (CV) with optimized GBR model predicted interaction energy, considering the loss function as mean absolute error ($\overline{\text{MAE}}$). All error units are in eV.

K-fold	$\overline{\text{MAE}}$	MAE (GBR)
2	0.373	0.326
3	0.356	
4	0.349	
5	0.356	
6	0.351	
7	0.353	
8	0.350	
9	0.330	
10	0.349	
Standard deviation of $\overline{\text{MAE}}$ = 0.011 eV		
$\overline{\text{MAE}}$ (CV) to MAE (GBR) = 0.025 eV		

It is expected that for a particular metal ion and solvent combination, with the increase in the number of solvents the interaction energy will increase as the metal ion will be coordinated with higher number of solvent molecules. From **Figure 4.6**, it has been observed that all the metals (Li, Na, Mg, Al, K and Ca) follow the expected trend i.e., the interaction energy is increasing with the increase of number of solvents thereby stabilizing the systems. The trend of interaction energy can be explained based on two effects, (a) ionic-potential effect and (b) inter-solvent repulsion effect. The ionic potential further depends on the charge to size ratio. Small size metal ion will generate high ionic potential as the charge to radius ratio will be higher compared to the large size metal ion. However, at the same time for the small size of metal ion the coordinated solvent molecules will remain close to each other leading to high inter-solvent repulsive interaction. Between these two oppositely moving deciding factors, it is expected that if the ionic-potential factor predominates over the inter-solvent repulsion,

the system will get stabilized. Whereas, if the repulsive interaction predominates over the ionic-potential factor, the system will get destabilized more with the increase in the number of solvents.

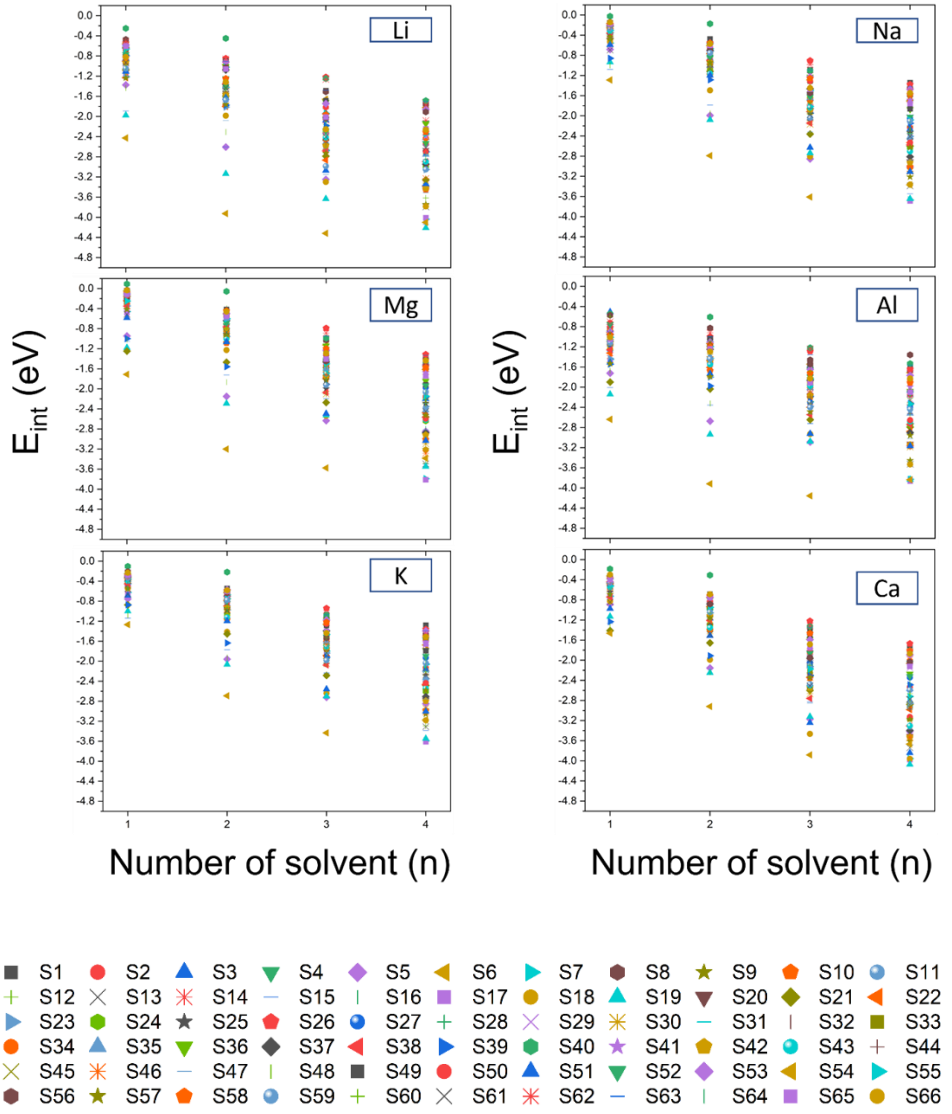


Figure 4.6: Interaction energy vs number of solvent for all considered 66 solvents where the number of a particular solvent around a metal ion (n) varies from 1 to 4. The 66 solvents have been represented as Si where 'i' varies from 1 to 66. The solvent corresponds to Si has been given in **Figure 4.1.**

In our case, the solvated metal ions are more stabilized with the increase in number of solvents which indicates that the high ionic potential is

dominating over the inter-solvent repulsive interaction. Once the interaction energies of all the system have been predicted, the next goal is to relate it with the voltage i.e., how the interaction energies affect the voltage. The voltage part has been discussed in detail in the next section.

4.3.3. Voltage

To determine the contribution of the solvent-metal interaction on voltage determination, voltage of each combination has been determined by utilizing the GBR predicted interaction energies. In order to draw a comparison between the voltage of all the system, a fixed graphite anode has been chosen for the voltage calculation of anodic half-cell. Since, Gaussian09 package does not fully support periodic calculations we have limited the graphite structure to a total of 48 C atoms in two layers with edge atoms saturated by H atoms.

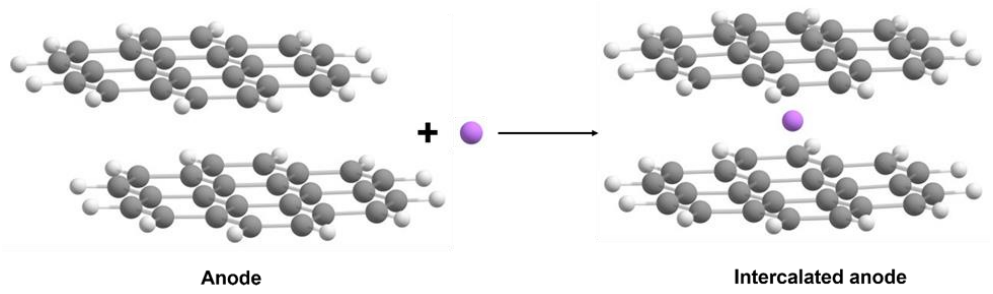


Figure 4.7: Schematic diagram showing the considered graphite anode and intercalation of metal ion during working of the half-cell.

During charging process, metal ion gets intercalated at the anode (A) as shown in **Figure 4.7.** Therefore, the reaction between anode and solvated metal ion can be written as **R2**.

$$A + MS_n \leftrightarrow A - M + nS \qquad (R2)$$

From the **R2**, the anodic voltage expression can be written as **equation 2**.

$$V_{Anode} = \frac{1}{z}\left[(E_{A-M} + n \times E_S) - (E_A + E_{MS_n})\right] \qquad (2)$$

Where z, $\mathrm{E_{A-M}}$, $\mathrm{E_S}$, $\mathrm{E_A}$, and $\mathrm{E_{MS_n}}$ represents the number of electron transfer, total energy of metal intercalated anode, solvent molecule, unintercalated anode, and metal-solvent system, respectively. The anodic voltage can be expressed in terms of $\mathrm{E_{int}}$ **(equation 3)** by substituting the value of $((n \times E_s) - E_{MS_n})$ form the **equation 1** in **equation 2**.

$$\mathrm{V_{Anode}} = \frac{1}{Z}[\mathrm{E_{A-M}} - \mathrm{E_A} - \mathrm{E_M} - \mathrm{E_{int}}] \quad (3)$$

Using the **equation 3** we have calculated the voltage for all the considered MIBs. Since the anodic voltage has been considered, lesser the anodic voltage higher will be the overall cell voltage. To compare how the number of coordinated solvents around a particular metal ion affects the voltage, we have plotted the voltage with respect to number of solvents for each individual metal ion **(Figure 4.8)**.

From the **Figure 4.8**, it has been observed that with increase in number of solvents around a particular metal ion, the anodic voltage increases (reverse trend of interaction energy) which is expected since strongly solvated metal ion will require more voltage to overcome metal-solvent interactions for the intercalation of metal ion in the anode. However, the voltage scale of all the metal ions is not same which is obvious due to different interaction of different metal ions with the solvents. The overall voltage trend for all solvents can be explained from the **equation 3** and the interaction energy trend in terms of E_{int} and E_{A-M}. For a particular metal ion, the first three terms i.e., E_{A-M}, E_A and E_M are constant. With the increase in number of solvents, E_{int} becomes more negative leading to increase in V_{Anode}. Voltage of all the 1584 systems have been provided in the Chapter-4 of Github repository (https://github.com/Souvik-ml/Thesis_data). We further determine the average voltage for all the 66 solvents and six metal combination, so that we can comment on the voltage of a MIB corresponding to each solvent. The voltage has been determined by taking the average of all four voltage for each solvent for a particular MIB. To

understand the change in average voltage with respect to change in average interaction energy, a combined average voltage-interaction energy plot for six MIBs and 66 solvents have been shown in **Figure 4.9**. As is evident from **equation 3**, the variation in average interaction energy and average voltage also shows an inverse trend in **Figure 4.9**.

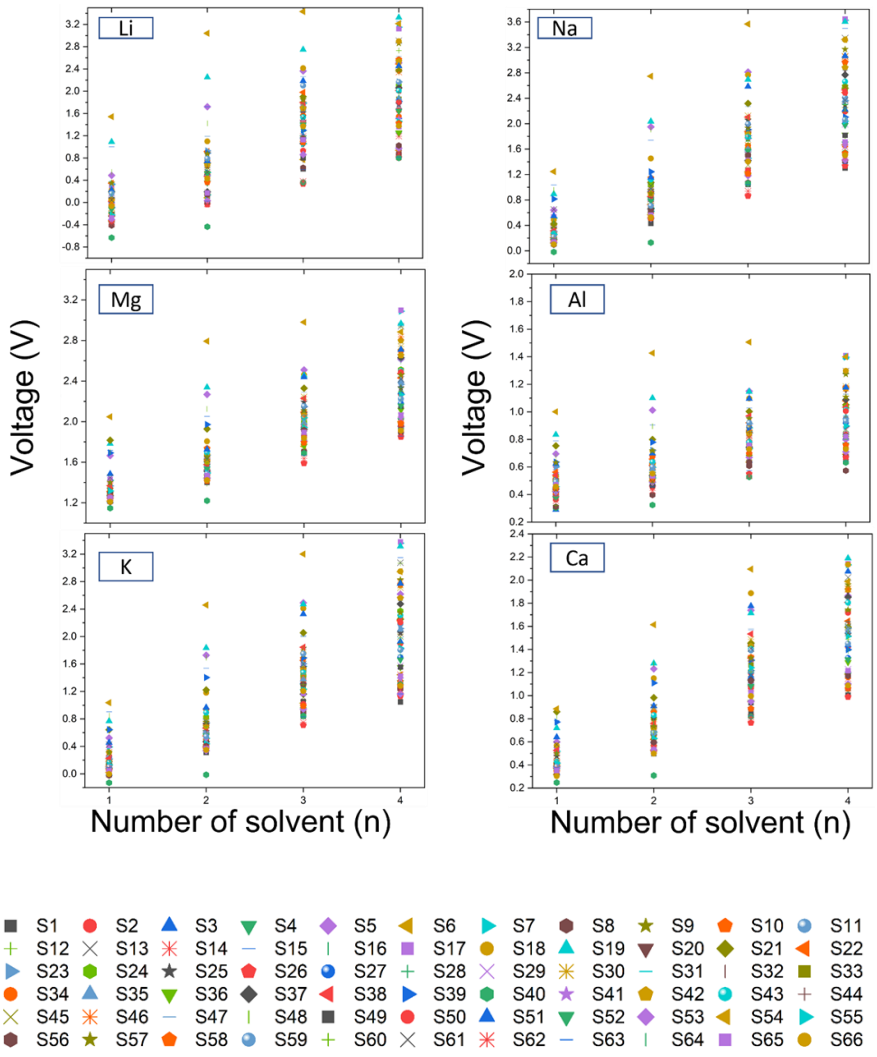


Figure 4.8: Voltage of each MIBs for all considered 66 solvents where the number of a particular solvent around a metal ion (n) varies from 1 to 4.

Each MIB has 66 average voltages corresponding to 66 solvents. Based on the average voltage we have proposed five most optimum solvent systems for minimum half-cell voltage of each MIB, and the result has been tabulated in **Table 4.4**. Average voltage of all the metal-solvent

combination has been provided in the Chapter-4 of Github repository (https://github.com/Souvik-ml/Thesis_data).

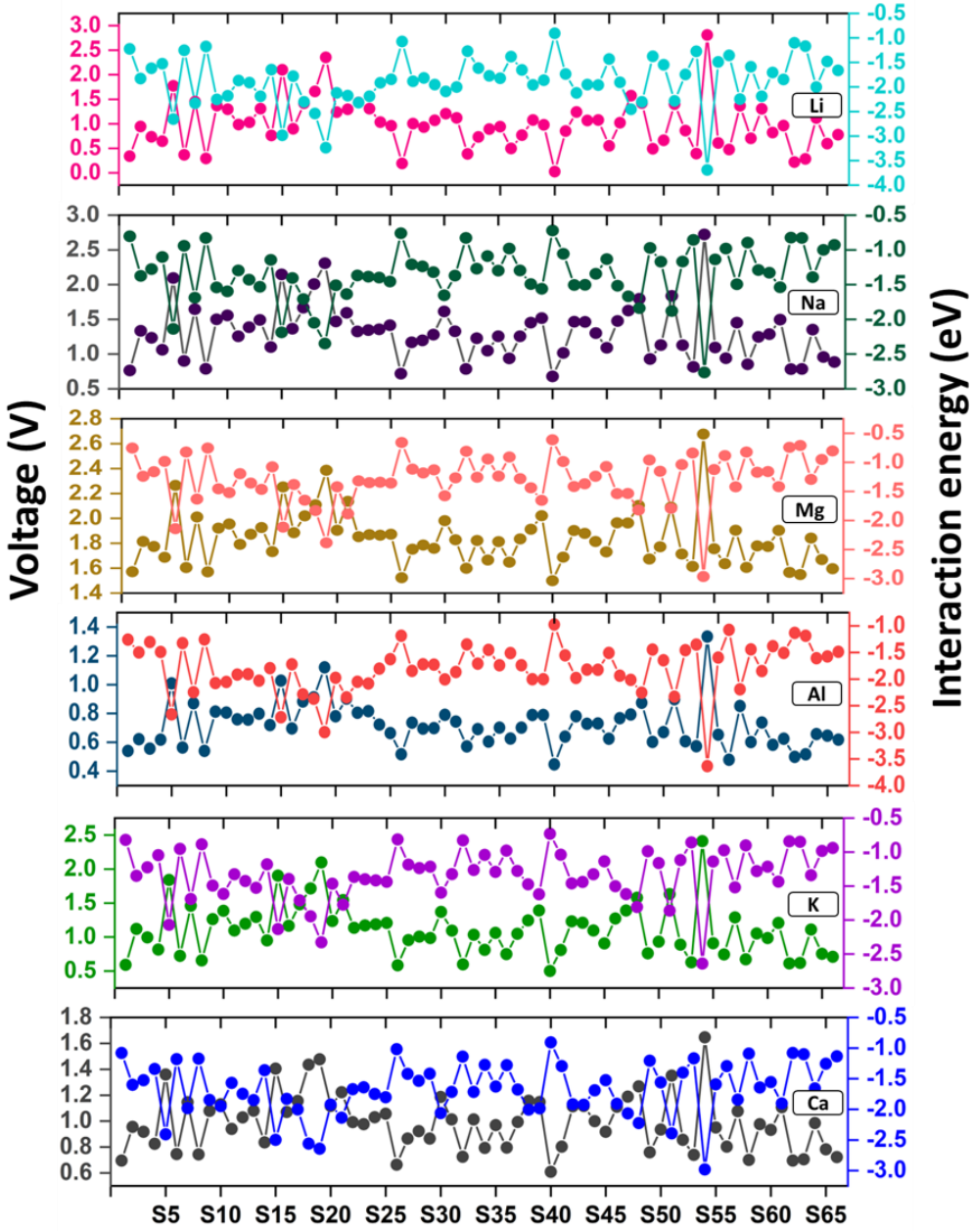


Figure 4.9: Average voltage and interaction energy plot for the combination of all considered six metals and 66 solvents.

Table 4.4: Proposed five best metal-solvent combinations for each MIB and their corresponding average voltage.

Metals	Solvent				
Li	Furan 0.024 V	Diethyl ether 0.190 V	Tetrahydro furan 0.217 V	Tetrahydro pyran 0.284 V	2-Methyl tetrahydrof uran 0.292 V
Na	Furan 0.679 V	Diethyl ether 0.718 V	1 ,3-dioxolane 0.763 V	Tetrahydro furan 0.783 V	Ethyl acetate 0.786 V
Mg	Furan 1.499 V	Diethyl ether 1.523 V	Tetrahydro pyran 1.550 V	Tetrahydro furan 1.565 V	2-Methyl tetrahydrof uran 1.571 V
Al	Furan 0.447 V	Propyle ne carbonat e 0.479 V	Tetrahydro furan 0.499 V	Tetrahydro pyran 0.516 V	Diethyl ether 0.516 V
K	Furan 0.501 V	Diethyl ether 0.584 V	1,3-dioxolane 0.591 V	Ethyl acetate 0.598 V	Tetrahydro furan 0.612 V
Ca	Furan 0.609 V	Diethyl ether 0.664 V	Tetrahydro furan 0.695 V	1,3-dioxolane 0.695 V	Propiolacto ne 0.701 V

We have also compared our predicted average voltage with some of the experimentally reported anodic voltage. For Li-ion battery with ethylene carbonate electrolyte and graphite anode, our ML predicted average voltage of 0.89 V is comparable to experimentally reported anodic voltage of 0.9 V at low specific capacity.*[61]* For Na-ion battery with propylene carbonate electrolyte and graphite anode, our ML predicted average voltage of 0.94 V

is also close to experimentally reported anodic voltage (~1 V) for propylene carbonate with hard carbon electrode.*[62]* Our ML predicted voltage cannot be compared exactly with theoretically reported voltage since in most of the cases only intercalation energy is considered without metal-solvent interaction energy. Therefore, we propose that inclusion of metal-solvent interactions can give more accurate anodic half-cell voltage values closer to experimental scenario.

4.3.4. Local Feature Analysis

The utilized ML models are black box models which are very hard to explain. Though the explanation of these models is difficult, still the learning ability of these models are high compared to simple interpretable ML models. Here, we have implemented a model dependent machine interpretable shapash library to extract knowledge from the black box model, GBR. Cooperative game theory based shapash library helps us to understand the contribution of local and global feature importance towards the predicted output. Shapash uses Shapley Additive explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) in backend to compute the feature contribution. The Shapley value (ϕ_i) signifies the importance of each feature and it can be computed using **equation 4**.*[63]*

$$\phi_i = \sum_{S \subseteq F, \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right] \quad (4)$$

Where, F represents all set of features and S indicates the subset of all features obtained from F after removing of ith feature. $f_{S \cup \{i\}}$ and f_S are the prediction model of with and without ith feature, respectively. x_S represents the value of the input features in the S set. In LIME, the prediction of the model is explained by approximating the decision boundary locally around the input data. It creates a sample perturbed instance which can be explained by **equation 5**.*[64]*

$$explanation\ (x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (5)$$

Where f is the probability for an instance x, g is the explanation model, π_x is the proximity measure between different instances, and L is the loss function which needs to be minimized for better model performance. Ω(g) measure the complexity of the model g. For tree-based algorithm $\Omega(g)$ may be depth or trees whereas for linear model it can be number of many non-zero weights. Hence, LIME is more prone to explain the local parameters whereas the SHAP is expert for the understanding of global behavior of a model. Both methods have been implemented in the shapash model to gain a comprehensive understanding for a particular model. The local feature importance is necessary as it can comment on the ML predicted result as accurate or not based on the feature contribution of each label explicitly. First utilizing shapash, the global contribution of feature towards the target variable has been determined **(Figure 4.10)**.

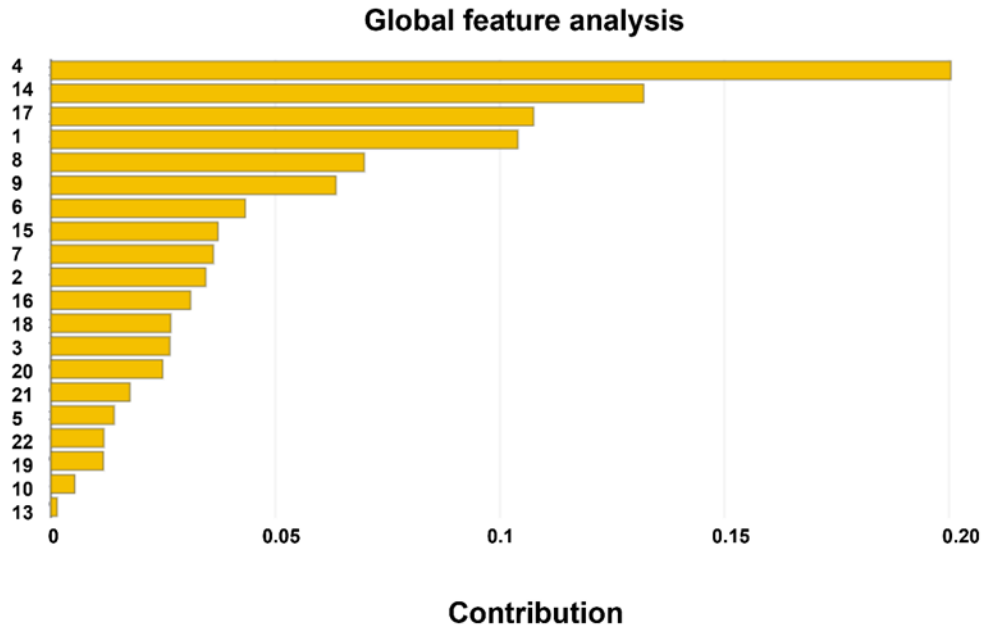


Figure 4.10: Global feature analysis of each feature towards the target variable (interaction energy) utilizing GBR model. The ticks in the y axis represent the feature number **(Table 4.1)**.

Figure 4.10 shows feature 4 (number of solvent) as the most contributing feature followed by feature 14 (formula weight) and feature 17 (LUMO) for

the prediction of interaction energy. This feature contribution has been determined on the whole result leading us to global feature importance. However, the feature contribution of each level can also be determined through shapash. Here, first we have determined the mean absolute deviation of prediction value from the true value. From the deviation, three systems, least deviated (most accurate prediction), most positively deviated, and most negatively deviated systems have been selected and feature contributions of each feature for these systems have been determined explicitly. Hence, we have plotted the feature importance of these three systems separately for local feature analysis through shapash **(Figure 4.11, 4.12a, 4.11b)**. In the inset of **Figure 4.11, 4.12a, 4.12b** the details of each system have been provided. The ticks in y axis shown in these plots correspond to the feature numbers provided in **Table 4.1**. It has been observed that the feature contribution trend in global feature analysis **(Figure 4.10)** and local feature analysis of the least deviated system (**Figure 4.11)** is almost same. For example, in both cases feature 4 (number of solvent) and 14 (formula weight) are the most contributing features towards the target property (interaction energy). **Figure 4.11** shows the most accurate prediction of interaction energy by GBR model and the feature importance of these features particularly for this system. The result shows contribution of feature 4 (number of solvent) is high and in the positive direction followed by the feature 14 (formula weight) and 1(ionic radius). Here, the positive and negative contributions are considered as coefficient of those features, not necessarily reflecting that they contribute positively and negatively towards the error of the property.

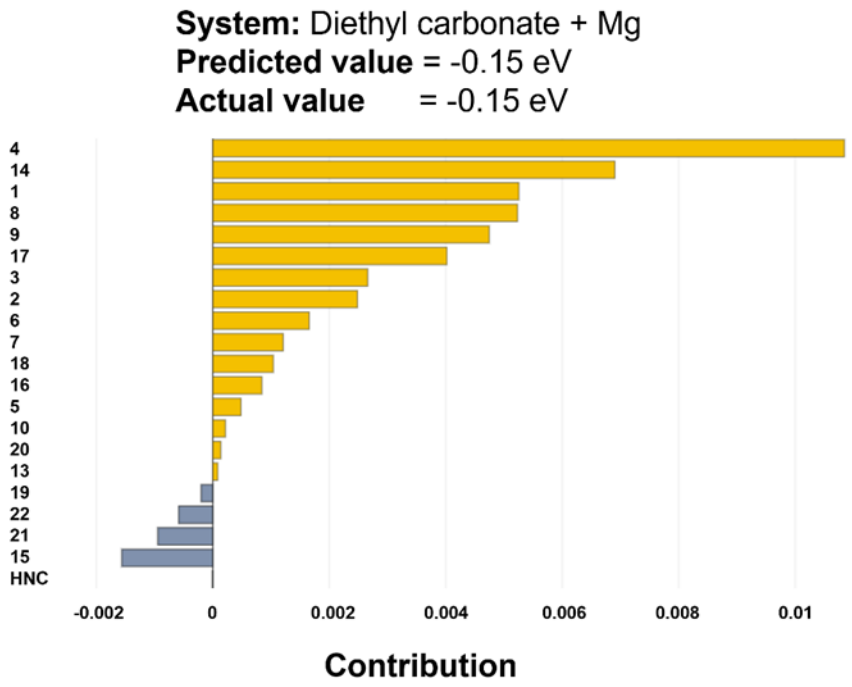


Figure 4.11: Feature importance of the least deviated system (Diethyl carbonate + Mg). Here, the HNC is the hidden negative contribution.

Similarly, we have plotted the feature importance of most positively deviated system and most negatively deviated system (**4.12a, 4.12b**). From **Figure 4.12a**, it has been observed that feature 1 (ionic radius) and 17 (LUMO) are the most contributed features. **Figure 4.12b** shows feature 14 (formula weight) and 4 (number of solvent) are the most contributing features in the case of the negatively deviated system. However, the contribution in this case is in the negative direction leading to the most negatively deviated system. Though in global feature analysis, the feature importance trend is identical with the least deviated system, the same is not true for deviated systems. Therefore, we believe that our overall ML predicted interaction energies are more accurate, indicating the stability and validity of considered ML model.

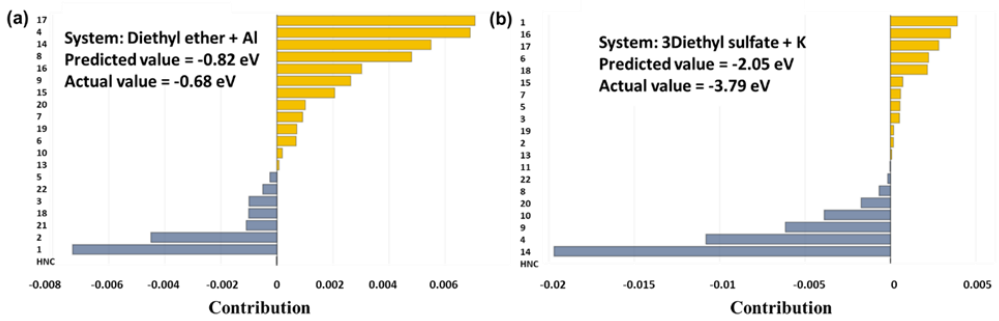


Figure 4.12: Feature importance of (a) most positively deviated system, and (b) most negatively deviated system.

Thus, analysis through shapash technique has given us an idea about which features dominate in case of an accurate or inaccurate prediction by the considered model. Hence, for a system whose actual value is not known, by the local feature importance plot we can understand whether the predicted value is least deviated or highly deviated by comparing the dominating features in that case. Thus, a system with feature 4 (number of solvent) as the most contributing feature importance is expected to have a more accurate predicted value by our ML model. Similarly, for unknown systems which have feature importance like **Figures 4.12a and 4.12b** are expected to deviate significantly from actual values. Hence, shapash technique can be used to validate ML predicted result without doing any quantum mechanical DFT calculation.

4.4. Conclusion

In this work, we have applied ML techniques for the screening of solvents based on metal-solvent interaction. The ML techniques speed up the process of examining the metal-solvent interaction property for all the metal-solvent systems. From the ML predicted result, GBR model is found to be the best suited algorithm for the prediction of interaction energy having RMSE of 0.489 eV and MAE of 0.326 eV. The interaction energy results show that the considered metals (Li, Na, Mg, Al, K, and Ca) are more stabilized with the increase in number of solvent molecules around the metal ion. Further,

to find out the effect of interaction energy on voltage determination, we have considered the graphite anode half-cell to calculate the anodic voltage of all considered systems utilizing the GBR predicted interaction energy. The voltage and interaction energy trend are found to be inversely related. Weaker metal-solvent interaction is found to result in preferably low anodic half-cell voltage. In this context, five optimum solvents for each metal have been presented based on minimum anodic half-cell voltage of the MIB. We have also found our ML predicted result are in good agreement with the experimental value. Further, to interpret the feature contribution, global and local feature analysis has been implemented utilizing shapash library. Thus, features with high feature importance have been determined which can lead to accurately predicted values for unknown metal-solvent systems. The shapash model can be used to validate any ML predicted result for any unknown systems by comparing the global feature contribution with the feature contribution of the least deviated system thereby avoiding DFT calculations. We believe, the ML approach has the potential to accelerate the process of voltage calculation based on ML predicted interaction energy for a large number of systems. The results also ascertain the effect of metal-solvent interaction energy in determining the voltage which can guide the experimental researchers to choose suitable solvents for metal ion battery designing.

4.5. References

(1) Nitta N., Wu F., Lee J. T., Yushin G. (2015), Li-ion battery materials: present and future, Mater. Today, 18, 252–264 (DOI: 10.1016/j.mattod.2014.10.040)

(2) Winter M., Barnett B., Xu K. (2018), Before Li ion batteries, Chem. Rev., 118, 11433–11456 (DOI: 10.1021/acs.chemrev.8b00422)

(3) Lombardo T., Duquesnoy M., El-Bouysidy H., Årén F., Gallo-Bueno A., Jørgensen P. B., Bhowmik A., Demortière A., Ayerbe E., Alcaide F., Reynaud M., Carrasco J., Grimaud A., Zhang C., Vegge T., Johansson P.,

Franco A. A. (2022), Artificial intelligence applied to battery research: hype or reality?, Chem. Rev., 122, 10899–10969 (DOI: 10.1021/acs.chemrev.1c00990)

(4) Louis S. Y., Siriwardane E. M. D., Joshi R. P., Omee S. S., Kumar N., Hu J. (2022), Accurate prediction of voltage of battery electrode materials using attention-based graph neural networks, ACS Appl. Mater. Interfaces, 14, 26587–26594 (DOI: 10.1021/acsami.2c06388)

(5) Joshi R. P., Ozdemir B., Barone V., Peralta J. E. (2015), Hexagonal BC_3: a robust electrode material for Li, Na, and K ion batteries, J. Phys. Chem. Lett., 6, 2728–2732 (DOI: 10.1021/acs.jpclett.5b01073)

(6) Bhauriyal P., Mahata A., Pathak B. (2017), Hexagonal BC_3 electrode for a high-voltage Al-ion battery, J. Phys. Chem. C, 121, 9748–9756 (DOI: 10.1021/acs.jpcc.7b02329)

(7) Posada J. O. G., Rennie A. J. R., Villar S. P., Martins V. L., Marinaccio J., Barnes A., Glover C. F., Worsley D. A., Hall P. J. (2017), Aqueous batteries as grid scale energy storage solutions, Renew. Sustain. Energy Rev., 68, 1174–1182 (DOI: 10.1016/j.rser.2016.02.009)

(8) Dunn B., Kamath H., Tarascon J. M. (2011), Electrical energy storage for the grid: a battery of choices, Science, 334, 928–935 (DOI: 10.1126/science.1212741)

(9) Eames C., Islam M. S. (2014), Ion intercalation into two-dimensional transition-metal carbides: global screening for new high-capacity battery materials, J. Am. Chem. Soc., 136, 16270–16276 (DOI: 10.1021/ja507382e)

(10) Thackeray M. M., Wolverton C., Isaacs E. D. (2012), Electrical energy storage for transportation—approaching the limits of, and going beyond, lithium-ion batteries, Energy Environ. Sci., 5, 7854–7863 (DOI: 10.1039/c1ee02874j)

(11) Kulish V. V., Koch D., Manzhos S. (2017), Ab initio study of Li, Mg and Al insertion into rutile VO_2: fast diffusion and enhanced voltages for multivalent batteries, Phys. Chem. Chem. Phys., 19, 22538–22545 (DOI: 10.1039/c7cp03980b)

(12) Guduru R. K., Icaza J. C. (2016), A brief review on multivalent intercalation batteries with aqueous electrolytes, Nanomaterials, 6, 41 (DOI: 10.3390/nano6030041)

(13) Butler K. T., Davies D. W., Cartwright H., Isayev O., Walsh A. (2018), Machine learning for molecular and materials science, Nature, 559, 547–555 (DOI: 10.1038/s41586-018-0337-2)

(14) Schmidt J., Marques M. R. G., Botti S., Marques M. A. L. (2019), Recent advances and applications of machine learning in solid-state materials science, npj Comput. Mater., 5, 83 (DOI: 10.1038/s41524-019-0221-0)

(15) Wang A. Y. T., Murdock R. J., Kauwe S. K., Oliynyk A. O., Gurlo A., Brgoch J., Persson K. A., Sparks T. D. (2020), Machine learning for materials scientists: an introductory guide toward best practices, Chem. Mater., 32, 4954–4965 (DOI: 10.1021/acs.chemmater.0c01907)

(16) Agrawal A., Choudhary A. (2016), Perspective: machine learning potentials for atomistic simulations, J. Chem. Phys., 145, 170901 (DOI: 10.1063/1.4966192)

(17) Isayev O., Oses C., Toher C., Gossett E., Curtarolo S., Tropsha A. (2017), Universal fragment descriptors for predicting properties of inorganic crystals, Nat. Commun., 8, 15679 (DOI: 10.1038/ncomms15679)

(18) Coley C. W., Jin W., Rogers L., Jamison T. F., Jaakkola T. S., Green W. H., Barzilay R., Jensen K. F. (2019), A graph-convolutional neural network model for the prediction of chemical reactivity, Chem. Sci., 10, 370–377 (DOI: 10.1039/c8sc04228d)

(19) Chen C., Ye W., Zuo Y., Zheng C., Ong S. P. (2019), Graph networks as a universal machine learning framework for molecules and crystals, Chem. Mater., 31, 3564–3572 (DOI: 10.1021/acs.chemmater.9b01294)

(20) Vandermause J., Torrisi S. B., Batzner S., Xie Y., Sun L., Kolpak A. M., Kozinsky B. (2020), On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events, npj Comput. Mater., 6, 20 (DOI: 10.1038/s41524-020-0303-4)

(21) Mailoa J. P., Kornbluth M., Batzner S., Samsonidze G., Lam S. T., Vandermause J., Ablitt C., Molinari N., Kozinsky B. (2019), A fast neural network approach for direct covariant forces prediction in complex multi-element extended systems, Nat. Mach. Intell., 1, 471–479 (DOI: 10.1038/s42256-019-0105-1)

(22) Granda J. M., Donina L., Dragone V., Long D. L., Cronin L. (2018), Controlling an organic synthesis robot with machine learning to search for new reactivity, Nature, 559, 377–381 (DOI: 10.1038/s41586-018-0307-8)

(23) Raccuglia P., Elbert K. C., Adler P. D. F., Falk C., Wenny M. B., Mollo A., Zeller M., Friedler S. A., Schrier J., Norquist A. J. (2016), Machine-learning-assisted materials discovery using failed experiments, Nature, 533, 73–76 (DOI: 10.1038/nature17439)

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

(25) Balachandran P. V., Kowalski B., Sehirlioglu A., Lookman T. (2018), Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning, Nat. Commun., 9, 1668 (DOI: 10.1038/s41467-018-04061-6)

(26) Carrete J., Li W., Mingo N., Wang S., Curtarolo S. (2014), Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors

via high-throughput materials modeling, Phys. Rev. X, 4, 011019 (DOI: 10.1103/physrevx.4.011019)

(27) Kim C., Pilania G., Ramprasad R. (2016), Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX_3 perovskites, J. Phys. Chem. C, 120, 14575–14580 (DOI: 10.1021/acs.jpcc.6b02964)

(28) Seko A., Hayashi H., Nakayama K., Takahashi A., Tanaka I. (2017), Representation of compounds for machine-learning prediction of physical properties, Phys. Rev. B, 95, 144110 (DOI: 10.1103/physrevb.95.144110)

(29) Chen C., Zuo Y., Ye W., Li X., Deng Z., Ong S. P. (2020), A critical review of machine learning of energy materials, Adv. Energy Mater., 10, 1903242 (DOI: 10.1002/aenm.201903242)

(30) Zhang Y., He X., Chen Z., Bai Q., Nolan A. M., Roberts C. A., Banerjee D., Matsunaga T., Mo Y., Ling C. (2019), Unsupervised discovery of solid-state lithium ion conductors, Nat. Commun., 10, 1–7 (DOI: 10.1038/s41467-019-09777-8)

(31) Tshitoyan V., Dagdelen J., Weston L., Dunn A., Rong Z., Kononova O., Persson K. A., Ceder G., Jain A. (2019), Unsupervised word embeddings capture latent knowledge from materials science literature, Nature, 571, 95–98 (DOI: 10.1038/s41586-019-1335-8)

(32) Umer M., Umer S., Zafari M., Ha M., Anand R., Hajibabaei A., Abbas A., Lee G., Kim K. S. (2022), Machine learning assisted high-throughput screening of transition metal single atom based superb hydrogen evolution electrocatalysts, J. Mater. Chem. A, 10, 6679–6689 (DOI: 10.1039/d1ta10917k)

(33) Zafari M., Nissimagoudar A. S., Umer M., Lee G., Kim K. S. (2021), First principles and machine learning based superior catalytic activities and selectivities for N_2 reduction in MBenes, defective 2D materials and 2D π-

conjugated polymer-supported single atom catalysts, J. Mater. Chem. A, 9, 9203–9213 (DOI: 10.1039/d0ta11308b)

(34) Zafari M., Kumar D., Umer M., Kim K. S. (2020), Machine learning-based high throughput screening for nitrogen fixation on boron-doped single atom catalysts, J. Mater. Chem. A, 8, 5209–5216 (DOI: 10.1039/c9ta12431j)

(35) Ha M., Kim D. Y., Umer M., Gladkikh V., Myung C. W., Kim K. S. (2021), Tuning metal single atoms embedded in N_xC_γ moieties toward high-performance electrocatalysis, Energy Environ. Sci., 14, 3455–3468 (DOI: 10.1039/d0ee03365b)

(36) Joshi R. P., Eickholt J., Li L., Fornari M., Barone V., Peralta J. E. (2019), Machine learning the voltage of electrode materials in metal-ion batteries, ACS Appl. Mater. Interfaces, 11, 18494–18503 (DOI: 10.1021/acsami.9b02051)

(37) Moses I. A., Joshi R. P., Ozdemir B., Kumar N., Eickholt J., Barone V. (2021), Machine learning screening of metal-ion battery electrode materials, ACS Appl. Mater. Interfaces, 13, 53355–53362 (DOI: 10.1021/acsami.1c15053)

(38) Louis S. Y., Siriwardane E. M. D., Joshi R. P., Omee S. S., Kumar N., Hu J. (2022), Accurate prediction of voltage of battery electrode materials using attention-based graph neural networks, ACS Appl. Mater. Interfaces, 14, 26587–26594 (DOI: 10.1021/acsami.2c06388)

(39) Zhang X., Zhou J., Lu J., Shen L. (2022), Interpretable learning of voltage for electrode design of multivalent metal-ion batteries, npj Comput. Mater., 8, 1–8 (DOI: 10.1038/s41524-022-00727-6)

(40) Sun Y., Ayalasomayajula S. M., Deva A., Lin G., García R. E. (2022), Artificial intelligence inferred microstructural properties from voltage–capacity curves, Sci. Rep., 12, 1–11 (DOI: 10.1038/s41598-022-10831-5)

(41) Ling C. (2022), A review of the recent progress in battery informatics, npj Comput. Mater., 8, 1–22 (DOI: 10.1038/s41524-022-00730-x)

(42) Ha M., Hajibabaei A., Kim D. Y., Singh A. N., Yun J., Myung C. W., Kim K. S. (2022), Al-doping driven suppression of capacity and voltage fadings in 4d-element containing Li-ion-battery cathode materials: machine learning and density functional theory, Adv. Energy Mater., 12, 2201018 (DOI: 10.1002/aenm.202201018)

(43) Manthiram A. (2017), An outlook on lithium ion battery technology, ACS Cent. Sci., 3, 1063–1069 (DOI: 10.1021/acscentsci.7b00288)

(44) Ishikawa A., Sodeyama K., Igarashi Y., Nakayama T., Tateyama Y., Okada M. (2019), Machine learning prediction of coordination energies for alkali group elements in battery electrolyte solvents, Phys. Chem. Chem. Phys., 21, 26399–26405 (DOI: 10.1039/c9cp04689k)

(45) Carpenter J. E., Weinhold F. (1988), Analysis of the geometry of the hydroxymethyl radical by the “different hybrids for different spins” natural bond orbital procedure, J. Mol. Struct. THEOCHEM, 169, 41–62 (DOI: 10.1016/0166-1280(88)85007-5)

(46) Hehre W. J., Ditchfield K., Pople J. A. (1972), Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules, J. Chem. Phys., 56, 2257 (DOI: 10.1063/1.1677527)

(47) Hariharan P. C., Pople J. A. (1973), The influence of polarization functions on molecular orbital hydrogenation energies, Theor. Chim. Acta, 28, 213–222 (DOI: 10.1007/bf00533485)

(48) Krishnan R., Binkley J. S., Seeger R., Pople J. A. (1980), Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions, J. Chem. Phys., 72, 650 (DOI: 10.1063/1.438955)

(49) Grimme S., Antony J., Ehrlich S., Krieg H. (2010), A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu, J. Chem. Phys., 132 (15), 154104 (DOI: 10.1063/1.3382344)

(50) Roy D., Mandal S. C., Pathak B. (2021), Machine learning-driven high-throughput screening of alloy-based catalysts for selective CO_2 hydrogenation to methanol, ACS Appl. Mater. Interfaces, 13, 56151–56163 (DOI: 10.1021/acsami.1c13999)

(51) Boev A. O., Fedotov S. S., Stevenson K. J., Aksyonov D. A. (2021), High-throughput computational screening of cathode materials for Li–O_2 battery, Comput. Mater. Sci., 197, 110614 (DOI: 10.1016/j.commatsci.2021.110614)

(52) Zhu X., Yan J., Gu M., Liu T., Dai Y., Gu Y., Li Y. (2019), Activity origin and design principles for oxygen reduction on dual-metal-site catalysts: a combined density functional theory and machine learning study, J. Phys. Chem. Lett., 10, 7760–7766 (DOI: 10.1021/acs.jpclett.9b02859)

(53) Ren C., Lu S., Wu Y., Ouyang Y., Zhang Y., Li Q., Ling C., Wang J. (2022), A universal descriptor for complicated interfacial effects on electrochemical reduction reactions, J. Am. Chem. Soc., 144, 48 (DOI: 10.1021/jacs.1c09962)

(54) Raghunathan S., Priyakumar U. D. (2022), Molecular representations for machine learning applications in chemistry, Int. J. Quantum Chem., 122, e26870 (DOI: 10.1002/qua.26870)

(55) Pandit N. K., Roy D., Mandal S. C., Pathak B. (2022), Rational designing of bimetallic/trimetallic hydrogen evolution reaction catalysts using supervised machine learning, J. Phys. Chem. Lett., 13, 12 (DOI: 10.1021/acs.jpclett.1c03746)

(56) Duan C., Nandy A., Adamji H., Roman-Leshkov Y., Kulik H. J. (2022), Machine learning models predict calculation outcomes with the transferability necessary for computational catalysis, J. Chem. Theory Comput., 18, 4282–4292 (DOI: 10.1021/acs.jctc.2c00245)

(57) Nandy A., Duan C., Kulik H. J. (2021), Using machine learning and data mining to leverage community knowledge for the engineering of stable metal-organic frameworks, J. Am. Chem. Soc., 143, 17535–17547 (DOI: 10.1021/jacs.1c06600)

(58) Fischer J. M., Hunter M., Hankel M., Searles D. J., Parker A. J., Barnard A. S. (2020), Accurate prediction of binding energies for two-dimensional catalytic materials using machine learning, ChemCatChem, 12, 5109–5120 (DOI: 10.1002/cctc.202000385)

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

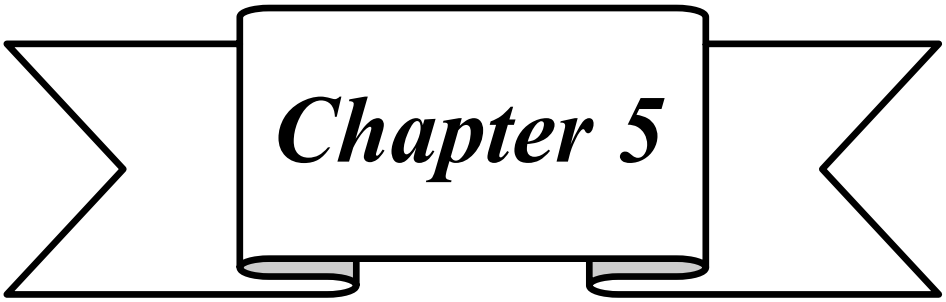
(60) Lever J., Krzywinski M., Altman N. (2016), Points of significance: model selection and overfitting, Nat. Methods, 13, 703–704 (DOI: 10.1038/nmeth.3945)

(61) Asenbauer J., Eisenmann T., Kuenzel M., Kazzazi A., Chen Z., Bresser D. (2020), The success story of graphite as a lithium-ion anode material – fundamentals, remaining challenges, and recent developments including silicon (oxide) composites, Sustain. Energy Fuels, 4, 5387–5416 (DOI: 10.1039/d0se00175a)

(62) Akkisetty B., Dimogiannis K., Searle J., Rogers D., Newton G. N., Johnson L. R. (2022), Enflurane additive for sodium negative electrodes, ACS Appl. Mater. Interfaces, 14, 36551–36556 (DOI: 10.1021/acsami.2c06898)

(63) Lundberg S. M., Allen P. G., Lee S. I. (2017), A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst., 30

(64) Ribeiro M. T., Singh S., Guestrin C. (2016), "Why should I trust you?" Explaining the predictions of any classifier, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 1135–1144 (DOI: 10.1145/2939672.2939778)



Chapter 5

Screening and Clustering of High ECW Solvent Electrolytes for Battery Applications

5.1. Introduction

The development of high energy density and high voltage batteries are crucial to sustain the high-tech growth of the current scenario. The renewable energy sources such as solar, wind, and hydro power, are inherently variable, and energy storage systems are needed to ensure uninterrupted electricity.*[1–4]* Recent advances in theoretical algorithms, modeling, simulations, and computer technologies are driving the rational design of lithium-ion batteries. These advances enable the integration of computational calculations with experimental data in shared databases, accelerating development across the industrial chain.*[5]* Multi-scale modelling and simulations complement experimental efforts by predicting path-independent properties and enhancing our understanding of battery performance across various scales. This integrative approach is essential for optimizing rechargeable battery technology. In the context of practical rechargeable batteries along with cathode and anode, electrolytes also a crucial part which serve as a medium for the conveyance of ions between two electrodes. To understand the battery stability a crucial condition is required where, the Fermi energy of the anode must be lower than the energy level associated with the lowest unoccupied molecular orbital (LUMO) of the electrolyte, and the Fermi of the cathode must exceed the energy level corresponding to the highest occupied molecular orbital (HOMO) of the electrolyte.*[6–8]* This condition ensures the prevention of undesirable reactions, unless a protective and stable solid layer known as the solid electrolyte interphase (SEI) forms on the electrode surface.*[9]* Electrochemical windows (ECWs) are typically measured through linear voltammetry, wherein the individual anodic and cathodic potentials govern the oxidation and reduction processes of electrolytes. However, the determination of limiting potential is influenced by several factors, including electrode characteristics and the arbitrary selection of current cutoffs to identify onset redox potentials. Consequently, the reported ECWs in literature exhibit considerable variability. To address these complexities,

various computational techniques have been developed to ascertain the oxidation and reduction potentials of solvent electrolytes.

Recently, Shi and co-workers have described different DFT methods for the calculation of ECW and proposed a novel thermodynamics cycle-based approach to determine the ECW accurately.*[10]* They have also described the limitations in the traditional approach of ECW determination using the difference between the HOMO and LUMO energies. Because of the consideration of HOMO/LUMO derived from electronic structure theory, it only represents the electronic properties of isolated neutral molecules and does not reflect the real ECWs of species participating in redox reactions.*[8,11–16]* Shi and co-workers have reported that the advantages of the thermodynamic cycle-based method over HOMO/LUMO method is its consideration of reorganization energy and solvation energy correction.*[10]* Solvation energy refers to the energy required to transfer molecules or ions from a vacuum into a solution phase. The inclusion of this solvation energy correction, where the correction term adjusts the substance from standard conditions (1 atm) to 1 mol L^{-1}, provides a more practical and accurate representation compared to the HOMO/LUMO approach. Reorganization energy, which reflects the energy changes due to geometric property alterations (such as bond lengths, bond angles, and dihedral angles) between the neutral and cationic states, varies significantly regardless of the solvent category.*[17]* They have shown that the reorganization energy for a range of 68 experimentally reported solvents varies from 0 to 0.58 eV.*[10]* The primary factor influencing the magnitude of reorganization energy is not the solvent category but the geometric changes during structural relaxation. Furthermore, Hutchison et al. have proposed that reorganization energy is a linear combination of changes in bond length and dihedral angle during oxidation from the neutral state to the corresponding cation.*[18]* This finding underscores the importance of geometric changes over solvent type in determining reorganization energy. Shi and colleagues have also highlighted the significance of incorporating

an implicit solvation model for chemical kinetics studies, such as those between I_2 and LiOH in Li-air batteries.*[19]* This integrated ab initio and ML investigation emphasizes the critical role of solvation effects in achieving more reliable and practical results. Thus, the thermodynamic cycle-based approach offers substantial advantages over the HOMO/LUMO method by accounting for reorganization energy and solvation effects, leading to more accurate and practical predictions of redox potentials. Recently Wang and his co-worker has also demonstrated the highest accuracy in determining the ECW can be achieved by redox reaction process (thermodynamic method) compared to other methods.*[20]* The traditional non-aqueous electrolytes primarily rely on organic carbonates, which exhibit ECWs of less than 5.0 V. These carbonates include both linear and cyclic solvent electrolytes such as dimethyl carbonate, diethyl carbonate, propylene carbonate, and ethylene carbonate. Considering recent advancements, the development of high-voltage batteries (>5 V) necessitates the engineering of high ECW based electrolytes.*[21–24]*

However, the number of solvents that are yet to be tested as electrolytes for rechargeable batteries based on their ECW values are very high. Determining the ECWs for all unknown solvents poses a considerable challenge, both experimentally and computationally. This difficulty arises from the diverse measuring conditions required for each solvent in experimental settings and the expensive as well as time-consuming nature of computational calculations. Recently, machine learning (ML) has emerged as a transformative force in energy research, showcasing rapid growth with the potential to revolutionize the design of batteries through its swift progress and high-level accuracy. ML algorithms can be used in the field of both catalysis as well as energy storage and conversion to analyse large datasets of physicochemical properties and performance data, enabling researchers to identify correlations and patterns that may not be apparent through traditional methods.*[25–28]* Previously we have

successfully reported the application of ML for the prediction of capacity and voltage for metal ion battery.*[29,30]* Currently ML has emerged as an essential tool for determining electrochemical performance of different kind of batteries also by solving complex structure-property relationship.*[29–34]* Application of ML is not only limited for material search but also applicable to the domain of solvent electrolyte screening. Pathak and co-workers have previously reported discovery of ionic liquid-based electrolytes through thermodynamic cycle method as well as ML driven approach for dual-ion battery.*[35,36]* Thus, ML can be particularly useful for predicting the ECW of electrolytes, which is a critical parameter in battery design.

In this study, we delved into the realm of ML, employing algorithms to predict oxidation and reduction potentials across a vast, unfamiliar dataset using molecular descriptors. Leveraging an optimized ML model, we have predicted the oxidation and reduction potential for all unknown solvent electrolytes, proposing optimal solvents poised for exploration in higher voltage electrode materials-an avenue yet untouched for unknown solvent space by experimental and DFT approaches in battery applications. Furthermore, our scientific exploration extended to an unsupervised clustering method applied to the entire solvent database. Molecules were intricately categorized based on their molecular descriptors, with a keen emphasis on ECW as a pivotal feature. By applying clustering method, we have been able to sub-group a vast solvent database consisting of similar kind of solvent electrolytes along with their ECW. Unravelling patterns within this molecular landscape, our approach not only offers insights into untapped potential but also provides an aesthetically compelling road map for future exploration. This comprehensive research underscores the vast potential of ML in predicting ECW values, thus facilitating the design and optimization of high-voltage batteries.

5.2. Methods and Materials

In our quest for better rechargeable batteries, we followed a simple yet powerful roadmap **(Figure 5.1)**. First, we have collected the reported reduction and oxidation potentials data calculated using the thermodynamic cycle method.*[10]* Utilizing RDKit, we converted complex solvent electrolyte structures into manageable and interpretable features. These features were subsequently employed to establish a mapping between the input variables and the target red-ox potentials. Given the high dimensionality of the dataset, which can pose challenges for small datasets, we adopted the Select-K-Best statistical feature selection method to identify the most relevant features. Initially various ML models have been considered through two different robust cross-validation methods such as repeated-K-fold cross-validation (RKFCV) and leave-one-out cross-validation (LOOCV) for the identification of suitable ML models to map the input variables with the red-ox potentials. From the cross-validation results, three models, XGBoost Regressor (XGBR), Random Forest Regressor (RFR), and Gradient Boosting Regressor (GBR) demonstrated as strong performance ML models for the prediction of red-ox potentials. We further refined these models through hyperparameter tuning, focusing on the features with the greatest influence on the target variable. Following optimization, we assessed the train-test error for each model to evaluate their performance in red-ox potential prediction. This workflow process is summarized as model evaluation in **Figure 5.1**. Mean Absolute Error (MAE) has been used as the performance metric for evaluating the ML models. Upon obtaining the electrochemical window (ECW) of all solvent electrolytes, we screened for novel solvent electrolytes suitable for battery applications based on predefined criteria. Additionally, we employed clustering techniques to categorize the large number of solvent electrolytes into distinct subclasses. Utilizing ML as predictive tool, we swiftly explored a vast database of unknown solvents followed by utilization of clustering

method to group similar kinds of solvents, revealing new possibilities for high voltage electrode materials.

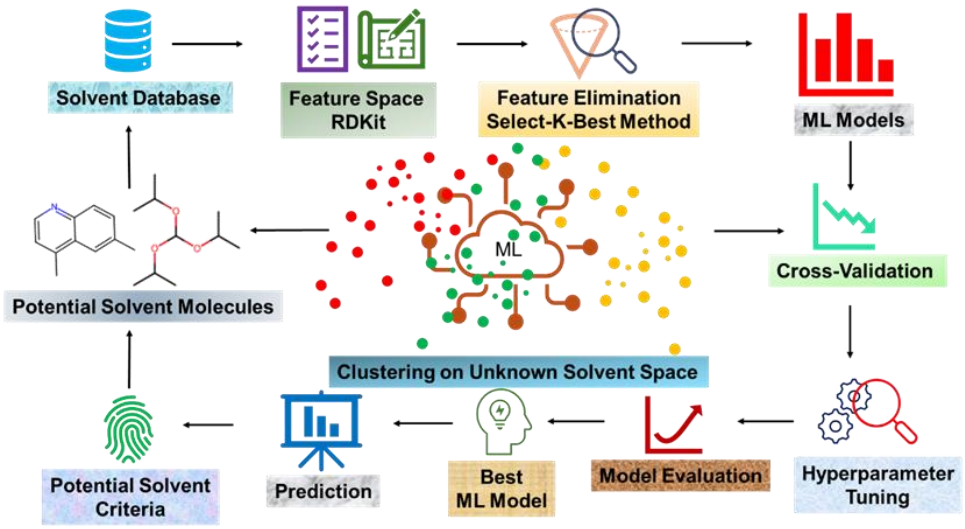


Figure 5.1: A visual representation of the multi-step ML workflow for novel solvent electrolyte design for rechargeable batteries.

The reduction and oxidation potentials of various solvent electrolytes were chosen as target variables for this study, extracted from previous reports utilizing the thermodynamic cycle method for their calculation.*[10]* The variation of oxidation and reduction potentials and the ECW values with respect to various functional groups present in different solvent electrolytes represented through a violin plot in **Figure 5.2**.

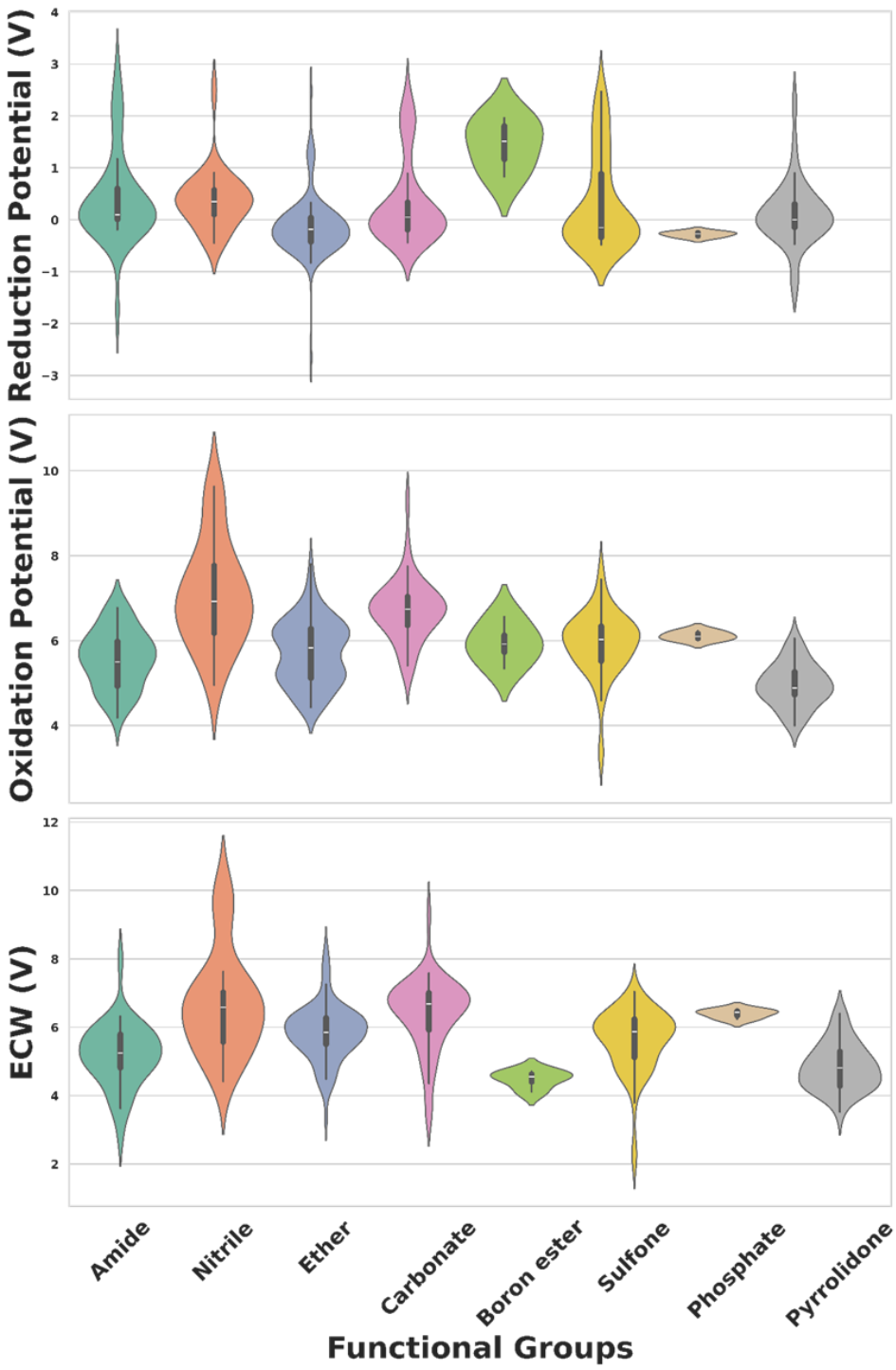


Figure 5.2: Violin plot represents the variation of reduction potential, oxidation potential and ECW with respect to various functional group present in the solvent electrolytes.

The plot reveals that solvent electrolytes containing amide and ether groups exhibit a wide range of reduction potentials, spanning approximately -2.8 V to 3.8 V and -3.2 V to 3 V, respectively. Notably, solvent electrolytes with nitrile groups demonstrate the highest ECW values, attributed to their significantly high oxidation potentials. In violin plots, the middle section indicates a higher density of data points, while the narrow terminal sections indicate fewer data points, which can act as potential outliers. **The Figure 5.2.** illustrates the overall pictorial diagram of the whole dataset where few data points with very high or low potential values, highlighting the presence of outliers in dataset. Our final objective is to leverage trained models to predict the ECW for approximately 4500 unexplored solvents. This unexplored solvent space includes a wide variety of solvent electrolytes with potential ECWs ranging from low to very high values. If we discard anomalous data points determined through DFT calculations, we will risk limiting the model's ability to predict red-ox potentials across the full spectrum of solvent electrolytes. This is because excluding these data points would narrow the range of training data, potentially leading to biased models that might not generalize well to the diverse and extensive range of unexplored solvents. By retaining these anomalous data points, we aim to build more robust ML models capable of handling and predicting such variability. Reducing the training dataset by discarding these data points could also hinder the model's robustness and accuracy. A comprehensive training dataset that encompasses the full range of potential redox behaviours is crucial for developing models that not only perform well on the training and testing sets but also generalize effectively to the large number of unexplored solvent electrolytes.

5.3. Feature Space

It is imperative to represent the solvent electrolytes in numerical and readable format for the development of efficient ML models to find out quantitative or qualitative structure property relationship. To get accurately trained ML model, consideration of suitable descriptors, which can

represent and distinguish each solvent electrolyte with remarkable precision is highly important. Here, we have utilized the RDKit, an open-source cheminformatics library to generate 210 molecular descriptors where the SMILES (simplified molecular-input line-entry system) string of each solvent has been provided to represent the solvent electrolytes.*[37–41]* RDKit offers a comprehensive suite of tools for generating a wide array of molecular descriptors, which are essential for numerically representing the structural and chemical properties of solvent electrolytes. In our study, we used the SMILES string of each solvent as input for RDKit to generate these descriptors. Specifically, we utilized the MoleculeDescriptors.MolecularDescriptorCalculator() function to compute a variety of descriptors for each solvent molecule.*[42]* These descriptors encompass various physicochemical properties, topological indices, and electronic features, providing a detailed numerical representation of the solvent electrolytes. The names of these descriptors can be accessed using the GetDescriptorNames() function, with a comprehensive list available in **Table 5.1**.

Table 5.1: List of extracted descriptors considered for the learning of reduction and oxidation potential of the solvent electrolytes. All the features have been extracted using RDKit library.

Symbols	Descriptors
MaxAbsEStateIndex	Maximum absolute value of E-State index, which characterizes the electron distribution in a molecule.
MaxEStateIndex	Maximum E-State index, indicating the electrophilicity or electron-attracting ability of a molecule.
MinAbsEStateIndex	Minimum absolute value of E-State index.
MinEStateIndex	Minimum E-State index.

qed	Quantitative Estimate of solvent-likeness, a measure of how "solvent-like" a molecule is based on certain criteria.
MolWt	Molecular weight of a molecule, the sum of the atomic weights of all atoms in the molecule.
HeavyAtomMolWt	Molecular weight considering only the heavy (non-hydrogen) atoms.
ExactMolWt	Exact molecular weight, accounting for isotopic masses of atoms.
NumValenceElectrons	Total number of valence electrons in the molecule.
NumRadicalElectrons	Number of unpaired electrons or radicals in the molecule.
MaxPartialCharge	Maximum partial atomic charge within the molecule.
MinPartialCharge	Minimum partial atomic charge within the molecule.
MaxAbsPartialCharge	Maximum absolute value of partial atomic charge.
MinAbsPartialCharge	Minimum absolute value of partial atomic charge.
FpDensityMorgan1	Fingerprint Density for Morgan circular fingerprints with radius 1. It measures the density of substructures in the molecule.
FpDensityMorgan2	Fingerprint Density for Morgan circular fingerprints with radius 2. It measures the density of larger substructures than FpDensityMorgan1.
FpDensityMorgan3	Fingerprint Density for Morgan circular fingerprints with radius 3. It measures the density of even larger substructures than FpDensityMorgan2.
BCUT2D_MWHI	Burden modified molecular weight for atoms with high atomic numbers.

BCUT2D_MWLOW	Burden modified molecular weight for atoms with low atomic numbers.
BCUT2D_CHGHI	Burden modified charge for atoms with high atomic numbers.
BCUT2D_CHGLO	Burden modified charge for atoms with low atomic numbers.
BCUT2D_LOGPHI	Burden modified logarithm of the partition coefficient for polarizability.
BCUT2D_LOGPLOW	Burden modified logarithm of the partition coefficient for lipophilicity.
BCUT2D_MRHI	Burden modified molar refractivity for atoms with high atomic numbers.
BCUT2D_MRLOW	Burden modified molar refractivity for atoms with low atomic numbers.
AvgIpc	Average Information Content of the Physicochemical Properties. It quantifies the diversity of physicochemical properties in a molecule.
BalabanJ	Balaban J Index, a topological index used to describe the molecular structure and branching.
BertzCT	Bertz Chemical Topological Index, a topological index used to characterize the complexity of the molecular structure.
Chi0	Zeroth-order molecular connectivity index, which quantifies molecular branching.
Chi0n	Chi0 without hydrogen contributions.
Chi0v	Chi0 including hydrogen contributions.
Chi1	First-order molecular connectivity index, related to molecular shape.
Chi1n	Chi1 without hydrogen contributions.
Chi1v	Chi1 including hydrogen contributions.

Chi2n	Chi2 without hydrogen contributions.
Chi2v	Chi2 includes hydrogen contributions.
Chi3n	Chi3 without hydrogen contributions.
Chi3v	Chi3 including hydrogen contributions.
Chi4n	Chi4 without hydrogen contributions.
Chi4v	Chi4 including hydrogen contributions.
HallKierAlpha	Hall-Kier Alpha shape index, used to describe molecular shape.
Ipc	Information content of topological indices, a measure of molecular complexity.
Kappa1	First-order kappa shape index, characterizing molecular shape.
Kappa2	Second-order kappa shape index, characterizing molecular shape.
Kappa3	Third-order kappa shape index, characterizing molecular shape.
LabuteASA	Labute's Approximated Surface Area, an estimate of molecular surface area.
PEOE_VSA1 to PEOE_VSA14	Partial Equalization of Orbital Electronegativities - Van der Waals Surface Area descriptors for different substructures.
SMR_VSA1 to SMR_VSA10	Solvent Accessible Surface Area (SASA) Molecular Surface Representations (SMR) for different substructures.
SlogP_VSA1 to SlogP_VSA12	SlogP (logarithm of the partition coefficient) contributions for different substructures.
TPSA	Topological Polar Surface Area, a measure of the total polar surface area in a molecule.
EState_VSA1	Estate Van der Waals Surface Area descriptor for a specific substructure

EState_VSA2 to EState_VSA11	E-State Van der Waals Surface Area descriptors for specific substructures.
VSA_EState1 to VSA_EState10	Van der Waals Surface Area descriptors based on the E-State indices for specific substructures.
FractionCSP3	Fraction of carbons with sp3 hybridization in the molecule.
HeavyAtomCount	Count of heavy atoms (non-hydrogen atoms) in the molecule.
NHOHCount	Count of N-H and O-H groups in the molecule.
NOCount	Count of nitro groups (N=O) in the molecule.
NumAliphaticCarbocycles	Number of aliphatic (non-aromatic) carbocycles.
NumAliphaticHeterocycles	Number of aliphatic (non-aromatic) heterocycles.
NumAliphaticRings	Total number of aliphatic (non-aromatic) rings.
NumAromaticCarbocycles	Number of aromatic carbocycles.
NumAromaticHeterocycles	Number of aromatic heterocycles.
NumAromaticRings	Total number of aromatic rings.
NumHAcceptors	Number of hydrogen bond acceptor sites in the molecule.
NumHDonors	Number of hydrogen bond donor sites in the molecule.
NumHeteroatoms	Number of heteroatoms (non-carbon and non-hydrogen atoms) in the molecule.
NumRotatableBonds	Number of rotatable bonds in the molecule.

NumSaturatedCarbocycles	Number of saturated (non-aromatic) carbocycles.
NumSaturatedHeterocycles	Number of saturated (non-aromatic) heterocycles.
NumSaturatedRings	Total number of saturated (non-aromatic) rings.
RingCount	Total number of rings in the molecule.
MolLogP	Logarithm of the partition coefficient (logP), a measure of a molecule's lipophilicity.
MolMR	Molecular refractivity, a measure of the total polarizability of a molecule.
fr_Al_COO	Presence of an aliphatic carboxylic acid group.
fr_Al_OH	Presence of an aliphatic alcohol group.
fr_Al_OH_noTert	Presence of an aliphatic alcohol group without tertiary carbon adjacent to it.
fr_ArN	Presence of an aromatic amine group.
fr_Ar_COO	Presence of an aromatic carboxylic acid group.
fr_Ar_N	Presence of an aromatic nitrogen (not in an amine group).
fr_Ar_NH	Presence of an aromatic amine group.
fr_Ar_OH	Presence of an aromatic alcohol group.
fr_COO	Presence of a carboxylic acid group.
fr_COO2	Presence of a ketene group (carbonyl group attached to a carbonyl group).
fr_C_O	Presence of a carbonyl group.
fr_C_O_noCOO	Presence of a carbonyl group without adjacent carboxylic acid groups.
fr_C_S	Presence of a carbon-sulfur bond.
fr_HOCCN	Presence of an N-substituted hydroxylamine.

fr_Imine	Presence of an imine group.
fr_NH0, fr_NH1, fr_NH2	Presence of zero, one, or two nitrogen-hydrogen (amine) groups.
fr_N_O	Presence of a nitrogen-oxygen bond.
fr_Ndealkylation1, fr_Ndealkylation2	Presence of potential sites for dealkylation reactions.
fr_Nhpyrrole	Presence of a nitrogen atom in a pyrrole ring.
fr_SH	Presence of a thiol (sulfhydryl) group.
fr_aldehyde	Presence of an aldehyde group.
fr_alkyl_carbamate	Presence of an alkyl carbamate group.
fr_alkyl_halide	Presence of an alkyl halide group.
fr_allylic_oxid	Presence of an allylic oxidation site.
fr_amide	Presence of an amide group.
fr_amidine	Presence of an amidine group.
fr_aniline	Presence of an aniline group.
fr_aryl_methyl	Presence of an aryl methyl group.
fr_azide	Presence of an azide group.
fr_azo	Presence of an azo group.
fr_barbitur	Presence of a barbiturate group.
fr_benzene	Presence of a benzene ring.
fr_benzodiazepine	Presence of a benzodiazepine group.
fr_bicyclic	Presence of a bicyclic ring system.
fr_diazo	Presence of a diazo group.
fr_dihydropyridine	Presence of a dihydropyridine group.
fr_epoxide	Presence of an epoxide group.
fr_ester:	Presence of an ester group.

fr_ether	Presence of an ether group.
fr_furan	Presence of a furan ring.
fr_guanido	Presence of a guanido group.
fr_halogen	Presence of a halogen atom.
fr_hdrzine	Presence of a hydrazine group.
fr_hdrzone:	Presence of a hydrazone group.
fr_imidazole:	Presence of an imidazole group.
fr_imide	Presence of an imide group.
fr_isocyan	Presence of an isocyanate group.
fr_isothiocyan	Presence of an isothiocyanate group.
fr_ketone	Presence of a ketone group.
fr_ketone_Topliss	Presence of a ketone group (Topliss approach).
fr_lactam	Presence of a lactam group.
fr_lactone	Presence of a lactone group.
fr_methoxy	Presence of a methoxy group.
fr_morpholine	Presence of a morpholine group.
fr_nitrile	Presence of a nitrile group.
fr_nitro	Presence of a nitro group.
fr_nitro_arom	Presence of a nitro group on an aromatic ring.
fr_nitro_arom_nonortho	Presence of a nitro group on an aromatic ring in a non-ortho position.
fr_nitroso:	Presence of a nitroso group.
fr_oxazole	Presence of an oxazole ring.
fr_oxime	Presence of an oxime group.
fr_para_hydroxylation	Presence of a para-hydroxylation site
fr_phenol	Presence of a phenol group.

fr_phenol_noOrtho Hbond	Presence of a phenol group without ortho-hydrogen bonding.
fr_phos_acid	Presence of a phosphonic acid group.
fr_phos_ester	Presence of a phosphoester group.
fr_piperdine:	Presence of a piperidine group.
fr_piperzine	Presence of a piperazine group.
fr_priamide	Presence of a primary amide group.
fr_prisulfonamd	Presence of a primary sulfonamide group.
fr_pyridine	Presence of a pyridine ring.
fr_quatN	Presence of a quaternary nitrogen atom.
fr_sulfide:	Presence of a sulfide (thioether) group.
fr_sulfonamd:	Presence of a sulfonamide group.
fr_sulfone	Presence of a sulfone group.
fr_term_acetylene	Presence of a terminal acetylene group.
fr_tetrazole	Presence of a tetrazole ring.
fr_thiazole	Presence of a thiazole ring.
fr_thiocyan	Presence of a thiocyanate group.
fr_urea	Presence of a urea group.
Sum_fp	Sum of binary bits of fingerprints
Average_fp	Average of binary bits of fingerprints
Deviation_fp	deviation of binary bits of fingerprints

To ensure that the features used in our models are both relevant and distinctive, we excluded elemental properties, as they do not differentiate between isomers. Accurate differentiation between such isomers is crucial for the effective training of ML models. The considered molecular descriptors represent the chemical and physical properties of the solvent

electrolytes. Along with the molecular descriptors, statistical sum, average and standard variation of molecular fingerprints have also been considered for the solvent electrolytes. The molecular descriptors involved both the structural features as well as electronic features. For example, NumRotatableBonds, MolWt, NumHDonors, NumHAcceptors etc. represent the structural features whereas NumValenceElectrons, MaxPartialCharge, MinPartialCharge etc., represent the electronic features. Moreover, the molecular descriptors also involved those features which represent whether a particular functional group is present or not. For example, fr_ester, and fr_ether represent the presence of ester and ether group in a solvent electrolyte.

5.4. Results and Discussion

5.4.1. ML Models

To map the molecular descriptors of the solvent electrolytes with their corresponding oxidation and reduction potential, we have started with various linear and non-linear ML models, based on the extracted oxidation and reduction potential data computed via thermodynamic cycle method.*[10]* We have employed twelve different ML algorithms, namely Lasso (LS), Partial Least Squares (PLS), Ridge Regression (RDG), Kernel Ridge Regression (KRR), Elastic Net Regression (ENR), K-Neighbors Regression (KNR), Support Vector Regression (SVR), AdaBoost Regression (ABR), Gradient Boosting Regression (GBR), Decision Tree Regression (DTR), Extreme Gradient Boosting Regression (XGBR), and Random Forest Regression (RFR). In the initial stage, we evaluated these nine ML algorithms for predicting red-ox potentials using cross-validation with default parameters to screen for the most promising model types (linear, non-linear, tree-based, bagging-boosting). This preliminary screening allowed us to avoid overfitting and provided a fair assessment of each model's baseline performance without extensive computational cost. After identifying the most suitable models, we performed detailed

hyperparameter tuning on these selected models. To ensure the reliability and generalization of the ML models, we started with examining the stability of the considered ML models through three different cross-validation (CV) methods, namely, K-fold CV, repeated K-fold CV (RKFCV), and leave-one-out-CV (LOOCV) methods. The bar plot of RKFCV and LOOCV for oxidation and reduction potential has been depicted in Figure 1 and the K-fold CV data has been tabulated in **Table 5.2**.

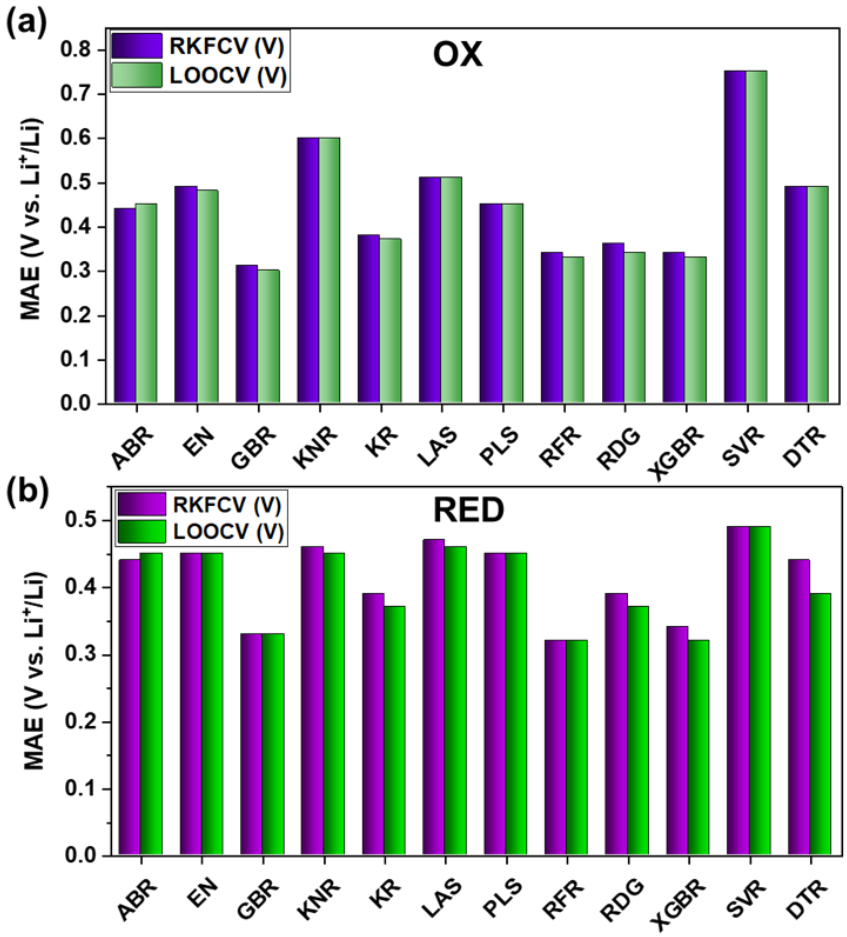


Figure 5.3: Bar plot of mean absolute error (MAE) calculated using repeated K-fold cross-validation (RKFCV) and leave one out cross-validation (LOOCV). Error bar for (a) Oxidation potential, and (b) Reduction potential. For RKFCV, 5 times repeated 10-fold CV has been considered.

Table 5.2: The cross validated MAE of the ML algorithms for the prediction of oxidation potential and reduction potential of various solvents.

ML Models	MAE (V)/OX			MAE (V)/Red		
	10-fold CV	R-K-fold CV	LOOCV	10-fold CV	R-K-fold CV	LOOCV
LS	0.50	0.51	0.51	0.47	0.47	0.47
PLS	0.45	0.45	0.45	0.45	0.45	0.45
RR	0.35	0.37	0.35	0.39	0.40	0.37
KRR	0.37	0.38	0.37	0.38	0.39	0.37
ENR	0.49	0.49	0.48	0.45	0.45	0.45
KNR	0.61	0.60	0.60	0.47	0.46	0.45
SVR	0.75	0.75	0.75	0.49	0.49	0.49
ABR	0.42	0.43	0.44	0.45	0.45	0.46
GBR	0.30	0.31	0.30	0.34	0.32	0.33
DTR	0.49	0.50	0.50	0.43	0.43	0.39
RFR	0.33	0.34	0.33	0.32	0.32	0.32
XGBR	0.32	0.33	0.33	0.35	0.34	0.32

From the CV result **(Figure 5.3)**, it has been observed that linear ML models are not suitable to capture the underlying pattern of the data for the prediction of both oxidation and reduction potential as the MAE is relatively high compared to the other ML models which reflects the complexity of the dataset. Low MAE has been observed for the tree-based algorithms indicating the high potential of learning the underlying factors affecting the oxidation and reduction potential of the solvent electrolytes. However, the consistency in MAE through two different CV methods have been remain conserved showcasing the stability of the applied ML models. Among the tree-based algorithms GBR, RFR, and XGBR have been found to produce higher accuracy for the prediction of red-ox (Reduction and oxidation)

potential. Hence, from the primary CV assessment, we proceed with these three algorithms.

5.4.2. Feature Engineering

The learning of ML models with such high dimensional (210) data is a highly challenging task and also computationally costly. Thus, the focus has been shifted to eliminate unnecessary features from the input space to make the training process more efficient. Numerous techniques are available to reduce high-dimensional datasets by eliminating redundant features. For example, a method named NCOR-FS is proposed, which incorporates materials domain knowledge through Non-Co-Occurrence Rules (NCORs).*[43]* NCORs quantify how well feature subsets adhere to domain knowledge, and the feature selection process is optimized using a swarm intelligence algorithm. Lin and co-workers have adopted recursive feature elimination process to eliminate the unnecessary features which do not contribute to the achievement of higher accuracy.*[44]* In our previous study we have successfully applied statistical Select-K-Best method to reduce the dimension of our data and got some excellence accuracy for the prediction of adsorption energy.*[34]* Here, in addition to Select-K-Best, model dependent feature selection has also been applied to make the feature selection process more robust. Hence, we have applied two-stepwise feature elimination methods, namely, Select-K-Best, which is a model independent feature selection method, followed by model dependent feature selection. The Select-K-Best measures the variance of target variable with respect to the input variables through ANOVA-F value analysis, a statistical method to compute the model independent feature importance. We have applied this method on the three selected ML models, GBR, RFR, and XGBR to reduce the dimension of the red-ox dataset based on the minimum MAE error. The higher the F-value, the higher will be the correlation between that input feature and the output. We have stridden the selection of features by 10 starting with 10 features, thus, generating 21 number of different feature sets. For each feature set the MAE error has been computed for all the three

models for red-ox potential data. The feature selection plot for XGBR and RFR are depicted in **Figure 5.4**, and the same for GBR has been shown on **Figure 5.5**.

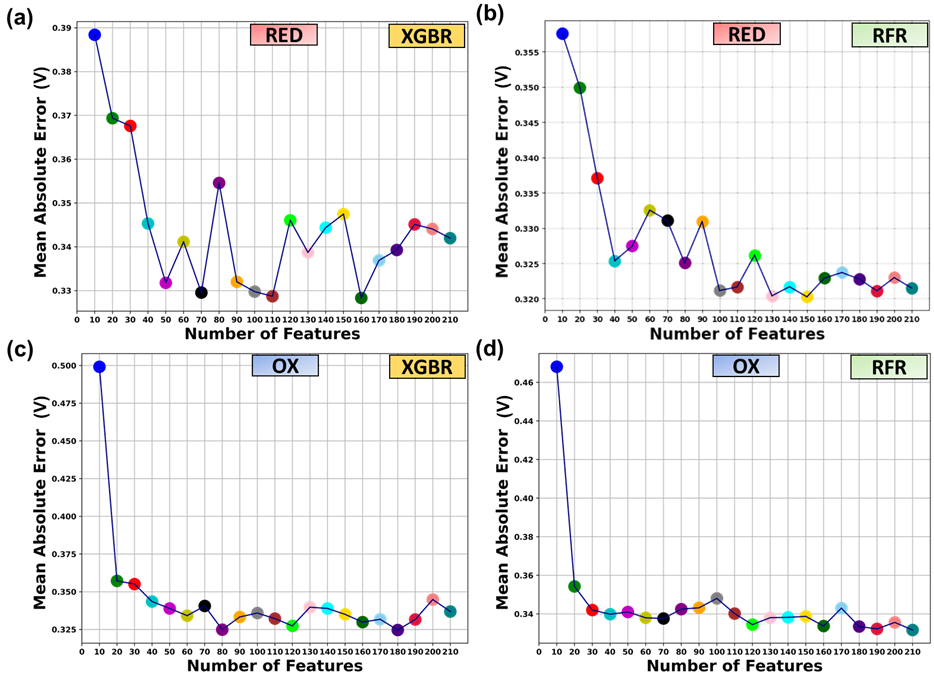


Figure 5.4: Feature selection plot based on the MAE for all the 21 feature sets. (a) XGBR/RED, (b) RFR/RED, (c) XGBR/OX, and (d) RFR/OX. All the different colour balls are different features sets consisting of top ranked features.

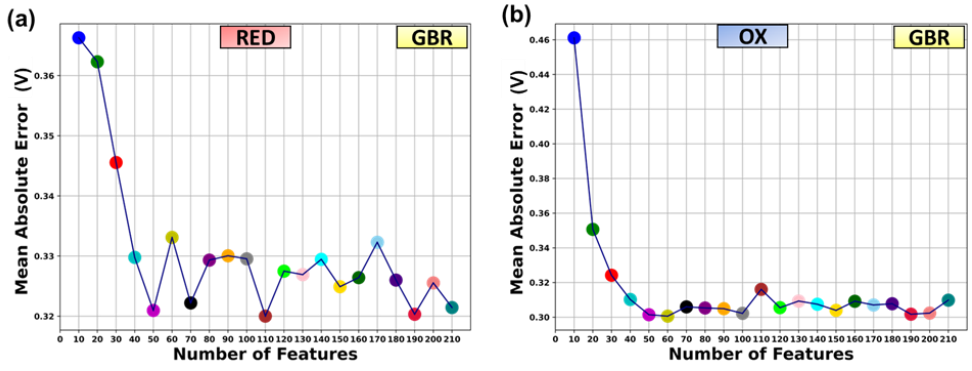


Figure 5.5: Feature selection plot based on the MAE for all the 21 feature sets. (a) GBR/RED, and (b) GBR/OX.

The error has been calculated through RKFCV method on the whole dataset to ensure the stability of the ML models and to avoid the stochasticity problem due to the automatically and randomly train-test split data set. The selection of best features is based on the ANOVA-F value and not selected sequentially. Hence, each set of features are comprised of top ranked features.

The number of features providing minimum error in case of reduction potential for GBR, XGBR, and RFR are 110, 160, and 130 respectively whereas for oxidation potential 60, 80, and 190 respectively (**Figure 5.4, and Figure 5.5**). The different number of features for different algorithms indicates the various learning capability of these ML models. Further, we have processed the data with these statistically selected top ranked features for each model to calculate the model dependent feature importance. The feature importance of each feature extracted from Select-K-Best method for each algorithm was then calculated, and the results were sorted in descending order. Our approach involved starting with the top 10 most influential features and employing RKFCV to compute the MAE. Subsequently, we incrementally increased the number of features by 10 and re-evaluated the MAE for each feature set. This process was iteratively performed for three ML algorithms (XGBR, RFR, and GBR). The optimization of the feature selection method was determined by identifying the minimum MAE across different feature sets for each algorithm. We have discarded all those features with very less contribution towards the prediction of target variable for the corresponding ML models. The selection of the most contributed features is based on the comparing MAE value and thus we have further reduced the dimensionality the dataset without any compromise in the accuracy for the prediction of red-ox potential. The feature importance plot for reduction potential and oxidation potential of XGBR and RFR models has been depicted in **Figure 5.6** and the same for GBR has shown in **Figure 5.7**.

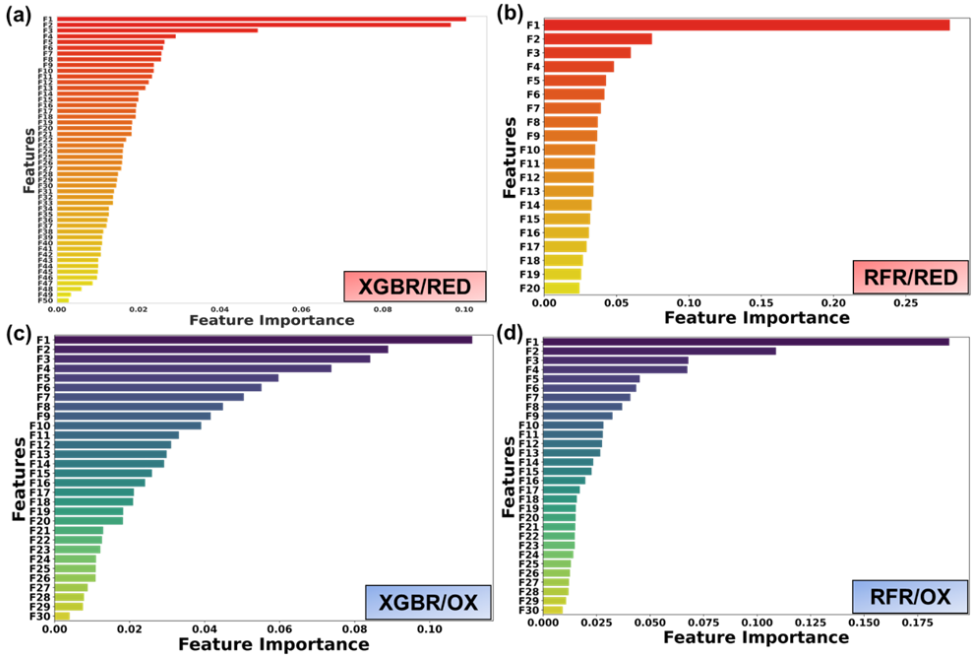


Figure 5.6: Model dependent feature importance plot for (a) XGBR/RED, (b) RFR/RED (c) XGBR/OX, and (d) RFR/OX. Features corresponding to each algorithm for reduction and oxidation potential have been tabulated in **Table 5.3 and 5.4** respectively.

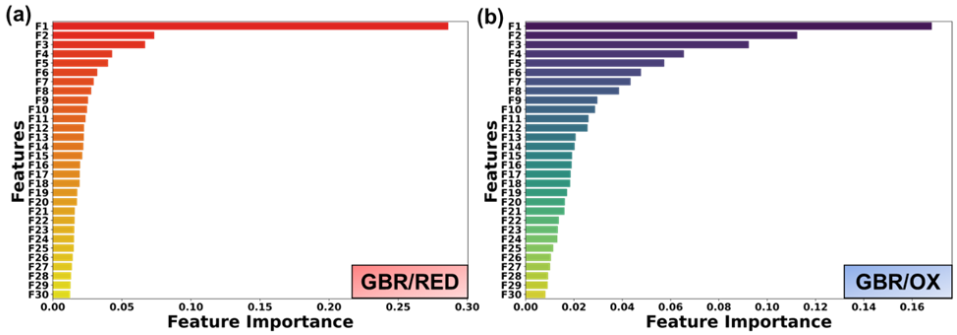


Figure 5.7: Model dependent feature importance plot for (a) GBR/RED, and (b) GBR/OX.

Table 5.3: The list of most contributed features for reduction potential for GBR, RFR, and XGBR.

GBR/RED		RFR/RED		XGBR/RED	
symbol	**Features Name**	**symbol**	**Features Name**	**symbol**	**Features Name**
F1	SMR_VSA10	F1	SMR_VSA10	F1	SMR_VSA10
F2	fr_halogen	F2	BCUT2D_MWHI	F2	fr_halogen
F3	VSA_EState7	F3	MinAbsEStateIndex	F3	FractionCSP3
F4	HallKierAlpha	F4	PEOE_VSA5	F4	HallKierAlpha
F5	BCUT2D_CHGLO	F5	FractionCSP3	F5	EState_VSA10
F6	BCUT2D_MRHI	F6	BCUT2D_MRHI	F6	VSA_EState7
F7	BCUT2D_MWHI	F7	HallKierAlpha	F7	VSA_EState4
F8	MinAbsEStateIndex	F8	BCUT2D_CHGLO	F8	SMR_VSA7
F9	qed	F9	PEOE_VSA8	F9	PEOE_VSA5
F10	FpDensityMorgan2	F10	VSA_EState10	F10	Chi3v
F11	SMR_VSA6	F11	fr_halogen	F11	EState_VSA1
F12	MinPartialCharge	F12	BCUT2D_MRLOW	F12	BCUT2D_LOGPHI
F13	FpDensityMorgan1	F13	VSA_EState7	F13	PEOE_VSA12
F14	EState_VSA10	F14	MinPartialCharge	F14	SlogP_VSA2
F15	Kappa2	F15	FpDensityMorgan1	F15	PEOE_VSA6
F16	BCUT2D_LOGPLOW	F16	qed	F16	SlogP_VSA3
F17	PEOE_VSA5	F17	VSA_EState4	F17	PEOE_VSA14
F18	VSA_EState4	F18	SlogP_VSA3	F18	BCUT2D_CHGHI

F19	SlogP_VSA2	F19	SlogP_VSA2	F19	Chi2v
F20	VSA_EState3	F20	Chi3v	F20	BCUT2D_MWHI
F21	Chi2v			F21	Ipc
F22	MaxPartialCharge			F22	PEOE_VSA4
F23	PEOE_VSA14			F23	BCUT2D_CHGLO
F24	FractionCSP3			F24	PEOE_VSA8
F25	SMR_VSA7			F25	fr_NH0
F26	Chi3v			F26	Chi0
F27	PEOE_VSA9			F27	SlogP_VSA10
F28	MaxAbsPartialCharge			F28	NumValenceElectrons
F29	BCUT2D_MWLOW			F29	Chi1n
F30	SlogP_VSA3			F30	qed
				F31	FpDensityMorgan1
				F32	Chi0n
				F33	SMR_VSA6
				F34	AvgIpc
				F35	VSA_EState1
				F36	SMR_VSA3
				F37	EState_VSA8
				F38	PEOE_VSA7
				F39	VSA_EState3
				F40	Kappa1
				F41	SlogP_VSA4
				F42	TPSA
				F43	EState_VSA2
				F44	BCUT2D_MRHI
				F45	LabuteASA
				F46	VSA_EState8
				F47	MaxPartialCharge
				F48	PEOE_VSA1

		F49	MinAbsEStateIndex
		F50	BertzCT

Table 5.4: The list of most contributed features for oxidation potential for GBR, RFR, and XGBR.

GBR/OX		RFR/OX		XGBR/OX	
Symbol	**Feature Name**	**Symbol**	**Feature Name**	**Symbol**	**Feature Name**
F1	TPSA	F1	TPSA	F1	PEOE_VSA4
F2	PEOE_VSA4	F2	PEOE_VSA4	F2	SMR_VSA7
F3	MaxPartialCharge	F3	MaxPartialCharge	F3	TPSA
F4	BCUT2D_MWHI	F4	BCUT2D_MWHI	F4	SlogP_VSA10
F5	VSA_EState7	F5	MinPartialCharge	F5	PEOE_VSA14
F6	PEOE_VSA14	F6	VSA_EState7	F6	EState_VSA10
F7	MinAbsPartialCharge	F7	BCUT2D_LOGPLOW	F7	fr_Ndealkylation1
F8	AvgIpc	F8	PEOE_VSA14	F8	SMR_VSA3
F9	HallKierAlpha	F9	MinAbsEStateIndex	F9	VSA_EState7
F10	PEOE_VSA8	F10	MinAbsPartialCharge	F10	SlogP_VSA1
F11	VSA_EState3	F11	MaxAbsPartialCharge	F11	fr_amide
F12	SMR_VSA7	F12	PEOE_VSA8	F12	VSA_EState3
F13	SlogP_VSA6	F13	Chi3v	F13	MaxPartialCharge
F14	SMR_VSA6	F14	MolLogP	F14	NumAromaticHeterocycles
F15	SlogP_VSA1	F15	FpDensityMorgan3	F15	MinPartialCharge

F16	FpDensityMorgan3	F16	SlogP_VSA10	F16	BCUT2D_MWHI
F17	Kappa1	F17	AvgIpc	F17	BCUT2D_LOGPHI
F18	EState_VSA1	F18	VSA_EState6	F18	SMR_VSA2
F19	MinEStateIndex	F19	Kappa1	F19	fr_alkyl_halide
F20	VSA_EState6	F20	VSA_EState3	F20	SlogP_VSA11
F21	EState_VSA10	F21	FpDensityMorgan1	F21	PEOE_VSA8
F22	fr_alkyl_halide	F22	SMR_VSA7	F22	EState_VSA2
F23	SMR_VSA3	F23	BCUT2D_MRHI	F23	EState_VSA1
F24	SlogP_VSA10	F24	EState_VSA10	F24	fr_furan
F25	Chi0n	F25	fr_alkyl_halide	F25	Chi0n
F26	BCUT2D_LOGPHI	F26	FpDensityMorgan2	F26	Chi1n
F27	Chi1n	F27	Kappa2	F27	fr_lactone
F28	fr_amide	F28	BCUT2D_LOGPHI	F28	SlogP_VSA7
F29	BCUT2D_MRHI	F29	MaxAbsEStateIndex	F29	SMR_VSA6
F30	fr_halogen	F30	Chi4v	F30	MinAbsPartialCharge

In case of reduction potential, there are 30, 50, and 20 number of features have been found to be the most contributing factor for GBR, XGBR, and RFR ML models respectively whereas for oxidation potential we have got 30 features for all the ML models. All the most contributed features corresponding to each algorithm for reduction and oxidation potential have been tabulated in **Table 5.3-5.4**. It has been found that the feature F1 (SMR_VSA10) is the most contributed feature for the learning of reduction

potential for all three ML models. It suggests that this specific aspect of solvent accessibility is highly related to the surface area, can capture the information about the distribution of electron density on the surface of the solvent molecules. High values of this descriptor might indicate regions of the molecule that are more accessible to electrons, potentially influencing the reduction potential. However, for oxidation potential prediction TPSA (Topological Polar Surface Area) is the most contributing feature for both GBR and RFR algorithms and PEOE_VSA4 (Partial Equalization of Orbital Electronegativities and the Variable Shape Descriptors for Atoms) for XGBR. TPSA is related to the distribution of polar atoms and polar bonds on a molecule. In the context of oxidation potential, it captures the information about electron donation or withdrawal capabilities thus highly related to oxidation potential. PEOE_VSA4 specifically represents the Potential Energy for the fourth bin in the Variable Shape Descriptor (VSA) scheme. The PEOE_VSA method divides a molecule into bins based on its van der Waals surface, and each bin corresponds to a specific range of electrostatic potential values.

5.4.3. Hyperparameter Tuning

After selecting the most contributed features for each ML model for both reduction and oxidation potential, we have processed the data further for hyperparameter tuning to improve the accuracy of these ML models. We have utilized the RandomizedSearchCV with 500 iterations as implemented in the scikit-learn library to find out the optimum hyperparameters for each ML model.45 All the optimized hyperparameters are tabulated in **Table 5.5**.

Table 5.5: Optimized hyperparameters used for the testing of ML models for the prediction of reduction and oxidation potential.

ML Models	**Optimized Hyperparameters**	
	Reduction Potential	Oxidation Potential
GBR	n_estimators=300, max_depth=5, min_samples_split=10	n_estimators=1800, min_samples_leaf=2, min_samples_split=5

XGBR	n_estimators=800, max_depth=11, learning_rate=0.05, gamma=0.1, reg_alpha=0.5, reg_lambda=0.5, min_child_weight=5	n_estimators=100, max_depth=13, learning_rate=0.1, gamma=0.1, reg_alpha=0.5, reg_lambda=0, min_child_weight=3
RFR	n_estimators=200, max_depth=9, min_samples_leaf=2	n_estimators=60, max_depth=11

With the optimized hyperparameters, the MAE of each ML model has been calculated on 80% train data and 20% test data. In the scatter plot **(Figure 5.8 and Figure 5.9)**, we have demonstrated the solvent electrolytes with respect to different functional group.

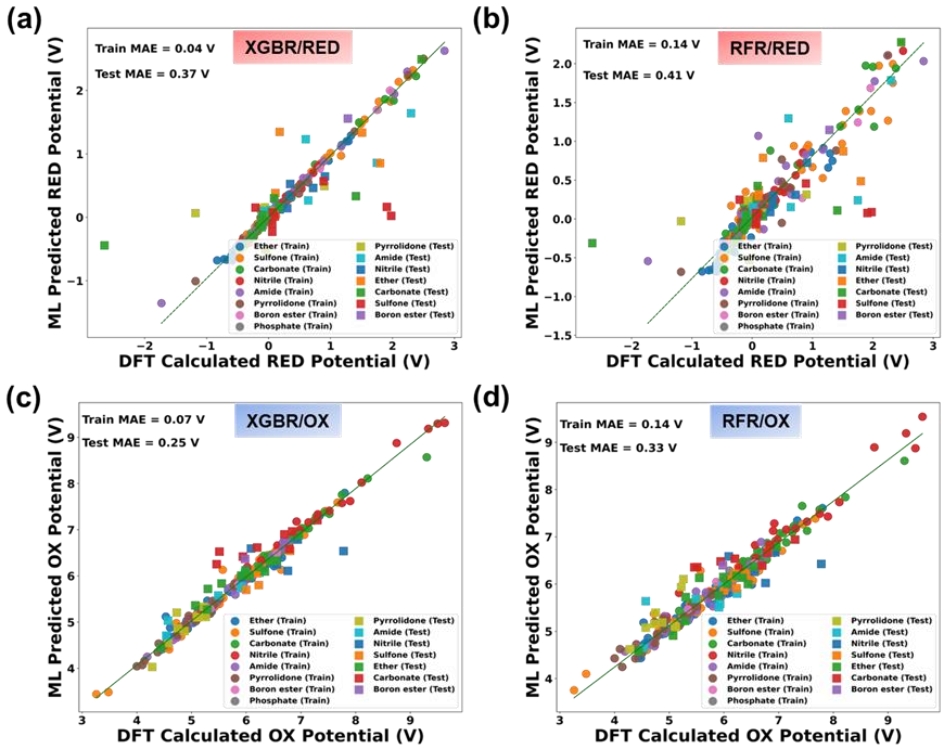


Figure 5.8: Parity plot to compare the DFT calculated and ML predicted result for (a) XGBR/RED, (b) RFR/RED, (c) XGBR/OX, and (d) RFR/OX.

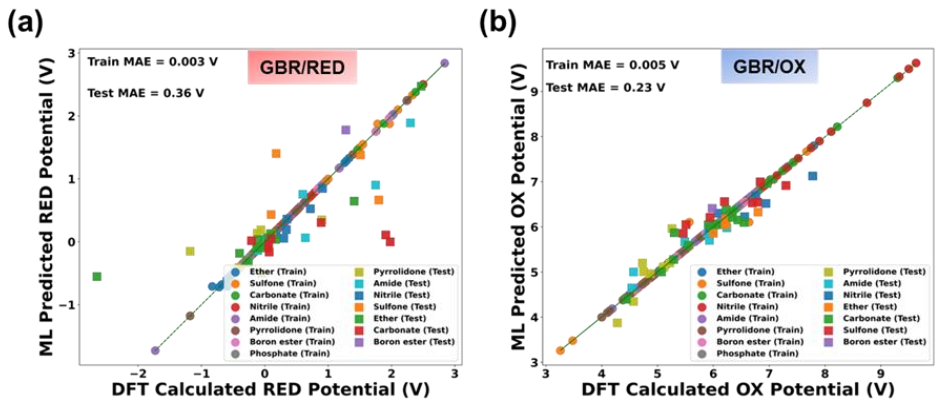


Figure 5.9: Parity plot to compare the DFT calculated and ML predicted result for (a) GBR/RED, and (b) GBR/OX.

The names of the functional groups are represented in the inset of the plot. There is total eight different functional groups present in the data. For both the reduction potential and oxidation potential XGBR has been found to be performed well with a test MAE of 0.37 V and 0.25 V respectively. The train and test MAE for RFR is much higher compared to XGBR and GBR which indicates the inefficient training of the RFR model. However, for GBR, though the test errors for both reduction (0.36 V) and oxidation (0.23 V) potential are comparable with XGBR, the training errors are 0.003 V, and 0.005 V indicates the overfitting issue where the ML model is able to learn the training data accurately while unable to predict the test data. Thus, we have excluded both the GBR and RFR models for the prediction of reduction and oxidation potential. The accuracy of the XGBoost Regressor (XGBR) model is substantiated by the outcomes of both repeated RKFCV and LOOCV, as illustrated in **Figure 5.3**. These results serve as compelling evidence of the model's stability and generalizability, affirming its robust predictive capabilities. Thus, for both reduction and oxidation potential, we have selected XGBR model as the most suited ML model for further application in unknown solvent space. This could be attributed to its exceptional scalability and proficiency in managing extensive datasets characterized by intricate non-linear correlations between input features and output variables.*[46]* From the reduction and oxidation potential, we have

further calculated the ECW values and compared it with the DFT calculated ECW values. The predicted results show the high absolute error in ECW for few solvents is mainly originates from the reduction potential value. Learning the reduction potential through the trained ML model with 308 number solvent data covering various functional groups is highly challenging compared to oxidation potential. This could be because of presence of both negative and positive reduction potential value present in the data set as well as very high variance with respect to change in the structure of the solvent molecules. For example, though there are lot of similarity between the structure of 4-methyl-1,3-dioxolane (4ME13DOL) and 2-methyl-1,3-dioxolane (2ME13DOL), the reduction potentials are not even close to each other (-2.65 V and -0.42 V) whereas the oxidation potentials are very close to each other (5.29 V and 5.35 V). Similar high error trend has been observed for ethyl methanesulfonylacetate (EMSA) and 2-(methylsulfonyl)ethyl acetate (MSEA) because of high variance. Thus, though ML model is able to predict the oxidation potential accurately, is unable to predict the reduction potential with such high accuracy. Increasing the number of solvents in the training set with similar type of solvent molecules can mitigate the issue of unable to distinguish between almost two similar structures. However, as the models are trained on the extracted DFT calculated data, we have not included any experimental data to maintain the homogeneity of the data and to ensure the method of calculated ECW remain same for each data point.

5.4.4. Validation of ML Prediction

In addition to cross-validation, we assessed the ML-predicted oxidation and reduction potentials for certain solvents in the validation set, which were not present in either the training or testing sets. Specifically, we compared these predictions against experimentally reported values for at least one oxidation and reduction potential. The results of this comparative analysis are presented in **Table 5.6**. The observed conformity between the ML-predicted oxidation and reduction potentials and the corresponding

experimental values, consistent with the trends identified through DFT, underscores the accurate prediction ability of the ML model. Consequently, the XGBR model exhibits the capability to efficiently determine the ECW of solvent electrolytes within a short timeframe and with minimal resource requirements. This predictive ability positions the model as a valuable tool for guiding experimental researchers in the exploration of novel solvent electrolytes for rechargeable metal-ion batteries.

Table 5.6: Comparison of DFT and experimentally measured oxidation potential, reduction potential, and ECW of solvent electrolytes belongs to validation set with the XGBR model predicted oxidation potential, reduction potential, and ECW value.

Solvents	**Oxidation Potential** (V *vs.* Li^+/Li)			**Reduction Potential** (V *vs.* Li^+/Li)			**ECW** (V *vs.* Li^+/Li)		
	DFT	**ML**	**Exp.**	**DFT**	**ML**	**Exp.**	**DFT**	**ML**	**Exp.**
Ethyl isobutyl sulfone (EiBS)	5.75	5.89	5.6	-0.34	-0.31	/	6.09	6.2	/
Ethyl methyl sulfone (EMS)	6.01	6.01	5.9	-0.27	-0.17	/	6.28	6.18	/
2-Fluoropropyl methyl carbonate (2FPMC)	6.6	6.53	6.4	-0.14	-0.23	/	6.74	6.76	/
Dimethoxymethane (DMM)	5.4	5.2	3.5	-0.19	-0.39	/	5.59	5.59	/

2-Fluoroethyl acetate (2FEA)	6.37	6.48	5.8	-0.06	0.09	/	6.43	6.39	/
Dipropyl sulfone (DPS)	5.79	5.74	5.7	0.16	-0.39	/	5.63	6.13	/
Methyl propyl carbonate (MPC)	6.16	6.47	6.4	-0.14	-0.25	/	6.3	6.72	/
n-Methylacetamide (NMA)	5.32	5.48	5.3	0.03	0.02	/	5.29	5.46	/
1-(Ethanesulfonyl)-2-methoxyethane (EMES)	5.53	5.50	/	-0.35	-0.3	/	5.88	5.8	5.6
2-Fluoroethyl propionate (2FEP)	6.13	6.15	6.2	-0.22	0.23	/	6.35	5.92	/
Fluoromethyl methyl carbonate (MFDMC)	6.98	7.15	6.57	-0.2	0.1	0.19	7.18	7.05	6.38
Methyl 2,2,3,3,3-	7.05	6.98	6.6	0.08	0.15	/	6.97	6.83	/

pentafluoropropyl carbonate (PFPMC)									
1-Ethoxy-2-(2,2,2-trifluoroethoxy) ethane (ETFEE)	5.29	5.32	5.75	-0.52	-0.52	/	5.81	5.84	/

Once the XGBR model was validated, we applied SHAP (SHapley Additive exPlanations) analysis to gain an in-depth understanding of the model through local feature analysis. Given that XGBR is a black-box model, SHAP analysis is particularly valuable in identifying which features significantly influence the accuracy of predictions and which are responsible for less accurate predictions. We conducted SHAP analysis for both reduction and oxidation potential predictions to enhance the interpretability and robustness of our results. Specifically, we selected two systems for each type of prediction: one with the most accurate predictions and another with the most deviated predictions. The SHAP waterfall plots for the systems with the most accurate predictions are depicted in **Figure 5.10**, while those for the most deviated systems are shown in **Figure 5.11**. For reduction potential predictions, features such as MinAbsEStateIndex, SMR_VSA10, and IPC had the most significant influence, contributing negatively to the accurate predictions **(Figure 5.10a)**. For oxidation potential predictions, TPSA, MaxPartialCharge, and Estate-VSA10 were found to be crucial for accurate predictions **(Figure 5.10b)**.

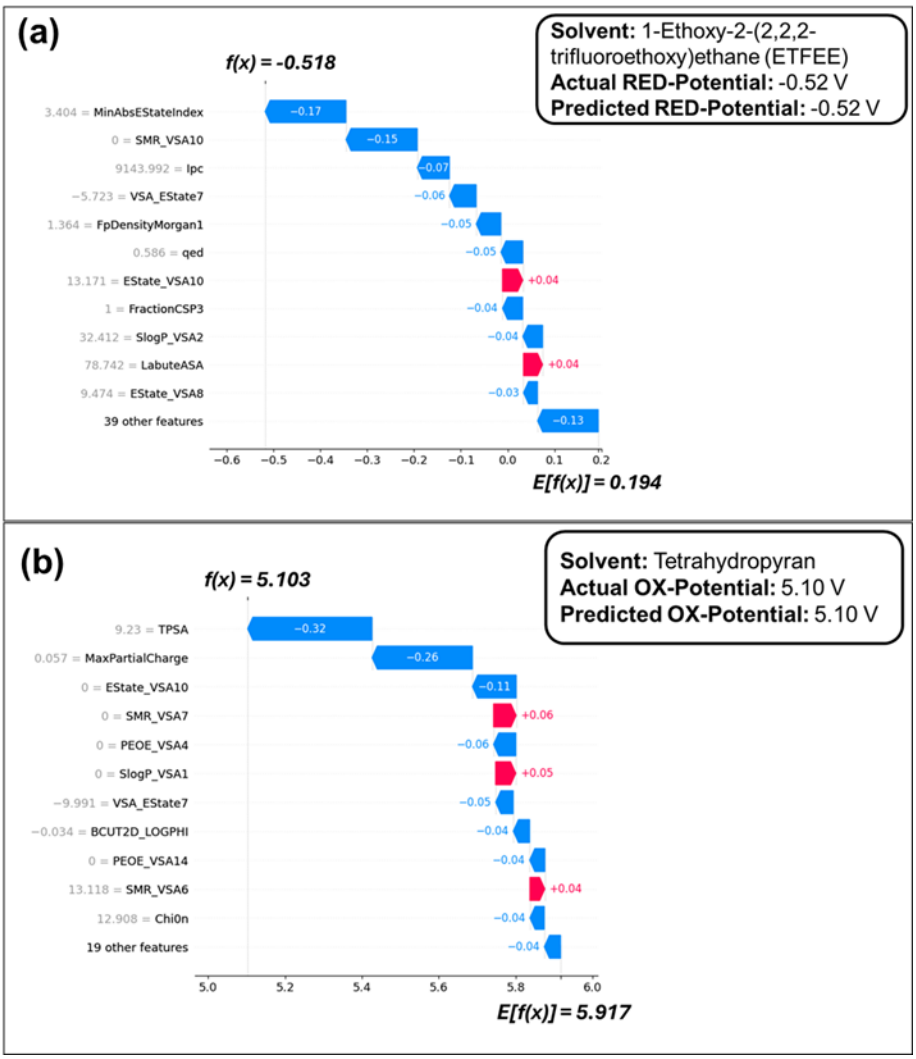


Figure 5.10: SHAP waterfall plot for the most accurately predicted (a) oxidation potential system, and (b) reduction potential system.

For reduction potential predictions, MinAbsEStateIndex highlights the most electron-dense regions, SMR_VSA10 reflects polarizable surface areas, and IPC indicates structural complexity. These factors collectively influence how readily a molecule accepts electrons. For oxidation potential predictions, TPSA represents the molecule's polar surface area, MaxPartialCharge points to the most electron-deficient site, and Estate-VSA10 combines electronic state with surface accessibility.

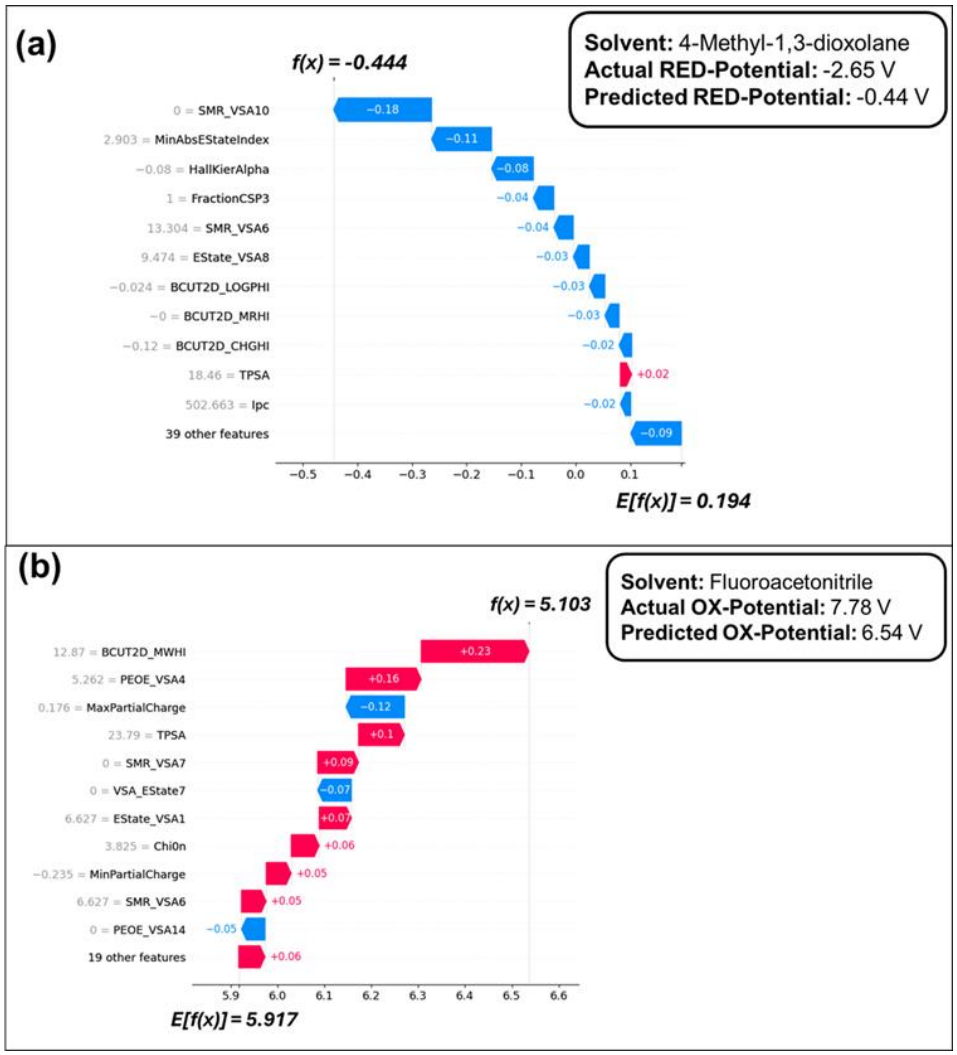


Figure 5.11: SHAP waterfall plot for the most deviated (a) oxidation potential system, and (b) reduction potential system.

These features affect how easily a molecule can lose electrons, stabilizing the positive charge generated during oxidation. In the case of the most deviated reduction potential system, although MinAbsEStateIndex and SMR_VSA10 remained influential, their order and contribution differed, and the absence of IPC significantly impacted the prediction accuracy. For the most deviated oxidation potential system, the deviation was attributed to the absence of key features present in the most accurate system. Additionally, several features contributed negatively, underscoring the differences between the most accurate and most deviated systems. This

local feature analysis using SHAP provides a clear picture of each feature's contribution to redox potential predictions. By applying SHAP to our XGBR model, we establish a baseline, identifying important features along with their magnitude and direction for accurate redox potential predictions, thus enhancing the model's transparency and interpretability.

5.5. Unknown Solvent Space Exploration

Following the finalization of the XGBR model for determining ECW, we applied the model to systematically explore an extensive unknown solvent space concerning ECW, leveraging optimized features. We amassed a dataset comprising of 4882 solvents, each accompanied by its SMILES string, sourced from the PARIS III database, widely utilized across various industries.*[47]* Employing RDkit, we generated the requisite features for applying the optimized XGBR model, enabling the prediction of ECW for each solvent electrolyte. Utilizing selected features and fine-tuned hyperparameters, we initially predicted the oxidation and reduction potentials of the solvent molecules. Subsequently, we subjected the resultant data to a clustering model, categorizing the solvent database into smaller subgroups. This clustering approach offers a strategic means for experimental researchers to navigate the vast solvent landscape, facilitating the identification of promising candidates for battery electrolyte testing while circumventing the inefficiencies associated with trial-and-error methodologies. The outcomes of these predictions, coupled with the data analysis and clustering methodologies, are elaborated upon in the subsequent sections.

5.5.1. Clustering of Solvent Electrolytes

Utilizing the XGBR model with optimized hyperparameters we have determine the ECW of 4882 solvent electrolytes. The lowest ECW is observed for 4-Methylbenzyl chloride (2.29 V) where Methyl glycidyl ether shows highest ECW (9.67 V). The ML-predicted ECW of Methyl glycidyl ether, also known by its IUPAC name 2-(Methoxymethyl)oxirane is an

epoxide, characterized by a three-membered ring containing an oxygen atom. The molecular structure of 2-(Methoxymethyl)oxirane can be found in **Figure 5.12**.

Figure 5.12: Structure of 2-(Methoxymethyl)oxirane.

The presence of electronegative atoms can lead to a stabilized anodic limiting potential energy level, which results in a larger energy gap between the cathodic and anodic limiting potentials, thus yielding a higher ECW. Additionally, the high ECW prediction can be attributed to the unique presence of an epoxide group in the solvent molecule. It is important to note that the training set used for our XGBR model did not include any solvent electrolytes containing an epoxide group. This divergence arises because the training data spans a range of approximately 3.2 V to 9.6 V **(Figure 5.2)**. The extended region in the violin plot represents areas where the ML model predicts values that are at the extremes or even slightly beyond the actual observed range in your dataset. The oxidation potential prediction for this solvent is notably high at 9.31 V. However, this is the highest predicted ECW value, suggesting that it may represent an extreme point with an associated margin of error. Given that this is an ML-predicted value, it is inherently subject to some degree of uncertainty. The main objective is to identify suitable solvent electrolytes with optimum ECW that can be compatible with high working voltage electrode materials. We divided the ECW region in three equal range starting from minimum ECW value to maximum value. This has been done by subtracting the maximum ECW from the minimum ECW followed by divided by three. Thus, we have categorized the solvent molecules in three categories, Low ECW (LECW)

(2.29 – 4.75 V), Medium-ECW (MECW) (4.76 V – 7.21 V), and High-ECW (HECW) (7.22 V – 9.67 V). However, analysing the individual 4882 solvent electrolytes one by one is practically impossible. Thus, to gain chemical insights on the predicted data, first we have performed clustering using unsupervised method on the predicted ECWs of 4882 solvents and subgroup them in different clusters followed by categorized each cluster into the three categories. Clustering is a fundamental technique in unsupervised machine learning, used to group data points into distinct clusters or categories based on their inherent similarities. Unlike supervised learning, where the algorithm is provided with labelled training data to learn from, unsupervised learning operates on unlabelled data, seeking to discover hidden patterns, structures, or relationships within the data. There are various clustering algorithms such as k-Means, Hierarchical, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), Gaussian Mixture Model (GMM), and Cloud-Clustering, etc. Each algorithm has its own advantages and disadvantages depending on the specific problem. For example, the Cloud-Clustering method can quantify the degree of uncertainty during clustering, whereas the GMM model can recognize patterns such as speaker identification through clustering.*[48,49]* However, both the Cloud-Cluster method and GMM are in their infancy stage in the field of material science and can be more computationally expensive than k-Means due to the additional complexity of handling parameters such as expectation, entropy, hyper-Entropy for Cloud-Clustering, and means, variances, and mixing coefficients for GMM. This increased the computational cost for large dataset like our case. On the other hand, k-Means is a well-established algorithm widely validated and commonly used method in the field of chemistry and materials science, providing confidence in its applicability to our data.*[50,51]* Additionally, the interpretability of k-Means clustering is easier compared to other clustering algorithms, making it more straightforward to understand and explain the

results Given these advantages we have considered k-Means clustering for its balance of efficiency, simplicity, and robustness.

5.5.2. Optimization of Clusters

Here in this study, we have implemented widely applicable k-Means clustering technique to group down the similar types of data points present in our dataset. k-Means aims to partition data into k number of clusters by iteratively minimizing the variance. Each cluster is represented by its centroid, and data points are assigned to the cluster whose centroid is nearest to them. For the clustering method, we primarily used features from the oxidation potential dataset, which demonstrated significantly lower prediction errors to minimize the overall error in the clustering process. Post-clustering, we conducted PCA and found that the contribution of ECW to the principal components was minimal, indicating that ECW errors had a limited impact on clustering. Excluding it would hinder the effective linkage of clustering results with ECW classification. Choosing the optimized number of clusters i.e., the optimum value of k is highly important, and the parameter tuning is necessary. We have implemented Elbow method to find out the optimized number of clusters. It provides a visual aid for selecting the appropriate number of clusters based on the within-cluster sum of squares (WCSS), which measured the variability or dispersion also known as distortion of data points within each cluster. In the context of k-Means clustering, the WCSS for a given number of clusters (k) is calculated as the sum of the squared Euclidean distances between each data point in a cluster and its centroid, summed over all clusters. Mathematically, for k number of clusters:

$$Distortion = \sum_{i=1}^{k} \sum_{x \in c_i} ||x - \mu_i||^2 \quad (1)$$

Where k and c_i represent the number of clusters and the data points present in cluster "i" respectively. x is the individual data, μ_i is the centroid of cluster "i" and $||x - \mu_i||$ is the Euclidian distance between the data point x and the centroid μ_i.

The plot of the number of clusters (k) against the distortion has shown in **Figure 5.13a**. The graph typically resembles an elbow. The point at which this graph shows a noticeable bend or inflection is known as the elbow point. This point is where the rate of decrease in distortion starts to slow down significantly. We have fixed the maximum number of clusters as 20 and then calculated the distortion or WCSS for each cluster **(Figure 5.13a)**. However, from the elbow curve the point at which rate of decrease in distortion is slowing down is not clear and indistinguishable. Thus, to identify the elbow point, we further calculate Silhouette score for each cluster **(Figure 5.13b)**. The Silhouette Score is a metric used to evaluate the quantity of clustering in unsupervised learning. The highest Silhouette score has been observed with 11 number of clusters. The 11 optimized number of clusters indicates that machine is able to group down all the solvent in 11 subclasses where each subclass is expected to show similar kind of characteristics. Thus, we fixed $k = 11$ and performed the clustering method. While the anticipated number of clusters in the established DFT dataset aligns with the eight distinct categories of solvent sets featuring various functional groups, the identification of eleven clusters within the unknown solvent space underscores the remarkable capacity of k-Means clustering.

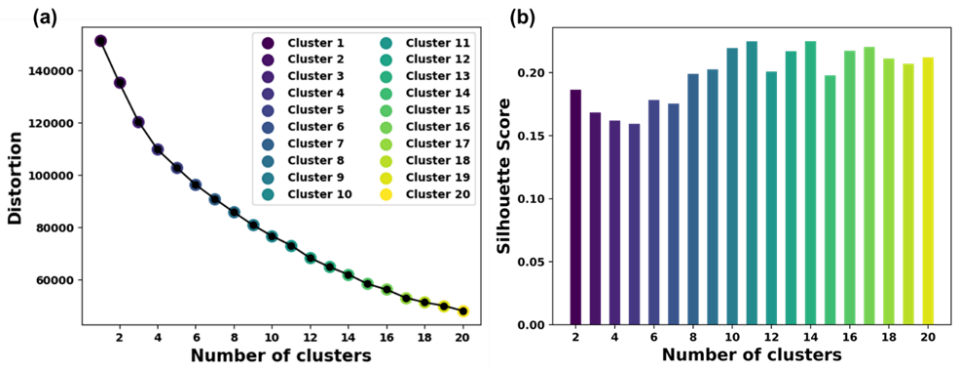


Figure 5.13: Optimization of number of clusters where (a) elbow curve, and (b) bar plot of Silhoutte score. We have considered the maximum number of clusters as 20.

5.5.3. Cluster Data Analysis

The visualization of all 11 clusters in a two-dimensional space poses a considerable challenge. To surmount this, we employed principal component analysis (PCA) to effectively reduce the dimensionality of the dataset scaled by Minmax scaler. Through this process, we distilled the features into two principal components—Principal Component 1 (PC1) and Principal Component 2 (PC2). Each component represents a linear combination of existing features, with PC1 exerting a greater influence, followed by PC2. The resulting scatter plot in **Figure 5.14a** illustrates the distribution of the 11 clusters in relation to PC1 and PC2. Each cluster is distinguished by a different colour ball, while the cross marks pinpoint the centroids of these clusters. Notably, the scatter plot reveals a substantial overlap among the clusters, indicative of the intricate and complex nature of the unknown solvent space. This observation underscores the inherent challenge in achieving clear separations into distinct clusters within this multifaceted domain.

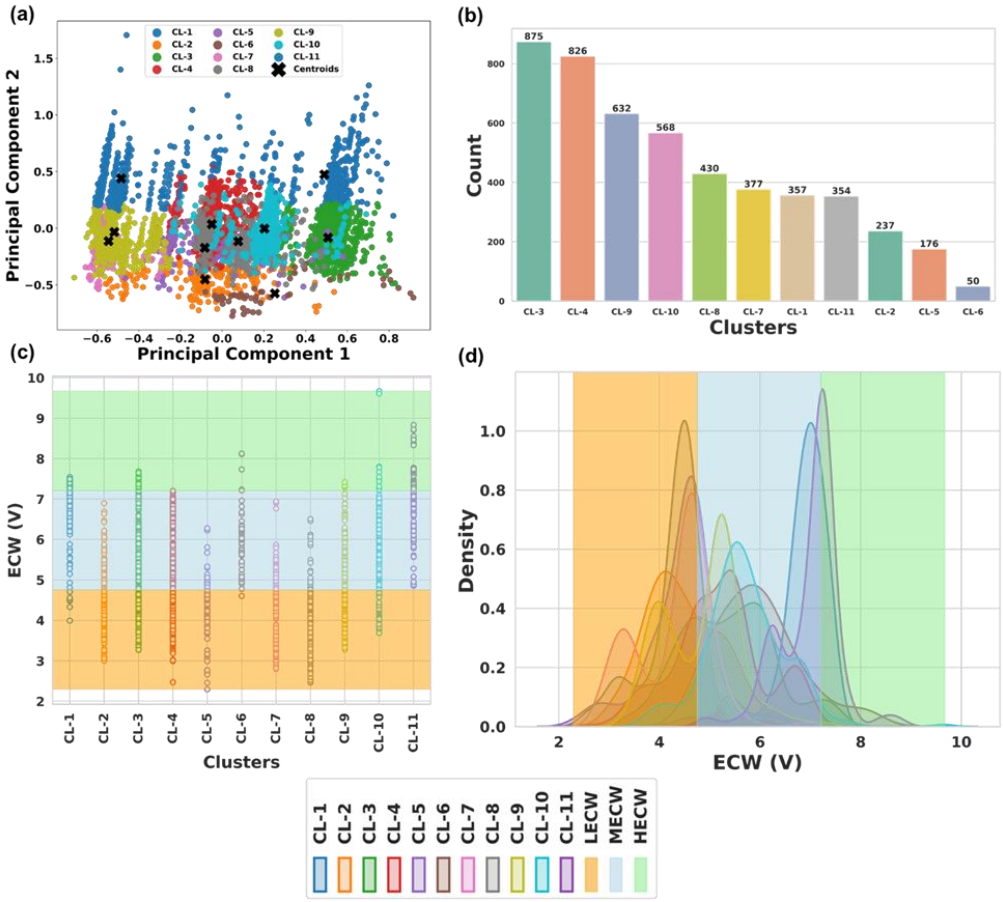


Figure 5.14: (a) Clustering of all unknown solvents with respect to PC1 and PC2, where the different colour ball and black cross represent each cluster and centroids of the optimum 11 clusters, respectively. (b) Bar plot showing number of solvents belongs to each cluster, (c) distribution plot of the optimized 11 clusters with respect to ECW, and (d) density plot of each cluster residing on various ECW range. The shaded regions of orange, blue, and green represents the LECW, MECW, and HECW, respectively.

In our in-depth analysis of cluster data, we initiated by quantifying the distribution of solvent ECW within each cluster depicted in **Figure 5.14b**. The ensuing bar plot highlights noteworthy trends, with clusters 3 and 4 demonstrating the highest solvent counts, while cluster 6 comprises of least solvents. For a more nuanced understanding, we delved into the ECW distribution within each cluster, employing distribution and density plots **(Figure 5.14c-d)**. To discern the distinct categories within each cluster, we

classified solvent electrolytes based on their ECW values. The categorization involved dividing the ECW range into three segments as mentioned above, LECW, MECW, and HECW. The resulting visualizations in **Figures 5.14c-d** illustrate these categorized regions through shaded areas, offering a comprehensive view of ECW distribution across the solvent clusters. The distribution plot underscores a notable pattern wherein solvents with high ECW (HECW) predominantly associate with cluster 1, 3, 10 and cluster 11, with a very smaller representation in cluster 6. The frequency of solvent electrolytes with different categories has been depicted through bar plot in **Figure 5.15**. It has been observed that there are only four clusters (CL-1, CL-3, CL-10, and CL-11) contains solvent electrolytes (77, 40, 14, and 152) with HECW whereas a large number of solvent molecules belongs to the MECW, and a moderately high number of solvent electrolytes present in the LECW across all the clusters.

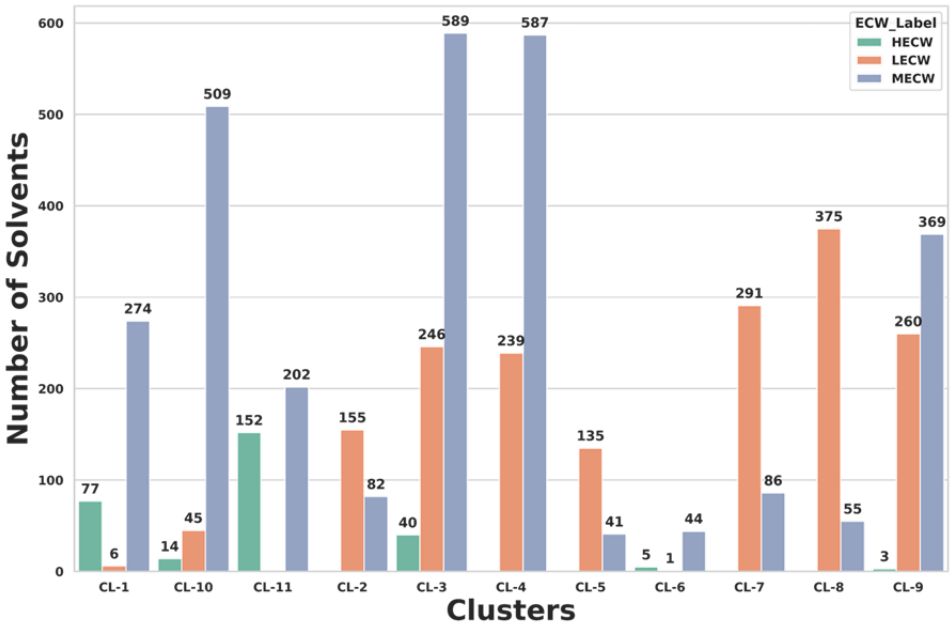


Figure 5.15: Bar plot showing the frequency of solvent electrolytes belonging to HECW, LECW, and MECW for each cluster.

We observed that solvents containing functional groups with oxygen (O) and fluorine (F) tend to exhibit high electrochemical windows (HECW). To

further analyses this phenomenon, we examined how various molecular features influence HECW values. Our analysis identified MaxPartialCharge (the maximum partial atomic charge within the molecule) and MinPartialCharge (the minimum partial atomic charge within the molecule) as critical determinants of HECW values. Specifically, ECW values peak at a certain MaxPartialCharge (~0.06), beyond which the ECW values decrease. In contrast, more negative MinPartialCharge values are associated with higher ECW, indicating that solvents with highly electronegative atoms or groups exhibit greater stabilization of anodic limiting potential energy levels, leading to higher ECW values. Additionally, structural analysis of solvents with HECW, particularly those in Cluster 11, revealed that most HECW solvents contain highly electronegative atoms (such as O, F, NO_2, and CN groups) or conjugated double bond systems. These structural features are likely to stabilize the solvents more effectively, resulting in a larger energy gap between anodic and cathodic limiting potentials and consequently higher ECW values. Cluster 6, the smallest cluster, comprises 50 solvent electrolytes. Within this cluster, 44 solvents are classified as MECW, 5 as HECW, and 1 as LECW. This distribution demonstrates a notable predominance of MECW solvents. To understand the grouping of these solvents, we analysed their oxidation and reduction potentials and compared them with those in Cluster 11. The average reduction potential in Cluster 6 is 1.03 V, and the average oxidation potential is 7.00 V. In contrast, Cluster 11 exhibits an average reduction potential of 0.13 V and an oxidation potential of 7.12 V. The distribution of these potentials across different ECW classes for these two clusters is illustrated through box plots **(Figure 5.16)**. The box plots reveal that, although the average oxidation potential is similar between the clusters, the higher average reduction potential in Cluster 6 leads to fewer HECW solvents. This indicates that Cluster 6 effectively groups solvents with similar electrochemical characteristics. Further structural analysis of Cluster 6 reveals that among the 5 HECW solvents, 4 contain electron-

withdrawing groups (e.g., CO, CN). This analysis aligns well with the classification of different classes of solvent electrolytes, suggesting that the solvents in different cluster share common electrochemical properties.

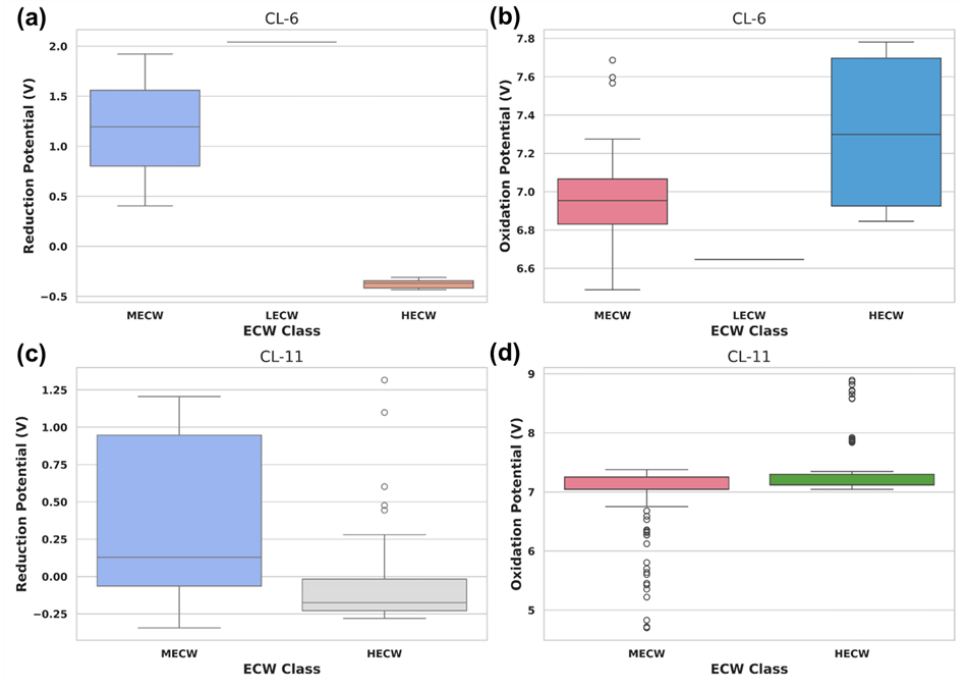


Figure 5.16: Box plot showing the variation of reduction and oxidation potential with respect to different solvent electrolyte classes (MECW, LECW, HECW). (a) Reduction potential of CL-6, (b) oxidation potential of CL-6, (c) reduction potential of CL-11, and (d) oxidation potential of CL-11.

Intriguingly, the majority of clusters exhibit a dual presence of both low ECW (LECW) and moderate ECW (MECW) ranges, implying that even solvents with apparently similar characteristics undergo variations in ECW due to subtle structural or compositional differences. This observation suggest that the clustering method effectively discerns solvents with HECW from those with LECW and MECW, it struggles to establish a definitive boundary between solvents ranging within the LECW and MECW categories. This inherent challenge is graphically represented in the scatter

plot **(Figure 5.14a)**, where overlapping clusters blur the distinction between LECW and MECW regions. Thus, we have provided an efficient method where ML has been used as both regression and clustering tool to explore a vast solvent space for better battery design.

5.6. Conclusion

This study endeavours to present a robust ML methodology, incorporating both supervised and unsupervised approaches, for the efficient determination of the electrochemical windows (ECWs) of solvent electrolytes. The primary objective is to ascertain the suitability of a solvent as an electrolyte in conjunction with high-voltage electrode materials. A systematic multistep roadmap is proposed, aiming to achieve rapid and accurate predictions of oxidation and reduction potentials through the synergistic utilization of supervised and unsupervised ML techniques. Following various ML algorithms, employing diverse cross-validation methods, and subsequent feature selection and hyperparameter tuning, we advocated an optimized XGBR model which demonstrates a high level of accuracy in predicting both oxidation and reduction potentials based on distinct feature sets. The exclusion of irrelevant features from the input set is achieved through ANOVA-F value analysis (Select-K-Best) and model-dependent feature importance. Utilizing the optimized ML models, an extensive solvent space is thoroughly explored. Beyond addressing the regression problem, a clustering method is implemented to categorize solvent molecules into smaller, more manageable subgroups or clusters within the expansive solvent space. Our findings reveal that 11 optimal clusters effectively represent the large solvent space. The overlap observed among different clusters underscores the inherent complexity within the solvent space. Each cluster is further classified into three categories based on boundary conditions, elucidating the specific types of solvent molecules within, characterized by lower, moderate, or higher ECWs. Overall, this study presents an efficacious methodology for condensing a vast solvent space, thereby facilitating the identification of optimal solvent electrolytes

capable of competing with high-voltage electrode materials in rechargeable batteries. Solvents with O and F atoms containing functional groups are found to show HECW belongs to those four clusters. This work not only showcases the remarkable predictive capabilities of the XGBR model but also underscores the efficiency of combining supervised and unsupervised approaches. The insights gained from this study hold significant implications for the development of novel solvent electrolytes, accelerating their application in batteries and contributing to advancements in energy storage technologies.

5.7. References

(1) Zheng Q., Yamada Y., Shang R., Ko S., Lee Y. Y., Kim K., Nakamura E., Yamada A. (2020), A cyclic phosphate-based battery electrolyte for high voltage and safe operation, Nat. Energy, 5 (4), 291–298 (DOI: 10.1038/s41560-020-0589-6)

(2) Van Noorden R. (2014), The rechargeable revolution: a better battery, Nature, 507 (7490), 26–28 (DOI: 10.1038/507026a)

(3) Dunn B., Kamath H., Tarascon J. M. (2011), Electrical energy storage for the grid: a battery of choices, Science, 334 (6058), 928–935 (DOI: 10.1126/science.1212741)

(4) Thackeray M. M., Wolverton C., Isaacs E. D. (2012), Electrical energy storage for transportation—approaching the limits of, and going beyond, lithium-ion batteries, Energy Environ. Sci., 5 (7), 7854–7863 (DOI: 10.1039/c1ee02717j)

(5) Shi S., Gao J., Liu Y., Zhao Y., Wu Q., Ju W., Ouyang C., Xiao R. (2016), Multi-scale computation methods: their applications in lithium-ion battery research and development, Chin. Phys. B, 25 (1), 018212 (DOI: 10.1088/1674-1056/25/1/018212)

(6) Manna S. S., Manna S., Pathak B. (2023), Molecular dynamics–machine learning approaches for the accurate prediction of electrochemical windows

of ionic liquid electrolytes for dual-ion batteries, J. Mater. Chem. A, 11, 21702–21712 (DOI: 10.1039/d3ta01228c)

(7) Xu K. (2004), Nonaqueous liquid electrolytes for lithium-based rechargeable batteries, Chem. Rev., 104 (10), 4303–4417 (DOI: 10.1021/cr030203g)

(8) Goodenough J. B., Kim Y. (2010), Challenges for rechargeable Li batteries, Chem. Mater., 22, 587–603 (DOI: 10.1021/cm901452z)

(9) Zhang S., Ma J., Hu Z., Cui G., Chen L. (2019), Identifying and addressing critical challenges of high-voltage layered ternary oxide cathode materials, Chem. Mater., 31 (16), 6033–6065 (DOI: 10.1021/acs.chemmater.9b00664)

(10) Wang D., He T., Wang A., Guo K., Avdeev M., Ouyang C., Chen L., Shi S. (2023), A thermodynamic cycle-based electrochemical windows database of 308 electrolyte solvents for rechargeable batteries, Adv. Funct. Mater., 33, 2212342 (DOI: 10.1002/adfm.202212342)

(11) Peljo P., Girault H. H. (2018), Electrochemical potential window of battery electrolytes: the HOMO–LUMO misconception, Energy Environ. Sci., 11 (9), 2306–2309 (DOI: 10.1039/c8ee01122c)

(12) Peljo P., Girault H. H. (2018), Electrochemical potential window of battery electrolytes: the HOMO–LUMO misconception, Energy Environ. Sci., 11 (9), 2306–2309 (DOI: 10.1039/c8ee01122c)

(13) Goodenough J. B., Park K. S. (2013), The Li-ion rechargeable battery: a perspective, J. Am. Chem. Soc., 135 (4), 1167–1176 (DOI: 10.1021/ja3091438)

(14) Wu K., Luo W., Xu B. (2023), An intercalation-type Li-free cathode with energy density exceeding 550 Wh kg^{-1}, Natl. Sci. Rev., 10, nwad032 (DOI: 10.1093/nsr/nwad032)

(15) Roohi H., Salehi R. (2020), Exploring the electrochemical windows of triazolium-based [PhMTZ][X1–7] ionic liquids (ILs) at MP2/Aug-cc-pVDZ level of theory by using thermochemical cycle in IL media, J. Electroanal. Chem., 877, 114606 (DOI: 10.1016/j.jelechem.2020.114606)

(16) Korth M. (2014), Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: evaluation of electronic structure theory methods, Phys. Chem. Chem. Phys., 16, 7919–7926 (DOI: 10.1039/c3cp55459g)

(17) Pan J. H., Chou Y. M., Chiu H. L., Wang B. C. (2007), Theoretical investigations of the molecular conformation and reorganization energies in the organic diamines as hole-transporting materials, J. Phys. Org. Chem., 20 (10), 743–753 (DOI: 10.1002/poc.1311)

(18) Hutchison G. R., Ratner M. A., Marks T. J. (2005), Hopping transport in conductive heterocyclic oligomers: reorganization energies and substituent effects, J. Am. Chem. Soc., 127 (7), 2339–2350 (DOI: 10.1021/ja045707l)

(19) Wang A., Zou Z., Wang D., Liu Y., Li Y., Wu J., Avdeev M., Shi S. (2021), Identifying chemical factors affecting reaction kinetics in Li–air battery via ab initio calculations and machine learning, Energy Storage Mater., 35, 595–601 (DOI: 10.1016/j.ensm.2021.01.014)

(20) Wang J., Wang Y. (2024), Strategies to improve the quantum computation accuracy for electrochemical windows of ionic liquids, J. Phys. Chem. B, 128 (8), 1943–1952 (DOI: 10.1021/acs.jpcb.3c08267)

(21) Du G., Nuli Y., Yang J., Wang J. (2008), Fluorine-doped $LiNi_{0.5}Mn_{1.5}O_4$ for 5 V cathode materials of lithium-ion battery, Mater. Res. Bull., 43, 3607–3613 (DOI: 10.1016/j.materresbull.2008.04.022)

(22) Wang F., Yang J., Nuli Y., Wang J. (2010), Highly promoted electrochemical performance of 5 V $LiCoPO_4$ cathode material by addition

of vanadium, J. Power Sources, 195, 6884–6887 (DOI: 10.1016/j.jpowsour.2009.12.119)

(23) Amine K., Yasuda H., Yamachi M. (2000), Olivine $LiCoPO_4$ as 4.8 V electrode material for lithium batteries, Electrochem. Solid State Lett., 3 (4), 178–179 (DOI: 10.1149/1.1393451)

(24) Xiao Y., Shi X., Zheng T., Yue Y., Shi S., Cheng Y. J., Xia Y. (2023), Dual role of bis(borate) additive in electrode/electrolyte interface layer construction for high-voltage NCM 523 cathode, ACS Appl. Energy Mater., 6 (9), 4817–4824 (DOI: 10.1021/acsaem.3c00466)

(25) Zafari M., Nissimagoudar A. S., Umer M., Lee G., Kim K. S. (2020), First principles and machine learning based superior catalytic activities and selectivities for N_2 reduction in MBenes, defective 2D materials and 2D π-conjugated polymer-supported single atom catalysts, J. Mater. Chem. A, 8, 5209–5216 (DOI: 10.1039/c9ta12639c)

(26) Kim C., Pilania G., Ramprasad R. (2016), Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX_3 perovskites, J. Phys. Chem. C, 120, 14575–14580 (DOI: 10.1021/acs.jpcc.6b03486)

(27) Zafari M., Kumar D., Umer M., Kim K. S. (2020), Machine learning-based high throughput screening for nitrogen fixation on boron-doped single atom catalysts, J. Mater. Chem. A, 8, 5209–5216 (DOI: 10.1039/d0ta00873d)

(28) Umer M., Umer S., Zafari M., Ha M., Anand R., Hajibabaei A., Abbas A., Lee G., Kim K. S. (2022), Machine learning assisted high-throughput screening of transition metal single atom based superb hydrogen evolution electrocatalysts, J. Mater. Chem. A, 10, 6679–6689 (DOI: 10.1039/d1ta10316e)

(29) Manna S., Roy D., Das S., Pathak B. (2022), Capacity prediction of K-ion batteries: a machine learning based approach for high throughput

screening of electrode materials, Mater. Adv., 3, 7833–7845 (DOI: 10.1039/d2ma00746k)

(30) Manna S., Manna S. S., Das S., Pathak B. (2023), Metal-solvent interaction contribution on voltage for metal ion battery: an interpretable machine learning approach, Electrochim. Acta, 467, 143148 (DOI: 10.1016/j.electacta.2023.143148)

(31) Wang A., Zou Z., Wang D., Liu Y., Li Y., Wu J., Avdeev M., Shi S. (2021), Identifying chemical factors affecting reaction kinetics in Li–air battery via ab initio calculations and machine learning, Energy Storage Mater., 35, 595–601 (DOI: 10.1016/j.ensm.2021.01.014)

(32) Das S., Manna S., Pathak B. (2023), Unlocking the potential of dual-ion batteries: identifying polycyclic aromatic hydrocarbon cathodes and intercalating salt combinations through machine learning, ACS Appl. Mater. Interfaces, 15, 54529–54541 (DOI: 10.1021/acsami.3c14063)

(33) Ha M., Hajibabaei A., Kim D. Y., Singh A. N., Yun J., Myung C. W., Kim K. S. (2022), Al-doping driven suppression of capacity and voltage fadings in 4d-element containing Li-ion-battery cathode materials: machine learning and density functional theory, Adv. Energy Mater., 12 (30), 2201034 (DOI: 10.1002/aenm.202201034)

(34) Manna S., Das A., Das S., Pathak B. (2024), Machine learning assisted screening of MXene with superior anchoring effect in Al–S batteries, ACS Mater. Lett., 6 (2), 572–582 (DOI: 10.1021/acsmaterialslett.3c01043)

(35) Manna S. S., Pathak B. (2023), Screening of ionic liquid-based electrolytes for Al dual-ion batteries: thermodynamic cycle and combined MD-DFT approaches, J. Phys. Chem. C, 127, 8924–8933 (DOI: 10.1021/acs.jpcc.2c08004)

(36) Manna S. S., Pathak B. (2023), Machine learning-driven ionic liquids as electrolytes for the advancement of high-voltage dual-ion battery, J. Phys. Chem. C, 127, 8924–8933 (DOI: 10.1021/acs.jpcc.2c08004)

(37) Bento A. P., Hersey A., Félix E., Landrum G., Gaulton A., Atkinson F., Bellis L. J., De Veij M., Leach A. R. (2020), An open source chemical structure curation pipeline using RDKit, J. Cheminform., 12, 51 (DOI: 10.1186/s13321-020-00456-1)

(38) Schneider N., Sayle R. A., Landrum G. A. (2015), Get your atoms in order—an open-source implementation of a novel and robust molecular canonicalization algorithm, J. Chem. Inf. Model., 55 (10), 2111–2120 (DOI: 10.1021/acs.jcim.5b00543)

(39) Capecchi A., Probst D., Reymond J. L. (2020), One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome, J. Cheminform., 12 (1), 1–15 (DOI: 10.1186/s13321-020-00445-4)

(40) Weininger D. (1988), SMILES, a chemical language and information system: 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci., 28 (1), 31–36 (DOI: 10.1021/ci00057a005)

(41) Weininger D., Weininger A., Weininger J. L. (1989), SMILES. 2. Algorithm for generation of unique SMILES notation, J. Chem. Inf. Comput. Sci., 29 (2), 97–101 (DOI: 10.1021/ci00062a008)

(42) rdkit.ML.Descriptors.MoleculeDescriptors module — The RDKit 2024.03.3documentation.rdkit.org/docs/source/rdkit.ML.Descriptors.MoleculeDescriptors.html (accessed 2024-06-02)

(43) Liu Y., Zou X., Ma S., Avdeev M., Shi S. (2022), Feature selection method reducing correlations among features by embedding domain knowledge, Acta Mater., 238, 118195 (DOI: 10.1016/j.actamat.2022.118195)

(44) Deng Q., Lin B. (2021), Exploring structure–composition relationships of cubic perovskite oxides via extreme feature engineering and automated machine learning, Mater. Today Commun., 28, 102590 (DOI: 10.1016/j.mtcomm.2021.102590)

(45) Pedregosa F., Michel V., Grisel O., Blondel M., Prettenhofer P., Weiss R., Vanderplas J., Cournapeau D., Varoquaux G., Gramfort A., Thirion B., Dubourg V., Passos A., Brucher M., Perrot M., Duchesnay É. (2011), Scikit-learn: machine learning in Python, J. Mach. Learn. Res., 12, 2825–2830 (https://www.jmlr.org/papers/v12/pedregosa11a.html)

(46) Chen T., Guestrin C. (2016), XGBoost: a scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 785–794 (DOI: 10.1145/2939672.2939785)

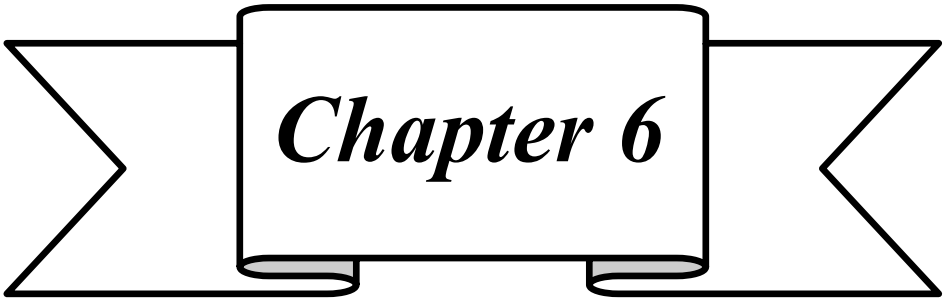
(47) Harten, P.; Martin, T.; Gonzalez, M.; Young, D. The Software Tool to Find Greener Solvent Replacements, PARIS III. Environ. Prog. Sustain. Energy 2020, 39 (1), 13331.

(48) Liu, Y.; Liu, Z.; Li, S.; Guo, Y.; Liu, Q.; Wang, G. Cloud-Cluster: An Uncertainty Clustering Algorithm Based on Cloud Model. Knowledge-Based Systems 2023, 263, 110261.

(49) Singh SIST-DIT, N.; Bhimrao, B.; Agrawal, A.; A Khan SIST-DIT, yahoocoin R. Gaussian Mixture Model: A Modeling Technique for Speaker Recognition and Its Component. Int. J. Comput. Appl. 2014, 975, 8887.

(50) Garikapati, P.; Balamurugan, & K.; Latchoumi, T. P.; Malkapuram, R. A Cluster-Profile Comparative Study on Machining AlSi 7 /63% of SiC Hybrid Composite Using Agglomerative Hierarchical Clustering and K-Means. Silicon 2021, 13, 961–972.

(51) Cohn, R.; Holm, E. Unsupervised Machine Learning Via Transfer Learning and K-Means Clustering to Classify Materials Image Data. Integr. Mater. Manuf. Innov. 2021, 10 (2), 231–244.



Chapter 6

Screening of MXene with Superior Anchoring Effect in Al–S Batteries

6.1. Introduction

Metal−sulfur batteries (Li−S, Na−S, and Al−S) have emerged as a prospective post lithium-ion battery energy storage technology due to abundance of sulfur, high energy density, and theoretical capacity.*[1-10]* Among the various available metal−sulfur batteries, Al−S batteries have garnered significant attention due to the attractive features of Al such as abundance, safety, high volumetric, and gravimetric capacity.*[7,8, 10-11]* However, the advancement of Al−S batteries faces roadblocks stemming from capacity decay over time, primarily caused by the formation of insulating polysulfides (e.g., Al_2S_{18}, Al_2S_{12}, Al_2S_6, and Al_2S_3) that are not reversibly transformed back into Al and S during discharge as they get diffused into the electrolyte.*[12, 13]* These deposits hinder the reversibility of sulfur reduction and contribute to the notorious "shuttle effect", thereby leading to the loss of electrical contact.*[14-16]* To surmount these obstacles and enhance the performance of Al−S batteries, an ingenious approach involves designing a sulfur host cathode with an exceptional ability to anchor Al_2S_n to bolster sulfur reversibility during charging/discharging cycles.

In this regard, MXenes have garnered considerable interest as electrodes in energy storage devices.*[17-20]* These classes of materials are known for their exceptional metal-like conductivity, rapid charge transfer capabilities, and remarkable surface charge accumulation.*[21-25]* These unique properties make MXenes highly promising for applications in catalysis and energy storage, where they can facilitate rapid charge transfer and ensure excellent electrode conductivity. Doped MXenes are reported as highly efficient bifunctional and multifunctional catalyst for water splitting as well as for metal−air batteries.*[26]* A distinct mechanistic pathway for water splitting using late transition metal doped MXene has also been reported.*[27]* Notably, properties of MXenes can be tuned according to requirements by using various terminating groups (O, OH, H, F, Cl, Br, NCO, SCN, NCS, etc.) that can be easily prepared by various etching

methods.*[28-31]* A number of studies have been reported on the utilization of MXene in Li−S batteries.*[24, 32-38]* Also, functionalized MXenes have been reported to have remarkable ability to anchor lithium polysulfides and effectively suppress the shuttle effect.*[39, 40]*

Although Al−S batteries share a working mechanism similar to that of Li−S batteries, research on designing and screening sulfur hosts for Al−S batteries is limited. Thus, it is essential to draw inspiration from the advancements in Li−S battery development to expedite progress in the field of Al−S battery research. The primary evaluation criterion for sulfur hosts is their nature of interaction with various possible polysulfide intermediates. However, the range of sulfur hosts tested for Al−S battery cathodes has been relatively limited. For instance, Zheng et al. proposed oxygen functionalized MXene ($Ti_3C_2O_2$) as a cathode material for Al−S batteries showing good anchoring capability.*[9]* Further, single atom doped $Ti_3C_2O_2$ has also been reported as suitable sulfur host cathode materials.*[41]* The number of possible MXenes is on the order of thousands considering different metal layers and terminal functional groups. Designing and testing of all of these MXenes for a particular application collectively exceeds the capabilities of traditional experimental or theoretical limits. In this regard, machine learning (ML) empowers us to discover novel patterns, make insightful predictions, and accelerate the discovery of materials with exceptional accuracies.*[42-44]* Kim and co-workers have resolved the capacity fading issues of unstable electrode materials by utilizing machine learning potential.*[45]* Significant progress has been reported for the inverse design of materials, direct property prediction and utilization of ML potential with the advance ML techniques.*[46]* With each iteration, our data-driven models become more refined, enhancing our ability to design MXenes tailored for specific applications. This synergistic combination of ML and materials design revolutionizes the process, unlocking unprecedented opportunities for the development of application specific high-performance MXene-based systems.

This work endeavors to address critical challenges in the development of high-performance aluminum-sulfur (Al-S) batteries by introducing anchoring materials to hinder the diffusion of Al polysulfide (Al_2S_n) intermediates into the electrolyte. As the adsorption energy is the deciding factor for the determination of the anchoring capability of the MXene materials, we have chosen that as the target variable. Therefore, the synergistic integration of density functional theory (DFT) calculations, data analysis, and ML algorithms has been implemented to propose potential cathode materials with strong anchoring, utilizing adsorption energies as the suitable descriptor. An optimum adsorption energy criterion has been set to identify MXenes capable of anchoring all the intermediates. This innovative DFT/ML-based multistep workflow not only unveils physically meaningful predictions but also paves the way for accelerated screening and rational design of MXenes as cathode host materials for Al−S batteries. The workflow adopted in this study is depicted in **Figure 6.1**.

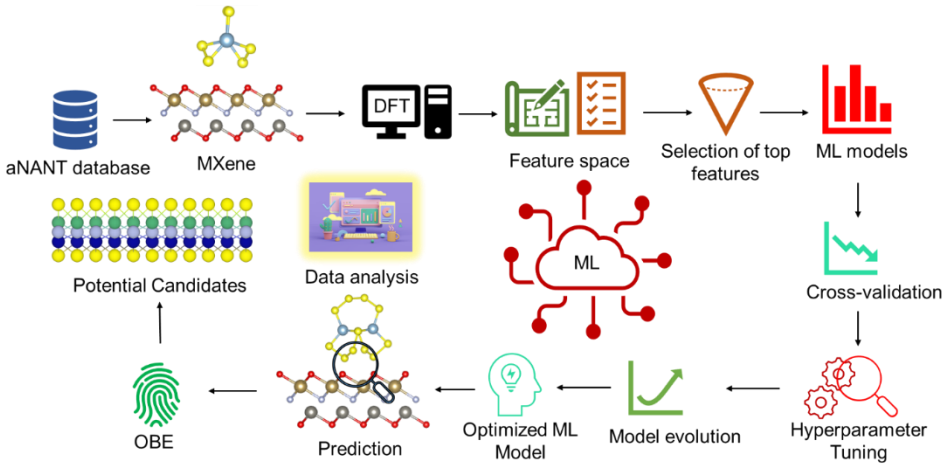


Figure 6.1: General workflow for the discovery of potential MXenes as sulfur host cathode materials using DFT+ML approach.

6.2. Data Generation

The possible MXene structures have been extracted from the aNANt, a functional material database along with their electronic properties (band gap and lattice constant).*[47, 48]* MXenes having the same terminal group on

both sides have been screened as a different terminal group on the other side is not expected to significantly affect the interaction in the intermediate adsorption side. Thus, 1846 different MXenes have been considered for the adsorption of polysulfides (S_8, Al_2S_3, Al_2S_6, Al_2S_{12}, Al_2S_{18}). The adsorption energies upon the adsorption of different polysulfides has been generated through DFT calculation for a fraction of the data set. Subsequently, construction of feature space followed by feature engineering was carried out to prepare a well-sampled adsorption energy database. Then various ML algorithms are trained, and the optimum ML model is finalized considering the mean absolute error (MAE) of test data as the performance evolution matrix. Further, the SelectKBest algorithm is applied on the overall data set to reduce the feature space without any compromise on the accuracy of the ML model. To improve the accuracy of the ML model and check the stability of the ML predicted result, hyperparameter tuning and different cross-validation (CV) methods have been performed for the selected ML models followed by the prediction of adsorption energies of all the MXenes for all of the intermediates. Further, based on the optimum adsorption energy criterion, potential MXene materials were screened out which could be best suitable as sulfur host cathode material.

6.3. Computational Details

The density functional theory-based calculations were conducted using the Vienna Ab initio Simulation Package (VASP). The Perdew−Burke−Ernzerhof generalized gradient approximation (GGA) functional was utilized to incorporate exchange−correlation effects, while the projected augmented wave (PAW) method was employed to account for ion-electron interactions.*[49-55]* For the structural relaxation, convergence thresholds of 10^{-2} eV $Å^{-1}$ force, and 10^{-4} eV energy were employed. The electronic wave functions were expanded up to 470 eV energy cutoff, and a Γ k-point grid of 2 x 2 x 1 has been considered to sample the Brillouin zone. To avoid periodic interactions, a vacuum of ∼10 Å was given along the z direction. To account for van der Waals interactions, Grimme's DFT-D3

has been considered.*[56]* The cell reactions involved in nonaqueous Al–S batteries can be given by eqs. 1 and 2.*[13,57,58]*

$$\text{Anode: } 2Al + 14AlCl_4^- \rightleftarrows 8Al_2Cl_7^- + 6e^- \quad (1)$$

$$\text{Cathode: } 8Al_2Cl_7^- + 6e^- + 3S \rightleftarrows Al_2S_3 + 14AlCl_4^- \quad (2)$$

The $AlCl_4^-$ and $Al_2Cl_7^-$ are generated in the electrolyte from the mixture of 1-ethyl 3-methyl imidazolium chloride and $AlCl_3$. The conversion of sulphur to Al_2S_3 goes through the formation of various Al-S intermediates.*[9, 59]* Thus, the charging process results in sequential reduction from S_8 to Al_2S_{18} to Al_2S_{12} to Al_2S_6 to Al_2S_3. The reverse happens during the discharging process. Initially we performed the structural optimization of the MXene configurations and polysulfides followed by their adsorptions.

The S_8 molecules adopt a crownlike orthorhombic structure, while Al_2S_3 exhibits a planar bent structure. As the number of sulfur atom increases, Al_2S_n structures tend to adopt 3D cluster shapes without overhanging bonds, showing excellent geometric stability. The geometrical structures of aluminum polysulfides and a general representation of considered terminal groups, metals and X layers of the MXene structure have been depicted in **Figure 6.2**.

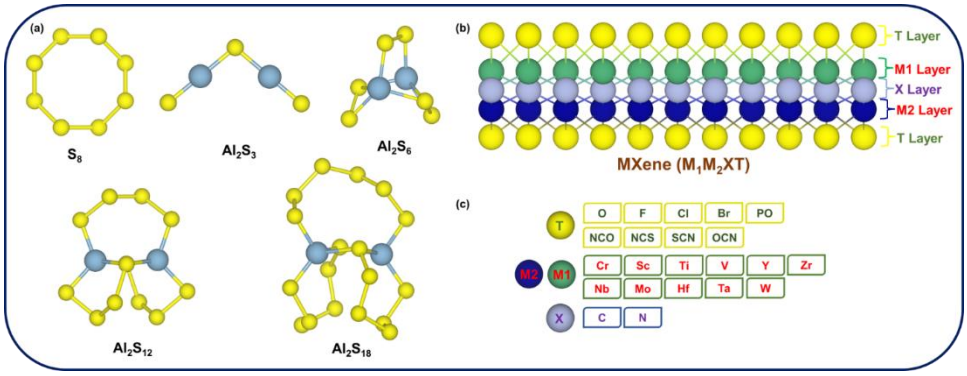


Figure 6.2: (a) Optimized geometries of considered polysulfide intermediates, (b) general representation of MXene structure, and (c) constituent elements and functional groups for MXenes.

The adsorption energy (E_{ads}) of various polysulfide intermediates can be calculated using,

$$E_{ads} = E_{MX-polysulfide} - E_{MX} - E_{polysulfide} \quad (5)$$

Where, $E_{MX-polysulfide}$, E_{MX}, and $E_{polysulfide}$ are the total energies of polysulfide intermediates adsorbed MXene, bare MXene and polysulfides, respectively.

6.4. Results and Discussion

6.4.1. Feature Space

It is imperative to establish a clear connection between materials and their relevant attributes through various features for the ML models to be accurately trained. Significant advancements in machine learning models have been achieved by integrating elemental and geometrical descriptors using column matrices and SOAP, which exhibit invariance with respect to translation, rotation, and permutation.*[60−63]* This necessitates the development of a set of features, representing material variables as well as different polysulfides that can be efficiently processed by computational methods and capture the physicochemical properties relevant to the adsorption process. In our quest to accurately distinguish each MXene material, we embarked on a comprehensive exploration of only elemental, structural, and electronic features rather than expensive DFT calculated features. These features play a crucial role in understanding the distinctive properties of MXenes and their interactions with polysulfides. In total, 105 numerical features have been considered as tabulated in **Tables 6.1−6.4**. The atomistic features that we considered delve into the unique properties of individual atoms within each layer of MXenes. Meanwhile, the structural features capture the overarching structural characteristics, encompassing vital information such as lattice constant, surface area, distances between adjacent atoms, and so on.

Table 6.1: List of elemental features considered for the prediction of adsorption energy.

Indicator	Elemental Features
WM1, WM2, WX, WT	Atomic mass
ANM1, ANM2, ANX, ANT	Atomic number
PM1, PM2, PX, PT	Period number
GM1, GM2, GX, GT	Group number
VM1, VM2, VX, VT	Valence electron
IEM1, IEM2, IEX, IET, IPTA	First ionization potential
M1OXs, M2OXs	Oxidation state
RM1, RM2, RX, RT	Atomic radius
MPM1, MPM2, MPX, MPT	Melting point
BPM1, BPM2, BPX, BPT	Boiling point
CEM1, CEM2, CEX, CET	Cohesive energy
χM1, χM2, χX, χT	Pauling electronegativity
X	C/N
T	Terminal group/atom type
TA	Terminal atom of terminal group

Table 6.2: List of structural features considered for the prediction of adsorption energy.

Indicator	Structural features
K	Lattice constant
SA	Surface area
SWT	Surface weight
DT1M1, DM1X, DXM2, DM2T2	Distance between T1-M1, M1-X, X-M2, M2-T2

Table 6.3: List of electronic features considered for the prediction of adsorption energy.

Indicator	Electronic features
E_g	Bandgap of the MXene
μ	Magnetic moment of MXene
ΔIETM1, ΔIEM1X, ΔIEXM2, ΔIEM2T	First ionization potential difference between T-M1, M1-X, X-M2, M2-T
ΔχTM1, ΔχM1X, ΔχXM2, ΔχM2T	Electronegativity difference between T-M1, M1-X, X-M2, M2-T
M1GVE, M2GVE, TGVE, XGVE	sum_gilmor_number_of_valence electron of M1, M2, T, and X
M1VS, M2VS, TVS	sum_valence_s electrons of M1, M2 and T
M1VP, M2VP, TVP, XVP	sum_valence_p electrons of M1, M2, T and X
M1VD, M2VD,	sum_valence_d electrons of M1 and M2
M1VF, M2VF,	sum_valence_f electrons of M1 and M2
M1USV, M2USV,	sum_Number_of_unfilled_s_valence electrons of M1 and M2
M1UPV, M2UPV, TUPV, XUPV	sum_Number_of_unfilled_p_valence_electrons of M1, M2, T, and X
M1UDV, M2UDV, TUDV	sum_Number_of_unfilled_d_valence electrons of M1, M2 and T
M1UFV, M2UFV, TUFV	sum_Number_of_unfilled_f_valence electrons of M1, M2 and T

For a more profound understanding of the adsorption process, we meticulously incorporated a multitude of electronic properties (Bandgap, Magnetic moment, first ionization potential difference between T-M1, M1-X, X-M2, and M2-T, electronegativity difference between T-M1, M1-X, X-

M2, M2-T, etc.) known to have significant influence. To effectively distinguish the adsorbed polysulfides, we dedicated careful attention to the features pertinent to these sulfur-rich species. We harnessed various statistical techniques, including averages and deviations, to establish meaningful connections between the features of MXenes and polysulfides.

Table 6.4: List of aluminum polysulfide features considered for the prediction of adsorption energy.

Indicator	**Aluminum polysulfide features**
Al-Fraction, S-Fraction	Fraction of aluminum and sulfur
Num_Al, Num_S	Number of Al and S
χAl_2S_n	Average electronegativity of polysulfide
$\Delta\chi Al_2S_nT$, $\Delta\chi Al_2S_nM1$	Electronegativity difference between Al_2S_n-T and Al_2S_n-M1
$IEAl_2S_n$	Average first ionization of Al_2S_n
$\Delta IEAl_2S_nT$, $\Delta IEAl_2S_nM1$, $\Delta IEAl_2S_nTA$	First ionization difference between Al_2S_n-T, Al_2S_n-M1, and Al_2S_n-TA
Al_2S_nGVE, Al_2S_nVP, Al_2S_nUPV	Valence electron, valence P electron and unfilled valence P electron of Al_2S_n

6.4.2. Machine Learning

In our study, we developed a robust machine learning (ML) approach to establish a regression relationship between the adsorption phenomena of polysulfides and the attributes of MXenes, based on data obtained from density functional theory (DFT) calculations. To achieve this, we employed 12 different ML algorithms, namely, Lasso (LS), Partial Least Squares (PLS), Ridge Regression (RR), Kernel Ridge Regressor (KRR), Elastic Net Regressor (ENR), K-Neighbors Regressor (KNR), Support Vector

Regression (SVR), AdaBoost Regressor (ABR), Gradient Boosting Regressor (GBR), Decision Tree Regressor (DTR), Extreme Gradient Boosting Regression (XGBR), and Random Forest Regressor (RFR). Utilizing an open-source Python distribution platform and the powerful scikit-learn libraries, we trained the ML models on the DFT calculated adsorption energy-based data set consisting of ~600 well-sampled polysulfide intermediate-MXene systems. To ensure the reliability and generalization of the supervised ML models, we started with cross-validation assessment followed by feature selection using the SelectKBest algorithm, which ranks the features based on the ANOVA Fvalue through the variance analysis of the data. The higher the F-value, the higher the correlation between that feature and the target variable. Moreover, the model's performance was enhanced through hyperparameter tuning, involving the random division of adsorption energy data derived from DFT calculations into training and testing sets for evaluation. To quantify the performance of our ML models, we used the coefficient of determination (R^2) and the mean absolute error (MAE), providing valuable insights into the models' predictive capabilities. Finally, we have predicted adsorption energies of all the intermediates (5930 configurations) using the best fitted model for the whole data set of MXene materials.

In our study, we utilized a data set comprising of 23,000 MXene structures extracted from the "aNANt" functional material database.*[47,48,64]* The MXenes of this database follow a sequential structure denoted as T1-M1-X-M2-T2, where terminal groups (T1 and T2) are present at both ends of the metal layer (M1 and M2), separated by an X layer (C or N). We focused on a subset of 1,846 MXenes with similar terminal groups (T1 = T2 = T). The general formula for these MXenes is T-M1-X-M2-T as illustrated in **Figure 6.2**.

Considering the adsorption energy of intermediates as a suitable descriptor, to generate the train data, we carried out the adsorption of intermediates on

the MXene surfaces using DFT. However, among all the terminal functional groups, it has been observed that during the adsorption of the polysulfides on MXenes having T = NO, H, OH, and CN, the system becomes unstable, leading to a distorted structure of MXenes or breaking of polysulfides. The instability of the H terminal group containing MXenes could be due to the small size of the H atom, which provides enough space for the S atoms of polysulfides to interact with the metal layer strongly leading to disintegration of the polysulfide. On the other hand, for the OH group containing MXenes, the O−H bond itself is broken during the adsorption of polysulfides. In the case of NO and CN as terminal groups, the bond angles are ~180° (∠M1NO and ∠M1CN). During the adsorption of polysulfides, the bond angle changes to ~90° resulting in an overall distortion of MXene structures.

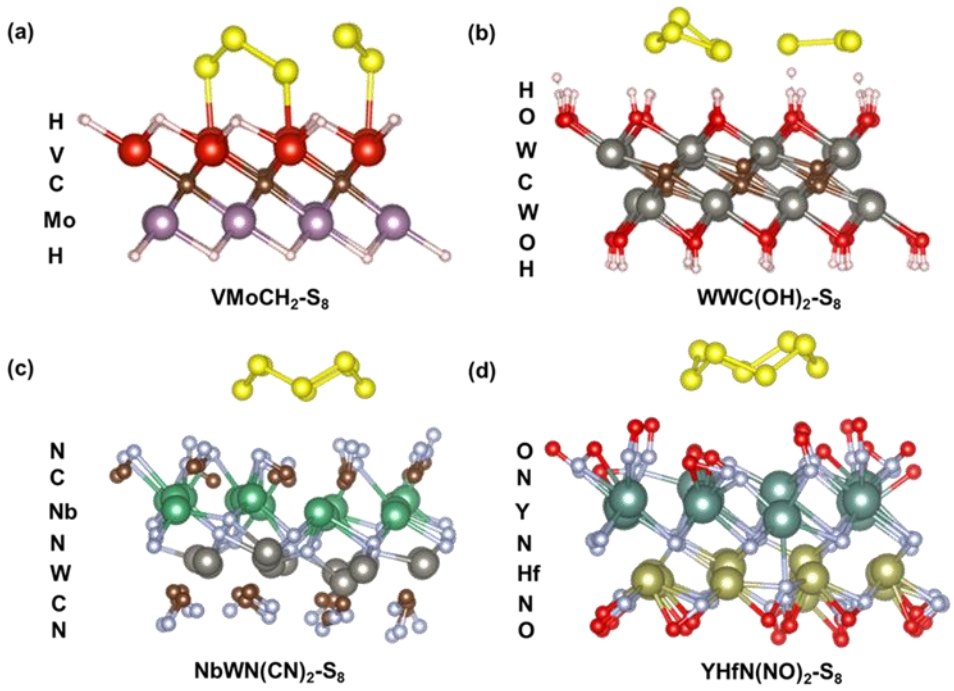


Figure 6.3: Distorted structure of MXenes upon the adsorption of S_8 having terminal groups (a) H, (b) OH, (c) CN, and (d) NO.

The variance in bond angles might stem from distinct charge distributions among the polysulfide atoms. Consequently, Al and S atoms tend to engage

with more electron-rich atoms (O in NO and N in CN) and fewer electron-rich atoms (N in NO and C in CN), respectively, on the MXenes. Examples of these excluded structures are illustrated in **Figure 6.3**. Thus, among the 13 terminal groups, we excluded these four terminal groups containing MXenes in our further study.

To create a well-sampled data set for training and testing purposes, we carefully incorporated various terminal groups and transition metals to generate ~600 data points to calculate adsorption energy of Al polysulfides. Our initial analysis involved examining the adsorption energy data separately in relation to terminal groups and X groups (C and N). It has been observed that for MXenes with O-containing terminal groups, the adsorption energy exhibited a wide range, spanning from low to high magnitudes. Conversely, very few MXenes containing OCN, and F terminal groups showcased higher adsorption energies. Nevertheless, a substantial portion of the adsorption energy data clustered within the −0.5 to −4.0 eV range. Notably, the distribution of adsorption energy concerning C and N as X groups appeared to be relatively uniform and consistent. However, for application as cathode host material in Al−S batteries, identifying MXene materials with an optimal adsorption energy becomes crucial. An excessively low adsorption energy could lead to polysulfide dissolution into the solvent. Conversely, an overly high adsorption energy might result in an irreversible electron transfer process. Thus, establishing a boundary criterion for an ideal adsorption energy is important to evaluate suitable cathode host materials. It is essential to ensure that the adsorption energy of polysulfides is higher than the solvation energy of Al−S intermediates into the EMIM+$AlCl_4$ – electrolyte to prevent the dissolution. In our previous work, solvation energies for dissolution of polysulfide into the electrolyte was found to be within the range of −1.1 to −1.5 eV from molecular dynamic simulations.*[65]* Thus, an adsorption energy of more than −1.5 eV would be essential to suppress the shuttle effect. Furthermore, during the DFT calculations in the current work, it has been observed that MXenes

with an adsorption energy greater than −5.0 eV tend to distort the MXene or polysulfide structure. Also, the too strong adsorption would make the conversion of lower order polysulfides into higher order polysulfides further irreversible as they are already reported to be insulating in nature.*[12]* Thus, MXenes having adsorption energy in the optimum range far away from both the boundaries would be the suitable choice as an optimum anchoring cathode host material for an Al−S battery. Thus, by calculating the median value between −1.8 and +5 eV and considering a range of ±0.7 eV, we set our optimum adsorption energy range to −2.8 to −4.2 eV. MXenes showing adsorption energies less than −2.8 eV and greater than −4.2 eV are categorized as having weak adsorption energy (WE_{ads}) and strong adsorption energy (SE_{ads}), respectively, whereas adsorption energies between −2.8 and −4.2 eV are categorized as optimum adsorption energies (OE_{ads}).

Table 6.5: The cross validated MAE of the ML algorithms for the prediction of adsorption energies.

ML Models	MAE (eV)		
	10-fold CV	Repeated K-fold CV	LOOCV
LS	0.84	0.83	0.84
PLS	0.84	0.84	0.84
RR	0.77	0.77	0.76
KRR	0.78	0.77	0.78
ENR	0.83	0.83	0.83
KNR	0.72	0.72	0.71
SVR	0.78	0.77	0.77
ABR	0.71	0.72	0.72
GBR	0.43	0.42	0.42
DTR	0.43	0.44	0.42
RFR	0.37	0.38	0.36
XGBR	0.36	0.35	0.35

The DFT generated adsorption energy data (~600) are processed further for the training and testing of various ML models to predict the adsorption energy for overall 5930 various configurations considering all the intermediates and MXenes. First, we applied various linear and tree-based ML algorithms to the DFT calculated data and evaluated the performance of each model in terms of mean absolute error (MAE). Capturing latent knowledge is highly challenging due to the diverse nature of MXene materials with varying components. Hence, to avoid the stochasticity of the divergent data and to ensure the stability of each model, we started with evaluation of each ML model's performance in three different cross-validation (CV) methods, namely, K-fold CV (K-fold CV, K = 10), Repeated K-fold CV (Repeated K-fold CV, Repetition = 5, K = 10), and Leave-one-out CV (LOOCV). All the CV results have been tabulated in **Table 6.5** and indicate that the ML models based on linear algorithms (LS, PLS, RR, KRR) are completely unable to map the input variables toward the target variable due to the complex nonlinear relationship between the features and target variable.

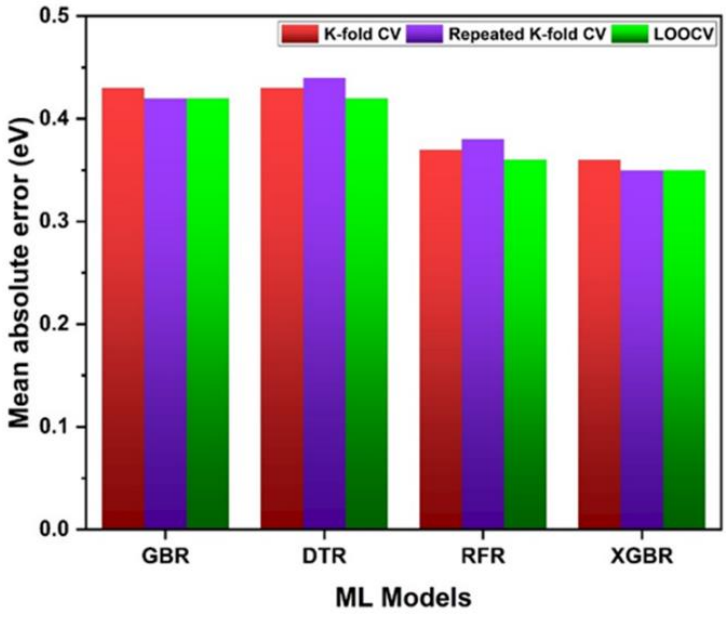


Figure 6.4: K-fold (K=10), repeated K-fold (Repetition = 5, K =10) and leave-one-out CV error bar plot for the selected ML models.

While shifting from the linear model to bagging boosting and tree-based ML models, a sharp jump in accuracy was observed for GBR, DTR, XGBR, and RFR **(Figure 6.4)**.

Moreover, the consistency in error for all three different cross-validation methods have been conserved, indicating the stability of the ML models for the prediction of the adsorption energy. Thus, from the primary ML model investigation, we have excluded all the linear model and considered only the GBR, DTR, XGBR, and RFR for further improvement of ML prediction along with 105 numerical attributes (elemental, structural, and electronic) to describe the adsorption phenomena of polysulfides on MXenes.

6.4.2. Optimization of Feature Space

In the realm of high-dimensional data analysis, feature selection plays a crucial role in identifying the most relevant predictors that significantly influence the target variable. By employing SelectKBest algorithm, a powerful and efficient method as implemented in the scikit-learn library, we reduce the feature space while retaining the essential information for accurate modeling. Our approach involved assessing each model's performance with different feature sets, beginning with the top five ranked features and progressively increasing the feature count in increments of five. This systematic process resulted in 21 distinct sets of features. The selection of features was not sequential; rather, it was based on the top ranked feature in each set. This ensured that all of the features chosen for each set were the most significant and influential attributes. Every feature set was evaluated through MAE for each ML model. The feature optimization plot, depicted in **Figure 6.5**, showcases the top-ranked features alongside their corresponding errors.

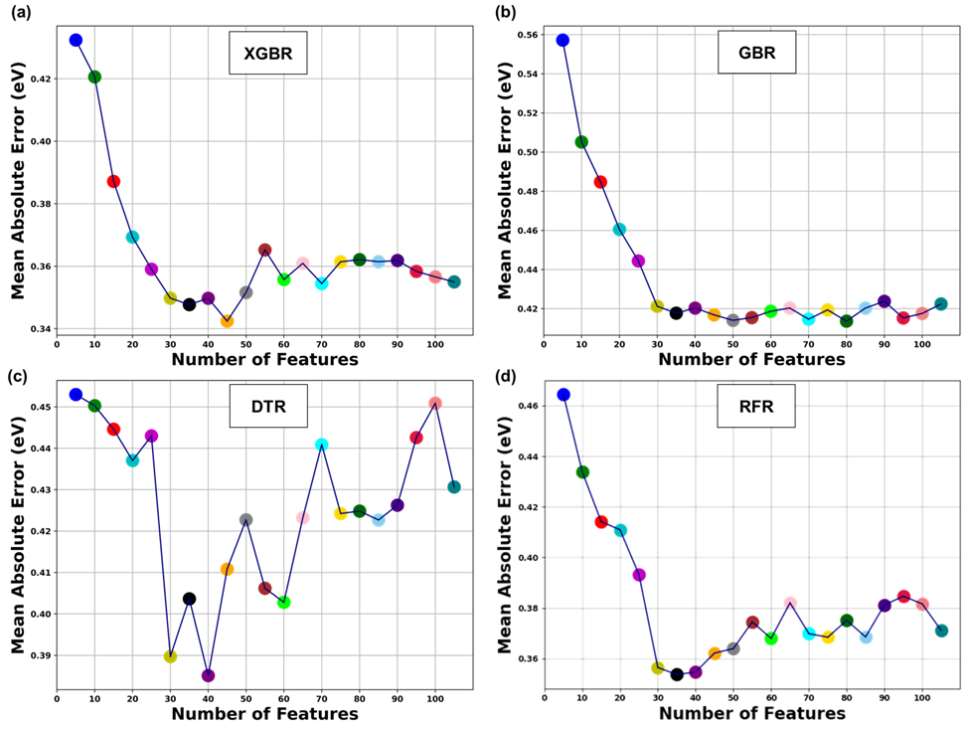


Figure 6.5: Feature optimization plots for considered (a) XGBR, (b) GBR, (c) DTR, and (d) RFR algorithms.

Each model is found to exhibit a distinct trend, with the maximum accuracy achieved at different feature counts and the error range varying continuously as we change the feature set. The GBR and DTR models have shown lowest errors when utilizing the top 80 and 40 features, respectively, while the RFR and XGBR models reached their peak performance with 35 and 45 best features, respectively **(Figure 6.5)**. These findings highlight the model-specific nature of the feature importance and the need for tailored feature selection for optimal performance. The top-ranked attributes for each model are presented in **Tables 6.6−6.9**.

However, it is worth mentioning that the DTR model exhibited some inconsistency, displaying a considerable deviation in error as we varied the number of features. We hypothesized that subsampling of feature sets across different estimators could contribute to this variability during the fitting process. Thus, using SelectKBest we have reduced the dimension

space from 105 without compromising accuracy, which ultimately helps to discard a large number of irrelevant features during the fitting and as well as reduce the computational cost for exhaustive large space hyperparameter tuning.

Table 6.6: List of selected best features using SelectKBest method for GBR model. The feature names of the following indicators are provided in Table 6.1-6.4.

Top ranked 80 features selected based on SelectKBest to fit the data on GBR model
IPTA, Eg, SA, WX, WT, SWT, ANX, ANT, PT, GM1, GM2, GX, GT, VM1, VM2, VX, VT, IEM1, IEX, IET, ΔIETM1, ΔIEM1X, ΔIEXM2, ΔIEM2T, RM1, RM2, RX, RT, MPM1, MPM2, MPX, MPT, BPX, BPT, χM1, χM2, χX, χT, ΔχTM1 , ΔχXM2 , ΔχM2T, CEMX, CEMT, DT1M1, DM1X, DXM2, DM2T2, M1GVE, M1VS, M1VD, M1OXs, M1USV, M1UDV, M2GVE, M2VS, M2VD, M2USV, M2UDV, TGVE, TVS, TVP, TUPV, TUDV, TUFV, XGVE, XVP, XUPV, Al-Fraction, S-Fraction, Num_Al, χAl_2S_n, $\Delta\chi Al_2S_nT$, $\Delta\chi Al_2S_nM1$, $IEAl_2S_n$, $\Delta IEAl_2S_nT$, $\Delta IEAl_2S_nM1$, $\Delta IEAl_2S_nTA$, Al_2S_nGVE, Al_2S_nVP, Al_2S_nUPV

Table 6.7: List of selected best features using SelectKBest method for DTR model. The feature names of the following indicators are provided in Table 6.1-6.4.

Top ranked 40 features selected based on SelectKBest to fit the data on DTR model
IPTA, SA, WX, WT, SWT, ANX, ANT, PT, GM2, GX, GT, VM2, VX, VT, IEX, ΔIETM1, ΔIEM2T, RX, RT, MPX, MPT, BPX, BPT, χM2, ΔχTM1, ΔχM2T, CEMX, DT1M1, DM2T2, M2GVE, M2VD, M2UDV, TGVE, TVS, TVP, TUDV, TUFV, XGVE, XUPV, $\Delta IEAl_2S_nTA$

Table 6.8: List of selected best features using SelectKBest method for RFR model. The feature names of the following indicators are provided in Table 6.1-6.4.

Top ranked 35 features selected based on SelectKBest to fit the data on RFR model
IPTA, SA, WT, SWT, ANX, ANT, PT, GM2, GX, GT, VM2, VX, VT, ΔIETM1, ΔIEM2T, RX, RT, MPT, BPT, χM2, ΔχTM1, ΔχM2T, CEMX, DT1M1, DM2T2, M2GVE, M2VD, M2UDV, TGVE, TVS, TVP, TUDV, TUFV, XGVE, $\Delta IEAl_2S_nTA$

Table 6.9: List of selected best features using SelectKBest method for XGBR model. The feature names of the following indicators are provided in Table 6.1-6.4.

Top ranked 45 features selected based on SelectKBest to fit the data on XGBR model
IPTA, SA, WX, WT, SWT, ANX, ANT, PT, GM2, GX, GT, VM2, VX, VT, IEX, IET, ΔIETM1, ΔIEM2T, RX, RT, MPX, MPT, BPX, BPT, χM2, χX, ΔχTM1, ΔχM2T, CEMX, CEMT, DT1M1, DM2T2, M2GVE, M2VD, M2UDV, TGVE, TVS, TVP, TUDV, TUFV, XGVE, XVP, XUPV, $\Delta IEAl_2S_nT$, $\Delta IEAl_2S_nTA$

6.4.3. Hyperparameter Tuning

Further, to improve the model's accuracy, we have implemented hyperparameter tuning for all four ML models, leveraging the best-selected features. The optimized hyperparameters of each model have been tabulated in **Table 6.10**. With the optimized hyperparameters, we proceeded to evaluate the performance of each model on both the training and test data sets as illustrated by the parity plots in **Figure 6.6**. Notably, the training accuracy for both XGBR (0.07 eV) and GBR (0.06 eV) models is very much

the same, reflecting their efficient training. However, the XGBR model (0.24 eV) exhibits slightly higher test accuracy than the GBR model (0.27 eV), showcasing its better predictive ability on unseen data. Conversely, the DTR and RFR models demonstrated comparatively higher train and test errors, signaling suboptimal training and consequent difficulties in accurately predicting adsorption energy **(Figure 6.6)**. In light of these findings, the XGBR model emerged as the most suitable model for adsorption energy prediction, outperforming the other models.

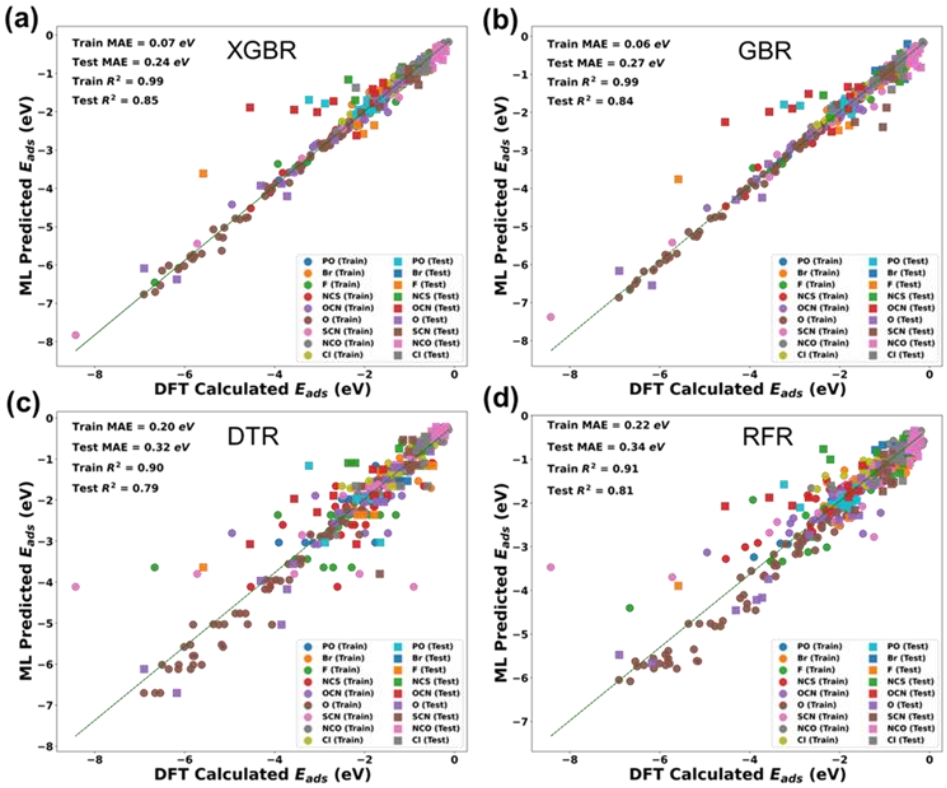


Figure 6.6: Parity plot between the ML predicted adsorption energy vs DFT calculated adsorption energy for the selected four models, (a) XGBR, (b) GBR, (c) DTR, and (d) RFR.

Table 6.10: List of optimized hyperparameters for ML models used to predict the adsorption energy of the polysulfide intermediates.

ML Models	Optimized Hyperparameters
GBR	max_depth=5, min_samples_leaf=2, min_samples_split=10, n_estimators=200
DTR	max_depth=10, max_features='auto', min_samples_leaf=2, min_samples_split=5
RFR	max_depth=15, min_samples_split=5, n_estimators=20
XGBR	base_score=0.5, gamma=0, learning_rate=0.01, max_depth=9, min_child_weight=1, n_estimators=1000, reg_alpha=0.5, reg_lambda=0.5,

6.4.4. Identification of Potential Anchoring Material

Utilizing the optimized XGBR model, we successfully predict the adsorption energies of all polysulfide intermediates for each MXene system. On the basis of adsorption energy values, we have classified the data set into three classes (SE_{ads}, OE_{ads}, and WE_{ads}) as mentioned before for all intermediates. Notably, MXenes within the −2.8 to −4.2 eV range can be considered as effective sulfur host cathode materials. These MXenes address both the shuttle effect issue, preventing polysulfide dissolution in the electrolyte by optimum anchoring, and the irreversibility of electron flow arising from strong adsorption with intermediates. We have analyzed the data with a focus on various components of the MXenes as well as polysulfide intermediates, as presented in **Figure 6.7**.

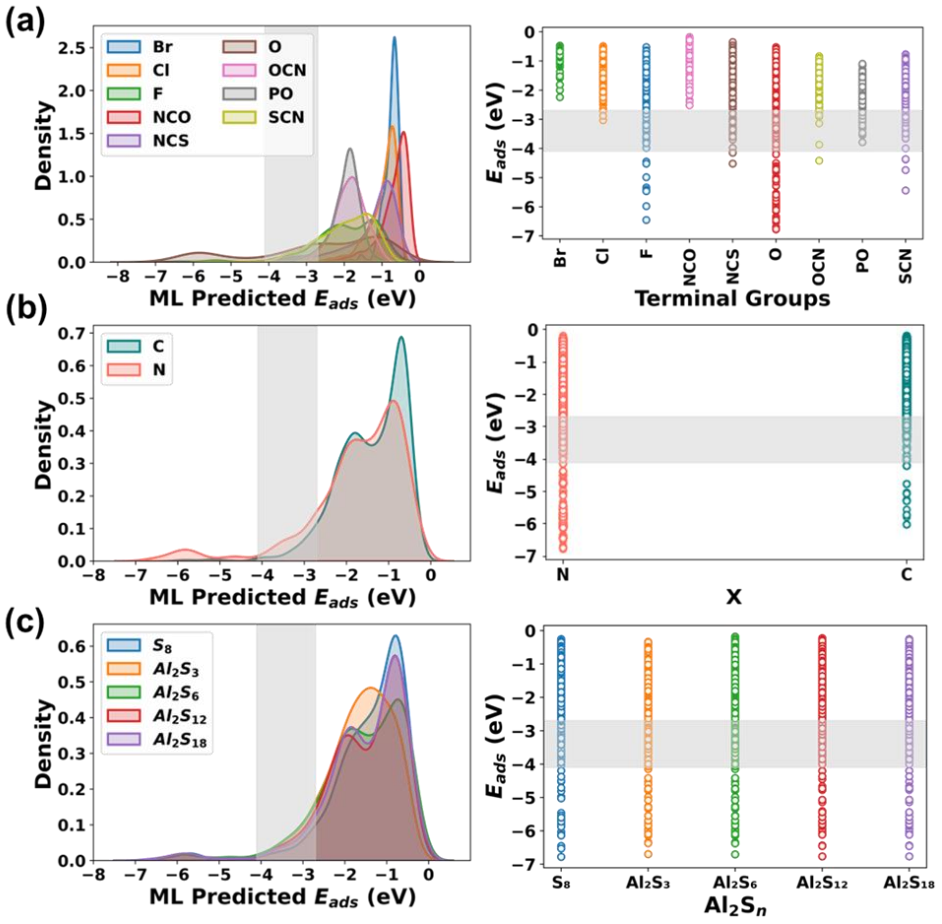


Figure 6.7: Adsorption energy density and their corresponding distribution plot with respect to (a) terminal groups, (b) X (C and N) groups, and (c) polysulfide intermediates for all the MXenes. The shaded area in the plots represents the optimum adsorption region.

From the density and distribution plots **(Figure 6.7a)**, it is observed that the predominant concentration of MXenes is present in the WE_{ads} region, while a smaller fraction occupies the OE_{ads}/SE_{ads} region. The SE_{ads} region is predominantly composed of MXenes bearing O terminal groups, with a sparse representation of F and SCN terminal groups for all of the intermediates **(Figure 6.7a)**. Upon studying the distribution of MXene adsorption energies for different polysulfide intermediates, the MXenes with terminal groups NCS and OCN are found to be distributed widely for

Al_2S_3 and Al_2S_6 compared to S_8, Al_2S_{12}, and Al_2S_{18}, which are concentrated in the WE_{ads} region **(Figure 6.8)**.

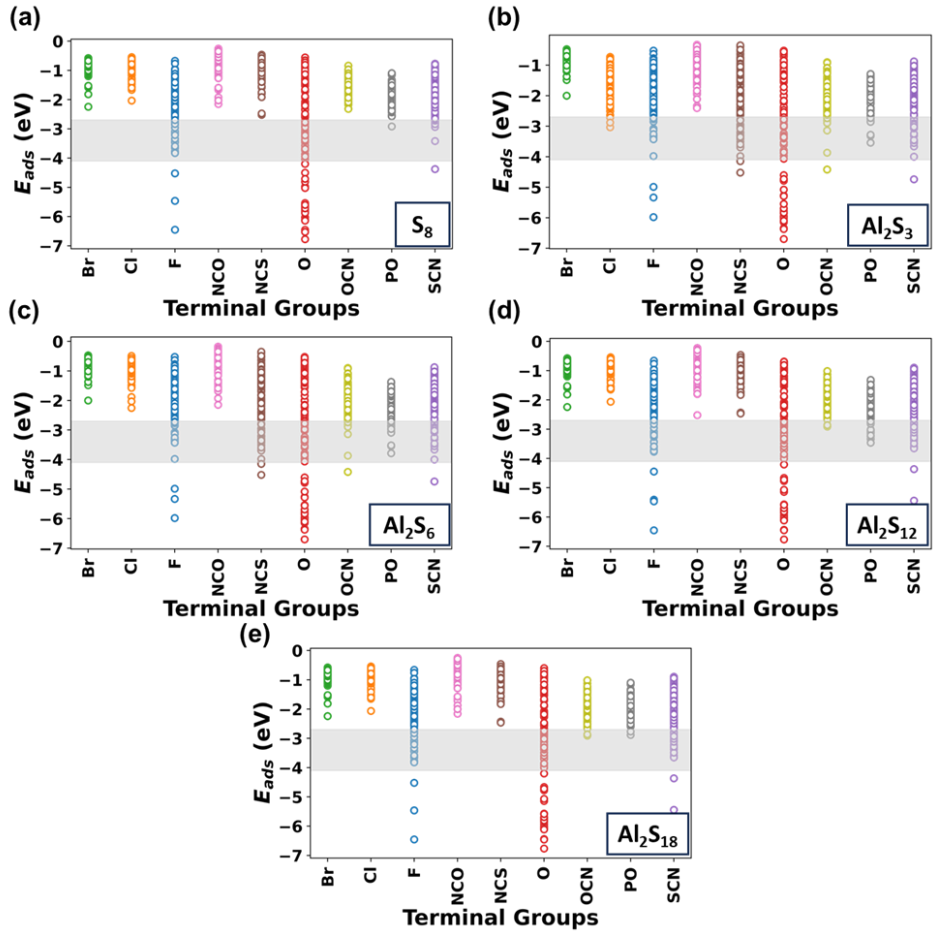


Figure 6.8: Distribution plot of adsorption energy with respect to terminal groups for (a) S_8, (b) Al_2S_3, (c) Al_2S_6, (d) Al_2S_{12}, and (e) Al_2S_{18}. The shaded area in the plots represents the optimum adsorption region.

The MXenes having C as the X are clustered predominantly in the WE_{ads} and OE_{ads} regions **(Figure 6.7b)**, whereas MXenes containing N as X display a broader span of adsorption energy, which signifies a more prominent role played by M1/M2 compositions and terminal groups in this case. The distribution of adsorption energy shows significant difference on going from S_8 to Al_2S_6 **(Figure 6.7c)**. However, the adsorption energy distribution is similar for Al_2S_{12} and Al_2S_{18}.*[65]* The density and distribution plot with respect to M1 and M2 groups revealed that Sc and Y

are preferable as the M1 group and Mo, W, and Nb are preferable as the M2 group to obtain adsorption energies of the polysulfide intermediates in the OE_{ads} region as is evident from **Figure 6.9**. On the other hand, Zr is not found to be suitable as either M1 or M2 metal for optimum adsorption of polysulfides.

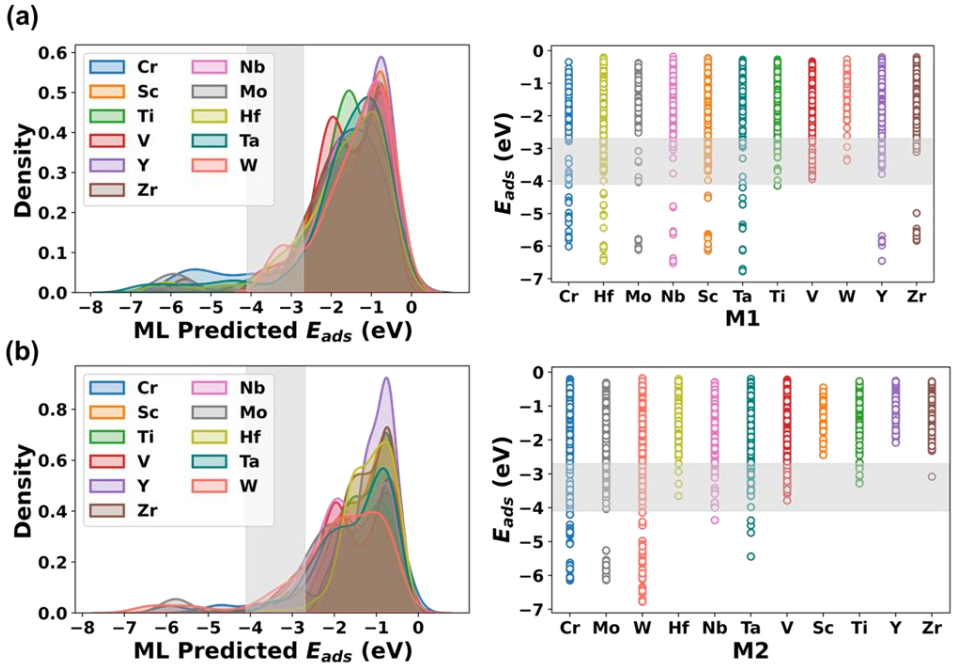


Figure 6.9: Distribution and density plot of adsorption energy with respect to (a) M1, and (b) M2. The shaded area in the plots represents the optimum adsorption region.

Further, we have constructed alluvial plots to understand how the various combinations of M1, M2, and T contribute to the respective adsorption energy regions **(Figure 6.10)**.

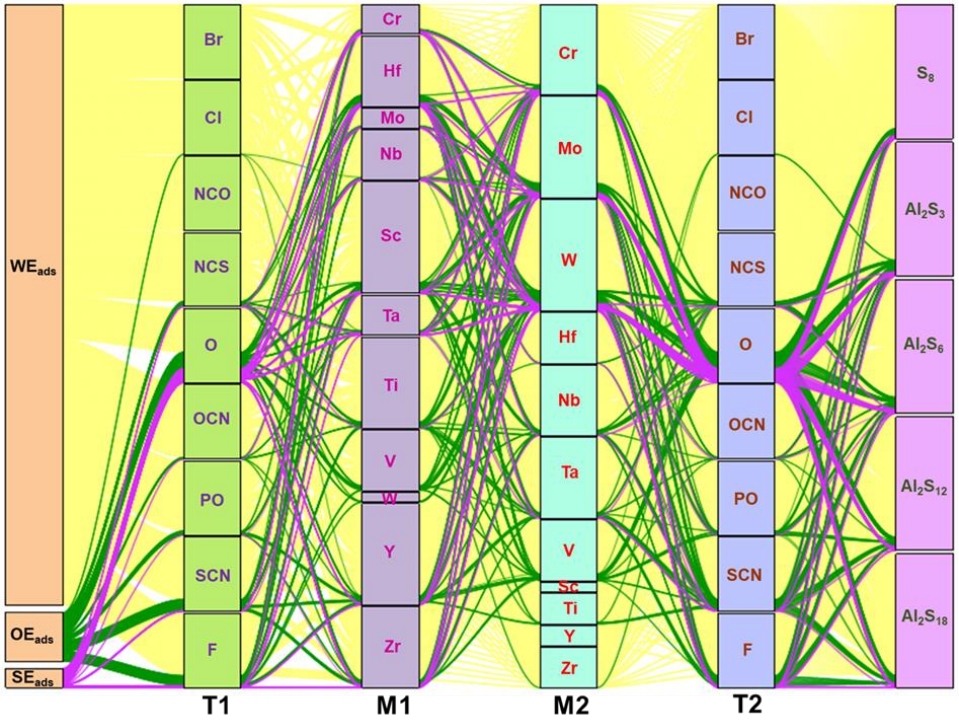


Figure 6.10: Alluvial plot of three categories of adsorption energy (WE_{ads}, OE_{ads}, and SE_{ads}) for various combination of T, M1, and M2 groups of MXene. The yellow-, green- and magenta-colored lines represent weak, optimum and strong adsorption energies for the polysulfide intermediates on the $M1M2XT_2$ MXenes.

The MXenes having O, SCN, and F terminal groups are found to majorly contribute toward the OE_{ads} region for polysulfide adsorption. On the other hand, terminal groups like Br, Cl, NCO, and OCN lead to weaker adsorption region. A significant number of strong adsorption energies for polysulfide also originate from MXenes with O terminal groups. Overall, the alluvial plot gives a visual representation and is very helpful to understand which combination of T, M1 and M2 can lead to optimum adsorption for each of the polysulfide intermediates separately. We have not considered the X layer (C, N) in the plot as it does not majorly contribute in deciding the adsorption energy. The number of MXenes residing in the OE_{ads} region for S_8, Al_2S_3, Al_2S_6, Al_2S_{12}, and Al_2S_{18} are found to be 62, 106, 106, 86, and

76, respectively. The different number of MXenes clearly indicates that all of the MXenes are not suitable for adsorption of all polysulfides.

Hence, we further screen out the MXene materials that show optimum adsorption energy for all five considered intermediates and extract 42 such MXene materials **(Table 6.11)**, which could be potential sulfur host cathode material for Al−S batteries.

Table 6.11: ML predicted adsorption energy of all the Al polysulfides for 42 best MXenes with optimum adsorption.

MXenes					**Polysulfide Adsorption Energy (E_{ads}) (eV)**				
M1	**M2**	**X**	**T1**	**T2**	**S_8**	**Al_2S_3**	**Al_2S_6**	**Al_2S_{12}**	**Al_2S_{18}**
Cr	Cr	C	O	O	-3.87	-3.97	-3.97	-3.93	-3.93
Hf	Mo	C	O	O	-3.73	-3.4	-3.4	-3.69	-3.69
Hf	V	C	O	O	-2.91	-2.74	-2.74	-2.88	-2.88
Hf	W	C	O	O	-3.66	-3.28	-3.28	-3.63	-3.63
Mo	Mo	C	O	O	-4.04	-3.88	-3.88	-4.01	-4.01
Sc	Cr	C	O	O	-3.45	-3.61	-3.6	-3.71	-3.7
Sc	Nb	N	O	O	-2.83	-2.83	-2.83	-2.83	-2.82
Sc	V	C	O	O	-3.14	-2.96	-2.95	-3.12	-3.12
Sc	V	N	O	O	-3.57	-3.54	-3.54	-3.59	-3.59
Sc	W	C	O	O	-3.12	-3.03	-3.02	-3.25	-3.25
Ti	Mo	N	O	O	-3.11	-2.94	-2.94	-3.1	-3.1
Ti	V	C	O	O	-3.28	-3.02	-3.02	-3.27	-3.27
V	Cr	C	O	O	-3.43	-3.4	-3.4	-3.49	-3.49
V	Cr	N	O	O	-3.69	-3.79	-3.79	-3.85	-3.85
V	Nb	N	O	O	-3.95	-3.85	-3.85	-3.85	-3.86
V	V	N	O	O	-3.36	-3.24	-3.24	-3.38	-3.38
V	W	N	O	O	-3.27	-3.01	-3.02	-3.2	-3.21
W	W	N	O	O	-3.38	-3.31	-3.31	-3.37	-3.37
Y	Cr	C	O	O	-3.22	-3.36	-3.34	-3.32	-3.31
Y	V	N	O	O	-2.74	-2.75	-2.75	-2.74	-2.73
Zr	Cr	C	O	O	-2.95	-2.99	-2.98	-3.03	-3.02
Zr	V	N	O	O	-2.77	-2.78	-2.78	-2.75	-2.74
Sc	V	C	PO	PO	-2.92	-2.83	-3.05	-3.33	-2.88
Sc	Nb	N	SCN	SCN	-2.84	-3.5	-3.5	-3.59	-3.59
Sc	Ta	N	SCN	SCN	-2.79	-3.55	-3.55	-3.65	-3.65

Y	Ta	N	SCN	SCN	-2.94	-3.45	-3.45	-3.49	-3.49
Y	Ti	C	SCN	SCN	-2.84	-3.03	-3.03	-3.04	-3.04
Y	Ti	N	SCN	SCN	-2.84	-3.25	-3.25	-3.29	-3.29
Zr	Nb	C	SCN	SCN	-2.72	-2.9	-2.9	-2.94	-2.94
Cr	W	N	F	F	-3.79	-3.32	-3.32	-3.71	-3.78
Mo	W	N	F	F	-3.43	-3.08	-3.08	-3.43	-3.43
Nb	W	N	F	F	-3.77	-3.3	-3.3	-3.77	-3.77
Sc	Mo	N	F	F	-3.09	-3.08	-3.08	-3.08	-3.08
Sc	Ta	N	F	F	-3.46	-3.31	-3.31	-3.44	-3.44
Sc	W	C	F	F	-3.34	-2.76	-2.76	-3.26	-3.33
Ti	Nb	N	F	F	-2.74	-2.73	-2.73	-2.74	-2.74
Ti	Ta	N	F	F	-2.77	-2.73	-2.73	-2.77	-2.77
Ti	W	N	F	F	-3.65	-3.13	-3.13	-3.61	-3.65
V	W	N	F	F	-3.83	-3.26	-3.26	-3.75	-3.82
Y	Mo	N	F	F	-3.59	-3.43	-3.43	-3.59	-3.59
Y	Ta	N	F	F	-2.83	-2.71	-2.71	-2.83	-2.83
Y	W	C	F	F	-3.21	-2.73	-2.73	-3.2	-3.2

All of these best MXene materials are presented through an alluvial plot to represent the combination of various components (M1, M2, X, T), as illustrated in **Figure 6.11**. However, after analysis of the common OE_{ads} region, MXenes containing O and F are found to be the majority in the list and SCN is the minority. The O and F containing MXenes are also reported to be stable among other functional groups present as a terminal group.*[66]* O-terminated MXenes are found to show higher adsorption energy compared to F-terminated MXenes, which are well supported with the previously report.*[67]* Among all the metals considered (M1 and M2), we end up with 10 M1 and 7 M2 groups containing MXenes that reside in the OE_{ads} region. Altogether it has been found that the terminal groups have the major influence on the activity of MXenes.

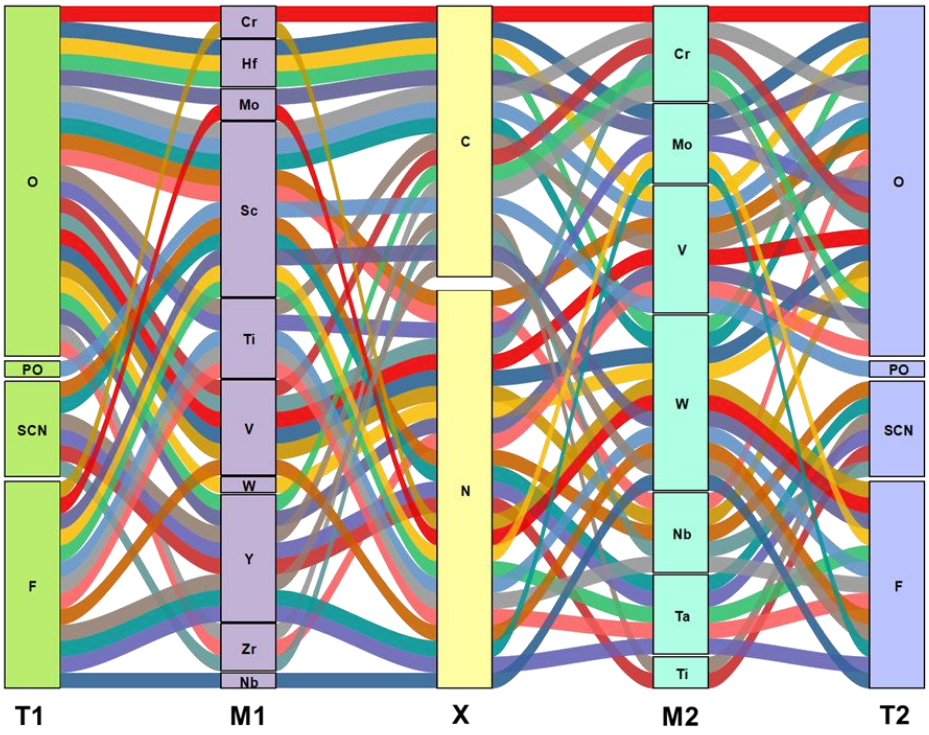


Figure 6.11: Alluvial plot of three categories of adsorption energy (WE_{ads}, OE_{ads}, and SE_{ads}) for various combination of T, M1, and M2 groups of MXene. The yellow-, green- and magenta-colored lines represent weak, optimum and strong adsorption energies for the polysulfide intermediates on the $M1M2XT_2$ MXenes.

MXenes containing transition metal elements mostly exhibit metallic properties, which could enable them to enhance electronic conductivity in intermediates, thus facilitating the conversion back to higher-order Al−S intermediates.*[68]* To explore this further, we have conducted an investigation into the electronic structure of one of our most promising MXene materials, $ScCrCO_2$, both in its pristine state and when adsorbing Al−S intermediates. To do this, we conducted density of states calculations for each of these systems **(Figure 6.12)**.

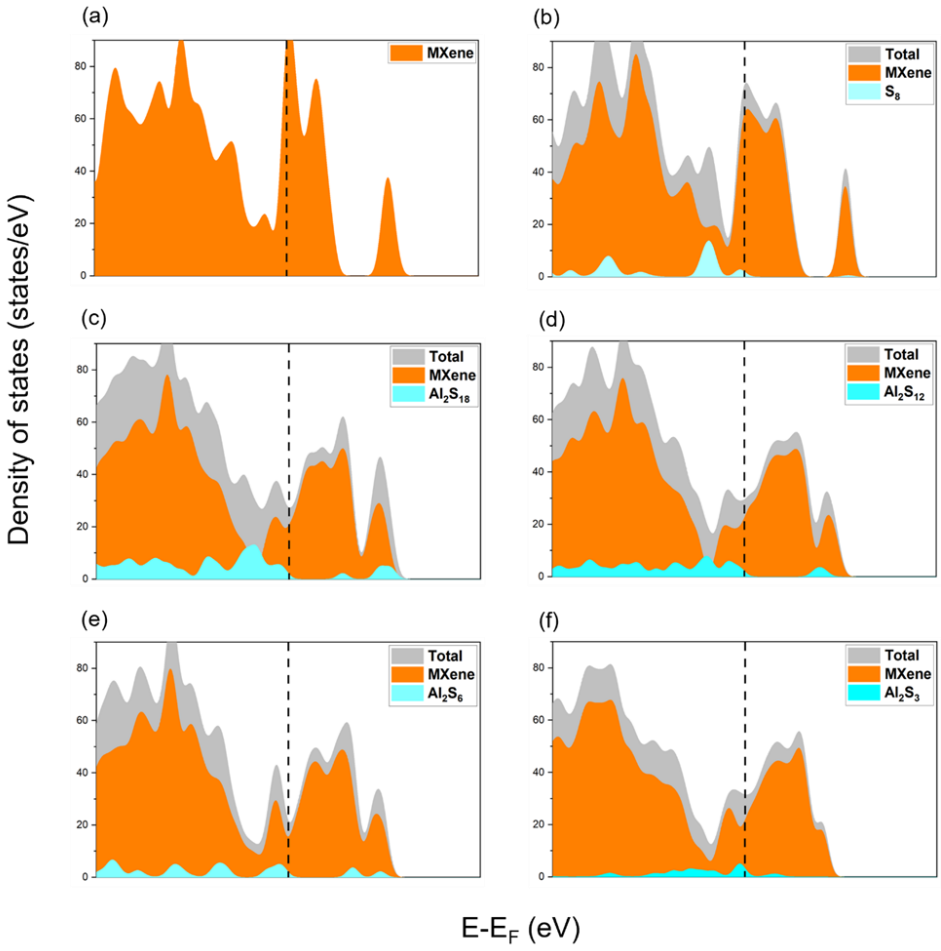


Figure 6.12: Density of state plots of (a) pristine MXene ($ScCrCO_2$), and adsorbed with (b) S_8, (c) Al_2S_{18}, (d) Al_2S_{12}, (e) Al_2S_6, (f) Al_2S_3. The Fermi level is set at zero denoted by dash line. The density of states calculations has been carried out using a Γ k-point grid of 6 × 6 × 1.

Our findings reveal that the MXene material is inherently metallic, owing to the presence of d-electronic states from Sc and Cr, in addition to p-states from C and O. Furthermore, when Al−S intermediates are adsorbed, there is a notable overlap between the electronic states of MXene and the p-states of the Al−S intermediates. Importantly, there are electronic states from all of the components of Al−S and MXene at the Fermi level, preserving the metallic character of the adsorbed systems. Such an intrinsic metallic nature would be conducive in facilitating electron conduction throughout the entire

system, establishing a pathway for electrons to engage in the redox reactions involved in polysulfide conversions on the MXene surface. The insights gathered from this study hold promise for the future development of high-performance Al−S batteries, contributing to the advancement of sustainable and efficient energy technologies.

6.5. Conclusion

In the present work, we have strived to present suitable sulfur host cathode materials for Al−S batteries from the large family of $M1M2XT_2$ MXenes, which can address both the dissolution of intermediates due to weak adsorption and irreversible reactions due to strong adsorption of intermediates. A multistep workflow has been developed for rapid and accurate adsorption energy predictions by combining DFT and machine learning (ML). The trained ML models were improved using ANOVA F-value analysis, by which top ranked features were selected from the overall feature space. Among various implemented ML models, boosting tree-based algorithms were found to outperform the other ML models and XGBR emerged as the most accurate model. From the predicted results, 42 MXene materials have been identified as potential anchoring agents with optimal adsorption for all Al−S intermediate species. Terminal functional groups are found to be the most dominating factors for determining the anchoring capability. MXenes containing O and F terminal groups exhibit greater suitability for achieving optimal adsorption of all the intermediates. Overall, this study represents an efficient method to reduce a large material space to find out the best MXene materials that can anchor the polysulfide intermediates thereby addressing the shuttle effect in Al−S batteries. We believe this study will guide the experimental researchers to choose and design suitable sulfur host materials from a large material space and encourage them to explore novel MXene materials for the Al−S battery application.

6.6. References

(1) Liu Y., Elias Y., Meng J., Aurbach D., Zou R., Xia D., Pang Q. (2021), Electrolyte solutions design for lithium-sulfur batteries, Joule, 5, 2323–2364 (DOI: 10.1016/j.joule.2021.06.002)

(2) Liu Y. T., Liu S., Li G. R., Gao X. P. (2021), Strategy of enhancing the volumetric energy density for lithium-sulfur batteries, Adv. Mater., 33 (8), 2003955 (DOI: 10.1002/adma.202003955)

(3) Ng S. F., Lau M. Y. L., Ong W. J. (2021), Lithium-sulfur battery cathode design: tailoring metal-based nanostructures for robust polysulfide adsorption and catalytic conversion, Adv. Mater., 33 (50), 2008654 (DOI: 10.1002/adma.202008654)

(4) Zhang S., Pollard T. P., Feng X., Borodin O., Xu K., Li Z. (2020), Altering the electrochemical pathway of sulfur chemistry with oxygen for high energy density and low shuttling in a Na/S battery, ACS Energy Lett., 5 (4), 1070–1076 (DOI: 10.1021/acsenergylett.0c00142)

(5) Huang X. L., Wang Y.-X., Chou S.-L., Dou S. X., Wang Z. M. (2021), Materials engineering for adsorption and catalysis in room-temperature Na-S batteries, Energy Environ. Sci., 14, 3757 (DOI: 10.1039/d1ee00669g)

(6) Ye X., Ruan J., Pang Y., Yang J., Liu Y., Huang Y., Zheng S. (2021), Enabling a stable room-temperature sodium-sulfur battery cathode by building heterostructures in multichannel carbon fibers, ACS Nano, 15 (3), 5639–5648 (DOI: 10.1021/acsnano.0c09746)

(7) Guo Y., Jin H., Qi Z., Hu Z., Ji H., Wan L. J. (2019), Carbonized-MOF as a sulfur host for aluminum-sulfur batteries with enhanced capacity and cycling life, Adv. Funct. Mater., 29 (7), 1807676 (DOI: 10.1002/adfm.201807676)

(8) Li H., Meng R., Guo Y., Chen B., Jiao Y., Ye C., Long Y., Tadich A., Yang Q.-H., Jaroniec M., Qiao S.-Z. (2021), Reversible electrochemical

oxidation of sulfur in ionic liquid for high-voltage Al-S batteries, Nat. Commun., 12, 5714 (DOI: 10.1038/s41467-021-25910-2)

(9) Zheng X., Wang Z., Li J., Wei L. (2022), Binder-free S@Ti3C2Tx sandwich structure film as a high-capacity cathode for a stable aluminum-sulfur battery, Sci. China. Mater., 65 (6), 1463–1475 (DOI: 10.1007/s40843-021-1981-4)

(10) Guo Y., Hu Z., Ji H., Wan L. J., Wang J., Peng Z., Zhu J. (2020), Rechargeable aluminium-sulfur battery with improved electrochemical performance by cobalt-containing electrocatalyst, Angew. Chem., Int. Ed., 59, 22963–22967 (DOI: 10.1002/anie.202008675)

(11) Zheng X., Wang Z., Li J., Wei L. (2022), Binder-free S@Ti3C2Tx sandwich structure film as a high-capacity cathode for a stable aluminum-sulfur battery, Sci. China. Mater., 65 (6), 1463–1475 (DOI: 10.1007/s40843-021-1981-4)

(12) Yu X., Manthiram A. (2017), Electrochemical energy storage with a reversible nonaqueous room-temperature aluminum-sulfur chemistry, Adv. Energy Mater., 7 (18), 1700561 (DOI: 10.1002/aenm.201700561)

(13) Smajic J., Wee S., Simoes F. R. F., Hedhili M. N., Wehbe N., Abou-Hamad E., Costa P. M. F. J. (2020), Capacity retention analysis in aluminum-sulfur batteries, ACS Appl. Energy Mater., 3 (7), 6805–6814 (DOI: 10.1021/acsaem.0c01036)

(14) Zhang D., Zhang X., Wang B., He S., Liu S., Tang M., Yu H. (2021), Highly reversible aluminium-sulfur batteries obtained through effective sulfur confinement with hierarchical porous carbon, J. Mater. Chem. A, 9, 8966–8974 (DOI: 10.1039/d1ta01279g)

(15) Sungjemmenla, Soni C. B., Kumar V. (2021), Recent advances in cathode engineering to enable reversible room-temperature aluminium-

sulfur batteries, Nanoscale Adv., 3 (6), 1569–1581 (DOI: 10.1039/d1na00113c)

(16) Hu Z., Guo Y., Jin H., Ji H., Wan L. J. (2020), A rechargeable aqueous aluminum-sulfur battery through acid activation in water-in-salt electrolyte, ChemComm., 56 (13), 2023–2026 (DOI: 10.1039/c9cc09419g)

(17) Tang X., Guo X., Wu W., Wang G. (2018), 2D metal carbides and nitrides (MXenes) as high-performance electrode materials for lithium-based batteries, Adv. Energy Mater., 8 (33), 1801897 (DOI: 10.1002/aenm.201801897)

(18) Lukatskaya M. R., Mashtalir O., Ren C. E., Dall'Agnese Y., Rozier P., Taberna P. L., Naguib M., Simon P., Barsoum M. W., Gogotsi Y. (2013), Cation intercalation and high volumetric capacitance of two-dimensional titanium carbide, Science, 341 (6153), 1502–1505 (DOI: 10.1126/science.1241488)

(19) Naguib M., Halim J., Lu J., Cook K. M., Hultman L., Gogotsi Y., Barsoum M. W. (2013), New two-dimensional niobium and vanadium carbides as promising materials for Li-ion batteries, J. Am. Chem. Soc., 135 (43), 15966–15969 (DOI: 10.1021/ja405735d)

(20) Zhou J., Zha X., Zhou X., Chen F., Gao G., Wang S., Shen C., Chen T., Zhi C., Eklund P., Du S., Xue J., Shi W., Chai Z., Huang Q. (2017), Synthesis and electrochemical properties of two-dimensional hafnium carbide, ACS Nano, 11 (4), 3841–3850 (DOI: 10.1021/acsnano.6b08123)

(21) Zhang C., Cui L., Abdolhosseinzadeh S., Heier J. (2020), Two-dimensional MXenes for lithium-sulfur batteries, InfoMat, 2 (4), 613–638 (DOI: 10.1002/inf2.12110)

(22) Balach J., Giebeler L. (2021), MXenes and the progress of Li-S battery development—a perspective, J. Phys. Energy, 3 (2), 021002 (DOI: 10.1088/2515-7655/abe63d)

(23) Giebeler L., Balach J. (2021), MXenes in lithium-sulfur batteries: scratching the surface of a complex 2D material - a minireview, Mater. Today Commun., 27, 102323 (DOI: 10.1016/j.mtcomm.2021.102323)

(24) Sim E.S., Yi G.S., Je M., Lee Y., Chung Y.C. (2017), Understanding the anchoring behavior of titanium carbide-based MXenes depending on the functional group in LiS batteries: a density functional theory study, J. Power Sources, 342, 64–69 (DOI: 10.1016/j.jpowsour.2016.12.003)

(25) Zhao Q., Zhu Q., Liu Y., Xu B. (2021), Status and prospects of MXene-based lithium-sulfur batteries, Adv. Funct. Mater., 31 (21), 2100457 (DOI: 10.1002/adfm.202100457)

(26) Anand R., Ram B., Umer M., Zafari M., Umer S., Lee G., Kim K.S. (2022), Doped MXene combinations as highly efficient bifunctional and multifunctional catalysts for water splitting and metal-air batteries, J. Mater. Chem. A, 10, 22500–22511 (DOI: 10.1039/d2ta02063e)

(27) Anand R., Nissimagoudar A.S., Umer M., Ha M., Zafari M., Umer S., Lee G., Kim K.S. (2021), Late transition metal doped MXenes showing superb bifunctional electrocatalytic activities for water splitting via distinctive mechanistic pathways, Adv. Energy Mater., 11 (48), 2102388 (DOI: 10.1002/aenm.202102388)

(28) Xie Y., Dall'Agnese Y., Naguib M., Gogotsi Y., Barsoum M.W., Zhuang H.L., Kent P.R.C. (2014), Prediction and characterization of MXene nanosheet anodes for non-lithium-ion batteries, ACS Nano, 8 (9), 9606–9615 (DOI: 10.1021/nn503921j)

(29) Xie Y., Naguib M., Mochalin V.N., Barsoum M.W., Gogotsi Y., Yu X., Nam K.W., Yang X.Q., Kolesnikov A.I., Kent P.R.C. (2014), Role of surface structure on Li-ion energy storage capacity of two-dimensional transition-metal carbides, J. Am. Chem. Soc., 136 (17), 6385–6394 (DOI: 10.1021/ja501520b)

(30) Li X., Guan Q., Zhuang Z., Zhang Y., Lin Y., Wang J., Shen C., Lin H., Wang Y., Zhan L., Ling L. (2023), Ordered mesoporous carbon grafted MXene catalytic heterostructure as Li-ion kinetic pump toward high-efficient sulfur/sulfide conversions for Li-S battery, ACS Nano, 17, 1653–1662 (DOI: 10.1021/acsnano.2c10742)

(31) Wang Z., Yu K., Feng Y., Qi R., Ren J., Zhu Z. (2019), VO2(p)-V2C(MXene) grid structure as a lithium polysulfide catalytic host for high-performance Li-S battery, ACS Appl. Mater. Interfaces, 11 (47), 44282–44292 (DOI: 10.1021/acsami.9b15050)

(32) Zhang Q., Wang Y., Seh Z.W., Fu Z., Zhang R., Cui Y. (2015), Understanding the anchoring effect of two-dimensional layered materials for lithium-sulfur batteries, Nano Lett., 15 (6), 3780–3786 (DOI: 10.1021/acs.nanolett.5b00814)

(33) Guo X., Wang Z., Deng Z., Li X., Wang B., Chen X., Ong S.P. (2019), Water contributes to higher energy density and cycling stability of Prussian blue analogue cathodes for aqueous sodium-ion batteries, Chem. Mater., 31 (15), 5933–5942 (DOI: 10.1021/acs.chemmater.9b02047)

(34) Shen Z., Zhang Z., Li M., Yuan Y., Zhao Y., Zhang S., Zhong C., Zhu J., Lu J., Zhang H. (2020), Rational design of a Ni3N0.85 electrocatalyst to accelerate polysulfide conversion in lithium-sulfur batteries, ACS Nano, 14 (6), 6673–6682 (DOI: 10.1021/acsnano.0c02288)

(35) Van Der Ven A., Deng Z., Banerjee S., Ong S.P. (2020), Rechargeable alkali-ion battery materials: theory and computation, Chem. Rev., 120 (14), 6977–7019 (DOI: 10.1021/acs.chemrev.9b00609)

(36) Jana M., Xu R., Cheng X.B., Yeon J.S., Park J.M., Huang J.Q., Zhang Q., Park H.S. (2020), Rational design of two-dimensional nanomaterials for lithium-sulfur batteries, Energy Environ. Sci., 13 (4), 1049–1075 (DOI: 10.1039/c9ee03299c)

(37) Li J., Qu Y., Chen C., Zhang X., Shao M. (2021), Theoretical investigation on lithium polysulfide adsorption and conversion for high-performance Li-S batteries, Nanoscale, 13 (1), 15–35 (DOI: 10.1039/d0nr07187c)

(38) Zhang H., Wang Z., Ren J., Liu J., Li J. (2021), Ultra-fast and accurate binding energy prediction of shuttle effect-suppressive sulfur hosts for lithium-sulfur batteries using machine learning, Energy Storage Mater., 35, 88–98 (DOI: 10.1016/j.ensm.2021.01.012)

(39) Naguib M., Mochalin V.N., Barsoum M.W., Gogotsi Y. (2014), MXenes: a new family of two-dimensional materials, Adv. Mater., 26 (7), 992–1005 (DOI: 10.1002/adma.201304138)

(40) Naguib M., Mashtalir O., Carle J., Presser V., Lu J., Hultman L., Gogotsi Y., Barsoum M.W. (2012), Two-dimensional transition metal carbides, ACS Nano, 6 (2), 1322–1331 (DOI: 10.1021/nn204153h)

(41) Wang Z., Zheng X., Chen A., Han Y., Wei L., Li J. (2022), Unraveling the anchoring effect of MXene-supported single atoms as cathodes for aluminum-sulfur batteries, ACS Mater. Lett., 4, 1436–1445 (DOI: 10.1021/acsmaterialslett.2c00394)

(42) Zhang X., Zhou J., Lu J., Shen L. (2022), Interpretable learning of voltage for electrode design of multivalent metal-ion batteries, npj Comput. Mater., 8 (1), 175 (DOI: 10.1038/s41524-022-00865-4)

(43) Louis S.Y., Siriwardane E.M.D., Joshi R.P., Omee S.S., Kumar N., Hu J. (2022), Accurate prediction of voltage of battery electrode materials using attention-based graph neural networks, ACS Appl. Mater. Interfaces, 14 (23), 26587–26594 (DOI: 10.1021/acsami.2c04391)

(44) Sun Y., Ayalasomayajula S.M., Deva A., Lin G., García R.E. (2022), Artificial intelligence inferred microstructural properties from voltage-

capacity curves, Sci. Rep., 12 (1), 13421 (DOI: 10.1038/s41598-022-17271-w)

(45) Ha M., Hajibabaei A., Kim D.Y., Singh A.N., Yun J., Myung C.W., Kim K.S. (2022), Al-doping driven suppression of capacity and voltage fadings in 4d-element containing Li-ion-battery cathode materials: machine learning and density functional theory, Adv. Energy Mater., 12 (30), 2201497 (DOI: 10.1002/aenm.202201497)

(46) Ng M.F., Sun Y., Seh Z.W. (2023), Machine learning-inspired battery material innovation, Energy Adv., 2 (4), 449–464 (DOI: 10.1039/d3ya00022j)

(47) Mishra A., Satsangi S., Rajan A.C., Mizuseki H., Lee K.R., Singh A.K. (2019), Accelerated data-driven accurate positioning of the band edges of MXenes, J. Phys. Chem. Lett., 10 (4), 780–785 (DOI: 10.1021/acs.jpclett.8b03742)

(48) Rajan A.C., Mishra A., Satsangi S., Vaish R., Mizuseki H., Lee K.R., Singh A.K. (2018), Machine-learning-assisted accurate band gap predictions of functionalized MXene, Chem. Mater., 30 (12), 4031–4038 (DOI: 10.1021/acs.chemmater.8b00861)

(49) Kresse G., Furthmüller J. (1996), Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B, 54 (16), 11169 (DOI: 10.1103/physrevb.54.11169)

(50) Kresse G., Furthmüller J. (1996), Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci., 6 (1), 15–50 (DOI: 10.1016/0927-0256(96)00008-0)

(51) Kresse G., Hafner J. (1994), Ab initio molecular-dynamics simulation of the liquid-metal-amorphous-semiconductor transition in germanium, Phys. Rev. B, 49 (20), 14251 (DOI: 10.1103/physrevb.49.14251)

(52) Kresse G., Hafner J. (1993), Ab initio molecular dynamics for liquid metals, Phys. Rev. B, 47 (1), 558 (DOI: 10.1103/physrevb.47.558)

(53) Kresse G., Joubert D. (1999), From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B, 59 (3), 1758 (DOI: 10.1103/physrevb.59.1758)

(54) Blöchl P.E. (1994), Projector augmented-wave method, Phys. Rev. B, 50 (24), 17953 (DOI: 10.1103/physrevb.50.17953)

(55) Perdew J.P., Burke K., Ernzerhof M. (1996), Generalized gradient approximation made simple, Phys. Rev. Lett., 77 (18), 3865 (DOI: 10.1103/physrevlett.77.3865)

(56) Grimme S., Antony J., Ehrlich S., Krieg H. (2010), A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, J. Chem. Phys., 132 (15), 154104 (DOI: 10.1063/1.3382344)

(57) Gao T., Li X., Wang X., Hu J., Han F., Fan X., Suo L., Pearse A.J., Lee S.B., Rubloff G.W., Gaskell K.J., Noked M., Wang C. (2016), A rechargeable Al/S battery with an ionic-liquid electrolyte, Angew. Chem., Int. Ed., 55 (34), 9898–9901 (DOI: 10.1002/anie.201603245)

(58) Yang H., Yin L., Liang J., Sun Z., Wang Y., Li H., He K., Ma L., Peng Z., Qiu S., Sun C., Cheng H.M., Feng L. (2018), An aluminum-sulfur battery with a fast kinetic response, Angew. Chem., Int. Ed., 57 (7), 1898–1902 (DOI: 10.1002/anie.201711125)

(59) Zhang H., Wang Z., Ren J., Liu J., Li J. (2021), Ultra-fast and accurate binding energy prediction of shuttle effect-suppressive sulfur hosts for lithium-sulfur batteries using machine learning, Energy Storage Mater., 35, 88–98 (DOI: 10.1016/j.ensm.2021.01.012)

(60) Zafari M., Nissimagoudar A.S., Umer M., Lee G., Kim K.S. (2021), First principles and machine learning based superior catalytic activities and

selectivities for N2 reduction in MBenes, defective 2D materials and 2D π-conjugated polymer-supported single atom catalysts, J. Mater. Chem. A, 9, 9203–9213 (DOI: 10.1039/d0ta12291e)

(61) Zafari M., Kumar D., Umer M., Kim K.S. (2020), Machine learning-based high throughput screening for nitrogen fixation on boron-doped single atom catalysts, J. Mater. Chem. A, 8 (10), 5209–5216 (DOI: 10.1039/d0ta00430e)

(62) Umer M., Umer S., Zafari M., Ha M., Anand R., Hajibabaei A., Abbas A., Lee G., Kim K.S. (2022), Machine learning assisted high-throughput screening of transition metal single atom based superb hydrogen evolution electrocatalysts, J. Mater. Chem. A, 10, 6679–6689 (DOI: 10.1039/d1ta10323a)

(63) Kim C., Pilania G., Ramprasad R. (2016), Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX3 perovskites, J. Phys. Chem. C, 120 (27), 14575–14580 (DOI: 10.1021/acs.jpcc.6b03229)

(64) Chandrasekaran A., Mishra A., Singh A.K. (2017), Ferroelectricity, antiferroelectricity, and ultrathin 2D electron/hole gas in multifunctional monolayer MXene, Nano Lett., 17 (5), 3290–3296 (DOI: 10.1021/acs.nanolett.7b00231)

(65) Bhauriyal P., Pathak B. (2020), Superior anchoring effect of a Cu-benzenehexathial MOF as an aluminium-sulfur battery cathode host, Mater. Adv., 1, 3572 (DOI: 10.1039/d0ma00534a)

(66) Abraham B.M., Sinha P., Halder P., Singh J.K. (2023), Fusing a machine learning strategy with density functional theory to hasten the discovery of 2D MXene-based catalysts for hydrogen generation, J. Mater. Chem. A, 11, 8091–8100 (DOI: 10.1039/d2ta08988c)

(67) Wang Y., Guo T., Alhajji E., Tian Z., Shi Z., Zhang Y.Z., Alshareef H.N. (2023), MXenes for sulfur-based batteries, Adv. Energy Mater., 13 (4), 2202860 (DOI: 10.1002/aenm.202202860)

(68) Xu W., Ke Y., Wang Z., Zhang W., Thye A., Wee S. (2021), The metallic nature of two-dimensional transition-metal dichalcogenides and MXenes, Surf. Sci. Rep., 76, 100542 (DOI: 10.1016/j.surfrep.2021.100542)

Chapter 7

Scope for Future Works

7. Scope for Future Works

The present doctoral work demonstrates how machine learning (ML) can be strategically integrated into battery materials research to efficiently screen large chemical spaces and extract meaningful structure–property relationships. While this thesis addresses multiple bottlenecks in electrode and electrolyte discovery using supervised, unsupervised and interpretable ML models, it also opens the door to a broader set of possibilities. As the energy storage landscape continues to evolve, new directions are emerging where data-driven tools especially next-generation ML and AI frameworks can further accelerate discovery and innovation. The potential future extensions of this work are outlined below:

7.1. Data Generation Using Pretrained ML Potentials

A critical limitation in applying ML to emerging battery chemistries (e.g., aluminum dual-ion batteries, redox flow batteries, organic systems) is the lack of large, curated datasets. While Materials Project and related databases cover metal-ion batteries to some extent, there is a serious gap in data for less conventional systems. Using pretrained and fine-tuned ML potentials, large volumes of DFT-like data can be generated rapidly for underexplored systems. This opens up the possibility of automatically generating datasets for battery materials that are chemically meaningful and diverse, enabling robust model training without requiring expensive computations for every new material.

7.2. ML Potentials for Probing Long-Time Dynamics and Thermal Stability

Understanding how battery materials behave over time and under varying thermal conditions is critical for improving their performance, safety, and lifespan. However, traditional ab initio molecular dynamics (AIMD) simulations are severely limited by high computational cost and short accessible timescales, especially for large and complex systems. This makes

it extremely difficult to explore key dynamic phenomena—such as ion diffusion, phase transformations, structural degradation, and thermal stability under operating conditions. Machine-learned interatomic potentials (ML potentials) offer a promising solution by replicating DFT-level accuracy at a fraction of the computational cost. Models such as CHGNet, M3GNet, and MACE can enable large-scale, long-timescale simulations that were previously infeasible. These simulations make it possible to investigate temperature-driven effects, stability under cycling, and atomistic changes during thermal stress—all of which are crucial for understanding battery failure mechanisms and improving material design.

7.3. Automated Diffusion Barrier Prediction Pipelines

Ion diffusion is a critical parameter in determining battery rate performance. However, calculating diffusion barriers using DFT for each material and each pathway is computationally intensive and scales poorly with system size. With the integration of ML potentials, a fully automated pipeline for diffusion barrier prediction can be developed. By coupling this with existing ML models for capacity, voltage, and stability, a comprehensive screening platform could be constructed to identify multi-objective-optimized battery materials. This framework could also be extended to solid-state electrolytes, where similar ion-hopping mechanisms are present and where computational challenges are even more severe.

7.4. Generative AI for Electrode Material Design

One of the most exciting directions is the application of generative AI (Gen AI) for designing novel materials with user-defined properties. Unlike traditional ML models that predict properties based on existing data, generative models (e.g., diffusion models or generative adversarial networks) can create entirely new molecular or crystal structures conditioned on specific constraints. For instance, one could define a material that should deliver more than 4 V voltage, be composed only of

earth-abundant elements, maintain structural stability across multiple charge/discharge cycles, and exhibit high reversibility. Although this level of multi-objective generation is still a significant challenge, early research suggests that generative AI can dramatically reduce the design space and produce chemically plausible candidates that would otherwise remain undiscovered. Integrating Gen AI into battery materials research has the potential to reshape how we approach materials discovery from reactive exploration to intelligent design.

While this thesis lays the foundation for accelerating battery materials discovery using machine learning, the future lies in more intelligent, generative, and self-improving systems. These tools will not only reduce the time and cost of materials research but also enable the exploration of entirely new chemistries that are beyond the current bounds of human intuition and conventional experimentation. The convergence of ML, generative models, and domain knowledge promises a new era of data-driven innovation in energy storage.