Data-Driven Innovations in Material Science and Chemical Engineering: Enabling Energy Material Design and Industrial Process Optimization

Zhuo Wang (20322762)

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

31st July 2024

Contents

Acknowledgements

I would like to express my deepest gratitude to all those who have supported and guided me throughout the course of my research and the preparation of this thesis.

First and foremost, I am deeply grateful to my supervisors, Prof. Cheng Heng Pang, Prof. Kam Loon Fow, Prof. Edward Lester, and Prof. Siew Shee Lim, for their invaluable guidance and unwavering support. Their mentorship has not only shaped my academic journey but also inspired me with their professionalism and humanity. I have learned so much from them, both academically and personally, and I will always cherish their wisdom and kindness.

I am immensely thankful for the invaluable help and unwavering support provided by Prof. Hainam Do and Prof. Xue Zhang. Their expert guidance and assistance have been instrumental in overcoming numerous challenges encountered during our research work, and their trust in me has filled me with immense pride. I sincerely appreciate their contributions to our research work.

Lastly, and most importantly, I would like to thank my family and friends for their unconditional love, patience, and support. To my parents, thank you for believing in me and providing me with the foundation to pursue my dreams.

This work would not have been possible without the support of all these wonderful people. Thank you all from the bottom of my heart.

List of Publications

The following peer-reviewed journal articles have been published or are in preparation to be published as a result of the work undertaken as part of this thesis:

1. **WANG, Z.**, SUN, Z., YIN, H., LIU, X., WANG, J., ZHAO, H., PANG, C. H., WU, T., LI, S., YIN, Z. and YU, X.-F. (2022). "Data-Driven Materials Innovation and Applications." *Advanced Materials* 34(36): 2104113.

2. ZHANG, X.[#], **WANG, Z**.[#], LAWAN, A. M., WANG, J., HSIEH, C.-Y., DUAN, C., PANG, C. H., CHU, P. K., YU, X.-F. and ZHAO, H. (2023). "Data-driven structural descriptor for predicting platinum-based alloys as oxygen reduction electrocatalysts." *InfoMat.* 2023; 5(6): e12406.

3. **WANG, Z.**, SUN, Z., YIN, H., WEI, H., PENG, Z., PANG, Y. X., JIA, G., ZHAO, H., PANG, C. H. and YIN, Z. (2023). "The role of machine learning in carbon neutrality: catalyst property prediction, design, and synthesis for carbon dioxide reduction." *eScience*: 100136.

4. YIN, H.[#], SUN, Z.[#], **WANG, Z.**[#], TANG, D., PANG, C. H., YU, X., BARNARD, A. S., ZHAO, H. and YIN, Z. (2021). "The data-intensive scientific revolution occurring where two-dimensional materials meet machine learning." *Cell Reports Physical Science* 2(7): 100482.

5. **WANG, Z.**, MENG, Y., FOW, K. L., WU, T. and PANG, C. H. "A Deep-Learning-Assisted Approach for Fault Detection and Real-Time Monitoring for Steam Boilers." *To be submitted.*

6. **WANG, Z.**, HU Y., LIU Z., FOW, K. L., WU, T. and PANG, C. H. "Optimizing Synthesis of High-Performance Lithium Iron Phosphate Using a Data-Driven Active Learning Framework." *To be submitted.*

The following peer-reviewed journal articles have been published or are in preparation to be published but not as part of this thesis:

7. ZHAO, H., CHEN, W., HUANG, H., SUN, Z., CHEN, Z., WU, L., ZHANG, B., LAI, F., WANG, Z., ADAM, M. L., PANG, C. H., CHU, P. K., LU, Y., WU, T., JIANG, J., YIN, Z. and YU, X.-F. (2023). "A robotic platform for the synthesis of colloidal nanocrystals." *Nature Synthesis*.

8. MOSES, O. A., GAO, L., ZHAO, H., WANG, Z., LAWAN ADAM, M., SUN, Z., LIU, K., WANG, J., LU, Y., YIN, Z. and YU, X. (2021). "2D materials inks toward smart flexible electronics." *Materials Today* 50: 116-148.

9. MOSES, O. A., CHEN, W., ADAM, M. L., WANG, Z., LIU, K., SHAO, J., LI, Z., LI, W., WANG, C., ZHAO, H., PANG, C. H., YIN, Z. and YU, X. (2021). "Integration of data-intensive, machine learning and robotic experimental approaches for accelerated discovery of catalysts in renewable energy-related reactions." *Materials Reports: Energy* 1(3): 100049.

The following conference paper have been published:

10. **WANG, Z.**, SUN, Z., YIN, Z., and PANG, C. H. (2023). "Mechanism of carbon dioxide conversion into acetic acid on the dual-metal atom doped two-dimensional Molybdenum Trioxide: A first-principle study." *Applied Energy Symposium 2023.*

11. **WANG, Z.**, YEOH, J. X., WONG, C. D. S., and PANG, C. H. (2022). "Fault Detection and Diagnosis of Steam Boiler Operation Process with Multi-way Principal Components Analysis" *Applied Energy Symposium 2022.*

List of Figures

List of Tables

List of Abbreviations

ABC	Ada Boost Classifier
AFLOW	Automatic-FLOWLIB
ANN	Artificial neural networks
APG	Alkyl Polyglucosides
AUC	The Area Under the Curve
$\mathbf{BET}_{\mathbf{FePO4}}$	BET Specific Surface Area of $FePO_4$
C_{1C}	Initial discharge capacity at 1C rate of LFP samples
CC-CV	Constant Current and Constant Voltage
CN	Coordination number of Pt
CNN	Convolutional Neural Network
COD	Crystallography Open Database
CRM	Cluster Ranking Model
CRR	Carbon Dioxide Reduction Reaction
CSD	Cambridge Structural Database
CV	Cross Validation
$D10_{S}$	The particle diameter at which 10% of the slurry's particles are smaller.
$D50_{S}$	The median particle diameter, where 50% of the particles are smaller and 50% are larger.
$D90_{S}$	The particle diameter at which 90% of the slurry's particles are smaller.
DFT	Density Functional Theory
$Dmax_{S}$	The maximum particle diameter observed in the sample.
DMC	Dimethyl Carbonate
DMSCs	Dual-Metal-Site Catalysts
DNN	Deep neural network
DT	Decision Tree

EC	Ethylene Carbonate
$\boldsymbol{E}_{\textbf{formaiton}}$	Formation Energy
FN	False Negatives
FP	False Positive
FPR	False Positive Rate
FPs	Iron Phosphate
GB	Gradient Boosting
GBC	Gradient Boosting Classification
GBR	Gradient Boosting Regression
GBRT	Gradient Boosting Regression Tree
GCLP	Grand Canonical Linear Programming
GPR	Gaussian process regression
HER	Hydrogen Evolution Reaction
HOIP	Hybrid Organic-Inorganic Perovskites
HOMO	Highest Occupied Molecular Orbital
HTVS	High-Throughput Virtual Screening
ICDD	International Centre for Diffraction Data
ICSD	Inorganic Crystal Structure Database
***k*NN**	*k*-Nearest Neighbor
KPI	Key Performance Indicator
KRR	Kernel ridge regression
LASSO	Least Absolute Shrinkage and Selection Operator.
LFP	Lithium Iron Phosphate
LOOCV	Leave-one-out cross-validation
LR	Linear Regression
LSTM	Long-Short-Term Memory
LUMO	Lowest Unoccupied Molecular Orbital

MAE	Mean Absolute Error
$\boldsymbol{m}_{\mathbf{APG}}$	The mass of Alkyl Polyglucosides added
$\boldsymbol{m}_{\mathbf{CHO}}$	The mass of Glucose added
$\boldsymbol{m}_{\mathbf{H2O}}$	The mass of Deionized Water added
MIV	Mean Impact Value
ML	Machine Learning
$\boldsymbol{m}_{\mathbf{LC}}$	The mass of $LiCO_3$ added
MN	Number of heteroatom around Pt
MN	Number of heteroatom around Pt
MN/CN	Ratio of heteroatom around Pt
MN/CN	Ratio of heteroatom around Pt
MP	Materials Project
MPCA	Multiway Principal Components Analysis
$\boldsymbol{m}_{\mathbf{PEG}}$	The mass of Polyethylene Glycol added
MQSPR	Material Quantitative Structure And Property Relationship
MSE	Mean Square Error
$\boldsymbol{m}_{\mathbf{TiO2}}$	The mass of TiO_2 added
MWCNT	Multi-Walled Carbon Nanotubes
NMP	N-methyl-2-pyrrolidone
NNP	Neural Network Potential
OER	Oxygen Evolution Reaction
OQMD	Open Quantum Materials Database
ORR	Oxygen Reduction Reaction
PC	Principal Component
PCA	Principal Components Analysis
PEG	Polyethylene Glycol
PLS	Partial Least Squares

P_M	Period of a heteroatom in the periodic table
Pt/M	Atomic ratio of Pt and heteroatom in the alloy
Q	Squared Prediction Error
R^2	Coefficient of Dependence
REST	Representational State Transfer
RF	Radom Forest
RFC	Radom Forest Classification
RFR	Radom Forest Regression
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic Curve
RR1	LFP sample recommended by the first round of active learning
RR1B	Control group of RR1 without addition of APG
RR2	LFP sample recommended by the second round of active learning
RR2B	Control group of RR2 without addition of APG
RRF	Regularized Random Forests
SEM	Scanning Electron Microscopy
SGDC	Stochastic gradient descent classifier
SHE	Standard Hydrogen Electrode
SISSO	Sure Independence Screening and Sparsifying Operator
SVM	Support Vector Machine
SVR	Support Vector Regression
T^2	Hotelling's T^2 statistics
TN	True Negative
TP	True Positive
TPR	True Positive Rate
T_s	The highest temperature reached during the sintering process.

t-SNE	t-Distributed Stochastic Neighbor Embedding
VAE	Variational Autoencoder
VE$_{\mathrm{M\text{-}d}}$	Number of valence electrons in the d orbital
VE$_{\mathrm{M\text{-}s}}$	Number of valence electrons in the s orbital
WGAN	Wasserstein Generative Adversarial Network
XAS	X-Ray Absorption Spectra
XRD	X-ray diffraction
$\boldsymbol{Z}_{\mathbf{M}}$	Atomic number of the heteroatom
$\boldsymbol{\Delta A}_{\mathbf{Pt\text{-}M}}$	Difference of relative atomic mass between Pt and heteroatom
$\boldsymbol{\Delta E}_{\mathbf{CO}}$	CO adsorption energy on active sites
ΔEn$_{\mathbf{Pt\text{-}M}}$	Difference of the electronegativity between Pt and heteroatom
$\boldsymbol{\Delta E}_{\mathbf{OH}}$	OH adsorption energy on active sites
$\boldsymbol{\Delta r}_{\mathbf{Pt\text{-}M}}$	Difference of atomic radius between Pt and heteroatom
$\boldsymbol{\eta}$	Overpotentials
$\boldsymbol{\rho}_{\mathbf{30kN}}$	Compacted density of LFP under 30000N
$\boldsymbol{WL}$	The water level of the drum
$\boldsymbol{C}$	Covariance Matrix
$\boldsymbol{\varphi}$	Combined Index of T^2 and Q Statistics
$\boldsymbol{P_s}$	The pressure of the generated steam
$\boldsymbol{T_e}$	The fuel temperature at the inlet of the economizer
$\boldsymbol{T_f}$	The temperature of the inlet fuel
$\boldsymbol{\hat{P}}$	Loading Matrix
$\boldsymbol{\hat{T}}$	Score Matrix
$\underline{\boldsymbol{X}}$	Three-Dimensional Dataset

Abstract

This thesis explores the integration of data-driven methodologies into material science and chemical engineering, focusing on the design and optimization of energy materials and industrial processes. The research is structured into three interconnected areas: catalyst design, energy material synthesis, and industrial process optimization. Chapter 4 investigates the structure-activity relationship of Pt-based alloys for oxygen reduction reaction (ORR) catalysis using high-throughput density functional theory (DFT) and the SISSO algorithm, demonstrating how computational techniques can predict optimal catalyst candidates. Chapter 5 extends the data-driven approach to optimize the synthesis of lithium iron phosphate (LFP) using machine learning (ML) models, where active learning-based optimization enhances the electrochemical performance of battery materials. Finally, Chapter 6 shifts focus to macro-scale industrial process monitoring, applying long short-term memory (LSTM) networks and multivariate statistical process control (MPCA) for real-time monitoring and prediction of steam boiler operations in industrial settings.

While these three chapters address distinct aspects of material science and chemical engineering, they share a unified methodological framework that employs data-driven techniques to solve complex problems across different scales. From micro-scale catalyst design to material synthesis at the meso-scale and real-time process optimization at the macro-scale, the common philosophy of iterative optimization,

integration of computational predictions with experimental validation, and data-driven innovation provides a cohesive strategy. By seamlessly bridging the scales and methodologies, this work demonstrates the broad-reaching impact of data-driven tools in the fourth paradigm of material science and chemical engineering. This thesis highlights the transformative potential of data-driven approaches, underscoring their applicability in accelerating the design of advanced materials, improving process efficiency, and contributing to the advancement of green chemical technology and sustainability.

Chapter 1

Introduction

1.1 Background

Data-driven innovation has transformed all aspects of our life. It typically involves the invention of novel products and systems based on the knowledge extracted from data by using advanced analysis tools. The adoption of data-driven approaches has led to data-based decision-making innovations in commerce and technology, such as autonomous vehicles, MuZero, and Alphafold (artificial intelligence for mastering games and predicting protein folding, respectively) (Senior et al., 2020, Silver et al., 2018, Vinyals et al., 2019, Schrittwieser et al., 2020). In particular, the massive amounts of data generated by employing both computational and experimental methods, in combination with advanced machine-learning (ML) techniques, have led the field of materials science and chemical engineering into the fourth paradigm of scientific research (**Figure 1.1**) (Schleder et al., 2019). This data-driven paradigm has resulted in the advancement of experimental tools, computational techniques, and big-data analysis (de Pablo et al., 2014, Green et al., 2017). The transformation from the trial-and-error to the data-driven paradigm requires a combination of authoritative and updated knowledge from the three domains of mathematics and statistics, computer science, and materials science and chemical engineering (Sun et al., 2016). The advancement and appropriate integration of these three domains will contribute to chemical data generation and analysis, uncertainty characterization, and efficient exploration of structure-property relationships, providing insights and promoting innovations in material science and chemical engineering.

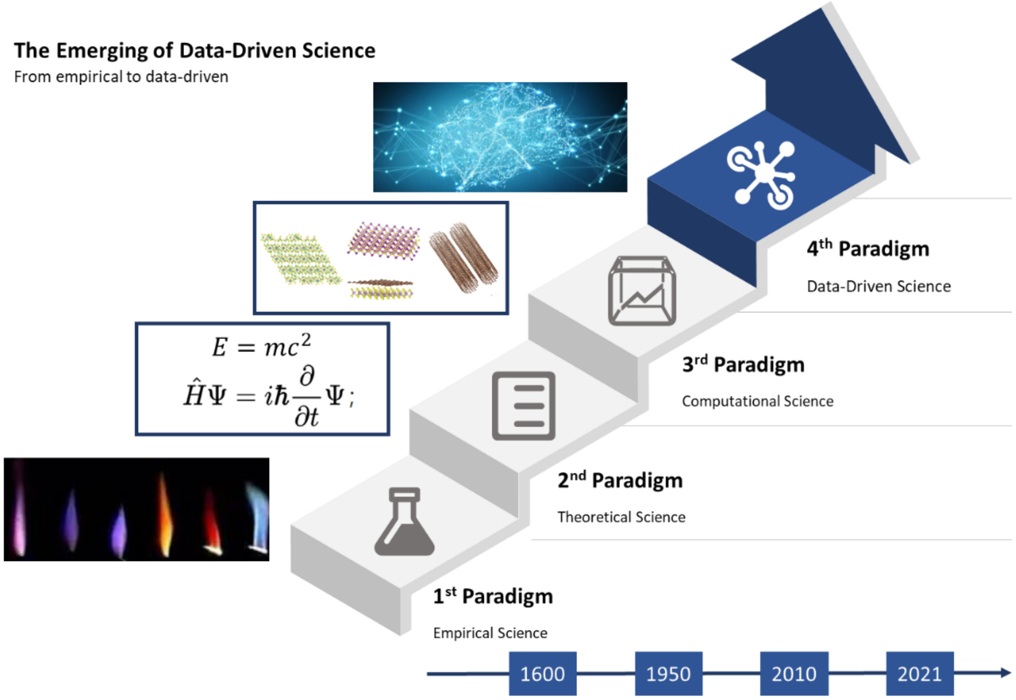

Figure 1.1 The four paradigms of science evolved along with time, including empirical science, theoretical science, computational science and data-driven science.

Data-driven innovations are essential and indispensable to breakthroughs in numerous applications, from energy conversion and storage to flexible electronics and optoelectronics (Chen et al., 2020a, Wexler et al., 2018a, Zhang et al., 2020a, Back et al., 2019, Tran and Ulissi, 2018, Dondapati and Chen, 2020, Ma et al., 2019, Zhang et al., 2020b). For instance, novel photovoltaic materials that are cheap, stable, and environmentally friendly, easy to synthesize, and exhibit a high power conversion efficiency are being investigated (Jin et al., 2020). Moreover, researchers are identifying highly active electrocatalysts that are selective towards the reduction of carbon dioxide (Zhong et al., 2020b). The development of effective data-driven approaches is essential to meet the rapidly growing demand for innovative materials with improved and robust performance (Rück et al., 2020, Ulissi et al., 2017). A basic data-driven framework involves three fundamental stages: employment of data-intensive strategies and ML algorithms (Elton et al., 2019, Lee et al., 2020), development of a comprehensive database and data generation approaches (Kirklin et al., 2015c, Jain et al., 2013), and construction of descriptors that can link data-intensive and experimental strategies (Chen et al., 2019a, Ouyang et al., 2018).

Data-driven approaches for enabling material science and chemical engineering have certain advantages: (1) they outperform conventional trial-and-error approaches in terms of efficiency and accuracy (Zunger, 2018, Hautier et al., 2012, Liu et al., 2017b); (2) they can rapidly learn and extract the complex and implicit inner correlations and knowledge from the massive amounts of chemical data (Chen et al., 2020b, Schleder et al., 2020, Jablonka et al., 2020, Ward et al., 2016a); (3) they can achieve tailored material design based on desired functionalities because of their ability to obtain composition-structure-process-property relations (Zunger, 2018, Sanchez-Lengeling and Aspuru-Guzik, 2018); (4) they use ML models and descriptors to utilize complex features such as electron density and molecular graphs for improving the performance of combinatorial generalization and relational reasoning (Tabor et al., 2018, Chen et al., 2019a). Because of these advantages, many data-driven approaches exhibit high accuracy and efficiency in the prediction of properties and the exploration of property relationships (Lu et al., 2018). Furthermore, the potential of dynamic and iterative meta-optimization data-driven processes, which represent an active learning loop that incorporates the fundamental stages, has been shown in some recent studies (Zhong et al., 2020b, Yuan et al., 2018). Comprehensive reviews have detailed the applicability of data-driven approaches to energy materials (Chen et al., 2020b, Gu et al., 2019, Chen et al., 2020a), structural materials (Sparks et al., 2020), polymeric materials(Cencer et al., 2021), and porous materials(Jablonka et al., 2020), with the help of high-throughput approaches such as density functional theory (DFT) and ML (Schleder et al., 2019). The applications of ML in synthetic chemistry (Strieth-Kalthoff et al., 2020) and the prediction of material properties (Liu et al., 2017b) have also been published.

The development of effective data-driven approaches is essential to meet the rapidly growing demand for superior materials and intelligent technologies with improved and robust performance. The main objective of the thesis is about exploring and developing methodology for applying data-driven approaches for rapid and efficient discovery of high-performance innovative materials at micro level and the effective establishment of intelligent and reliable system at macro level, thereby prompting and enhancing the development of material science and chemical engineering. To achieve this goal, the fundamental stages of the data-driven framework must be utilized and integrated, highlighting and the relationships between raw data and target properties or functions. Besides, those applications are mainly about energy materials design and discovery at micro level, and energy saving and efficiency improvements at macro level.

1.2 Thesis Outline

Chapter 2 provides an overview of data-driven innovation in material science and chemical engineering, covering four key areas: commonly used frameworks such as direct design, inverse design, and active learning; an introduction to widely used chemical databases; an exploration of descriptors that transfer chemical data to ML models; and a discussion on the application of data-driven techniques in areas like ORR, CRR, and battery materials.

Chapter 3 details the methodologies used throughout the thesis, including data generation, preparation, and collection. The chapter starts with the DFT calculation method for designing ORR catalysts, followed by the synthesis of LFP materials and associated characterization techniques. It concludes with a discussion on using data-driven techniques for real-time process monitoring, including control limits for fault detection.

Chapter 4 integrates high-throughput DFT and ML techniques to identify innovative descriptors for screening Pt-based alloy catalysts for ORR. This data-driven strategy highlighted five promising candidates from 77 materials, with further refinement through active learning. The use of the SISSO algorithm created highly predictive feature combinations, contributing to both ORR performance predictions and the rational design of electrocatalysts.

Chapter 5 develops an active learning framework to optimize lab-scale LFP synthesis via the solid-state reaction. A dataset of 80 LFP samples was used to train ensemble ML models, which successfully identified synthesis parameters that led to high-performance LFP samples. This study demonstrates the potential of integrating machine learning into material synthesis to enhance battery material properties.

Chapter 6 presents a generalized data-driven framework combining LSTM and MPCA for real-time fault detection in batch steam boilers. Using historical data, the system predicts future behavior and detects faults early, preventing failures and economic loss. This framework, validated with simulated data, paves the way for more intelligent monitoring systems in chemical engineering.

Chapter 7 summarizes the research findings and outlines potential future research directions.

1.3 Aims and Objectives

This thesis represents the studies on enhancing and improving the collaboration and integration between data-driven tools and material science and chemical engineering, recognizing this as a key strategy to advance green chemical technology and enable carbon neutrality. Specifically, this research focused on implementing data-driven innovation at various scale, from micro to macro level, and across different sub-disciplines in material science and chemical engineering:

from DFT computation validation, to experimental synthesis optimization, to industrial operation monitoring. Chapter 4 proposes a strategy combining high-throughput DFT calculations and machine learning to explore the descriptor used for screening Pt-based alloy catalysts with high Pt utilization and low Pt consumption. Moreover, Chapter 5 demonstrates a novel data-driven active learning framework to optimize the synthesis of high-performance lithium iron phosphate (LFP) materials. Furthermore, Chapter 6 develops a generalized framework incorporating conventional long-short-term memory (LSTM) network and multi-way principal components analysis (MPCA) is developed to apply fault detection and monitoring techniques to the dynamical steam boiler operation process. These data-driven based applications are worthy to be concerned for advancing the design and discovery of energy materials and optimizing industrial operations, thereby shed light on enabling energy saving and using efficiency, and reducing CO_2 emissions. The details of the exploration about innovations in collaboration and integration between data-driven tools and material science and chemical engineering are as follows:

1. The research in Chapter 4 aims to employ data-driven-based strategy in micro level material design and discovery. By integrating high-throughput DFT computations and ML techniques, innovative descriptors that can effectively screen Pt-based alloy catalysts with high Pt utilization and low Pt consumption for oxygen reduction reaction (ORR), are identified. By employing the data-driven strategy, 5 out of 77 materials are discovered as potential candidates that can catalytic ORR with low overpotential.

Moreover, with the established structure-property relationship, second and third round of active learning further recommend Pt-based alloys with high activity. Additionally, the utilized ML algorithm highlighted the critical features, as well as their combinations, which can be effectively employed for predicting the activity of Pt-based alloys.

2. The research in Chapter 5 aims to utilize data-driven active learning framework to optimize the lab-scale synthesis parameters of high-temperature solid-state reaction for the preparation of LFP materials. By learning from the dataset that contains the synthesis data of 80 LFP samples, an active learning-based framework incorporating with two ensemble ML models is developed. Specifically, a classification model works in series with a regression model. For a given synthesis recipe, the classification model will first predict the potential category of the sample: low, medium, and high compacted density of LFP under 30000N (ρ_{30kN}). If the classifier identifies that the synthesis parameter will results a sample with high ρ_{30kN}, the recipe is further sent into the regressor to accurately predict the value of initial discharge capacity at 1C rate of LFP samples (C_{1C}). The recipe will be used as the synthesis parameter of the next experiment and therefore newly generated data will augment the original dataset, dynamically updating the ML models for giving next-round recommendations.

3. The research in Chapter 6 aims to employ data-driven techniques on macro-level chemical industrial processes. Specifically, a method based on the

integration between deep learning model and MPCA is proposed to conduct fault detection and online monitoring for steam boilers work in batches in the real industry. The proposed deep-learning-based method in this work can predict the future behavior of steam boilers, evaluate the process condition, prevent further fault development, and avoid safety issues and economic loss, only using a historical database of past normal operations. The proposed method employed simulated operation data to establish a framework with several critical stages including data pre-processing, establishment of a historical database, calculation of statistical control limit, fault detection and online monitoring, which are intuitive and straightforward to understand and identify faults.

As is shown in Figure 1.2, the three main parts of this thesis: catalyst design (Chapter 4), material optimization (Chapter 5), and process monitoring (Chapter 6), may seem distinct in their application, yet they are unified by a shared methodological framework rooted in data-driven innovation. Each chapter employs computational and experimental approaches, iterative optimization, and feature analysis to refine material properties and industrial processes at different scales. Specifically, data-driven methodologies underpin each part: high-throughput DFT calculations and the Sure Independence Screening and Sparsifying Operator (SISSO) algorithm for catalyst screening (Chapter 4), ensemble machine learning models for active learning-driven LFP synthesis optimization (Chapter 5), and LSTM-MPCA frameworks for real-time industrial process monitoring (Chapter 6). These approaches highlight the scalability of data-driven techniques, from atomistic modeling at the micro-level to industrial-scale process optimization at the macro-level.

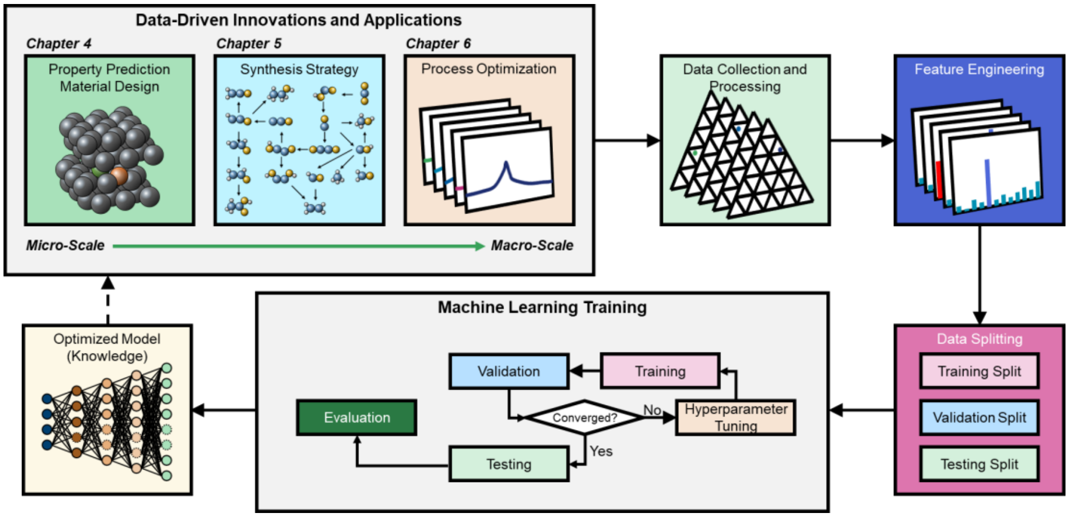

Figure 1.2 The research framework of data-driven innovation employed in this thesis.

Despite their differing applications, the three parts are connected by their shared focus on integrating computational predictions with experimental validation, and their commitment to iterative optimization. This unified approach demonstrates the transformative potential of data-driven methodologies in addressing complex challenges within material science and chemical engineering. By systematically bridging theory and experiment, the research not only advances material performance, such as in electrocatalysis and battery materials, but also enhances operational efficiency in industrial settings. These interconnections reinforce the overarching goal of advancing green chemical technology and supporting carbon neutrality, ensuring a more structured and coherent presentation of the research.

1.4 Publication List

The following peer-reviewed journal articles have been published or are in preparation to be published as a result of the work undertaken as part of this thesis:

1. **WANG, Z.**, SUN, Z., YIN, H., LIU, X., WANG, J., ZHAO, H., PANG, C. H., WU, T., LI, S., YIN, Z. and YU, X.-F. (2022). "Data-Driven Materials Innovation and Applications." *Advanced Materials* 34(36): 2104113.

2. ZHANG, X.[#], **WANG, Z**.[#], LAWAN, A. M., WANG, J., HSIEH, C.-Y., DUAN, C., PANG, C. H., CHU, P. K., YU, X.-F. and ZHAO, H. (2023). "Data-driven structural descriptor for predicting platinum-based alloys as oxygen reduction electrocatalysts." *InfoMat.* 2023; 5(6): e12406.

3. **WANG, Z.**, SUN, Z., YIN, H., WEI, H., PENG, Z., PANG, Y. X., JIA, G., ZHAO, H., PANG, C. H. and YIN, Z. (2023). "The role of machine learning in carbon neutrality: catalyst property prediction, design, and synthesis for carbon dioxide reduction." *eScience*: 100136.

4. YIN, H.[#], SUN, Z.[#], **WANG, Z.[#]**, TANG, D., PANG, C. H., YU, X., BARNARD, A. S., ZHAO, H. and YIN, Z. (2021). "The data-intensive scientific revolution occurring where two-dimensional materials meet machine learning." *Cell Reports Physical Science* 2(7): 100482.

5. **WANG, Z.**, MENG, Y., FOW, K. L., WU, T. and PANG, C. H. "A Deep-Learning-Assisted Approach for Fault Detection and Real-Time Monitoring for Steam Boilers." *To be submitted.*

6. **WANG, Z.**, HU Y., LIU Z., FOW, K. L., WU, T. and PANG, C. H. "Optimizing Synthesis of High-Performance Lithium Iron Phosphate Using a Data-Driven Active Learning Framework." *To be submitted.*

The following peer-reviewed journal articles have been published or are in preparation to be published but not as part of this thesis:

7. ZHAO, H., CHEN, W., HUANG, H., SUN, Z., CHEN, Z., WU, L., ZHANG, B., LAI, F., WANG, Z., ADAM, M. L., PANG, C. H., CHU, P. K., LU, Y., WU, T., JIANG, J., YIN, Z. and YU, X.-F. (2023). "A robotic platform for the synthesis of colloidal nanocrystals." *Nature Synthesis*.

8. MOSES, O. A., GAO, L., ZHAO, H., WANG, Z., LAWAN ADAM, M., SUN, Z., LIU, K., WANG, J., LU, Y., YIN, Z. and YU, X. (2021). "2D materials inks toward smart flexible electronics." *Materials Today* 50: 116-148.

9. MOSES, O. A., CHEN, W., ADAM, M. L., WANG, Z., LIU, K., SHAO, J., LI, Z., LI, W., WANG, C., ZHAO, H., PANG, C. H., YIN, Z. and YU, X. (2021). "Integration of data-intensive, machine learning and robotic experimental approaches for accelerated discovery of catalysts in renewable energy-related reactions." *Materials Reports: Energy* 1(3): 100049.

The following conference paper have been published:

10. **WANG, Z.**, SUN, Z., YIN, Z., and PANG, C. H. (2023). "Mechanism of carbon dioxide conversion into acetic acid on the dual-metal atom doped two-dimensional Molybdenum Trioxide: A first-principle study." *Applied Energy Symposium 2023.*

11. **WANG, Z.**, YEOH, J. X., WONG, C. D. S., and PANG, C. H. (2022). "Fault Detection and Diagnosis of Steam Boiler Operation Process with Multi-way Principal Components Analysis" *Applied Energy Symposium 2022.*

Chapter 2

Background Studies and Literature Review

Part of the content in this chapter has been published in:

WANG, Z., SUN, Z., YIN, H., LIU, X., WANG, J., ZHAO, H., PANG, C. H., WU, T., LI, S., YIN, Z. and YU, X.-F. (2022). "Data-Driven Materials Innovation and Applications." Advanced Materials 34(36): 2104113.

WANG, Z., SUN, Z., YIN, H., WEI, H., PENG, Z., PANG, Y. X., JIA, G., ZHAO, H., PANG, C. H. and YIN, Z. (2023). "The role of machine learning in carbon neutrality: catalyst property prediction, design, and synthesis for carbon dioxide reduction." *eScience*: 100136.

2.1 Synopsis

This chapter begins with a discussion of the recent advances in data-driven discovery in material science and innovation in chemical engineering. First, the various components of the conceptual framework, including the important stages that guide the data-driven process, are introduced. The typical data-driven frameworks and workflow, such as direct design, inverse design and active learning, as well as their critical stages, including data preparation, feature engineering and model training and applications, are described. Then for each critical stage, the chapter proceeds to give a more detailed review. The most wildly used databases, including computational and experimental databases, are introduced. Then this chapter critically reviews and summaries the commonly descriptors used in data-driven applications, and the discussed the descriptors' importance and the mechanism about how they transfer chemical information into the language that ML model can understand. Then this chapter also presents a critical discussion on how data-driven processes are applied in relevant material science and chemical engineering fields, including ORR, CRR and rechargeable alkali-ion batteries. Finally, this chapter ends with the description of the aims and objectives of the thesis, along with a brief outline of the thesis chapters.

2.2 Overview of Data-Driven Innovations and Frameworks in Material Science and Chemical Engineering

The development of the data-driven framework for material innovation has been extensively studied by using ML algorithms (Chen et al., 2020b), material databases (Grazulis et al., 2009, Kirklin et al., 2015c, Jain et al., 2013), and molecular descriptors (Yap, 2011, Ward et al., 2016a). A classical data-driven framework for innovative material discovery typically consists of five fundamental stages: goal identification, data processing, feature engineering, ML and analysis, and application (Wang et al., 2020a). This section describes commonly used frameworks for data-driven processes, including direct design (Liu and Yu, 2020, Zunger, 2018), inverse design (Zunger, 2018, Sanchez-Lengeling and Aspuru-Guzik, 2018), and active learning (Yuan et al., 2018, Zhong et al., 2020b, Zunger, 2018). Critical stages such as data processing, feature engineering, and ML model training facilitate the utilization and processing of material data and molecular descriptors and the effective implementation of ML algorithms (Barnard, 2020, Kotsiantis; et al., 2007).

The design and selection of the data-driven framework depend on the application and the material. Although ML can be potent and effective in a data-driven process, it is not the panacea to solve all challenges in materials science (Barnard, 2020). ML models cannot find solutions to questions that are ill-posed or not appropriately expressed. An in-depth and comprehensive understanding of the chemistry phenomena is necessary to accurately describe the question and relate it to a clear goal. The goal of a data-driven process should be specific, measurable, attainable, relevant, and timely (Schleder et al., 2019). Different ways of defining the goal will lead to varying outcomes of the data-driven process. For example, for the discovery of high-performance photovoltaic materials, Lu et al. (Lu et al., 2018) employed ML to predict the bandgap of candidate materials, whereas Padula et al. (Padula et al., 2019) predicted the power conversion efficiency. The nature of the question is also vital for designing the data-driven framework; using a classification model to explore the correlation between target properties and input features or a regression model to distinguish between several categories of materials is difficult. For instance, Jin et al. (Jin et al., 2020) applied a classification ML model to screen two-dimensional photovoltaic materials with suitable power conversion efficiencies, whereas Sahu et al. (Sahu et al., 2019) employed a regression ML model to predict the power conversion efficiency of candidate photovoltaic materials. Thus, the design of a suitable data-driven framework requires the customization of data processing, feature engineering, and ML model deployment based on the questions being posed.

2.2.1 Frameworks for the Overall Data-Driven Process

The data-driven process framework organizes and integrates the fundamental stages of processing data (Pankajakshan et al., 2017), generating molecular descriptors (Ward et al., 2016a) and deploying the ML model (Wang et al., 2020a). Such frameworks determine the data flow and the interaction style between the theory and experiments or computations (Zunger, 2018, Sanchez-Lengeling and Aspuru-Guzik, 2018, Gu et al., 2019). In this section, we introduced the most commonly employed frameworks to support the discovery of innovate materials including direct design, inverse design and active learning. As illustrated in **Figure 2.1**a and b, direct and inverse design differ from one another in terms of the direction assumed by predictions between material structure and target functionality. Active learning (**Figure 2.1**c) focuses primarily on data flow in the dynamic iteration loop to improve and accelerate the search and prediction process (Zhong et al., 2020b, Yuan et al., 2018). Alternative data-driven frameworks have also been reported in light of specific material phenomena to be addressed. For example, regression and classification models could be assembled into a single framework to enable high-throughput materials screening (Schleder et al., 2020). A transfer learning model could also be integrated into the framework to solve for a small data problem data within the broader the data-driven process (Frey et al., 2020).

It is worth noting that the ML techniques in such data-driven frameworks extend far beyond property prediction and pattern recognition (Ouyang et al., 2018, Shenai et al., 2012, Maaten and Hinton, 2008). They can be utilized in other fundamental stages to generate features,(Yosipof et al., 2015) evaluate feature importance (Wexler et al., 2018a), and visualize data (Zhong et al., 2020b), In both direct and inverse design, the selection of ML algorithms influences the framework architecture (Sanchez-Lengeling and Aspuru-Guzik, 2018).

2.2.1.1. Direct Design

Direct design is the conventional approach to material discovery and primarily involves measurement and theoretical interpretation of the target property (Zunger, 2018). This trial-and-error approach involves searching for the material demonstrating the targeted functionality within the chemical space, which the prior knowledge can help constrain (Schleder et al., 2019). Analogous to the structure-property relations derived by data-driven approaches, the direct design approach typically employs the structural features of known materials to predict target properties. Though direct design is widely employed, it presents obstacles to deliberate discovery. For example, as the direct design initiates from a known structure, it is unable arrive at materials whose structure is not known *a priori* but may possess the desired properties (Zunger, 2018). The case-by-case searching characteristic of direct design is both time- and cost-intensive when extensive structure screening is employed to involve as many materials as possible (Weymuth and Reiher, 2014, Freeze et al., 2019).

As asserted by Zunger (Zunger, 2018), the direct design could be classified into descriptive and predictive approaches. Descriptive direct design employs both modeling and theory to interpret and confirm experimental observations. The predictive direct design, however, can be sub-divided into property prediction for a specific material, or candidate material search in a material space. For example, Jin et al. (Jin et al., 2020) applied a data-driven predictive direct design framework, screening 26 out of 187,093 inorganic crystal structures as potential photovoltaic candidates. The blue squares at the bottom of the graph of **Figure 2.1**a illustrate known compounds with specified compositions (presented by atom numbers Z_A and

Z_B), while question mark-labeled region corresponds to unreported compounds. The upper plot of **Figure 2.1**a represents the value of specific material properties as a function of Z_A and Z_B. In a direct-design-based data-driven framework, the materials discovery journey follows the path from the bottom part of the graph to the top part.

2.2.1.2. Inverse Design

Inverse design can be regarded as the opposite of direct design.(Freeze et al., 2019) In an inverse-design-based data-driven framework, the workflow is initiated in the functional space and terminates in the chemical space (Zunger, 2018). Its objective is to discover tailored materials with desired properties without the exploration of large material space (Freeze et al., 2019). In the inverse design framework, the target functionality is used as the input to predict the corresponding material structure. Rather than arriving at a unique structure with the desired functionality, the goal is to determine a distribution of probable structures. For instance, Dudiy et al. (Dudiy and Zunger, 2006a) employed inverse design in conjunction with specified target properties (e.g. deepest nitrogen level), followed by a search for a desirable material structure.

High-throughput virtual screening (HTVS) is one of the earliest employed methods in inverse design. However, HTVS analysis is generally applied to a smaller number of structures in the course of exploring various functionalities (Zunger, 2018). More recently, generative models, a class of ML method involving the implementation advanced algorithms, including variational autoencoders (VAEs),(Kramer, 1991) generative adversarial networks (GANs) (Goodfellow et al., 2014), recurrent neural network (RNN) (Abiodun et al., 2018), and reinforcement learning (Sutton and Barto, 2018), are commonly employed in inverse design to determine the molecular structure and the probability distribution both of material elemental parameters and desired target properties (**Figure 2.1**b). For example, Jin et al. (Jin et al., 2018) propose a VAE-based inverse design framework to generate graphs of molecular structure. Inverse design represents an advanced, effective data-driven framework for the discovery of novel materials; open research questions remain, including formulation of the molecular presentation in the inverse design process (Sanchez-Lengeling and Aspuru-Guzik, 2018).

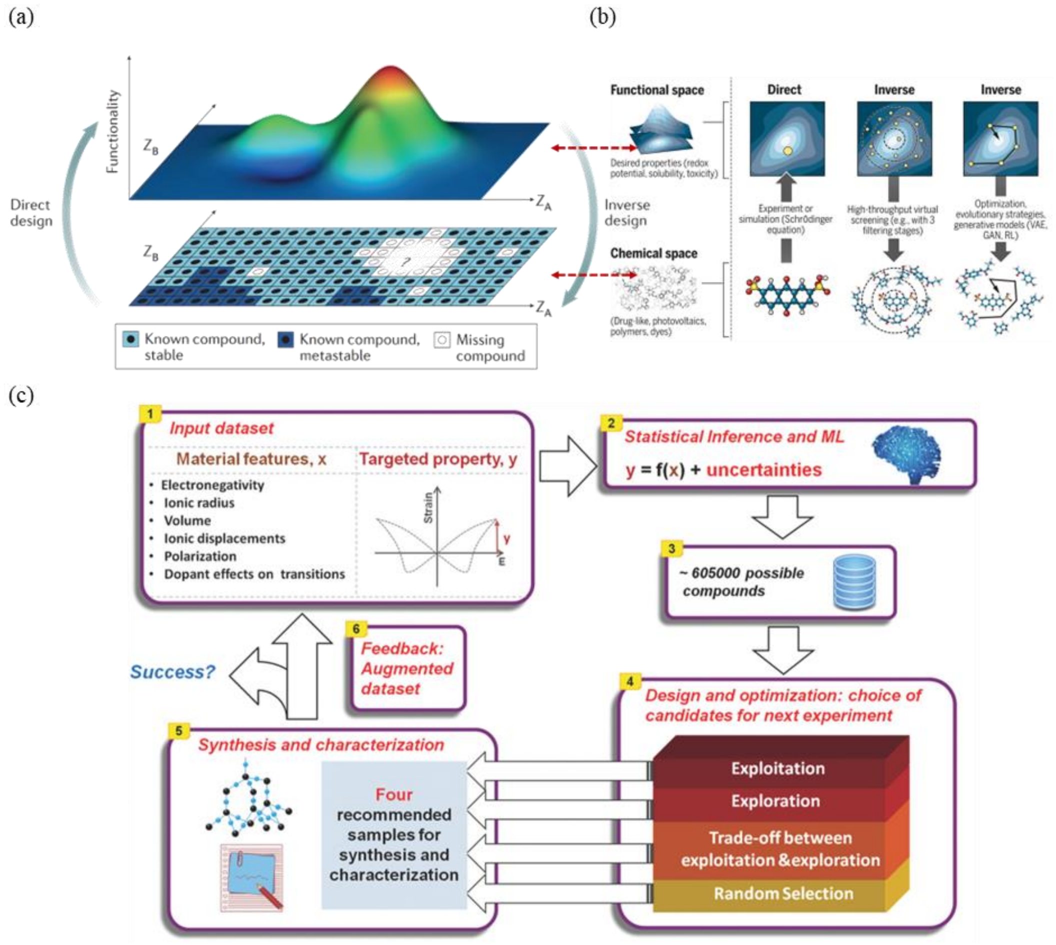
(a)
Functionality
Z_B
Z_A
Direct design
Known compound, stable
Known compound, metastable
Missing compound
(b)
Functional space
Desired properties (redox potential, solubility, toxicity)
Inverse design
Chemical space
(Drug-like, photovoltaics, polymers, dyes)
Direct
Inverse
Inverse
Experiment or simulation (Schrödinger equation)
High-throughput virtual screening (e.g., with 3 filtering stages)
Optimization, evolutionary strategies, generative models (VAE, GAN, RL)
(c)
1
Input dataset
Material features, x
Targeted property, y
Electronegativity
Ionic radius
Volume
Ionic displacements
Polarization
Dopant effects on transitions
Strain
E
y
2
Statistical Inference and ML
y = f(x) + uncertainties
3
~ 605000 possible compounds
4
Design and optimization: choice of candidates for next experiment
Exploitation
Exploration
Trade-off between exploitation &exploration
Random Selection
5
Synthesis and characterization
Four recommended samples for synthesis and characterization
6
Feedback: Augmented dataset
Success?

Figure 2.1 (a) Direct and inverse methods for the design and discovery of materials. Reproduced with permission (Zunger, 2018). Copyright 2018, Springer Nature Publications. (b) The schematic of direct design and inverse design with different targets in material design and discovery. Reproduced with permission (Sanchez-Lengeling and Aspuru-Guzik, 2018). Copyright 2018, AAAS Publications. (c) The active learning framework for the discovery of materials with high electrostrains. Reproduced with permission (Yuan et al., 2018). Copyright 2018, Wiley Publications.

2.2.1.3. Active Learning

The essential idea of active-learning-based data-driven frameworks is to provide high-performance ML models with less training; the machine selects its own training dataset(Settles, 2012) In an active learning framework, the stages of ML training, data processing, and the generation of new training sets are iteratively combined (Yuan et al., 2018, Smith et al., 2018, Zhong et al., 2020b). For instance, Zhong et al. (Zhong et al., 2020b) proposed a random-forest-based active ML framework that iteratively trained more than 300 ML models to predict the binding energy of carbon monoxide on the surface of catalyst for the carbon dioxide reduction reaction (CRR). The trained ML model indicated promising adsorption sites during their active learning workflow, which guided the DFT computation for the subsequent iteration. The DFT results evaluated in the latest iteration were combined with the original data to construct a new training dataset, which would yield an updated ML model.

In general, an active learning framework contains an inquiry loop to guide further experiments or computations (Settles, 2012, Smith et al., 2018). Active learning is most applicable when numerous data instances and their labels are easily collected, synthesized or computed to address queries in iterative training (Settles, 2012). In an active learning framework proposed by Yuan *et al*. (Yuan et al., 2018), the electrostrain of piezoelectric candidates were iteratively queried. Such active learning frameworks are suitable for dynamic optimization problems and sequential design in innovative material discovery.

2.2.2 Fundamental Stages in Data-Driven Framework

A complete data-driven material discovery framework involves fundamental stages including raw data processing (Jablonka et al., 2020, Pankajakshan et al., 2017), feature engineering (Ward et al., 2016a), and ML model training (Wang et al., 2020a). In the data processing stage, there are two major steps: data acquisition and data pre-processing (Cai et al., 2020). Generally, there are two types of data utilized in a data-driven material discovery process: experimental data and computational data (Schleder et al., 2019). Both could be either self-generated or queried from existing databases. Relevant, sufficient, consistent and complete data is the foundation of a successful data-driven process (Jablonka et al., 2020). Collected data may contain a number of issues including missing, redundant, abnormal or imbalanced data (Jablonka et al., 2020). Data pre-processing ensures that the ML model performs satisfactorily. Data pre-processing generally consists of four main stages: outlier detection, data complementation, discretization, and normalization (Kotsiantis; et al., 2007). Data may exist in various forms, including numerical values, structure graphs, images, text, or signals. For example, Lee et al. (Lee et al., 2020) trained a deep learning model to predict potential defects in electron microscopy images with aberration-corrected scanning transmission taken as the model input. Both the quantity and quality of data influence the selection and performance of ML models. For instance, neural network models typically require more data to be reliably implemented (Wang et al., 2020a). It is critical to acquire material data from reliable sources; commonly used material databases and relevant data management tools are systematically discussed in following sections.

Feature engineering is the process of constructing the descriptor space, which mainly consists of two steps: the selection or generation of descriptors; construction of the descriptor space (Wei et al., 2019). The selection of descriptors depends on the goal of the data-driven process and is characterized by the greatest extent of human intervention. The target of this step is to identify and extract the most appropriate and critical descriptors from the pre-processed data to construct descriptor space. Problem-specific domain knowledge is essential here, for example, to specify the relevant properties and determine the proper scale length (atomistic, coarse-grained, and global) (Jablonka et al., 2020). However, there may be situations in which no suitable descriptor is available, or the basic descriptors are not sufficient to describe the environment or frame the materials with respect specific targets. Thus, an alternative is to generate high-performance descriptors from the original ML training dataset. A good descriptor space is one that is sufficient for the prediction and resolution of the target functional space (Schleder et al., 2019). Therefore, an in-depth review of molecular descriptors is presented in Section 5 to offer insights on the construction of descriptor space.

ML model training, which follows the construction of the descriptor space, includes model selection, evaluation, and optimization (Yin et al., 2021). The implementation of the majority ML algorithms requires the specification of hyperparameters which determine the ML model configuration of ML (Raschka, 2018). Various hyperparameters result in different model formulations; model selection aims at identifying with the appropriate hyperparameter formulation which results in the best model performance. Therefore, hyperparameter tuning is critical to model optimization; it controls the complexity and flexibility of the model to identify the balance between overfitting and underfitting by handling the variance-bias trade-off (Jablonka et al., 2020). More complex models tend to fit training data better but also exhibit a higher variance on the test data, whereas a simpler models (such as regularized linear regression) tends to exhibit a higher bias on the test data. Hyperparameter tuning and model selection can be classified as a meta-optimization task (Raschka, 2018), where validation techniques are employed to evaluate the performance in terms of the ML algorithm objective function.

2.2.3 Model Performance Evaluation and Uncertainty Quantification

The ultimate goal of the ML model deployment stage is to train the model such that offers accurate predictions for both test and unseen data; therefore, it becomes essential to effectively assess the performance while characterizing the inherent uncertainty of the model (Morgan and Jacobs, 2020, Jablonka et al., 2020). A review by Morgan and Jacob (Morgan and Jacobs, 2020) gives an excellent overview and sample cases of best practices in ML model development, assessment and uncertainty quantification. In this subsection, we will discuss model performance evaluation methods and uncertainty quantification in the context of model deployment, focusing on commonly employed validation techniques and performance evaluation metrics.

2.2.3.1. Performance Evaluation Techniques

Three techniques are commonly employed for model performance evaluation: holdout (Raschka, 2018), cross-validation (CV) (Hawkins et al., 2003), and bootstrap (Jablonka et al., 2020). In most ML deployment processes, the data are divided into training data, validation data, and test data (Morgan and Jacobs, 2020). The holdout approach statically splits the available data for training, validation, and testing at a fixed ratio. Though the holdout approach is straightforward, it may introduce pessimistic bias when the size of the original dataset is small; such splitting further reduces the size while potentially impacting the statistics of the training data. CV represents a continuous, iterative, crossing-over training and validation process that can be regarded as the ensemble of the holdout approach, sampling data without replacement (Raschka, 2018). For a typical k-fold CV process, the dataset is divided equally into k parts, one of which is adopted as the validation set; the remaining $k - 1$ parts are combined into a new training subset. When the number of folds is equal to the data points ($k = n$), a special case of CV is manifested (the leave-one-out cross-validation (LOOCV), which, though computationally expensive, is useful when the dataset is small (Jablonka et al., 2020). Sahu et al. (Sahu et al., 2018a) applied the LOOCV to 280 data points of small molecule OPV systems to evaluate ML model predictions of power conversion efficiency. Unlike CV, bootstrap samples data with replacement result in only approximately 63.2% of the data points being sampled (Efron and Tibshirani, 1986) and potentially a high bias given that the sampled data is not representative of the complete dataset. To correct this bias, Efron (Efron and Tibshirani, 1997) has proposed a 0.632(+) bootstrap approach. In general, CV provides a nearly unbiased estimator with high variance, while bootstrap approaches tend to yield estimators with low variance for small datasets (Kim, 2009, Efron and Tibshirani, 1997).

2.2.3.2. Performance Evaluation Metrics

The determination of performance metrics is essential for ML model evaluation and optimization. For regression models, commonly employed metrics are the mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE) and coefficient of dependence (R^2), which are expressed as follows (Schleder et al., 2019, Liu et al., 2017b):

$$\boldsymbol{MAE = \frac{1}{N}\sum_{i=0}^{N}|y_i - \hat{y}_i|} \qquad (2\text{-}1)$$

$$\boldsymbol{MSE = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2\, , RMSE = \sqrt{MSE}} \qquad (2\text{-}2)$$

$$\boldsymbol{R^2 = 1 - \frac{\sum_{i=0}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{N}(y_i - \bar{y})^2}} \qquad (2\text{-}3)$$

where N refers to the number of sample data points, $y_i, \hat{y}_i, and\, \overline{y}$ represent the actual value, predicted value, and mean value, respectively. The MAE treats the errors equally, whereas larger errors are allocated a higher weight in the MSE and RMSE. The MSE and RMSE are differentiable and commonly used to identify minima optimization processes. R^2 represents the proportion of the variance in true values relative to the predicted values.

The predictivity of classification models can be described by the value of four indicators: true positive (TP), true negative (TN), false positive (FP), and false negatives (FN) (Lever et al., 2016). Frequently employed evaluation metrics, including Accuracy, Precision, Recall, and F1, can be derived based on the four indicators. Numerous misjudgments resulting in false positives contribute to low precisions, whereas missing of positives correspond to low recalls. A combined metric, called the F1 score, balances these two metrics and is beneficial for cases in which the data is imbalanced.

$$\boldsymbol{Accuracy = \frac{TP+TN}{TP+TN+FP+FN}} \tag{2-4}$$

$$\boldsymbol{Precision = \frac{TP}{TP+FP}} \tag{2-5}$$

The receiver operating characteristic (ROC) curve and the area under the curve (AUC) are also effective performance metrics in binary classification. The ROC represents the plot of the true positive rate (TPR) versus the false positive rate (FPR), where the formulas for TPR and FPR are presented as follows:

$$\boldsymbol{TPR\ =\ \frac{TP}{TP+FN}} \tag{2-6}$$

$$\boldsymbol{FPR\ =\ \frac{FP}{FP+TN}}. \tag{2-7}$$

A perfect binary classifier would demonstrate an AUC=1; AUC = 0.5 indicates that the binary classifier is no better than random guessing (Chen et al., 2020b).

2.2.3.3. Domain of Applicability and Uncertainty Quantification

The reliability and accuracy of the trained models must be evaluated by considering domain applicability and quantifying uncertainties (Morgan and Jacobs, 2020). to the determination of domain applicablity relates to distance metrics between the potential and training data points. Though many methods have been proposed to measure such distances (Sahigara et al., 2012, Schwaighofer et al., 2009), they are relatively difficult to implement to obtain qualitative guidance on model applicability. All such methods rely upon calculated distance metrics whose validity has not been determined for the particular problem, while also requiring the definition of suitable thresholds (Morgan and Jacobs, 2020).

Predicted value uncertainties are more intuitive and readily quantified to enable the evaluation of model performance. Evaluating error bars is an important tool to support model comparisons, stability estimation and of the reliability of model predictions (Jablonka et al., 2020). Ensemble approaches are commonly employed to quantify uncertainties; a popular methodology involves training the same model via bootstrap or CV, and then treating the ensemble variance as a surrogate for the error bars (Peterson et al., 2017). An alternative approach involves utilizing the same training data while refitting the model by adjusting the model architecture (Morgan and Jacobs, 2020). A large variance between these predictions in a specific chemical domain indicates that the ML models are still tangling and require additional training data (Behler, 2014). The two types of ensemble methods can also be combined in random forest decision tree models, for which Morgan and Jacobs provide an in-depth example (Morgan and Jacobs, 2020). The ensemble approaches are more computationally expensive; however, their flexibility enables them to be employed in numerous models.

Prediction uncertainty can also be quantified by distance-based approaches, which are based on the concept that such uncertainties correlate with the distance between the potential corresponding training data points. Hirschfeld et al. (Hirschfeld et al., 2020) employed log-scaled Tanimoto distance (Bajusz et al., 2015) and Euclidean distance (Janet et al., 2019) to quantify the displacement between potential points from training data and predictions of molecular properties, respectively. Bayesian approaches (Smith, 2013) can also automatically quantify uncertainty while potentially avoiding iterations, though this requires the adoption of specific ML models making it less generally applicable (Morgan and Jacobs, 2020, Jablonka et al., 2020).

2.3 Overview of Chemical Database for Material Science and Chemical Engineering

Recent developments in data-centric approaches are expected to dramatically accelerate the progress in materials science because experimental and computational methods generate massive amounts of data, causing increasing complexity (Draxl and Scheffler, 2020). Databases pertaining to both computational and experimental materials have been established to serve various specialized activities, rather than for dissemination or to enable contributions from the broader community (Hill et al., 2018). The primary challenge in choosing and comparing databases is identifying the specific function that the database uniquely support, while also being able to compare various databases on the same structural basis (Hegde et al., 2020b). Error! Reference source not found. lists the properties of dominant databases and their various attributes including data types, materials of focus, number of entries, data source, license, and a simple database descriptor.

Relatively simple analytical tasks pose challenges unique to the data-driven era because we are unable to capture, curate, store, search, share, analyze, and visualize the data in the absence of proper tools (Zhou et al., 2017). Thus, the identification of large numbers of correlations and patterns complex datasets has necessarily been carried out by high-throughput implementations of ML algorithms for decades to generate predictive and classification models for targeted physical properties. We have summarized representative high throughput tools (pymatgen (Ong et al., 2013), qmpy (Kirklin et al., 2015b), ASE (Larsen et al., 2017), and atomate (Mathew et al., 2017)) and workflow management tools (FireWorks (Jain et al., 2015), AFLOWπ (Supka et al., 2017), matminer (Ward et al., 2018a), and AiiDA (Yakutovich et al., 2021, Huber et al., 2020)). This class of high-throughput and workflow management tools is generally available in an open-source, Python infrastructure, with data connectivity implemented in RESTful API. These components aid in automating, managing, persisting, sharing, and reproducing the complex workflows associated with modern computational science and all associated data, reducing the cost and enhancing the efficiency of data summarization approaches with respect to the popular "*five V's*": volume, velocity, variety, veracity, and value (Nguyen, 2018). Representative databases and the high-throughput management toolkits have been summarized in **Figure 2.2**.

More specifically, individual databases each solve one specific problem by relaying the specific descriptors which have been extracted from other existing databases. For instance, database formulation may be motivated by the need to synthesize specific materials for a specific application, such as the accelerated discovery of stable lead-free hybrid organic-inorganic perovskites (HOIP) (Lu et al., 2018), accurate prediction of battery life (Severson et al., 2019), and various catalysis applications (Kitchin, 2018). The potential of data-driven strategies to uncover complex phenomena and design novel, high-performance materials is dependent on the quality and accessibility of databases and high-throughput tools, and which would otherwise not be possible with conventional trial-and-error approaches.

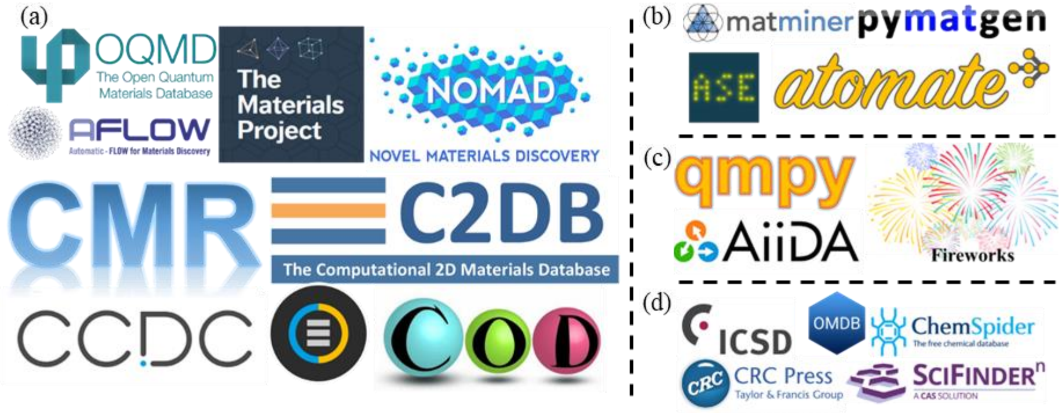

Figure 2.2 The representative (a) theoretical dominant databases (b) experimental dominant databases, (c) high-throughput tools with (d) workflow management tool.

Table 2.1 The database including the name, data type, materials types, number of the entries, and data sources.

Database	Types	Materials	No. Entries	Data Source	Ref.
Open Quantum Materials Database (OQMD)	Computational	Inorganic Solids	~300,000	ICSD, Hypothesis	(Kirklin et al., 2015b, Saal et al., 2013, Kirklin et al., 2015a)
Materials Project (MP)	Computational	Inorganic Solids; Nanoporous Materials	>130,000 ~530,000	ICSD	(Jain et al., 2013)
Automatic-FLOW (AFLOW)	Computational	Inorganic Solids, Alloys	3,312,125	ICSD	(Curtarolo et al., 2012)
Novel Material Discovery (NOMAD)	Computational	Inorganic Solids	--	Literarues	(Draxl and Scheffler, 2018)
The Computational Materials Repository (CMR)	Computational	Perovskites, 2D Materials	--	OQMD	(Landis et al., 2012)
Inorganic Crystal Structure Database (ICSD)	Experimental	Inorganic Crystal Structures	>232.012	Literarues	(Belsky et al., 2002)
Cambridge Structural Database (CSD)	Experimental	Metal Organic Frameworks, Orgaincs Molecure	>800.239	Literatures, ICDD	(Groom et al., 2016a)
Crystallography Open Database (COD)	Experimental, Computational		>385,000	Literatures,	(Grazulis et al., 2009)
The Computational 2D Materials Database (C2DB)	Computational	2D Materials	~4,000	MP, CMR	(Haastrup et al., 2018)

Database	Types	Materials	No. Entries	Data Source	Ref.
Clean Energy Project (CEP)	Computational	Organic Photovoltaics	>2,000,000	Literatures, Hypothesis	(Hachmann et al., 2011)
Organic Materials Database (OMDB)	Computational	Organic Materials	~12,500	COD	(Borysov et al., 2017)
Joint Automated Repository For Various Integrated Simulations (JARVIS)-DFT	Computational	2D/Solid Inorganics	~40,000	MP, OQMD, AFLOW, Literatures.	(Choudhary et al., 2020a)
Citrination	Experimental, Computational	Inorganic Solids, Molecules	--	Literatures,	(Hill et al., 2018)
Materials Cloud	Experimental, Computational	All Materials	--	ICSD, COD, Literatures	(Mounet et al., 2018)
Alloy Database	Computational	Intermetallics	--	ISCD	(Widom and Mihalkovic, 2005)
CatApp	Computational	Molecules on Surfaces	--	--	(Hummelshoj et al., 2012)
Computational Chemistry Comparison and Benchmark DataBase (CCCBDB)	Computational	Atoms, Moleculres	~2069	--	(Johnson III, 1999)
Computational Electronic Structure Database (CompES-X)	Computational	Inorganic Solids	>100	--	--
Crystalium	Computational	Elemental Solids	>145	Literatures	(Tran et al., 2016)
Phonondb	Computational	Inorganic Solids	--	MP	

Database	Types	Materials	No. Entries	Data Source	Ref.
TE Design Lab	Computational	Semiconductors	~2701	Literatures	(Gorai et al., 2016)
AIST Research Information Databases	Experimental	General Materials Data	--	Literatures	(Kouchi and Mochimaru, 2005)
American Mineralogist Crystal Structure Database	Experimental	Minerals	2627	Literatures	(Downs and Hall-Wallace, 2003)
ASM Alloy Center Database	Experimental	Alloys	--	Literatures	--
ASM Phase Diagrams	Experimental	Alloys	6200	Literatures	--
CALPHAD databases	Experimental	Alloys	--	Literatures	--
ChemSpider	Experimental	Chemical Materials	99,000,000	Literatures	(Pence and Williams, 2010)
CINDAS High-Performance Alloys Database	Experimental	Alloys	298	Literatures	--
CRC Handbook	Experimental	General Materials Data	--	--	--
CrystMet	Experimental	Metals	70,000	Literatures	(White et al., 2002)
DOE Hydrogen Storage Materials Database	Experimental	General Materials Data	--	Literatures	--
Granta CES Selector	Experimental	Metals, Polymers, Composites, Medical Materials, Coatings, Aerospace Materials	>4000	Literatures	--

Database	Types	Materials	No. Entries	Data Source	Ref.
Handbook of Optical Constants of Solids, Palik	Experimental	General Materials Data	--	Hard-copy sources	(Palik, 1998)
International Glass Database System (INTERGLAD)	Experimental	Glass	350,000	--	--
Knovel	Experimental	General Materials Data	--	Literatures	(Kress-Rogers and Brimelow, 2000)
Matbase	Experimental	General Materials Data	--	Literatures	--
MatDat	Experimental	General Materials Data	>4000	Literatures	--
MatNavi (NIMS)	Experimental	Polymers, Inorganic and Metallic Materials	--	Literatures	(Ogata and Yamazaki, 2012)
MatWeb	Experimental	Carbon, Ceramis, Fluid, Metal, Polymer, Wood and Natural Products	140,000	Literatures	(MatWeb, 1996)
Mindat	Experimental	Minerals, rocks, Meteorites	--	Literatures	--
NanoHUB	Experimental	Nanomaterials	--	Literatures	(Klimeck et al., 2008)
NIST Materials Data Repository (DSpace)	Experimental, Computational	General Materials Data	--	Literatures	--
NIST Interatomic Potentials Repository	Computational	Metals, Semiconductors, Oxides, and Carbon-containing systems	--	Literatures	(Becker et al., 2013, Hale et al., 2018)

Database	Types	Materials	No. Entries	Data Source	Ref.
NIST Standard Reference Database 3 (NIST SRD 3)	Experimental, Computational	Inorganic Solids	210,000	Literatures	--
Open Knowledge Database Of Interatomic Models (Open KIM)	Computational	Molecular	--	--	(Tadmor et al., 2011)
Pauling File	Experimental, Computational	Inorganic Solids	357,612	Literatures	(Villars et al., 2004)
Pearson’s Crystal Data (PCD)	Experimental	Inorganic Solids	350,000	Literatures	(Villars and Cenzual, 2009)
Pearson’s Handbook: Crystallographic Data	Experimental	Intermetallic phases	--	Hard-copy sources	--
Powder Diffraction File (PDF)	Experimental	Inorganic Solids	--	Literatures	(Faber and Fawcett, 2002)
PubChem	Experimental	Molecules	32,000	Literatures	(Kim et al., 2019)
Reaxys	Experimental	Chemical data	>118,000	Literatures, Patents	(Goodman, 2009)
SciFinder	Experimental	Chemical data	47,000,000	Literatures, Patents	(Gabrielson, 2018)
SciGlass	Experimental	Glasses	360,293	Literatures, Patents	--
SpringerMaterials	Experimental	General Materials Data	--	Literatures, Patents	--
Total Materia	Experimental	Metallic Materials Data	350,000	Literatures, Patents	--
UCSB-MRL thermoelectric database	Experimental	Thermoelectric Materials	18,000	Literatures	(Gaultois et al., 2013)

Database	Types	Materials	No. Entries	Data Source	Ref.
NRELMatDB	Computational	Inorganic Solids	--	Literatures, Patents	(Stevanović et al., 2012)
Metallurgical Thermochemistry, Kubaschewski	Experimental	Thermoelectric Materials	--	Hard-copy sources	--
3D Materials Atlas	Experimental	General Materials Data	--	--	--
Inorganic Material Database (AtomWork)	Experimental	Inorganic Solids, Metals	82,000	Literatures,	--
Mineralogy Database	Experimental	Minerals	4714	Literatures	(Barthelmy, 2007)
CSD Teaching Database	Experimental	Organic Materials	>750	CSD	--
Database of Zeolite Structures	Computational	zeolites	--	Literatures, Hypothesis	(Baerlocher, 2008)
RCSB Protein Data Bank	Experimental	biological macromolecular structures	>173,005	Literatures,	

2.3.1 Computational Databases

2.3.1.1. Open Quantum Materials Database (OQMD)

The OQMD (Saal et al., 2013, Kirklin et al., 2015a) is a DFT database containing calculated thermodynamic and structural properties of 815,654 materials, developed by Chris Wolverton's group at Northwestern University. The OQMD contains approximately 300,000 calculated structures, mainly from two sources: ~10% from the Inorganic Crystal Structure Database (ICSD) (Belsky et al., 2002) and ~90% from the iteration of many chemistries for some of simple prototypes. For the crystal structures in the ICSD, ~44,000 structures are calculable, of which the OQMD contains DFT calculations of 32,559 ICSD structures. The remaining calculable ICSD structures are continually being calculated and added to the OQMD. Additionally, 259,511 hypothetical compounds have been generated based on 16 elemental prototypes, 12 binary prototypes with their compositions, and three ternary prototypes with their compositions (Emery et al., 2016, Wang et al., 2018, Kirklin et al., 2015a). Moreover, OQMD provides a qhull algorithm for establishing DFT ground-state phase diagrams at ambient (high) pressure and Grand Canonical Linear Programming (GCLP) to analyze the complex ground state thermodynamics of metal hydrides (R. Akbarzadeh et al., 2007, Hegde et al., 2020a, Amsler et al., 2018). The OQMD provides the entirety of the underlying database to be freely downloaded at oqmd.org/download/, in addition to a Representational State Transfer (REST) Application Programming Interface (RESTful API) for programmatic access, which allows scientists and engineers to use simple Hyper Text Transfer Protocol (HTTP) requests to access all living data (Hegde et al., 2020b). For instance, Hu et al. (Hu et al., 2020) used the Wasserstein GAN model in conjunction with the OQMD database to generate novel hypothetical materials (**Figure 2.3**a). Fung et al. (Fung et al., 2021) predicted adsorption energies using the density of state data from the OQMD and Materials Project (MP) database combined with CNNs, targeting the accelerated discovery of catalytic materials (**Figure 2.3**b).

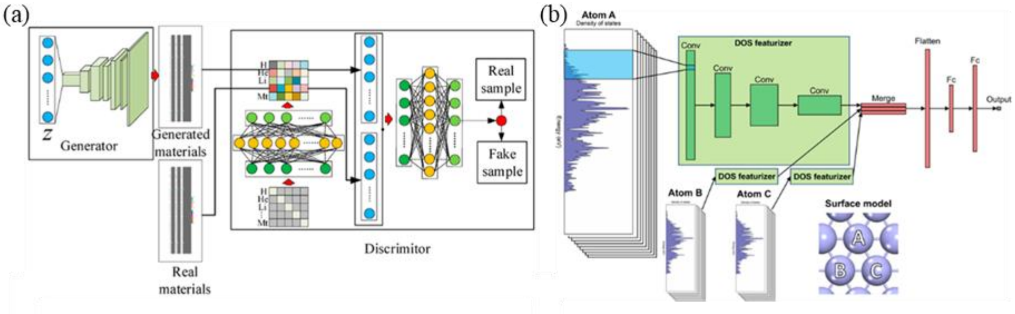

Figure 2.3 (a) The Wasserstein Generative Adversarial Network (WGAN) model using the OQMD database to generate novel hypothetical materials. Reproduced with permission **(Hu et al., 2020)**. Copyright 2020, MDPI Publications. (b) Using the density of state data from the OQMD and MP database by convolutional neural networks (CNNs) for the accelerated discovery of catalytic materials. Reproduced with permission **(Fung et al., 2021)**. Copyright 2021, Springer Nature Publications.

2.3.1.2. Materials Project (MP)

The Materials Project (MP) provides open web-based access to computed information on known and predicted materials to inspire and design novel materials (Jain et al., 2013). Most of the MP data pertain to chemical compounds in the ICSD

(Belsky et al., 2002, Bergerhoff et al., 1983). A significant challenge is the generation of novel compositions and compounds to perform calculations (Jain et al., 2013) even though there already exist multiple algorithmic, e.g., Optimization-based (Bergerhoff et al., 1983, Dudiy and Zunger, 2006b, Oganov and Glass, 2006, d'Avezac et al., 2012), and data-driven approaches (Hautier et al., 2011b, Hautier et al., 2010, Fischer et al., 2006) to tackle this problem. For materials included in the MP database, selected properties such as total energies (Jain et al., 2011a), electronic structure (Jain et al., 2011a), thermodynamic equations of state parameters (Latimer et al., 2018), phonons (Petretto et al., 2018), piezoelectricity (de Jong et al., 2015b), elasticity (de Jong et al., 2015a), dielectricity (Petousis et al., 2017), and thermoelectricity (Chen et al., 2016) have been calculated and included. In addition, MP includes apps to visualize phase diagrams (Jain et al., 2011c, Ong et al., 2008) and Pourbaix diagrams (Persson et al., 2012). Several other convenient applications such as Materials Explorer (de Jong et al., 2015a, de Jong et al., 2015b), Battery Explorer (Zhou et al., 2004), Reaction Explorer(Jain et al., 2011c), Structure Predictor (Hautier et al., 2011a), Crystal Toolkit (Ong et al., 2013), Nanoporous Materials Explorer (Ong et al., 2013), Molecules Explorer (Qu et al., 2015, Cheng et al., 2015b), Redox Flow Battery Dashboard (Dmello et al., 2016), X-Ray Absorption Spectra (XAS) (Mathew et al., 2018), Interface Reactions (Richards et al., 2016), and Synthesis Description Explorer (Kim et al., 2017) have also been included in MP. Both Python Materials Genomics (pymatgen) (Ong et al., 2013) and FireWorks (Jain et al., 2015) open-source libraries are available for materials analysis and high-throughput application. Note that all the underlying data for the calculations of ~530,000 nanoporous materials and 130,000 inorganic compounds are accessible via the Materials API (Ong et al., 2015) based on REST principles. Although the MP database was originally developed to predict the adsorption energy of the catalytic materials (Fung et al., 2021), it has supported many other applications such as the accelerated discovery of stable spinel material (Wang et al., 2021c) and carbon dioxide electrocatalysis (Zhong et al., 2020b). Additionally, the MP and OQMD databases' magnetization properties are nearly comparable (Hegde et al., 2020b).

2.3.2 Experimental Database

2.3.2.1. ICSD

The ICSD (Belsky et al., 2002) is the world's largest database of fully evaluated and published data containing inorganic crystal structures primarily derived from experimental results. Currently, the ICSD (Zagorac et al., 2019) has more than 232,012 entries, including ~2,902 elemental crystal, ~38,506 binary compounds, ~73,048 ternary compounds, and ~73,688 quarternary and quintenary compounds. The database is updated twice a year based on over 80 leading scientific journals and more than 1,400 other scientific journals; data sources have been expanded to include experimental inorganic structures, experimental metal-organic structures, and theoretical inorganic structures.

To be included in the database, the structure must be fully characterized. For instance, atomic coordinates can be determined or derived from known structure types, and the composition must be fully specified. Typical entries include chemical names, formulas, unit cells, space groups, complete atomic parameters (including atomic displacement parameters if available), site occupancy, titles, authors, and literature citations. For published data, many items (such as Wykov sequences, molecular formulas, weights, ANX formulas, and mineral groups) are introduced through expert evaluation or generated by computer programs.

The keyword-based search in the ICSD can be specified in terms of physical properties, analytical methods used, and technical application. Note that the ISCD data has been used to indicate promising novel applications of new ionic conductors, solar cell adsorbers, advanced ceramic materials, nature’s missing compounds, and structural relations between the crystalline compounds. In addition, ICSD data have been included in almost all other computational databases, such as OQMD, MP, and AFLOW. Organic and inorganic compounds are two of the main categories of chemical materials. Thus, we introduce the Cambridge Structural Database (CSD) for organic materials.

2.3.2.2. CSD

The CSD (Groom et al., 2016b) is the world's largest and most comprehensive collection for small-molecule organic and organometallic crystal structures, containing over one million structures from X-ray and neutron diffraction analyses. For comprehensive coverage of single-crystal data, cell parameters and all available data are included even if no coordinates are available. Similarly, powder structures are available from the International Centre for Diffraction Data (ICDD) (Kabekkodu et al., 2002) even though the coordination information is missing. Note that there is a slight overlap between the CSD and the ICSD in the area of molecular inorganics, but that purely inorganic structure is not contained in the CSD.

The CSD database has provides data in two distinct ways. The first is pertains only to structural aggregation and standardization, making it easier to access individual entries. The second is based on further study of data collection and the discovery of new knowledge transcending the results from individual experiments. Python-based API (Cole et al., 2019) has also been introduced to enable end-users to query CSD using customized script. Accessing data via scripts in conjunction with other packages such as RDKit (Coley et al., 2019) is very useful for more advanced structural data analysis. For instance, users will be able to use ML more conveniently in conjunction with APIs for solvate prediction, implementing fragment pocket analysis using structural information, and supporting crystal (co-crystal) structure prediction (Connor et al., 2019). More detailed insights could be developed as the scale of data increased, having a profound impact across the scientific community with specific consequences for drug discovery and development (Cole et al., 2019). However, the ICSD and CSD have paid licenses (as shown in **Error! Reference source not found.**), affecting a number of institutions or members who cannot access the data.

2.4 Overview of Key Descriptors Bridging Data Intensive Discoveries and Experimental Strategies for Material Science and Chemical Engineering

The key premise of the ML framework is that learning can be viewed as a reasonable model to explain the observed data (Ghahramani, 2015). Descriptors are the carriers of information exchange between humans and machines. In the context of materials science, they deliver information about molecular properties to machines in digital form. Key to the efficient use of ML in the field of chemical materials is the "descriptor selection" tool, which takes the entire descriptor set as an input, or combines it into a new reduced, but more reliable, descriptor set through correlation analysis while providing a mapping to a Key Performance Indicator (KPI) fingerprint (Pankajakshan et al., 2017). In this section, the strategy of transforming material data to ML through descriptors is introduced; descriptors can be divided into five main types: constitutional descriptors (Zhong et al., 2020a, Zhu et al., 2019a, Wexler et al., 2018b, Friederich et al., 2020, Sun et al., 2020b, Davies et al., 2019, Lu et al., 2018, Ward et al., 2016b, Hu et al., 2019, Ulissi et al., 2017, Ruck et al., 2020); geometric descriptors (Hu et al., 2019, Wexler et al., 2018b, Sun et al., 2020b, Ge et al., 2020b, Pankajakshan et al., 2017, Ma et al., 2015b, Ruck et al., 2020, Ward et al., 2016b, Friederich et al., 2020, Davies et al., 2019, Lu et al., 2018);

quantum chemistry descriptors (Ward et al., 2016b, Zhu et al., 2019a, Artrith et al., 2020, Friederich et al., 2020, Sun et al., 2020b, Zhong et al., 2020a, Lu et al., 2018, Davies et al., 2019, Pankajakshan et al., 2017, Ma et al., 2015b, Bai et al., 2019, Sahu et al., 2018b, Ge et al., 2020b, Fathinia et al., 2016, Ma et al., 2015a, Peterson and Nørskov, 2012, Hussain et al., 2018, Bagger et al., 2017, Ulissi et al., 2017, Ruck et al., 2020, Kang et al., 2018b, Zhang et al., 2020a, Hu et al., 2019, Wexler et al., 2018b, Back et al., 2019, Hammer and Nørskov, 2000, Masood et al., 2019); electrostatic descriptors (Pankajakshan et al., 2017, Ma et al., 2015a, Kang et al., 2018b, Sun et al., 2020b, Hammer and Nørskov, 2000, Ma et al., 2015b, Lu et al., 2018, Zhu et al., 2019a, Sahu et al., 2018b); combinational descriptors. These will be elaborated upon in the relevant subsections. Finally, we describe some of the extension packages of descriptors in the field of AI for materials science.

2.4.1 Information Bridging: from Chemical Structures to ML Models

2.4.1.1. Descriptor Importance

The selection of descriptors directly determines the feasibility of introducing ML to solve the posed question. When the scientific connection between the descriptor and the actuation mechanism is not clear, the causal relationship of the learned descriptor-attribute relationship is uncertain. Therefore, the reliable prediction, identification, and scientific development of new materials are called into question. Analyzing the problem and defining a suitable descriptor is a meaningful and necessary step (Ghiringhelli et al., 2015).

A number of studies have emphasized the importance of material descriptors in accelerating the calculation of material properties or material design. Ghiringhelli, L. M. et al. (Ghiringhelli et al., 2015) detail the required characteristics of a set of descriptors: the calculation of descriptors should not be as intensive as that of KPIs; they uniquely characterize materials and the basic processes which pertain to properties; very different materials should be characterized by very different descriptor values (and vice versa); their size should be as small as possible. Sahu et al. (Sahu et al., 2018a), utilized 13 microscopic properties of organic materials as descriptors to build a PCE prediction model. The results indicated that such descriptors can effectively be applied in the context of promising high-throughput virtual screening of new donor molecules for efficient organic photovoltaics. Implementing descriptors with appropriate features plays an important role in accelerating outcomes of material design, or the study of material characteristics.

2.4.1.2. Bridging and Transferring Process

Data bridging and transfer processes often introduce uncertainty to ML predictions. The evaluation of this uncertainty indicates whether the required prediction accuracy has been satisfied. The MGI (Jain et al., 2013) aims to capture, manage, and utilize material structure/property information on a large scale to enable the rapid, cost-effective, and efficient development of new materials with predictable properties. Although the use of such "genome" methods (to promote attribute prediction, virtual design, and material discovery) is relatively new, the concepts driving the development of materials informatics are firmly grounded previous lessons learned from the fields of chemoinformatics and bioinformatics.

The management and utilization of material structure/attribute information have increased the significance of cheminformatics to ML; a number of new methods have emerged for information and data conversion. Behler describes some of the ways in which chemoinformatics and ML methods have been adapted for materials science and engineering applications, including methodologies to create, verify, and use material quantitative structure and property relationship (MQSPR) models (Behler, 2011). Friederich et al. (Friederich et al., 2020) used full autocorrelation (FA) functions to transfer the features of chemical complexes. Combining DFT and ML methods, the obtained predictions of reactivity within large chemical spaces containing thousands of complexes. Affordable descriptors were transferred as functions and demonstrated as fingerprints for each complex by considering a specified product of atomic properties (PiPj) calculated in terms of all atoms. Compound compositions were guided by the properties of atoms i and j (**Figure** 2.4a). These atomic properties include electronegativity, atomic number, identity topology, and size. Each descriptor is multiplied as a function of Diracδδ to encode the structure and properties of the compound.

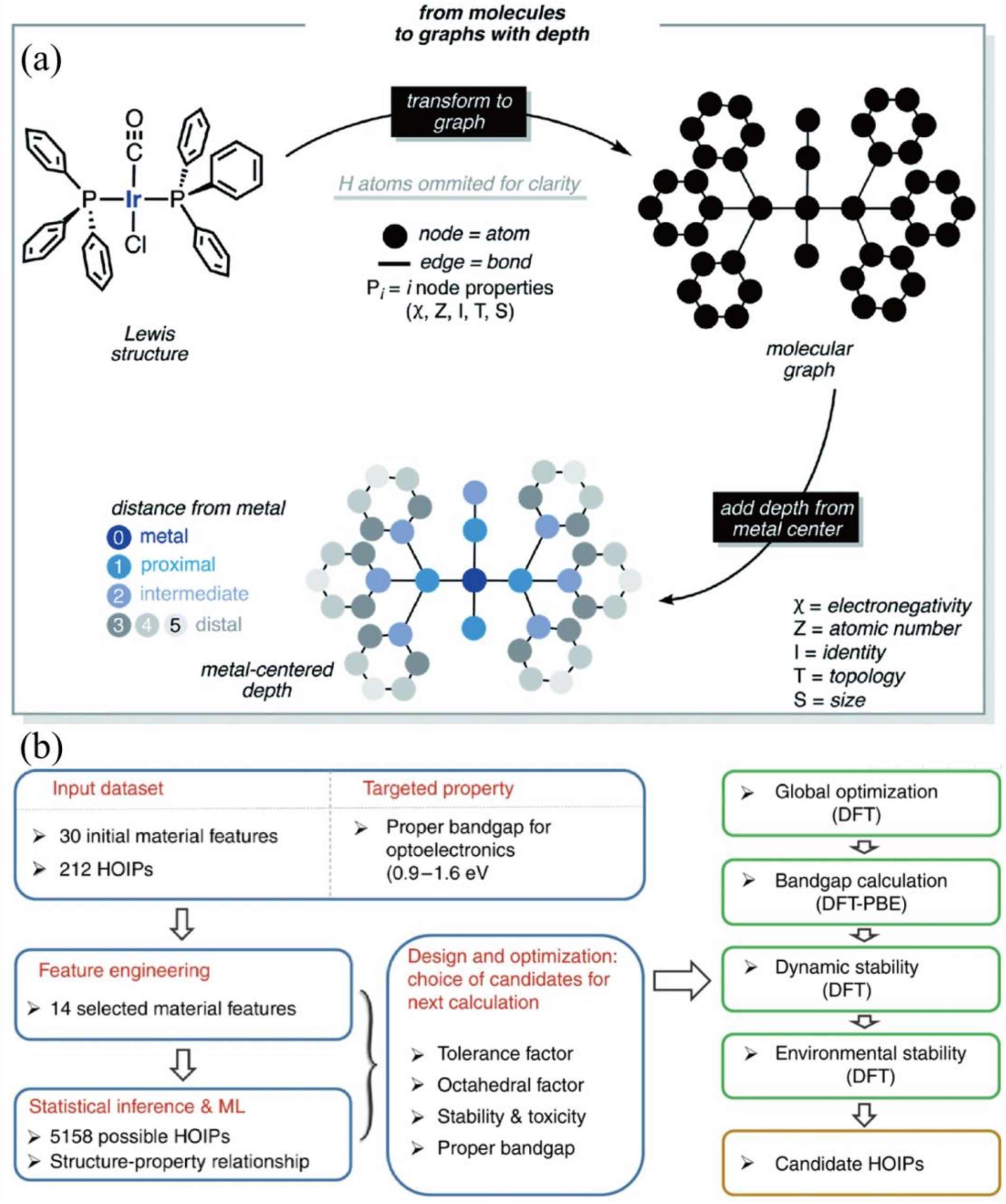
(a)
from molecules to graphs with depth
transform to graph
H atoms ommited for clarity
node = atom
edge = bond
P_i = *i* node properties (χ, Z, I, T, S)
Lewis structure
molecular graph
O
C
P
Ir
P
Cl
add depth from metal center
distance from metal
0 metal
1 proximal
2 intermediate
3 4 5 distal
metal-centered depth
χ = electronegativity
Z = atomic number
I = identity
T = topology
S = size
(b)
Input dataset
30 initial material features
212 HOIPs
Targeted property
Proper bandgap for optoelectronics (0.9−1.6 eV
Feature engineering
14 selected material features
Statistical inference & ML
5158 possible HOIPs
Structure-property relationship
Design and optimization: choice of candidates for next calculation
Tolerance factor
Octahedral factor
Stability & toxicity
Proper bandgap
Global optimization (DFT)
Bandgap calculation (DFT-PBE)
Dynamic stability (DFT)
Environmental stability (DFT)
Candidate HOIPs

Figure 2.4 (a) Schematic diagram of molecular graph in the calculation of autocorrelation and deltametric functions. Reproduced with permission.(Friederich et al., 2020) Copyright 2020, RSC Publications. (b) The schematic diagram of designing lead-free HOIP based on ML combined with DFT. The blue box represents the process of screening through the ML algorithm from the HOIP database. The green box indicates the use of DFT to calculate the electronic performance and stability evaluation of the candidate. Reproduced with permission.(Lu et al., 2018) Copyright 2018, Springer Nature Publications.

The selection of the descriptor, removal of redundant features, and establishment of relationships are crucial to the process of transferring information. As shown in **Figure** 2.4b, the prediction strategy integrates input HOIP data with the ML algorithm and DFT calculation (Lu et al., 2018). Based on the ML program, an input HOIP dataset is established; each input item is described by a signature that is used to train and test the ML model. Element design analysis is required as a prerequisite to remove redundant features and establish structure-attribute relationships. After the input feature set is fixed, grid search technology and 5-fold CV are utilized to select the best descriptor. The network is subsequently trained to predict the electronic performance and stability of the HOIPs. In this work, the 14 most important descriptors were sorted and selected to collectively describe HOIPs in the chemical space. These descriptors included structural features and elemental properties of A-, B-, and X-site ions. Based on linear correlations for features analysis, redundant or irrelevant features could to improve the accuracy and efficiency of the ML model and achieve accurate predictions based on relatively small training datasets. This work successfully predicted the bandgaps of thousands of HOIPs by using the trained ML model. The evaluation of the bridging and transfer process of characteristic information represented by the descriptor is key to successful ML model predictions. n the process of information transfer, it is also essential to provide more accurate descriptors without losing the original information characteristics. Some descriptors, though assigned a large weight, do not contribute to reliable model predictions (i.e. the phenomenon of over-egging the pudding).

2.4.1.3. Properties of Ideal Descriptors

Descriptors that can train predictive models to adapt to target attributes are highly desirable. **Figure 2.5**a presents a representative graphical summary of the workflow of the descriptor design, which is usually applicable throughout the development of a novel strategy. This summary represents a general processing method suitable for any application involving the main dataset, descriptor, training model, etc. Traditional methods rely on chemical intuition to determine the key descriptors for a specific application and develop a relationship which best represents observed material properties. It is more desirable, however, to automate the generation of interesting chemical insights through a rational design approach which does not rely chemical intuition.

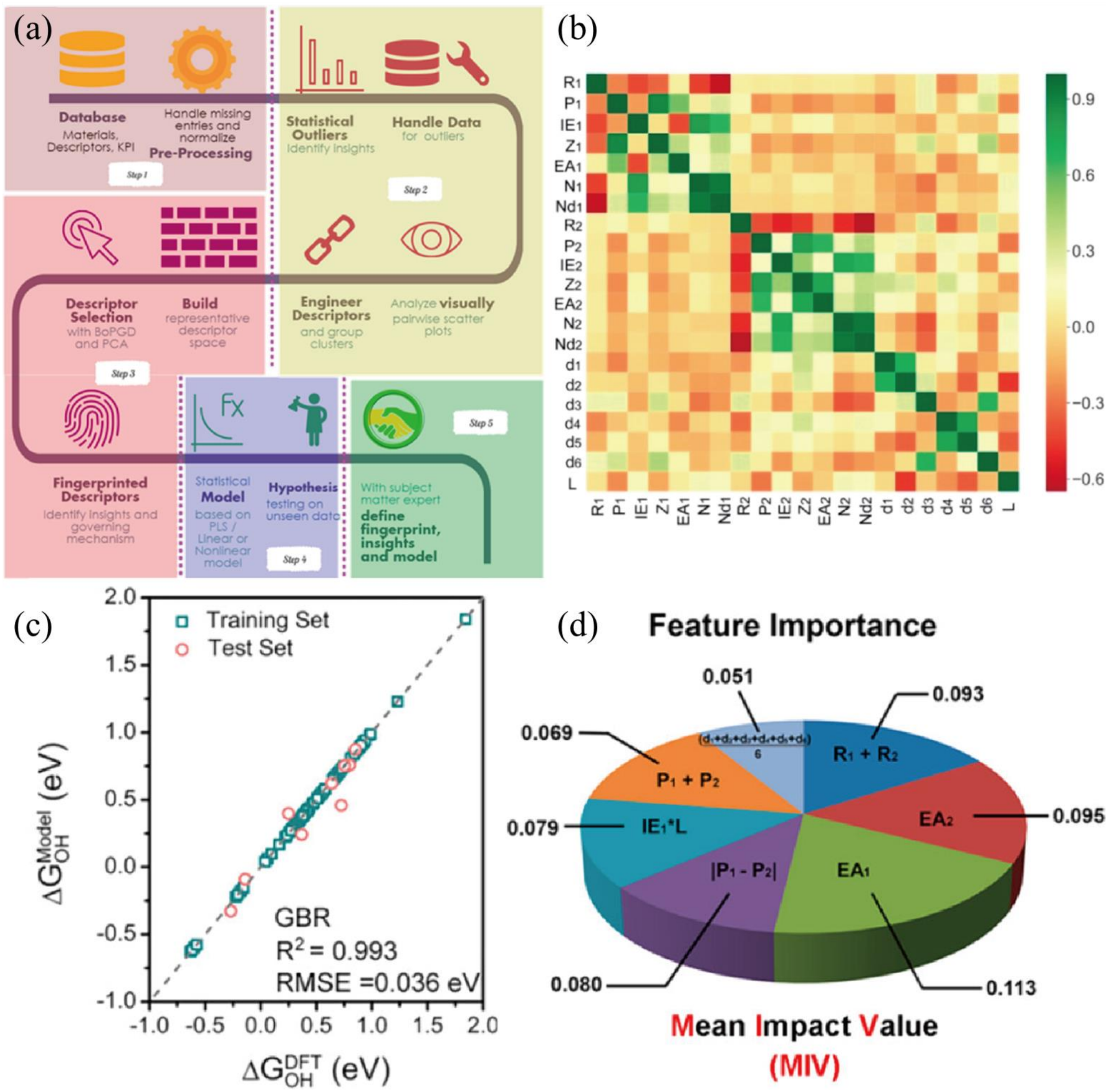

Figure 2.5 (a) The relationship between data, descriptors, and models. Reproduced with permission (Pankajakshan et al., 2017). Copyright 2017, ACS Publications. It involves the following steps: preprocessing, data analysis, fingerprinting descriptors, statistical model or linear/nonlinear model building and validations, and insights from a subject matter expert. (b) Heat map of the Pearson correlation coefficient matrix among the selected features for DMSCs. (c) Comparison of DFT-computed ΔG_{OH*} values with those predicted by GBR algorithm. (d) Feature importance based

on the Mean Impact Value (MIV). (b-d) Reproduced with permission (Zhu et al., 2019b). Copyright 2019, ACS Publications.

Regression fitting, correlation coefficient statistics, dataset partitioning, the establishment of new functions, and other methods have been widely applied to locate and rank ideal descriptors which correspond to the most relevant performance features. Meredig and Wolverton (Meredig and Wolverton, 2014) introduced a "cluster ranking model" (CRM) framework to identify unique descriptors that can predict the properties of new dopants. They used the X-means algorithm to cluster various dopants together, followed by regression fitting to rank the descriptors, ultimately utilizing the unique descriptors to model the behavior within each cluster. The existence of clusters in various sample datasets (four dopant clusters were present in this study) improves the effectiveness of the method. Given that all descriptors are ranked by using a regression model, they must necessarily fit to the prediction model of the target attribute. Selected descriptors are those that can best predict the target attributes; they are not necessarily indicative of the phenomenological mechanism. Ward et al (Ward et al., 2016a). generated an extensible set of attributes that can be used for materials with any number of constituent elements. This set of attributes can broadly capture enough diverse physical/chemical properties of materials to form the basis of accurate predictive models. The group used a total of 145 attribute sets, including stoichiometric attributes, elemental property statistics, electronic structure attributes, and ionic compound attributes. They proved that these attributes are sufficient for describing various properties, while also proposing a novel method to divide the dataset into groups of similar materials to improve prediction accuracy. This work demonstrated the applicability of this novel method to the prediction of various physical properties of crystalline and amorphous materials. Zhu et al.(Zhu et al., 2019b) employed DFT calculations, with the assistance of ML, to screen highly efficient dual-metal-site catalysts (DMSCs) for oxygen reduction reaction (ORR). They evaluated the correlation coefficient for selected DMSC features, as shown in Figure 18b. The performance of the ML model can be significantly improved by selecting features that are independent from one another (i.e., not redundant), based on an analysis of linear correlations of several features. The speed at which ML-based approaches can be used to arrive at valuable material property insights, including the identification of descriptors, has significantly improved in recent years. To obtain accurate descriptor relevant to the catalytic activity of DMSC, this work reported the seven characteristics which were deemed most relevant to the catalytic performance of DMSCs in terms of Mean Impact Value (MIV) (Figure 18d). These characteristics include: the electron affinity between two metal atoms; Van der Waals radius; Pauling electronegativity difference; the product of ionization energy and the distance between two metal atoms; the relationship between Pauling electronegativity and atomic distance.

2.4.2 Categories of Descriptors

In recent years, a large number of articles have demonstrated the importance of material descriptors in accelerating the discovery and design of novel materials. When identifying descriptors which are compatible with ML methods for material discovery, the initial set of descriptors should generally be broad/diverse. Both the choice of fingerprint descriptors and the methods employed to discover/estimate unique mappings are critical, especially when dealing with small datasets. From the perspective of ML, fingerprint descriptors are a subset (or offspring) of a superset of parent descriptors; they are unique to attributes and materials. The dimensionality or cardinality of the descriptor should be kept as low as possible, while the original descriptor space should be sufficient. This mathematical mapping is also unique to the construction model that maps fingerprint descriptors to attributes or KPIs (Pankajakshan et al., 2017). The key descriptors used in recent studies for training models in materials science are summarized in **Table 2.2** and are detailed further in subsequent sections.

Table 2.2 Key descriptors used for the model training in material science and chemical engineering

Description	Class	Ref
Atomic Number	Constitutional	(Zhong et al., 2020a, Zhu et al., 2019a, Wexler et al., 2018b, Friederich et al., 2020, Sun et al., 2020b, Davies et al., 2019)
Atomic Weight	Constitutional	(Ward et al., 2016b, Wexler et al., 2018b, Davies et al., 2019)
Numbers of and orbital electron	Constitutional	(Lu et al., 2018, Zhu et al., 2019a, Sun et al., 2020b)
Numbers of and valence electron	Constitutional	(Ward et al., 2016b, Zhu et al., 2019a, Wexler et al., 2018b, Sun et al., 2020b,

Description	Class	Ref
		Davies et al., 2019) (Lu et al., 2018)
Mendeleev number	Constitutional	(Ward et al., 2016b, Davies et al., 2019)
Melting Temperature		(Ward et al., 2016b, Davies et al., 2019)
Bond Number	Constitutional	(Hu et al., 2019)
Space Group Number	Constitutional	(Ward et al., 2016b, Davies et al., 2019)
the number of atoms of that element coordinated	Constitutional	(Zhong et al., 2020a, Ulissi et al., 2017, Ruck et al., 2020, Friederich et al., 2020)
Pauling electronegativity	Quantum chemical	(Zhong et al., 2020a) (Lu et al., 2018, Davies et al., 2019) (Ward et al., 2016b, Zhu et al., 2019a, Artrith et al., 2020, Friederich et al., 2020, Sun et al., 2020b)
The median monometallic adsorption energy	Quantum chemical	(Zhong et al., 2020a)
Ionic Charge	Quantum chemical	(Lu et al., 2018)
Electron Affinity	Quantum chemical	(Lu et al., 2018, Pankajakshan et al., 2017, Ma et al., 2015b, Zhu et al., 2019a, Bai et al., 2019, Sun et al., 2020b)
Ionization Energy	Quantum chemical	(Lu et al., 2018, Pankajakshan et al., 2017, Ma et al., 2015b, Zhu et al., 2019a, Bai et al., 2019, Sahu et al., 2018b)
The highest occupied molecular orbital	Quantum chemical	(Lu et al., 2018, Davies et al., 2019, Sahu et al., 2018b)

Description	Class	Ref
The lowest unoccupied molecular orbital	Quantum chemical	(Lu et al., 2018, Davies et al., 2019, Sahu et al., 2018b)
Bandgap Energy	Quantum chemical	(Ward et al., 2016b, Ge et al., 2020b, Fathinia et al., 2016, Davies et al., 2019)
Work Function	Quantum chemical	(Pankajakshan et al., 2017, Ma et al., 2015b)
Binding Energy	Quantum chemical	(Pankajakshan et al., 2017, Ma et al., 2015a, Peterson and Nørskov, 2012, Hussain et al., 2018, Bagger et al., 2017, Ulissi et al., 2017, Artrith et al., 2020, Sahu et al., 2018b)
Adsorption Energy	Quantum chemical	(Pankajakshan et al., 2017, Ma et al., 2015a, Peterson and Nørskov, 2012, Hussain et al., 2018, Bagger et al., 2017, Ulissi et al., 2017, Ruck et al., 2020, Kang et al., 2018b, Artrith et al., 2020, Zhang et al., 2020a, Hu et al., 2019, Wexler et al., 2018b, Sun et al., 2020b)
Local Pauling electronegativity	Quantum chemical	(Pankajakshan et al., 2017, Ma et al., 2015b)
Cohesive energy	Quantum chemical	(Sun et al., 2020b)
Density of states	Quantum chemical	(Hammer and Nørskov, 2000, Masood et al., 2019)
Partial Density of states	Quantum chemical	(Hu et al., 2019)
Bader Charge Transfer	Quantum chemical	(Sun et al., 2020b)

Description	Class	Ref
Fermi Energy	Quantum chemical	(Pankajakshan et al., 2017, Hammer and Nørskov, 2000, Masood et al., 2019)
Gibbs Free Energy	Quantum chemical	(Back et al., 2019, Hu et al., 2019, Wexler et al., 2018b, Sun et al., 2020b)
Surface Energy Density	Quantum chemical	
Total energy of surface slab obtained	Quantum chemical	(Back et al., 2019)
Bulk energy per atom	Quantum chemical	(Back et al., 2019)
Over potential	Quantum chemical	(Back et al., 2019, Hoar et al., 2020b, Hammer and Nørskov, 2000, Ge et al., 2020b)
Current density		(Hoar et al., 2020b)
Activation energy	Quantum chemical	(Artrith et al., 2020, Friederich et al., 2020, Hammer and Nørskov, 2000)
Transition-state energy	Quantum chemical	(Artrith et al., 2020, Friederich et al., 2020, Hammer and Nørskov, 2000, Sahu et al., 2018b)
Atomic nearest-neighbor distances		(Artrith et al., 2020)
Optical gap energy	Quantum chemical	(Bai et al., 2019, Sahu et al., 2018b)
Width of a band	Electrostatic	(Pankajakshan et al., 2017, Ma et al., 2015a)
Centre of a band	Electrostatic	(Pankajakshan et al., 2017, Ma et al., 2015a, Kang et al., 2018b, Sun et al., 2020b, Hammer and Nørskov, 2000)

Description	Class	Ref
Skewness of a band	Electrostatic	(Pankajakshan et al., 2017, Ma et al., 2015a)
Kurtosis of a band	Electrostatic	(Pankajakshan et al., 2017, Ma et al., 2015a)
Filling of a band	Electrostatic	(Pankajakshan et al., 2017, Ma et al., 2015a)
Spatial Extent of -orbitals	Electrostatic	(Pankajakshan et al., 2017, Ma et al., 2015b)
Adsorbate-metal coupling matrix element	Electrostatic	(Pankajakshan et al., 2017, Ma et al., 2015b, Hammer and Nørskov, 2000)
Metal -metal coupling matrix element	Electrostatic	(Hammer and Nørskov, 2000)
Partial distribution function	Geometric	(Ruck et al., 2020)
Polarizability	Electrostatic	(Lu et al., 2018, Sahu et al., 2018b)
First ionization potential	Electrostatic	(Lu et al., 2018, Zhu et al., 2019a, Sun et al., 2020b)
Magnetic Moment	Electrostatic	(Ward et al., 2016b)
Bond Length Position	Geometric	(Hu et al., 2019)
Atomic Identity	Geometric	(Friederich et al., 2020)
Optical Transmittance		(Bai et al., 2019)
Lattice parameters	Geometric	(Sun et al., 2020b)
Molar Ratio		(Sun et al., 2020b)
Dipole moment	Electrostatic	(Sahu et al., 2018b)
Atomic Radius	Geometric	(Pankajakshan et al., 2017, Ma et al., 2015b, Wexler et al., 2018b, Sun et al., 2020b)
Rotational angles	Geometric	(Ge et al., 2020b)
Distance between two layers	Geometric	(Ge et al., 2020b)
Bond Length	Geometric	(Hu et al., 2019, Wexler et al., 2018b, Sun et al., 2020b, Ge et al., 2020b)

Description	Class	Ref
Bond Angle	Geometric	(Wexler et al., 2018b)
Distance to alloy atoms	Geometric	(Ruck et al., 2020)
Estimation for the interatomic distance using Vegard's law	Geometric	(Ruck et al., 2020)
Covalent Radius	Geometric	(Ward et al., 2016b, Friederich et al., 2020, Davies et al., 2019)
Specific Volume	Geometric	(Ward et al., 2016b, Davies et al., 2019)
Van der Waals radii	Geometric	(Zhu et al., 2019a)
Tolerance Factor	Geometric	(Lu et al., 2018)
Octahedral Factor	Geometric	(Lu et al., 2018)
Iron Radii	Geometric	(Lu et al., 2018)
Sum of the of and orbital radii	Geometric	(Lu et al., 2018)
Atomic Radius	Geometric	(Pankajakshan et al., 2017, Ma et al., 2015b, Wexler et al., 2018b, Sun et al., 2020b)
Cutoff radius		(Zhang et al., 2020a, Jäger et al., 2018)
Bond distance		(Zhang et al., 2020a, Friederich et al., 2020)
Atom pair distance		(Zhang et al., 2020a)

2.5 Applications of Data-Driven Innovative Materials

The success of a large number of ML applications in materials science and chemical engineering has preliminarily demonstrated the capability of data-driven approaches in the discovery of innovative materials. By appropriately integrating ML techniques, databases, descriptors, target material properties, and engineering parameters, predictions for the focused design of materials and chemical processes can be efficiently and accurately made. Such approaches represent a synergy between materials science, chemical engineering, computer science, and mathematics. Recent advances in the applications of such synergies to the development of chemical innovation for energy conversion and storage (Chen et al., 2020a, Wexler et al., 2018a, Zhang et al., 2020a, Back et al., 2019, Tran and Ulissi, 2018, Xu et al., 2020, Ge et al., 2020a), environmental decontamination (Dondapati and Chen, 2020), flexible electronics (Zhang et al., 2020b), optoelectronics (Saeki, 2020) superconductors (Stanev et al., 2018), metallic glasses (Ward et al., 2016a), and magnet materials are investigated.

An overview of the applications of data-driven, innovative material discovery and chemical engineering is represented in **Table 2.3**. The integration of ML in the development of materials for key electrochemical reactions such as the oxygen reduction reaction (ORR) (Rück et al., 2020, Zhu et al., 2019b, Kang et al., 2018a, Groenenboom et al., 2020), carbon dioxide reduction reaction (CRR) (Zhong et al., 2020b, Ulissi et al., 2017, Batchelor et al., 2019), hydrogen evolution reaction (HER) (Wexler et al., 2018a, Zhang et al., 2020a), and oxygen evolution reaction (OER) (Back et al., 2019, Xu et al., 2020) is vital for advancing green chemical engineering. These reactions, along with the optimization of battery technologies (Sendek et al., 2017, Ahmad et al., 2018), are fundamental to sustainable energy systems, aiming to reduce reliance on fossil fuels and minimize environmental impact. A specific overview about Some typical examples will be discussed in the subsequent sections. cases where ML has been successfully applied will be discussed in detail, followed by an overall future outlook.

Table 2.3 Data-driven innovative applications for material design and chemical processes

Applications	Materials/Processes	Target Properties	ML Model/Algorithms	Data Source	Most Related Descriptors	Ref.
HER	Ni$_3$P$_2$(0001) of Ni$_2$P	Adsorption free energy of H* (ΔGH)	RRFs	DFT computation	Ni-Ni bond length Ni-Ni-Ni bond angle Hollow site area Hollow site perimeter	(Wexler et al., 2018a)
	Amorphous Ni$_2$P	Frozen adsorption energy (Efrozen) Relax adsorption energy (Erelax)	ANN GB DT GA	DFT computation	Bond length Symmetry functions	(Zhang et al., 2020a)
OER	IrO2 and IrO3 Polymorphs	Biding free energy (ΔG) for coverage calculations Biding free energy (ΔG) for OER calculations	CNN	DFT computation Material Project	Atomic structures	(Back et al., 2019)
	Doped RuO2 and IrO2	Identify new descriptors for calculation of adsorption enthalpy of O* (EO*)	SISSO	DFT computation	SISSO Features Width of the d-band Charge transfer energy Filling of the d-band Kurtosis of the d-band	(Xu et al., 2020)
OWS	Transition Metal Dichalcogenides (TMDC): MoS2, WS2, WSe2, MoSe2, MoTe2, and WTe2	HER Overpotentials (ηHER) OER Overpotentials (ηOER)	LASSO	DFT computation	Cosine of the rotational angle The distance between two secondary parts The average mx2 bond length The bandgap ratio of the two components	(Ge et al., 2020a)
PVs	Lead-free hybrid organic-inorganic perovskites	Bandgap	GBR	ICSD	Tolerance factor Number of ionic charges Octahedral factor p-orbital electron	(Lu et al., 2018)

Applications	Materials/Processes	Target Properties	ML Model/Algorithms	Data Source	Most Related Descriptors	Ref.
	Small molecule organic photovoltaic materials	Power conversion efficiency (PCE)	LR kNN ANN RF GBRT	Experiment data DFT computation	Hole–electron binding energy in donor molecules The reorganization energy for holes in donor molecules The unsaturated atom number in the main conjugation path of donor molecules Polarizability of donor molecules	(Sahu et al., 2018a)
	Organic photovoltaics materials	PCE	ANN kNN GBRT	Experiment data DFT computation	Hole–electron binding energy in donor molecules The reorganization energy for holes in donor molecules The unsaturated atom number in the main conjugation path of donor molecules The number of hetero atoms	(Sahu et al., 2019)
	Organic photovoltaics materials	PCE Open circuit voltage (VOC) Short circuit current (JSC)	kNN KRR	DFT computation Literature data	HOMO energy for the donor LUMO energy for the donor LUMO energy for the acceptor The total internal reorganisation energy Daylight fingerprint Morgan fingerprint	(Padula et al., 2019)
	Metal oxides photovoltaic materials	VOC JSC Internal quantum efficiency (IQE)	PCA kNN Genetic programming	Literature data Experiment data	The thickness of the absorber layer Thickness of the window layer Bandgap of abosorb layer The distance between the cell and the center of deposition plume Resistance of the absorber layer Maximum value of calculated theoretical photocurrent	(Yosip of et al., 2015)
	Two-dimensional photovoltaic materials	Applicability in PV applications	GBC SVM RFC Ada boosting (Ada), LR SGDC,	ICSD	Packing factor (Pf), Average sublattice neighbour count (SNC), Mulliken electronegativity maximum and minimum value average atomic volume Lattice parameter	(Jin et al., 2020)

Applications	Materials/Processes	Target Properties	ML Model/Algorithms	Data Source	Most Related Descriptors	Ref.
			DT		Average bond ionicity of sublattice Anion framework coordination	
	Kesterite I2-II-IV-V4 quaternary compounds	Bandgap	LR SVR-linear kernel SVR-radial bias function kernel Boosted regression tree RF Logistic regression	DFT computation MP	Electronegativity Ionic radius Row in the periodic table	(Weston and Stampfl, 2018)
	16-atom constructed wurtzite nitrides in an orthorhombic cell	Bandgap Band offset	LR SVR-linear kernel SVR-poly kernel SVR-radial kernel ANN DNN	DFT computation	Electronegativity Covalent radius Valence First ionization energy	(Huang et al., 2019)
ORR	Dual-metal-site catalysts(DMSC)	Adsorption free energy of OH* (ΔGOH)	Gradient Boosted Regression (GBR)	DFT computation	Electron affinity Electronegativity Sum of the van der Waals (vdW) radius of the two transition-metal atoms. Absolute value of the difference between the two transition-metal atoms Sum of the Pauling negativity of the two transition-metal atoms.	(Zhu et al., 2019b)
	Binary and ternary nanocatalysts: PtCu, PtNi, CuNi, PtCuNi	Energy contribution of atom i. (Ei)	Neural Network Potential (NNP) with Monte Carlo	DFT computation Molecular dynamics simulations	Gaussian descriptor on the symmetry functions of radial (G2) Gaussian descriptor on the symmetry functions of angular (G4)	(Kang et al., 2018a)

Applications	Materials/Processes	Target Properties	ML Model/Algorithms	Data Source	Most Related Descriptors	Ref.
			Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm			
	Bimetallic Pt core-shell nanocatalysts	Strained coordination number (cn*(j))	KRR	DFT computation EMT Calculations	Coordination number Generalized coordination number Partial distribution function Distance to alloy atoms Interatomic distance from Vegard's law	(Rück et al., 2020)
	Titanium alloys: TiAl2O5	Kohn–Sham density functional theory energy of $TiAl_2O_5$ structures per atom (EBPNN) Kohn–Sham density functional theory force of $TiAl_2O_5$ structures (FBPNN)	Behler−Parrinello neural networks (BPNNs)	DFT computation	Behler−Parrinello descriptors (Behler, 2011)	(Groenenboom et al., 2020)
CRR	Intermetallics	CO and H adsorption energy on active sites	RFR PCA t-SNE	Materials Project DFT computation	Atomic number, Coordinated number Electronegativity Adsorption energy	(Tran and Ulissi, 2018)
	Bimetallic: Ni, NiGa, Ni3Ga, Ni5Ga3	CO adsorption energy on active sites	NNP	Materials Project DFT computation	Adsorption energy relative to the unrelaxed slab The gas-phase CO energy	(Ulissi et al., 2017)
	Bimetallic or multimetallic: Cu-Al alloy	ΔE_{CO}	RF t-SNE	Materials Project DFT computation	Atomic number, Coordinated number Electronegativity Adsorption energy	(Zhong et al., 2020b)
	Bimetallic or multimetallic: (100)-	ΔE_{CO}	ANN	DFT computation	Filling (f) of d-band Center (εd) of d-band Width (wd) of d-band	(Ma et al., 2015c)

Applications	Materials/Processes	Target Properties	ML Model/Algorithms	Data Source	Most Related Descriptors	Ref.
	terminated Cu multimetallic alloys				Skewness (γ1) d-band Kurtosis (γ2) of d-band, Local Pauling electronegativity (χl)	
	High-entropy alloys CoCuGaNiZn and AgAuCuPdPt	ΔECO ΔEH	GPR	DFT computation	CO and H adsorption energy on local atomatic environment	(Batchelor et al., 2019)
NRR	IrO2, MoS2	Free energy of all possible adsorbate coverages	GPR	DFT computation	Surface coverage configurations	(Ulissi et al., 2016)
	NRR electrocatalytic electrode	Total current density (|itotal|) Faradaic efficiency (F.E)	ANN	Self-generation	Overpotential Electrode morphorlogy The kinetic predisposition of NRR	(Hoar et al., 2020a)
Thermoelectricity	Ba(MgX)2, (X = P, As, Bi), X2YZ6 (X = K, Rb, Y=Pd, Pt, Z = Cl, Br), K2PtX2 (X = S, Se), NbCu3X4 (X = S, Se, Te), Sr2XYO6 (X = Ta, Zn, Y=Ga, Mo), TaCu3X4 (X = S, Se, Te), and XYN (X = Ti, Zr, Y=Cl, Br).	Types of Seebeck factors (S) Types of Power factors (σS2)	GBDT DT RF kNN ANN	JARVIS-DFT BoltzTrap calculations	CFID descriptors Chemical descriptor Radial distribution function Angle-distribution up to first neighbourDihedral angle distribution	(Choudhary et al., 2020b)
	100 single crystal inorganic materials	The lattice thermal conductivity (kl)	GPR	MP	Bulk modulus Space group number Maximum atomic radius Volume per atom	(Chen et al., 2019b)

Applications	Materials/Processes	Target Properties	ML Model/Algorithms	Data Source	Most Related Descriptors	Ref.
	Off-stoichiometric samples (namely, Al23.5+xFe36.5Si40–x) of the Al2Fe3Si3 compound	σS2	GPR	Experiment measurements	Al/Si Ratio Temperature	(Hou et al., 2019)
Piezoelectricity	Pb-free BaTiO3	Electrostrains	GB	Experiment measurements	Electronegativity Ionic radius, volume Ionic displacements Polarization and Dopant effects on transition	(Yuan et al., 2018)
	(Ba0.50Ca0.50)TiO3-Ba(Ti0.70Zr0.30Sn0.30)O3	Morphotropic phase boundary	Bayesian learning SVR radial bias function SVR linear regression LR	Experimental measurements	Unit cell volume difference The ratio of average ionic radii Ionic displacements The ratio of the effective nuclear charge Ratio of electronegativities	(Xue et al., 2016)
Rechargeable Alkali-Ion Battery	Li containing crystalline solids	Is it a superionic material	Logistic regression	MP ICSD	The average number of Li neighbors for each Li The average sublattice bond ionicity The average anion coordination number in the anion framework The average shortest Li–anion and Li-Li distance in angstroms	(Sendek et al., 2017)
	Li metal anode	Shear moduli Bulk moduli Elastic constants C11 Elastic constants C12 Elastic constants C44	Graph convolutional neural network GBR KRR	DFT computation MP	Crystal structure Mass density Ratio of bond iconicity between Li and sublattice Sublattice electronegativity Volume per atom	(Ahmad et al., 2018)
	Electrolyte solvents	Coordination energy (Ecoord)	MLR LASSO	KISHIDA Chemical Database	Ionic radius NBO charge of O atom Atomic weight	(Ishikawa et

Applications	Materials/Processes	Target Properties	ML Model/Algorithms	Data Source	Most Related Descriptors	Ref.
			Exhaustive search with linear regression	Experiment measurements	Bolling point of solvent HOMO LUMO	al., 2019)
	Silicate-based cathodes with the composition of Li–Si–(Mn, Fe, Co)–O.	Types of crystal system	ANN SVM k-NN RF	MP	Formation energy Energy above the hull Bandgap Number of sites	(Attarian Shandiz and Gauvin, 2016)
	Electrode materials for metal-ion batteries	Electrode voltage	ANN SVM KRR	MP	Working ion in the battery The concentration of the active metal ion in a given compound crystal lattice types Space group numbers.	(Joshi et al., 2019)
Supercapactior	Carbon-based electrodes	Capacitance	LR LASSO ANN	Published literature	pore size, ID/IG, Specific surface area N-doping level	(Zhu et al., 2018)
Environmental Decontamination	TiO2	Degradation rate	MLR	Experiment measurements	Bond Lipophilicity Dipole Bond dipole Bond molar refractivity	(Dondapati and Chen, 2020)
Flexible Electronics	Ag/poly amic acid (Ag/PAA) composites	Sheet resistance Processing time	ANN	Experiment measurements	Concentration of PAA Concentration of NaBH4, Reduction time of NaBH4, The ion exchange time of AgNO3.	(Zhang et al., 2020b)

Applications	Materials/Processes	Target Properties	ML Model/Algorithms	Data Source	Most Related Descriptors	Ref.
Optoelectronics	2D octahedral oxyhalides	Bandgap	GBR PCA	DFT computation	Distorted stacked octahedral factors	(Ma et al., 2019)
Superconductors	12,000+ known superconductors	Critical temperature (TC)	RF	SuperCon ICSD	Stoichiometric descriptors Elemental property statistics Electronic structure descriptors Ionic compound descriptors	(Stanev et al., 2018)
	Superconductors in the SuperCon data set	TC	DNN	SuperCon COD Published literature	Composition of materials	(Konno et al., 2021)
Metallic glasses	Ternary amorphous alloys	Volumen per atom Fromation energy Bandgap energy	RF LR RF Reduced-errpr Pruning Tree	Nonequilibrium Phase Diagram of Ternary Amorphous Alloys	Stoichiometric descriptors Elemental property statistics Electronic structure descriptors Ionic compound descriptors	(Ward et al., 2016a)
	Bulk metallic glasses	The existence ability in an amorphous state: glass-forming ability (GFA) The critical casting diameter (Dmax) The supercooled liquid range (ΔTx)	RF	Experiment measurements	Composition of materials	(Ward et al., 2018b)
Magnetic materials	Ferromagnetic materials and antiferromagnetic	Curie temperature Magnetic ground state	RF	AtomWork	Magpie descriptors SOAP Space group number	(Long et al., 2021)

Applications	Materials/Processes	Target Properties	ML Model/Algorithms	Data Source	Most Related Descriptors	Ref.
	Permanent magnets	Uniaxial magneto-crystalline anisotropy constant (K1) The magnetization (μ0M) The relative phase stability energy (Ef)	SVR DT	Published literature	Crystal configuration	(Möller et al., 2018)
	Soft magnetic materials	Magnetic saturation Coercivity Magnetostriction	GBDT LR SVM DT, RF, *k*NN	Published literature	Annealing temperature Annealing Time Primary Crystallization Onset Primary Crystallization Peak	(Wang et al., 2020b)

2.5.1 Data-Driven Innovtion for ORR

Using data-driven technology to discover innovative, economical and efficient electrocatalysts has gradually become the focus of oxygen reduction reaction (ORR) research. ORR plays a vital role in chemical-electrical energy conversion in fuel cells and metal-air batteries, which is a promising and indispensable field in the development of renewable energy (Kulkarni et al., 2018). Recently, a new frontier ORR catalyst has emerged referred to as dual-metal-site catalysts (DMSCs). By employing ML techniques, Zhu et al. (Zhu et al., 2019b) identify the origin of ORR activity and reveal design principles that offer a universal description of the activity in relation to intrinsic properties for graphene-based DMSCs. In this research, they used DFT simulations to screen potential catalyst candidates by considering the two criteria of geometric structure and free energy for the reaction. Each candidate's catalytic performance was quantified based on the theoretical potential of the rate-limiting step (U_L); a value larger than 0.7V was regarded as favorable ORR activity. Their U_L of such DMSCs can only be higher than 0.7V when the rate-limiting step is either the first or fourth electrochemical step. A linear scaling relationship between ΔG_{OOH*} and ΔG_{OH*} for the evaluated DMSCs were determined via regression ($\Delta G_{OOH*} = 0.92\ \Delta G_{OH*} + 3.01$); thus, the trends in ORR activity with the variations in ΔG_{OOH*} and ΔG_{OH*} can be plotted (**Figure 2.6**a). Based on the DFT computations, numerous primary physiochemical parameters were enumerated as possible descriptors for ML training. As the activity of catalysts is essentially dominated by electronic strictures, properties of localized d-orbital and continuum s- and p- orbitals were selected as the primary descriptors. Additionally, considering interactions between two transition-metal atoms, some geometric structure-related properties were set as descriptors. The Person correlation coefficient matrix was used to identify the inner correlation between random descriptor pairs to eliminate redundant descriptors. With some simple mathematical transformations, the descriptor space was extended and optimized in accordance with the ML model's prediction accuracy. Finally, a gradient boosting regression (GBR) model with an R^2 of 0.993 and RMSE of 0.036 eV was obtained. The mean impact value (MIV) (Jiang et al., 2013) method was coupled with the trained ML model to evaluate each descriptor's influence on the ORR activity. The seven most related descriptors are: the electron affinity (EA_1 and EA_2); the sum of the van der Waals (vdW) radius ($R_1 + R_2$); the absolute value of the difference between and the sum of the Pauling negativity ($|P_1-P_2|$, P_1+P_2) of the two transition-metal atoms; the product ($IE_1 \times L$) of the ionization energy of the first transition-metal atom (IE_1); the distance (L) between the two transition-metal atoms; the average distance between the two transition-metal atoms and the surrounding N atoms ((d1+d2+d3+d4+d5+d6)/6). Among the seven descriptors, five are electronics properties. However, isolated individual descriptors may have their limitations and may not be sufficient to describe the effects of atoms on catalytic performance. In contrast, too many descriptors would lead to the dimensionality curse and disrupt the model's predictive performance. Hence, it is essential to discover and identify new, high-dimensional descriptors which are highly related to the target results and carry the most information. Based on the data generated from DFT computations and microkinetic simulations, the trained ML model can accurately describe the ORR catalytic activity of DMSCs via fundamental parameters with acceptable error.

To study the electrocatalytic performance of more complex, larger structures, traditional DFT calculations are limited due to their large computational expense and time. Researchers have gradually developed new strategies that combine ML with DFT and other computing methods. Kang et al. (Kang et al., 2018a) used Gaussian descriptors (Behler and Parrinello, 2007, Behler, 2011) to characterize local atomic structure. The authors applied an ML-based framework to explore the thermo-electrochemical properties of ternary nano-electrocatalysts. A model of high-dimensional neural network potentials (NNPs) was trained with the employment of the atomistic ML package (AMP)(Khorshidi and Peterson, 2016) to describe the interactions between components (**Figure 2.6**b). The NNP method was then implemented in conjunction with Monte Carlo (MC) methods and molecular dynamics (MD) simulation to identify the effect of strain originating from surface segregation of selective components at the surface of the catalyst. 13,877 DFT calculated data for PtNi, PtCu, CuNi, and PtCuNi nanoparticles were used for the training sample. The training set of the model system was composed of nanoscale icosahedrons with transition-metal species mixed randomly. To distinguish the local structural environment, Gaussian descriptors on radial (G^2) and angular (G^4) symmetry functions were employed as the main parameters. The RMSE of the NNP model on the prediction of single-atom energy contribution converged to less than 7 meV with the implementation of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Broyden, 1970, Fletcher, 1970). The proposed candidate PtCuNi ternary that contains 60% Pt possesses a size of 2.6 nm demonstrates outstanding electrocatalytic ability toward ORR. According to the thermal-electrochemical stability analysis via MC and MD simulations under the canonical ensemble, the candidate is also consistently more stable than binary nanoparticle and pure Pt. The design principle which emerges from the ML model is that those ternary nanoparticles with 60% Pt composition and icosahedron configurations in which Cu/Ni and Pt as assume the core and shell, respectively, possess superior ORR catalysis performance in terms of both activity and stability.

The electrocatalytic performance of the core-shell catalysts is very attractive, but due to their impractical size, there still remain an insufficient number of mechanism studies having been reported. ML could further support the future development and exploration of core-shell catalysts. Rück(Rück et al., 2020) and his co-workers have further studied strained Pt-based core-shell electrocatalysts. They propose an ML-based framework for the prediction, with site-specific strain precision, to investigate how effect of strain on Pt core-shell nanocatalysts towards the ORR activity. The strained coordination number (cn*(j)), which describes the compressive and tensile strain on atom j with the variation of atomic coordination, was set as the target property of the ML model. The ML model was trained with a kernel ridge regression (KRR) algorithm, which applies a radial basis function (RBF) kernel to test nanoparticles whose structures are optimized for the minimum energy. The effective medium theory (EMT)(Jacobsen et al., 1996) was used to calculate the structure energy by employing the ASAP calculator in the Atomic Simulation Environment (Hjorth Larsen et al., 2017). The EMT-calculated energy was validated by DFT calculations on 1.9 nm sized core-shell nanoparticles. As is shown in **Figure 2.6**c, for each core, the ML model was trained with nanoparticle sizes from 1.6 nm to 5.4 nm at 0.2nm intervals. Five descriptors were selected: the coordination number (cn(j)) and generalized coordination number(CN(j)) to describe local-site structure, which has significant impact on the adsorption energy of the intermediates; the partial distribution function (PDF(j, r)); distance to alloy atoms(d_{alloy}(j)); the interatomic distance from Vegard's law (d_{veg}(j)). The MAE of the ML prediction of the strain on single atoms varied from 0.0007 to 0.0159 with respect to different catalyst cores. In this study, the relation established by the ML model indicates that the size of the nanoparticle determines the optimal strain. The mass activities could be enhanced by weakening compressive strain on PtAg and PtAu of sizes of 2.83 nm or by strengthening compressive strain on PtCu and PtNi of sizes of 1.92 nm.

To summarize, data-driven techniques are primarily implemented to establish the relation between the intrinsic properties and catalytic activity in the field of ORR. Some fundamental factors, including electronegativity, electron affinity and radii of the embedded transition-metal atoms, exhibit a high correlation with the ORR activity of DMSCs. Furthermore, in the design of core-shell ORR nanocatalysts, ML models indicate that the bimetallic material composition, size, and shell thickness of nanoparticles control the mass activity. In addition to catalytic activity, the thermal-electrochemical properties could also be predicted by ML models trained on descriptors generated by symmetric functions.

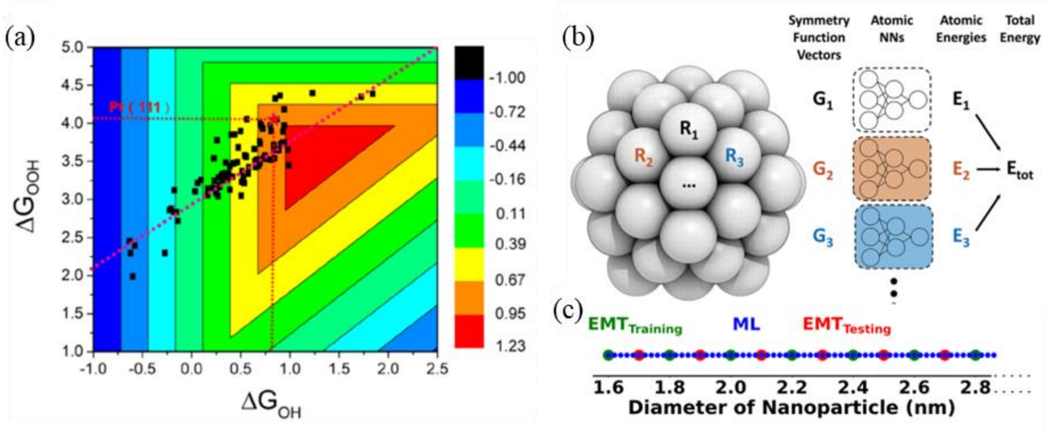

Figure 2.6 (a) The trends plot of ORR activity with the variation of ΔG_{OOH*} and ΔG_{OH*} of DMSCs. Reproduced with permission **(Rück et al., 2020)**. Copyright 2020, ACS Publications. (b) The scheme of the high dimension NNP method. The symmetry functions are transformed from the Cartesian to represent chemical environments. The NN then predicts the contribution of energy based on the symmetry functions and the total energy is obtained by adding up all of the energy contributions. (c) The size of nanoparticles used for training, testing and ML prediction, which are represented in green, red and blue colour, respectively. (b-c) Reproduced with permission **(Kang et al., 2018a)**. Copyright 2018, RSC Publications.

2.5.2 Data-Driven Innovtion for CRR

CRR is considered to be a promising, clean, and environmentally friendly strategy to reduce greenhouse gas emissions and resolve the energy crisis; it has been broadly studied to improve reaction efficiency and selectivity (Liu et al., 2017a, Lu et al., 2014). The introduction of ML for accelerating the discovery of CRR catalysts has been widely implemented in this domain, including the prediction of adsorption energies (Zhong et al., 2020b), identification of active sites on the surface of catalysts (Ulissi et al., 2017), optimization of reaction conditions for improving selectivity (Siebert et al., 2019), carbon dioxide capture ability, and design of catalysts.

Tran and Ulissi (Tran and Ulissi, 2018) employed an active ML model to guide the DFT simulation to identify optimal intermetallic electrocatalysts for CO_2 reduction and H_2 evolution. A workflow was established to screen a chemical space of 1499 candidates across 31 different elements (33% p-block and 50% d-block) of intermetallic materials acquired from the Materials Project. The open-source code pymatgen was implemented, by which 17,507 adsorption surfaces and 1,684,908 adsorption sites were enumerated. The vector employed to represent the environment of the coordination site contained four descriptors: atomic number (Z), Pauling electronegativity (χ) of the element, number of atoms of the element that coordinate (CN) with CO, and crude estimate of the adsorption energy on the site (ΔE) (**Figure 2.7**a). A framework of continuous, alternating iterations between ML screening and DFT computation was constructed, where the results of DFT simulation were fed back to the ML model, and newly predicted potential adsorption sites with near-optimal values (ΔE_{CO} = -0.67 eV and ΔE_H = -0.27 eV) were sent back for DFT calculations to generate new training data. **Figure 2.7**b represents the normalized distribution for the low coverage, DFT computed CO adsorption energies (ΔE_{CO}) of all of the DFT researched surfaces. The low coverage ΔE_{CO} computed by DFT for surface (131) and predicted by the ML model for surface

(844) are shown in **Figure 2.7**c and d, respectively. The RMSE, MAE, and MAD of the active learning model's prediction were 0.46, 0.29 and 0.17 eV, respectively. One reason for this considerable error could be the use of ideal structures rather than relaxed structures for DFT calculation, as it is faster and less computationally expensive, though with the trade-off of the prediction accuracy.

Zhong et al. (Zhong et al., 2020b) used an ML model to predict the CO adsorption energies (ΔE_{CO}) on the adsorption sites of copper-containing intermetallic crystals, among which Cu-Al alloy was found to be the most promising electrocatalyst. The ML-predicted CO adsorption energy combined with the volcano scaling relationships (Liu et al., 2017a) revealed the highest number of catalytic adsorption sites, where the CO adsorption value energies were near the optimal value of -0.67 eV (**Figure 2.7**e) (Zhong et al., 2020b, Tran and Ulissi, 2018). A similar descriptor space was applied for each element type-coordinate with CO to characterize the first and second neighbouring shell of CO for each active site, with the difference that ΔE is replaced by the median adsorption energy ($\Delta\tilde{E}$) between the pure element and CO, yielded from the prior DFT simulation. The constructed vector space was then sent to an automated ML tool called the Tree-based Pipeline Optimization Tool (TPOT)(Olson et al., 2016) to implement the random forest regression (RFR) model. By using 19,644 DFT simulated data points of ΔE_{CO} and an extra tree regressor with 5-fold CV, the RFR model demonstrated both a median absolute deviation (MAD) and mean absolute error (MAE) of about 0.1 eV in predicting the ΔE_{CO} on the test data (5% of the whole data size), which is comparable to the accuracy of DFT simulation. The trained ML model was then coupled with the quantum chemical computation framework to construct an active ML system. The ML model predicted the ΔE_{CO} of all the adsorption sites enumerated by the DFT framework from Materials Project (MP); those sites whose predicted ΔE_{CO} was close to -0.6 eV were automatically collected and sent to the next stage. DFT simulations of ΔE_{CO} were subsequently executed for these sites, and the additional yielded data of ΔE_{CO} were then added in the training dataset to iterate a new ML model. The further optimized and improved ML model would identify new promising adsorption sites based on the value of predicted ΔE_{CO}, which could be fed back to the DFT framework to provide new ML training data. Thus an automatically, iteratively and systematically active ML workflow was established and a DFT database of ΔE_{CO} on promising adsorption sites was constructed. In this work, the structures established from MP were managed by Atomic Simulation Environment (ASE); (Hjorth Larsen et al., 2017) the Python Materials Genomics (pymatgen), which currently powers the MP, was used to enumerate all the surfaces and adsorption sites. DFT calculations were performed with VASP, while software including Lungi and FireWorks were used to manage the computation framework and workflow. The active ML workflow finally trained more than 300 RFR models, and the guided DFT simulations were ultimately conducted for 4000 different candidates of adsorption sites with a near-optimal value of ΔE_{CO} on the Cu-containing surface quarter of which the majority were on Cu-Al Surfaces (**Figure 2.7**f). The integration of the volcano relationship, DFT simulation, and active ML achieved efficient and accurate prediction ideal electrocatalysts for active and selective CO_2 reduction to C_2H_2. Based on the ML results, the author concluded that those Cu-Al alloys that contain higher Cu composition are more promising for CRR. A follow-up experimental validation was performed and the CO_2-to-C_2H_4 performance achieved ~55% PCE under 150 mA cm^{-2} at the cathode side. Although numerous DFT-calculated adsorption energies are required for the training of ML model, this approach reveals the importance of the data-driven and active-ML-guided experimental exploration in overcoming the limitations of the conventional single-component catalysts in CRR.

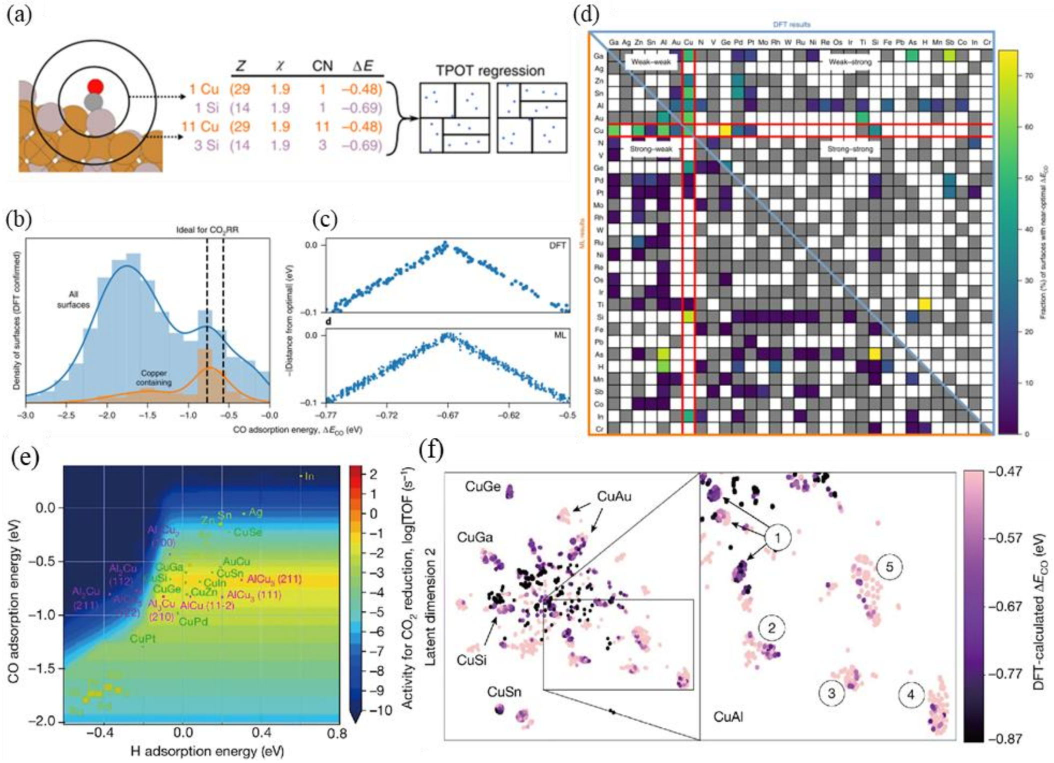

Figure 2.7 (a) The sample of the numerical encoding for the adsorption site. The constructed descriptor space is employed as model input by the Tree-based Pipeline Optimization Tool (TPOT) to predict ΔE_{CO}. (b) The normalized distribution of the low coverage, DFT derived ΔE_{CO} for all of the DFT computed surfaces. The sub-distribution for cooper containing surface is marked in orange, and the black dashed lines indicate the range of for the optimal ΔE_{CO} (-0.67 eV). (c) The low coverage

ΔE_{CO} computed by DFT for surface (131) and (d) predicted by ML model for surface (844). (a-d) Reproduced with permission (Tran and Ulissi, 2018). Copyright 2018, Springer Nature Publications. (e) A two-dimensional activity volcano plot for CO_2 reduction. TOF, turnover frequency. (f) *t*-SNE representation of approximately 4,000 adsorption sites on performed DFT calculations with Cu-containing alloys. The Cu-Al clusters are labelled numerically. (e-f) Reproduced with permission (Zhong et al., 2020b). Copyright 2020, Springer Nature Publications.

Pedersen et al.(Pedersen et al., 2020) have explored a probabilistic and unbiased method to research high-entropy alloy performance as the electrocatalysts for the reduction of CO_2 and CO. The authors integrated the quantum chemical simulations and ML model to predict the ΔE_{CO} and adsorption energy of hydrogen (ΔE_H) of all the adsorption sites on the surface of the disordered CoCuGaNiZn and AgAuCuPdPt HEAs. The disordered surface consists of different metal atoms that would naturally provide many distinct adsorption sites with each adsorbate's unique adsorption properties, as determined by the site's microstructure. Hence, a Gaussian process regression (GPR) model was established that uses the adsorption energy of CO and H in the local atomic environment around the adsorption sites (computed by DFT) to predict the ΔE_{CO} and ΔE_H. The training data size was ca. 1000, where 5-fold CV was applied with MAEs of 46-64 meV (Figure 2.8a). The predictive model allows the optimization of HEA compositions to increase the probability of catalyzing performance improvement. Every local adsorption site contributes to the HEAs' global catalytic properties; some of the local optimal compositions such as $Co_9Ga_{42}Ni_7Zn_{42}$, $Ga_{83}Ni_{17}$, $Ag_{69}Cu_{31}$, and $Ag_{84}Pd_{16}$/$Au_{84}Pd_{16}$ were predicted. The best five-metal alloy candidates that contain at least of 10% of each elements are $Co_{10}Cu_{10}Ga_{60}Ni_{10}Zn_{10}$ and $Ag_{30}Au_{33}Cu_{17}Pd_{10}Pt_{10}$. A concurrent and independent work published by Nellaiappan et al.(Nellaiappan S; Kumar N; Kumar R; Parui A, 2019) have experimentally investigated the CRR performance on the AgAuCuPdPt HEA, where the results are in favorable agreement with the predictions in this work.

Important descriptors for the performance of CRR catalysts are also a necessary means to improve the accuracy of ML. Ma et al.(Ma et al., 2015c) pioneered the use of a feed-forward ANN ML model via open-source PyBrain code to establish a nonlinear correlation between the descriptor vector and the ΔE_{CO}. The descriptor vector consisted of 13 electronic properties which were determined theoretically, among which characterize the properties of the clean adsorption surface (such as d-states distribution) including the filling (f), center (ε_d), width (W_d), skewness (γ_1) and kurtosis (γ_2) of a d-band, in conjunction with the local Pauling electronegativity (χ_l) determined by delocalized sp-states, were taken as the primary descriptors. The secondary descriptors such as work function (W), atomic radius (r_0), the spatial extent of d-orbitals (r_d), ionization potential (IE), electron affinity (EA), Pauling electronegativity (χ) and the square of adsorbate-metal interatomic d coupling matrix element (V^2_{ad}), were also fed into the ML model. All the input features were standardized to improve the performance of the ANN model, and a 10-fold CV was performed; the ML-predicted adsorption energy of CO was shown to agree well with the DFT simulations, where the average RMSE achieved a value of 0.13 eV (**Figure 2.8**b). The outperformed candidate [100]-terminated Cu multimetallic alloys were discovered to have lower overpotentials but potentially higher selectivity towards the reduction of CO_2 to C_2 species. After a perturbation to the input descriptors was performed and the model responses were compared, the importance of the descriptors was examined (**Figure 2.8**c). The developed ML model demonstrated a novel methodology for capturing complexity in electrocatalytic CRR and acquiring accurate values of adsorption energies without expensive quantum chemical computations, providing in-depth understanding and strategies for catalysts design.

The majority of applications of data-driven innovation in CRR are for predicting the adsorption energy of CO and H to evaluate the activity and selectivity of the catalyst candidates. The atom environment of the local adsorption site plays a dominant role in the catalyst performance, and descriptors, such as electronegativity and coordination numbers, have high impact on adsorption energy. The exploration of the catalytic performance and material structure by using data-driven techniques provides the possibility of a rational design of high-performance materials to boost the CRR.

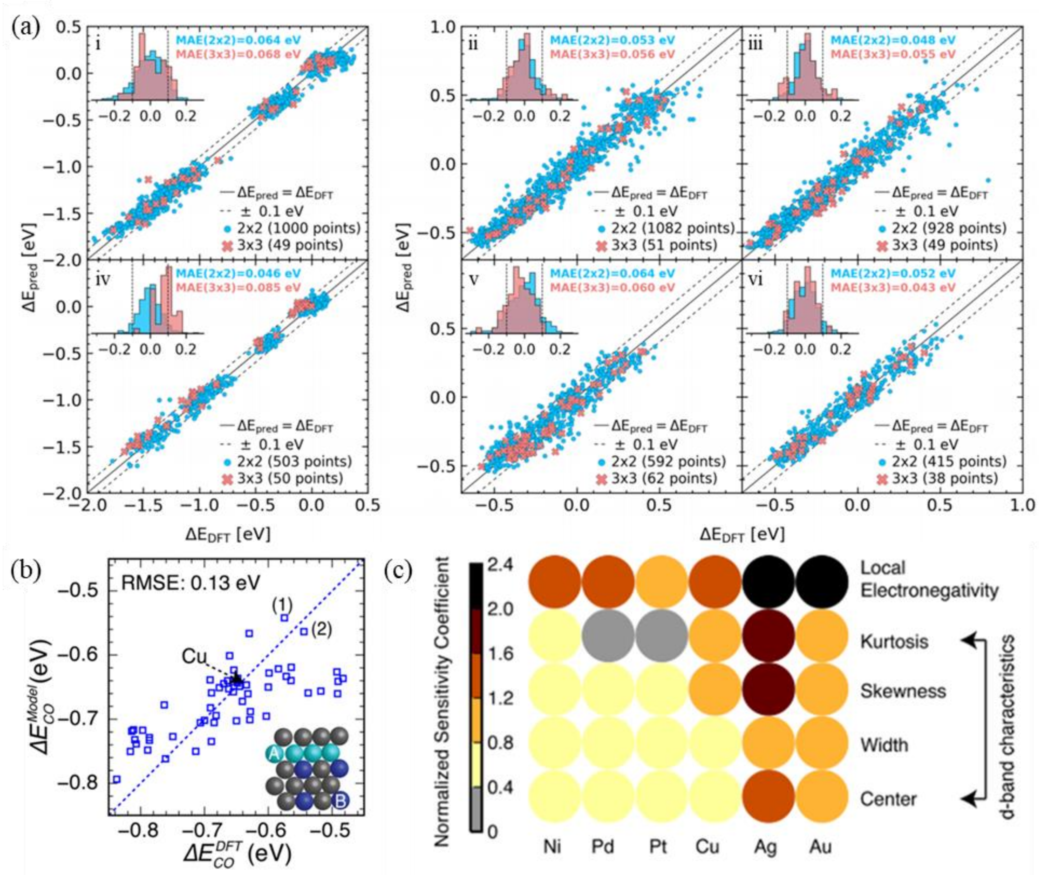

Figure 2.8 (a) The performance of the GBR ML model for adsorption energy prediction. The ML predicted and DFT computed adsorption energies for on-top CO (i,iv), fcc-hollow H (ii,v) and hcp-hollow H (iii,vi) on the CoCuGaNiZn (i-iii) and AgAuCuPdPt (iv-vi) HEAs. Reproduced with permission **(Pedersen et al., 2020)**. Copyright 2020, ACS Publications. (b) The performance of the NN ML model for adsorption energy prediction Cu monolayer alloys. (c) The nominalized sensitivity coefficient of the d-band descriptors. Reproduced with permission **(Ma et al., 2015c)**. Copyright 2015, ACS Publications.

2.5.3 Rechargeable Alkali-Ion Battery

This sub-section will discuss the accelerated discovery of potential material candidates for electrolytes and electrodes based on data-driven strategies. As a key component of electrochemical energy storage, rechargeable batteries are extremely vital for various applications, including new energy vehicles, consumer electronics, and aerospace. To meet the growing needs of these applications, larger volumes of rechargeable batteries are being demanded with higher energy density, higher power density, longer cycle life, greater safety, and at an acceptable cost. Thus, it is essential to develop key rechargeable battery materials, including those for electrodes and electrolytes, to improve the performance of rechargeable batteries. Data-driven screening of electrolytes often quickly identifies promising electrolytes through indicators such as chemical and structural stability (Sendek et al., 2017), electronic properties (Sendek et al., 2017), mechanical properties (Ahmad et al., 2018), and coordination energy (Ishikawa et al., 2019). For electrodes, voltage (Joshi et al., 2019), volume (LeCun et al., 2015) and redox potential (Attarian Shandiz and Gauvin, 2016) are essential for ML to successfully predict and evaluate the performance of electrodes.

2.5.3.1. Electrolytes

Electrolytes are vital components of rechargeable batteries; it is essential to find high-performance electrolytes in the development of advanced rechargeable batteries (Xu, 2004, Bhatt and O'Dwyer, 2015). With the significant advance of quantum chemical computations and ML learning techniques, some researchers have applied high-throughput data-driven approaches to discover innovative, next-generation battery electrolytes (Cheng et al., 2015a, Halls and Tasaki, 2010, Hautier et al., 2012, Curtarolo et al., 2013, Korth, 2014, Husch et al., 2015). Sendek et al. (Sendek et al., 2017) have proposed a workflow of large scale computational material screening for solid electrolytes in lithium-ion batteries (Figure 2.9a-c). The authors first acquired atomistic and electronic structure parameters for 12,831 lithium-containing candidates from the MP database, including the equilibrium atom position, the energy above the convex hull, the bandgaps, and the Gibbs free energy, utilizing the Python package Pymatgen (Ong et al., 2013). This was followed by a primary screening stage using four prerequisite criteria: low electronic conductivity, high chemical and structural stability, and low material cost. A logistic regression model was trained to identify the structures that are most likely to exhibit excellent lithium conduction based on five features including the average number of Li neighbors for each Li, the average sublattice bond ionicity, the average anion coordination number in the anion framework, the average shortest Li–anion and Li–Li distance in angstroms. The training set consisted of 40 crystal structures whose ionic conductivity values were available in the literature. The threshold of superionic conductive behavior was set as 0.1 mS/c; finally, 21 structures demonstrated potential as high-performance electrolytes, some of which have been experimentally investigated (Wada et al., 1983, Tomita et al., 2008, Yamada et al., 2006, Court-Castagnet et al., 1993). This method is applicable to confirming the ionic conductivity of unreported inorganic materials.

Similarly, Ahmad et al. (Ahmad et al., 2018) conducted a high-throughput data-driven search over for solid electrolytes with outstanding dendrite suppression capability of Li on the anode. A crystal graph-based convolutional neural network (CGCNN) (Xie and Grossman, 2018) was trained to predict the moduli of shear and bulk given a large, available, low noise dataset obtained from low uncertainty first-principle-calculated values. The CGCNN model was trained by only structural descriptors, which bypass first-principles calculations. Additional ML models based on GBR and KRR were also employed to predict the elastic constants of cubic materials (**Figure 2.9**d). Those predicted mechanical properties are critical in stabilizing the interface and computationally expensive to obtain via first-principle methods. Those properties were taken as the input of the theoretical framework utilizing the stability parameter (Ahmad and Viswanathan, 2017b, Ahmad and Viswanathan, 2017a) to figure out the dendrite initiation on the Li metal anode. The stiffness of the material was found to be positively correlated with the mass density and the ratio of bond iconicity between Li and the sublattice, whereas a negative correlation was obtained with the sublattice electronegativity and volume per atom. Further investigations of thermodynamic stability and electronic conductivity were performed. Additionally, the method proposed by Sendek et al. (Sendek et al., 2017) was employed to confirm the ionic conductivity. Over 20 mechanically anisotropic interfaces and 4 electrolytes including Li_2WS_4-$P\bar{4}2m$, Li_2WS_4-$I\bar{4}2m$, $LiBH_4$-$\bar{P}1$ and LiOH-P_4/nmm were predicted as promising to be employed to suppress dendrite growth. The screened candidates were highly anisotropic and generally soft, which indicate opportunities for acquiring innovative solid electrolytes with both high ionic conductivity and dendrite suppression. The R^2 on the predictions of elastic constants C_{11}, C_{12}, and C_{44} were 0.60, 0.79, and 0.6, respectively; this might be due to the uncertainty inherent in the DFT-calculated values (Ahmad and Viswanathan, 2016, Deng et al., 2015); the use of low uncertainity might improve the model performance. With the ability of data handling and feature generation, the proposed methodology in this study is readily applicable the screening of other inorganic materials for properties of interest.

Existing studies have mainly concentrated on solid electrolytes. Investigations of liquid electrolytes hav barely been reported (Chen et al., 2018b, Chen et al., 2018a), mainly because the molecular structure of a liquid system is more flexible, which makes it challenging to extract structural information. Ishikawa et al. (Ishikawa et al., 2019) integrated a data-driven method with quantum chemistry computations to predict the coordination energy (E_{coord}) (Okoshi et al., 2013, Okoshi et al., 2016) of alkali group metal ions (Li, Na, K, Rb, and Cs) in battery electrolyte solvents. The E_{coord} is closely related to ion transfer at the interface of electrolyte/electrode, which is first obtained by quantum chemical computations. The calculated E_{coord} for 5 alkali ions is shown in **Figure 2.9**e. Three ML regression methods, namely, MLR, LASSO, and exhaustive search with linear regression (ES-LiR),(Sodeyama et al., 2018, Igarashi et al., 2016, Igarashi et al., 2018) were implemented to identify the relationship between E_{coord} and selected descriptors. The descriptor space consists of both ion and solvent properties, such as the ions' atomic weight and boiling point of the solvents. The results revealed that the most critical descriptors are the ionic radius and the oxygen atom's charge connected to the metal ion. The ES-LiR model yielded a CV error (Sodeyama et al., 2018, Igarashi et al., 2016, Igarashi et al., 2018) of 0.127 eV for the prediction accuracy of E_{coord} (**Figure 2.9**f). By implementing the exhaustive search with Gaussian process (ES-GP) (**Figure 2.9**g), a further improvement of the prediction accuracy with a CV error of 0.016 eV was achieved. This study demonstrated that the integrated data-driven techniques and quantum chemistry calculations can accurately predict E_{coord} of any alkali metal ion coordination. The trained ML model could be employed to search for battery electrolyte materials, where several descriptors including ionic radius and NBO charge of the O atom are identified as critical in developing next-generation post-Li batteries.

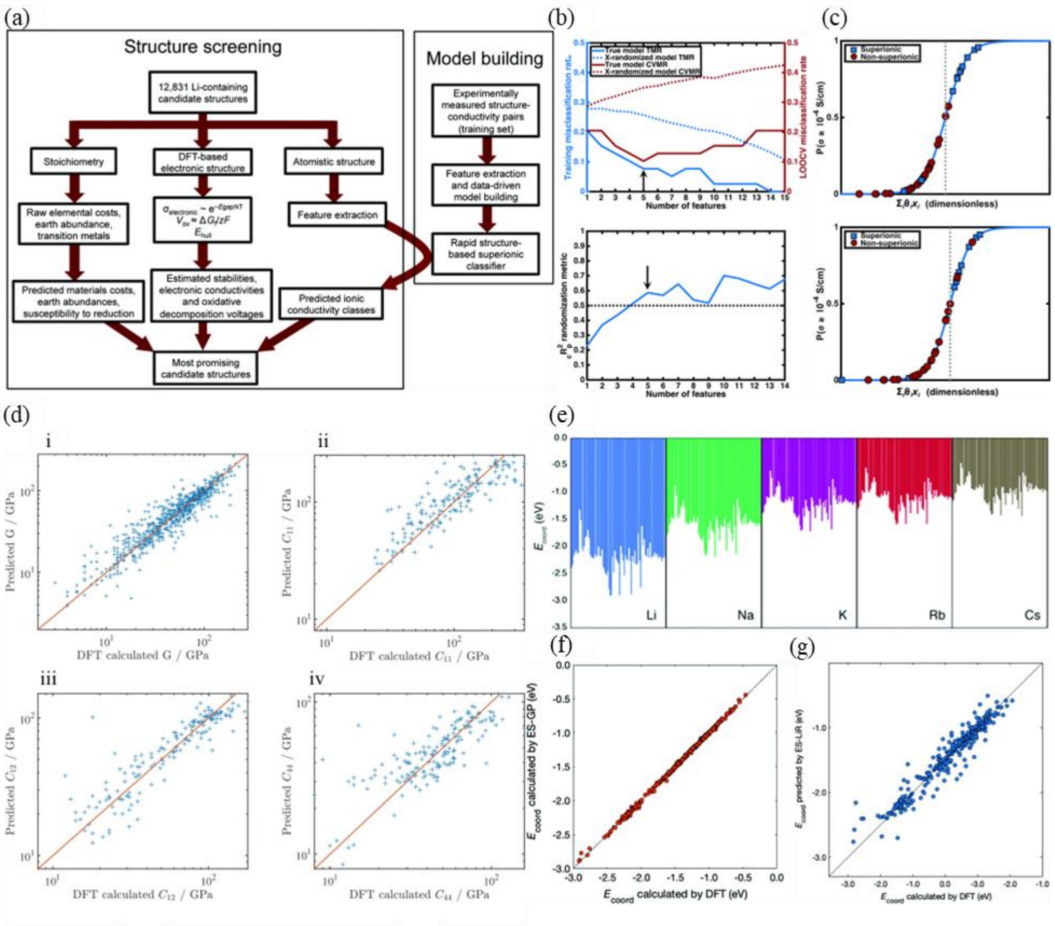

Figure 2.9 (a) Flow diagram of the ML assisted material screening process for Li-contained candidates. (b) (top) The training misclassification rate (TMR) and cross-validation misclassification rate (CVMR) via LOOCV. The dashed lines in the top diagram describe the mean value of the performed *X*-randomization analysis which is applied to ensure the model is not built on chance correlation. (bottom) The performance of ML models compare with chance correlations, the black dashed line indicates the threshold. (c) The performance of the training data using logistical regression with LOOCV. (a-c) Reproduced with permission (Sendek et al., 2017). Copyright 2017, RSC Publications. (d) The comparison diagram of elastic properties between the ML predicted and DFT computed value: (i) shear modulus

and elastic constants (ii) C_{11} (iii) C_{12} and (iv) C_{44}. Reproduced with permission (Ahmad et al., 2018). Copyright 2018, ACS Publications. (e) E_{coord} of 70 solvents and the five alkali metal ions. (f) The performance of the ES-GP model for the prediction of E_{coord}. (g) The performance of the ES-LiR model for the prediction of E_{coord}. Reproduced with permission (Ishikawa et al., 2019). Copyright 2019, RSC Publications.

2.5.3.2. Electrodes

Accelerating the discovery of suitable materials for high-power, safe, and stable electrodes is essential for developing improved rechargeable batteries. Because of the development of first-principles computations, the properties of unknown electrode materials can be obtained to support the research of complex phenomena (Meng and Arroyo-de Dompablo, 2009, Nishijima et al., 2014, Meng and Arroyo-de Dompablo, 2013, Yan et al., 2014). Nevertheless, the advancement of ML techniques can enable more efficient discovery of innovative materials to identify the complex, implicit correlations between crystal structure and various properties of electrode materials such as voltage, capacity, ionic and electronic mobility, stability, redox potential, and volume changes in the battery (Meng and Arroyo-de Dompablo, 2009, Nishijima et al., 2014, Meng and Arroyo-de Dompablo, 2013, Yan et al., 2014, Liu et al., 2020b).

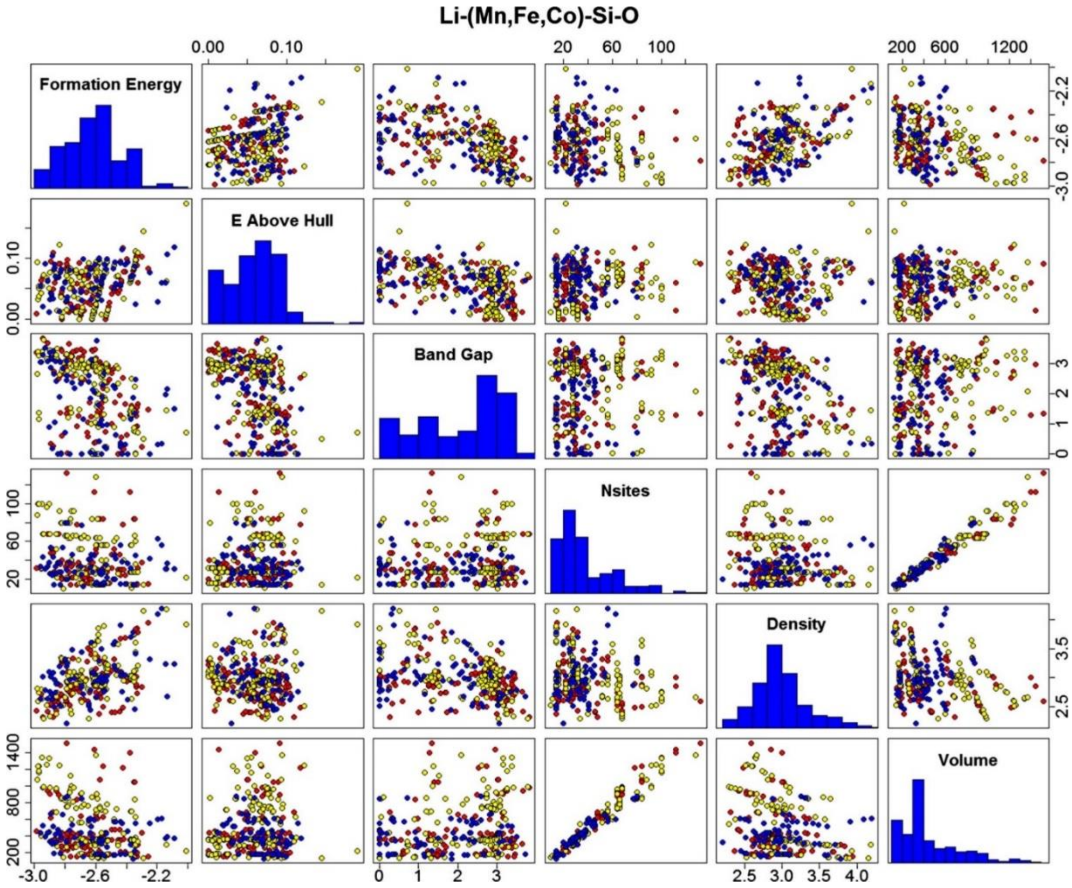

Figure 2.10 The plot of different properties pairs of Li–(Mn, Fe, Co)–Si–O cathodes according to the extracted data from MP database. The red, yellow and blue dots indicate the monoclinic, orthorhombic and triclinic crystal systems, respectively. Reproduced with permission (Attarian Shandiz and Gauvin, 2016). Copyright 2016, Elsevier Publications.

Five ML classification models, including ANN, SVM, k-NN, RF, and extremely randomized trees (ERT) were implemented by Shandiz et al. (Attarian Shandiz and Gauvin, 2016) to categorize the crystal systems of silicate-based cathodes with the composition of Li–Si–(Mn, Fe, Co)–O into three major types: monoclinic; triclinic; orthorhombic. The training dataset contained 339 cathode material data points obtained from MP (Jain et al., 2013, Jain et al., 2011b), with 5 descriptors including formation energy (E_f), energy above hull (E_H), bandgap (E_g), number of sites (N_s), and volume of unit cell (V_{uc}) (**Figure 2.10**). The prediction results indicated that the ensemble methods (RF and ERT) gave the highest accuracy of over 75% under Monte Carlo validation (Xu et al., 2004), where the N_s and V_{uc} were dominant in determining the crystal system type. More recently, Joshi et al. (Joshi et al., 2019) developed an ML-based tool to predict the voltage of electrode materials in metal-ion batteries. A total of 3,977 samples were collected from the MP database, where 237 features, such as the elemental properties of their constituents(Ward et al., 2016a) and properties of chemical compounds (Ghiringhelli et al., 2015), were initially added to the descriptor vector. A PCA (Pearson, 1901, Jolliffe, 2011) model was then performed to reduce the dimensionality of the descriptor vector to 80. The deep neural network (**Figure 2.11**a) (LeCun et al., 2015), SVM(Noble, 2006), and KRR(Vovk, 2013) model yield an R^2 value of 0.84, 0.86, and 0.86, respectively, on the prediction of voltage, therefore offering an alternative way to generate voltage profile diagrams instead of DFT methods (Zhang et al., 2018) (**Figure 2.11**b). Additionally, nearly 5,000 electrode material candidates were proposed for Na- and Ki-ion batteries via these ML models, some of which were comparable with published experimental and DFT values (Ong et al., 2011, Billaud et al., 2014, Nisar et al., 2018). Further improvement of the model performance could be implemented via the employment of different algorithms, more data, and novel ways of characterizing intercalation reactions.

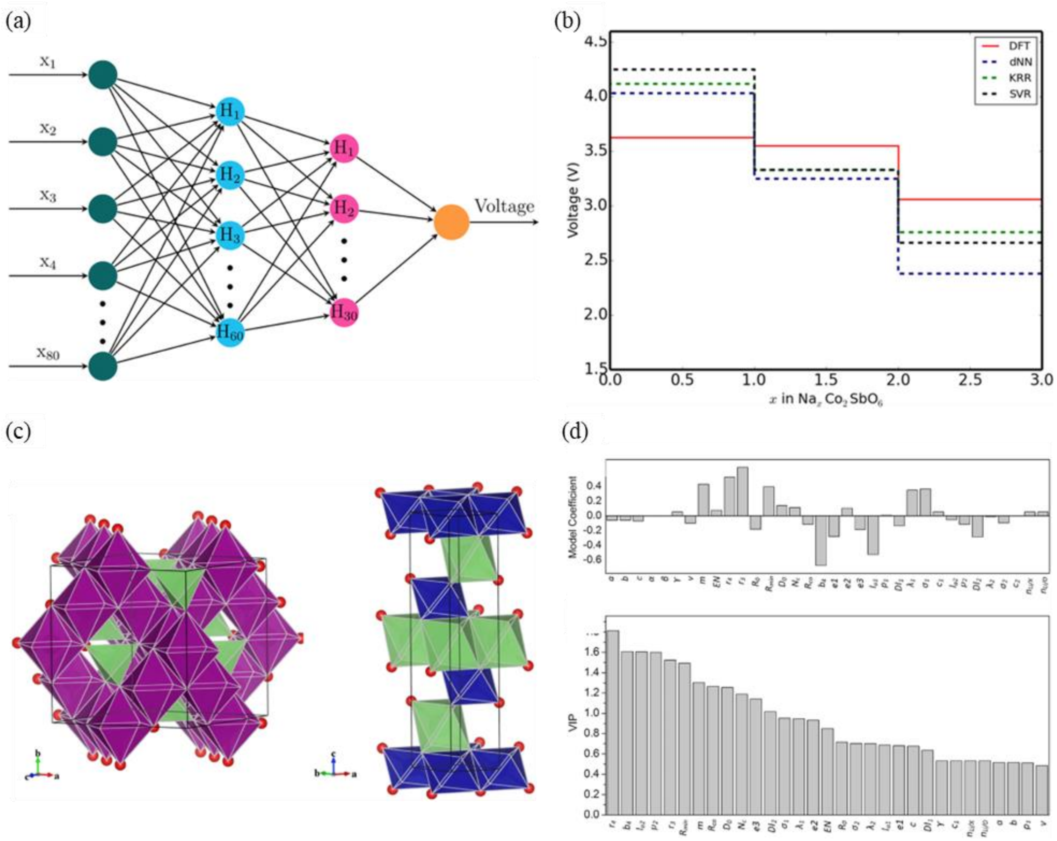

Figure 2.11 (a) Schematic diagram of the neural network employed in this study. x_i represents the input of the NN and H_i represents the nodes in the hidden layers. (b) The obtained voltage profile diagram from several ML models and DFT computation for $Na_xCo_2SbO_6$. a-b) Reproduced with permission **(Joshi et al., 2019)**. Copyright 2019, ACS Publications. (c) The scheme of crystal structures for (left) spinel LiX_2O_4 and (right) layered $LiXO_2$. (d) (top) The model coefficient plot and (bottom) variable importance plot of the independent variables for the modeling PLS. (c-d) Reproduced with permission **(Wang et al., 2017)**. Copyright 2017, Elsevier Publications.

Small volume changes of cathodes are critical for extending the cycle life of batteries.(Chen et al., 2003) Wang et al. (Wang et al., 2017) reported a methodology integrating first-principles calculations and partial least square (PLS) regression to formulate the quantitative structure-activity relationship (QSAR) of the volume change for cathode materials in Li-ion batteries. The scheme of crystal structures of the material is shown in **Figure 2.11**c. A total of 34 descriptors in five types, including element, crystal structure, composition, local distortion and electronic level, were selected to acquire the QSAR formulation (**Figure 2.11**d). It was found that the radius of X^{4+} ion and the octahedron descriptors of X contributed the most to cathode volume change. The established QSAR could be applied to a broader range of real or simulated materials. It is still challenging to design the low-strain cathode with the determined optimal combination of the descriptors, which might be realized via codoping at various atomic sites.

Data-driven innovation has emerged as a significant driver of material discovery and fundamental knowledge exploration in rechargeable alkali-ion batteries. This is typically accompanied by the integration of first-principles computation and ML techniques, which reveal implicit structure-property correlations and accelerate the high-throughput screening of electrolyte and electrode materials (Ahmad et al., 2018). Although some of the trained ML models discussed earlier might have relatively weak prediction accuracy (Ahmad et al., 2018), which might be caused by the uncertainty of DFT computation (Ahmad and Viswanathan, 2016), they still reflect the correct trends in target properties with with respect to material parameters. The selection of descriptors is of great significance to model performance. In some cases, geometric attributes and electronic properties can sufficiently describe the material and are relatively (computationally) cheap to obtain (Wang et al., 2017, Sendek et al., 2017, Ahmad et al., 2018).

2.6 Conclusion

In conclusion, this chapter provides a critical review of the recent advances in data-driven innovation of materials science and chemical engineering are elaborated. First, several data-driven frameworks, along with direct design, inverse design, and active learning, are discussed based on the flow of data and information in the data-driven process. Then, the chemical databases that store and manage material data are systematically discussed. Furthermore, the descriptors that carry the chemical information in the data-driven process are introduced. Finally, a critical discussion on how the data-driven approach is applied for various energy materials innovation and green chemical engineering is provided. The development of novel and intelligent algorithms, the capability of computational and experimental material databases to generate and store data, and the design and validation of accurate and efficient descriptors have many outcomes. Their synergistic integration is promising and effective for material discovery and industrial optimization. This thesis aims to provide a deeper understanding of these advanced technologies and to explore innovative approaches for enhancing the synergistic integration in this community at various scale, from micro to macro level, and across different sub-disciplines in material science and chemical engineering to enable carbon neutrality.

Chapter 3

Methodology and Materials

3.1 Synopsis

Chapter 3 details the methodology used in this thesis. The chapter starts with a section on the method of DFT calculations for ORR, generating the necessary data for the ML training in Chapter 4. This is followed by the information of raw materials and their corresponding suppliers, which were taken for LFP synthesis in Chapter 5. This chapter then describes the synthesis process of LFP materials using the high-temperature solid-state method. Next, various characterization measurements conducted on LFP materials are briefly described, including the characterization of the LFPs' electrochemical and physicochemical properties. Then, a discussion on the data simulation process and the ML training strategy used in Chapter 6, as well as the real-time monitoring methodology applied in the research, is presented.

3.2 DFT Calculations for ORR

All DFT calculations were performed using the Vienna ab initio simulation program, along with projected augmented wave (Blöchl, 1994) pseudopotentials and the Perdew–Burke–Ernzerhof functional (Perdew et al., 1996). According to the different crystal systems screened by the critera, the (111), (100), and (211) surfaces were cleavaged to simulate the planar and step sites. Neighboring slabs were separated by a vacuum of 15 Å to avoid spurious self-interactions. The surface irreducible Brillouin zone was sampled on the k-point mesh generated by the Monkhorst–Pack scheme. An energy cutoff of 400 eV was employed for the plane-wave basis set. The convergence threshold for electronic steps in geometry optimization was 1×10^{-5} eV. Geometries were deemed converged when the forces on each atom were below 0.02 eV $Å^{-1}$. A frequency analysis was carried out on the stable states to confirm that these represent genuine minima. All the electronic energies were corrected for zero-point energy contributions.

3.3 Materials for the synthesis of LFP

Table 3.1 The suppliers of the mainly used materials

Materials	Suppliers
$FePO_4$	Sichuan Development Lomon Co., Ltd
$LiCO_3$	Aorislithium Co., Ltd
Glucose	Aladdin Co., Ltd
PEG	Jilin Ruiji Co., Ltd
TiO_2	Ningbo Xinfu Co., Ltd
APG	Aladdin Co., Ltd

3.4 Preparation Process of LFP

The preparation method for lithium iron phosphate (LFP) is the high-temperature solid-state method. The main chemicals, including $FePO_4$, Li_2CO_3 (LC), glucose, polyethylene glycol (PEG), TiO_2, and alkyl polyglucosides (APG) with a total carbon content of 0.05wt%, are added to a ball milling jar. Zirconia beads, 3-5 times the amount of the material, are also added for ball milling. After ball milling, the mixture undergoes spray drying to obtain a yellow pre-mixed precursor. This pre-mixed precursor is then sintered at 800°C for 8 hours under an inert atmosphere to produce the APG-modified LFP. The final step involves air jet milling to achieve the desired particle size and distribution.

3.5 Characterizations of LFP

3.5.1 Characterization of electrichemical properties

The electrochemical performance characterization of the prepared LFP was conducted using coin cells and the LANHE battery testing system (Wuhan, China). During the characterization process, the active material LFP, conductive agent Super P, and binder were mixed in a weight ratio of 80:10:10. These three substances were dissolved in N-methyl-2-pyrrolidone (NMP) and homogenized using a THINKY mixer. After homogenization, the slurry was coated onto aluminum foil and vacuum dried at 120 °C for 3 hours to obtain the electrode. Lithium metal (RoHS) was used as the counter electrode, Celgard 2325 polypropylene membrane (Celgard) served as the separator, and the electrolyte consisted of 1 mol/L LiPF6 dissolved in a mixture of ethylene carbonate (EC) and dimethyl carbonate (DMC) in a 1:1 volume ratio. The entire battery assembly process was carried out in an argon-filled Mikrouna glove box. The constant current charge-discharge curves and rate performance tests were conducted using the LANHE battery testing system (Wuhan, China) at a temperature of 25°C and a voltage range of 2.25-3.75V (vs. Li/Li$^+$).

3.5.2 Characterization of physicalchemical properties

The phase and crystal structure of the four materials were determined using X-ray diffraction (PANalytical). The microstructure was characterized by field emission electron microscopy (Bruker). The powder compaction density of the materials was characterized using an electronic pressure testing machine (UTM7305).

3.6 The Simulation of Steam Boiler Operation Data for Fault Detection and Real-Time Monitoring

For the generation of the simulation dataset, Aspen Plus V10 and Aspen Plus Dynamic V10 were used as the simulator for fault simulation of the boiler. Components of the boiler were simulated using blocks, as shown in **Figure 3.1**, such as mixer (as MIXER), RGibss reactor (as combustor), pump (as soft water pump), separator (as flash drum), and three heat exchangers (as smoke tube, economizer, and condenser) and were denoted as MIXER, COMB, PUMP, DRUM, SMKTUBE, ECOMM, and COND, respectively. Assumptions of the simulation includes no heat loss for all equipment except COMB.

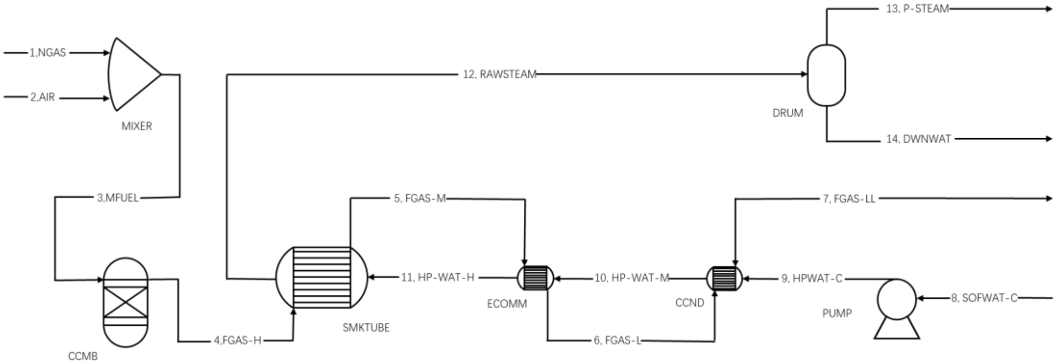

Figure 3.1 The schematic of the simulated boiler operation system.

3.7 ML-Based Data-Driven Techniques for Fault Detection and Real-Time Monitoring

3.7.1 Principal Components Analysis (PCA)

PCA is an unsupervised ML algorithm, which is regarded as a linear multivariate statistic technique and an exploratory data analysis tool for dimensionality reduction, decorrelation, denoising, and pattern recognition of a dataset. PCA seeks a linear transformation to map the original data to low-dimension latent variables, regarded as principal components (PCs), with the least loss of information (Wang et al., 2022). The target of PCA on a given dataset $\boldsymbol{X} \in \mathbf{R}^{q \times p}$ which has q sample points with p observed variables, is to determine the linear combination $\boldsymbol{Xa}$ of the p variables given by $\sum_{i=1}^{p} a_i \boldsymbol{X}_i = \boldsymbol{Xa}$ that provides the maximum variances where $\boldsymbol{a}$ is a vector that is consisted of constants a_i. The variance of the linear combination $\mathrm{var}(\boldsymbol{Xa})$ can be calculated from the covariance matrix $\boldsymbol{C} \in \mathbf{R}^{p \times p}$ of the original dataset $\boldsymbol{X}$:

$$\mathbf{var}(\boldsymbol{Xa}) = \mathbf{cov}(\boldsymbol{Xa}, \boldsymbol{Xa}) = \boldsymbol{a}^{\boldsymbol{T}} \boldsymbol{Ca} \tag{3-1}$$

The problem of seeking the $\boldsymbol{Xa}$ with maximum variance is equivalent to finding the specific $p \times 1$ vector $\boldsymbol{a}$ that maximum $\boldsymbol{a}^T\boldsymbol{Ca}$. With the additional restriction that $\boldsymbol{a}$ is a unit-norm vector, the maximum problem could be written as:(Jolliffe and Cadima, 2016)

$$\mathbf{arg}\ \max_{\boldsymbol{a}}[\boldsymbol{a^T Ca} - \boldsymbol{\lambda}(\boldsymbol{a^T a} - \mathbf{1})] \qquad (3\text{-}2)$$

where the λ is the Lagrange multiplier of the constrained optimization problem. The maximum value of variance occurs when the differentiation of equation (**3-2**) is equal to the null vector:

$$\boldsymbol{Ca} - \boldsymbol{\lambda a} = \mathbf{0} \qquad (3\text{-}3)$$

and the $\boldsymbol{a}$ is the eigenvector of $\boldsymbol{C}$**,** and λ is the corresponding eigenvalue whose value is equal to $\mathrm{var}(\boldsymbol{Xa})$. The application of the Lagrange approach obtains all of the *p* eigenvalues as the solution of new linear combinations:

$$\boldsymbol{Xa_n} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{a_{ni} X_i}\, , \boldsymbol{for\ n} = \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{p} \qquad (3\text{-}4)$$

where these linear combination $\boldsymbol{Xa}_n$ are regarded as PCs, whose elements are called PC scores. Those PCs are uncorrelated with each other due to the orthogonality of eigenvectors of $\boldsymbol{a}_n$, whose elements are called PC loadings.

For a dataset denoted as $\boldsymbol{X}^*$**,** where each column of the variables has a mean of zero, its covariance matrix can be calculated by:

$$\boldsymbol{C} = \frac{\mathbf{1}}{\boldsymbol{q}-\mathbf{1}} \boldsymbol{X}^{*\boldsymbol{T}} \boldsymbol{X}^* \qquad (3\text{-}5)$$

The eigendecomposition of $\boldsymbol{C}$ could be transformed into the singular value decomposition of $\boldsymbol{X}^*$. Let $\boldsymbol{X}^* = \boldsymbol{U\Sigma A}^T$, the eigenvectors of $\boldsymbol{X}^*\boldsymbol{X}^{*T}$ and $\boldsymbol{X}^{*T}\boldsymbol{X}^*$are the columns of unitary matrices $\boldsymbol{U} \in \mathbf{R}^{q\times q}$ and $\boldsymbol{A} \in \mathbf{R}^{p\times p}$, respectively. The main diagonal of the non-negative diagonal matrix $\boldsymbol{\Sigma}^2 \in \mathbf{R}^{p\times p}$ are the squared singular values (σ_n^2) in decreasing order, which are also the real eigenvalues (λ_n) of the $(q-1)\boldsymbol{C}$. The right singular vector $\boldsymbol{A}$ is the loading matrix, and the score matrix $\boldsymbol{X}^*\boldsymbol{A}$ can be expressed as:

$$\boldsymbol{X}^*\boldsymbol{A} = \boldsymbol{U\Sigma A^T A} = \boldsymbol{U\Sigma} \tag{3-6}$$

The *m*th PC is the *m*th column of $\boldsymbol{X}^*\boldsymbol{A}$ and $\mathrm{var}(\boldsymbol{X}^*\boldsymbol{A}_m) = \lambda_m$, which is the *m*th largest eigenvalue of $(q-1)\boldsymbol{C}$. The tr($\boldsymbol{\Sigma}^2$) is the total sum of the variances of the *p* original variables, which equals to the total variances of the *p* PCs. For the *m*th PC, its proportion of tr($\boldsymbol{\Sigma}^2$) is regarded as its contribution c_m to the total variance and the accumulated contribution ac_m of the top *m* PCs are normally the pre-defined hyperparameters to specify how many PCs are expected to be retained. The dataset $\boldsymbol{X}^*$ could be expressed as the sum of residuals and the outer products of their score matrix and loading matrix (**Figure 3.2**):

$$\boldsymbol{X}^* = \widehat{\boldsymbol{X}} + \widetilde{\boldsymbol{X}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \widehat{\boldsymbol{t}_\iota}\widehat{\boldsymbol{p}_\iota} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{p-m}} \widetilde{\boldsymbol{t}_\iota}\widetilde{\boldsymbol{p}_\iota} = \widehat{\boldsymbol{T}}\widehat{\boldsymbol{P}}^{\boldsymbol{T}} + \widetilde{\boldsymbol{T}}\widetilde{\boldsymbol{P}}^{\boldsymbol{T}} \tag{3-7}$$

where $\widehat{\boldsymbol{P}} \in \mathbf{R}^{p\times m}$ stand for the loading matrix in the PC space and consist of the largest m eigenvalues of $(q-1)\boldsymbol{C}$; $\widehat{\boldsymbol{T}} \in \mathbf{R}^{q\times m}$ is the corresponding score matrix whose columns are the selected PCs. The matrix $\widetilde{\boldsymbol{X}}$ is the residual matrix, and the columns of the $\widetilde{\boldsymbol{T}}$ and $\widetilde{\boldsymbol{P}}$ are the left $(p-m)$ PCs and eigenvectors, respectively.

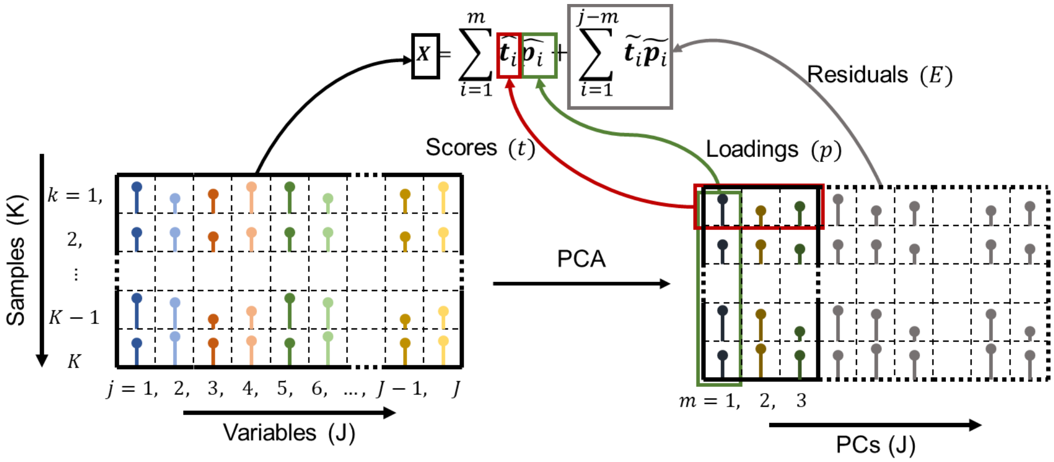
$\boxed{\boldsymbol{X}} = \sum_{i=1}^{m} \boxed{\widehat{\boldsymbol{t}_i}}\boxed{\widehat{\boldsymbol{p}_i}} + \boxed{\sum_{i=1}^{j-m} \widetilde{\boldsymbol{t}_i}\widetilde{\boldsymbol{p}_i}}$
Residuals (E)
Scores (t)
Loadings (p)
Samples (K)
$k = 1,$
2,
⋮
$K - 1$
K
$j = 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ \dots,\ J-1,\ J$
Variables (J)
PCA
$m = 1,\ 2,\ 3$
PCs (J)

Figure 3.2 The typical PCA decomposition on a dataset with J variables of K samples.

3.7.2 Multiway PCA (MPCA)

The MPCA is an extension of traditional PCA to analyze a multi-dimensional dataset. Traditional PCA can only analyze a single two-dimensional matrix (Figure 3.2) containing plenty of features, generating a small number of principal components (PCs). For a series of batch processes, the data is normally collected in a three-dimensional form $\underline{\boldsymbol{X}}$ (I×J×K), where I is the number of batches the boiler completed, J is the observed variables, and K is the time intervals (Figure 3.3a). To process the batch dataset, MPCA unfolds $\underline{\boldsymbol{X}}$ along with one of three possible directions ($\boldsymbol{X_1}(J \times IK), \boldsymbol{X_2}(I \times JK)$ and $\boldsymbol{X_3}(I \times KJ)$) and arranges it into a two-dimensional dataset to perform projection and decomposition. Every horizontal slice of the $\underline{\boldsymbol{X}}$ is a $(J \times K)$ matrix (Figure 3.3b) that records the J observed variables during the time period K of a single batch i. Similarly, a vertical slice along the variables is a $(I \times K)$ matrix (Figure 3.3c) that records the variation of a single variable j during the time period K in I different batches, and a $(I \times J)$ matrix (Figure 3.3d) is a vertical slide alone another direction that records the values of J variables of I batches at time point k. In this study, with the consideration of the variations of variables among different batches, the matrix $\underline{\boldsymbol{X}}$ is unfolded in the form of $\boldsymbol{X_3}$, where each $(I \times J)$ vertical slice of $\underline{\boldsymbol{X}}$ is put side by side along with the time series.

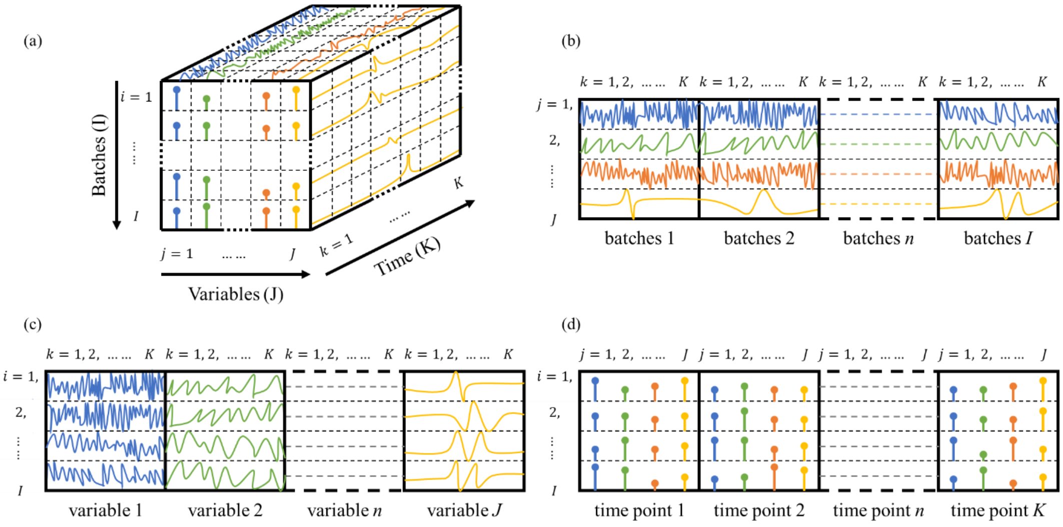

Figure 3.3 The graphic presentation of the decomposition of the three-dimensional data tensor. (a) The illustration of three-dimension data tensor $\underline{X}$ of the steam boiler working in batches. (b) Unfolding $\underline{X}$ along with batches, $X_1(J \times IK)$. (c) Unfolding $\underline{X}$ along with variables, $X_2(I \times JK)$. (d) Unfolding $\underline{X}$ along with time points, $X_3(I \times KJ)$.

The $\boldsymbol{X_3}$ unfolding allows the analysis of the variabilities among batches via arranging the data with respect to variables and time points. Figure 3.4 illustrates the relationship between the MPCA on $\underline{\boldsymbol{X}}$ and by PCA on the unfolded $\boldsymbol{X_3}$. In the $\boldsymbol{X_3}$ unfolding, each score vector element characterizes a single batch's behavior and its variability compared to other batches in X over time. Loading vectors, on the other hand, signify the maximum variance direction and offer a clear view of the batch data. Each loading vector outlines the variable changes at each time point, with its elements used as weights applied to batch variables when calculating the batch score. With the joint covariance matrix detailing variable deviations, MPCA can track variable shifts in magnitude and their correlations within batches over time.

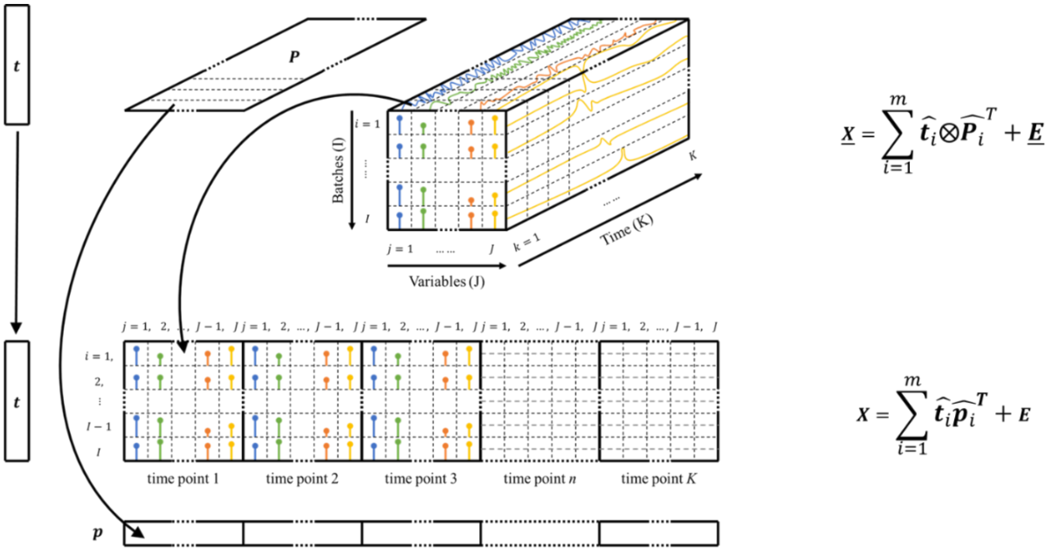

Figure 3.4 The graphic presentation of the decomposition on the three-dimensional batch data by MPCA (top) and the equivalent form of PCA on the unfolded data to obtain the $\hat{\boldsymbol{t}}$ and $\hat{\boldsymbol{p}}$.

3.7.3 Long-Short Term Memory (LSTM)

Developed from the recurrent neural network (RNN), LSTM is proposed with the addition of a forgetting mechanism to overcome the gradient explosion and vanishing problem.(Man et al., 2022, Liu et al., 2020a, Hochreiter and Schmidhuber, 1997) The architecture of LSTM is shown in Figure 3.5, where a particular memory cell similar to an accumulator and a gated neuron is encoded in. The weight of next time step is computed parallelly and the actual value of the state is copied and accumulated. A self-connection mechanism controlled by a multiplication gate which is employed to determine the moment to clear the memory content by another unit, is added in LSTM. A typical LSTM architecture consists of input gate, output gate and forget gate. Denote time interval as subscript k, input as x, cell state as C, activation function as σ, the hidden state as h, layer weights as W, bias as b, the output of input gate, output gate, forget gate and the reserved portion of the original loop layer as i, f, o and $\hat{C}$, respectively. The forward propagation process of LSTM can be expressed as:

$$\boldsymbol{i_k = \sigma(W_i[h_{k-1}, x_k] + b_i)} \tag{3-8}$$

$$\boldsymbol{f_k = \sigma\left(W_f[h_{k-1}, x_k] + b_f\right)} \tag{3-9}$$

$$\boldsymbol{o_k = \sigma(W_o[h_{k-1}, x_k] + b_o)} \tag{3-10}$$

and the update of cell information and hidden information are presented by:

$$\boldsymbol{\widehat{C}_k = \mathbf{tanh}(W_C[h_{k-1}, x_k] + b_C)} \tag{3-11}$$

$$\boldsymbol{C_k = f_k \times C_{k-1} + i_k \times \widehat{C}_k} \tag{3-12}$$

$$\boldsymbol{h_k = o_k \times \mathbf{tanh}(C_k)} \tag{3-13}$$

Different from RNN, the LSTM does not process all historical data but only selected length of data backward within the forgetting mechanism, and add the input information of the next time point to the backward transfer.(Man et al., 2022) Then the backpropagation will be employed to updated the hyperparameter to optimize the model.(Rumelhart et al., 1986) The LSTM have been widely applied to process sequence data such as natural language processing,(Gers and Schmidhuber, 2001) river daily runoff,(Man et al., 2022) and heart rate signals.(G et al., 2018)

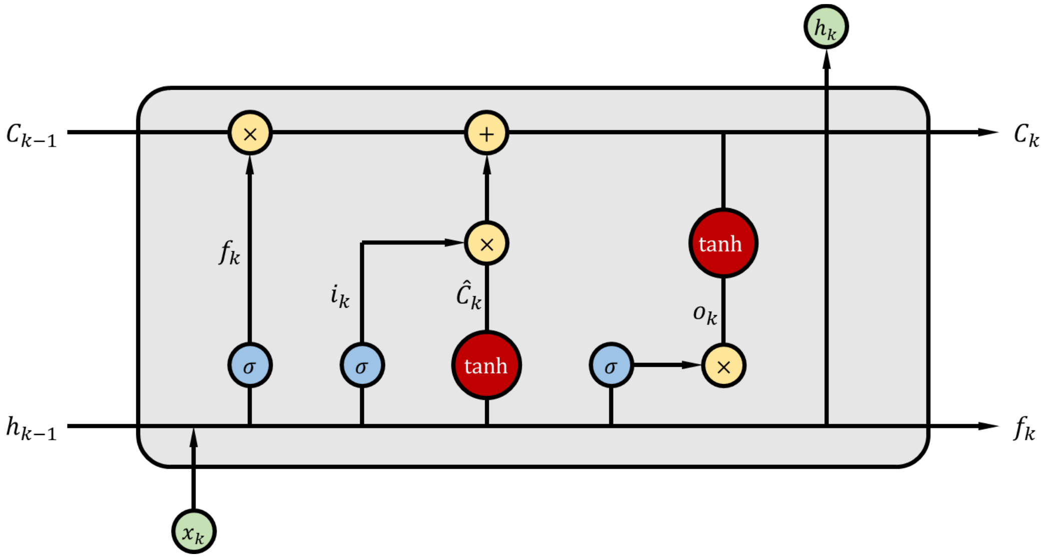

Figure 3.5 The Architecture of LSTM. $\boldsymbol{C_{k-1}}$ and $\boldsymbol{h_{k-1}}$ represent the state of the cell and hidden layer of last time point, respectively.

3.7.4 ML-Assisted Real-Time Monitoring of Industrial Process.

The loading matrix ($\widehat{\boldsymbol{P}}$) from the previous MPCA analysis encapsulates the data structure information, highlighting deviations of process variables from the mean trajectories under standard operating conditions. By capturing the pattern of the data at each time point, it enables the real-time examination of device behavior, making it possible to identify and respond to deviations as they occur. When testing a new batch for abnormal behavior detection, the existing loading matrix will be utilized to derive corresponding PCs and residuals for the test batch $\underline{\boldsymbol{X}}_{test}(1 \times K \times J)$. The $\underline{\boldsymbol{X}}_{test}$ could then be pre-processed and unfolded into a one-dimension vector $x_{test}^{*}(1 \times KJ)$ and the corresponding PC vector t_{test} and residuals $\tilde{x}_{test}$ could be calculated by:

$$\boldsymbol{t}_{\boldsymbol{test}} = \boldsymbol{x}_{\boldsymbol{test}}^{*}\widehat{\boldsymbol{P}} \tag{3-14}$$

$$\widetilde{\boldsymbol{x}}_{\boldsymbol{test}} = \boldsymbol{x}_{\boldsymbol{test}}^{*} - \boldsymbol{t}_{\boldsymbol{test}}\widehat{\boldsymbol{P}}^{\boldsymbol{T}} \tag{3-15}$$

For an ongoing test batch, only $\boldsymbol{k}$ rows exist in $\underline{\boldsymbol{X}}_{\boldsymbol{test}}$ due to the absence of future measurements. Several solutions are proposed to address this, such as assuming the variables are multi-normally distributed and then using T^2 statistics to perform an F test, or establishing a series of PCA models at each time point (Nomikos and MacGregor, 1995). The $\boldsymbol{t}_{\boldsymbol{test},\boldsymbol{k}}$ can be expressed as the cross product of the observation vector $\boldsymbol{x}_{\boldsymbol{test},\boldsymbol{k}}^{*}$ and the corresponding rows of the loading matrix $\widehat{\boldsymbol{P}}_{\boldsymbol{k}}$. The use of $\widehat{\boldsymbol{P}}_{\boldsymbol{k}}$ assumes that the future data will maintain a consistent data structure due to the inner correlation encoded by the MPCA model. Therefore, the problem becomes finding the value of $\boldsymbol{t}_{\boldsymbol{test},\boldsymbol{k}}$ that provides the least squared error of:

$$\boldsymbol{x}_{\boldsymbol{test},\boldsymbol{k}}^{*} = \boldsymbol{t}_{\boldsymbol{tesk},\boldsymbol{k}}\widehat{\boldsymbol{P}}_{\boldsymbol{k}}{}^{\boldsymbol{T}} \tag{3-16}$$

whose the least square solution could be expressed as:

$$\boldsymbol{t}_{\boldsymbol{tesk,k}} = \boldsymbol{x}^{*}_{\boldsymbol{test,k}} \widehat{\boldsymbol{P}}_{\boldsymbol{k}}{}^{\boldsymbol{T}} \left(\widehat{\boldsymbol{P}}_{\boldsymbol{k}} \widehat{\boldsymbol{P}}_{\boldsymbol{k}}{}^{\boldsymbol{T}}\right)^{-\mathbf{1}} \tag{3-17}$$

where $\boldsymbol{x}^{*}_{test,k}$ is the scaled and centralized data according to the reference dataset.

The matrix $\widehat{\boldsymbol{P}}_k$ has the dimensionality of $(k \times J \times m)$, including all the loading vectors of J variables up to time point k with m selected PCs. Due to the orthogonality property of loading vectors, the matrix $\widehat{\boldsymbol{P}}_k \widehat{\boldsymbol{P}}_k{}^{T}$ is well-conditioned.(Nomikos and MacGregor, 1995) Similar to the MPCA implementation, the score plot and residual of the test batch are used to determine whether it is normal. The $\boldsymbol{t}_{tesk,k}$ represents the project of $\boldsymbol{x}^{*}_{test}$ to the reduced m-dimensional space determined by the MPCA model.

3.7.5 LSTM-MPCA employed for Real-Time Monitoring and Early Warning

While MPCA is effective for real-time monitoring, it has limitations in predicting future behavior. To address this, MPCA is enhanced with Long Short-Term Memory (LSTM) networks for time series forecasting. The proposed ensemble method combines four LSTM sub-models, each with three hidden layers, to predict the next time step of a given time series. Each hidden layer consists of an LSTM unit with 256 memory cells, selected activation functions, and dropout layers. The four models have different activation functions and dropout rates and are trained independently. A dense layer with the same number of neurons as the input features is added to the output of the third LSTM layer to make the final prediction.

During the prediction phase, the outputs of the model are concatenated, and the average of the outputs is computed as the final prediction using a voting mechanism. This voting mechanism leverages the strengths of each individual model and mitigates their weaknesses. By integrating LSTM with MPCA, the system not only performs real-time monitoring but also effectively predicts future boiler behaviors, combining MPCA's strengths for immediate anomaly detection with LSTM's capabilities for forecasting, providing a comprehensive solution for steam boiler fault detection and monitoring.

The proposed model architecture was employed on the simulated dataset, with the training, validation, and testing sets split in a ratio of 8.5:1:0.5. During training, each LSTM sub-model takes in a sequence of 16 timesteps to predict the variables and MPCA score values of the next time points. The model was trained using the Mean Squared Error (MSE) loss function and the Adamax optimizer with a learning rate of 0.001 (Kingma and Ba, 2014). The performance of the model was evaluated using the R-squared metric, which measures the proportion of the variance in the target variable explained by the model.

3.8 Statistics and Control Limits for Fault Detection and Real-Time Monitoring

Once the MPCA model is developed based on the historical reference batch data, the structural information of the reference data is contained in the generated $\hat{\boldsymbol{P}}$, which can indicate how the measured variables would vary at one specific time point. For a new batch sample, its deviation information over the time history is contained by the corresponding $\hat{\boldsymbol{t}}$ and residuals $\boldsymbol{e}$, which will be compared against the reference distribution to evaluate its behavior. If the future behavior of the new batch predicted by LSTM is consistent with the reference distribution, its corresponding $\hat{\boldsymbol{t}}$ should fall in the control region with acceptable variation and the $\boldsymbol{e}$ is sufficiently small (Nomikos and MacGregor, 1994). In this study, Hotelling's T^2 statistics (Anderson, 2003, Tracy et al., 1992), squared prediction error and a combined statistic(Yue and Qin, 2001) are applied to define the control range and evaluate the behavior of samples.

3.8.1 Hotelling's T^2 statistics

Hotelling's T^2 statistics are used to measure the variability of the observation vectors by evaluating the distance between the sample points and the origin of the PC space. It can be calculated by:

$$\boldsymbol{T^2}(\boldsymbol{x}) = \boldsymbol{x}\widehat{\boldsymbol{P}}\boldsymbol{\Sigma}_{\boldsymbol{m}}^{-2}\widehat{\boldsymbol{P}}^{\boldsymbol{T}}\boldsymbol{x}^{\boldsymbol{T}} = \hat{\boldsymbol{t}}\boldsymbol{\Sigma}_{\boldsymbol{m}}^{-2}\hat{\boldsymbol{t}}^{\boldsymbol{T}} \tag{3-18}$$

where $\boldsymbol{\Sigma}_m^2 \in \mathbf{R}^{m\times m}$is a diagonal matrix that concludes the mth row and column of $\boldsymbol{\Sigma}^2$ and $\boldsymbol{x}$ is a $1 \times p$ observation vector that represents a sample point of the dataset $\boldsymbol{X}^*$. Based on the assumption that the variables of the samples follow the multivariable normal distribution, the corresponding T^2 statistics of normal samples follow the chi-square (χ^2) distribution with m degrees of freedom and should not be greater than the threshold. Therefore, the control limit related to $\hat{\boldsymbol{t}}$ is based on multivariate distribution and the threshold T_α of T^2 statistics with a given level of significance α could be computed as (Anderson, 2003):

$$\boldsymbol{T^2(x) \leq T^2_\alpha = \frac{m(q^2-1)}{q(q-m)} F_{m,q-m;\alpha}} \qquad \text{(3-19)}$$

where $F_\alpha(m, q-m)$ is the critical value of *F*-distribution with *m* and *q-m* under α significance level.

3.8.2 *Q* statistics

In the context of multivariate process control, the residual term $\boldsymbol{e}$ of a batch sample indicates the degree of unexplained variability that remains after the MPCA model has been applied. To effectively monitor the residual term, the *Q* statistics, also known as squared prediction error (SPE), is often utilized. This approach quantifies the variation of the observation vectors projected onto the residual space, and is widely recognized for its accuracy and effectiveness, it can be expressed as:

$$\boldsymbol{Q(x) = \|e\|^2 = \left\|x\widetilde{P}\widetilde{P}^T\right\|^2 = \left\|x\left(I - \widehat{P}\widehat{P}^T\right)\right\|^2} \qquad \text{(3-20)}$$

If their *Q* statistics are under the threshold for observation vectors, then the process could be regarded as a normal state. Jackson and Mudholkar (Jackson and Mudholkar, 1979) give out the formula to calculate the threshold Q_α when the residual vector follows the normal distribution at a given significance level α:

$$\boldsymbol{Q^2_\alpha = \theta_1 \left[\frac{c_\alpha\sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0(h_0-1)}{\theta_1^2}\right]^{1/h_0}} \qquad \text{(3-21)}$$

where

$$\boldsymbol{\theta_i = \sum_{j=m+1}^{q} \lambda_j^i\, , for\ i = 1, 2, 3} \tag{3-22}$$

$$\boldsymbol{h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}} \tag{3-23}$$

and λ_j is the eigenvalue of the covariance matrix of the original dataset $\boldsymbol{X}^*$, c_α is the critical value of a standard normal deviate at the $(1 - \alpha)$ percentile, *m* is the number of the selected PCs and q is the dimension of the original dataset. Besides, Qin (Joe Qin, 2003) gives a simplified form of the Q_α^2:

$$\boldsymbol{Q_\alpha^2 = g\chi_{h;\,\alpha}^2} \tag{3-24}$$

where $g = \theta_2/\theta_1$ and $h = \theta_1^2/\theta_2$. Besides, Nomikos et al.(Nomikos and MacGregor, 1995) gives the control limit on the SPE at significance level a for time point *k*:

$$\boldsymbol{Q_{\alpha,k}^2 = (v/2m)_k\chi_{(2m^2/v),k;\,\alpha}^2} \tag{3-25}$$

where m and v are the mean and variance of observation vector $\boldsymbol{x}_{test,k}^*$ of the test batch, respectively. **Figure 3.6** illustrates the case where PCA is applied to a dataset with three variables, generating a two-dimensional PC space, and the control limits defined by T^2 and Q statistics, respectively.

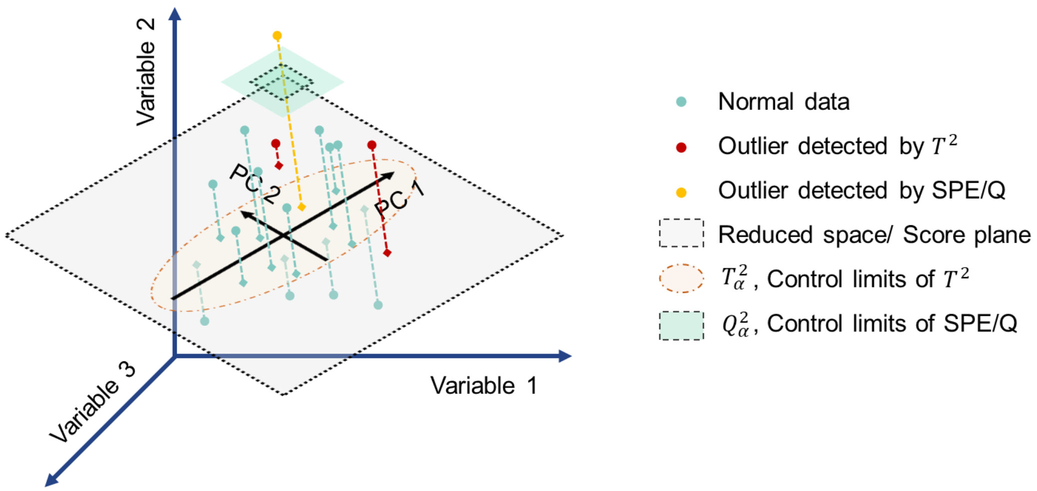
Variable 2
Variable 1
Variable 3
PC 1
PC 2
● Normal data
● Outlier detected by T^2
● Outlier detected by SPE/Q
Reduced space/ Score plane
T^2_α, Control limits of T^2
Q^2_α, Control limits of SPE/Q

Figure 3.6 The control limits defined by T^2 and Q statistics in the scenario where PCA is applied to a dataset with three variables, generating two principal components.

3.8.3 The combined index φ

In practice, the T^2 and Q statistics may yield inconsistent outcomes when used together to detect faults, as they evaluate different aspects of data deviation. To address this issue, Yue and Qin (Yue and Qin, 2001) introduced a new index (φ), which combines the T^2 and Q statistics for more effective fault detection. As φ is utilized for online monitoring, for the data recorded at time point k at a test batch, its index $\varphi(\boldsymbol{x}^*_{test,k})$ could be expressed as:

$$\boldsymbol{\varphi}(\boldsymbol{x}^*_{test,k}) = \frac{\boldsymbol{SPE}(\boldsymbol{x}^*_{test,k})}{\boldsymbol{Q}^2_{\alpha,k}} + \frac{\boldsymbol{T}^2(\boldsymbol{x}^*_{test,k})}{\boldsymbol{T}^2_{\alpha,k}} = \boldsymbol{x}^*_{test,k}\boldsymbol{\Phi}_k\boldsymbol{x}^{*\ \boldsymbol{T}}_{test,k}$$

(3-26)

where $\varphi\left(\boldsymbol{x}_{test,k}^{*}\right)$ is a quadratic function with respect to the observation vector $\boldsymbol{x}_{test,k}^{*}$, and $\boldsymbol{\Phi}_{k}$ is a symmetric and positive definite matrix and could be expressed as:

$$\boldsymbol{\Phi}_{\boldsymbol{k}}=\frac{\boldsymbol{I}-\widehat{\boldsymbol{P}}_{\boldsymbol{k}}\widehat{\boldsymbol{P}}_{\boldsymbol{k}}{}^{\boldsymbol{T}}}{\boldsymbol{Q}_{\boldsymbol{\alpha},\boldsymbol{k}}^{2}}+\frac{\widehat{\boldsymbol{P}}_{\boldsymbol{k}}\boldsymbol{\Sigma}_{\boldsymbol{m},\boldsymbol{k}}^{-2}\widehat{\boldsymbol{P}}_{\boldsymbol{k}}{}^{\boldsymbol{T}}}{\boldsymbol{T}_{\boldsymbol{\alpha},\boldsymbol{k}}^{2}} \tag{3-27}$$

The computation of $\boldsymbol{\Phi}_{k}$ requires the corresponding thresholds of both $T_{\alpha,k}^{2}$ and $Q_{\alpha,k}^{2}$ statistics for a given confidence limit α. The $\widehat{\boldsymbol{P}}_{k}$ stand for the loading matrix in the principal components space at the specific time point, which contains the structure information of the historical reference dataset about how would the variable measurement varies from the average trajectories under normal operation. Based on the assumption that the quadratic form of $\boldsymbol{x}_{test,k}^{*}$ is multi-normally distributed. The corresponding threshold $\varphi_{\alpha,k}$ is given by:(Yue and Qin, 2001)

$$\boldsymbol{\varphi}_{\boldsymbol{\alpha},\boldsymbol{k}}=\boldsymbol{g}_{\boldsymbol{k}}\boldsymbol{\chi}_{\boldsymbol{h}_{\boldsymbol{k}};\,\boldsymbol{\alpha},\boldsymbol{k}}^{\boldsymbol{2}} \tag{3-28}$$

and the coefficient g and degree of freedom h could be expressed as:

$$\boldsymbol{g}_{\boldsymbol{k}}=\frac{\boldsymbol{tr}\left((\boldsymbol{C}_{\boldsymbol{k}}\boldsymbol{\Phi}_{\boldsymbol{k}})^{\boldsymbol{2}}\right)}{\boldsymbol{tr}(\boldsymbol{C}_{\boldsymbol{k}}\boldsymbol{\Phi}_{\boldsymbol{k}})} \quad \boldsymbol{and} \quad \boldsymbol{h}_{\boldsymbol{k}}=\frac{\left(\boldsymbol{tr}(\boldsymbol{C}_{\boldsymbol{k}}\boldsymbol{\Phi}_{\boldsymbol{k}})\right)^{\boldsymbol{2}}}{\boldsymbol{tr}\left((\boldsymbol{C}_{\boldsymbol{k}}\boldsymbol{\Phi}_{\boldsymbol{k}})^{\boldsymbol{2}}\right)} \tag{3-29}$$

where the $\boldsymbol{C}_{k}$ denotes the covariance matrix of the historical dataset $\boldsymbol{X}_{k}^{*}$. For a confidence level of α, a fault is detected by φ if

$$\boldsymbol{\varphi}\left(\boldsymbol{x}_{\boldsymbol{test},\boldsymbol{k}}^{*}\right)\geq\boldsymbol{\varphi}_{\boldsymbol{\alpha},\boldsymbol{k}}=\boldsymbol{g}_{\boldsymbol{k}}\boldsymbol{\chi}_{\boldsymbol{h}_{\boldsymbol{k}};\,\boldsymbol{\alpha},\boldsymbol{k}}^{\boldsymbol{2}} \tag{3-30}$$

where χ^{2} denotes the chi-square distribution. Despite the higher computational demand of the combined index, which generates a large covariance matrix for every time point during online monitoring, this study chose to use it because it provides a simplified means to detect outliers, and control lines can be calculated offline in advance.

3.9 Conclusion

This chapter has outlined the comprehensive methodology employed throughout this thesis. The chapter provided a detailed description of the DFT calculations used to generate essential data for machine learning training in Chapter 4, forming the foundation for catalyst optimization. Additionally, it covered the sourcing of raw materials and their suppliers for the synthesis of LFP, setting the stage for the experimental work discussed in Chapter 5. The high-temperature solid-state method, utilized for LFP synthesis, was also explained, along with the corresponding characterization techniques employed to evaluate the electrochemical and physicochemical properties of the materials.

Furthermore, this chapter introduced the data simulation process and machine learning strategy that were crucial for optimizing LFP synthesis parameters, as discussed in Chapter 6. Lastly, the real-time monitoring methodology, based on machine learning tools, was presented, highlighting its significance for the broader goal of industrial process optimization.

The methodologies described in Chapter 3 provide the technical framework for advancing the understanding of material properties and the application of data-driven tools to optimize energy materials and industrial processes, integrating computational, experimental, and ML techniques to contribute to innovations in material science and chemical engineering.

Chapter 4

Data-Driven Structural Descriptor for Predicting Platinum-Based Alloys as Oxygen Reduction Electrocatalysts

Part of the content in chapter has been published in:

ZHANG, X.[#], **WANG, Z**.[#], LAWAN, A. M., WANG, J., HSIEH, C.-Y., DUAN, C., PANG, C. H., CHU, P. K., YU, X.-F. and ZHAO, H. (2023). "Data-driven structural descriptor for predicting platinum-based alloys as oxygen reduction electrocatalysts." *InfoMat.* 2023; 5(6): e12406.

4.1 Synopsis

Due to the increasing global demand for carbon-neutral and fossil-free energy systems, extensive research is being conducted on efficient and inexpensive electrocatalysts for catalyzing the kinetically sluggish oxygen reduction reaction (ORR) at the cathode of fuel cells. Platinum (Pt)-based alloys are considered promising candidates for replacing expensive Pt catalysts. However, the current screening process of Pt-based alloys is time-consuming and labor-intensive, and the descriptor for predicting the activity of Pt-based catalysts is generally inaccurate. A strategy was proposed combining high-throughput first-principles calculations and machine learning to explore the descriptor used for screening Pt-based alloy catalysts with high Pt utilization and low Pt consumption. Among the 77 prescreened candidates, 5 potential candidates were identified for catalyzing ORR with low overpotential. Furthermore, during the second and third rounds of active learning, more Pt-based alloy ORR candidates were identified based on the relationship between the structural features of Pt-based alloys and their activity. The role of structural features in Pt-based alloys was highlighted, and it was found that the difference between the electronegativity of Pt and the heteroatom, the number of valence electrons of the heteroatom, and the ratio of heteroatoms around Pt are also the main factors affecting the activity of ORR. More importantly, the combination of these structural features can be used as a structural descriptor for predicting the activity of Pt-based alloys. It is believed that the findings of this study will provide new insights for predicting ORR activity and contribute to exploring

Pt-based electrocatalysts with high Pt utilization and low Pt consumption experimentally.

4.2 Introduction

The oxygen reduction reaction (ORR) at the cathode of proton exchange membrane fuel cells presents a significant challenge due to its slow electrochemical kinetics. To address this issue, a robust and effective catalyst that can enhance the electrochemical kinetics of ORR is necessary (Wang et al., 2019). Currently, platinum (Pt) is regarded as a promising catalyst; however, its high cost limits its widespread application. A key strategy to mitigate this limitation involves the nanostructuring or alloying of pure Pt-based catalysts with nonprecious metals (Jaouen et al., 2011, Beermann et al., 2017). Pt-based alloys have exhibited remarkable stability and excellent electrocatalytic performance, making them a viable alternative to pure Pt catalysts (Chung et al., 2015, Fan et al., 2022, Gong et al., 2022, Hwang et al., 2011, Liu et al., 2022a, Stamenkovic et al., 2006, Vej-Hansen et al., 2017, Wu et al., 2020). Nonetheless, the presence of different elements with varying mixing ratios in alloyed catalysts introduces numerous chemical features and adsorption sites, which are absent in their pristine form (Li et al., 2022b, Peng et al., 2022). Furthermore, the dynamic chemical space of these alloyed catalysts complicates the screening process for functional catalysts (Goldsmith et al., 2018).

The trial-and-error method is commonly used to design ORR catalysts in traditional experiments. However, this approach is limited by cumbersome synthetic procedures and the need for in situ characterization (Wang et al., 2021b). With advancements in first-principles methods and computational resources, theoretical modeling offers new opportunities for rational catalyst design (Chen et al., 2020a, Wei et al., 2019). Creating extensive databases based on first-principles results and identifying materials with desired properties from these databases is an efficient and powerful approach for material design. Nevertheless, this method typically requires a reliable descriptor model that can easily evaluate and correlate a candidate material's intrinsic properties with target properties such as activity and selectivity (Wu et al., 2022). Thus, accurately identifying such descriptors can accelerate and improve the catalyst selection process.

Extensive research has been conducted to identify and utilize descriptors for establishing property correlations in materials. For instance, the linear relationship between the reaction free energy and activation energy in heterogeneous catalysis (van Santen et al., 2010), and the linear relationship between the d-band center of a clean surface and the adsorption energy of molecules on that surface (Nørskov et al., 2009, Hammer and Norskov, 1995), are examples of such descriptors. Although these descriptors are well-studied and widely used due to their simplicity and clear physical meaning, they can be imprecise. Consequently, an increasing number of researchers are focusing on overcoming the limitations imposed by these approximations (Zhang et al., 2019, Ding et al., 2021).

Furthermore, structural factors such as the diversity of catalyst structures and the local environment around adsorption sites have become significant in determining catalyst performance (Yang et al., 2023, Kim et al., 2022). In real-world conditions, the intricate relationship between catalyst performance and structure complicates the reliable characterization of catalytic performance (Xu et al., 2022, Wu et al., 2021). Fortunately, ML techniques and high-throughput calculations can uncover this complex relationship between catalytic activity and structure. These techniques not only enable accurate and efficient structure optimization (Chen et al., 2022) but also offer insights into the catalytic properties of materials and predict the catalytic properties of unknown materials (Li et al., 2017b, Noh et al., 2018, Andersen et al., 2017, Chen et al., 2021, Zhong et al., 2020b). Therefore, for precise modeling of alloyed catalyst properties, more reliable methods should be employed, incorporating data-driven descriptors and chemical descriptors (e.g., adsorption energy, coordination numbers) (Andersen and Reuter, 2021, Jinnouchi and Asahi, 2017, Zhou et al., 2020, Weng et al., 2020, Andersen et al., 2019).

This chapter adopts a workflow that utilizes ML and high-throughput calculations to accelerate the discovery of Pt-based alloy catalysts. By combining first-principles calculations with compressed-sensing data-analytics methodology, Pt-based alloys for ORR applications are prescreened by identifying descriptors based on the properties of their different compositions and structures. The ratio of heteroatoms around Pt, the difference in electronegativity between Pt and heteroatoms, and the valence electrons of Pt and heteroatoms are predicted to be indicative of the ORR activity in the alloy. The recently developed Sure Independence Screening and

Sparsifying Operator (SISSO), a state-of-the-art compressed-sensing-based approach, is used to identify key descriptive parameters (Ouyang et al., 2019). Through this method, the ratio of heteroatoms around Pt, the difference in electronegativity between Pt atoms and heteroatoms, and their valences were identified as structural descriptors capable of predicting the ORR activity in alloyed catalysts. These results are expected to provide a valuable dataset for experimentalists to further investigate the predicted ORR activity and for data scientists to construct ML models for ORR performance predictions.

4.3 Results and Discussion

4.3.1 Data Collection and Material Prescreening

A data-driven scheme is proposed to explore potential Pt-based alloys as highly efficient ORR electrocatalysts. Initially, datasets of Pt-based alloys are curated from the created materials database (ICSD and MP). High-throughput screening techniques are then employed to identify Pt-based binary alloys exhibiting highly efficient ORR performance. High-throughput density functional theory (DFT) calculations are utilized to study the ORR reaction mechanism and identify the rate-determining step. Based on the redox potential of this step, machine learning is used to derive descriptors capable of identifying and classifying highly efficient catalysts from the datasets, utilizing SISSO to pinpoint the best low-dimensional descriptors from numerous candidates. Finally, a structure–activity relationship and prediction model are established, which can be used to screen candidate materials with suitable properties for ORR. Active learning and reverse design are included for subsequent rounds of screening. **Figure 4.1** summarizes this process. The overall goal is to identify Pt$_n$M$_m$ alloys that are stable and possess good ORR activity.

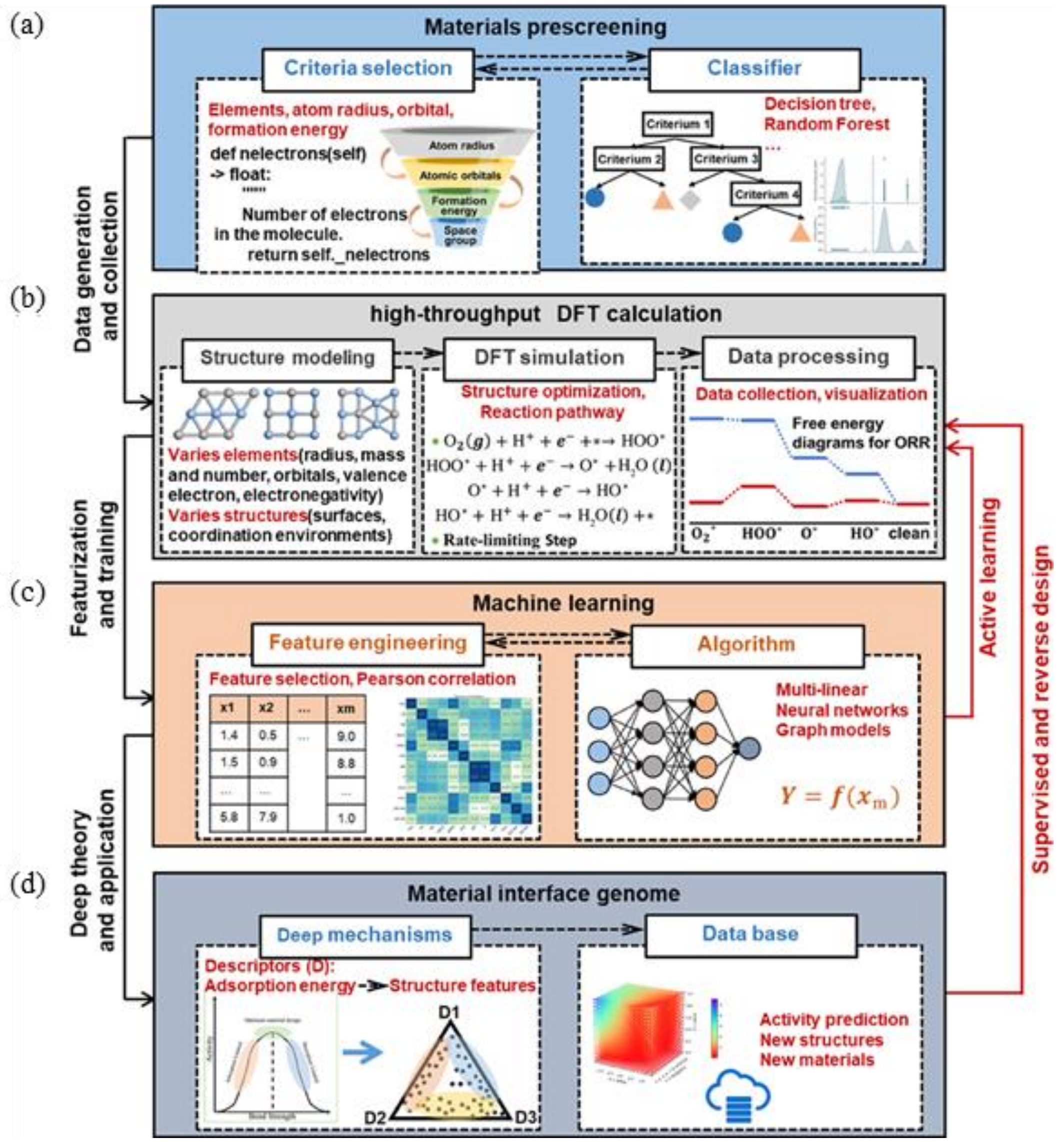
(a)
Materials prescreening
Criteria selection
Classifier
Elements, atom radius, orbital, formation energy
def nelectrons(self)
-> float:
"""
Number of electrons in the molecule.
return self._nelectrons
Atom radius
Atomic orbitals
Formation energy
Space group
Decision tree, Random Forest
...
Criterium 1
Criterium 2
Criterium 3
Criterium 4
Data generation and collection
(b)
high-throughput DFT calculation
Structure modeling
DFT simulation
Data processing
Varies elements(radius, mass and number, orbitals, valence electron, electronegativity)
Varies structures(surfaces, coordination environments)
Structure optimization, Reaction pathway
$O_2(g) + H^+ + e^- + * \rightarrow HOO^*$
$HOO^* + H^+ + e^- \rightarrow O^* + H_2O(l)$
$O^* + H^+ + e^- \rightarrow HO^*$
$HO^* + H^+ + e^- \rightarrow H_2O(l) + *$
Rate-limiting Step
Data collection, visualization
Free energy diagrams for ORR
O_2^*
HOO^*
O^*
HO^*
clean
Featurization and training
(c)
Machine learning
Feature engineering
Algorithm
Feature selection, Pearson correlation
x1 x2 ... xm
1.4 0.5 ... 9.0
1.5 0.9 8.8
... ...
5.8 7.9 1.0
Multi-linear
Neural networks
Graph models
$Y = f(x_m)$
Active learning
Supervised and reverse design
Deep theory and application
(d)
Material interface genome
Deep mechanisms
Data base
Descriptors (D):
Adsorption energy → Structure features
D1
D2
D3
Activity prediction
New structures
New materials

Figure 4.1 A workflow for constructing machine learning (ML) models for predicting Platinum-based alloys as oxygen reduction reaction (ORR) electrocatalysts. Four major steps are involved in this workflow: (a) material prescreening, (b) high-throughput density functional theory (DFT) calculation, (c)

machine learning, and (d) material interface genome. Based on these steps, the data are generated and collected, as well as featured and trained to produce the deep theory and further application. Various databases and model packages have enabled a much easier experience of model construction.

To curate the Pt$_n$M$_m$ alloys datasets, a high-throughput screening approach is adopted to screen elements from the periodic table that should be alloyed with Pt while being as stable as possible. We focus on only the materials that have been reported at the current stage from experiments and theoretical calculations. There are ~160 000 and ~140 000 materials in ICSD and MP, respectively. And ~1500 Pt-based binary alloys were found in the database. As shown in **Figure 4.2**, five criteria were applied to prescreen alloy formation by Pt and the other metal elements in the ICSD and the MP database based on the radius, orbital configuration of the transition metal atom, formation energy, crystal system, and the atom ratio between the transition metal and Pt atom of the compounds. These criteria are considered while training ML models based on the label of suitability for catalyzing ORR. The difference of atomic radius between Pt and the other elements is the first prescreening criterion, which can be used to evaluate the stability of an alloy during the ORR reaction process (Deshmukh et al., 2018, Guo et al., 2011). In this step, the upper bound for the screening is set to 0.3 Å; a higher difference of atomic radius would make it difficult to keep the morphology of the alloy during the electrochemical tests. The second criterion is the atom orbital of the metal elements, where the 3d-5d metals are chosen. The selection of 3d and 4d transition metals in the alloy formation with platinum was based on their unique electronic properties, which are crucial in optimizing the catalytic performance of Pt-based alloys. Transition metals, particularly from the 3d and 4d series, have distinct d-orbitals that interact with platinum, modifying its electronic structure and improving its catalytic activity. These metals are known to influence the binding energies of intermediates, which is especially important in catalytic reactions like the oxygen reduction reaction (ORR). The use of these transition metals, as well as their atomic orbitals, has been supported by numerous computational and experimental studies that demonstrate their effectiveness in tuning the catalytic properties of platinum alloys.

Furthermore, the formation energy of the alloy is the third criterion, because a stable material is needed for high-performance electrochemical catalysts. Considering the possible uncertainties/errors of the formation energy associated with DFT in the database, we slightly loosen the restriction to 50 meV as the threshold value. The fourth criterion is the crystal system, and cube crystal is chosen, which is the same as that of Pt unit cell. By restricting our calculations, the PtM, Pt_3M, and PtM_3 alloys are screened since they are widely studied and experimental evidence suggests that more excellent activity of ORR over the surfaces of these alloys was observed than that over the pure Pt surface (Li et al., 2017a, Bing et al., 2010, Greeley et al., 2009). The atom ratios between platinum and transition metals, 1:3, 1:1, and 3:1, were chosen based on their representation of well-established ordered alloy phases observed in Pt-based alloys. These ratios cover a range of compositional extremes, each offering distinct catalytic behaviors that are ideal for exploring alloy properties in catalytic applications. The 1:3 ratio (Pt_3M) represents platinum-rich alloys, which typically exhibit high stability and durability, making them suitable for long-term catalytic processes like oxygen reduction reaction (ORR). The 1:1 ratio (PtM) corresponds to equiatomic alloys, which balance catalytic activity and stability by optimizing the electronic interaction between platinum and the transition metal, enhancing the catalyst's performance. The 3:1 ratio (PtM_3) focuses on transition-metal-rich alloys, where the transition metal plays a dominant role in modifying the alloy's electronic structure. While these alloys may have lower stability, they offer unique catalytic properties that are valuable for investigating the upper limits of electronic modification in platinum. Together, these ratios represent a comprehensive approach to studying the catalytic behavior of Pt-based alloys, enabling a broad exploration of their performance across different phases. It is noteworthy that the Pt alloy containing lanthanide and actinide metals is not excluded in this chapter.

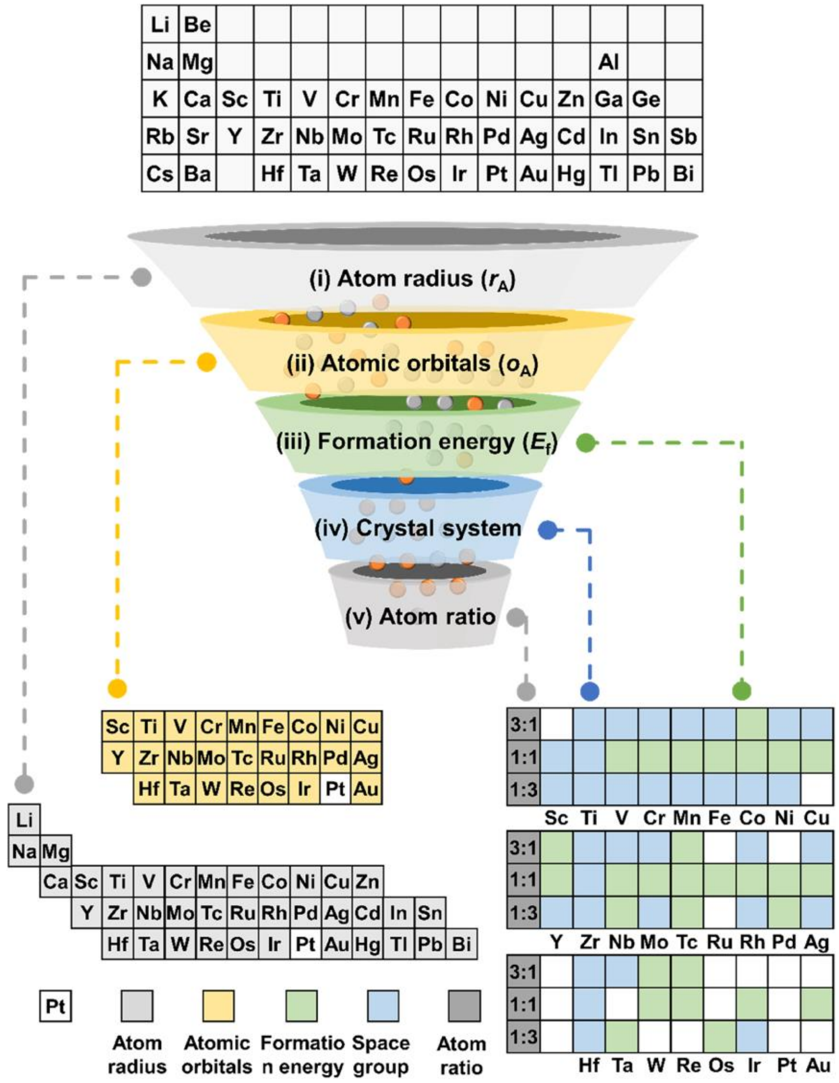
Li Be
Na Mg Al
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb
Cs Ba Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi
(i) Atom radius (r_A)
(ii) Atomic orbitals (o_A)
(iii) Formation energy (E_f)
(iv) Crystal system
(v) Atom ratio
Sc Ti V Cr Mn Fe Co Ni Cu
Y Zr Nb Mo Tc Ru Rh Pd Ag
Hf Ta W Re Os Ir Pt Au
Li
Na Mg
Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn
Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn
Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi
3:1
1:1
1:3
Sc Ti V Cr Mn Fe Co Ni Cu
Y Zr Nb Mo Tc Ru Rh Pd Ag
Hf Ta W Re Os Ir Pt Au
Pt
Atom radius
Atomic orbitals
Formatio n energy
Space group
Atom ratio

Figure 4.2 The approach for materials prescreening. From a large number of possible compounds with the criteria of (i) the radius and (ii) the orbital configuration of the transition metal atom, (iii) formation energy, (iv) crystal system, and (v) the atom ratio between the transition metal and the Pt atom, the materials considered to be calculated by density functional theory (DFT) are generated. Using machine learning techniques, we classify materials based on the decision tree.

The trained models are designed to make predictions and classify unknown materials into two categories. The training dataset consists of 555 potential material candidates with various structural configurations, divided into training and test data in a 9:1 ratio. Due to the small quantity of training data and to prevent distribution variation during validation, a 10-fold cross-validation scheme is chosen for hyperparameter tuning. In the cross-validation stage, the training data are evenly split into 10 groups, with each group used as the validation data to assess and evaluate the model trained on the remaining 9 groups of data. The performance of each validation is recorded, and the mean value of these 10 performances is considered the final score of the trained model. Decision tree (DT) and random forest classifiers are selected to implement the classification application. Using the aforementioned prescreening criteria, 77 materials are identified as catalyst candidates for ORR. It should be noted that in the prescreening procedure, the reference values of the prescreening criteria are tunable parameters, which can be adjusted to achieve different sizes of screened datasets. Loosening these criteria may allow for the inclusion of more materials, such as PtBi, which has relatively low stability (Zhang et al., 2008). Considering the requirement on the formation energy, decreasing the $E_{\mathrm{formaiton}}$ criteria threshold from 0.01 to 0.001 eV would filter out only three more 2D materials. For an alloy, the more open the surfaces, the stronger the intermediates bind, and eventually the surface get blocked. Although the utilization efficiency of step sites is much higher compared to the conventional basal plane, the basal plane sites are dominant. Therefore, another implicit assumption is that these dominate the activity of polycrystalline Pt, and the step sites and basal sites are considered. All surfaces are shown in **Figure 4.3**.

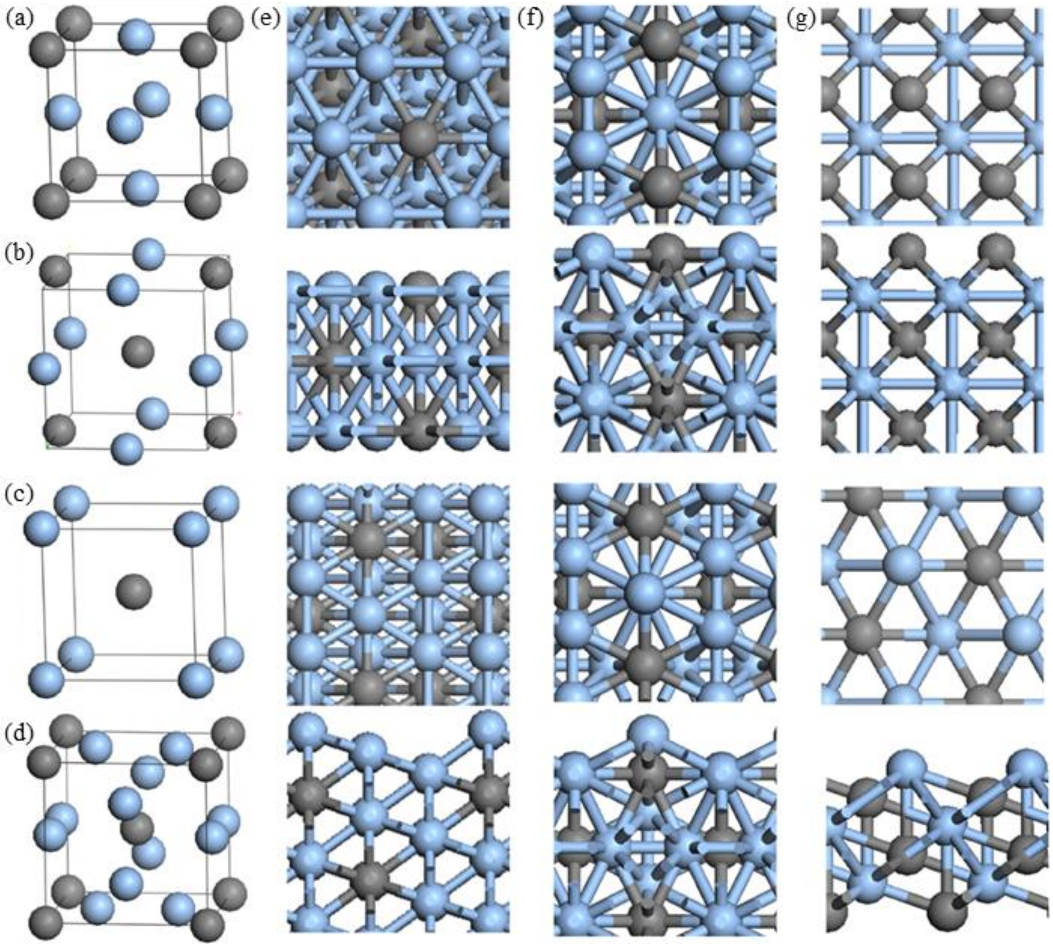

Figure 4.3 The unite cell of the platinum (Pt)-based alloy and surface structures used in this study. (a–d) The crystal structures of the Pt-based alloy prescreened from the database. (e) The (111), (211), (100) surfaces of crystal structure (a). (f) The (111), (211), (100) surfaces of crystal structure (b). (g) The (111), (211), (100) surfaces of crystal structure (c).

In **Figure 4.3**, the (111), (211), and (100) crystal planes were selected for their stability and well-known catalytic properties in Pt-based alloys. These low-index surfaces were chosen to simplify the computational model and align with common experimental results. While higher-index planes like (110) could also be effective catalytically, they were not included due to computational limitations. In data-driven high-throughput screening, the surface planes are typically predefined and cannot be dynamically adjusted to account for varying alloy compositions. This constraint standardizes the approach but also limits the exploration of all possible catalytic surfaces. Experimental work would be necessary to validate these predictions and investigate the role of higher-index planes.

4.3.2 First-Principal Calculations and Feature Engineering

The overall ORR can be described as the following equation:

$$\boldsymbol{O_2(g) + 4e^- + 4H^+ \rightarrow 2H_2O(l)} \tag{4-1}$$

The four-electron reaction pathway (2–5) of O_2 reduction in acidic media is considered (pH = 0) (Nørskov et al., 2004):

$$\boldsymbol{O_2(g) + e^- + H^+ \rightarrow OOH^*} \tag{4-2}$$

$$\boldsymbol{OOH^* + e^- + H^+ \rightarrow O^* + H_2O(l)} \tag{4-3}$$

$$\boldsymbol{O^* + e^- + H^+ \rightarrow OH^*} \tag{4-4}$$

$$\boldsymbol{OH^* + e^- + H^+ \rightarrow H_2O(l) + {}^*} \tag{4-5}$$

where the asterisk denotes an adsorbed site on the surface, (*g*) and (*l*) refer to the gas and liquid phases, respectively. The ORR mechanism on more than 62 Pt-based alloys is calculated; this large dataset comprised the adsorption energy of the intermediates and the free energy of the element steps at the surfaces of the Pt alloyed with 62 transition metal. The intermediates are placed at different sites and the energy for the most favorable site is included in the dataset. The changes of the free energy calculated by DFT during the ORR show that two reaction steps are sluggish that involve a positive change in free energy: the third electron and proton transfer for forming the adsorbed OH (ΔG_3) and the last transfer for removing OH from the surface to form water (ΔG_4). To screen the Pt-based alloy more easily, we use the values of overpotential (η) transferred from the ΔG for these steps and equilibrium potential as a measure of the activity. The overpotential for ORR is calculated by the equation:

$$\boldsymbol{\eta = \frac{\max\{\Delta G_1,\Delta G_2,\Delta G_3,\Delta G_4\}}{e} + 1.23} \qquad (4\text{-}6)$$

where ΔG_1, ΔG_2, ΔG_3, and ΔG_4 denote the reaction free energies in (4-2), (4-3) (4-4), and (4-5), respectively. According to thermodynamics, the smaller the overpotential, the higher the corresponding activity of ORR is; therefore, the performance of candidates is better for catalyzing ORR. The implicit assumption in this analysis is that the kinetic relationship is closely related to thermodynamics and can be simplified by thermodynamics. Because there will be an activation free energy in four elemental steps at the equilibrium potential of 1.23 eV, which is at least equal to the largest of the reaction free energies and the corresponding step is therefore likely the rate-limiting step (Chen et al., 2020a, Nørskov et al., 2004, Parada et al., 2019, Mayer, 2011). Forming a Pt alloy is one way to modify the electronic structure of the Pt surface to tune the stability of these critical intermediates. The stabilities of OH intermediates ($E_{\mathrm{ads(OH)}}$), in turn, scale roughly with the stability of adsorbed O ($E_{\mathrm{ads(O)}}$). Therefore, this parameter is particularly useful for characterizing both

ΔG_1 and ΔG_2. The ORR activity trends on different metal surfaces are summarized in **Figure 4.4**a. Plotting measured activities (overpotential of the rate-determining step) for a series of different catalysts as a function of the calculated OH adsorption energy results in a simple "volcano" relationship (**Figure 4.4**a). If ΔE_{OH} becomes increasingly positive, adsorbed H_2O is destabilized and can desorb from the surface more easily. However, if ΔE_{OH} keeps increasing in the positive range, it becomes easier to break the Pt-OH bonds, which makes the OH^* formation difficult. This appears to be a reasonable conjecture, given that more open surfaces tend to bind intermediates considerably stronger and become blocked.

As shown in **Figure 4.4**a, a surface that binds OH with the adsorption energy of 1.1 eV exhibits an optimal ORR activity. The decreasing of overpotential indicates the increasing ORR activity and the activity is closely related to the behavior of OH adsorption. The weaker OH adsorption on the surface results in the lower ORR activity. The rate-determining step of the left half branch is H_2O formation. In this branch, the ORR activity becomes better as the adsorption strength of OH decreases. When the OH adsorption is too strong, the H_2O formed after OH hydrogenation is difficult to desorb from the catalyst surface; therefore, weaker OH adsorption strength is beneficial for ORR. The rate-determining step of the right half branch is OH formation. In this branch, the ORR activity becomes worse as the OH adsorption strength decreases. When the OH adsorption is too weak, it is difficult to form OH from O hydrogenation. With the calculated overpotential for the catalyst candidates, it is desirable to leverage such data to examine whether the OH adsorption is simply correlated with a certain intrinsic property of a given material.

Such simple correlations are usually established on a series of adsorption systems bearing similar atomic structures, leading to the predominance of the electronic effect. Although all computational results are perfect on the monocrystalline surfaces in experiments, a mixture of single crystal, vacuum-annealed polycrystalline and Ar-sputtered polycrystalline surfaces is always used. The resulting structural differences introduce deviations from our single-crystal models, thereby indicating modest changes in the ORR activity. However, despite on the polycrystalline of the Pt-based alloy, the site corresponding to the most stable configuration can always be model on the single crystal surface and the differences do not substantially alter the trends described above (Stamenkovic et al., 2007, Tian et al., 2007). Although computationally based electrocatalyst discovery is the principal aim of this approach, more generally and, perhaps, more importantly, we probe our present understanding of the ORR. In the field of catalysis, there is no stronger evidence for accuracy of a theoretical framework than the ability to use that framework to identify new active materials.

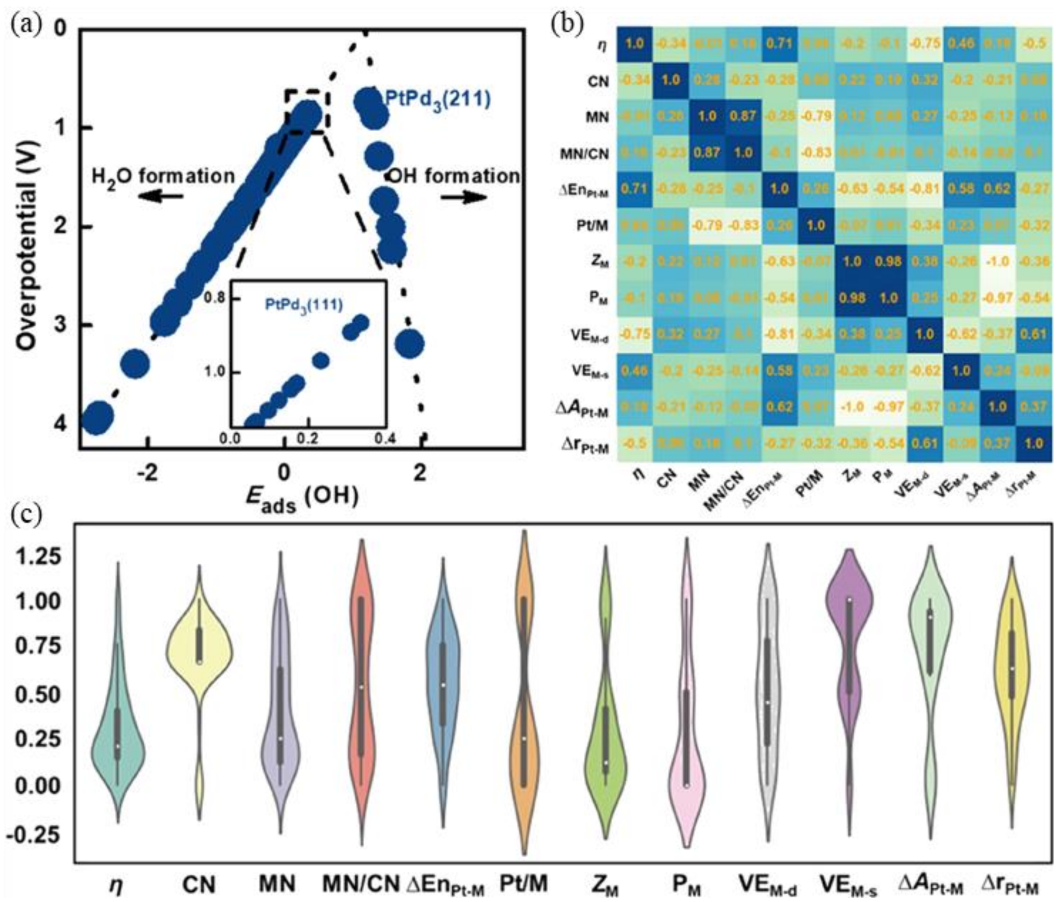

Figure 4.4 The activity of oxygen reduction reaction (ORR) and potential features. (a) The ORR activity trends on different metal surfaces, (b) the in-pair Pearson correlation coefficients of the selected potential features and the overpotential, and (c) a violin plot of the distribution of the variables after min–max normalization. Every colorful area is the density plot of corresponding variables. The black box in each density plot represents the range from 25% to 75% percentiles where the white point marks the mean value, and the whiskers denote 95% and 5%.

Additionally, a correlation study was conducted on all the factors affecting O_2 reduction activity on Pt-based alloys. These factors include the coordination number of Pt, the number of heteroatoms around Pt, the electronegativity difference between

Pt and heteroatoms, the atomic number and period of the heteroatoms, the ratio of Pt to heteroatoms, the number of valence electrons (d and s electrons) of the heteroatoms, and the differences in relative mass and atomic radius between Pt and the heteroatoms. All the abbreviations of the features are summarized in **Table 4.1**, which indicate various important or typical factors of a catalyst to influence the activity of ORR. It should be noted that the structure features are based on the final optimized structures by DFT calculation. The electronegativity difference and valence electron number showed the greatest correlation with the ORR activity (**Figure 4.4**b). For the 12 features, we showed a violin plot of the distribution of the variables after min–max normalization in **Figure 4.4**c and the in-detail calculation is provided in the Supporting Information. The violin plot synergistically combines the box plot and the density trace (or smoothed histogram) into a single display that reveals structure found within the data, which can provide us a quick overview of the combination of the box plot and density trace. According to the density plot, we observed that the material with higher CN (CN = 8 and 9) and lower MN (MN = 3 and 4) tend to be more promising as candidates. Furthermore, materials with 1:3 and 3:1 ratios of Pt atom over transition metal atom are dominant in the screening stage, which is in agreement with the distribution of the MN/CN value. Variables such as $\Delta En_{Pt\text{-}M}$, $VE_{M\text{-}d}$, and $\Delta r_{Pt\text{-}M}$ show a relatively even distribution. The $VE_{M\text{-}s}$ and relative atomic mass of most materials are 2 and approximately 140, respectively.

Table 4.1 A set of the 11 least-correlated primary features used for the descriptor construction

Features	Description
CN	Coordination number of Pt
MN	Number of heteroatom around Pt
MN/CN	Ratio of heteroatom around Pt
ΔEn_{Pt-M}	Difference of the electronegativity between Pt and heteroatom
Pt/M	Atomic ratio of Pt and heteroatom in the alloy
Z_M	Atomic number of the heteroatom
P_M	Period of a heteroatom in the periodic table
VE_{M-d}	Number of valence electrons in the d orbital
VE_{M-s}	Number of valence electrons in the s orbital
ΔA_{Pt-M}	Difference of relative atomic mass between Pt and heteroatom
Δr_{Pt-M}	Difference of atomic radius between Pt and heteroatom

4.3.3 ML Discovery of Descriptors And Establishment of Structure–Activity Relationship

To obtain the best low-dimensional descriptors using SISSO, the descriptors with lower Root Mean Square Error (RMSE) are chosen. Here, the descriptor can be one feature or the combination of various features, which is used in a model of the relationship between the activity and the feature(s) of a catalyst. As shown in **Figure 4.5**a, the three-dimensional (3D) models were found to be more accurate. Therefore, we chose 3D descriptors. From the 3D model's descriptors, we obtained the following relationships as the formula:

For the rate-determining step of H_2O formation,

$$\mathbf{0.237} \times \frac{\mathbf{\Delta En^2_{Pt-M}} \times (\mathbf{VE_{M-d}} + \mathbf{VE_{M-s}})}{\mathbf{MN/CN}} + \mathbf{0.645} \qquad (4\text{-}7)$$

For the rate-determining step of OH formation,

$$\mathbf{8.995} \times \frac{(\mathbf{MN/CN}) \times \Delta\mathbf{En}_{\mathbf{Pt-M}}}{\mathbf{VE}_{\mathbf{M-d}} + \mathbf{VE}_{\mathbf{M-s}}} - \mathbf{0.0003724} \times \boldsymbol{e}^{\boldsymbol{P}_{\mathbf{M}}} \times (\mathbf{MN/CN}) \times \Delta\boldsymbol{r}_{\mathbf{Pt-M}} +$$

$\mathbf{1.1765}$In this work, the hold-out validation method was used to verify the model due to the relatively small and imbalanced training dataset for the SISSO model. The use of cross-validation might introduce extra bias if the random 10-fold splitting is uneven, despite efforts to ensure the data of every type of material is included in the training set. The trained SISSO model achieved an RMSE on the validation set far below 0.001 eV (Sun et al., 2020a). The activity predicted by this model was found to be consistent with the model calculated by DFT, regardless of whether H_2O formation or OH formation was the rate-determining step (**Figure 4.5**b, c).

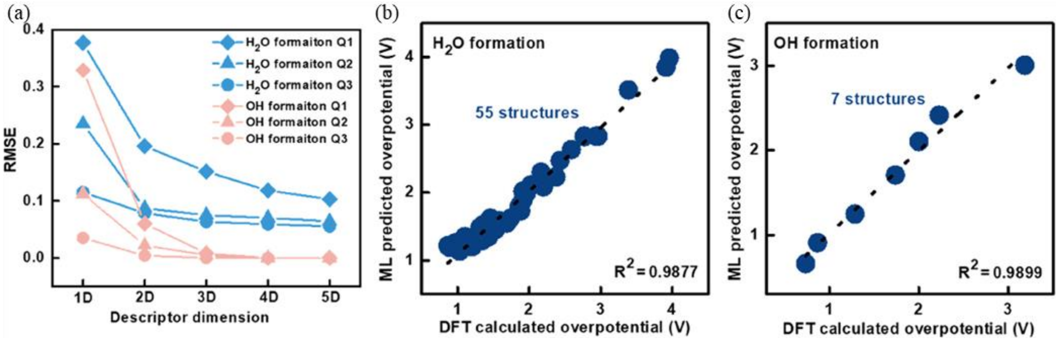

Figure 4.5 Validation of the model by machine learning. (a) Training and validation RMSE from descriptor dimension for all tested Sure Independence Screening and Sparsifying Operator (SISSO) parameters, (b) the calculated activity based on our density functional theory (DFT) model as well as a dashed line indicating predicted activity for the left branch with the rate-determining step of H2O formation, and (c) the calculated activity based on our DFT model as well as a dashed line indicating predicted activity for the left branch with the rate-determining step of OH formation.

Based on the above model, 11 structures were added as the second active learning. The new data followed the general trends well (**Figure 4.6**a), indicating that the model exhibited true predictive ability in describing the trends of ORR activity on Pt-based alloys. Furthermore, Pt_3Co (211) showed more optimized performance with the OH formation as the rate-determining step. Then, the third active learning was carried out, and the four structures were calculated. It was found that Pt_3Ni(111) was upshifted to the top along the left branch with H_2O formation as the rate-determining step. As the volcano plot shows, the Pt_3Co(211) showed the most optimized performance among the 77 structures. In **Figure 4.6**b, more detailed calculations on the Pt_3Co(211) surface were included. The free energy changes of the elemental steps were found to be negative at 0 V versus Standard Hydrogen Electrode (SHE). At the equilibrium potential of 1.23 V, the OOH formation, OH formation and H_2O formation were endothermic and the OH formation exhibited the highest free energy change. Therefore, the energy level diagram of ORR on the Pt_3Co(211) surface is the rate-determining step. The activity predicted by this model is still consistent with the model calculated by DFT, irrespective of whether $_{H2O}$ formation or OH formation was the rate-determining step (**Figure 4.6**c, d). Here, $|\eta| \leqslant 1$ eV was used as the final screening criterion and identified five potential catalyst candidates: Pt_3Co(211) (0.38 V), $PtPd_3$(211) (0.74 V), Pt_3Ni(111) (0.85 V), $PtPd_3$(111) (0.86 eV), and PtAu(111) (0.870 eV), on which OH adsorption is nearly thermoneutral for ORR at low overpotential. It is worth noting that Pt_3Co has been experimentally proven with an efficient ORR activity.

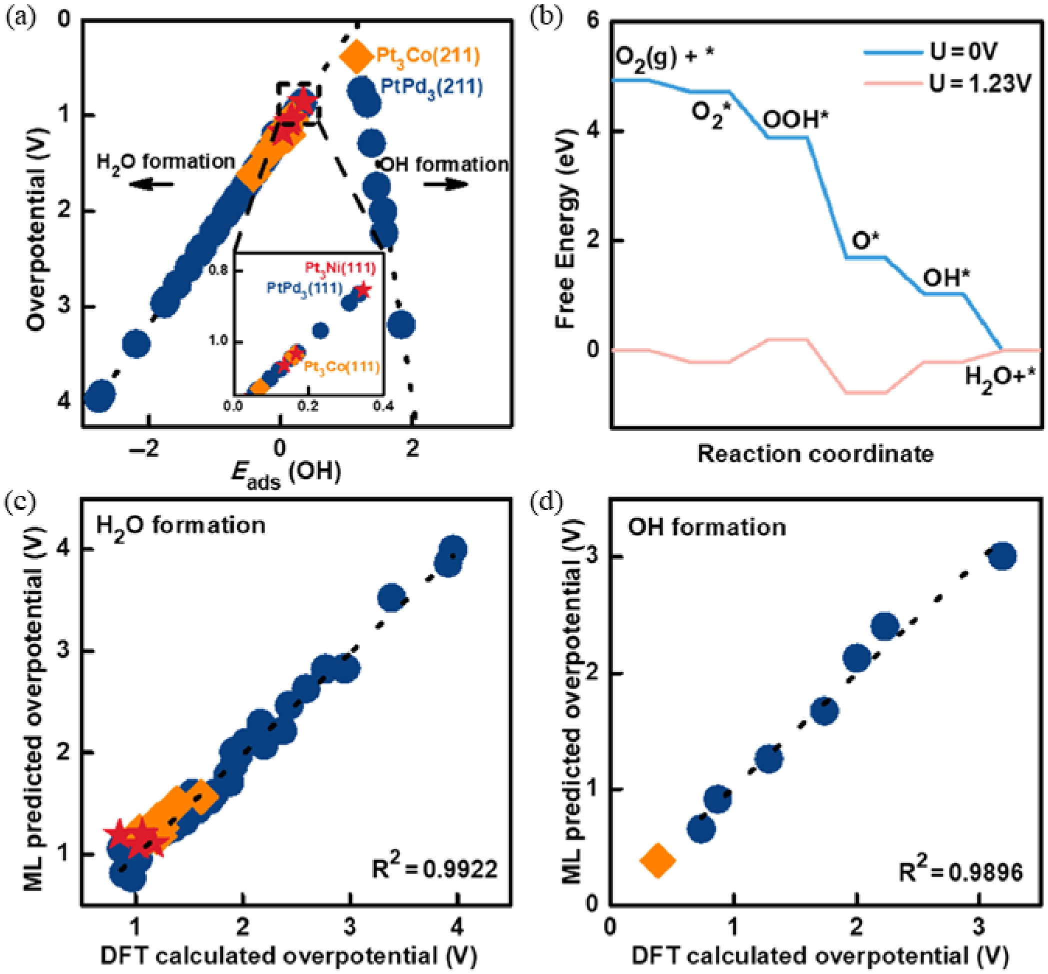

Figure 4.6 The active learning for platinum (Pt)-based alloy. (a) Active learning for another 11 (second round, orange squares) and 4 structures (third round, red stars); (b) the energy diagram of oxygen reduction reaction (ORR) on Pt3Co (211) at 0 and 1.23 V versus SHE; (c) the calculated activity based on our DFT model as well as a dashed line indicating predicted activity for the left branch with the rate-determining step of H$_2$O formation; and (d) The calculated activity based on our DFT model as well as a dashed line indicating the predicted activity for the left branch with the rate-determining step of OH formation.

It was observed that based on the model, the difference between the electronegativity of Pt and heteroatom, the valence electrons number of the heteroatom, and the ratio of heteroatoms around Pt have the most obvious effects on the ORR performance. This is because the difference in the electronegativity between Pt atoms and heteroatoms and the number of valence electrons of heteroatoms can distinguish the types of alloys, the coordination number of Pt atoms can reflect the different surfaces, and the ratio of heteroatoms around Pt atoms can reflect the doping ratio of the heteroatoms. The three factors in the model are scaled as x, y, and z, respectively. The 3D plot and cross-sectional view in the middle of each axis are shown in **Figure 4.7**, where the blank part is ascribed to the negative overpotential. As for the strong OH adsorption, when the electronegativity of Pt is lower than that of the heteroatom, the more the number of valence electrons of heteroatom is, and the lower the ratio of the number of heteroatoms around the Pt is, the higher the activity is. As for the weak OH adsorption, its activity is mainly determined by the electronegativity difference between Pt and heteroatoms, which is within 0.5; therefore, the activity is higher. It is worth noting that the activity can be predicted from the structural information for a given structure, which is very convenient and direct for new materials prediction without the need for tedious electronic structure calculations.

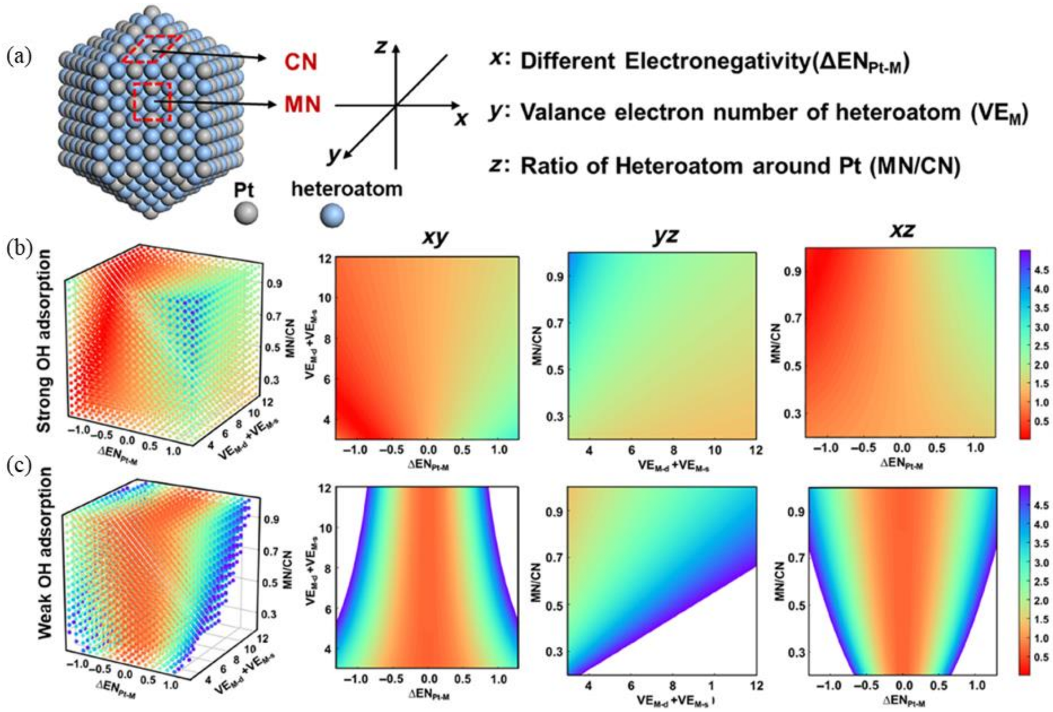

Figure 4.7 The representation of descriptors and the relationship of the structure–activity. (a) Schematic of descriptors on Pt-based alloys. The three-dimensional plot of oxygen reduction reaction (ORR) activity with the three descriptors scaled in x, y, z and the cross-sectional view at the middle of each axis, where (b) the rate-determining step is H_2O formation and (c) the rate-determining step is OH formation.

4.4 Conclusion

In summary, high-throughput first-principles calculations were conducted to screen high-performance catalysts for ORR, and the structure–activity relationship was determined. Based on this relationship, additional excellent Pt-based alloys were identified through the second and third rounds of active learning. Among the 77 prescreened candidates, five candidates demonstrated the thermodynamic capability for ORR with the lowest overpotential, indicating their potential for catalyzing ORR. Furthermore, the difference in electronegativity between Pt and heteroatoms, the number of valence electrons of the heteroatoms, and the ratio of heteroatoms around Pt were found to have the most significant effect on ORR performance according to the model. The results of this study are expected to provide a useful dataset for experimentalists to further examine the predicted ORR activity and for data scientists to develop ML models for ORR performance predictions. Additionally, this study may aid in the exploration of catalysts for other electrocatalytic processes, such as water electrolysis.

Chapter 5

Optimizing Synthesis of High-Performance Lithium Iron Phosphate Using a Data-Driven Active Learning Framework

Part of the content in chapter is ready to be submitted:

WANG, Z., HU Y., LIU Z., FOW, K. L., WU, T. and PANG, C. H. "Optimizing Synthesis of High-Performance Lithium Iron Phosphate Using a Data-Driven Active Learning Framework." *To be submitted.*

5.1 Synopsis

Lithium iron phosphate (LFP) has attracted significant interest due to its abundant raw materials, non-toxicity, environmental friendliness, and high theoretical capacity. However, its intrinsic low electrical conductivity and slow ion diffusion rate adversely impact its rate performance and low-temperature capabilities, limiting its broader application. This study proposes a novel data-driven active learning framework to optimize the synthesis of high-performance LFP materials. The framework integrates two ensembled ML models to iteratively refine synthesis parameters, aiming to enhance the physicochemical properties and electrochemical performance of the resulting LFP samples. The active learning loop not only guides the synthesis process but also significantly reduces the number of experiments required to identify optimal formulations. The approach begins with the creation of a dataset comprising various synthesis parameters and their corresponding material properties. Initial synthesis is performed, and data from these experiments are used to train the ML models. These models then predict the optimal synthesis conditions, which are tested experimentally. The results of these tests are fed back into the models, continuously improving their predictive accuracy. Electrochemical characterization of the synthesized samples, including those recommended by the ML models, reveals that the addition of alkyl polyglucosides (APG) significantly reduces polarization and enhances cycling performance. The second-round recommended sample (RR2) demonstrates the highest discharge capacities, best cycling stability, and highest compaction density, validating the efficacy of the active learning framework. This study underscores the potential of integrating machine learning into the material synthesis process, leading to the discovery of two high-performance LFP materials. The findings highlight the effectiveness of the active learning loop in identifying and optimizing critical synthesis parameters, paving the way for the development of advanced battery materials with enhanced properties.

5.2 Introduction

Lithium iron phosphate (LFP) has garnered significant attention due to its abundant raw materials, non-toxicity, environmental friendliness, and high theoretical capacity (170 mAh/g). However, its intrinsic low electrical conductivity and slow ion diffusion rate significantly limit its performance, particularly in high-rate and low-temperature conditions. To address these challenges, several modification strategies have been developed, including element doping (Zhang et al., 2020c, Wang et al., 2021a, Li et al., 2009), surface carbon coating (Hsieh et al., 2012, Liu et al., 2022b, Xi and Lu, 2020), particle nanization (Chen et al., 2013, Huang et al., 2018), and material compositing (Zhang et al., 2014, Medvedeva et al., 2019).

Among these, surface coating has proven particularly effective in enhancing ionic conductivity and reducing electrode polarization. Common coatings, such as graphite, graphene, metal powders, and conductive polymers, can prevent agglomeration, inhibit particle growth, and improve overall electrochemical performance (Prosini et al., 2001, Wang et al., 2010, Croce et al., 2002, Liu et al., 2008, Wang et al., 2005). Xie et al. (Xie and Zhou, 2006) explored Al doping in

$LiFePO_4/C$ cathodes, finding that Al^{3+} reduced particle size and shortened lithium-ion transport pathways, resulting in improved discharge capacity. Similarly, Ti doping has shown to enhance performance, as demonstrated by Li et al. (Li et al., 2009) and Wang et al. (Wang et al., 2012b).

Carbon coatings, particularly from organic sources like glucose and inorganic sources like carbon black, have been found to significantly improve the conductivity of $LiFePO_4$ (Armand et al., 2009), with composite carbon sources offering enhanced conductivity and particle size control (Liu et al., 2012). Advanced coatings like multi-walled carbon nanotubes (MWCNTs) (Qin et al., 2014). and graphene (Xu et al., 2015, Geng and Ohno, 2013, Song et al., 2016) further enhance performance by increasing electronic conductivity and structural flexibility. Graphene, in particular, improves the electrochemical performance due to its high conductivity and stability when bonded with the LFP surface (Fei et al., 2014). Wang et al. (Wang et al., 2016) demonstrated that combining Mn-doped LFP with graphene enhanced the material's electrochemical performance, as the graphene increased electronic conductivity and facilitated Li+ migration at the interface.

Despite the progress made using traditional methods, there is limited research on integrating machine learning (ML) to improve LFP performance. This study introduces a novel data-driven active learning framework to optimize the synthesis of high-performance LFP materials. The framework combines two ensembled ML models to iteratively refine synthesis parameters, enhancing both the physicochemical and electrochemical properties of the resulting LFP samples. By using this approach, we aim to reduce the number of experimental trials needed, speeding up the discovery of optimal synthesis conditions and facilitating the development of advanced LFP materials.

This active learning methodology represents a significant step forward in material synthesis, offering a more efficient, data-driven approach to optimizing LFP performance and accelerating innovation in battery technology.

5.3 Results and Discussion

5.3.1 Data PrePartion and Model Training

Although machine learning techniques have rapidly advanced in materials science and chemical engineering recently, fundamental challenges must be addressed before selecting specific algorithms. These include determining the workflow, establishing the datasets to learn from, choosing the vitally concerned properties as the target, preliminarily identifying the highly related features, and defining the type of tasks (Yin et al., 2021). In this chapter, a data-driven enabled LFP synthesis strategy is proposed, utilizing two ensembled ML models working in series within an active learning loop. The first step is to create a dataset that allows ML models to learn from. More specifically, in the first 4 months, 80 LFP samples are synthesized, with reaction conditions and mass of added chemicals varying. Then, a screening process is conducted on the raw dataset to exclude some samples: those with significant gaps in the target property values, those using different devices such as tube furnaces, and those with abnormal data. More specifically, in the current production of lithium iron phosphate (LFP) cathode materials, the two most critical performance metrics are the initial discharge capacity at a 1C rate (C_{1C}) and the compacted density under a pressure of 30,000 N (ρ_{30kN}). In addition to these, other key performance indicators include cycling performance, which measures the material's ability to maintain capacity over multiple charge-discharge cycles, and rate capability, which evaluates the performance at various discharge rates. These metrics significantly influence the energy density, long-term stability, and overall performance of batteries manufactured with this material. The screened sample dataset underwent further feature engineering, including feature selection and representation (Wang et al., 2022). Fixed or slightly varied parameters such as ball milling time and speed, as well as high-temperature sintering reaction time, were removed. The sample dataset with selected features are then sent to the first ML model to classify whether they have high, medium, or low ρ_{30kN}. Following this classification, a regression model is trained to accurately establish the relationship between the synthesis parameters and C_{1C}. By utilizing the established parameter-performance relationship, the screening and recommendation of potential synthesis recipes can be conducted to identify the optimal reaction conditions and chemical quantities for synthesizing LFP with desired properties. Furthermore, by employing active learning, the results of the recommended recipes can be added back to the dataset to iteratively augment it and dynamically update the model, enhancing its accuracy and providing better suggestions. The workflow of the data-driven assisted LFP synthesis is shown in **Figure 5.1**. The overall goal is to identify the recipe that maximizes both the C_{1C} and ρ_{30kN} of LFP.

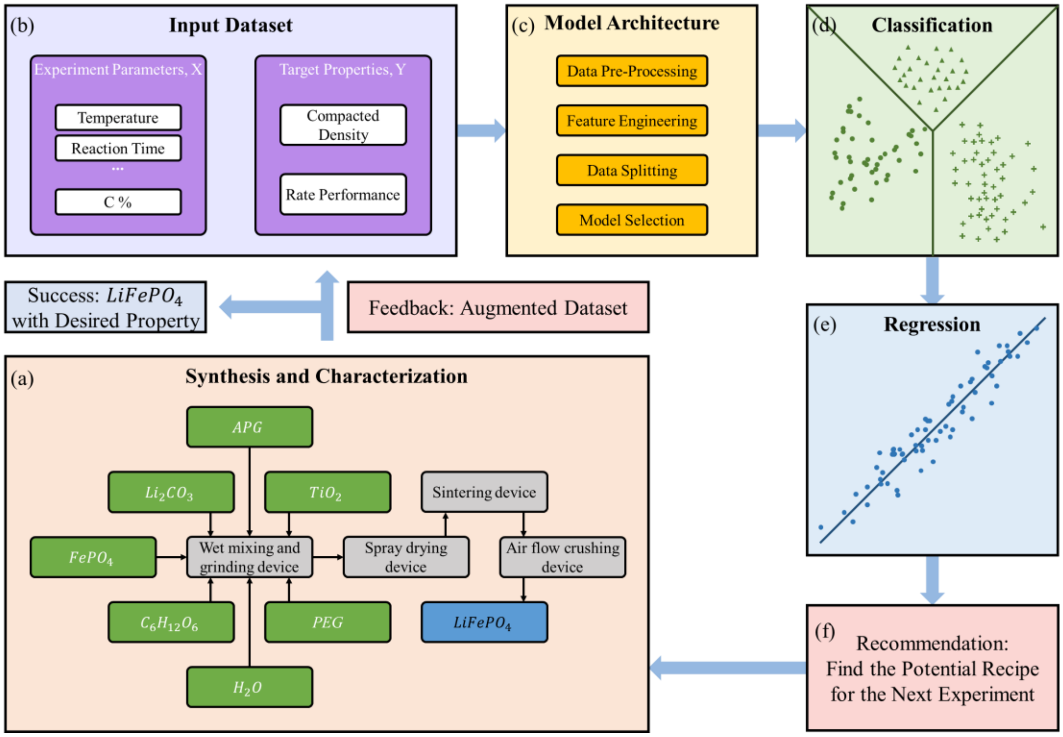

Figure 5.1 The workflow for utilizing ML models to suggest potential synthesis recipes via an active learning loop involves five main stages: (a) material synthesis and characterization, (b) data collection, augmentation, and dataset construction, (c) data preprocessing, feature engineering, and model selection, (d) classification for different categories of LFPs based on ρ_{30kN}, (e) regression for the prediction of C_{1C}. (f) Utilizing the well-trained model to make recommendations for the next experiment. Based on these stages, the synthesis of LFP is guided, their properties are characterized and added to the original dataset, and the model is further trained to update hyperparameters, thereby improving model performance.

The synthesis recipe primarily focuses on the properties of raw chemicals, including the Fe/P ratio and BET surface area of the FPs, the mass or volume of additives, and the set reaction conditions. During the synthesis of the 80 samples, all these synthesis parameters are recorded and stored in the primary dataset as potential features. Although there are numerous parameters that can be adjusted and influence the properties of the samples, only a subset of these parameters was selected as features. These were chosen because they can be easily adjusted and have a direct influence on the properties of LFPs based on preliminary experiments. These critical parameters were shown in **Table 5.1** and selected as features and used for subsequent ML training.

Table 5.1 The selected features for ML learning

Features	Description
T_s	The highest temperature reached during the sintering process.
m_{LC}	The mass of LC added
m_{CHO}	The mass of Glucose added
m_{PEG}	The mass of Polyethylene Glycol added
m_{TiO2}	The mass of TiO_2 added
m_{APG}	The mass of Alkyl Polyglucosides added
m_{H2O}	The mass of Deionized Water added
BET_{FePO4}	BET Specific Surface Area of $FePO_4$
$D10_S$	The particle diameter at which 10% of the slurry's particles are smaller.
$D50_S$	The median particle diameter, where 50% of the particles are smaller and 50% are larger.
$D90_S$	The particle diameter at which 90% of the slurry's particles are smaller.
$Dmax_S$	The maximum particle diameter observed in the sample.

The characterization of the LFPs includes recording a series of properties such as physicochemical properties, electrochemical properties, and morphological properties. In this study, ρ_{30kN} and C_{1C} are selected as the target properties, and the samples are labeled accordingly based on task type for further training. Specifically, the first step of the workflow involves reviewing the dataset and excluding outlier samples. As each ML model has its own domain of applicability, sample points that are significantly far from the desired target property values are excluded for better training performance. For this research, the ideal LFPs typically have a $\rho_{30\mathrm{kN}}$ higher than 2.55 g/cm³ and a $C_{1\mathrm{C}}$ higher than 135 mAh/g. Therefore, samples with $\rho_{30\mathrm{kN}}$ values lower than 2.3 g/cm³ and $C_{1\mathrm{C}}$ values lower than 135 mAh/g are excluded from the dataset. Additionally, samples that were sintered in different tube furnaces were removed to eliminate system errors. **Figure 5.2** provides a visual summary of the distribution of the scaled features in the dataset. The line in the middle of each box represents the median value of the corresponding feature. It should be noted that the added mass of glucose remains almost constant, influencing the carbon residue of the products and resulting in only limited variation in this feature.

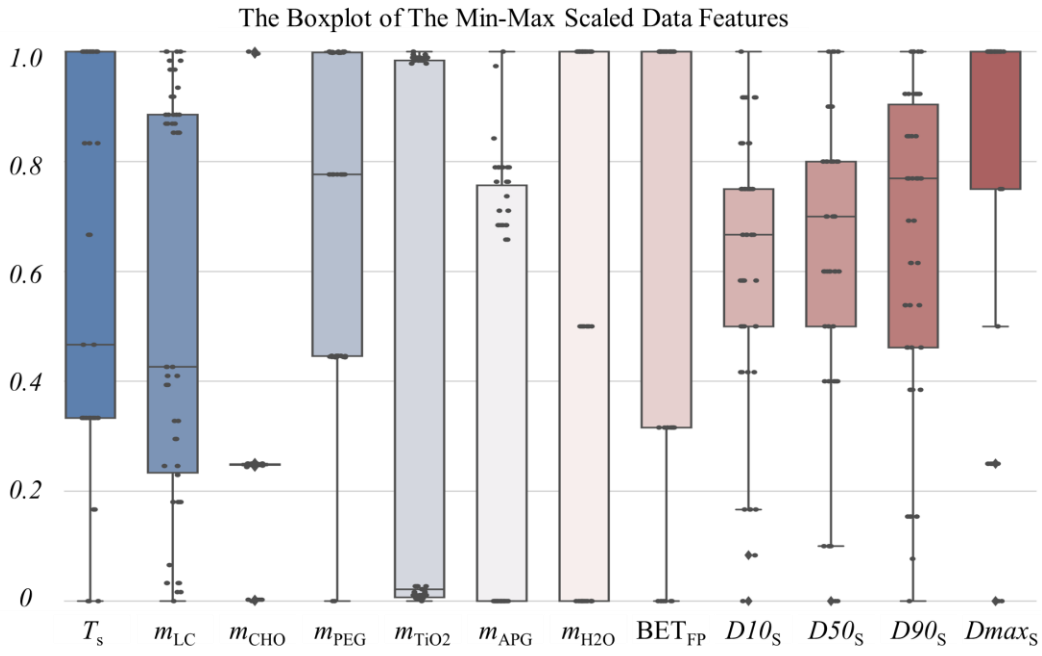

Figure 5.2 The box plot of the distribution of the features after min–max normalization. Every colorful area is the density plot of corresponding variables. The black box in each density plot represents the range from 25% to 75% percentiles where the line in the middle of the box represents the median value of the feature, and the whiskers denote 95% and 5%.

5.3.2 ML Training Strategy and Performance

The remaining LFP samples are labeled into three categories based on ρ_{30kN} values: high, medium, and low compacted density. LFPs with a ρ_{30kN} greater than 2.5 g/cm³ are labeled as "High" compacted density. Those with a ρ_{30kN} less than 2.4 g/cm³ are labeled as "Low" compacted density. The LFPs with ρ_{30kN} values between 2.4 and 2.5 g/cm³ are classified as having "Medium" compacted density. At this stage, the raw dataset consists of 60 samples, which are divided into a training set and a test set with a ratio of 9:1.

A classification model leveraging a sophisticated architecture that combines multiple ensemble methods through a stacking classifier is designed to handle the small-sized data. This design aims to enhance predictive performance by utilizing the strengths of various algorithms while mitigating their individual weaknesses. More specifically, the architecture comprises three primary components: base models, a meta-model, and a stacking classifier. The base models include a "Random Forest Classifier (RFC)", a "Gradient Boosting Classifier (GBC)", and an "Ada Boost Classifier (ABC)". The RFC constructs multiple decision trees during training and aggregates their outputs to determine the final classification, thus reducing overfitting and improving generalization. The GBC builds models sequentially, each correcting errors made by the previous ones, effectively reducing bias and variance. The ABC adjusts instance weights based on classification accuracy, focusing on difficult-to-classify instances in subsequent iterations. At the heart of the stacking classifier is the meta-model, a Logistic Regression model configured for multinomial classification with the 'lbfgs' solver. This logistic regression model combines the predictions of the base models to make the final decision. The overall stacking classifier integrates these base models and the meta-model to enhance the overall predictive performance. The ensemble model is encapsulated within a Pipeline, which ensures consistent preprocessing and model training. The pipeline comprises a preprocessing step, denoted as preprocessor, which likely includes transformations such as scaling, encoding, and imputation, followed by the stacking classifier.

The receiver operating characteristic (ROC) curve for the $\rho_{30\mathrm{kN}}$ classification model is shown in **Figure 5.3**a, presenting the model's true positive rate (TPR) against the false positive rate (FPR) with varying thresholds. Since the model addresses a multi-class classification problem, three ROC curves, calculated using the one-vs-all method, are generated. The curves for the low compacted density (red), medium compacted density (yellow), and high compacted density (green) categories achieve AUC scores of 1.00, 0.80, and 0.83 on the test set, respectively, indicating a significant degree of classification ability. As expected, the classification ability for identifying the medium compacted density is lower due to the nature of particle growth. Specifically, LFP particles tend to grow into larger particles once the energy absorbed or temperature is sufficiently high, creating significant boundaries between the low and medium compacted density classes. However, there are no clear features or phenomena to distinguish between the medium and high compacted density classes, resulting in overlapping boundaries between these categories. Upon completing the classification training, the hyperparameters of the model are determined, allowing for the evaluation of each feature's contribution. **Figure 5.3**b illustrates the importance of the 12 selected features. As anticipated, Ts plays the most critical role in influencing the compacted density. Additionally, the mass of the added carbon source such as Glucose, PEG and APG, and m_{LC} potentially affects the maximum particle size of LFPs at a fixed temperature. Therefore, the synergy between sintering temperatures and the mass of the added lithium source and carbon source should be carefully considered in future experiments.

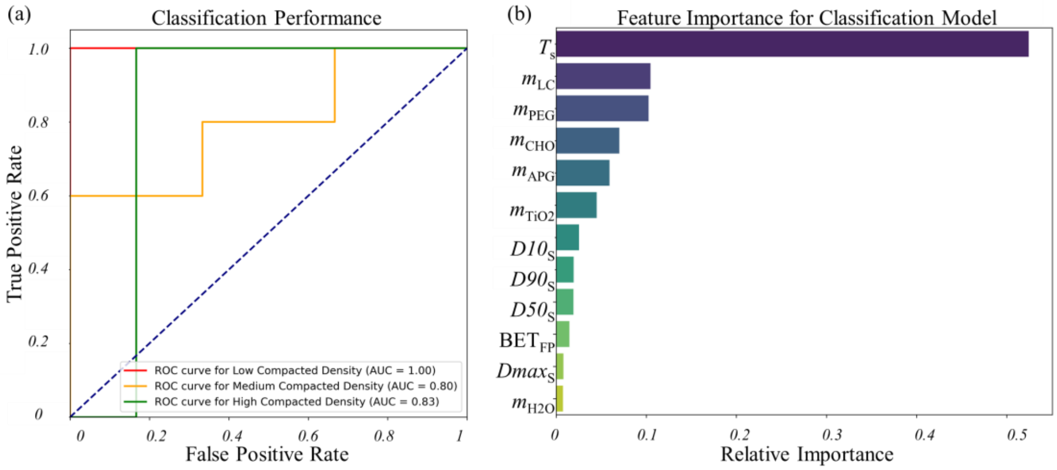

Figure 5.3 (a) The true positive rate (TPR) against the false positive rate (FPR) of the classfication model (b) The feature importance of the classification model.

Once the compacted density of an LFP sample is classified, its C_{1C} value is another critical property that influence the overall energy density. In the following regression task, the C_{1C} of each sample in the training set was designated as the target variable. This step involved reorganizing the training dataset to focus specifically on the C_{1C} values, thereby creating a new training set tailored for this purpose. The newly prepared dataset was then employed to train a regression model, aiming to accurately predict the C_{1C} values based on the given features. Specifically, a stacking regressor was employed to enhance predictive performance through the combination of multiple base models and a meta-model. The base models utilized included a Random Forest Regressor (RFR) and a Gradient Boosting Regressor (GBR). The RFR, an ensemble method, aggregates multiple decision trees (DTs) trained on different subsets of the data to improve predictive accuracy and mitigate overfitting. Similarly, the GBR sequentially builds models where each model corrects the errors of its predecessor, thereby enhancing overall prediction accuracy. The meta-model chosen was again an RFR, which integrates the predictions from the base models to generate the final output. A comprehensive hyperparameter grid was defined to optimize the performance of both the base models and the meta-model. This grid included parameters such as the number of estimators and maximum depth for the Random Forest, and the learning rate and maximum depth for the GBR To streamline the process, a pipeline was constructed that encompassed both data preprocessing steps and the stacking regressor, ensuring a systematic and efficient workflow. ‘GridSearchCV’ was utilized to conduct an exhaustive search over the specified hyperparameter grid, employing CV to evaluate model performance across different subsets of the training data.

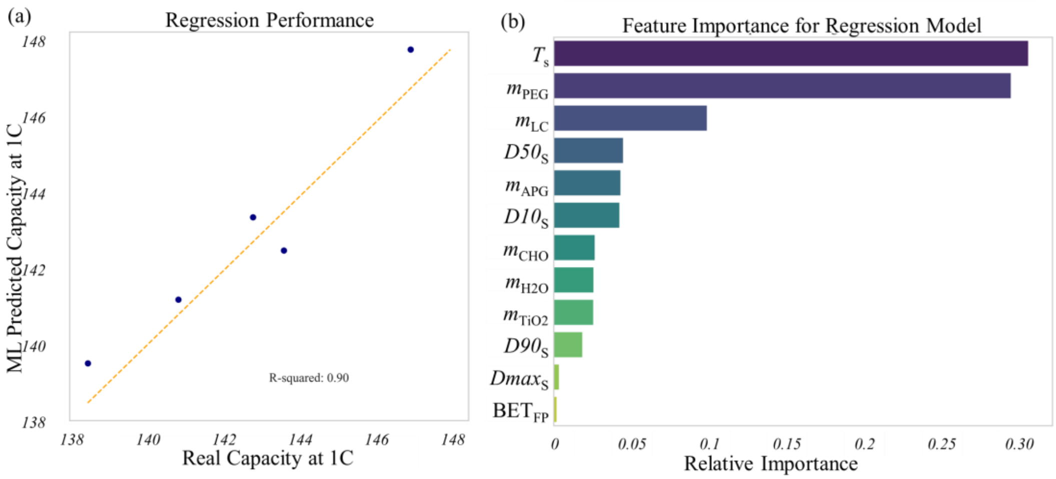

Figure 5.4 (a) The prediction performance of C_{1C} on test set. (b) The feature importance of the regression model.

For the regression model, the training set and test set were randomly generated with a ratio of 9:1. The model achieved an R-squared value of 0.90 and an RMSE of 0.79 mAh/g, indicating high predictive accuracy and low prediction error, respectively. The representative correlation plot of the predicted C_{1C} values is illustrated in **Figure 5.4**a, demonstrating the model's good prediction ability on the test set. Furthermore, the feature importance of the regression model in predicting C_{1C}, as shown in **Figure 5.4**b, highlights that T_s plays a dominant role in influencing the C_{1C} of LFPs. This is attributed to the tendency of high T_s to produce larger LFP particles, resulting in higher compacted density but a reduction in C_{1C}. Additionally, the impact of m_{PEG} on C_{1C} was relatively significant, which could be due to the improved suspension system facilitated by PEG during the wet ball milling and spray drying processes, leading to a more uniform precursor.

The two trained machine learning models constitute the core component of the recommendation stage within the active learning loop. In this process, a series of synthesis recipes, informed by feature importance, are input into the workflow. The recipe yielding the best predicted results is selected for the subsequent experiment. The first round of recommendations resulted in the synthesis of LFPs denoted as RR1, with the corresponding data detailed in the subsequent sections. The recipe and characterization data from RR1 are then incorporated into the dataset to augment it and update the hyperparameters of the models, facilitating the second round of recommendations. This iterative process yields the sample designated as RR2. It is important to note that both RR1 and RR2 were synthesized with the addition of APG. Corresponding control groups, RR1B and RR2B, were synthesized without the addition of APG, allowing for a comparative analysis of the impact of APG on the synthesis outcomes.

5.3.3 Morphology Characterization

The LFPs were synthesized according to the recipes recommended by the first and second rounds of the active learning loop (RR1 and RR2), while the corresponding blank samples without APG added were also synthesized as a control group (RR1B and RR2B). In the synthesis process, the RR1 and RR1B groups were sintered at 790°C, while the RR2 and RR2B groups were synthesized at a sintering temperature of 800°C. The primary difference between RR1 and RR1B, as well as between RR2 and RR2B, is the inclusion of APG in the RR1B and RR2B groups, which allows for a direct comparison of the influence of APG addition on the synthesis process and the resulting electrochemical performance. The crystal structures of the four materials were characterized using X-ray diffraction (XRD). The XRD patterns of the four samples are shown in **Figure 5.5**. As can be seen from the figure, the characteristic diffraction peaks of all four samples are consistent with the standard lithium iron phosphate (LFP) pattern (PDF 81-1173) with no impurity peaks, indicating a high purity of the LFP samples. All four samples exhibit narrow and strong diffraction peaks, suggesting high crystallinity, which is attributed to the orthorhombic structure of the material (Kadoma et al., 2010). Furthermore, the two samples with added APG also showed no significant impurity peaks, indicating that the addition of APG does not affect the crystallinity of LFP. The APG acts as a carbon source providing hard carbon rather than graphitized carbon, thus not introducing new impurities.

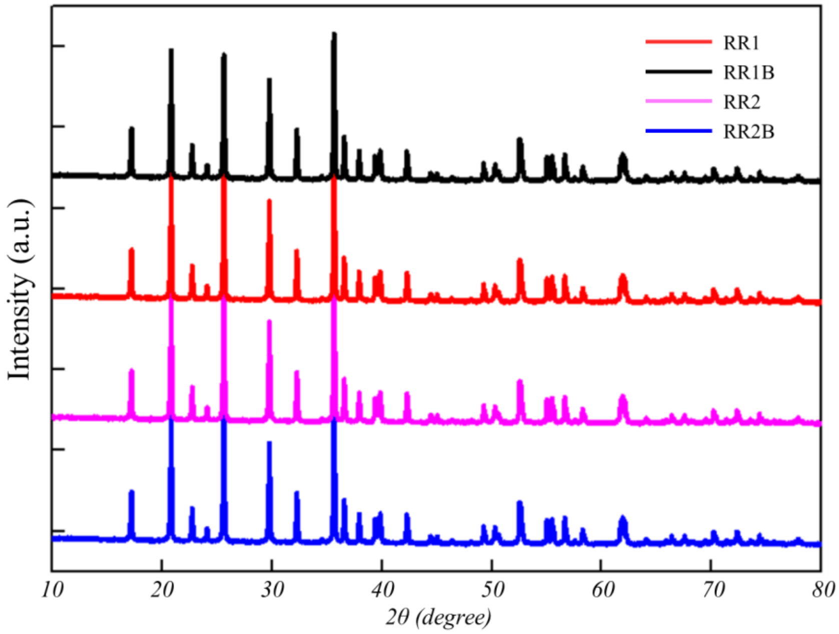

Figure 5.5 X-ray diffraction profiles of RR1, RR1B, RR2, and RR2B.

The $\rho_{30\mathrm{kN}}$ of RR2, RR2B, RR1, and RR1B are 2.599, 2.547, 2.588 and 2.587 g/cm^3, respectively. The SEM images of these four samples are shown in Figure 1 at magnifications of 10,000x and 50,000x. It can be observed that the particles exhibit a certain gradation in size, which helps enhance compaction density while ensuring electrochemical performance. In all four samples, micron-sized particles of LFP can be seen with nanosized particles growing on them. The boundaries between the different-sized particles are distinct, and the crystal structure is well-formed. The small particles range in size from 100 to 350 *nm* and exhibit high sphericity, which facilitates the extraction and insertion of Li$^+$ ions, thereby improving the effective utilization of the LFP material. These small particles adhere together to form larger secondary particles, significantly reducing particle agglomeration. In contrast, the large particles range from 1 to 4 μm in size and are morphologically regular, with smooth, rounded surfaces and no sharp edges. Additionally, the particle surfaces have a uniform carbon coating that forms a more complete carbon network between particles. This ensures good contact between particles and increases conductivity, resulting in excellent electrochemical and processing performance for the samples. The BET surface area of these four sample groups ranges between 14 and 16 m^2/g, which is moderate and conducive to sufficient contact between the active material and the electrolyte, thereby enhancing reaction efficiency.

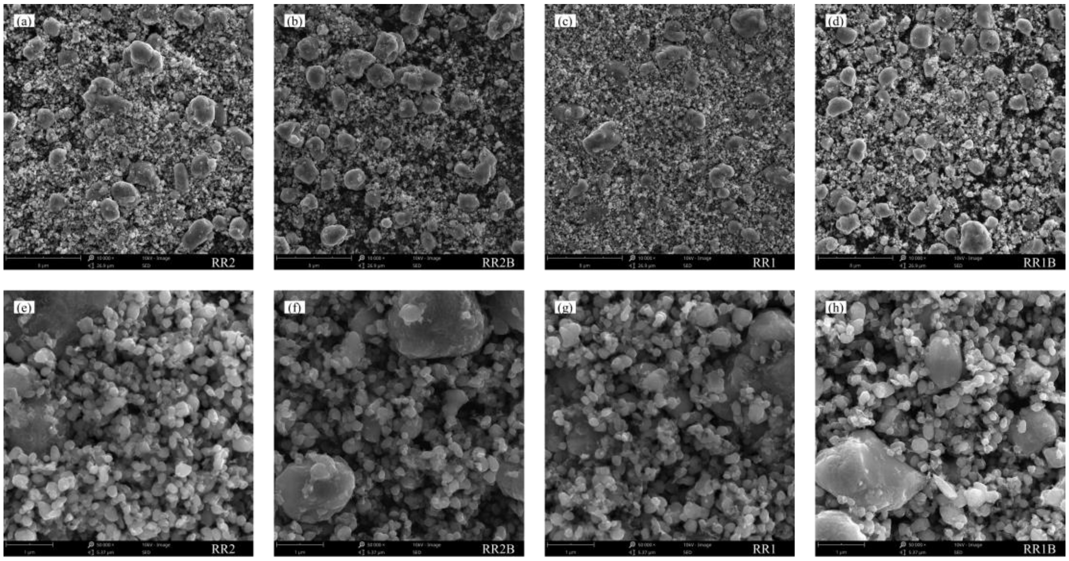

Figure 5.6 SEM images of the (a) the second-round-recommended sample RR2, (b) the control group RR2B, (c) the first-round-recommended sample RR1, and (d) its corresponding control group RR1B at magnifications of 10,000x. SEM images of the (e) the second-round-recommended sample RR2, (f) the control group RR2B, (g) the first-round-recommended sample RR1, and (h) its corresponding control group RR1B at magnifications of 50,000x.

Moreover, the SEM images reveal differences between the products prepared using the first-round recommended recipe and those from the second round. The number of micron-sized primary particles in the second-round products (RR2 and RR2B) is significantly higher than in the first-round products (RR1 and RR1B). An appropriate gradation of micron-sized primary particles with nanosized primary particles can yield high-compaction lithium iron phosphate products. The particle size distribution of the second-round products aligns with high-compaction gradation, which accounts for their superior performance compared to the first-round products. Additionally, the presence of nanosized primary particles in RR2 and RR2B ensures the stability of their electrochemical performance. The presence of APG tends to play a crucial role, as it potentially inhibits the growth of LFP primary particles. The two control groups (RR2B and RR1B) exhibit an excessive number of micron-sized primary particles, which fail to achieve optimal particle gradation for higher $\rho_{30\mathrm{kN}}$. The excessive number of micron-sized particles in RR2B and RR1B increases the Li^+ extraction/insertion distance, reducing the electrochemical performance of LFP products. Conversely, the presence of APG potentially inhibits the growth of some LFP particles, allowing for better particle gradation and shorter Li^+ extraction/insertion distances. This assumption has a high possibility of resulting in products with both higher $\rho_{30\mathrm{kN}}$ and $C_{1\mathrm{C}}$. The varying carbon source environments ensure that the micron-sized large particles do not grow excessively, maintaining them within a certain range and exhibiting an olivine structure, which further improves electrical performance.

In addition to the addition of APG forming a composite carbon source, which provides different carbon coating conditions resulting in more complete coverage of LFPs, metal doping also plays a vital role in the modification of LFPs. Specifically, all four samples were doped with aluminum (Al) to enhance the rate performance of LFPs. The test results indicated that the discharge capacity of the Al-doped cathode material initially increased and then decreased. This behavior can be attributed to the fact that an appropriate amount of Al^{3+} can enhance the material's conductivity, while excessive inert Al^{3+} occupying Li sites can reduce the active Li^+ content and discharge capacity. The Al^{3+} doping potentially improves the high-rate charge/discharge performance of LFP by inhibiting particle agglomeration. Additionally, doping at Li and Fe sites promotes the formation of mixed Fe^{3+}/Fe^{2+} redox couples and inhibits the formation of a single-phase $FePO_4$, both of which further enhance the material's conductivity (Zou et al. 2024). Besides Al, Ti was also added to the four samples to dope the LFPs and improve the discharge capacity and rate performance. This is because Ti^{4+} might occupy Li^+ sites, reducing interplanar spacing and particle size, and forming mixed-valence Fe^{2+}/Fe^{3+} states within the lattice, thus improving charge/discharge and cycling performance. However, excessive Ti^{4+} would lead to the formation of impurity phases like $Li_4P_2O_7$ and reduce the amount of active Li^+. This optimizes crystal morphology and prevents the growth of large LFP particles, thereby enhancing electrical performance.

5.3.4 Electrochemical Characterization

The electrochemical performance of the 4 samples is shown in **Figure 5.7**. **Figure 5.7**a shows the constant current and constant voltage (CC-CV) charge-discharge curves of the four samples at a rate of 0.2C, within the voltage range of 2.25-3.75V. The figure reveals that all four samples exhibit long and stable charge-discharge plateaus, indicating good stability. Among the samples with added APG (RR2 and RR1), the initial discharge capacity at 0.2C is slightly lower than the blank sample RR2B. However, the latter displays the largest voltage plateau difference, indicating the most severe polarization and the poorest cycling performance. Sample RR2, with added APG, has the smallest voltage plateau difference, indicating the least polarization and the best cycling performance, consistent with the machine learning model's second-round recommendations. Furthermore, the voltage plateau differences for the glycoside-added samples (RR2 and RR1) are smaller than those of the blank samples, suggesting that the addition of APG can reduce polarization and enhance battery performance.

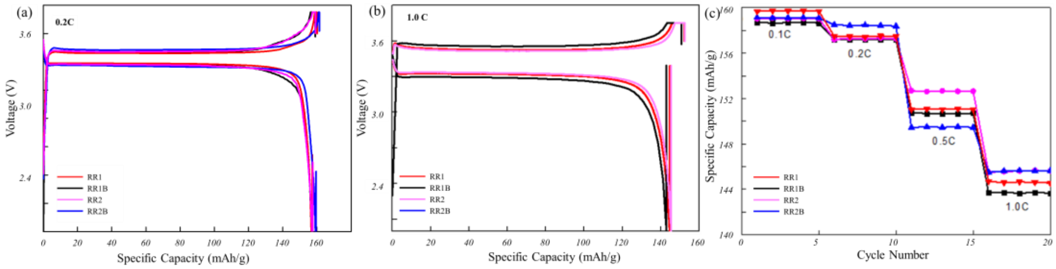

Figure 5.7 The initial voltage profiles of the four LFPs under (a) 0.2C and (b) 1.0C. (c) The discharge capacities of the four LFPs in continuous cycling at various rates of 0.1, 0.2, 0.5, and 1.0C.

Figure 5.7b presents the CC-CV charge-discharge curves at a rate of 1C, within the voltage range of 2.25-3.75V. The second-round recommended samples (RR2 and RR2B) exhibit similar 1C discharge capacities of 145.91 mAh/g and 145.68 mAh/g, respectively, which are higher than those of the first-round recommended samples (RR1 and RR1B), with 1C discharge capacities of 145.01 mAh/g and 143.36 mAh/g, respectively. Additionally, the second-round recommended samples have smaller voltage plateau differences and longer voltage plateaus than the first-round recommended samples, indicating smaller polarization values and better cycling performance, further validating the accuracy of the machine learning recommendations. The APG-added sample RR1 shows a higher 1C discharge capacity than the blank sample RR1B, with a smaller voltage plateau difference, proving that the addition of APG can reduce polarization and improve electrochemical performance.

Figure 5.7c illustrates the rate capability cycling curves for the four samples at 0.1C, 0.2C, 0.5C, and 1C rates. RR2 shows the highest discharge capacities at 0.5C and 1C, with values of 153.31 mAh/g and 145.91 mAh/g, respectively, indicating its suitability for use in power batteries. This sample also has a high powder compaction density of 2.58 g/cm^3, further demonstrating its potential for producing high energy density LFP. The second-round recommendations of ML model resulted in higher 1C discharge capacities of 145.91 mAh/g and 145.68 mAh/g, significantly outperforming the first-round recommendations, thus proving the feasibility of the machine learning approach for selecting optimal synthesis recipes. Additionally, RR1 shows higher discharge capacities at 0.1C, 0.2C, 0.5C, and 1C rates (159.91 mAh/g, 157.62 mAh/g, 151.5 mAh/g, and 145.01 mAh/g, respectively) compared to the blank sample RR1B with corresponding values of 158.91 mAh/g, 157.02 mAh/g, 150.6 mAh/g, and 143.36 mAh/g. This demonstrates that the addition of APG can optimize the electrochemical performance of LFP.

5.4 Conclusion

In summary, a data-driven active learning framework was proposed to recommend synthesis recipes for high-performance LFP material. This framework integrated two ensembled ML models to iteratively refine the synthesis parameters, aiming to enhance the physicochemical properties ($\rho_{30\mathrm{kN}}$) and electrochemical performance ($C_{1\mathrm{C}}$) of the resulting LFP samples. By employing this data-driven methodology, the synthesis process was guided, leading to the identification of high-performing samples in both the first and second rounds of active learning. The electrochemical characterization revealed that these samples, synthesized according to the machine learning recommendations, displayed superior properties. The addition of APG was found to significantly reduce polarization and improve cycling performance. Notably, the second-round recommended sample, RR2, demonstrated the highest discharge capacities, the best cycling stability, and the highest compaction density, validating the efficacy of the active learning loop. In conclusion, the active learning loop proposed in this study successfully identified and optimized critical synthesis parameters, resulting in the discovery of two high-performance LFP samples. This study underscores the potential of integrating machine learning into the materials synthesis process, paving the way for the development of advanced battery materials with enhanced properties.

Chapter 6

A Deep-Learning-Assisted Approach for Fault Detection and Real-Time Monitoring for Steam Boilers

Part of the content in chapter has been or is ready to be submitted:

WANG, Z., MENG, Y., FOW, K. L., WU, T. and PANG, C. H. "A Deep-Learning-Assisted Approach for Fault Detection and Real-Time Monitoring for Steam Boilers." *To be submitted.*

WANG, Z., YEOH, J. X., WONG, C. D. S., and PANG, C. H. (2022). "Fault Detection and Diagnosis of Steam Boiler Operation Process with Multi-way Principal Components Analysis" *Applied Energy Symposium 2022*.

6.1 Synopsis

Fault detection and online monitoring for steam boilers in the real industry are challenging tasks due to the complexity and nonlinearity of the operation process. As critical industrial equipment works at high temperatures and pressures, steam boilers are prone to faults such as leakage and break. Some of the faults are not obvious at the start but might cause severe issues without proper and timely maintenance, where the development of those faults would consume additional energy and lower the combustion efficiency, leading to extra carbon emissions and diminishing carbon neutrality. In this study, a generalized framework incorporating conventional long-short-term memory (LSTM) network and multi-way principal components analysis (MPCA) is developed to apply fault detection and monitoring techniques to the dynamical steam boiler operation process. The proposed deep-learning-based method in this work can predict the future behavior of steam boilers, evaluate the process condition, prevent further fault development, and avoid safety issues and economic loss, only using a historical database of past normal operations. The proposed method employed simulated operation data to establish a framework with several critical stages including data pre-processing, establishment of a historical database, calculation of statistical control limit, fault detection and online monitoring, which are intuitive and straightforward to understand and identify faults. The framework exhibited excellent fault prediction ability using actual data acquired from real industrial operations, indicating that accurate and timely support and suggestions could be effectively provided for monitoring and maintaining the operation of steam boilers in real industry.

6.2 Introduction

In the previous two chapters of this thesis, we focused on enhancing the effectiveness and efficiency of material design and synthesis through data-driven strategies. While these techniques have rapidly gained traction in recent micro-level research, their application in real industrial operations remains relatively underexplored. Specifically, steam boilers are integral to the power, chemical, and manufacturing sectors, where they convert water into steam for applications such as electricity generation (Król and Ocłoń, 2018), oil refinery (Yi et al., 1998), and food industry (Biglia et al., 2017). Operating at high temperatures and pressures with combustion processes, faults during steam boiler operation can lead to severe safety issues, significant economic losses, reduced combustion efficiency, excessive greenhouse gas emissions, and negative impacts on carbon neutrality (Swiercz and Mroczkowska, 2020, Xi et al., 2021, Wiryadinata et al., 2019, Li et al., 2022a).

The inefficiencies and pollution of small- and medium-sized boilers in underdeveloped regions are particularly severe. These boilers often suffer from high exhaust temperatures, low thermal efficiency, and lack automation, leading to poor performance and difficulty in adapting to changing conditions. Due to scattered geographical locations and limited data infrastructure, remote monitoring is a challenge, and existing systems struggle with processing real-time data that is nonlinear and influenced by operational load. This chapter introduces a deep-learning-based framework for fault detection and real-time monitoring of steam boilers operating in batches.

The community has widely investigated the method of processing batch data. Pioneering works by Jackson and Mudholkar (Jackson and Mudholkar, 1979), proposed residuals associated with PCA as control statistics for researching multivariate processes, while Kourti and MacGregor (MacGregor and Kourti, 1995) and Nomikos and MacGregor (Nomikos and MacGregor, 1994) introduced MPCA to handle high-dimensional, multi-way data structures for a defined duration. Subsequent studies led to numerous improvements and refinements, including dynamic PCA and dynamic PLS models by Chen and Liu (Chen and Liu, 2002) for real-time monitoring, and robust PCA by Hubert et al. (Hubert et al., 2005) for addressing outliers, and multi-phase MPCA by Wang et al. (Wang et al., 2012a) for specific injection molding processes.

Aside from MPCA, various other tensor decomposition techniques have been applied to fault detection in batch processes. For instance, Wise et al.(Wise et al., 2001) employed a model named parallel factor analysis 2 for detecting the fault induced by the changing of control variables such as chamber pressure and plasma power in semiconductor etch, and Deng et al.(Deng et al., 2019) proposed an outlier detection method based on tensor Tucker factorization. A Bayesian temporal factorization framework proposed by Chen et al.(Chen and Sun, 2022) showed superiority in processing large-scale and multi-dimensional spatiotemporal data sets. To comprehensively exam timely research on data-based process monitoring, please refer to the recent critical review.(Ge et al., 2013, Lu et al., 2006, van Sprang et al., 2002) Besides, various techniques employed in steam boiler fault detection and real-time monitoring have been proposed, such as model-based methods (Addel-Geliel et al., 2012), programming logical control and supervisory control and data acquisition (Mohod and Raut, 2019), which rely on pre-defined rules and expert knowledge, and data-driven approaches, including bagged auto-associative kernel regression-based fault detection for steam boilers in thermal power plants (Yu et al., 2017), and artificial neural network-based method for modeling complex relationships between process variables (Rakhshani et al., 2009). Sun et al. (Sun et al., 2005) applied PCA to leak detection in boilers, demonstrating its effectiveness in reducing data dimensionality but also revealing limitations in handling multi-batch and multi-way data structures.

Although advanced techniques for steam boiler fault detection and real-time monitoring have been proposed, they often face limitations such as high computational complexity, low interpretability, and the need for specific data structures. Supervised methods, for example, require pre-processed data, while tensor decomposition methods demand extensive hyperparameter tuning.

Steam boiler operation is a complex, nonlinear process with time-varying features and disturbances. While MPCA has been widely used for fault detection in batch processes, its application to steam boilers is limited. LSTM, a type of recurrent neural network (RNN), can handle vanishing gradients and selectively output information, offering great potential for accurately predicting steam boiler states in both current and future time steps.

This study proposes a generalized framework combining LSTM and MPCA for early fault detection and real-time monitoring of horizontal steam boilers operating in batches. Unlike supervised methods, this framework requires only historical operational data and does not depend on fault-specific data, which is often hard to obtain. LSTM-MPCA outperforms traditional MPCA by processing time-series data and predicting future behavior, allowing for faster, more sensitive fault detection with dynamically adjusted thresholds and low computational cost. The framework demonstrated effectiveness using both simulated and industrial datasets, identifying anomalous batches and fault locations with high accuracy. Overall, the LSTM-MPCA framework offers a practical, interpretable method for detecting faults in steam boiler operations, helping operators prevent severe industrial failures.

6.3 Data Generation and Collection

6.3.1 The Generation of Simulation Dataset.

The design parameters and equipment sizing (**Table 6.1**) were set in Aspen Plus. Subsequently, the process variables were transferred to Aspen Plus Dynamic to determine the dynamic effect of each stream from time 0 minute up to 720 minutes. To evaluate the feasibility of the proposed generalized MPCA framework for fault detection and real-time monitoring in steam boilers, a simulated dataset consisting of 160 normal batches, each with 145 time points and 135 variables, is selected. Employing a dataset with a relatively large number of variables for training an MPCA model offers several advantages, such as enhanced feature capturing, identification of higher-order correlations between variables, improved robustness of the MPCA model to noise and outliers, and potentially increased generalization. These benefits ultimately result in superior performance and stability, ensuring the framework's effectiveness in detecting and monitoring industrial data. At the base case, the soft water input and heat loss of COMB were 2000 kg/h and -7000 W, respectively. The load is varied from 1200 kg/h to 3000 kg/h with a step size of 200 kg/h, and the heat loss is changed from 0 W to -15000 W with a step size of 1000 W. Fault data was generated for the machine learning algorithm to learn and identify possible future faults. Faulty situations are defined as abnormal values of input variables which may lead to variation in boiler output. The faults in the boiler were simulated by varying the natural gas inlet flow rate for cases L1H1 and L10H16 (Table S1). The increment in natural gas was set at 50% and 100% higher than the initial value. The specifications of these sets of abnormal data generation are shown in Table S2.

In total, 135 variables (Table S3) obtained from Aspen Plus Dynamic for streams after combustor include molar flow, mass flow, volume flow, temperature, pressure, mole fraction and mass fraction of all components (CH_4-methane, O_2-oxygen, N_2-nitrogen, CO-carbon monoxide, CO_2-carbon dioxide, N_2O-nitrous oxide, NO_2-nitrogen dioxide, H_2O-water, NO-nitrogen oxide). Similarly, results for three heat exchangers include duty, LMTD correction temperature, hot side and cold side pressure drop; for the combustor include temperature, pressure, and heat loss; for drum include temperature, pressure, and liquid level.

Table 6.1 The design parameters of the steam boiler in Aspen Plus for the generation of the simulated dataset

Parameters	Value	Unit	Parameters	SMK TUBE	ECOMM	COND
Work Capacity	2000.00	kg/hr	Shell Inner Diameter (in)	19.25	23.25	23.25
Steam Pressure	1.25	MPa	Shell Outer Diameter (in)	22.31	25.79	24
Steam Temperature	190.00	°C	Length (in)	98.00	29.81	47.24
Air to Natural Gas Ratio	10.38	-	Baffle spacing (in)	8.00	3.50	5.12
Excess Air %	9	-	Tube Outer Diameter (in)	0.75	0.75	0.75
Flue Gas Economizer Outlet Temperature	113.00	°C	Tube Pitch (in)	0.9375	0.9375	0.9375
			Number of tubes	300	570	250
Flue Gas Condenser Outlet Temperature	62.00	°C	Number of passes	1	1	1
			Location of hot fluid	Tube side	Tube side	Tube side
Pump Efficiency	75	%	Area for heat exchange (m^2)	31.30	22.30	16.44

6.3.2 Real Industrial Data Collection

For the preparation of the real industrial dataset, the specific industrial operation 2021/22 dataset for a steam boiler is utilized to perform the analysis, and several process variables measured during the control process are used. **Figure 6.1** illustrates the structure of the one-drum and three-pass horizontal steam boiler researched in this work, and the description of the recorded process parameters is shown in **Table 6.2**. The dimensions of this unit are 5.2 meters long, 3.3 meters wide, and 2.5 meters high, with a total weight of 7551 kg and a full water volume of 4.65 m^3. The rated amount of evaporated steam is two tons per hour with a rated pressure of one MPa. This steam boiler is fired with natural gas and equipped with a burner operating between 1100 kW and 1950 kW.

The core parts of the horizontal steam boiler include the drum, the burner, the flue, the combustion chamber, the steam-water system, meters, the economizer, and the supporting base. The steam from the drum is supplied directly to the demand end, and the air and natural gas are supplied together to the combustion chamber. An economizer is located at the exhaust gas outlet to collect heat and pre-heat the water supply.

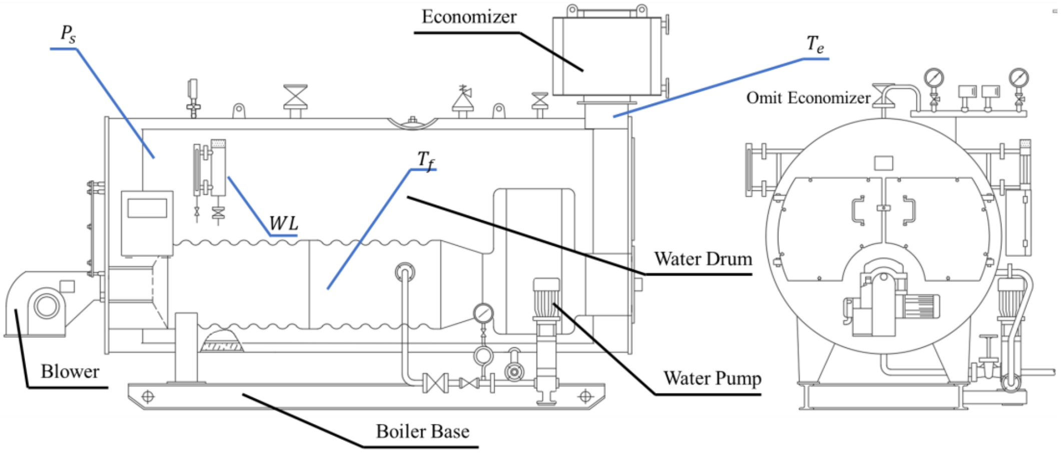

Figure 6.1 The schematic view of the horizontal steam boiler. $\boldsymbol{T_f}, \boldsymbol{P_s}, \boldsymbol{T_e}$ and $\boldsymbol{WL}$ represent (1) the temperature of the inlet fuel (2) the pressure of the generated steam, (3) the fuel temperature at the inlet of the economizer and (4) the water level of the drum, respectively.

Table 6.2 The process parameters recorded by the installed sensors for the generation of the real industrial dataset.

Parameters	Unit	Type	Description
T_f	°C	Continuous	The temperature of the inlet fuel
P_s	MPa	Continuous	The pressure of the generated steam
T_e	°C	Continuous	The fuel temperature at the inlet of the economizer
WL	%	Discrete	The water level of the drum

In the specific plant where the data was collected, only a limited number of sensors were installed, allowing for the recording of only three continuous variables and one discrete variable. For each time point, there are four variables measured: (1) the temperature of the inlet fuel (T_f), (2) the pressure of the generated steam (P_s), (3) the fuel temperature at the inlet of the economizer (T_e) and (4) the water level of the drum (WL). The first three variables are collected in the form of continuous values, whereas the water level data is collected discretely as the sensor only has seven levels from 0 to 100 (0, 1/6, 1/3, 1/2, 2/3, 5/6 and 1). **Table 6.2** shows the type and unit of the collected process parameters.

The first stage involves creating a reference database and training MPCA models on twelve operation batches, each containing 4 variables and 3200 time points, to identify abnormal batches. Eleven normal batches (Batch 1, 3-12 in **Figure 6.2**) display consistent variable variations, establishing the baseline for tolerance amplitude. The scatter plot of pairwise correspondence between variables are shown in **Figure 6.3**, whose diagonal plot shows the distribution of the variables. The T_e of batch 2 is higher than that of the rest eleven batches and the P_s of batch 8, 9, and 10 are relatively lower. Although batches 8, 9, and 10 exhibit lower steam pressure, they are still considered normal but with different loads. In the validation stage, batch 2 and 4 served as the test batch, while batch 2, characterized by higher and irregular fluctuations and significant shifts in all variables occurring between time points 1750 and 1950, is labeled abnormal. The batch data is organized into a tensor and then unfolded along the time series. Standardization is applied to the dataset, ensuring that each column has a mean of 0 and a standard deviation of 1. This preprocessing step helps to normalize the data and improve the performance of the MPCA models during the training and fault detection process.

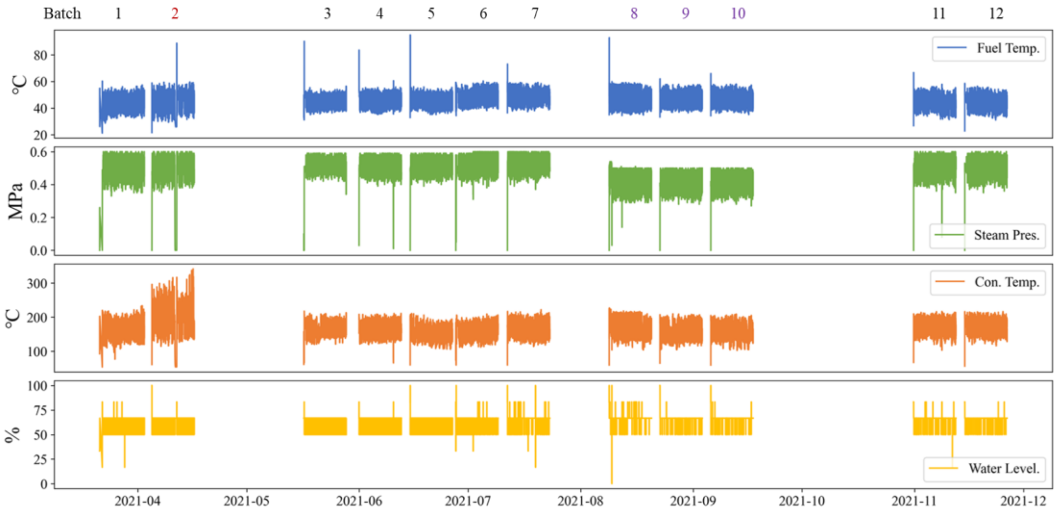

Figure 6.2 The trajectories of the four variables of the selected 12 batches of the real industrial data. Batch 2 has a relatively higher fluctuation amplitude in $\boldsymbol{T_f}$, $\boldsymbol{P_s}$ and $\boldsymbol{T_e}$. The fluctuation range of batch 8, 9, and 10 in $\boldsymbol{P_s}$ are lower than other batches.

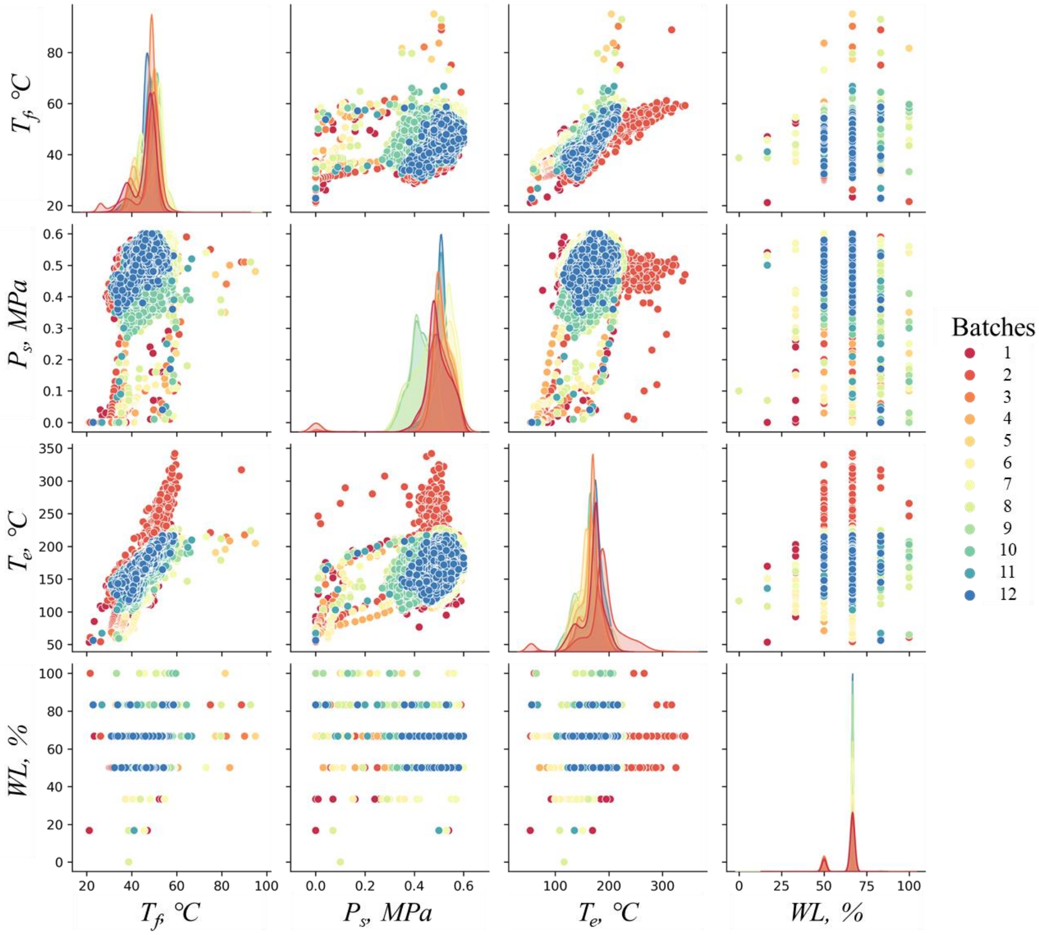

Figure 6.3 The scatter plot of pairwise correspondence between the four variables. The diagonal plot is the distribution of the variables.

6.4 Results and Discussion

6.4.1 Abnormal Batch Detection with MPCA

To deploy the proposed framework to a new system, the primary step is to learn the data patterns from the historical database of normal batches. The statistics mentioned in Chapter 3 will determine the confidence region, where batches within this region can be regarded as normal batches. Therefore, it is essential first to identify the normal batches for learning and exclude the abnormal batches from the database. This step can be efficiently executed using MPCA individually, which clusters and classifies the batches to indicate whether the score vectors contain sufficient information to measure the similarity of a batch to typical normal batch operations. This determination is based on the statistical consistency of variable measurements of the test batch with the statistical benchmark from normal batches, as summarized by the MPCA model.

6.4.1.1. Abnormal Batch Detection on Simulated Data

Utilizing MPCA on 160 normal batches allows for the summarization of data patterns and information essential for confident limits determination. The results show that only two PCs can explain over 85% total variance of the simulated dataset. More specifically, the plane of the first two principal components (PC1, PC2) demonstrates that there are discrete variables that contribute the most variance (**Figure 6.4**a), which caused the layer-by-layer arrangement of the points on the plot, with each layer containing 16 points, reflecting that the data is generated with 10 levels of load and 16 levels of heat loss, representing 16 batches with the same load. The confidence ellipses (**Figure 6.4**a), based on T^2 statistics, reveal that batches with higher heat loss are positioned closer to the ellipse edge, while abnormal batches are distant from the score plane origin.

The *Q* statistics of the batches falling (**Figure 6.4**b) below the confident limits indicate that the model adequately explains the normal batches in an academic context. Moreover, the Q statistics of the four abnormal batches (161-164) are far above the confidence limits, further validating the model's ability to differentiate between normal and abnormal batches. As the MPCA seeks to explain all the predictable variable variations in the normal batch dataset, it is informative to investigate the percentage of the explained variation in each variable and at each time point. The importance of principal components (PCs) in explaining process variability is highlighted by the significant contribution of PC1 (as indicated by the blue bar) in **Figure 6.4**c, while PC2 and PC3 exhibit a more dominant impact on specific variables, with the three PCs combined explaining over 80% of the variance in 90% of the simulated process measurements. Therefore, in the process that employs MPCA on simulated data, three PCs are selected based on the percentage of explained variance to comprise the PC space. **Figure 6.4**d uses the terms PC1, PC1+2, and PC1+2+3 to show the ability of PCs to explain variable variance as a function of time, with the diagram plotted on a cumulative basis. The amount of variance accounted for over time, shown in **Figure 6.4**d, indicates that PC1 consistently plays a dominant role in variability throughout the simulated process.

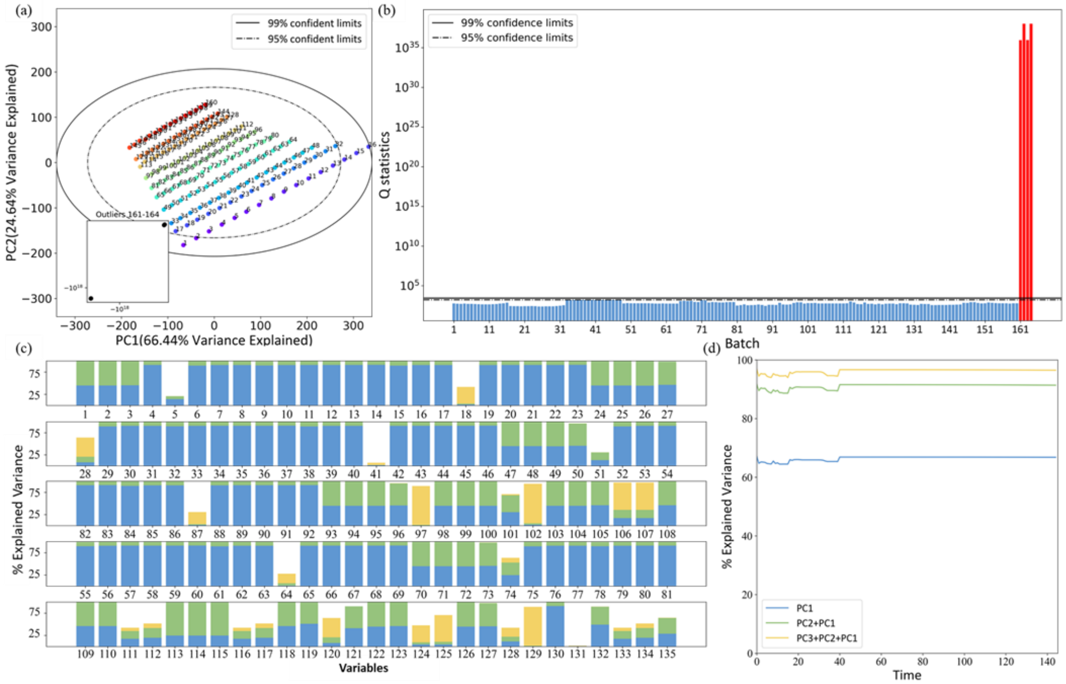

Figure 6.4 (a). The 164 batches on the plane of the first two PCs. The 160 normal batches define the normal operation region in the PC space. (b). The sum of squares of the residual of the 164 batches, with the defined 95% and 99% confident limits. The red bar represents the four abnormal batches. Percentage of the explained variance with respect to (c). variables (blue, green, and yellow represent PC1, PC2, and PC3, respectively) and (d) time, plotted on a cumulative basis.

6.4.1.2. Abnormal Batch Detection on Industrial Data

The MPCA model is then applied to the industrial dataset. The projection of the 12 batches onto the plane of PC1 and PC2 is shown in **Figure 6.5**a, where the suspected abnormal batch is located far from the cluster of normal batches. This clustering clearly indicates that batches with similar process parameters will cluster closer together on the PC plane, reflecting different operating conditions. More specifically, as shown in **Figure 6.2**, batches 8, 9, and 10 have slightly higher $\boldsymbol{P_s}$ are closer to each other on the PC plane. The 90%, 95%, and 99% Hotelling confidence ellipsoids in **Figure 6.5**a highlight that batch 2 exhibits relatively unusual behavior. Besides the T^2 statistics, the Q statistics for each batch are illustrated in Figure 8b, representing the squared distance of each batch in the principal space to the plane.

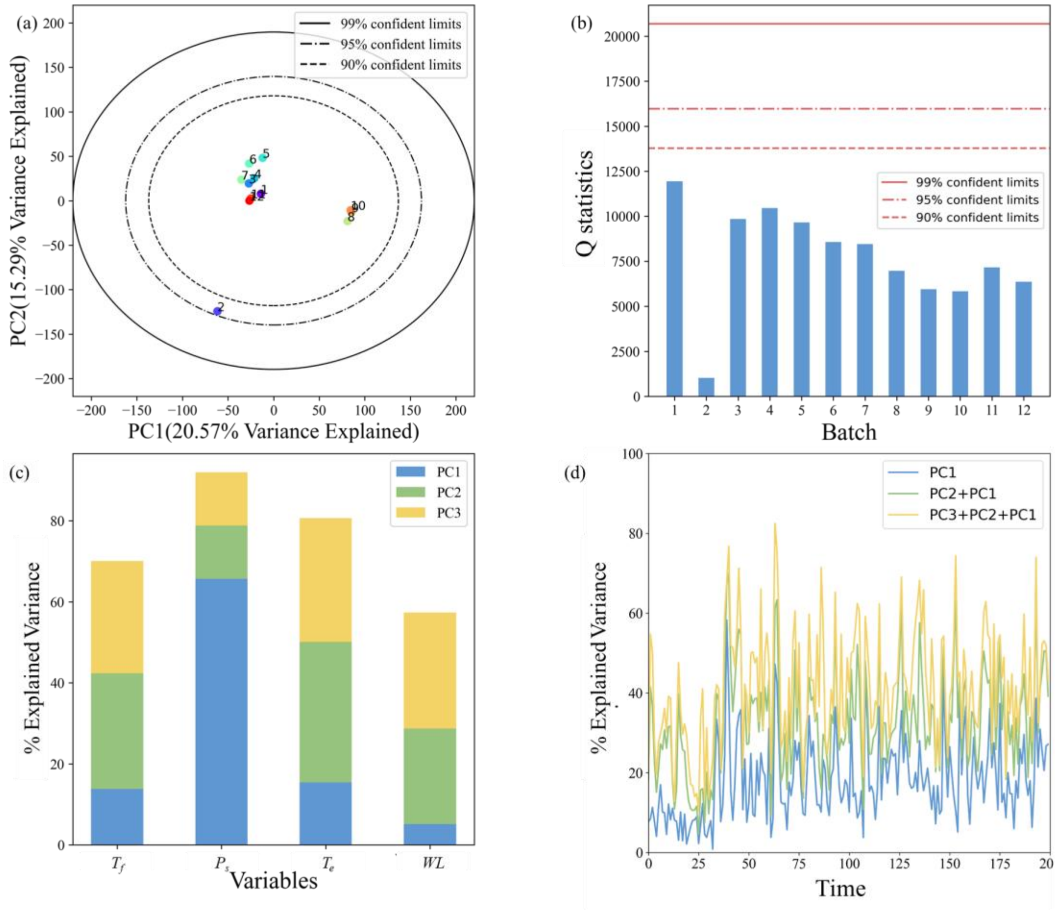

Figure 6.5 (a) The 12 batches on the plane of the first two PCs. Batch 2 is located far away from others. (b) The plot Q statistics of the 12 batches, with 90%, 95% and 99% confidence limits. The percentage of the variance explained with respect to (c)

the measured variables and (d) the time points for each of the three PCs, plotted on a cumulative basis.

Figure 6.5c demonstrates that three PCs are capable of explaining 70% of the variance of the first three variables of the industrial data. As the last process parameter, water level (*WL*) is recorded discretely, the ability of the PCs to explain its variation is relatively lower. More specifically, steam pressure plays the dominant role in PC1. All three PCs explain the variation of temperature at the fuel ($\boldsymbol{T_f}$) and the inlet of the economizer ($\boldsymbol{T_e}$), while PC2 and PC3 mostly explain the *WL*. The results in **Figure 6.5**c indicate that three PCs are sufficient to retain the data structure and explain the variance of variables. The amount of variance explained by each PC with respect to time points is plotted in **Figure 6.5**d. It can be concluded that PC1 explains much of the variability at most time points, while at some time points, PC2 also captures significant variability.

6.4.2 Real-Time Monitoring of Boiler Behaviors

The essence of early warning and real-time monitoring is the ability to predict the PC value of the next time point, allowing control limits to determine its state. However, as mentioned above, the real measurements of the future haven't been taken, and hence accurate prediction is critical. The LSTM-MPCA model demonstrates excellent prediction ability for both variable and PC values of the next time point, with R-squared values of 0.9204 and 0.9645 on the testing set, respectively. A data comparison between the real and predicted PC1 and PC2 in one test simulated batch is shown in **Figure 6.6**, indicating high accuracy. This high accuracy suggests that the predicted PCs can be used to determine the process behavior of the next point.

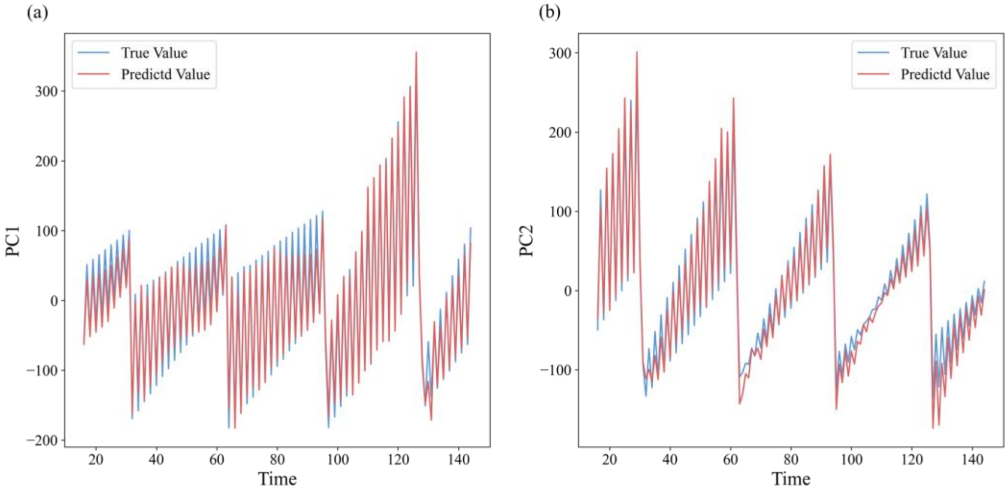

Figure 6.6 The comparison between the real MPCA (a) PC1 and (b) PC2 values and predictive values by the ensemble LSTM model.

While the LSTM model can make accurate predictions in time series, it requires a sequence of time points and, therefore, cannot fully replace MPCA at the starting stage of a batch process. Furthermore, the R-squared values indicate that the LSTM model for predicting MPCA scores performs better than only predicting variable values. This suggests that MPCA captures important data features and has the potential to improve LSTM performance. However, when handling data from different processes, the hyperparameters of the LSTM should be adjusted accordingly.

6.4.2.1. Real-Time Monitoring of Simulated Steam Boiler Behaviors

The LSTM-MPCA model is then employed for investigating the state of boilers along with the time series, which is helpful for early warning and pinpointing the time point when faults might occur. The predicted future behavior of the operation process is compared against the reference distribution defined by the historical normal batch dataset. The results of applying such a time-series-based fault detection strategy to a normal batch (blue scatter) and an abnormal batch (red scatter) are illustrated in **Figure 6.7**. The control limits shown in **Figure 6.7** are derived based on Q and T^2 and statistics, respectively. A normal batch should behave similarly to the batches in the reference database. The red scatter located below the control limit before the time point of 240 minutes aligns well with the generated fault data, indicating that the LSTM-MPCA model can effectively detect faults and trace the specific time point where abnormal behavior occurs.

Given that multivariate statistical process control incorporating LSTM-MPCA can detect simulated faults, the employed method uses joint covariance and is sensitive to finding not only magnitude shifts of variables but also breaks in the inner correlations among variables. The LSTM-MPCA model detects faults by capturing large shift variables and extracting the relationship information among them during the multivariate process. By using only a historical dataset of normal batches, the LSTM-MPCA model is capable of detecting abnormal behaviors in new batches.

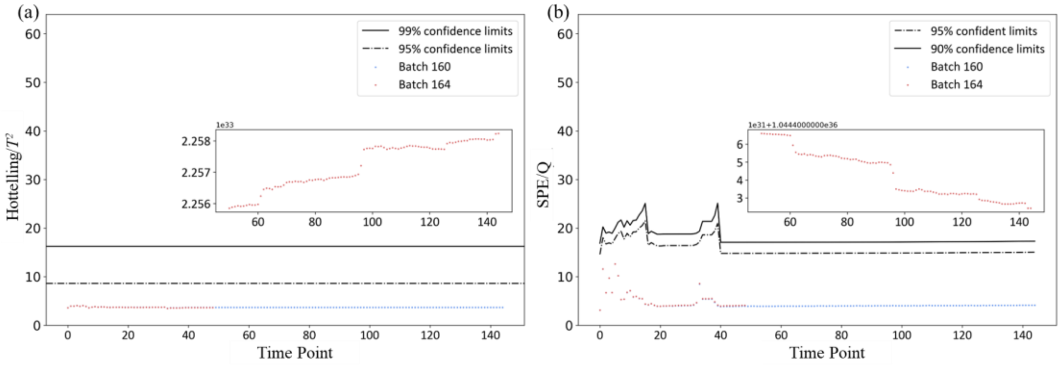

Figure 6.7 Fault detection charts with 95% and 99% confidence limits based on (a) Hottling's T^2 and (b) SPE. The fault occurs at the 49th time point (240 minutes) of batch 164 (red scatter).

6.4.2.2. Prediction of Future Industrial Steam Boiler Behaviors

The result of implementing this monitoring strategy on real industrial data is shown in **Figure 6.8**. The variable measurements of the abnormal batch (batch 2) are first pre-processed and then unfolded to obtain the observation vector $\boldsymbol{x}^{*}_{test,k}$. **Figure 6.8**a shows a dynamic plot of the combined index of batch 2 at every time point with 90%, 95%, and 99% confidence limits. The combined index plot can detect faults correlated with both PCs and residuals in a complementary way. Although the sudden shutdown around time point 1800 can be easily detected via the significant shift of single variables, the monitoring scheme can also detect abnormal time points where the inner correlations of variables are broken without introducing considerable deviation. For example, from the start to time point 2000, more than 10% of sample points are determined as abnormal points, which are difficult to detect by measuring the deviation of T_e. **Figure 6.8**b shows the monitoring process of a normal batch. It can be observed that there are several single points located above the control limit, caused by sensor offset, indicating that the LSTM-MPCA model is also capable of detecting abnormal behaviors at specific time points in a normal batch.

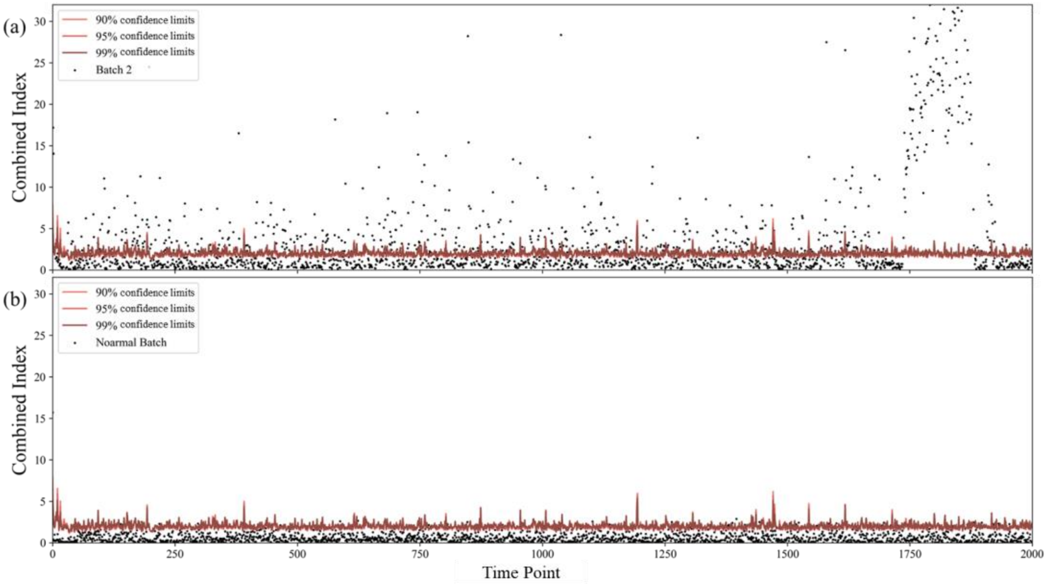

Figure 6.8 The performance of LSTM-MPCA-based real-time monitoring on real industrial data. Confidence limits (90%, 95%, and 99%) of the combined index for (a) an abnormal batch (batch 2), and (b) a normal batch (batch 4).

The proposed generalized MPCA model incorporates implicit constraints, considering both the magnitude and trends of deviations of variables and the correlations among all variables derived from the historical benchmark. The encoded information about data structure and relationships is crucial for fault detection. When the relationships among the deviations of variables change during the process, the likelihood of fault occurrence is typically high, even if the fluctuation magnitude is not substantial. It is important to note that analyzing historical benchmark data is essential for estimating confidence limits in monitoring plots and defining the normal region. The calculation algorithm is not complex but computationally intensive, as a series of covariance matrices must be generated and stored. However, once a sufficiently large benchmark database is established and the resource-intensive offline processing is completed, the online monitoring scheme demands relatively low computational resources.

The efficacy of this deep learning-based, data-driven approach lies in its multivariate consideration of measurements, where fluctuations of deviations and trends are meticulously accounted for, and inter-variable correlations are accurately evaluated. The primary aim of this method is early fault detection to prevent severe batch performance disruption or shutdown.

6.5 Conclusion

This chapter introduces a LSTM-MPCA framework for fault detection and real-time monitoring of steam boiler operations. Developed using simulated boiler operation data and validated with actual industrial data, the framework utilizes LSTM to forecasting future behavior and MPCA to effectively distinguish and categorize different batch types through clustering patterns in a reduced space, achieved by unfolding the three-dimensional data tensor of completed normal batches along the time sequence. Upon establishing a benchmark database of standard batches, the statistical reference control limits can be determined for normal operation at each time point.

Despite the progress made, there remain opportunities for further research and enhancement. These include using specific decomposition techniques tailored to dataset characteristics for more efficient and accurate fault detection, and incorporating advanced machine learning techniques for fault diagnosis and prevention. Moreover, the approach presented leverages machine learning to predict the time point of a potential future anomaly based on the monitoring data from the four parts of the steam boiler system. While the model successfully predicts when an anomaly may occur, it does not currently identify the specific part of the system responsible for the anomaly. However, this methodology can be further enhanced to provide more detailed feedback by integrating fault detection and classification models into the predictive framework. By analyzing sensor data from each individual component of the boiler, machine learning techniques could be used to pinpoint which part is most likely to cause the impending failure. This would allow for more targeted maintenance efforts and early interventions, ultimately improving system reliability and safety. Future work will focus on refining the model to isolate and identify critical failure points within the system, thereby extending the utility of this approach for both predictive and diagnostic purposes.

This study demonstrates the remarkable adaptability and generalizability of the LSTM-MPCA framework in monitoring batch steam boiler processes when integrated with advanced data-driven techniques, which shed light on potential to broaden its application for fault diagnosis in various scenarios.

Chapter 7

Conclusion

7.1 Conclusion

This thesis has demonstrated the successful integration of data-driven innovations into material science and chemical engineering, showcasing significant advancements in catalyst design, material synthesis, and industrial process monitoring. Our results show that employing data-driven technologies can significantly promote the efficiency catalytic ORR, which plays a vital role in chemical-electrical energy conversion in fuel cells and metal-air batteries and is a promising and indispensable field in the development of renewable energy. Besides, the combination of two ensemble ML models in an active learning loop can dynamically optimize the synthesis parameters of LFP materials, recommending two LFP samples with superior energy density. Furthermore, the incorporation of LSTM and MPCA can effectively predict the behavior of steam boiler, achieving early warning and fault detection, preventing low energy efficiency and serve safety issue, and lowing CO_2 emissions.

In the first part of this work, the structure-activity relationship of Pt-based alloy catalysts for ORR was determined using high-throughput DFT computations combined with the SISSO algorithm. This integration enabled the identification of innovative descriptors based on primary structural features, facilitating the screening of materials without the need for time-consuming electronic structure calculations. Out of 77 potential candidates, five alloys—Pt_3Co(211), $PtPd_3$(211), Pt_3Ni(111), $PtPd_3$(111), and PtAu(111)—were identified as the most promising for ORR, demonstrating nearly thermoneutral OH adsorption at low overpotentials. The use of structural information as a predictor for catalytic activity marks a significant advance in catalyst design. It provides a direct and efficient approach for identifying new materials with enhanced performance potential, which can greatly accelerate the development of new electrocatalysts for ORR and other energy-related processes.

The second part of this thesis focuses on optimizing the lab-scale synthesis parameters for $LiFePO_4$ (LFP) materials using a data-driven framework built on active learning. Through two rounds of active learning (RR1 and RR2), the system recommended optimal synthesis parameters that led to the production of LFP samples with exceptional electrochemical properties. Specifically, the C_1C value of the RR2 sample exceeded 145 mAh/g with a $\rho_{30}kN$ of 2.599 g/cm^3, outperforming all other samples in the original dataset. Feature importance analysis revealed that sintering temperature (Ts) was the most influential factor, guiding future LFP material design by focusing on synthesis conditions that maximize performance. This approach demonstrates the power of integrating machine learning techniques into the material synthesis process, enhancing both the quality and efficiency of energy material development. The successful optimization of LFP synthesis via active learning highlights the potential of data-driven methodologies to revolutionize material design, providing a pathway to more efficient and cost-effective battery materials.

The third part of this work explores the integration of data-driven tools, specifically LSTM (Long Short-Term Memory) networks and MPCA (Multivariate Process Control Analysis), into traditional chemical engineering systems, with a focus on batch steam boiler operations. The results of this case study illustrate how data-driven techniques, such as deep learning models and statistical control limits, can be leveraged to enhance the safety, reliability, and efficiency of industrial processes. The LSTM-MPCA model demonstrated outstanding predictive performance, with R-squared values of 0.9204 and 0.9645 for the variable and principal component (PC) values, respectively, on the testing set. These results show the exceptional capability of the framework to predict and monitor system behavior, providing real-time insights that can preemptively address potential faults and prevent costly downtimes. The approach offers a scalable solution for optimizing a wide range of non-linear industrial processes, making it highly adaptable to other batch operations in the chemical and manufacturing sectors.

In conclusion, this thesis illustrates the significant impact of integrating data-driven innovations into material science and chemical engineering. By optimizing electrocatalyst design for ORR, enhancing the synthesis of LFP energy materials, and improving the operational efficiency of conventional chemical processes, this research demonstrates the potential of data-driven tools to advance the development of green chemical technologies. These findings provide valuable insights for future research in energy materials, green chemical engineering, and industrial process optimization, and they underscore the transformative role of data science in accelerating the transition towards a more sustainable, efficient, and environmentally friendly chemical industry. The integration of data-driven frameworks in these domains will continue to promote innovation and pave the way for more intelligent, automated, and scalable solutions in materials and industrial processes.

7.2 Major Contributions

The major contributions of this thesis lie in the successful application of data-driven innovations to material science and chemical engineering, aimed at enhancing energy material design and improving energy efficiency and safety in conventional chemical industrial processes.

1. Chapter 4 focused on high-throughput first-principles calculations to screen high-performance catalysts for ORR. The study established a structure–activity relationship, identifying five promising catalyst candidates: $Pt_3Co(211)$, $PtPd_3(211)$, $Pt_3Ni(111)$, $PtPd_3(111)$, and PtAu(111) alloys were identified. Key factors influencing ORR performance, such as the electronegativity difference between Pt and heteroatoms, the number of valence electrons, and the ratio of heteroatoms around Pt, were determined. This work provides valuable data for experimentalists to validate ORR activity and offers insights for data scientists to refine ML models for catalyst performance prediction.

2. Chapter 5 proposed a data-driven active learning framework to recommend optimal synthesis recipes for high-performance LFP material. The framework utilized two ensembled ML models to iteratively refine synthesis parameters, targeting enhanced physicochemical properties ($\rho_{30\text{kN}}$) and

electrochemical performance (C_{1C}) of the resulting LFP samples. The RR2 and RR1 sample, synthesized according to the parameters suggested by second and first round active learning, exhibit high C_{1C} (145.91 and 145.01 mAh/g) and high ρ_{30kN} (2.599 and 2.588 g/cm3), outperforming all 80 synthesized LFP samples. Besides, by comparing RR2 and RR1 with their APF-free control groups, it can be observed that APG have a positive promoting effect on both ρ_{30kN} and C_{1C}. Further, the feature importance analysis also reveals that T_s is the most critical reaction condition that has significant impact on the two target properties. The framework highlighted the positive impact of APG on both properties, and feature importance analysis confirmed the critical role of T_s as the most influential reaction condition.

3. Chapter 6 introduced a combined deep learning model, LSTM and MPCA, to improve fault detection and real-time monitoring in traditional chemical industrial processes. Trained initially on simulated data, the framework was validated using a real industrial dataset, achieving R-squared values of 0.9204 and 0.9645 for predicting variable and PC values, respectively. The LSTM-MPCA method demonstrated significant potential for application in various industrial batch processes, requiring only historical normal data for training.

In summary, this thesis presents a comprehensive workflow for applying data-driven techniques to advance energy material design and discovery, as well as to enhance energy efficiency and safety in conventional chemical processes. The proposed strategies span from micro to macro levels, incorporating stages such as data collection, feature engineering, ML model selection and training, and innovative applications.

7.3 Future Work

1. The results of Chapter 4 are expected to provide a useful dataset for experimentalists to further examine the predicted ORR activity and for data scientists to develop ML models for ORR performance predictions. While only one of the discovered 5 potential candidates was experimentally examined, further experiments would be necessary to eventually validate the suggested material. Besides, the proposed data-driven workflow only considered binary Pt-based alloys, while more complex material systems are worthy of being further explored. Additionally, this study may aid in the exploration of catalysts for other electrocatalytic processes, such as water electrolysis.

2. In Chapter 5, while the synthesized sample recommended by developed active learning show superior performance on $\rho_{30\mathrm{kN}}$ and $C_{1\mathrm{C}}$, more properties such as the cycle number, constant current charge ratio and initial Coulombic efficiency should be focused as well. Besides, appropriate descriptors that can transfer all chemical information of additives including APG, Glucose and PEG are yet to be further explored. Due to the nature of data generation of this study, only one sample can be collected, leading to

the limited size of dataset and potential system errors of data. High-throughput experimentation is worthy to be designed for the future study.

3. Despite the progress made in Chapter 6, opportunities for further research and enhancement remain. These include using specific decomposition techniques tailored to dataset characteristics for more efficient and accurate fault detection and incorporating advanced machine learning techniques for fault diagnosis and prevention. The adaptability and generalizability of the LSTM-MPCA framework in monitoring batch steam boiler processes, when integrated with advanced data-driven techniques, suggest its potential for broader application in fault diagnosis across various scenarios.

Bibliography

ABIODUN, O. I., JANTAN, A., OMOLARA, A. E., DADA, K. V., MOHAMED, N. A. & ARSHAD, H. 2018. State-of-the-art in artificial neural network applications: A survey. *Heliyon,* 4.

ADDEL-GELIEL, M., ZAKZOUK, S. & SENGABY, M. E. Application of model based fault detection for an industrial boiler. 2012 20th Mediterranean Conference on Control & Automation (MED), 3-6 July 2012 2012. 98-103.

AHMAD, Z. & VISWANATHAN, V. 2016. Quantification of uncertainty in first-principles predicted mechanical properties of solids: Application to solid ion conductors. *Physical Review B,* 94**,** 064105.

AHMAD, Z. & VISWANATHAN, V. 2017a. Role of anisotropy in determining stability of electrodeposition at solid-solid interfaces. *Physical Review Materials,* 1**,** 055403.

AHMAD, Z. & VISWANATHAN, V. 2017b. Stability of Electrodeposition at Solid-Solid Interfaces and Implications for Metal Anodes. *Physical Review Letters,* 119**,** 056003.

AHMAD, Z., XIE, T., MAHESHWARI, C., GROSSMAN, J. C. & VISWANATHAN, V. 2018. Machine Learning Enabled Computational Screening of Inorganic Solid Electrolytes for Suppression of Dendrite Formation in Lithium Metal Anodes. *ACS Central Science,* 4**,** 996-1006.

AMSLER, M., HEGDE, V. I., JACOBSEN, S. D. & WOLVERTON, C. 2018. Exploring the High-Pressure Materials Genome. *Physical Review X,* 8**,** 041021.

ANDERSEN, M., LEVCHENKO, S. V., SCHEFFLER, M. & REUTER, K. 2019. Beyond Scaling Relations for the Description of Catalytic Materials. *ACS Catalysis,* 9**,** 2752-2759.

ANDERSEN, M., MEDFORD, A. J., NØRSKOV, J. K. & REUTER, K. 2017. Scaling-Relation-Based Analysis of Bifunctional Catalysis: The Case for Homogeneous Bimetallic Alloys. *ACS Catalysis,* 7**,** 3960-3967.

ANDERSEN, M. & REUTER, K. 2021. Adsorption Enthalpies for Catalysis Modeling through Machine-Learned Descriptors. *Accounts of Chemical Research,* 54**,** 2741-2749.

ANDERSON, T. W. 2003. *An Introduction to Multivariate Statistical Analysis, 3rd Edition*, A JOHN WILEY & SONS, INC., PUBLICATION

ARMAND, M., GAUTHIER, M., MAGNAN, J.-F. & RAVET, N. 2009. Method for synthesis of carbon-coated redox materials with controlled size. Google Patents.

ARTRITH, N., LIN, Z. & CHEN, J. G. 2020. Predicting the Activity and Selectivity of Bimetallic Metal Catalysts for Ethanol Reforming using Machine Learning. *ACS Catalysis,* 10**,** 9438-9444.

ATTARIAN SHANDIZ, M. & GAUVIN, R. 2016. Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries. *Computational Materials Science,* 117**,** 270-278.

BACK, S., TRAN, K. & ULISSI, Z. W. 2019. Toward a Design of Active Oxygen Evolution Catalysts: Insights from Automated Density Functional Theory Calculations and Machine Learning. *ACS Catalysis,* 9**,** 7651-7659.

BAERLOCHER, C. 2008. Database of zeolite structures. *http://www. iza-structure. org/databases/.*

BAGGER, A., JU, W., VARELA, A. S., STRASSER, P. & ROSSMEISL, J. 2017. Electrochemical CO2 Reduction: A Classification Problem. *Chemphyschem,* 18**,** 3266-3273.

BAI, Y., WILBRAHAM, L., SLATER, B. J., ZWIJNENBURG, M. A., SPRICK, R. S. & COOPER, A. I. 2019. Accelerated Discovery of Organic Polymer Photocatalysts for Hydrogen Evolution from Water through the Integration of Experiment and Theory. *J Am Chem Soc,* 141**,** 9063-9071.

BAJUSZ, D., RÁCZ, A. & HÉBERGER, K. 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics,* 7**,** 20.

BARNARD, A. S. 2020. Best Practice Leads to the Best Materials Informatics. *Matter,* 3**,** 22-23.

BARTHELMY, D. 2007. Mineralogy database. *http://webmineral. com/.*

BATCHELOR, T. A. A., PEDERSEN, J. K., WINTHER, S. H., CASTELLI, I. E., JACOBSEN, K. W. & ROSSMEISL, J. 2019. High-Entropy Alloys as a Discovery Platform for Electrocatalysis. *Joule,* 3**,** 834-845.

BECKER, C. A., TAVAZZA, F., TRAUTT, Z. T. & DE MACEDO, R. A. B. 2013. Considerations for choosing and using force fields and interatomic potentials in materials science and engineering. *Current Opinion in Solid State and Materials Science,* 17**,** 277-283.

BEERMANN, V., GOCYLA, M., KÜHL, S., PADGETT, E., SCHMIES, H., GOERLIN, M., ERINI, N., SHVIRO, M., HEGGEN, M., DUNIN-BORKOWSKI, R. E., MULLER, D. A. & STRASSER, P. 2017. Tuning the Electrocatalytic Oxygen Reduction Reaction Activity and Stability of Shape-Controlled Pt–Ni Nanoparticles by Thermal Annealing − Elucidating the Surface Atomic Structural and Compositional Changes. *Journal of the American Chemical Society,* 139**,** 16536-16547.

BEHLER, J. 2011. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics,* 134**,** 074106.

BEHLER, J. 2014. Representing potential energy surfaces by high-dimensional neural network potentials. *Journal of Physics: Condensed Matter,* 26**,** 183001.

BEHLER, J. & PARRINELLO, M. 2007. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical Review Letters,* 98**,** 146401.

BELSKY, A., HELLENBRANDT, M., KAREN, V. L. & LUKSCH, P. 2002. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica Section B: Structural Science,* 58**,** 364-369.

BERGERHOFF, G., HUNDT, R., SIEVERS, R. & BROWN, I. D. 1983. The Inorganic Crystal-Structure Data-Base. *Journal of Chemical Information and Computer Sciences,* 23**,** 66-69.

BHATT, M. D. & O'DWYER, C. 2015. Recent progress in theoretical and computational investigations of Li-ion battery materials and electrolytes. *Physical Chemistry Chemical Physics,* 17**,** 4799-4844.

BIGLIA, A., COMBA, L., FABRIZIO, E., GAY, P. & RICAUDA AIMONINO, D. 2017. Steam batch thermal processes in unsteady state conditions: Modelling and application to a case study in the food industry. *Applied Thermal Engineering,* 118**,** 638-651.

BILLAUD, J., CLÉMENT, R. J., ARMSTRONG, A. R., CANALES-VÁZQUEZ, J., ROZIER, P., GREY, C. P. & BRUCE, P. G. 2014. β-NaMnO2: A High-Performance Cathode for Sodium-Ion Batteries. *Journal of the American Chemical Society,* 136**,** 17243-17248.

BING, Y., LIU, H., ZHANG, L., GHOSH, D. & ZHANG, J. 2010. Nanostructured Pt-alloy electrocatalysts for PEM fuel cell oxygen reduction reaction. *Chemical Society Reviews,* 39**,** 2184-2202.

BLÖCHL, P. E. 1994. Projector augmented-wave method. *Physical Review B,* 50**,** 17953-17979.

BORYSOV, S. S., GEILHUFE, R. M. & BALATSKY, A. V. 2017. Organic materials database: An open-access online database for data mining. *PloS one,* 12**,** e0171501.

BROYDEN, C. G. 1970. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics,* 6**,** 76-90.

CAI, J., CHU, X., XU, K., LI, H. & WEI, J. 2020. Machine learning-driven new material discovery. *Nanoscale Advances,* 2**,** 3115-3130.

CENCER, M. M., MOORE, J. S. & ASSARY, R. S. 2021. Machine learning for polymeric materials: an introduction. *Polymer International,* n/a.

CHEN, A., ZHANG, X. & ZHOU, Z. 2020a. Machine learning: Accelerating materials development for energy storage and conversion. *InfoMat,* 2**,** 553-576.

CHEN, B. W. J., XU, L. & MAVRIKAKIS, M. 2021. Computational Methods in Heterogeneous Catalysis. *Chemical Reviews,* 121**,** 1007-1048.

CHEN, C., LIU, G. B., WANG, Y., LI, J. L. & LIU, H. 2013. Preparation and electrochemical properties of LiFePO4/C nanocomposite using FePO4·2H2O nanoparticles by introduction of Fe3(PO4)2·8H2O at low cost. *Electrochimica Acta,* 113**,** 464-469.

CHEN, C., YE, W., ZUO, Y., ZHENG, C. & ONG, S. P. 2019a. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials,* 31**,** 3564-3572.

CHEN, C., ZUO, Y., YE, W., LI, X., DENG, Z. & ONG, S. P. 2020b. A Critical Review of Machine Learning of Energy Materials. *Advanced Energy Materials,* 10**,** 1903242.

CHEN, J. & LIU, K.-C. 2002. On-line batch process monitoring using dynamic PCA and dynamic PLS models. *Chemical Engineering Science,* 57**,** 63-75.

CHEN, L., TIAN, Y., HU, X., YAO, S., LU, Z., CHEN, S., ZHANG, X. & ZHOU, Z. 2022. A Universal Machine Learning Framework for Electrocatalyst Innovation: A Case Study of Discovering Alloys for Hydrogen Evolution Reaction. *Advanced Functional Materials,* 32**,** 2208418.

CHEN, L., TRAN, H., BATRA, R., KIM, C. & RAMPRASAD, R. 2019b. Machine learning models for the lattice thermal conductivity prediction of inorganic materials. *Computational Materials Science,* 170**,** 109155.

CHEN, W., PÖHLS, J.-H., HAUTIER, G., BROBERG, D., BAJAJ, S., AYDEMIR, U., GIBBS, Z. M., ZHU, H., ASTA, M. & SNYDER, G. J. 2016. Understanding thermoelectric properties from high-throughput calculations: trends, insights, and comparisons with experiment. *Journal of Materials Chemistry C,* 4**,** 4414-4426.

CHEN, X., LI, H.-R., SHEN, X. & ZHANG, Q. 2018a. The Origin of the Reduced Reductive Stability of Ion–Solvent Complexes on Alkali and Alkaline Earth Metal Anodes. *Angewandte Chemie International Edition,* 57**,** 16643-16647.

CHEN, X., SHEN, X., LI, B., PENG, H.-J., CHENG, X.-B., LI, B.-Q., ZHANG, X.-Q., HUANG, J.-Q. & ZHANG, Q. 2018b. Ion–Solvent Complexes Promote Gas Evolution from Electrolytes on a Sodium Metal Anode. *Angewandte Chemie International Edition,* 57**,** 734-737.

CHEN, X. & SUN, L. 2022. Bayesian Temporal Factorization for Multidimensional Time Series Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 44**,** 4659-4673.

CHEN, Z., CHRISTENSEN, L. & DAHN, J. R. 2003. Large-volume-change electrodes for Li-ion batteries of amorphous alloy particles held by elastomeric tethers. *Electrochemistry Communications,* 5**,** 919-923.

CHENG, L., ASSARY, R. S., QU, X., JAIN, A., ONG, S. P., RAJPUT, N. N., PERSSON, K. & CURTISS, L. A. 2015a. Accelerating Electrolyte Discovery for Energy Storage with High-Throughput Screening. *The Journal of Physical Chemistry Letters,* 6**,** 283-291.

CHENG, L., ASSARY, R. S., QU, X., JAIN, A., ONG, S. P., RAJPUT, N. N., PERSSON, K. & CURTISS, L. A. 2015b. Accelerating Electrolyte Discovery for Energy Storage with High-Throughput Screening. *J Phys Chem Lett,* 6**,** 283-91.

CHOUDHARY, K., GARRITY, K. F., REID, A. C. E., DECOST, B., BIACCHI, A. J., WALKER, A. H. R., TRAUTT, Z., HATTRICK-SIMPERS, J., KUSNE, A. G., CENTRONE, A., DAVYDOV, A., JIANG, J., PACHTER, R., CHEON, G., REED, E., AGRAWAL, A., QIAN, X. F., SHARMA, V., ZHUANG, H. L., KALININ, S. V., SUMPTER, B. G., PILANIA, G., ACAR, P., MANDAL, S., HAULE, K., VANDERBILT, D., RABE, K. & TAVAZZA, F. 2020a. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *Npj Computational Materials,* 6.

CHOUDHARY, K., GARRITY, K. F. & TAVAZZA, F. 2020b. Data-driven discovery of 3D and 2D thermoelectric materials. *Journal of Physics: Condensed Matter,* 32**,** 475501.

CHUNG, D. Y., JUN, S. W., YOON, G., KWON, S. G., SHIN, D. Y., SEO, P., YOO, J. M., SHIN, H., CHUNG, Y.-H., KIM, H., MUN, B. S., LEE, K.-S., LEE, N.-S., YOO, S. J., LIM, D.-H., KANG, K., SUNG, Y.-E. & HYEON, T. 2015. Highly Durable and Active PtFe Nanocatalyst for Electrochemical Oxygen Reduction Reaction. *Journal of the American Chemical Society,* 137**,** 15478-15485.

COLE, J. C., WIGGIN, S. & STANZIONE, F. 2019. New insights and innovation from a million crystal structures in the Cambridge Structural Database. *Struct Dyn,* 6**,** 054301.

COLEY, C. W., GREEN, W. H. & JENSEN, K. F. 2019. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. *J Chem Inf Model,* 59**,** 2529-2537.

CONNOR, L. E., VASSILEIOU, A. D., HALBERT, G. W., JOHNSTON, B. F. & OSWALD, I. D. H. 2019. Structural investigation and compression of a co-crystal of indomethacin and saccharin. *CrystEngComm,* 21**,** 4465-4472.

COURT-CASTAGNET, R., KAPS, C., CROS, C. & HAGENMULLER, P. 1993. Ionic conductivity-enhancement of LiCl by homogeneous and heterogeneous dopings. *Solid State Ionics,* 61**,** 327-334.

CROCE, F., D' EPIFANIO, A., HASSOUN, J., DEPTULA, A., OLCZAC, T. & SCROSATI, B. 2002. A Novel Concept for the Synthesis of an Improved

LiFePO4 Lithium Battery Cathode. *Electrochemical and Solid-State Letters,* 5**,** A47.

CURTAROLO, S., HART, G. L. W., NARDELLI, M. B., MINGO, N., SANVITO, S. & LEVY, O. 2013. The high-throughput highway to computational materials design. *Nature Materials,* 12**,** 191-201.

CURTAROLO, S., SETYAWAN, W., WANG, S., XUE, J., YANG, K., TAYLOR, R. H., NELSON, L. J., HART, G. L., SANVITO, S. & BUONGIORNO-NARDELLI, M. 2012. AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science,* 58**,** 227-235.

D'AVEZAC, M., LUO, J. W., CHANIER, T. & ZUNGER, A. 2012. Genetic-algorithm discovery of a direct-gap and optically allowed superstructure from indirect-gap Si and Ge semiconductors. *Phys Rev Lett,* 108**,** 027401.

DAVIES, D. W., BUTLER, K. T. & WALSH, A. 2019. Data-Driven Discovery of Photoactive Quaternary Oxides Using First-Principles Machine Learning. *Chemistry of Materials,* 31**,** 7221-7230.

DE JONG, M., CHEN, W., ANGSTEN, T., JAIN, A., NOTESTINE, R., GAMST, A., SLUITER, M., KRISHNA ANDE, C., VAN DER ZWAAG, S., PLATA, J. J., TOHER, C., CURTAROLO, S., CEDER, G., PERSSON, K. A. & ASTA, M. 2015a. Charting the complete elastic properties of inorganic crystalline compounds. *Sci Data,* 2**,** 150009.

DE JONG, M., CHEN, W., GEERLINGS, H., ASTA, M. & PERSSON, K. A. 2015b. A database to enable discovery and design of piezoelectric materials. *Sci Data,* 2**,** 150053.

DE PABLO, J. J., JONES, B., KOVACS, C. L., OZOLINS, V. & RAMIREZ, A. P. 2014. The Materials Genome Initiative, the interplay of experiment, theory and computation. *Current Opinion in Solid State and Materials Science,* 18**,** 99-117.

DENG, X., JIANG, P., PENG, X. & MI, C. 2019. An Intelligent Outlier Detection Method With One Class Support Tucker Machine and Genetic Algorithm Toward Big Sensor Data in Internet of Things. *IEEE Transactions on Industrial Electronics,* 66**,** 4672-4683.

DENG, Z., WANG, Z., CHU, I.-H., LUO, J. & ONG, S. P. 2015. Elastic Properties of Alkali Superionic Conductor Electrolytes from First Principles Calculations. *Journal of The Electrochemical Society,* 163**,** A67-A74.

DESHMUKH, A. A., KUTHE, S. A. & PALIKUNDWAR, U. A. 2018. Understanding the effect of compositions on electronegativity, atomic radius and thermal stability of Mg-Ni-Y amorphous alloy. *AIP Conference Proceedings,* 1953**,** 090016.

DING, R., CHEN, Y., CHEN, P., WANG, R., WANG, J., DING, Y., YIN, W., LIU, Y., LI, J. & LIU, J. 2021. Machine Learning-Guided Discovery of Underlying Decisive Factors and New Mechanisms for the Design of Nonprecious Metal Electrocatalysts. *ACS Catalysis,* 11**,** 9798-9808.

DMELLO, R., MILSHTEIN, J. D., BRUSHETT, F. R. & SMITH, K. C. 2016. Cost-driven materials selection criteria for redox flow battery electrolytes. *Journal of Power Sources,* 330**,** 261-272.

DONDAPATI, J. S. & CHEN, A. 2020. Quantitative structure–property relationship of the photoelectrochemical oxidation of phenolic pollutants at modified nanoporous titanium oxide using supervised machine learning. *Physical Chemistry Chemical Physics,* 22**,** 8878-8888.

DOWNS, R. T. & HALL-WALLACE, M. 2003. The American Mineralogist crystal structure database. *American Mineralogist,* 88**,** 247-250.

DRAXL, C. & SCHEFFLER, M. 2018. NOMAD: The FAIR concept for big data-driven materials science. *MRS Bulletin,* 43**,** 676-682.

DRAXL, C. & SCHEFFLER, M. 2020. Big Data-Driven Materials Science and Its FAIR Data Infrastructure. *In:* ANDREONI, W. & YIP, S. (eds.) *Handbook of Materials Modeling: Methods: Theory and Modeling.* Cham: Springer International Publishing.

DUDIY, S. V. & ZUNGER, A. 2006a. Searching for Alloy Configurations with Target Physical Properties: Impurity Design via a Genetic Algorithm Inverse Band Structure Approach. *Physical Review Letters,* 97**,** 046401.

DUDIY, S. V. & ZUNGER, A. 2006b. Searching for alloy configurations with target physical properties: impurity design via a genetic algorithm inverse band structure approach. *Phys Rev Lett,* 97**,** 046401.

EFRON, B. & TIBSHIRANI, R. 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science,* 1**,** 54-75.

EFRON, B. & TIBSHIRANI, R. 1997. Improvements on Cross-Validation: The 632+ Bootstrap Method. *Journal of the American Statistical Association,* 92**,** 548-560.

ELTON, D. C., BOUKOUVALAS, Z., FUGE, M. D. & CHUNG, P. W. 2019. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering,* 4**,** 828-849.

EMERY, A. A., SAAL, J. E., KIRKLIN, S., HEGDE, V. I. & WOLVERTON, C. 2016. High-throughput computational screening of perovskites for thermochemical water splitting applications. *Chemistry of Materials,* 28**,** 5621-5634.

FABER, J. & FAWCETT, T. 2002. The powder diffraction file: present and future. *Acta Crystallographica Section B: Structural Science,* 58**,** 325-332.

FAN, C., WEN, P., LI, G., LI, G., GU, J., LI, Q. & LI, B. 2022. Facile synthesis of Pt5La nanoalloys as the enhanced electrocatalysts for oxygen reduction reaction and methanol oxidation reaction. *Journal of Alloys and Compounds,* 894**,** 161892.

FATHINIA, M., KHATAEE, A., ABER, S. & NASERI, A. 2016. Development of kinetic models for photocatalytic ozonation of phenazopyridine on TiO2 nanoparticles thin film in a mixed semi-batch photoreactor. *Applied Catalysis B: Environmental,* 184**,** 270-284.

FEI, H., PENG, Z., YANG, Y., LI, L., RAJI, A.-R. O., SAMUEL, E. L. G. & TOUR, J. M. 2014. LiFePO4 nanoparticles encapsulated in graphene nanoshells for high-performance lithium-ion battery cathodes. *Chemical Communications,* 50**,** 7117-7119.

FISCHER, C. C., TIBBETTS, K. J., MORGAN, D. & CEDER, G. 2006. Predicting crystal structure by merging data mining with quantum mechanics. *Nat Mater,* 5**,** 641-6.

FLETCHER, R. 1970. A new approach to variable metric algorithms. *The Computer Journal,* 13**,** 317-322.

FREEZE, J. G., KELLY, H. R. & BATISTA, V. S. 2019. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chemical Reviews,* 119**,** 6595-6612.

FREY, N. C., AKINWANDE, D., JARIWALA, D. & SHENOY, V. B. 2020. Machine Learning-Enabled Design of Point Defects in 2D Materials for Quantum and Neuromorphic Information Processing. *ACS Nano,* 14**,** 13406-13417.

FRIEDERICH, P., DOS PASSOS GOMES, G., DE BIN, R., ASPURU-GUZIK, A. & BALCELLS, D. 2020. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chemical Science,* 11**,** 4584-4601.

FUNG, V., HU, G., GANESH, P. & SUMPTER, B. G. 2021. Machine learned features from density of states for accurate adsorption energy prediction. *Nature Communications,* 12**,** 88.

G, S., KP, S. & R, V. 2018. Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. *Procedia Computer Science,* 132**,** 1253-1262.

GABRIELSON, S. W. 2018. SciFinder. *Journal of the Medical Library Association: JMLA,* 106**,** 588.

GAULTOIS, M. W., SPARKS, T. D., BORG, C. K., SESHADRI, R., BONIFICIO, W. D. & CLARKE, D. R. 2013. Data-driven review of thermoelectric materials: performance and resource considerations. *Chemistry of Materials,* 25**,** 2911-2920.

GE, L., YUAN, H., MIN, Y., LI, L., CHEN, S., XU, L. & GODDARD, W. A. 2020a. Predicted Optimal Bifunctional Electrocatalysts for the Hydrogen Evolution Reaction and the Oxygen Evolution Reaction Using Chalcogenide Heterostructures Based on Machine Learning Analysis of in Silico Quantum Mechanics Based High Throughput Screening. *The Journal of Physical Chemistry Letters,* 11**,** 869-876.

GE, L., YUAN, H., MIN, Y., LI, L., CHEN, S., XU, L. & GODDARD, W. A., 3RD 2020b. Predicted Optimal Bifunctional Electrocatalysts for the Hydrogen Evolution Reaction and the Oxygen Evolution Reaction Using Chalcogenide Heterostructures Based on Machine Learning Analysis of in Silico Quantum Mechanics Based High Throughput Screening. *J Phys Chem Lett,* 11**,** 869-876.

GE, Z., SONG, Z. & GAO, F. 2013. Review of Recent Research on Data-Based Process Monitoring. *Industrial & Engineering Chemistry Research,* 52**,** 3543-3562.

GENG, W. T. & OHNO, T. 2013. Carbon Coating of LiFePO4 Can Be Strengthened by Sc and Ti. *The Journal of Physical Chemistry C,* 117**,** 276-279.

GERS, F. A. & SCHMIDHUBER, J. 2001. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks,* 12**,** 1333-1340.

GHAHRAMANI, Z. 2015. Probabilistic machine learning and artificial intelligence. *Nature,* 521**,** 452-459.

GHIRINGHELLI, L. M., VYBIRAL, J., LEVCHENKO, S. V., DRAXL, C. & SCHEFFLER, M. 2015. Big Data of Materials Science: Critical Role of the Descriptor. *Physical Review Letters,* 114**,** 105503.

GOLDSMITH, B. R., ESTERHUIZEN, J., LIU, J.-X., BARTEL, C. J. & SUTTON, C. 2018. Machine learning for heterogeneous catalyst design and discovery. *AIChE Journal,* 64**,** 2311-2323.

GONG, L., LIU, J., LI, Y., WANG, X., LUO, E., JIN, Z., GE, J., LIU, C. & XING, W. 2022. An ultralow-loading platinum alloy efficient ORR electrocatalyst based on the surface-contracted hollow structure. *Chemical Engineering Journal,* 428**,** 131569.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. & BENGIO, Y. 2014.

Generative adversarial nets. *Advances in Neural Information Processing Systems,* 27**,** 2672-2680.

GOODMAN, J. 2009. Computer software review: Reaxys. ACS Publications.

GORAI, P., GAO, D., ORTIZ, B., MILLER, S., BARNETT, S. A., MASON, T., LV, Q., STEVANOVIĆ, V. & TOBERER, E. S. 2016. TE Design Lab: A virtual laboratory for thermoelectric material design. *Computational Materials Science,* 112**,** 368-376.

GRAZULIS, S., CHATEIGNER, D., DOWNS, R. T., YOKOCHI, A. F., QUIROS, M., LUTTEROTTI, L., MANAKOVA, E., BUTKUS, J., MOECK, P. & LE BAIL, A. 2009. Crystallography Open Database - an open-access collection of crystal structures. *J Appl Crystallogr,* 42**,** 726-729.

GREELEY, J., STEPHENS, I. E. L., BONDARENKO, A. S., JOHANSSON, T. P., HANSEN, H. A., JARAMILLO, T. F., ROSSMEISL, J., CHORKENDORFF, I. & NØRSKOV, J. K. 2009. Alloys of platinum and early transition metals as oxygen reduction electrocatalysts. *Nature Chemistry,* 1**,** 552-556.

GREEN, M. L., CHOI, C. L., HATTRICK-SIMPERS, J. R., JOSHI, A. M., TAKEUCHI, I., BARRON, S. C., CAMPO, E., CHIANG, T., EMPEDOCLES, S., GREGOIRE, J. M., KUSNE, A. G., MARTIN, J., MEHTA, A., PERSSON, K., TRAUTT, Z., VAN DUREN, J. & ZAKUTAYEV, A. 2017. Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Applied Physics Reviews,* 4**,** 011105.

GROENENBOOM, M. C., ANDERSON, R. M., WOLLMERSHAUSER, J. A., HORTON, D. J., POLICASTRO, S. A. & KEITH, J. A. 2020. Combined Neural Network Potential and Density Functional Theory Study of TiAl2O5 Surface Morphology and Oxygen Reduction Reaction Overpotentials. *The Journal of Physical Chemistry C,* 124**,** 15171-15179.

GROOM, C. R., BRUNO, I. J., LIGHTFOOT, M. P. & WARD, S. C. 2016a. The Cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials,* 72**,** 171-179.

GROOM, C. R., BRUNO, I. J., LIGHTFOOT, M. P. & WARD, S. C. 2016b. The Cambridge Structural Database. *Acta Crystallogr B Struct Sci Cryst Eng Mater,* 72**,** 171-9.

GU, G. H., NOH, J., KIM, I. & JUNG, Y. 2019. Machine learning for renewable energy materials. *Journal of Materials Chemistry A,* 7**,** 17096-17117.

GUO, X., MAO, D., LU, G., WANG, S. & WU, G. 2011. The influence of La doping on the catalytic behavior of Cu/ZrO2 for methanol synthesis from CO2 hydrogenation. *Journal of Molecular Catalysis A: Chemical,* 345**,** 60-68.

HAASTRUP, S., STRANGE, M., PANDEY, M., DEILMANN, T., SCHMIDT, P. S., HINSCHE, N. F., GJERDING, M. N., TORELLI, D., LARSEN, P. M. & RIIS-JENSEN, A. C. 2018. The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals. *2D Materials,* 5**,** 042002.

HACHMANN, J., OLIVARES-AMAYA, R., ATAHAN-EVRENK, S., AMADOR-BEDOLLA, C., SANCHEZ-CARRERA, R. S., GOLD-PARKER, A., VOGT, L., BROCKWAY, A. M. & ASPURU-GUZIK, A. 2011. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *Journal of Physical Chemistry Letters,* 2**,** 2241-2251.

HALE, L. M., TRAUTT, Z. T. & BECKER, C. A. 2018. Evaluating variability with atomistic simulations: the effect of potential and calculation methodology on the modeling of lattice and elastic constants. *Modelling and Simulation in Materials Science and Engineering,* 26**,** 055003.

HALLS, M. D. & TASAKI, K. 2010. High-throughput quantum chemistry and virtual screening for lithium ion battery electrolyte additives. *Journal of Power Sources,* 195**,** 1472-1478.

HAMMER, B. & NORSKOV, J. K. 1995. Why gold is the noblest of all the metals. *Nature,* 376**,** 238-240.

HAMMER, B. & NØRSKOV, J. K. 2000. Theoretical surface science and catalysis—calculations and concepts. *Advances in catalysis,* 45**,** 71-129.

HAUTIER, G., FISCHER, C., EHRLACHER, V., JAIN, A. & CEDER, G. 2011a. Data mined ionic substitutions for the discovery of new compounds. *Inorg Chem,* 50**,** 656-63.

HAUTIER, G., FISCHER, C., EHRLACHER, V., JAIN, A. & CEDER, G. 2011b. Data mined ionic substitutions for the discovery of new compounds. *Inorganic chemistry,* 50**,** 656-663.

HAUTIER, G., FISCHER, C. C., JAIN, A., MUELLER, T. & CEDER, G. 2010. Finding Nature's Missing Ternary Oxide Compounds Using Machine

Learning and Density Functional Theory. *Chemistry of Materials,* 22**,** 3762-3767.

HAUTIER, G., JAIN, A. & ONG, S. P. 2012. From the computer to the laboratory: materials discovery and design using first-principles calculations. *Journal of Materials Science,* 47**,** 7317-7340.

HAWKINS, D. M., BASAK, S. C. & MILLS, D. 2003. Assessing Model Fit by Cross-Validation. *Journal of Chemical Information and Computer Sciences,* 43**,** 579-586.

HEGDE, V. I., AYKOL, M., KIRKLIN, S. & WOLVERTON, C. 2020a. The phase stability network of all inorganic materials. *Science Advances,* 6**,** eaay5606.

HEGDE, V. I., BORG, C. K., DEL ROSARIO, Z., KIM, Y., HUTCHINSON, M., ANTONO, E., LING, J., SAXE, P., SAAL, J. E. & MEREDIG, B. 2020b. Reproducibility in high-throughput density functional theory: a comparison of AFLOW, Materials Project, and OQMD. *arXiv preprint arXiv:2007.01988*.

HILL, J., MANNODI-KANAKKITHODI, A., RAMPRASAD, R. & MEREDIG, B. 2018. Materials Data Infrastructure and Materials Informatics. *In:* SHIN, D. & SAAL, J. (eds.) *Computational Materials System Design.* Cham: Springer International Publishing.

HIRSCHFELD, L., SWANSON, K., YANG, K., BARZILAY, R. & COLEY, C. W. 2020. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *Journal of Chemical Information and Modeling,* 60**,** 3770-3780.

HJORTH LARSEN, A., JØRGEN MORTENSEN, J., BLOMQVIST, J., CASTELLI, I. E., CHRISTENSEN, R., DUŁAK, M., FRIIS, J., GROVES, M. N., HAMMER, B., HARGUS, C., HERMES, E. D., JENNINGS, P. C., BJERRE JENSEN, P., KERMODE, J., KITCHIN, J. R., LEONHARD KOLSBJERG, E., KUBAL, J., KAASBJERG, K., LYSGAARD, S., BERGMANN MARONSSON, J., MAXSON, T., OLSEN, T., PASTEWKA, L., PETERSON, A., ROSTGAARD, C., SCHIØTZ, J., SCHÜTT, O., STRANGE, M., THYGESEN, K. S., VEGGE, T., VILHELMSEN, L., WALTER, M., ZENG, Z. & JACOBSEN, K. W. 2017. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter,* 29**,** 273002.

HOAR, B. B., LU, S. & LIU, C. 2020a. Machine-Learning-Enabled Exploration of Morphology Influence on Wire-Array Electrodes for Electrochemical Nitrogen Fixation. *The Journal of Physical Chemistry Letters,* 11**,** 4625-4630.

HOAR, B. B., LU, S. & LIU, C. 2020b. Machine-Learning-Enabled Exploration of Morphology Influence on Wire-Array Electrodes for Electrochemical Nitrogen Fixation. *J Phys Chem Lett,* 11**,** 4625-4630.

HOCHREITER, S. & SCHMIDHUBER, J. 1997. Long Short-Term Memory. *Neural Comput.,* 9**,** 1735–1780.

HOU, Z., TAKAGIWA, Y., SHINOHARA, Y., XU, Y. & TSUDA, K. 2019. Machine-Learning-Assisted Development and Theoretical Consideration for the Al2Fe3Si3 Thermoelectric Material. *ACS Applied Materials & Interfaces,* 11**,** 11545-11554.

HSIEH, C.-T., CHEN, I. L., CHEN, W.-Y. & WANG, J.-P. 2012. Synthesis of iron phosphate powders by chemical precipitation route for high-power lithium iron phosphate cathodes. *Electrochimica Acta,* 83**,** 202-208.

HU, J., CAO, X., ZHAO, X., CHEN, W., LU, G. P., DAN, Y. & CHEN, Z. 2019. Catalytically Active Sites on Ni5P4 for Efficient Hydrogen Evolution Reaction From Atomic Scale Calculation. *Front Chem,* 7**,** 444.

HU, T., SONG, H., JIANG, T. & LI, S. 2020. Learning Representations of Inorganic Materials from Generative Adversarial Networks. *Symmetry,* 12.

HUANG, X., YAO, Y., LIANG, F. & DAI, Y. 2018. Concentration-controlled morphology of LiFePO4 crystals with an exposed (100) facet and their enhanced performance for use in lithium-ion batteries. *Journal of Alloys and Compounds,* 743**,** 763-772.

HUANG, Y., YU, C., CHEN, W., LIU, Y., LI, C., NIU, C., WANG, F. & JIA, Y. 2019. Band gap and band alignment prediction of nitride-based semiconductors using machine learning. *Journal of Materials Chemistry C,* 7**,** 3238-3245.

HUBER, S. P., ZOUPANOS, S., UHRIN, M., TALIRZ, L., KAHLE, L., HAUSELMANN, R., GRESCH, D., MULLER, T., YAKUTOVICH, A. V., ANDERSEN, C. W., RAMIREZ, F. F., ADORF, C. S., GARGIULO, F., KUMBHAR, S., PASSARO, E., JOHNSTON, C., MERKYS, A., CEPELLOTTI, A., MOUNET, N., MARZARI, N., KOZINSKY, B. &

PIZZI, G. 2020. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci Data,* 7**,** 300.

HUBERT, M., ROUSSEEUW, P. J. & VANDEN BRANDEN, K. 2005. ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics,* 47**,** 64-79.

HUMMELSHOJ, J. S., ABILD-PEDERSEN, F., STUDT, F., BLIGAARD, T. & NORSKOV, J. K. 2012. CatApp: a web application for surface chemistry and heterogeneous catalysis. *Angew Chem Int Ed Engl,* 51**,** 272-4.

HUSCH, T., YILMAZER, N. D., BALDUCCI, A. & KORTH, M. 2015. Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: computing infrastructure and collective properties. *Physical Chemistry Chemical Physics,* 17**,** 3394-3401.

HUSSAIN, J., JÓNSSON, H. & SKÚLASON, E. 2018. Calculations of Product Selectivity in Electrochemical CO2 Reduction. *ACS Catalysis,* 8**,** 5240-5249.

HWANG, S. J., YOO, S. J., JANG, S., LIM, T.-H., HONG, S. A. & KIM, S.-K. 2011. Ternary Pt−Fe−Co Alloy Electrocatalysts Prepared by Electrodeposition: Elucidating the Roles of Fe and Co in the Oxygen Reduction Reaction. *The Journal of Physical Chemistry C,* 115**,** 2483-2488.

IGARASHI, Y., NAGATA, K., KUWATANI, T., OMORI, T., NAKANISHI-OHNO, Y. & OKADA, M. 2016. Three levels of data-driven science. *Journal of Physics: Conference Series,* 699**,** 012001.

IGARASHI, Y., TAKENAKA, H., NAKANISHI-OHNO, Y., UEMURA, M., IKEDA, S. & OKADA, M. 2018. Exhaustive Search for Sparse Variable

Selection in Linear Regression. *Journal of the Physical Society of Japan,* 87**,** 044802.

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. *Physical Chemistry Chemical Physics,* 21**,** 26399-26405.

JABLONKA, K. M., ONGARI, D., MOOSAVI, S. M. & SMIT, B. 2020. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chemical Reviews,* 120**,** 8066-8129.

JACKSON, J. E. & MUDHOLKAR, G. S. 1979. Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics,* 21**,** 341-349.

JACOBSEN, K. W., STOLTZE, P. & NØRSKOV, J. K. 1996. A semi-empirical effective medium theory for metals and alloys. *Surface Science,* 366**,** 394-402.

JÄGER, M. O. J., MOROOKA, E. V., FEDERICI CANOVA, F., HIMANEN, L. & FOSTER, A. S. 2018. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials,* 4.

JAIN, A., HAUTIER, G., MOORE, C. J., ONG, S. P., FISCHER, C. C., MUELLER, T., PERSSON, K. A. & CEDER, G. 2011a. A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science,* 50**,** 2295-2310.

JAIN, A., HAUTIER, G., MOORE, C. J., PING ONG, S., FISCHER, C. C., MUELLER, T., PERSSON, K. A. & CEDER, G. 2011b. A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science,* 50**,** 2295-2310.

JAIN, A., HAUTIER, G., ONG, S. P., MOORE, C. J., FISCHER, C. C., PERSSON, K. A. & CEDER, G. 2011c. Formation enthalpies by mixing GGA and GGA plus U calculations. *Physical Review B,* 84**,** 045115.

JAIN, A., ONG, S. P., CHEN, W., MEDASANI, B., QU, X. H., KOCHER, M., BRAFMAN, M., PETRETTO, G., RIGNANESE, G. M., HAUTIER, G., GUNTER, D. & PERSSON, K. A. 2015. FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurrency and Computation-Practice & Experience,* 27**,** 5037-5059.

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 Materials,* 1**,** 011002.

JANET, J. P., DUAN, C., YANG, T., NANDY, A. & KULIK, H. J. 2019. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chemical Science,* 10**,** 7913-7922.

JAOUEN, F., PROIETTI, E., LEFÈVRE, M., CHENITZ, R., DODELET, J.-P., WU, G., CHUNG, H. T., JOHNSTON, C. M. & ZELENAY, P. 2011. Recent advances in non-precious metal catalysis for oxygen-reduction reaction in polymer electrolyte fuel cells. *Energy & Environmental Science,* 4**,** 114-130.

JIANG, J.-L., SU, X., ZHANG, H., ZHANG, X.-H. & YUAN, Y.-J. 2013. A Novel Approach to Active Compounds Identification Based on Support Vector

Regression Model and Mean Impact Value. *Chemical Biology & Drug Design,* 81, 650-657.

JIN, H., ZHANG, H., LI, J., WANG, T., WAN, L., GUO, H. & WEI, Y. 2020. Discovery of Novel Two-Dimensional Photovoltaic Materials Accelerated by Machine Learning. *The Journal of Physical Chemistry Letters,* 11, 3075-3081.

JIN, W., BARZILAY, R. & JAAKKOLA, T. 2018. Junction Tree Variational Autoencoder for Molecular Graph Generation. *In:* JENNIFER, D. & ANDREAS, K. (eds.) *Proceedings of the 35th International Conference on Machine Learning.* Proceedings of Machine Learning Research: PMLR.

JINNOUCHI, R. & ASAHI, R. 2017. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *The Journal of Physical Chemistry Letters,* 8, 4279-4283.

JOE QIN, S. 2003. Statistical process monitoring: basics and beyond. *Journal of Chemometrics,* 17, 480-502.

JOHNSON III, R. D. 1999. NIST 101. Computational chemistry comparison and benchmark database.

JOLLIFFE, I. 2011. Principal Component Analysis. *In:* LOVRIC, M. (ed.) *International Encyclopedia of Statistical Science.* Berlin, Heidelberg: Springer Berlin Heidelberg.

JOLLIFFE, I. T. & CADIMA, J. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* 374, 20150202.

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 Applied Materials & Interfaces,* 11**,** 18494-18503.

KABEKKODU, S. N., FABER, J. & FAWCETT, T. 2002. New Powder Diffraction File (PDF-4) in relational database format: advantages and data-mining capabilities. *Acta Crystallogr B,* 58**,** 333-7.

KADOMA, Y., KIM, J.-M., ABIKO, K., OHTSUKI, K., UI, K. & KUMAGAI, N. 2010. Optimization of electrochemical properties of LiFePO4/C prepared by an aqueous solution method using sucrose. *Electrochimica Acta,* 55**,** 1034-1041.

KANG, J., NOH, S. H., HWANG, J., CHUN, H., KIM, H. & HAN, B. 2018a. First-principles database driven computational neural network approach to the discovery of active ternary nanocatalysts for oxygen reduction reaction. *Physical Chemistry Chemical Physics,* 20**,** 24539-24544.

KANG, J., NOH, S. H., HWANG, J., CHUN, H., KIM, H. & HAN, B. 2018b. First-principles database driven computational neural network approach to the discovery of active ternary nanocatalysts for oxygen reduction reaction. *Phys Chem Chem Phys,* 20**,** 24539-24544.

KHORSHIDI, A. & PETERSON, A. A. 2016. Amp: A modular approach to machine learning in atomistic simulations. *Computer Physics Communications,* 207**,** 310-324.

KIM, E., HUANG, K., SAUNDERS, A., MCCALLUM, A., CEDER, G. & OLIVETTI, E. 2017. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials,* 29**,** 9436-9444.

KIM, J.-H. 2009. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis,* 53**,** 3735-3745.

KIM, M., FIRESTEIN, K. L., FERNANDO, J. F. S., XU, X., LIM, H., GOLBERG, D. V., NA, J., KIM, J., NARA, H., TANG, J. & YAMAUCHI, Y. 2022. Strategic design of Fe and N co-doped hierarchically porous carbon as superior ORR catalyst: from the perspective of nanoarchitectonics. *Chemical Science,* 13**,** 10836-10845.

KIM, S., CHEN, J., CHENG, T., GINDULYTE, A., HE, J., HE, S., LI, Q., SHOEMAKER, B. A., THIESSEN, P. A. & YU, B. 2019. PubChem 2019 update: improved access to chemical data. *Nucleic acids research,* 47**,** D1102-D1109.

KINGMA, D. P. & BA, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

KIRKLIN, S., SAAL, J. E., MEREDIG, B., THOMPSON, A., DOAK, J. W., AYKOL, M., RUHL, S. & WOLVERTON, C. 2015a. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *Npj Computational Materials,* 1**,** 1-15.

KIRKLIN, S., SAAL, J. E., MEREDIG, B., THOMPSON, A., DOAK, J. W., AYKOL, M., RÜHL, S. & WOLVERTON, C. 2015b. The Open Quantum

Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials,* 1.

KIRKLIN, S., SAAL, J. E., MEREDIG, B., THOMPSON, A., DOAK, J. W., AYKOL, M., RÜHL, S. & WOLVERTON, C. 2015c. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials,* 1**,** 15010.

KITCHIN, J. R. 2018. Machine learning in catalysis. *Nature Catalysis,* 1**,** 230-232.

KLIMECK, G., MCLENNAN, M., BROPHY, S. P., ADAMS III, G. B. & LUNDSTROM, M. S. 2008. nanohub. org: Advancing education and research in nanotechnology. *Computing in Science & Engineering,* 10**,** 17-23.

KONNO, T., KUROKAWA, H., NABESHIMA, F., SAKISHITA, Y., OGAWA, R., HOSAKO, I. & MAEDA, A. 2021. Deep learning model for finding new superconductors. *Physical Review B,* 103**,** 014509.

KORTH, M. 2014. Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: evaluation of electronic structure theory methods. *Physical Chemistry Chemical Physics,* 16**,** 7919-7926.

KOTSIANTIS;, S. B., KANELLOPOULOS;, D. & PINTELAS, P. E. 2007. Data Preprocessing for Supervised Leaning. *Zenodo*.

KOUCHI, M. & MOCHIMARU, M. 2005. AIST Research Information Database (http://riodb. ibase. aist. go. jp/riohomee. html). *H16PRO287*.

KRAMER, M. A. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal,* 37**,** 233-243.

KRESS-ROGERS, E. & BRIMELOW, C. 2000. *Knovel solvents-a properties database*, ChemTec Publishing.

KRÓL, J. & OCŁOŃ, P. 2018. Economic analysis of heat and electricity production in combined heat and power plant equipped with steam and water boilers and natural gas engines. *Energy Conversion and Management,* 176**,** 11-29.

KULKARNI, A., SIAHROSTAMI, S., PATEL, A. & NØRSKOV, J. K. 2018. Understanding Catalytic Activity Trends in the Oxygen Reduction Reaction. *Chemical Reviews,* 118**,** 2302-2312.

LANDIS, D. D., HUMMELSHOJ, J. S., NESTOROV, S., GREELEY, J., DULAK, M., BLIGAARD, T., NORSKOV, J. K. & JACOBSEN, K. W. 2012. The computational materials repository. *Computing in Science & Engineering,* 14**,** 51-57.

LARSEN, A. H., MORTENSEN, J. J., BLOMQVIST, J., CASTELLI, I. E., CHRISTENSEN, R., DUŁAK, M., FRIIS, J., GROVES, M. N., HAMMER, B. & HARGUS, C. 2017. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter,* 29**,** 273002.

LATIMER, K., DWARAKNATH, S., MATHEW, K., WINSTON, D. & PERSSON, K. A. 2018. Evaluation of thermodynamic equations of state across chemistry and structure in the materials project. *Npj Computational Materials,* 4**,** 1-7.

LECUN, Y., BENGIO, Y. & HINTON, G. 2015. Deep learning. *Nature,* 521**,** 436-444.

LEE, C.-H., KHAN, A., LUO, D., SANTOS, T. P., SHI, C., JANICEK, B. E., KANG, S., ZHU, W., SOBH, N. A., SCHLEIFE, A., CLARK, B. K. & HUANG, P. Y. 2020. Deep Learning Enabled Strain Mapping of Single-Atom Defects in Two-Dimensional Transition Metal Dichalcogenides with Sub-Picometer Precision. *Nano Letters,* 20**,** 3369-3377.

LEVER, J., KRZYWINSKI, M. & ALTMAN, N. 2016. Classification evaluation. *Nature Methods,* 13**,** 603-604.

LI, J., ALSUDAIRI, A., MA, Z.-F., MUKERJEE, S. & JIA, Q. 2017a. Asymmetric Volcano Trend in Oxygen Reduction Activity of Pt and Non-Pt Catalysts: In Situ Identification of the Site-Blocking Effect. *Journal of the American Chemical Society,* 139**,** 1384-1387.

LI, L., LI, X., WANG, Z., WU, L., ZHENG, J. & GUO, H. 2009. Stable cycle-life properties of Ti-doped LiFePO4 compounds synthesized by co-precipitation and normal temperature reduction method. *Journal of Physics and Chemistry of Solids,* 70**,** 238-242.

LI, W., ZHANG, F., PAN, L. & LI, Z. 2022a. Gas or electricity? Regional pathway selection under carbon neutrality target: A case study of industrial boilers. *Journal of Cleaner Production,* 349**,** 131313.

LI, X., LI, B., YANG, Z., CHEN, Z., GAO, W. & JIANG, Q. 2022b. A transferable machine-learning scheme from pure metals to alloys for predicting adsorption energies. *Journal of Materials Chemistry A,* 10**,** 872-880.

LI, Z., WANG, S., CHIN, W. S., ACHENIE, L. E. & XIN, H. 2017b. High-throughput screening of bimetallic catalysts enabled by machine learning. *Journal of Materials Chemistry A,* 5**,** 24131-24138.

LIU, B.-E. & YU, W. 2020. On-demand Direct Design of Polymeric Thermal Actuator by Machine Learning Algorithm. *Chinese Journal of Polymer Science,* 38**,** 908-914.

LIU, H., WANG, G. X., WEXLER, D., WANG, J. Z. & LIU, H. K. 2008. Electrochemical performance of LiFePO4 cathode material coated with ZrO2 nanolayer. *Electrochemistry Communications,* 10**,** 165-169.

LIU, M., HUANG, Y., LI, Z., TONG, B., LIU, Z., SUN, M., JIANG, F. & ZHANG, H. 2020a. The Applicability of LSTM-KNN Model for Real-Time Flood Forecasting in Different Climate Zones in China. *Water* [Online], 12.

LIU, Q.-B., LIAO, S.-J., SONG, H.-Y. & LIANG, Z.-X. 2012. High-performance LiFePO4/C materials: Effect of carbon source on microstructure and performance. *Journal of Power Sources,* 211**,** 52-58.

LIU, X., HAO, S., ZHENG, G., SU, Z., WANG, Y., WANG, Q., LEI, L., HE, Y. & ZHANG, X. 2022a. Ultrasmall Pt2Sr alloy nanoparticles as efficient bifunctional electrocatalysts for oxygen reduction and hydrogen evolution in acidic media. *Journal of Energy Chemistry,* 64**,** 315-322.

LIU, X., XIAO, J., PENG, H., HONG, X., CHAN, K. & NØRSKOV, J. K. 2017a. Understanding trends in electrochemical carbon dioxide reduction rates. *Nature Communications,* 8**,** 15438.

LIU, Y., GUO, B., ZOU, X., LI, Y. & SHI, S. 2020b. Machine learning assisted materials design and discovery for rechargeable batteries. *Energy Storage Materials,* 31**,** 434-450.

LIU, Y., ZHANG, H., HUANG, Z., WANG, Q., GUO, M., ZHAO, M., ZHANG, D., WANG, J., HE, P., LIU, X., TERRONES, M. & WANG, Y. 2022b. Understanding the influence of nanocarbon conducting modes on the rate performance of LiFePO4 cathodes in lithium-ion batteries. *Journal of Alloys and Compounds,* 905**,** 164205.

LIU, Y., ZHAO, T., JU, W. & SHI, S. 2017b. Materials discovery and design using machine learning. *Journal of Materiomics,* 3**,** 159-177.

LONG, T., FORTUNATO, N. M., ZHANG, Y., GUTFLEISCH, O. & ZHANG, H. 2021. An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature. *Materials Research Letters,* 9**,** 169-174.

LU, N., WANG, F., GAO, F. & WANG, S. 2006. Statistical modeling and online monitoring for batch processes. *Zidonghua Xuebao/Acta Automatica Sinica,* 32**,** 400.

LU, Q., ROSEN, J., ZHOU, Y., HUTCHINGS, G. S., KIMMEL, Y. C., CHEN, J. G. & JIAO, F. 2014. A selective and efficient electrocatalyst for carbon dioxide reduction. *Nature Communications,* 5**,** 3242.

LU, S., ZHOU, Q., OUYANG, Y., GUO, Y., LI, Q. & WANG, J. 2018. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nature Communications,* 9**,** 3405.

MA, X.-Y., LEWIS, J. P., YAN, Q.-B. & SU, G. 2019. Accelerated Discovery of Two-Dimensional Optoelectronic Octahedral Oxyhalides via High-Throughput Ab Initio Calculations and Machine Learning. *The Journal of Physical Chemistry Letters,* 10**,** 6734-6740.

MA, X., LI, Z., ACHENIE, L. E. & XIN, H. 2015a. Machine-learning-augmented chemisorption model for CO2 electroreduction catalyst screening. *The journal of physical chemistry letters,* 6**,** 3528-3533.

MA, X., LI, Z., ACHENIE, L. E. & XIN, H. 2015b. Machine-Learning-Augmented Chemisorption Model for CO2 Electroreduction Catalyst Screening. *J Phys Chem Lett,* 6**,** 3528-33.

MA, X., LI, Z., ACHENIE, L. E. K. & XIN, H. 2015c. Machine-Learning-Augmented Chemisorption Model for CO2 Electroreduction Catalyst Screening. *The Journal of Physical Chemistry Letters,* 6**,** 3528-3533.

MAATEN, L. V. D. & HINTON, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research,* 9**,** 2579-2605.

MACGREGOR, J. F. & KOURTI, T. 1995. Statistical process control of multivariate processes. *Control Engineering Practice,* 3**,** 403-414.

MAN, Y., YANG, Q., SHAO, J., WANG, G., BAI, L. & XUE, Y. 2022. Enhanced LSTM Model for Daily Runoff Prediction in the Upper Huai River Basin, China. *Engineering*.

MASOOD, H., TOE, C. Y., TEOH, W. Y., SETHU, V. & AMAL, R. 2019. Machine Learning for Accelerated Discovery of Solar Photocatalysts. *ACS Catalysis,* 9**,** 11774-11787.

MATHEW, K., MONTOYA, J. H., FAGHANINIA, A., DWARAKANATH, S., AYKOL, M., TANG, H. M., CHU, I. H., SMIDT, T., BOCKLUND, B., HORTON, M., DAGDELEN, J., WOOD, B., LIU, Z. K., NEATON, J., ONG, S. P., PERSSON, K. & JAIN, A. 2017. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Computational Materials Science,* 139**,** 140-152.

MATHEW, K., ZHENG, C., WINSTON, D., CHEN, C., DOZIER, A., REHR, J. J., ONG, S. P. & PERSSON, K. A. 2018. High-throughput computational X-ray absorption spectroscopy. *Sci Data,* 5**,** 180151.

MATWEB, L. 1996. MatWeb. *Material property data, Data base of materials data sheets*.

MAYER, J. M. 2011. Simple Marcus-Theory-Type Model for Hydrogen-Atom Transfer/Proton-Coupled Electron Transfer. *The Journal of Physical Chemistry Letters,* 2**,** 1481-1489.

MEDVEDEVA, A. E., PECHEN, L. S., MAKHONINA, E. V., RUMYANTSEV, A. M., KOSHTYAL, Y. M., PERVOV, V. S. & EREMENKO, I. L. 2019. Synthesis and Electrochemical Properties of Lithium-Ion Battery Cathode Materials Based on LiFePO4–LiMn2O4 and LiFePO4–LiNi0.82Co0.18O2 Composites. *Russian Journal of Inorganic Chemistry,* 64**,** 829-840.

MENG, Y. S. & ARROYO-DE DOMPABLO, M. E. 2009. First principles computational materials design for energy storage materials in lithium ion batteries. *Energy & Environmental Science,* 2**,** 589-609.

MENG, Y. S. & ARROYO-DE DOMPABLO, M. E. 2013. Recent Advances in First Principles Computational Research of Cathode Materials for Lithium-Ion Batteries. *Accounts of Chemical Research,* 46**,** 1171-1180.

MEREDIG, B. & WOLVERTON, C. 2014. Dissolving the Periodic Table in Cubic Zirconia: Data Mining to Discover Chemical Trends. *Chemistry of Materials,* 26**,** 1985-1991.

MOHOD, S. & RAUT, A. PLC SCADA Based Fault Detection System for Steam Boiler In Remote Plant. 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 5-6 July 2019 2019. 1007-1010.

MÖLLER, J. J., KÖRNER, W., KRUGEL, G., URBAN, D. F. & ELSÄSSER, C. 2018. Compositional optimization of hard-magnetic phases with machine-learning models. *Acta Materialia,* 153**,** 53-61.

MORGAN, D. & JACOBS, R. 2020. Opportunities and Challenges for Machine Learning in Materials Science. *Annual Review of Materials Research,* 50**,** 71-103.

MOUNET, N., GIBERTINI, M., SCHWALLER, P., CAMPI, D., MERKYS, A., MARRAZZO, A., SOHIER, T., CASTELLI, I. E., CEPELLOTTI, A. & PIZZI, G. 2018. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nature nanotechnology,* 13**,** 246-252.

NELLAIAPPAN S; KUMAR N; KUMAR R; PARUI A, M. K. D. P. K. G. S., A. K.; SHARMA, S.; TIWARY, C. S.; BISWAS, K. 2019. Nobel Metal Based High

Entropy Alloy for Conversion of Carbon Dioxide (CO2) to Hydrocarbon. . *ChemRxiv. Cambridge: Cambridge Open Engage*.

NGUYEN, T. L. A Framework for Five Big V's of Big Data and Organizational Culture in Firms. 2018 IEEE International Conference on Big Data (Big Data), 10-13 Dec. 2018 2018. 5411-5413.

NISAR, U., SHAKOOR, R. A., ESSEHLI, R., AMIN, R., ORAYECH, B., AHMAD, Z., KUMAR, P. R., KAHRAMAN, R., AL-QARADAWI, S. & SOLIMAN, A. 2018. Sodium intercalation/de-intercalation mechanism in Na4MnV(PO4)3 cathode materials. *Electrochimica Acta,* 292**,** 98-106.

NISHIJIMA, M., OOTANI, T., KAMIMURA, Y., SUEKI, T., ESAKI, S., MURAI, S., FUJITA, K., TANAKA, K., OHIRA, K., KOYAMA, Y. & TANAKA, I. 2014. Accelerated discovery of cathode materials with prolonged cycle life for lithium-ion battery. *Nature Communications,* 5**,** 4553.

NOBLE, W. S. 2006. What is a support vector machine? *Nature Biotechnology,* 24**,** 1565-1567.

NOH, J., BACK, S., KIM, J. & JUNG, Y. 2018. Active learning with non-ab initio input features toward efficient CO2 reduction catalysts. *Chemical Science,* 9**,** 5152-5159.

NOMIKOS, P. & MACGREGOR, J. F. 1994. Monitoring batch processes using multiway principal component analysis. *AIChE Journal,* 40**,** 1361-1375.

NOMIKOS, P. & MACGREGOR, J. F. 1995. Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics,* 37**,** 41-59.

NØRSKOV, J. K., BLIGAARD, T., ROSSMEISL, J. & CHRISTENSEN, C. H. 2009. Towards the computational design of solid catalysts. *Nature Chemistry,* 1**,** 37-46.

NØRSKOV, J. K., ROSSMEISL, J., LOGADOTTIR, A., LINDQVIST, L., KITCHIN, J. R., BLIGAARD, T. & JÓNSSON, H. 2004. Origin of the Overpotential for Oxygen Reduction at a Fuel-Cell Cathode. *The Journal of Physical Chemistry B,* 108**,** 17886-17892.

OGANOV, A. R. & GLASS, C. W. 2006. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *J Chem Phys,* 124**,** 244704.

OGATA, T. & YAMAZAKI, M. 2012. New stage of MatNavi, materials database at NIMS.

OKOSHI, M., YAMADA, Y., KOMABA, S., YAMADA, A. & NAKAI, H. 2016. Theoretical Analysis of Interactions between Potassium Ions and Organic Electrolyte Solvents: A Comparison with Lithium, Sodium, and Magnesium Ions. *Journal of The Electrochemical Society,* 164**,** A54-A60.

OKOSHI, M., YAMADA, Y., YAMADA, A. & NAKAI, H. 2013. Theoretical Analysis on De-Solvation of Lithium, Sodium, and Magnesium Cations to Organic Electrolyte Solvents. *Journal of The Electrochemical Society,* 160**,** A2160-A2165.

OLSON, R. S., URBANOWICZ, R. J., ANDREWS, P. C., LAVENDER, N. A., KIDD, L. C. & MOORE, J. H. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization. *In:* SQUILLERO, G. &

BURELLI, P., eds. Applications of Evolutionary Computation, 2016 2016 Cham. Springer International Publishing, 123-137.

ONG, S. P., CHEVRIER, V. L., HAUTIER, G., JAIN, A., MOORE, C., KIM, S., MA, X. & CEDER, G. 2011. Voltage, stability and diffusion barrier differences between sodium-ion and lithium-ion intercalation materials. *Energy & Environmental Science,* 4**,** 3680-3688.

ONG, S. P., CHOLIA, S., 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. *Computational Materials Science,* 97**,** 209-215.

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. *Computational Materials Science,* 68**,** 314-319.

ONG, S. P., WANG, L., KANG, B. & CEDER, G. 2008. Li− Fe− P− O2 phase diagram from first principles calculations. *Chemistry of Materials,* 20**,** 1798-1807.

OUYANG, R., AHMETCIK, E., CARBOGNO, C., SCHEFFLER, M. & GHIRINGHELLI, L. M. 2019. Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO. *Journal of Physics: Materials,* 2**,** 024002.

OUYANG, R., CURTAROLO, S., AHMETCIK, E., SCHEFFLER, M. & GHIRINGHELLI, L. M. 2018. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials,* 2**,** 083802.

PADULA, D., SIMPSON, J. D. & TROISI, A. 2019. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Materials Horizons,* 6**,** 343-349.

PALIK, E. D. 1998. *Handbook of optical constants of solids*, Academic press.

PANKAJAKSHAN, P., SANYAL, S., DE NOORD, O. E., BHATTACHARYA, I., BHATTACHARYYA, A. & WAGHMARE, U. 2017. Machine Learning and Statistical Analysis for Materials Science: Stability and Transferability of Fingerprint Descriptors and Chemical Insights. *Chemistry of Materials,* 29**,** 4190-4201.

PARADA, G. A., GOLDSMITH, Z. K., KOLMAR, S., PETTERSSON RIMGARD, B., MERCADO, B. Q., HAMMARSTRÖM, L., HAMMES-SCHIFFER, S. & MAYER, J. M. 2019. Concerted proton-electron transfer reactions in the Marcus inverted region. *Science,* 364**,** 471-475.

PEARSON, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,* 2**,** 559-572.

PEDERSEN, J. K., BATCHELOR, T. A. A., BAGGER, A. & ROSSMEISL, J. 2020. High-Entropy Alloys as Catalysts for the CO2 and CO Reduction Reactions. *ACS Catalysis,* 10**,** 2169-2176.

PENCE, H. E. & WILLIAMS, A. 2010. ChemSpider: an online chemical information resource. ACS Publications.

PENG, J., TAO, P., SONG, C., SHANG, W., DENG, T. & WU, J. 2022. Structural evolution of Pt-based oxygen reduction reaction electrocatalysts. *Chinese Journal of Catalysis,* 43**,** 47-58.

PERDEW, J. P., BURKE, K. & ERNZERHOF, M. 1996. Generalized Gradient Approximation Made Simple. *Physical Review Letters,* 77**,** 3865-3868.

PERSSON, K. A., WALDWICK, B., LAZIC, P. & CEDER, G. 2012. Prediction of solid-aqueous equilibria: Scheme to combine first-principles calculations of solids with experimental aqueous states. *Physical Review B,* 85**,** 235438.

PETERSON, A. A., CHRISTENSEN, R. & KHORSHIDI, A. 2017. Addressing uncertainty in atomistic machine learning. *Physical Chemistry Chemical Physics,* 19**,** 10978-10985.

PETERSON, A. A. & NØRSKOV, J. K. 2012. Activity Descriptors for CO2 Electroreduction to Methane on Transition-Metal Catalysts. *The Journal of Physical Chemistry Letters,* 3**,** 251-258.

PETOUSIS, I., MRDJENOVICH, D., BALLOUZ, E., LIU, M., WINSTON, D., CHEN, W., GRAF, T., SCHLADT, T. D., PERSSON, K. A. & PRINZ, F. B. 2017. High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials. *Sci Data,* 4**,** 160134.

PETRETTO, G., DWARAKNATH, S., H, P. C. M., WINSTON, D., GIANTOMASSI, M., VAN SETTEN, M. J., GONZE, X., PERSSON, K. A., HAUTIER, G. & RIGNANESE, G. M. 2018. High-throughput density-functional perturbation theory phonons for inorganic materials. *Sci Data,* 5**,** 180065.

PROSINI, P. P., ZANE, D. & PASQUALI, M. 2001. Improved electrochemical performance of a LiFePO4-based composite cathode. *Electrochimica Acta,* 46**,** 3517-3523.

QIN, G., MA, Q. & WANG, C. 2014. A porous C/LiFePO4/multiwalled carbon nanotubes cathode material for Lithium ion batteries. *Electrochimica Acta,* 115**,** 407-415.

QU, X. H., JAIN, A., RAJPUT, N. N., CHENG, L., ZHANG, Y., ONG, S. P., BRAFMAN, M., MAGINN, E., CURTISS, L. A. & PERSSON, K. A. 2015. The Electrolyte Genome project: A big data approach in battery materials discovery. *Computational Materials Science,* 103**,** 56-67.

R. AKBARZADEH, A., OZOLIŅŠ, V. & WOLVERTON, C. 2007. First-principles determination of multicomponent hydride phase diagrams: application to the Li-Mg-N-H system. *Advanced Materials,* 19**,** 3233-3239.

RAKHSHANI, E., SARIRI, I. & ROUZBEHI, K. Application of data mining on fault detection and prediction in Boiler of power plant using artificial neural network. 2009 International Conference on Power Engineering, Energy and Electrical Drives, 18-20 March 2009 2009. 473-478.

RASCHKA, S. 2018. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *CoRR,* abs/1811.12808.

RICHARDS, W. D., MIARA, L. J., WANG, Y., KIM, J. C. & CEDER, G. 2016. Interface Stability in Solid-State Batteries. *Chemistry of Materials,* 28**,** 266-273.

RUCK, M., GARLYYEV, B., MAYR, F., BANDARENKA, A. S. & GAGLIARDI, A. 2020. Oxygen Reduction Activities of Strained Platinum Core-Shell Electrocatalysts Predicted by Machine Learning. *J Phys Chem Lett,* 11**,** 1773-1780.

RÜCK, M., GARLYYEV, B., MAYR, F., BANDARENKA, A. S. & GAGLIARDI, A. 2020. Oxygen Reduction Activities of Strained Platinum Core–Shell Electrocatalysts Predicted by Machine Learning. *The Journal of Physical Chemistry Letters,* 11**,** 1773-1780.

RUMELHART, D. E., HINTON, G. E. & WILLIAMS, R. J. 1986. Learning representations by back-propagating errors. *Nature,* 323**,** 533-536.

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.

SAEKI, A. 2020. Evaluation-oriented exploration of photo energy conversion systems: from fundamental optoelectronics and material screening to the combination with data science. *Polymer Journal,* 52**,** 1307-1321.

SAHIGARA, F., MANSOURI, K., BALLABIO, D., MAURI, A., CONSONNI, V. & TODESCHINI, R. 2012. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules,* 17.

SAHU, H., RAO, W., TROISI, A. & MA, H. 2018a. Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. *Advanced Energy Materials,* 8**,** 1801032.

SAHU, H., RAO, W., TROISI, A. & MA, H. 2018b. Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. *Advanced Energy Materials,* 8.

SAHU, H., YANG, F., YE, X., MA, J., FANG, W. & MA, H. 2019. Designing promising molecules for organic solar cells via machine learning assisted virtual screening. *Journal of Materials Chemistry A,* 7**,** 17480-17488.

SANCHEZ-LENGELING, B. & ASPURU-GUZIK, A. 2018. Inverse molecular design using machine learning: Generative models for matter engineering. *Science,* 361**,** 360.

SCHLEDER, G. R., ACOSTA, C. M. & FAZZIO, A. 2020. Exploring Two-Dimensional Materials Thermodynamic Stability via Machine Learning. *ACS Applied Materials & Interfaces,* 12**,** 20149-20157.

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. *Journal of Physics: Materials,* 2**,** 032001.

SCHRITTWIESER, J., ANTONOGLOU, I., HUBERT, T., SIMONYAN, K., SIFRE, L., SCHMITT, S., GUEZ, A., LOCKHART, E., HASSABIS, D., GRAEPEL, T., LILLICRAP, T. & SILVER, D. 2020. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature,* 588**,** 604-609.

SCHWAIGHOFER, A., SCHROETER, T., MIKA, S. & BLANCHARD, G. 2009. How Wrong Can We Get? A Review of Machine Learning Approaches and Error Bars. *Combinatorial Chemistry & High Throughput Screening,* 12**,** 453-468.

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 & Environmental Science,* 10**,** 306-320.

SENIOR, A. W., EVANS, R., JUMPER, J., KIRKPATRICK, J., SIFRE, L., GREEN, T., QIN, C., ŽÍDEK, A., NELSON, A. W. R., BRIDGLAND, A., PENEDONES, H., PETERSEN, S., SIMONYAN, K., CROSSAN, S., KOHLI, P., JONES, D. T., SILVER, D., KAVUKCUOGLU, K. & HASSABIS, D. 2020. Improved protein structure prediction using potentials from deep learning. *Nature,* 577**,** 706-710.

SETTLES, B. 2012. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning,* 6**,** 1-114.

SEVERSON, K. A., ATTIA, P. M., JIN, N., PERKINS, N., JIANG, B., YANG, Z., CHEN, M. H., AYKOL, M., HERRING, P. K., FRAGGEDAKIS, D., BAZANT, M. Z., HARRIS, S. J., CHUEH, W. C. & BRAATZ, R. D. 2019. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy,* 4**,** 383-391.

SHENAI, P. M., XU, Z. & ZHAO, Y. 2012. Applications of principal component analysis (PCA) in materials science. *Principal component analysis-engineering applications.* IntechOpen.

SIEBERT, M., KRENNRICH, G., SEIBICKE, M., SIEGLE, A. F. & TRAPP, O. 2019. Identifying high-performance catalytic conditions for carbon dioxide reduction to dimethoxymethane by multivariate modelling. *Chemical Science,* 10**,** 10466-10474.

SILVER, D., HUBERT, T., SCHRITTWIESER, J., ANTONOGLOU, I., LAI, M., GUEZ, A., LANCTOT, M., SIFRE, L., KUMARAN, D., GRAEPEL, T., LILLICRAP, T., SIMONYAN, K. & HASSABIS, D. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science,* 362**,** 1140.

SMITH, J. S., NEBGEN, B., LUBBERS, N., ISAYEV, O. & ROITBERG, A. E. 2018. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics,* 148**,** 241733.

SMITH, R. C. 2013. *Uncertainty Quantification: Theory, Implementation, and Applications*, Society for Industrial and Applied Mathematics.

SODEYAMA, K., IGARASHI, Y., NAKAYAMA, T., TATEYAMA, Y. & OKADA, M. 2018. Liquid electrolyte informatics using an exhaustive search with linear regression. *Physical Chemistry Chemical Physics,* 20**,** 22585-22591.

SONG, J., SUN, B., LIU, H., MA, Z., CHEN, Z., SHAO, G. & WANG, G. 2016. Enhancement of the Rate Capability of LiFePO4 by a New Highly Graphitic

Carbon-Coating Method. *ACS Applied Materials & Interfaces,* 8**,** 15225-15231.

SPARKS, T. D., KAUWE, S. K., PARRY, M. E., TEHRANI, A. M. & BRGOCH, J. 2020. Machine Learning for Structural Materials. *Annual Review of Materials Research,* 50**,** 27-48.

STAMENKOVIC, V., MUN, B. S., MAYRHOFER, K. J. J., ROSS, P. N., MARKOVIC, N. M., ROSSMEISL, J., GREELEY, J. & NØRSKOV, J. K. 2006. Changing the Activity of Electrocatalysts for Oxygen Reduction by Tuning the Surface Electronic Structure. *Angewandte Chemie International Edition,* 45**,** 2897-2901.

STAMENKOVIC, V. R., MUN, B. S., ARENZ, M., MAYRHOFER, K. J. J., LUCAS, C. A., WANG, G., ROSS, P. N. & MARKOVIC, N. M. 2007. Trends in electrocatalysis on extended and nanoscale Pt-bimetallic alloy surfaces. *Nature Materials,* 6**,** 241-247.

STANEV, V., OSES, C., KUSNE, A. G., RODRIGUEZ, E., PAGLIONE, J., CURTAROLO, S. & TAKEUCHI, I. 2018. Machine learning modeling of superconducting critical temperature. *npj Computational Materials,* 4**,** 29.

STEVANOVIĆ, V., LANY, S., ZHANG, X. & ZUNGER, A. 2012. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Physical Review B,* 85**,** 115104.

STRIETH-KALTHOFF, F., SANDFORT, F., SEGLER, M. H. S. & GLORIUS, F. 2020. Machine learning the ropes: principles, applications and directions in synthetic chemistry. *Chemical Society Reviews,* 49**,** 6154-6168.

SUN, B., FERNANDEZ, M. & BARNARD, A. S. 2016. Statistics, damned statistics and nanoscience – using data science to meet the challenge of nanomaterial complexity. *Nanoscale Horizons,* 1**,** 89-95.

SUN, X., MACMANUS-DRISCOLL, J. L. & WANG, H. 2020a. Spontaneous Ordering of Oxide-Oxide Epitaxial Vertically Aligned Nanocomposite Thin Films. *Annual Review of Materials Research,* 50**,** 229-253.

SUN, X., MARQUEZ, H. J., CHEN, T. & RIAZ, M. 2005. An improved PCA method with application to boiler leak detection. *ISA Transactions,* 44**,** 379-397.

SUN, X., ZHENG, J., GAO, Y., QIU, C., YAN, Y., YAO, Z., DENG, S. & WANG, J. 2020b. Machine-learning-accelerated screening of hydrogen evolution catalysts in MBenes materials. *Applied Surface Science,* 526.

SUPKA, A. R., LYONS, T. E., LIYANAGE, L., D'AMICO, P., AL RAHAL AL ORABI, R., MAHATARA, S., GOPAL, P., TOHER, C., CERESOLI, D., CALZOLARI, A., CURTAROLO, S., NARDELLI, M. B. & FORNARI, M. 2017. AFLOWπ: A minimalist approach to high-throughput ab initio calculations including the generation of tight-binding hamiltonians. *Computational Materials Science,* 136**,** 76-84.

SUTTON, R. S. & BARTO, A. G. 2018. *Reinforcement learning: An introduction*, MIT press.

SWIERCZ, M. & MROCZKOWSKA, H. 2020. Multiway PCA for Early Leak Detection in a Pipeline System of a Steam Boiler-Selected Case Studies. *Sensors (Basel, Switzerland),* 20**,** 1561.

TABOR, D. P., ROCH, L. M., SAIKIN, S. K., KREISBECK, C., SHEBERLA, D., MONTOYA, J. H., DWARAKNATH, S., AYKOL, M., ORTIZ, C., TRIBUKAIT, H., AMADOR-BEDOLLA, C., BRABEC, C. J., MARUYAMA, B., PERSSON, K. A. & ASPURU-GUZIK, A. 2018. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature Reviews Materials,* 3**,** 5-20.

TADMOR, E. B., ELLIOTT, R. S., SETHNA, J. P., MILLER, R. E. & BECKER, C. A. 2011. The potential of atomistic simulations and the knowledgebase of interatomic models. *Jom,* 63**,** 17.

TIAN, N., ZHOU, Z.-Y., SUN, S.-G., DING, Y. & WANG, Z. L. 2007. Synthesis of Tetrahexahedral Platinum Nanocrystals with High-Index Facets and High Electro-Oxidation Activity. *Science,* 316**,** 732-735.

TOMITA, Y., MATSUSHITA, H., KOBAYASHI, K., MAEDA, Y. & YAMADA, K. 2008. Substitution effect of ionic conductivity in lithium ion conductor, LI3INBR6−xCLx. *Solid State Ionics,* 179**,** 867-870.

TRACY, N. D., YOUNG, J. C. & MASON, R. L. 1992. Multivariate Control Charts for Individual Observations. *Journal of Quality Technology,* 24**,** 88-95.

TRAN, K. & ULISSI, Z. W. 2018. Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution. *Nature Catalysis,* 1**,** 696-703.

TRAN, R., XU, Z., RADHAKRISHNAN, B., WINSTON, D., SUN, W., PERSSON, K. A. & ONG, S. P. 2016. Surface energies of elemental crystals. *Scientific Data,* 3**,** 160080.

ULISSI, Z. W., SINGH, A. R., TSAI, C. & NØRSKOV, J. K. 2016. Automated Discovery and Construction of Surface Phase Diagrams Using Machine Learning. *The Journal of Physical Chemistry Letters,* 7**,** 3931-3935.

ULISSI, Z. W., TANG, M. T., XIAO, J., LIU, X., TORELLI, D. A., KARAMAD, M., CUMMINS, K., HAHN, C., LEWIS, N. S., JARAMILLO, T. F., CHAN, K. & NØRSKOV, J. K. 2017. Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO2 Reduction. *ACS Catalysis,* 7**,** 6600-6608.

VAN SANTEN, R. A., NEUROCK, M. & SHETTY, S. G. 2010. Reactivity Theory of Transition-Metal Surfaces: A Brønsted−Evans−Polanyi Linear Activation Energy−Free-Energy Analysis. *Chemical Reviews,* 110**,** 2005-2048.

VAN SPRANG, E. N. M., RAMAKER, H.-J., WESTERHUIS, J. A., GURDEN, S. P. & SMILDE, A. K. 2002. Critical evaluation of approaches for on-line batch process monitoring. *Chemical Engineering Science,* 57**,** 3979-3991.

VEJ-HANSEN, U. G., ESCUDERO-ESCRIBANO, M., VELÁZQUEZ-PALENZUELA, A., MALACRIDA, P., ROSSMEISL, J., L. STEPHENS, I. E., CHORKENDORFF, I. & SCHIØTZ, J. 2017. New Platinum Alloy Catalysts for Oxygen Electroreduction Based on Alkaline Earth Metals. *Electrocatalysis,* 8**,** 594-604.

VILLARS, P., BERNDT, M., BRANDENBURG, K., CENZUAL, K., DAAMS, J., HULLIGER, F., MASSALSKI, T., OKAMOTO, H., OSAKI, K. & PRINCE, A. 2004. The pauling file. *Journal of Alloys and Compounds,* 367**,** 293-297.

VILLARS, P. & CENZUAL, K. 2009. Pearson's Crystal Data. *Crystal Structure Database for Inorganic Compounds (Materials Park (OH): ASM International, 2012)*.

VINYALS, O., BABUSCHKIN, I., CZARNECKI, W. M., MATHIEU, M., DUDZIK, A., CHUNG, J., CHOI, D. H., POWELL, R., EWALDS, T., GEORGIEV, P., OH, J., HORGAN, D., KROISS, M., DANIHELKA, I., HUANG, A., SIFRE, L., CAI, T., AGAPIOU, J. P., JADERBERG, M., VEZHNEVETS, A. S., LEBLOND, R., POHLEN, T., DALIBARD, V., BUDDEN, D., SULSKY, Y., MOLLOY, J., PAINE, T. L., GULCEHRE, C., WANG, Z., PFAFF, T., WU, Y., RING, R., YOGATAMA, D., WÜNSCH, D., MCKINNEY, K., SMITH, O., SCHAUL, T., LILLICRAP, T., KAVUKCUOGLU, K., HASSABIS, D., APPS, C. & SILVER, D. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature,* 575**,** 350-354.

VOVK, V. 2013. Kernel Ridge Regression. *In:* SCHÖLKOPF, B., LUO, Z. & VOVK, V. (eds.) *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik.* Berlin, Heidelberg: Springer Berlin Heidelberg.

WADA, H., MENETRIER, M., LEVASSEUR, A. & HAGENMULLER, P. 1983. Preparation and ionic conductivity of new B2S3-Li2S-LiI glasses. *Materials Research Bulletin,* 18**,** 189-193.

WANG, A. Y.-T., MURDOCK, R. J., KAUWE, S. K., OLIYNYK, A. O., GURLO, A., BRGOCH, J., PERSSON, K. A. & SPARKS, T. D. 2020a. Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. *Chemistry of Materials,* 32**,** 4954-4965.

WANG, D. S., AMSLER, M., HEGDE, V. I., SAAL, J. E., ISSA, A., ZHOU, B. C., ZENG, X. Q. & WOLVERTON, C. 2018. Crystal structure, energetics, and phase stability of strengthening precipitates in Mg alloys: A first-principles study. *Acta Materialia,* 158**,** 65-78.

WANG, G. X., YANG, L., CHEN, Y., WANG, J. Z., BEWLAY, S. & LIU, H. K. 2005. An investigation of polypyrrole-LiFePO4 composite cathode materials for lithium-ion batteries. *Electrochimica Acta,* 50**,** 4649-4654.

WANG, H., LAI, A., HUANG, D., CHU, Y., HU, S., PAN, Q., LIU, Z., ZHENG, F., HUANG, Y. & LI, Q. 2021a. Y–F co-doping behavior of LiFePO4/C nanocomposites for high-rate lithium-ion batteries. *New Journal of Chemistry,* 45**,** 5695-5703.

WANG, H., ZHAO, N., SHI, C., HE, C., LI, J. & LIU, E. 2016. Interface and Doping Effect on the Electrochemical Property of Graphene/LiFePO4. *The Journal of Physical Chemistry C,* 120**,** 17165-17174.

WANG, L., WANG, H., LIU, Z., XIAO, C., DONG, S., HAN, P., ZHANG, Z., ZHANG, X., BI, C. & CUI, G. 2010. A facile method of preparing mixed conducting LiFePO4/graphene composites for lithium-ion batteries. *Solid State Ionics,* 181**,** 1685-1689.

WANG, S., CHANG, Y.-Q., ZHAO, Z. & WANG, F.-L. 2012a. Multi-phase MPCA modeling and application based on an improved phase separation method. *International Journal of Control, Automation and Systems,* 10**,** 1136-1145.

WANG, X., LI, Z., QU, Y., YUAN, T., WANG, W., WU, Y. & LI, Y. 2019. Review of Metal Catalysts for Oxygen Reduction Reaction: From Nanoscale Engineering to Atomic Design. *Chem,* 5**,** 1486-1511.

WANG, X., XIAO, R., LI, H. & CHEN, L. 2017. Quantitative structure-property relationship study of cathode volume changes in lithium ion batteries using ab-initio and partial least squares analysis. *Journal of Materiomics,* 3**,** 178-183.

WANG, Y., TIAN, Y., KIRK, T., LARIS, O., ROSS, J. H., NOEBE, R. D., KEYLIN, V. & ARRÓYAVE, R. 2020b. Accelerated design of Fe-based soft magnetic materials using machine learning and stochastic optimization. *Acta Materialia,* 194**,** 144-155.

WANG, Y., WANG, C., LI, M., YU, Y. & ZHANG, B. 2021b. Nitrate electroreduction: mechanism insight, in situ characterization, performance evaluation, and challenges. *Chemical Society Reviews,* 50**,** 6720-6733.

WANG, Z.-H., PANG, Q.-Q., DENG, K.-J., YUAN, L.-X., HUANG, F., PENG, Y.-L. & HUANG, Y.-H. 2012b. Effects of titanium incorporation on phase and electrochemical performance in LiFePO4 cathode material. *Electrochimica Acta,* 78**,** 576-584.

WANG, Z., SUN, Z., YIN, H., LIU, X., WANG, J., ZHAO, H., PANG, C. H., WU, T., LI, S., YIN, Z. & YU, X. 2022. Data-Driven Materials Innovation and Applications. *Advanced Materials,* n/a**,** 2104113.

WANG, Z., ZHANG, H. & LI, J. 2021c. Accelerated discovery of stable spinels in energy systems via machine learning. *Nano Energy,* 81**,** 105665.

WARD, L., AGRAWAL, A., CHOUDHARY, A. & WOLVERTON, C. 2016a. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials,* 2**,** 16028.

WARD, L., AGRAWAL, A., CHOUDHARY, A. & WOLVERTON, C. 2016b. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Computational Materials,* 2**,** 16028.

WARD, L., DUNN, A., FAGHANINIA, A., ZIMMERMANN, N. E. R., BAJAJ, S., WANG, Q., MONTOYA, J., CHEN, J. M., BYSTROM, K., DYLLA, M., CHARD, K., ASTA, M., PERSSON, K. A., SNYDER, G. J., FOSTER, I. & JAIN, A. 2018a. Matminer: An open source toolkit for materials data mining. *Computational Materials Science,* 152**,** 60-69.

WARD, L., O'KEEFFE, S. C., STEVICK, J., JELBERT, G. R., AYKOL, M. & WOLVERTON, C. 2018b. A machine learning approach for engineering bulk metallic glass alloys. *Acta Materialia,* 159**,** 102-111.

WEI, J., CHU, X., SUN, X.-Y., XU, K., DENG, H.-X., CHEN, J., WEI, Z. & LEI, M. 2019. Machine learning in materials science. *InfoMat,* 1**,** 338-358.

WENG, B., SONG, Z., ZHU, R., YAN, Q., SUN, Q., GRICE, C. G., YAN, Y. & YIN, W.-J. 2020. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. *Nature Communications,* 11**,** 3513.

WESTON, L. & STAMPFL, C. 2018. Machine learning the band gap properties of kesterite ${\mathrm{I}}_{2}\text{\ensuremath{-}}\mathrm{II}\text{\ensuremath{-}}\mathrm{IV}\text{\ensuremath{-}}{\mathrm{V}}_{4}$ quaternary compounds for photovoltaics applications. *Physical Review Materials,* 2**,** 085407.

WEXLER, R. B., MARTIREZ, J. M. P. & RAPPE, A. M. 2018a. Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni2P from Nonmetal Surface Doping Interpreted via Machine Learning. *Journal of the American Chemical Society,* 140**,** 4678-4683.

WEXLER, R. B., MARTIREZ, J. M. P. & RAPPE, A. M. 2018b. Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni2P from Nonmetal Surface Doping Interpreted via Machine Learning. *J Am Chem Soc,* 140**,** 4678-4683.

WEYMUTH, T. & REIHER, M. 2014. Inverse quantum chemistry: Concepts and strategies for rational compound design. *International Journal of Quantum Chemistry,* 114**,** 823-837.

WHITE, P. S., RODGERS, J. R. & LE PAGE, Y. 2002. CRYSTMET: a database of the structures and powder patterns of metals and intermetallics. *Acta Crystallographica Section B: Structural Science,* 58**,** 343-348.

WIDOM, M. & MIHALKOVIC, M. 2005. Stability of Fe-based alloys with structure type C 6 Cr 23. *Journal of materials research,* 20**,** 237-242.

WIRYADINATA, S., MOREJOHN, J. & KORNBLUTH, K. 2019. Pathways to carbon neutral energy systems at the University of California, Davis. *Renewable Energy,* 130**,** 853-866.

WISE, B. M., GALLAGHER, N. B. & MARTIN, E. B. 2001. Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch. *Journal of Chemometrics,* 15**,** 285-298.

WU, D., SHEN, X., PAN, Y., YAO, L. & PENG, Z. 2020. Platinum Alloy Catalysts for Oxygen Reduction Reaction: Advances, Challenges and Perspectives. *ChemNanoMat,* 6**,** 32-41.

WU, N., WAN, S., SU, S., HUANG, H., DOU, G. & SUN, L. 2021. Electrode materials for brain–machine interface: A review. *InfoMat,* 3**,** 1174-1194.

WU, P., HE, T., ZHU, H., WANG, Y., LI, Q., WANG, Z., FU, X., WANG, F., WANG, P., SHAN, C., FAN, Z., LIAO, L., ZHOU, P. & HU, W. 2022. Next-generation machine vision systems incorporating two-dimensional materials: Progress and perspectives. *InfoMat,* 4**,** e12275.

XI, H., WU, X., CHEN, X. & SHA, P. 2021. Artificial intelligent based energy scheduling of steel mill gas utilization system towards carbon neutrality. *Applied Energy,* 295**,** 117069.

XI, Y. & LU, Y. 2020. Toward Uniform In Situ Carbon Coating on Nano-LiFePO4 via a Solid-State Reaction. *Industrial & Engineering Chemistry Research,* 59**,** 13549-13555.

XIE, H. & ZHOU, Z. 2006. Physical and electrochemical properties of mix-doped lithium iron phosphate as cathode material for lithium ion battery. *Electrochimica Acta,* 51**,** 2063-2067.

XIE, T. & GROSSMAN, J. C. 2018. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters,* 120**,** 145301.

XU, G., ZHONG, K., ZHANG, J.-M. & HUANG, Z. 2015. First-principles study of structural, electronic and Li-ion diffusion properties of N-doped LiFePO4 (010) surface. *Solid State Ionics,* 281**,** 1-5.

XU, K. 2004. Nonaqueous Liquid Electrolytes for Lithium-Based Rechargeable Batteries. *Chemical Reviews,* 104**,** 4303-4418.

XU, M., XU, M. & MIAO, X. 2022. Deep machine learning unravels the structural origin of mid-gap states in chalcogenide glass for high-density memory integration. *InfoMat,* 4**,** e12315.

XU, Q.-S., LIANG, Y.-Z. & DU, Y.-P. 2004. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *Journal of Chemometrics,* 18**,** 112-120.

XU, W., ANDERSEN, M. & REUTER, K. 2020. Data-Driven Descriptor Engineering and Refined Scaling Relations for Predicting Transition Metal Oxide Reactivity. *ACS Catalysis***,** 734-742.

XUE, D., BALACHANDRAN, P. V., YUAN, R., HU, T., QIAN, X., DOUGHERTY, E. R. & LOOKMAN, T. 2016. Accelerated search for BaTiO&lt;sub&gt;3&lt;/sub&gt;-based piezoelectrics with vertical morphotropic phase boundary using Bayesian learning. *Proceedings of the National Academy of Sciences,* 113**,** 13301.

YAKUTOVICH, A. V., EIMRE, K., SCHÜTT, O., TALIRZ, L., ADORF, C. S., ANDERSEN, C. W., DITLER, E., DU, D., PASSERONE, D., SMIT, B., MARZARI, N., PIZZI, G. & PIGNEDOLI, C. A. 2021. AiiDAlab – an ecosystem for developing, executing, and sharing scientific workflows. *Computational Materials Science,* 188**,** 110165.

YAMADA, K., KUMANO, K. & OKUDA, T. 2006. Lithium superionic conductors Li3InBr6 and LiInBr4 studied by 7Li, 115In NMR. *Solid State Ionics,* 177**,** 1691-1695.

YAN, L.-M., SU, J.-M., SUN, C. & YUE, B.-H. 2014. Review of the first principles calculations and the design of cathode materials for Li-ion batteries. *Advances in Manufacturing,* 2**,** 358-368.

YANG, F.-R., GAO, L., LAI, W.-C. & HUANG, H.-W. 2023. Recent advance on structural design of high-performance Pt-based nanocatalysts for oxygen reduction reaction. *Advanced Sensor and Energy Materials,* 2**,** 100022.

YAP, C. W. 2011. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry,* 32**,** 1466-1474.

YI, H.-S., YEO, Y.-K., KIM, J.-K., KIM, M. K. & KANG, S. S. 1998. A Rule-Based Steam Distribution System for Petrochemical Plant Operation. *Industrial & Engineering Chemistry Research,* 37**,** 1051-1062.

YIN, H., SUN, Z., WANG, Z., TANG, D., PANG, C. H., YU, X., BARNARD, A. S., ZHAO, H. & YIN, Z. 2021. The data-intensive scientific revolution occurring where two-dimensional materials meet machine learning. *Cell Reports Physical Science,* 2**,** 100482.

YOSIPOF, A., NAHUM, O. E., ANDERSON, A. Y., BARAD, H.-N., ZABAN, A. & SENDEROWITZ, H. 2015. Data Mining and Machine Learning Tools for Combinatorial Material Science of All-Oxide Photovoltaic Cells. *Molecular Informatics,* 34**,** 367-379.

YU, J., JANG, J., YOO, J., PARK, J. H. & KIM, S. 2017. Bagged auto-associative kernel regression-based fault detection and identification approach for steam boilers in thermal power plants. *Journal of Electrical Engineering and Technology,* 12**,** 1406-1416.

YUAN, R., LIU, Z., BALACHANDRAN, P. V., XUE, D., ZHOU, Y., DING, X., SUN, J., XUE, D. & LOOKMAN, T. 2018. Accelerated Discovery of Large Electrostrains in BaTiO3-Based Piezoelectrics Using Active Learning. *Advanced Materials,* 30**,** 1702884.

YUE, H. H. & QIN, S. J. 2001. Reconstruction-Based Fault Identification Using a Combined Index. *Industrial & Engineering Chemistry Research,* 40**,** 4403-4414.

ZAGORAC, D., MULLER, H., RUEHL, S., ZAGORAC, J. & REHME, S. 2019. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *J Appl Crystallogr,* 52**,** 918-925.

ZHANG, J.-F., SHEN, C., ZHANG, B., ZHENG, J.-C., PENG, C.-L., WANG, X.-W., YUAN, X.-B., LI, H. & CHEN, G.-M. 2014. Synthesis and performances of 2LiFePO4·Li3V2(PO4)3/C cathode materials via spray drying method with double carbon sources. *Journal of Power Sources,* 267**,** 227-234.

ZHANG, J., HU, P. & WANG, H. 2020a. Amorphous Catalysis: Machine Learning Driven High-Throughput Screening of Superior Active Site for Hydrogen Evolution Reaction. *The Journal of Physical Chemistry C,* 124**,** 10483-10494.

ZHANG, M., LI, J., KANG, L., ZHANG, N., HUANG, C., HE, Y., HU, M., ZHOU, X. & ZHANG, J. 2020b. Machine learning-guided design and development of multifunctional flexible Ag/poly (amic acid) composites using the differential evolution algorithm. *Nanoscale,* 12**,** 3988-3996.

ZHANG, X., LI, Y., GUO, P., LE, J.-B., ZHOU, Z.-Y., CHENG, J. & SUN, S.-G. 2019. Theory on optimizing the activity of electrocatalytic proton coupled electron transfer reactions. *Journal of Catalysis,* 376**,** 17-24.

ZHANG, X., ZHANG, Z., YAO, S., CHEN, A., ZHAO, X. & ZHOU, Z. 2018. An effective method to screen sodium-based layered materials for sodium ion batteries. *npj Computational Materials,* 4**,** 13.

ZHANG, Y., ALARCO, J. A., NERKAR, J. Y., BEST, A. S., SNOOK, G. A., TALBOT, P. C. & COWIE, B. C. C. 2020c. Observation of Preferential Cation Doping on the Surface of LiFePO4 Particles and Its Effect on Properties. *ACS Applied Energy Materials,* 3**,** 9158-9167.

ZHANG, Y., ZHOU, Y. J., LIN, J. P., CHEN, G. L. & LIAW, P. K. 2008. Solid-Solution Phase Formation Rules for Multi-component Alloys. *Advanced Engineering Materials,* 10**,** 534-538.

ZHONG, M., TRAN, K., MIN, Y., WANG, C., WANG, Z., DINH, C.-T., DE LUNA, P., YU, Z., RASOULI, A. S. & BRODERSEN, P. 2020a. Accelerated discovery of CO 2 electrocatalysts using active machine learning. *Nature,* 581**,** 178-183.

ZHONG, M., TRAN, K., MIN, Y., WANG, C., WANG, Z., DINH, C.-T., DE LUNA, P., YU, Z., RASOULI, A. S., BRODERSEN, P., SUN, S., VOZNYY, O., TAN, C.-S., ASKERKA, M., CHE, F., LIU, M., SEIFITOKALDANI, A., PANG, Y., LO, S.-C., IP, A., ULISSI, Z. & SARGENT, E. H. 2020b. Accelerated discovery of CO2 electrocatalysts using active machine learning. *Nature,* 581**,** 178-183.

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. *Physical Review B,* 70**,** 235121.

ZHOU, J., SHI, Q., ULLAH, S., YANG, X., BACHMATIUK, A., YANG, R. & RUMMELI, M. H. 2020. Phosphorus-Based Composites as Anode Materials for Advanced Alkali Metal Ion Batteries. *Advanced Functional Materials,* 30**,** 2004648.

ZHOU, L., PAN, S., WANG, J. & VASILAKOS, A. V. 2017. Machine learning on big data: Opportunities and challenges. *Neurocomputing,* 237**,** 350-361.

ZHU, S., LI, J., MA, L., HE, C., LIU, E., HE, F., SHI, C. & ZHAO, N. 2018. Artificial neural network enabled capacitance prediction for carbon-based supercapacitors. *Materials Letters,* 233**,** 294-297.

ZHU, X., YAN, J., GU, M., LIU, T., DAI, Y., GU, Y. & LI, Y. 2019a. 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.

ZHU, X., YAN, J., GU, M., LIU, T., DAI, Y., GU, Y. & LI, Y. 2019b. Activity Origin and Design Principles for Oxygen Reduction on Dual-Metal-Site Catalysts: A Combined Density Functional Theory and Machine Learning Study. *The Journal of Physical Chemistry Letters,* 10**,** 7760-7766.

ZUNGER, A. 2018. Inverse design in search of materials with target functionalities. *Nature Reviews Chemistry,* 2**,** 0121.