

Understanding ORR Activity of Nanoclusters Electrocatalysts: Combined DFT and Machine Learning Approach for Multiscale Modeling

M.Sc. Thesis

**By
SARTHAK MAITY**



**DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF TECHNOLOGY INDORE**

May 2025

Understanding ORR Activity of Nanoclusters Electrocatalysts: Combined DFT and Machine Learning Approach for Multiscale Modelling

A THESIS

*Submitted in partial fulfillment of the
requirements for the award of the degree*
of
Master of Science

by
SARTHAK MAITY



**DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF TECHNOLOGY INDORE**

May 2025

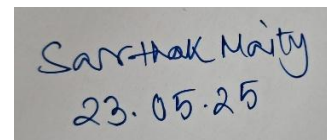


INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION


I hereby certify that the work which is being presented in the thesis entitled **“Understanding ORR Activity of Nanoclusters Electrocatalysts: Combined DFT and Machine Learning Approach for Multiscale Modelling”** in the partial fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE** and submitted in the **DEPARTMENT OF CHEMISTRY**, Indian Institute of Technology Indore, is an authentic record of my own work carried out during the time period from July 2024 to May 2025 under the supervision of **Dr. BISWARUP PATHAK**, Professor, Department of Chemistry, IIT Indore.

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



Signature of the student with date
Sarthak Maity

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



Signature of the Supervisor of M.Sc.
Thesis (with date)
Prof. Biswarup Pathak

Sarthak Maity has successfully given his M.Sc. Oral Examination, held on May 15, 2025.

ACKNOWLEDGEMENTS

I want to convey my sincere gratitude to my supervisor, **Prof. Biswarup Pathak**, Department of Chemistry, IIT Indore, for providing me with the opportunity to work on this M.Sc. project work under his esteemed supervision.

I am deeply indebted to Mr. Rahul Kumar Sharma and Ms. Harpriya Minhas for their invaluable guidance and mentorship throughout my learning journey. Their insightful feedback and unwavering support have been instrumental during my one-year-long M.Sc. research project. I would also like to express my sincere gratitude to my lab members for their constructive discussions, valuable suggestions, and continuous support at every stage of the project. Additionally, I thank other group members for fostering a supportive and collaborative environment. Finally, I am profoundly grateful to my family and friends for their constant encouragement and moral support.

Sarthak Maity

Dedicated to My Parents



Abstract

Unravelling the nature of ORR interaction with catalyst surfaces under reaction conditions is essential for designing next-generation electrocatalysts. In this work, we explore the coverage-dependent adsorption behavior of these intermediates on graphene-supported platinum subnano clusters (Pt_n , $n = 7 - 13$) using spin-polarized DFT and ab initio thermodynamic analysis. Our study uncovers a delicate balance of lateral interactions both attractive and repulsive that influence adsorption strength as surface coverage increases. By calculating differential average adsorption energies across various coverage scenarios, we reveal non-monotonic trends in stability that highlight the complex thermodynamic landscape of subnanometer clusters, demonstrating the pivotal role of multi-site interactions and structural fluxionality in catalytic performance. When combined with machine learning models built on geometric and electronic descriptors, our approach provides an effective strategy to predict active sites and break free from traditional scaling limitations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

Chapter 1: Introduction

- 1.1. Sustainable Energy Conversion Technologies
- 1.2. Comprehensive Review of Fuel Cell Technologies
- 1.3. Proton-Exchange Membrane Fuel Cell (PEMFC)
 - 1.3.1. Electrode Reactions
 - 1.3.1.1. Hydrogen Oxidation Reaction (HOR)
 - 1.3.1.2. Oxygen Reduction Reaction (ORR)
- 1.4. Electrocatalysts
 - 1.4.1. Heterogeneous Catalysts
 - 1.4.1.1. Nanocluster Catalysts

Chapter 2: Review of Past Work and Problem Formulation

Chapter 3: Theoretical Methods

- 3.1. Schrödinger Equation
 - 3.1.1. Many-Body Problem
 - 3.1.2. The Born-Oppenheimer (BO) Approximation
 - 3.1.3. Mean-Field Approximation
- 3.2. Hartree-Fock (HF) Theory
 - 3.2.1. Hartree Approximation: Independent Particle Model
 - 3.2.2. Antisymmetry and Slater Determinant
 - 3.2.3. Energy Hamiltonian
 - 3.2.4. The Self-Consistent Field (SCF) Method
 - 3.2.5. The Hartree-Fock Potential
- 3.3. Density Functional Theory (DFT)
 - 3.3.1. Electron Density in DFT
 - 3.3.2. The Hohenberg-Kohn Theorems
 - 3.3.3. Kohn-Sham Equations
 - 3.3.4. Exchange-Correlation Functionals
 - 3.3.4. Local Density Approximation (LDA)

- 3.3.4. Generalized Gradient Approximation (GGA)
- 3.3.5. Dispersion Corrected Density Functional Theory
- 3.3.6. Spin Polarized Density Functional Theory
- 3.3.7. Projector Augmented Wave (PAW) Method
- 3.3.8. Basis Sets
 - 3.3.8.1. Plane-Wave Basis Sets
- 3.3.9. Pseudopotentials
- 3.4. Computational Hydrogen Electrode (CHE) Model

Chapter 4: Machine Learning Methods

- 4.1 Artificial Intelligence (AI)
- 4.2. Machine Learning (ML)
 - 4.2.1. Supervised Learning
 - 4.2.1.1. Regression
 - 4.2.1.2. Classification
 - 4.2.2. Train and Test Data
 - 4.2.3. Feature Representations
 - 4.2.4. Performance Evaluations of ML Models
 - 4.2.4.1. Root Mean Square Error
 - 4.2.4.2. Mean Absolute Error
 - 4.2.4.3. R-Squared
 - 4.2.5. Hyperparameter Tuning
 - 4.2.5.1. GridSearchCV
 - 4.2.5.2. RandomSearchCV
 - 4.2.6. Cross-Validation (CV) Method
 - 4.2.6.1. K-Fold CV
 - 4.2.7. ML Algorithms
 - 4.2.7.1. Kernel Ridge Regression (KRR)
 - 4.2.7.2. Random Forest Regression (RFR)
 - 4.2.7.3. eXtreme Gradient Boosting Regression (XGBR)
 - 4.2.7.4. Gradient Boosting Regression (GBR)
 - 4.2.7.5. Extra Trees Regression (ETR)
 - 4.2.7.6. Adaptive Boosting (AdaBoost)
 - 4.2.7.7. Categorical Boosting (CatBoost)

4.2.8. Correlation Matrices

4.2.8.1. Pearson Correlation Coefficient (PCC)

4.2.8.2. Spearman's Correlation Coefficient (SCC)

Chapter 5: Results and Discussion

Chapter 6: Conclusions and Scope for Future Work

APPENDIX-A

REFERENCES

LIST OF FIGURES

Figure 1.1	Schematic representation of the working mechanism of a PEMFC.
Figure 1.2	Schematic illustration of the Oxygen Reduction Reaction (ORR) mechanism.
Figure 3.1	Isomeric distribution of Pt_n/G ($n = 7-13$) (a) distribution of isomers of isomers as a function of energy across the entire sampled energy landscape, (b) isomer distribution within 0.4 eV relative to the global minimum (GM) energy (GM energy taken as the reference). (c) Side and top views for the energetically most stable global minimum structures of each Pt_n/G SNCs identified through GO.
Figure 3.2	Adsorption energy landscape of ORR intermediates on graphene-supported Pt_n ($n = 7 - 13$) SNCs: (a) distribution of adsorption energies for all optimized intermediate-cluster configurations, (b) adsorption energy distribution for the most stable intermediate-cluster configurations. (c) Side views of the DFT-optimized most stable adsorption geometries of ORR intermediates on the global minimum structures of graphene-supported Pt SNCs.
Figure 3.3	Schematic representation of the screening workflow to identify active electrocatalysts from the adsorption energy dataset, guided by the Sabatier principle.
Figure 3.4	Pearson's correlation coefficient (PCC) matrices for adsorption energy datasets: Correlation matrices illustrating feature-feature and feature-target relationships for (a) E_{*O} , (b) E_{*OH} , and (c) E_{*OOH} datasets, computed using the initial 12 input features.
Figure 3.5	The predictive performance of five machine learning models: KRR, XGBR, RFR, ABR, ETR, GBR, CR was assessed using MAE, RMSE, and R^2 for the (a) E_{*O} , (b) E_{*OH} , and (c) E_{*OOH} datasets, reported in units of eV. For each intermediate, the model yielding the best performance metrics is highlighted with a pink rectangular box. Additionally, parity plots are presented to compare DFT-calculated versus ML-predicted values

	of (d) E_{*O} , (e) E_{*OH} , and (f) E_{*OOH} , as obtained from the respective best-performing models.
Figure 3.6	Construction of volcano plots for ORR activity using (a) DFT-calculated, and (b) ML-predicted η values in data set 3. Heat map corresponds to (c) DFT predicted η values, (d) ML predicted η values. Catalysts positioned at the apex outperform the Pt(111) surface (marked horizontally dark pink color) at the subnanometer regime with the lower η values.
Figure A1	LEME structures of Pt ₇ /G SNCs: LM1-LM9 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM.
Figure A2	LEME structures of Pt ₈ /G SNCs: LM1-LM12 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM.
Figure A3	LEME structures of Pt ₉ /G SNCs: LM1-LM7 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM.
Figure A4	LEME structures of Pt ₁₀ /G SNCs: LM1-LM15 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM.
Figure A5	LEME structures of Pt ₁₁ /G SNCs: LM1-LM6 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM.
Figure A6	LEME structures of Pt ₁₂ /G SNCs: LM1-LM5 depict the low-energy configurations along with their relative energies (in eV) with respect to the global minimum (GM). Using GO, three structures were identified within 0.4 eV of the GM and two additional structures slightly above this threshold.

Figure A7	LEME structures of Pt ₁₃ /G SNCs: LM1-LM8 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM
Figure A8	Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt ₇ supported on graphene.
Figure A9	Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt ₈ supported on graphene.
Figure A10	Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt ₉ supported on graphene.
Figure A11	Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt ₁₀ supported on graphene.
Figure A12	Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt ₁₁ supported on graphene.
Figure A13	Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt ₁₂ supported on graphene.
Figure A14	Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt ₁₃ supported on graphene.
Figure A15	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt ₇ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A16	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt ₈ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A17	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt ₉ /G

	within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A18	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt ₁₀ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A19	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt ₁₁ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A20	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt ₁₂ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A21	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt ₁₃ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A22	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt ₇ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A23	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt ₈ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A24	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt ₉ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A25	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt ₁₀ /G

	within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A26	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt ₁₁ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A27	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt ₁₂ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A28	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt ₁₃ /G within the LEME. Here, the asterisk * represents active sites of the catalyst.
Figure A29	Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt ₇ O _x (x =1–13) Pt _n /G SNCs.
Figure A3	Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt ₈ O _x (x =1–13) Pt _n /G SNCs.
Figure A31	Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt ₉ O _x (x =1–13) Pt _n /G SNCs.
Figure A32	Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt ₁₀ O _x (x =1–13) Pt _n /G SNCs.
Figure A33	Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a

	function of oxygen chemical potential for Pt_{11}O_x ($x=1-13$) Pt_n/G SNCs.
Figure A34	Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt_{12}O_x ($x=1-13$) Pt_n/G SNCs.
Figure A35	Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt_{13}O_x ($x=1-13$) Pt_n/G SNCs.
Figure A36	DFT optimized oxidized structures of the most active GM isomer of $\text{Pt}_7\text{O}_x/\text{G}$ ($x=2-6$) at varying oxygen coverage.
Figure A37	DFT optimized oxidized structure of the most active LM3 isomer of $\text{Pt}_8\text{O}_x/\text{G}$ ($x=2-8$, $x \neq 5$) at varying oxygen coverage.
Figure A38	DFT optimized oxidized structures of the most active LM3 isomer of $\text{Pt}_9\text{O}_x/\text{G}$ ($x=2-9$, $x \neq 7$) at varying oxygen coverage.
Figure A39	DFT optimized oxidized structures of the most active LM5 isomer of $\text{Pt}_{10}\text{O}_x/\text{G}$ ($x=2-9$) at varying oxygen coverage.
Figure A40	DFT optimized oxidized structures of the most active LM1 isomer of $\text{Pt}_{11}\text{O}_x/\text{G}$ ($x=2-11$, $x \neq 11$) at varying oxygen coverage.
Figure A41	DFT optimized oxidized structures of the most active LM1 isomer of $\text{Pt}_{12}\text{O}_x/\text{G}$ ($x=2-12$, $x \neq 11$) at varying oxygen coverage.
Figure A42	DFT optimized oxidized structures of the most active GM isomer of $\text{Pt}_{13}\text{O}_x/\text{G}$ ($x=2-13$, $x \neq 12$) at varying oxygen coverage.
Figure A43	Pearson's correlation coefficient (PCC) matrices for adsorption energy datasets: Correlation matrices illustrating feature-feature and feature-target relationships for (a) E_{*O} , (b) E_{*OH} , and (c) E_{*OOH} datasets, computed using the extracted 7 features.

LIST OF TABLES

Table 1.1	Classification of fuel cells and their operational characteristics.
Table 3.1	List of Descriptors capturing elemental, electronic, and geometric Properties.
Table A1	The total number of structural configurations sampled for each size-selected Pt _n /G (n = 7-13) isomer to identify the most stable adsorption geometries of ORR intermediates.
Table A2	Adsorption energies of ORR intermediates in the gas phase: adsorption energies (in eV) of ORR intermediates corresponding to their most stable intermediate–cluster adsorption configurations on Pt _n /G (n = 7-13), evaluated in the gas phase.
Table A3	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt ₇ /G within the LEME.
Table A4	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt ₈ /G within the LEME.
Table A5	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt ₉ /G within the LEME.
Table A6	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt ₁₀ /G within the LEME.
Table A7	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt ₁₁ /G within the LEME.
Table A8	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt ₁₂ /G within the LEME.
Table A9	Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt ₁₃ /G within the LEME.

Table A10	Identified rate-determining steps (RDS) for various metastable isomers of each Pt_n/G ($n = 7 - 13$) SNCs, evaluated under gas-phase conditions following the associative mechanism of the ORR.
Table A11	Refined set of descriptors retained for machine learning analysis following the elimination of highly correlated features.

ACRONYMS

AI	Artificial Intelligence
CHE	Computational Hydrogen Electrode
CN	Coordination Number
CSNE	Core-Shell Nanocluster
CV	Cross Validation
DFT	Density Functional Theory
DOS	Density of States
ETR	Extra Trees Regression
GBR	Gradient Boosting Regression
GCN	Generalized Coordination Number
GGA	Generalized Gradient Approximation
HOR	Hydrogen Oxidation Reaction
GO	Global Optimization
GM	Global Minimum
KRR	Kernel Ridge Regression
LDA	Local Density Approximation
LEME	Low-Energy Metastable Ensemble
LFESR	Linear Free Energy Scaling Relationship
LM	Local Minimum
MAE	Mean Absolute Error
MI	Metastable Isomer
ML	Machine Learning
NC	Nanocluster
ORR	Oxygen Reduction Reaction
PAW	Projected Augmented Wave
PBE	Perdew-Burke-Ernzerhof

PEMFC	Proton Exchange Membrane Fuel Cell
PGM	Putative Global Minimum
pMuTT	Python Multiscale Thermochemistry Toolbox
RFR	Random Forest Classifier
RHE	Reversible Hydrogen Electrode
RMSE	Root Mean Square Error
RDS	Rate Determining Step
SHE	Standard Hydrogen Electrode
SNC	Subnanocluster
TDSE	Time Dependent Schrodinger Equation
TISE	Time Independent Schrodinger Equation
VASP	Vienna Ab initio Simulation package
XGBR	eXtreme Gradient Boosting Regression
ZPE	Zero-Point Energy

Chapter 1

Introduction

1.1 Sustainable Energy Conversion Technologies

As the 21st century progresses, the global community faces a twofold imperative: satisfying increasing energy needs while stemming environmental degradation through fossil fuel dependence. Conventional energy systems are not only non-sustainable owing to resource exhaustion but also are significant emitters of greenhouse gases, particularly CO₂, NO₂. Consequently, the energy sector faces a revolutionary transition towards clean and renewable technologies. Some of the most hopeful substitutes include fuel cells, solar systems, hydrogen-based energy systems, thermoelectric, and biomass usage [1]. These systems are picking up speed because of their potential to be environmentally benign, sustainable in the long run, and very high in energy efficiency. At the center of this change is not merely the rollout of current technology but also highlighting the scientific swing in discovering and engineering superior materials capable of powering these systems more efficiently. At the heart of these disciplines lies materials science, which provides the tools and theoretical frameworks to engineer materials at the atomic and nanoscale level, including perovskites in solar cells, Pt-based and other transition metal-based catalysts in fuel cells, 2D materials for spintronics, and metal-organic frameworks (MOFs) for the storage of hydrogen [2,3].

By employing first-principles simulations, data-driven models, and methods such as Density Functional Theory (DFT) and machine learning (ML), researchers can forecast material behavior, model catalytic reactions, and speed up material discovery using high-throughput screening. Such computational evidence has enabled major breakthroughs in renewable energy, such as rational design of catalysts for ORR and HER, band

structure tailoring for photovoltaics, and surface engineering for enhanced hydrogen kinetics [4].

This study presents a computational study of Pt-based nano catalysts for proton exchange membrane fuel cells (PEMFCs) to elucidate and optimize ORR catalysis. It explores strategies such as alloying, atomic utilization, and surface engineering to enhance the performance of the catalysts with the ultimate objective of contributing towards efficient, stable, and large-scale clean energy technologies.

1.2. Comprehensive Review of Fuel Cell Technologies

While renewable energy sources are vital for a sustainable future, their intermittency poses challenges to maintaining a consistent power supply. This has intensified global interest in energy storage and conversion systems such as fuel cells, which generate electricity efficiently through electrochemical reactions without combustion. Fuel cells offer several advantages, including zero emissions, high energy density, and quiet operation, making them suitable for a wide range of applications from vehicles and trains to backup power systems and consumer electronics [5]. They operate by directly converting the chemical energy of hydrogen into electricity, thereby bypassing the inefficiencies associated with combustion engines.

Fuel cells are broadly classified based on their operating temperature and electrolyte composition, which influence their reaction mechanisms, material requirements, and practical applications. These include: (i) low-temperature fuel cells, such as proton exchange membrane fuel cells (PEMFC), alkaline fuel cells (AFC), and direct methanol fuel cells (DMFC), and (ii) high-temperature fuel cells, such as phosphoric acid fuel cells (PAFC), molten carbonate fuel cells (MCFC), and solid oxide fuel cells (SOFC) [6,7]. A comparative overview of their key features, such as electrolyte types, operating temperatures, and fuel compatibility, is summarized in **Table 1.1**.

Table 1.1: Classification of fuel cells and their operational characteristics.

Type of Fuel Cell	Working Temperature	Anodic Reaction	Cathodic Reaction	Applications
Proton Exchange Membrane Fuel Cell (PEMFC)	60 – 80°C	$\text{H}_2 \rightarrow 2\text{H}^+ + 2\text{e}^-$	$\frac{1}{2}\text{O}_2 + 2\text{H}^+ + 2\text{e}^- \rightarrow \text{H}_2\text{O}$	Automobiles Stationary Power
Phosphoric Acid Fuel Cell (PAFC)	160 – 200°C	$\text{H}_2 \rightarrow 2\text{H}^+ + 2\text{e}^-$	$\frac{1}{2}\text{O}_2 + 2\text{H}^+ + 2\text{e}^- \rightarrow \text{H}_2\text{O}$	Stationary Power
Alkaline Fuel Cell (AFC)	60 – 120°C	$\text{H}_2 + 2(\text{OH})^- \rightarrow 2\text{H}_2\text{O} + 2\text{e}^-$	$\frac{1}{2}\text{O}_2 + \text{H}_2\text{O} + 2\text{e}^- \rightarrow 2(\text{OH})^-$	Space Endeavors
Direct Methanol Fuel Cell (DMFC)	50 – 120°C	$\text{CH}_3\text{OH} + \text{H}_2\text{O} \rightarrow 6\text{H}^+ + 6\text{e}^- + \text{CO}_2$	$\frac{1}{2}\text{O}_2 + 2\text{e}^- + \text{CO}_2 \rightarrow \text{CO}_3^{2-}$	Electronic Devices Military Applications
Molten Carbonate Fuel Cell (MCFC)	500 – 650°C	$\text{CO}_3^{2-} + \text{H}_2 \rightarrow \text{CO}_2 + \text{H}_2\text{O} + 2\text{e}^-$	$\frac{1}{2}\text{O}_2 + 2\text{e}^- + \text{CO}_2 \rightarrow \text{CO}_3^{2-}$	Industrial Waste Heat

Solid-Oxide Fuel Cell (SOFC)	600 – 1000 °C	$\text{H}_2 + \text{O}^{2-} \rightarrow \text{H}_2\text{O} + 2\text{e}^-$	$\frac{1}{2} \text{O}^{2-} + 2\text{e}^- \rightarrow \text{O}^{2-}$	Automobile range Extenders
------------------------------	---------------	---	---	----------------------------

1.3. Proton-Exchange Membrane Fuel Cell (PEMFC)

Proton Exchange Membrane Fuel Cells (PEMFCs) are a cornerstone of sustainable energy conversion technologies. Operating efficiently at low temperatures (~60-80 °C), they are suitable for automotive, residential, and portable energy applications. Their high-power density, modular design, and zero on-site carbon emissions position them as a viable replacement for fossil-fuel-driven systems in both transportation and stationary power sectors [6]. At the heart of a PEMFC lies the membrane electrode assembly (MEA), which includes anode and cathode electrocatalyst layers, often made of platinum or Pt-based alloys, a proton-conductive polymer electrolyte membrane such as Nafion™, gas diffusion layers and flow fields to manage reactant transport and product removal [7]. The fundamental operating mechanism of PEMFCs are illustrated in **Figure 1.1**.

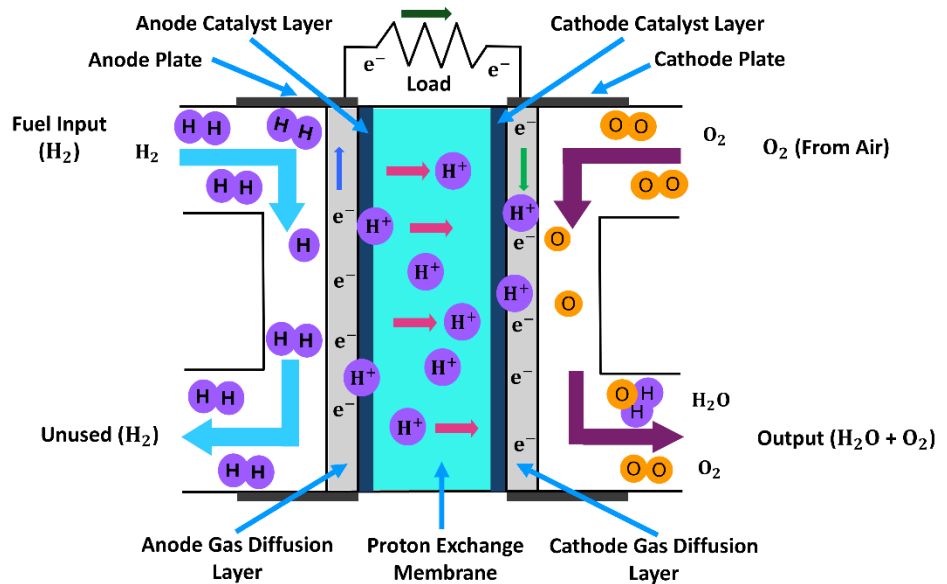


Figure 1.1: Schematic representation of the working mechanism of a PEMFC.

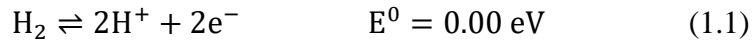
1.3.1. Electrode Reactions

PEMFC typically involves two fundamental electrochemical reactions, taking place at two different electrodes, the anode and the cathode, respectively. The anode is made up of a kind of material that can facilitate the hydrogen oxidation reaction, while the cathode consists of a catalyst which can promote the electrocatalytic reduction of oxygen.

1.3.1.1. Hydrogen Oxidation Reaction (HOR)

The hydrogen oxidation reaction generally facile on Pt surfaces. In PEMFC, at anode hydrogen fuel (H_2) adsorbs and splits into two H atoms on the surface of the Pt catalyst. Each H atom releases a proton and an electron.

Mechanism:

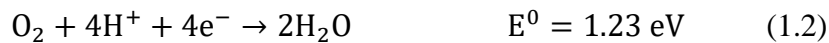


In the above **Equation 1.1**, E^0 represents the half-cell potential for the reaction with reference to the standard reversible hydrogen electrode (RHE).

On high-surface area Pt catalysts, this process is fast and reversible, with exchange current densities. It is essential for achieving the expected efficiency of a PEMFC [8].

1.3.1.2. Oxygen Reduction Reaction (ORR)

At the cathode of a PEMFC, oxygen from air (O_2) undergoes a reduction by combining with protons generated at the anode via HOR, and electrons are delivered through the external circuit. This reaction proceeds according to the following **Equation 1.2**:



This process, known as the oxygen reduction reaction (ORR), is initiated by the adsorption of O₂ molecules onto the surface of the cathode catalyst. Once adsorbed, O₂ can follow multiple mechanistic pathways, each involving distinct intermediates. In the dissociative pathway, O₂ dissociates into two *O, which subsequently undergo stepwise protonation to form *OH and finally H₂O [9]. Alternatively, in the associative pathway, *O₂ can undergo direct protonation to form *OOH. It may either follow the peroxy pathway, dissociating into *O and *OH, or proceed via the peroxide pathway, undergoing further protonation *H₂O₂, which then dissociates into two *OH species. These intermediates ultimately lead to water formation through subsequent protonation steps. A comprehensive mechanistic scheme for ORR, illustrating these pathways, is depicted in **Figure 1.2**. Additionally, a two-electron reduction route may occur, producing H₂O₂ from adsorbed O₂, as shown below in **Equation 1.3**.



While this two-electron process is commercially significant for the synthesis of H₂O₂, it is generally undesirable in fuel cell operation due to the oxidative degradation it causes to the proton-conducting membrane. Despite its importance, it remains the kinetically limiting step in PEMFCs. Challenges such as high activation energy for O₂ adsorption, competition between two- and four-electron pathways, and sluggish overall kinetics necessitate the development of highly active and durable catalysts. Enhancing ORR activity is, therefore, a central goal in catalyst research.

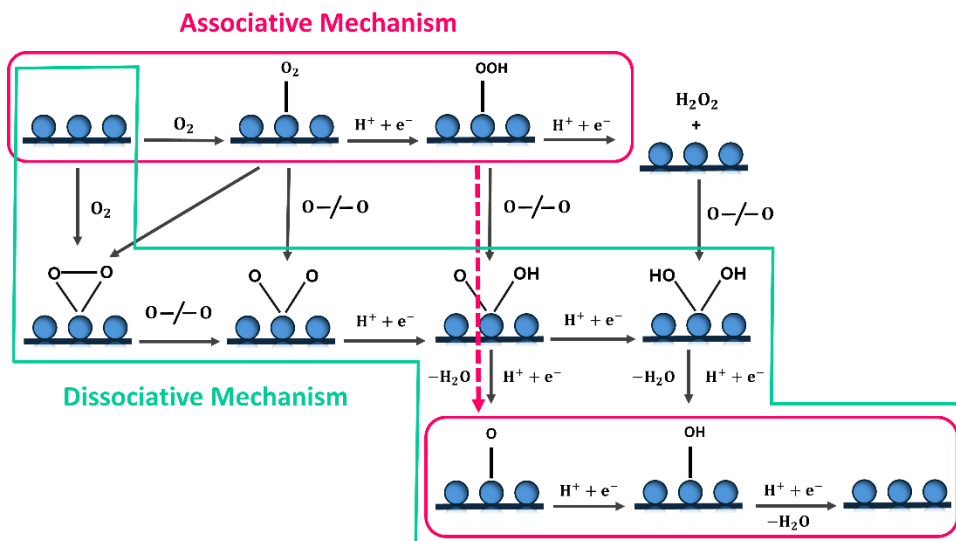


Figure 1.2: Schematic illustration of the Oxygen Reduction Reaction (ORR) mechanism.

1.4. Electrocatalysts

Electrocatalysts are substances that accelerate electrochemical reactions without being consumed in the reaction. They are crucial in energy conversion technologies, especially in processes that include: (a) Oxygen Reduction Reaction (ORR) [10], (b) Hydrogen Evolution Reaction (HER) [11], (c) Oxygen Evolution Reaction (OER) [10], (d) CO_2 Reduction Reaction (CO_2RR) [12], (e) Nitrogen Reduction Reaction (NRR) [13]. These reactions are central to clean energy systems, including PEMFCs, electrolyzers, and metal-air batteries. Based on their phase, electrocatalysts are classified as homogeneous, where catalysts are in the same phase as the reactants (typically liquid), and heterogeneous electrocatalysts, where catalysts are in a different phase than the reactants, usually a solid interacting with liquid or gases.

1.4.1. Heterogeneous Catalysts

Heterogeneous catalysis involves catalysts that exist in a different phase from the reactants. This field is undergoing rapid evolution, with current innovations focused on enhancing catalyst selectivity, recyclability, and

activity under mild reaction conditions. Recent advancements have underscored the emergence of catalyst systems such as bulk (single-phase) catalysts [14], nanocluster catalysts (NCs) [15], single-atom catalysts (SACs) [13], and various supported catalysts [16,17]. These developments are pivotal in driving improvements in catalytic efficiency, selectivity, and sustainability.

1.4.1.1. Nanocluster Catalysts

Nanoclusters occupy a unique category of materials that lie between discrete molecules and bulk solids. In these systems, properties diverge from bulk behavior, leading to distinctive phenomena such as under-coordination, finite-size effects, and pronounced surface-related features, endow nanoclusters with extraordinary optical, electronic, and catalytic capabilities, making them highly versatile in applications like electronics, drug delivery, sensors, and energy catalysis [18,19]. Generally, nanoclusters range from 0.2 to 2 nm in diameter, contrasting with larger nanoparticles (5-100 nm) and demonstrating unique behaviors due to their reduced dimensions. Their small size grants advantages such as minimal metal usage, enhanced surface-to-volume ratios, and adjustable surface configurations, which contribute to their catalytic efficiency [20]. Metals that are typically inert in their bulk state, such as gold, become catalytically active when downsized to the nanocluster regime [21]. This behavior is driven by their unique surface atom arrangement, discrete energy states, and the emergence of new active sites. Nanoclusters thus present an opportunity to tailor material properties to enhance catalytic performance. For example, bimetallic or alloy nanoclusters can be engineered to adjust surface atom distributions for better interaction with reaction intermediates [22]. Additionally, core-shell nanoclusters can dramatically reduce the demand for precious metals like platinum by incorporating a cheaper metal core. The relationship between cluster size and catalytic activity can be exploited to determine the optimal dimensions for peak performance. Altering the

shape or atomic configuration of nanoclusters also modulates their activity, evident in the ORR performance shift observed when Pt clusters transition from cuboctahedral to octahedral geometries at equal loadings [23].

Recent research has highlighted the exceptional behavior of subnanometer clusters (<1 nm), which often surpass the activity of their larger counterparts due to their "*every atom counts*" principle. These clusters exhibit extreme structural adaptability, allowing them to interconvert among several active geometries under reaction conditions. Various external influences, such as the nature of the support, ligand environment, and proximity effects, also play vital roles in governing nanocluster activity [24]. For instance, manipulating electronic metal-support interaction (EMSI) can align energy levels for optimal intermediate binding. Ligand-protected nanoclusters offer tunability through modifications in ligand identity, number, and spatial distribution [25].

Computational studies are indispensable for exploring nanocluster catalysis due to the challenges of precisely synthesizing atomically defined clusters. A robust theoretical framework must therefore integrate local atomic structure, electronic deviations from bulk behavior, and reaction dynamics. This demands comprehensive modeling strategies, including potential energy surface mapping and kinetic modeling tailored to coordinatively unsaturated environments. Finally, simulation approaches must also consider the multidimensional nature of catalytic systems, factoring in intermediate stability, reaction conditions, and support effects to accurately predict nanocluster behavior and guide rational catalyst design.

Chapter 2

Review of Past Work and Problem Formulation

The global transition toward sustainable energy technologies is driven by the progressive depletion of fossil fuel reserves and the imperative to mitigate environmental degradation [26]. This paradigm alteration has accelerated research and development efforts focused on harnessing renewable energy sources and enhancing the efficiency of energy conversion and storage systems, fostering a more sustainable energy infrastructure. In this context, electrocatalysts play a crucial role, as they form the foundation of numerous cutting-edge energy technologies, such as hydrogen production systems [27], photovoltaic devices [28], and fuel cells [29], like proton exchange membrane fuel cells (PEMFCs). Among these, PEMFCs have emerged as a highly promising clean energy technology to meet the growing global energy demand, offering high efficiency, low operating temperatures, and zero-emission output [30]. The oxygen reduction reaction (ORR) serves as a central electrochemical process at the cathode of PEMFCs. Despite its significance, the ORR is inherently sluggish owing to its complex multi-electron transfer mechanism and high activation energy barrier, which substantially constrains overall performance and widespread industrial deployment of PEMFC technology [31]. Therefore, expensive noble metals such as platinum (Pt) have been extensively investigated owing to their exceptional catalytic activity and stability in the acidic environment of PEMFCs. In recent years, Subnanoclusters (SNCs) have emerged as a frontier in catalyst design, receiving growing attention [32,33]. Unlike larger nanoparticles [34], SNCs exhibit a relatively shallow and anharmonic potential energy surface (PES) at finite temperatures, resulting in high structural flexibility and dynamic rearrangements, deviating from Arrhenius-type kinetic behavior [35]. This cluster dynamism facilitates the exploration of a wide ensemble of low-energy metastable surface states [36,37]. Moreover, the synergistic

effect of high electronic accessibility and under-coordinated metal atoms in SNCs fundamentally possesses electronic structure reconstruction, resulting in enhanced orbital overlap with reactant molecules and modified charge distribution at active sites [38]. Recent investigations have revealed that platinum subnano clusters (Pt SNCs), when anchored onto highly conductive substrates such as graphene or indium tin oxide (ITO), can catalyze ORR at rates exceeding those observed with conventional nanostructured platinum catalysts [39,40]. An intriguing aspect of Pt SNCs is their capacity to overcome conventional scaling relations that dictate the adsorption energetics of ORR intermediates [41]. These constraints, intrinsic to volcano plot frameworks, hinder the independent optimization of reaction steps, thus capping the theoretical limit on catalytic efficiency [42]. Furthermore, under reaction conditions involving elevated temperatures, reactive gases, and fluctuating chemical environments, catalytic interfaces undergo continuous restructuring [43]. These dynamic transformations give rise to a spectrum of transient structural and stoichiometric states, as the dynamic cluster catalysts should be viewed not as single static thermodynamic states, but as statistical ensembles of numerous metastable surface states [37]. Each of these states can participate in the catalytic process, contributing collectively to the observed activity, selectivity, and stability under operando conditions [36,41,42]. In electrochemical processes, the catalytic interface is highly responsive to the accumulation of adsorbed species. Borna et al., using first-principles calculations, demonstrated that for size-selected supported Pt_n clusters ($n = 1-6$) in the subnanometer regime, adsorbate coverage plays a critical role in modulating the catalytic activity of SNCs under gas-phase conditions [41]. Additionally, recent work by Zhang et al., based on first-principles calculations, highlights that the hydrogen evolution reaction (HER) activity on the WB (001) surface is significantly affected by surface restructuring [43]. Furthermore, Roldan et al. reported significant structural rearrangements of Cu NPs under CO_2 electro-reduction conditions,

associated with the dynamic cluster interfaces [44]. Despite significant advances, a comprehensive investigation into coverage-dependent activity of TMSNCs of varying sizes on the ORR remains unexplored. Moreover, the design of active catalysts and the elucidation of the relationship between their catalytic performance and descriptors, as well as how the high coverage of adsorbate, which is developed under ORR ambient conditions of subnanometer clusters, remains largely unresolved.

The screening of active electrocatalysts is traditionally a time-intensive and costly endeavor, spanning from experimental design to commercial deployment. Recently, the integration of theoretical approaches, particularly density functional theory (DFT) and machine learning (ML), has gained significant traction in the exploration of promising catalytic materials. This combined methodology has proven effective in accelerating the catalyst development and extraction while reducing research and development expenditures [45-51]. DFT remains one of the most widely used quantum-mechanical simulation techniques, offering deep insights into the properties of known materials and enabling the prediction of characteristics for novel candidates [52,53]. However, it is important to recognize that DFT alone may not be sufficient for efficiently screening a large number of potential catalysts or for identifying the key factors that govern catalytic performance. Moreover, the design space of Pt SNCs is extraordinarily vast, making it impractical to evaluate all potential candidates through experimental or purely theoretical approaches within a short timeframe and limited cost. ML, a powerful statistical tool, leverages algorithms to map input features to target properties. By utilizing data design techniques, ML can uncover complex relationships between catalyst attributes, such as geometric and electronic structures, and their performance [54]. Additionally, the rapid performance prediction across a wide range of candidates significantly accelerates the catalyst screening process.

In this study, we develop a ML framework to systematically screen active electrocatalysts for the ORR and the uncertainty quantification in the subnanometer regime, focusing on a size-selected chemical space of Pt_n ($n = 7-13$) clusters. By employing ML algorithms, we predict adsorption energies and establish correlations between catalytic activity and the underlying geometric and electronic characteristics of local atomic environments. Our approach is highlighted by the observed size-dependent shift in the apex of the volcano plot across different isomers of Pt SNCs, revealing intricate structure-activity relationships unique to this regime.

Chapter 3

Theoretical Methods

Electronic structure calculation methods have emerged as robust and unswerving tools for investigating the properties of molecular and solid-state systems by providing insights into the interactions between electrons and nuclei. Over the past few decades, these computational techniques have been extensively applied to gain fundamental insights into homogeneous and heterogeneous catalytic systems. In the following section, we provide a concise overview of quantum mechanical frameworks employed throughout our work.

3.1. Schrödinger Equation

3.1.1. Many-Body Problem

The time-dependent Schrodinger equation (TDSE) is utilized to provide a description of the electronic structures and intrinsic molecular properties of a solid-state system based on solving the system's Hamiltonian. Additionally, TDSE is the most general form of the Schrödinger equation that can be used to explain the temporal evolution of quantum mechanical systems. This might be gained by solving the Schrödinger equation, which directs the quantum behavior of particles, and it can be illustrated as shown in **Equation 3.1**.

$$\hat{H}\Psi(\vec{r}, t) = i\hbar \frac{\partial}{\partial t} \Psi(\vec{r}, t) \quad (3.1)$$

$$\hat{H} = -\frac{\hbar^2}{2m} \nabla^2 + V(\vec{r}) \quad (3.2)$$

In **Equation 3.1**, \hat{H} refers to Hamiltonian operator consisting of both kinetic energy operators (\hat{T}) and potential energy operators (\hat{V}), $\Psi(\vec{r}, t)$ is the time-dependent wave function demonstrates the quantum state of the system, $i =$

$\sqrt{-1}$, \hbar is the reduced Planks constant ($\hbar = \frac{h}{2\pi}$), m is the particle mass, and in **Equation 3.2**, ∇^2 is the Laplacian operator (second derivative with respect to space), acts on spatial coordinates. The wave function encodes all the information about the system. Despite the innocent look of the Schrödinger wave equation, solving it and calculating the systems properties is a very difficult task. It works nicely for simple systems such as an electron in well, hydrogen, or helium (although not exact). However, we are talking about materials up to numerous atoms that contain several thousand electrons. Then the calculations for these n -electron systems are completely out of our imagination. This is called “many-body problem” or we can say that n -electron problem. Dealing with n -electrons that interact with all the other electrons of that system is too complex to solve even numerically. For practical purposes in electronic structure theory, the time-independent Schrödinger equation (TISE) can be written is as follows (**Equation 3.3**)-

$$\hat{H}\Psi(\vec{r}) = E\Psi(\vec{r}) \quad (3.3)$$

In **Equation 3.3**, E indicates the energy eigenvalue (a constant for a stationary state), the time-independent wavefunction $\Psi(r)$, depends on the position. In 1D, it can be illustrated as (**Equation 3.4**):

$$-\frac{\hbar^2}{2m} \frac{d^2\Psi(x)}{dx^2} + V(x)\Psi(x) = E\Psi(x) \quad (3.4)$$

For a system containing multiple electrons and nuclei, the many electrons Hamiltonian becomes significantly complex. Therefore, TISE can be expressed as follows (**Equation 3.5**)-

$$\Psi = \Psi(r_1, r_2, \dots, r_n, R_1, R_2, \dots, R_N) \quad (3.5)$$

Where, r_1, r_2, \dots, r_n and R_1, R_2, \dots, R_N indicate the coordinates of n electrons and the coordinates of N nuclei, respectively. Therefore, for this type of complex system, the total Hamiltonian consisting of kinetic and

potential energy of the many-body Schrödinger equation can be illustrated as shown in **Equation 3.6**.

$$\begin{aligned}\hat{H} = & \sum_i^n -\frac{\hbar^2}{2m_e} \nabla_i^2 + \sum_I^N -\frac{\hbar^2}{2M_I} \nabla_I^2 + \frac{1}{2} \sum_{i \neq j} \frac{e^2}{4\pi\epsilon_0} \frac{1}{|r_i - r_j|} \\ & + \frac{1}{2} \sum_{I \neq J} \frac{e^2}{4\pi\epsilon_0} \frac{Z_I Z_J}{|R_I - R_J|} - \sum_{i,I} \frac{e^2}{4\pi\epsilon_0} \frac{Z_I}{|r_i - R_I|}\end{aligned}\tag{3.6}$$

The right-hand side terms of **Equation 3.6** represent the kinetic energy of the electrons, the kinetic energy of nuclei, the Coulombic interactions between electrons and nuclei, the Coulombic interelectronic repulsion, and the Coulombic internuclear repulsion, respectively. Additionally, the Z_I denote the number of protons in the I^{th} atom. The m_e and M_I are the mass of the i^{th} electron and I^{th} nuclei, respectively.

As I stated, the exact solution of the Hamiltonian in the Schrödinger equation (**Equation 3.6**) is feasible only for simple systems, such as hydrogen-like atoms. However, for many-body systems like solids consisting of numerous interacting atoms, an exact analytical solution becomes intractable. Therefore, various approximation methods have been developed to achieve approximate solutions to describe the quantum behavior of systems effectively.

3.1.2. Born-Oppenheimer (BO) Approximation

The Born-Oppenheimer approximation, also known as the clamped nuclei approximation [55], is a fundamental principle in quantum mechanics that simplifies the complex many-body Schrödinger equation of molecules. Within this approximation, nuclei can be treated as nearly stationary compared to the electron dynamics due to their substantial mass difference. It decouples the nuclear and electronic motions, allowing the total

wavefunction to be approximated as a product of wavefunctions of electrons and nuclei, as given in **Equation 3.7**.

$$\Psi(r_1, r_2, \dots, r_n, R_1, R_2, \dots, R_N) = \Psi(r_1, r_2, \dots, r_n) \Psi(R_1, R_2, \dots, R_N) \quad (3.7)$$

Considering the motion of nuclei independent of electronic motion, the kinetic energy of the nuclei and the potential energy for internuclear repulsion can be neglected without a substantial deficit of accuracy. With these considerations, the many-body Schrödinger equation (**Equation 3.6**) can be simplified and expressed as in **Equation 3.8**:

$$\left[\sum_i \frac{\nabla_i^2}{2} - \sum_{i,l} \frac{Z_l}{|r_i - R_l|} + \frac{1}{2} \sum_{i \neq j} \frac{1}{|r_i - r_j|} \right] \Psi(r, R) = E \Psi(r, R) \quad (3.8)$$

Here, in **Equation 3.8**, the Hamiltonian includes terms representing the kinetic energy of the electrons, their interaction with the relatively static nuclei (electron-nuclear coulombic attraction), and the interelectronic repulsion. By applying the Born–Oppenheimer approximation, the total degrees of freedom in the system are effectively reduced, allowing the focus to shift solely to the electronic problem. Nevertheless, solving the electronic Schrödinger equation remains a formidable task, due to the complexity introduced by electron-electron repulsion in systems. As a more practical approach, it is often convenient to perform with the electron density rather than tracking the coordinates of individual electron. In the upcoming section, we depict Density Functional Theory (DFT) framework, which reformulates the many-body Hamiltonian as a functional of the electron density, thereby avoiding the direct use of the complex many-electron wavefunction.

3.1.3. Mean-field Approximation

In the mean-field approximation, the complex many-body interactions between electrons are approximated by assuming that each electron moves independently in an average, effective electrostatic potential generated by

all other electrons. This approximation transforms the inherently many-body problem into a tractable one-electron problem. This effective potential arises from the overall electron density distribution and is commonly referred to as the Hartree potential, denoted as $V_H(r)$. It represents the self-consistent coulomb interaction experienced by an electron at position r , resulting from the smeared-out charge density of all other electrons in the system. Mathematically, the Hartree potential is illustrated as (**Equation 3.9**):

$$V_H(r) = \int \frac{n(r')}{|r-r'|} dr' \quad (3.9)$$

In the above **Equation 3.9**, $n(r')$ is the electron density at position r' , $|r - r'|$ is the distance between the observation point r and the source point r' , and the integral is taken over all space, accounting for contributions from the entire electron distribution. The many-body Schrödinger equation can be approximated by decomposing it into single-electron equations. These are known as the Hartree equations, or the mean-field equations for independent particles, which describe the motion of individual electrons under the influence of an effective average potential created by all other electrons in the system. The Hartree equation for a single electron can be expressed as shown in **Equation 3.10**.

$$\left[\sum_i \frac{\nabla_i^2}{2} - \sum_I \frac{Z_I}{|r_i - R_I|} + V_H(r) \right] \phi_i(r) = \epsilon_i \phi_i(r) \quad (3.10)$$

In **Equation 3.10**, $\phi_i(r)$ is the single-electron orbital wavefunction, ϵ_i is the corresponding orbital energy. Due to the inherently nonlinear nature of this formulation, the Hartree potential $V_H(r)$ depends on the electron density, which in turn is a function of the orbitals $\phi_i(r)$, solving the Hartree equations requires a self-consistent field (SCF) approach. Typically, these equations are solved iteratively using numerical methods until electron density convergence is reached. Once the self-consistent set of orbitals

$\phi_i(r)$ is obtained, the total electron density $n(r)$ can be computed using **Equation 3.11**.

$$n(r) = \sum_i |\phi_i(r)|^2 \quad (3.11)$$

Equation 3.11 sums up the squared magnitudes of all occupied single-electron orbitals, capturing the spatial distribution of the electron cloud. Furthermore, the approximations and computational techniques are often employed to reduce the complexity and improve the accuracy of the calculations. Notably, the Hartree approach does not account for electron correlation effects, arising from instantaneous interactions between electrons, as it treats electrons as moving independently in the average field of others. Despite this limitation, the Hartree equation provides a foundational framework for understanding many-electron systems and serves as a stepping stone for more accurate methods, such as the Hartree-Fock (HF) method and Density Functional Theory (DFT).

3.1.4. Hartree-Fock (HF) Method

The Hartree-Fock (HF) theory incorporates self-consistent field (SCF) methods to demonstrate the electronic structure of a quantum system using a single-reference Slater determinant. Within this framework, each electron is treated as moving independently in the average field conceived by all other electrons.

3.1.4.1. Hartree Approximation: Independent Particle Model

The Hartree approximation assumes that the total wavefunction of a multi-electron system can be expressed as a product of individual one-electron wavefunctions (**Equation 3.12**):

$$|\Psi(r_1, r_2, \dots, r_N)\rangle \approx \Psi_1(r_1)\Psi_2(r_2) \dots \Psi_N(r_N) \quad (3.12)$$

However, the above **Equation 3.12** fails to satisfy the antisymmetry requirement imposed by the Pauli exclusion principle for fermions, which

requires that the total wavefunctions changes sign upon exchanging any two electrons.

3.1.4.2. Antisymmetry and Slater Determinant

To enforce this, the wavefunction must be antisymmetric under the exchange of any two electrons. For a two-electron system, this can be achieved using (**Equation 3.13**):

$$\Psi(x_1, x_2) = \frac{1}{\sqrt{2}} [\chi_1(x_1)\chi_2(x_2) - \chi_1(x_2)\chi_2(x_1)] \quad (3.13)$$

For an N-electron system, **Equation 3.13** can be generalized by a Slater determinant (**Equation 3.14**):

$$\Psi_{el} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \varphi_{(1)} & \varphi_{(2)} & \varphi_{(3)} \\ \varphi_{(2)} & \varphi_{(2)} & \varphi_{(2)} \\ \dots & \dots & \dots \\ \varphi_{(N)} & \varphi_{(N)} & \varphi_{(N)} \end{vmatrix} \quad (3.14)$$

3.1.4.3. Energy Hamiltonian

Therefore, the Hamiltonian for an interacting electron system is expressed by **Equation 3.15**.

$$\hat{H} = \hat{H}^c + \frac{1}{2} \sum_{i \neq j} \frac{e^2}{|r_i - r_j|} \quad (3.15)$$

In above **Equation 3.15**, the central Hamiltonian (\hat{H}^c) component offering an exact solution is given by **Equation 3.16**.

$$\hat{H}^c = -\frac{\hbar^2}{2m_e} \sum_i \nabla_i^2 - \sum_{i,l} \frac{Z_l e^2}{|r_i - R_l|} \quad (3.16)$$

Despite advances computational methods, the electron-electron interaction term introduces a major source of complexity as it couples all electrons and makes the exact solution intractable.

3.1.4.4. The Self-Consistent Field (SCF) Method

The HF method introduces the Fock operator \hat{f}_i , which acts on each orbital (**Equation 3.17**):

$$\hat{f}_i \chi_i = \epsilon_i \chi_i \quad (3.17)$$

In **Equation 3.17**, χ_i is the spin-orbital eigenfunctions, correspond to the orbital energy eigenvalue of the Fock operator \hat{f}_i , respectively. Now, the Fock operator is composed of the core Hamiltonian and the HF potential, expressed as given in **Equation 3.18**.

$$\hat{f}_i = \hat{H}^C + V_{HF}(i) \quad (3.18)$$

3.1.4.5. Hartree- Fock Potential

The HF potential $V_{HF}(i)$ approximates the average interelectronic interaction as shown in **Equation 3.19**.

$$V_{HF}(i) = \sum_{j,l} \frac{Z_l e^2}{|r_i - R_l|} \quad (3.19)$$

The HF potential can be written in terms of the Coulomb operator J_i and the exchange operator K_i , as demonstrated by the following **Equation 3.20**.

$$V_{HF}(i) = \sum_j [J_i(x_i) - K_i(x_i)] \quad (3.20)$$

The term $J_i(x_i)$ represents the classical Coulomb repulsion between electron i and electron j, while $K_j(x_i)$ represents the exchange interaction due to antisymmetry of the wavefunction.

While HF theory extends potentially a good first approximation to the electronic structure, However, it neglects electronic correlation beyond the average field. The computational cost increases rapidly with system size. Each electron is described using three spatial coordinates and one spin coordinate, making the wavefunction increasingly complex for larger

systems. As a result, applying HF theory to more intricate molecular systems introduce significant computational challenges.

3.2 Density Functional Theory

The quantum mechanical method density functional theory (DFT) is based on the electron density of a system to investigate the electronic structure of atoms, molecules, and condensed matter systems. Unlike traditional wavefunction-based methods, which become computationally intensive for systems with many electrons, DFT reformulates the many-body problem in terms of the electron density, a function of just three spatial variables. The origin of DFT was introduced by Thomas and Fermi, known as the Thomas-Fermi Theory. This theory treated electron density as a non-interacting homogeneous electron gas and expressed the total energy as a functional of the electron density. Additionally, this theory had significant limitations as it could not accurately account for exchange-correlation effects and lacked the ability to demonstrate chemical bonding, particularly in systems with rapidly varying electron densities such as atoms and molecules. In 1964, Pierre Hohenberg and Walter Kohn amplified DFT further, instituting the skeleton of the two Hohenberg-Kohn (HK) Theorems [56,57]. These theorems laid the groundwork of reformulating the many-body problem of electron density rather than the many electron wave functions.

3.2.1. Electron Density in DFT

In DFT, the vital assumption is that our reference system is noninteracting in nature. That means the electrons do not interact with each other. Therefore, the electron density can be expressed as a sum of the squared magnitudes of occupied non-interacting orbitals ϕ_i , as shown in **Equation 3.21**.

$$\rho(r) = \sum_i |\phi_i(r)|^2 = 2 \sum_i^{\text{occ}} |\phi_i(r)|^2 \quad (3.21)$$

Note that the usual wave functions, ψ_i are substituted by the orbitals ϕ_i , indicating that are now the KS orbitals of that system. Integrating the electron density over the entire space resulting into the total number of electrons, n (**Equation 3.22**):

$$\int \rho(r) dr = n \quad (3.22)$$

If we know the electron density of an atom is known, can be used as initial approximation to construct the electron density of the solid system composed of that atom. Additionally, the electron density in a system not only encapsulates information about the wavefunction, orbitals, and the total number of electrons, but is also fundamentally linked to the potentials, energies, and consequently, all physical properties of the system.

3.2.1 Hohenberg-Kohn Theorems

The Hohenberg-Kohn theorems are two fundamental theorems, demonstrated electron density as the central variable in DFT. They also formalized the interdependence between the electron density, external potential, Hamiltonian, and the many-body wave function.

Theorem I – Existence Theorem

Statement:

“For any system of interacting particles in an external potential $V_{ext}(r)$, the external potential is uniquely determined, up to a constant, by ground-state electron density, $\rho_0(r)$.”

Implication:

This highlights that all the ground-state properties (including the many-body wave functions and energy) are functionals of the electron density $\rho_0(r)$. It implies an extensive relationship between the ground-state electron density and the ground-state energy. In other words:

$$\rho(r) \Rightarrow V_{ext}(r) \Rightarrow \hat{H} \Rightarrow \psi_0 \Rightarrow E_0$$

Theorem II – Variational Principle

Statement:

“A universal functional for the energy can be defined in terms of electron density within in an external potential, and the exact ground-state electron density minimizes this functional.”

Implication:

For a given external potential V_{ext} , if one systematically minimizes the total energy of the system with respect to the electron density by using variation principle, the minimum energy obtained corresponds to the true ground-state energy. Crucially, this minimum is no physically realizable density can be generated an energy below it. The electron density that achieves this minimum is the ground-state electron density, denoted by **Equation 3.23**.

$$E_{\text{HK}}[n] = F_{\text{HK}}[n] + \int dr V_{\text{ext}}(r)n(r) \quad (3.23)$$

Where the internal energies consisting of kinetic and potential energies, are expressed in terms of F_{HK} in **Equation 3.23**. The total internal energy functional can be written as follows (**Equation 3.24**)-

$$E_{\text{HK}}[n] = T[n] + E_{\text{int}}[n] \quad (3.24)$$

Interestingly, **Equation 3.24** is independent of the external potential (V_{ext}) and solely depends upon the density of electrons. The total internal energy is the sum of the kinetic energy functional ($T[n]$) and the internal energy functional ($E_{\text{int}}[n]$). This indicates that the ground state energy of a many-body system is uniquely determined by the electron densities. However, the precise form of the density functional remains undefined, necessitating additions, extended refinement, and approximation.

3.2.2 Kohn-Sham Formalism

The Hohenberg-Kohn (HK) theorems offer a foundational framework for addressing the many-body problem by utilizing the particle density function and the variational principle. However, for practical applications involving particles, the Density Functional Theory, as realized through the Kohn-

Sham (KS) approach, is more commonly used. The central concept supporting the HK theorems is to substitute the interacting electron system with an auxiliary system of non-interacting particles that possess an identical electron density distribution. This allows the total energy functional to be expressed as shown in **Equation 3.25**.

$$E[\rho(r)] = T_0[\rho(r)] + \frac{1}{2} \iint \frac{\rho(r)\rho(r') dr dr'}{|r-r'|} + \int V_{\text{ext}}(r)\rho(r) dr + E_{\text{xc}}[\rho(r) dr] \quad (3.25)$$

In **Equation 3.25**, the kinetic energy functional of a non-interacting electron gas system is represented by $T_S[n]$. The external potential is expressed as a contribution due to nuclei and another external potential $(\int dr V_{\text{ext}}(r)n(r))$, and the classical Coulomb potential for the interelectronic interaction is expressed as $(\frac{1}{2} \int dr dr' \frac{n(r)n(r')}{|r-r'|})$ which is called the Hartree potential. The final term, $E_{\text{xc}}[n]$, in the expression captures all the many-body effects arising from exchange and correlation interactions, which is known as the exchange-correlation functional. The exact analytical expression of these exchange-correlation functionals ($E_{\text{xc}}[n]$) is yet to be determined. In **Equation 3.6**, the coulomb repulsion term due to internuclear repulsion is contributed directly as a constant term in the final energy expression. According to the second H-K theorem, the solution for the Kohn-Sham (KS) auxiliary systems can be obtained by minimizing the KS energy functional with respect to the electron density $[n(r)]$. This minimization of the total energy is achieved by employing a Schrödinger-like equation, as presented in **Equation 3.26**.

$$\left[-\frac{1}{2} \nabla^2 + V_{\text{eff}}(r) \right] \psi_i(r) = E_i \psi_i(r) \quad (3.26)$$

where $\Psi_i(r)$ corresponds to the Kohn-Sham orbital, ε_i are the eigenvalues corresponding to the energy Hamiltonian, and V_{KS} is the effective potential of the system as defined in **Equation 3.27**.

$$V_{\text{eff}} = V_{\text{Hartree}} + V_{\text{ext}} + V_{\text{xc}} \quad (3.27)$$

Here, V_{xc} defines the exchange-correlation potential as shown in **Equation 3.28**.

$$V_{\text{xc}} = \frac{\delta E_{\text{xc}}[n]}{\delta n(r)} \quad (3.28)$$

The $\Psi_i(r)$ and Kohn-Sham orbitals do not represent the wave functions of electrons. These orbitals lack any direct physical significance for the system. The auxiliary functions are used to compute the electron density, as defined in **Equation 3.29**.

$$n(r) = \sum_i |\Psi_i(r)|^2 \quad (3.29)$$

The Kohn-Sham formalism can accurately determine the ground state of a system with many interacting particles, as long as the right expression for the exchange-correlation energy ($E_{\text{xc}}[n]$) is known. It should be noted that the effective potential of the system depends on the electron density (**Equation 3.10**). Hence, it is necessary to solve the KS equations in a self-consistent manner using an iterative approach. Ultimately, the self-consistent solution guarantees the attainment of the accurate ground-state density [58].

3.3 Exchange-Correlation Functionals

Transition from the wavefunction-based picture to the density-based picture, it is essential to incorporate the effects of electron–electron interactions, especially the quantum effects that cannot be described by classical electrostatics. The exchange-correlation functional $E_{\text{xc}}[\rho]$, captures all the missing mechanical effects in DFT that are not included in the following three parts:

1. Kinetic energy of non-interacting electrons ($T_s[\rho]$),
2. External potential energy (interactions with nuclei, $E_{\text{ext}}[\rho]$),
3. Hartree energy ($E_{\text{H}}[\rho]$, classical electron-electron Coulomb interactions).

In the Kohn-Sham DFT, the total energy functional can be illustrated as in **Equation 3.30**.

$$E[\rho] = T_s[\rho] + \int V_{\text{ext}}(r)\rho(r) dr + E_H[\rho] + E_{\text{xc}}[\rho] \quad (3.30)$$

Here $E_{\text{xc}}[\rho]$ is the correction term, can be separated into two distinct components: exchange energy part (due to Pauli exclusion principle) and correlation energy part (due to dynamic electron-electron correlations), **Equation 3.31**.

$$E_{\text{xc}}[\rho] = E_x[\rho] + E_c[\rho] \quad (3.31)$$

The terms $E_x[\rho]$ and $E_c[\rho]$ represent the exchange and correlation of that system. These non-classical effects are not captured by the first three terms, and they are grouped together and approximated by **Equation 3.32**.

$$E_{\text{xc}}[\rho] = (T[\rho] - T_s[\rho]) + (V_{\text{ee}}[\rho] - E_H[\rho]) \quad (3.32)$$

where $T[\rho]$ is the true kinetic energy, $V_{\text{ee}}[\rho]$ is the full electron-electron interaction energy, $T_s[\rho]$ and $E_H[\rho]$ are approximations. In the subsequent section, we will delve into several widely implemented approximations for the exchange-correlation functional in DFT calculations, including the Local Density Approximation (LDA) and the Generalized Gradient Approximation (GGA). These approaches are designed to progressively enhance the accuracy of exchange-correlation functionals by accounting for varying degrees of quantum mechanical interactions.

3.3.1 The Local Density Approximation (LDA)

Introduced as the pioneering approximation within the Kohn-Sham (KS) formalism during its initial development, the Local Density Approximation (LDA) laid the foundation for subsequent refinements. In this approach, the exchange-correlation energy density has been considered as a homogeneous electron gas [59,60]. The uniform electron gas model is employed due to its incorporation of the most fundamental form of the exchange-correlation functional, which has proven remarkably effective for various metallic

systems. The Local Density Approximation can be mathematically represented as shown in **Equation 3.33**.

$$E_{xc}^{LDA} = \int \rho(r) \epsilon_{xc}^{hom}(\rho(r)) d^3r \quad (3.33)$$

where ϵ_{xc}^{uni} represents the exchange-correlation energy functional for a uniform electron density $n(r)$ calculated at a distance r . This ϵ_{xc}^{uni} can further be sliced into two counterparts, which are exchange (ϵ_x) and correlation (ϵ_c) terms respectively. The exchange (ϵ_x) part is obtained from an analytical methodology, but the exact part of the correlation (ϵ_c) part is yet to be discovered. The LDA formalism reported working quite well in several model systems with slowly changing densities, such as the free electrons in metallic systems. There are some limitations associated with the LDA formalism:

- (i) The calculated cohesive and binding energy values are overestimated using this correlation functional.
- (ii) LDA is unsuitable for working with diffused d and f orbitals, unlike s and p orbitals, which are relatively localized.
- (iii) The long-range interactions (i.e., van der Waals interactions) cannot be addressed due to the local nature of the LDA formalism.

3.3.2 The Generalized-Gradient Approximation (GGA)

The systems exhibiting significant inhomogeneities in electron density, the LDA exchange-correlation formalism is generally inadequate. Therefore, GGA was developed by Hohenberg and Kohn, extending the LDA by incorporating the gradient of the electron density when calculating the exchange-correlation functional. Therefore, the exchange-correlation energy (ϵ_{xc}) per atom, is formulated as a functional that depends not only local electron density, $\rho(r)$, but also on its spatial gradient, $\nabla\rho(r)$. The GGA can be mathematically expressed by **Equation 3.34**.

$$E_{xc}^{GGA} = \int \rho(r) \varepsilon_{xc}^{GGA}(\rho(r), \nabla\rho(r)) d^3r \quad (3.34)$$

In **Equation 3.34**, ε_x^{uni} denotes the exchange energy density functional for a homogeneous electron gas with an electron density equal to $\rho(r)$. In contrast, the enhancement factor, F_{xc} is the function of both the local electron density $\rho(r)$ and its gradient $\nabla\rho(r)$, the latter being a dimensional quantity that accounts for spatial inhomogeneity. The F_{xc} can be decomposed into exchange and correlation contributions. Several exchange functionals have been proposed within this framework, among which the Becke (B88), LYP, and Perdew-Burke-Ernzerhof (PBE) functionals are extensively adopted due to their reliable performance across various systems. GGA, by incorporating the gradient of the electron density, generally provides a lower exchange-correlation energy than LDA, resulting in improved agreement with experimental binding energies, although in some instances it may lead to underestimation of binding strength. Moreover, the GGA formalism demonstrates a significant advancement over LDA by addressing its limitations in systems with non-uniform electron densities. Nevertheless, GGA still faces intrinsic challenges in accurately describing long-range dispersion interactions, which remain beyond the scope of conventional semi-local approximations [61-63].

3.4. Projector Augmented Wave (PAW) Method

The quantum wavefunctions of core and valence electrons display markedly different characteristics. Core electrons typically exhibit rapidly fluctuating wave patterns, whereas valence electrons tend to show much smoother behavior. Standard basis sets, such as those built from plane waves, are generally adequate for representing valence electron states. However, these sets often fail to capture the intricate nature of core electron wavefunctions with sufficient accuracy. To overcome this shortfall, the Projector Augmented Wave (PAW) method is introduced. This method employs a

partial wave expansion specifically in regions close to the atomic core, allowing for a more detailed and accurate reconstruction of both core and valence electron wavefunctions [64-67].

The core concept of PAW lies in using a linear transformation operator that relates the true all-electron wavefunction Ψ_n to a smoother pseudo wavefunction $\widetilde{\Psi}_n$. This relationship is expressed as:

$$|\Psi_n\rangle = T|\widetilde{\Psi}_n\rangle \quad (3.37)$$

Both the true Ψ_n and pseudo $\widetilde{\Psi}_n$ wavefunctions can be expanded using a linear combination of partial waves:

$$|\Psi_n\rangle = \sum_i c_i |\phi_i\rangle \quad (3.38)$$

$$|\widetilde{\Psi}_n\rangle = \sum_i c_i |\tilde{\phi}_i\rangle \quad (3.39)$$

The transformation operator T is defined by the following expression:

$$T = 1 + \sum_i (|\phi_n\rangle - |\tilde{\phi}_n\rangle) \langle \tilde{p}_i | \quad (3.40)$$

Here, $\langle \tilde{p}_i |$ is a projection operator, which is central to this methodology. It enables the PAW approach to capture the essential physics of core electrons without needing to resolve their oscillations directly. By translating the complex core behavior into smoother pseudo wavefunctions, the PAW technique effectively simplifies calculations while retaining high accuracy. This framework is particularly valuable in materials science and solid-state physics due to its ability to handle the electronic structure with precision. Furthermore, enhancements such as ultra-soft pseudopotentials and elements from the linear augmented plane-wave (LAPW) technique have further improved the performance of the PAW method [64].

3.4. Dispersion Corrected Density Functional Theory

Traditional density functional methods often fall short in effectively modeling dispersion forces, especially across large intermolecular

distances. Coulombic and exchange interactions in these methods depend mainly on electron transition densities between interacting fragments. However, for a reliable description of long-range dispersion forces, additional treatment is necessary. To capture these long-range effects, techniques like dispersion-corrected DFT (DFT-D), van der Waals (vdW) functionals, or empirical force field methods are commonly applied. These approaches are capable of accurately modeling dispersion forces in both molecular and condensed-phase systems. A representative expression for the second-order dispersion energy is:

$$E_{\text{Disp}}^{(2)} = \sum_{ia} \sum_{jb} \frac{(\langle ia|jb \rangle)(\langle ia|jb \rangle - \langle ja|ib \rangle)}{\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j} \quad (3.41)$$

This equation accounts for the inclusion of all particle-hole excitations between orbitals $i \rightarrow a$ and $j \rightarrow b$, where i and j are orbitals localized on different fragments (A and B). The orbital energy is denoted by ϵ . These contributions are typically ignored in standard DFT [68]. To include dispersion interactions in practical simulations, empirical corrections are widely used. Among such methods, Grimme's DFT-Dn family, especially the DFT-D3 version, is highly popular. The general formula for DFT-D dispersion energy correction is:

$$E_{\text{Disp}}^{\text{DFT-D}} = \sum_{AB} \sum_{n=6,8,10,\dots} S_n \frac{C_n^{AB}}{R_{AB}^n} f_{\text{damp}}(R_{AB}) \quad (3.42)$$

3.3. Spin Polarized Density Functional Theory

Spin-polarized Density Functional Theory (DFT) is a major computational method used to investigate band magnetism in itinerant electron systems, particularly solid-state materials. This method plays a critical role in spin magnetic moment prediction and examining the fundamental interactions responsible for magnetic behavior. von Barth and Hedin's pioneering work, followed by efforts by Pant and Rajagopal, was pivotal in the development of this method [69,70].

A key feature in defining a material's magnetic properties is the interaction between exchange and kinetic energy. For electrons with parallel spin alignment, there is an exchange interaction gain in energy but a concomitant loss of kinetic energy. Spin-polarized DFT includes these many-body electron interactions through the exchange-correlation functional, so selecting an appropriate functional is crucial.

In the framework of KS-DFT, the explicit nature of exchange and correlation functionals remains ill-defined outside the free-electron gas model in an idealized form. One of the most widely applied approximations is the Local Density Approximation (LDA), which builds up the functional based on the electron density at a specific spatial coordinate [71]. A generalization (**Equation 3.38**) of this is the Local Spin-Density Approximation (LSDA), which adds spin-dependent terms to the LDA formalism [72].

$$E_{xc}^{GGA}[n_{\uparrow}, n_{\downarrow}] = \int \varepsilon_{xc}(n_{\uparrow}, n_{\downarrow}) n(\vec{r}) d^3r \quad (3.38)$$

Terkura and co-workers highlighted the limitations of LSDA in accurately capturing the magnetic behavior of metal oxide insulators [73]. Additionally, both LDA and LSDA are built on the presumption of uniform electron density, which leads to an overestimation of exchange-correlation energy. The Generalized Gradient Approximation (GGA) refines this by including spatial gradients of electron density. When spin polarization is included, the GGA functional is expressed as (**Equation 3.39**):

$$E_{xc}^{GGA}[n_{\uparrow}, n_{\downarrow}] = \int \varepsilon_{xc}(n_{\uparrow}, n_{\downarrow}, \vec{\nabla}_{n_{\uparrow}}, \vec{\nabla}_{n_{\downarrow}}) \rho(\vec{r}) d^3r \quad (3.39)$$

Through these modifications, LSDA and GGA struggle to model strong electronic correlations in partially occupied d and f-orbitals. To mitigate this constraint, the GGA+U method introduces a Hubbard-like U potential to explicitly treat electron-electron repulsion in localized orbitals [73-80]. In this scheme, GGA continues to describe delocalized s- and p-electrons

effectively, while localized d-electrons are corrected through additional Coulomb and exchange terms. This enhances the reliability of calculated excited-state properties such as band gaps and ground-state features like magnetic moments and interatomic exchange interactions. However, the accuracy of such calculations hinges critically on the appropriate choice of the U parameter.

Hybrid functionals have emerged to address strong electron correlation further. A prominent example is the B3LYP functional, which incorporates 20% Hartree-Fock (HF) exchange [81,82]. This method is effective in predicting thermochemical properties and the electronic structure of systems with substantial electron correlation. Other popular hybrid methods, such as B1LYP, mPWO, and screened hybrid functionals like HSE03 and HSE06, are particularly suited for investigating magnetic systems, as they decouple exchange-correlation interactions into long- and short-range components, treating the latter with HF exchange.

An accurate determination of a material's magnetic anisotropy requires correct modeling of spin alignment. Many spin-polarized DFT studies assume collinear spin alignment, which may not adequately describe materials exhibiting ferromagnetic or antiferromagnetic ordering. In such cases, non-collinear spin configurations must be considered, as they are critical in capturing spin-flip transitions and magnetic excitation spectra. Hobbs et al. demonstrated the effectiveness of non-collinear magnetic DFT using the projector augmented-wave (PAW) method [83]. Modern DFT implementations often combine such approaches with plane-wave pseudopotentials to handle non-collinear effects.

In contrast to atoms, solid-state systems often lack intrinsic magnetism because the exchange energy gain fails to overcome the kinetic energy cost. Nevertheless, elements like Fe, Co, Ni, and Cr can manifest magnetism under sufficient electron localization. Moreover, magnetic properties can be amplified by reducing dimensionality, as observed in metallic surfaces,

multilayers, heterostructures, two-dimensional materials, interfaces, nanowires, and nanosheets. These low-dimensional systems have been extensively analyzed through spin-polarized DFT methodologies.

3.4. Basis Sets

Basis sets are collections of one-electron functions used to construct molecular orbitals of a quantum system. In general, each atom in a molecule is associated with its own basis set of functions that approximate its atomic orbitals. Basis sets are categorized based on the mathematical form of the functions used to describe them, including: (i) Slater-type orbitals (STOs), (ii) Gaussian-type orbitals (GTOs), (iii) Effective core potentials (ECPs), (iv) Plane wave basis functions, among others.

Each category represents a distinct approach to demonstrate the behavior of electrons in atoms and molecules. For instance, Slater functions closely resemble hydrogen-like orbitals but are computationally demanding, whereas Gaussian functions simplify integral evaluations and are widely used in practical computations. Additionally, these different types of basis sets have specific advantages and limitations, making them suitable for various types of systems and calculations. The choice of basis set significantly influences the accuracy and computational cost of electronic structure methods such as HF Theory and DFT.

3.4.1. Plane Wave Basis Sets

Plane wave basis sets consist of periodic functions that span the entire system. Their precision is governed by a single parameter known as the cutoff energy [84]. The mathematical representation for a plane wave basis function is:

$$X_i(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} \quad (3.40)$$

Here, \mathbf{k} denotes the momentum vector, and \mathbf{r} corresponds to a position within the Bravais lattice. Selecting an appropriate basis set is crucial for

accurate quantum mechanical simulations. It is important to be aware of basis set superposition error (BSSE), especially when evaluating interaction energies like dimerization energies, which tend to be overestimated. This overestimation stems from the fact that the basis functions representing the full system are often more comprehensive than those representing each isolated part.

One widely accepted method to reduce this error is the counterpoise correction technique introduced by Boys and Bernardi [85]. This approach calculates the interaction energy for each system component in isolation and then introduces a correction term to adjust for the BSSE-related overestimation. Applying this correction is critical to ensure more reliable interaction energies in quantum chemical calculations.

3.4 Pseudopotentials

Modelling core electrons in periodic systems using plane-wave basis sets is computationally intensive due to the highly oscillatory nature of the wavefunctions near atomic cores. However, in most practical cases, valence electrons are primarily responsible for the materials properties of interest, while core electrons contribute minimally. To address this challenge, a widely used strategy involves selectively employing plane-wave basis functions to represent valence electron states. Simultaneously, pseudopotentials are introduced to effectively represent the interaction between core electrons and atomic nuclei. This method is commonly referred to as the frozen-core approximation. Pseudopotentials are formulated as effective potentials that replace the explicit treatment of core-valence and core-nucleus interactions. They enable an accurate description of valence electron behavior while avoiding the computational expense associated with treating core electrons explicitly. By allowing calculations to use a smaller basis set, pseudopotentials significantly reduce computational cost. Through this method, core electrons are either “frozen” or absorbed into the effective potential, simplifying the electronic structure

problem. Several types of pseudopotentials are available for plane-wave-based DFT calculations, including: (a) Norm-conserving pseudopotentials, (b) Ultrasoft pseudopotentials (USPP). Each type offers different trade-offs between accuracy and efficiency, and the appropriate choice depends on the system and the required level of precision.

3.5 Computational Hydrogen Electrode (CHE) Model

In the first principle-based modelling of electrocatalysis, determining the thermodynamic reaction of the free energy of the electrochemical steps induced a constraint. Calculating the free energy of electrochemical reactions involving protons and electrons cannot be done directly in DFT. The chemical potential of a proton-electron pair is involved in a concerted protonation step that can be assumed to be equivalent to that of a gaseous H₂ at the equilibrium potential, as shown in **Equation 3.43**.

$$\mu_{\text{H}^+} + \mu_{\text{e}^-} = \frac{1}{2} \mu_{\text{H}_2} \quad (3.43)$$

Therefore, the energy of the proton-electron pair can now be determined from the DFT calculated energy of the gaseous H₂ molecule. Since the chemical potential of the proton-electro pair can be determined from the DFT-calculated energy of the gaseous molecule. Since the chemical potential of protons gets shifted -eU under an applied potential U, the free energy of an electrochemical reaction can be determined as (**Equation 3.44**) [89],

$$\Delta G = \Delta E + \Delta \text{ZPE} - T\Delta S - neU \quad (3.44)$$

Where, ΔE is the change in electronic energy, ΔZPE is the change in zero-point energy, T is the temperature, ΔS is the change in entropy, and e is the electronic charge.

Chapter 4

Machine Learning Methods

4.1 Artificial Intelligence (AI)

Artificial Intelligence (AI) is a field of computer science that is concerned with the development of systems capable of performing tasks that would typically require human intelligence, but with excellent computational capabilities. The activities involve identifying language, pattern recognition, problem-solving, and decision-making. AI drives voice assistants (Alexa, Siri), autonomous vehicles, recommendation systems (YouTube, Netflix), ChatGPT, and deepseek. It operates based on algorithms and information to make machines learn, recognize, and improve over time.

The term AI was initially defined by British mathematician and logician Alan Mathison Turing in his seminal 1950 paper *"Computing Machinery and Intelligence"*. Turing presented in this paper a core question, *"Can machines think?"*, and formulated the notion of the Turing Test to quantify a machine's capability to demonstrate intelligent behavior equivalent to, or not distinguishable from, that of a human. The name *"Artificial Intelligence"* itself was coined initially later in 1956 by John McCarthy, who is universally acknowledged as one of the discoverers of AI, at the well-known Dartmouth Conference, which is regarded as the inception of AI as a subject of study. The decades saw enormous progress in the science of AI, especially with advances in computational capacity and access to large data sets. These advances have given rise to significant subfields of AI; Machine Learning (ML), Deep Learning (DL), and Data Science (DS) techniques were developed.

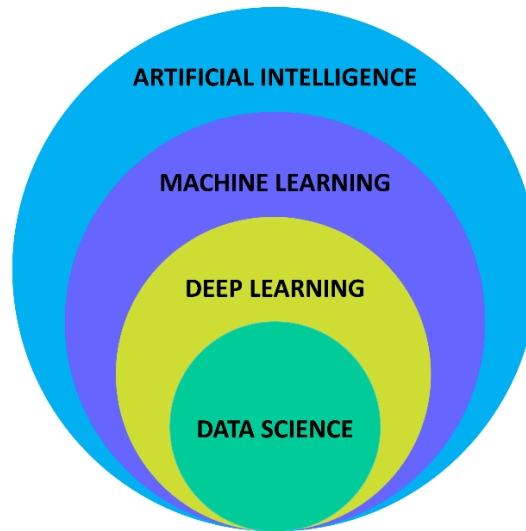


Figure 4: Hierarchical relationship between artificial intelligence, machine learning, deep learning, and data science.

4.2. Machine Learning (ML)

Machine Learning (ML) was first developed in 1952 by Arthur Lee Samuel, an American computer scientist. He had described ML as: *“The field of study that gives computers the ability to learn without being explicitly programmed.”* ML is specifically well-suited to tackle issues that involve vast fine-tuning, spontaneous improvement, and swift adaptation to changing conditions.

4.2.1 Supervised Learning

Supervised Machine Learning is among the most popular paradigms used in artificial intelligence, wherein the objective is to train a model on a labeled dataset to learn patterns between inputs and their respective outputs [90]. Every data point comes with an associated label so the model can identify patterns and make correct predictions on new, unseen data. The efficacy of supervised learning relies greatly on the quantity, quality, and variety of labeled data and the choice of informative features through which the model can differentiate between inputs efficiently and enhance prediction accuracy. Supervised learning is primarily concerned with two

issues: classification, which is to categorize input data into discrete pre-specified classes or categories, and regression, which is to predict a continuous value or quantity [90].

4.2.1.1. Regression Analysis

Regression analysis focuses on predicting a continuous numerical value based on input features. The target variable in regression tasks is always a real-valued number, making this approach crucial for problems where the output varies along a continuum [90]. The model learns to establish a relationship between the independent (input) variables and the dependent (output) variable by identifying trends or patterns. Standard techniques include Linear Regression, Polynomial Regression, Support Vector Regression (SVR), and ensemble methods like Gradient Boosting Regressors (GBR) and Neural Networks (NN). Regression plays a pivotal role in crucial massive tasks such as forecasting house prices based on location and size, predicting stock market trends, estimating temperature fluctuations over time, and predicting patient recovery times in healthcare. Moreover, assessing model performance in regression typically involves metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 Score to understand the quality of the predictions.

4.2.1.2. Classification Analysis

Classification involves categorizing input data into one of several predefined, discrete classes or categories [90]. Here, the target variable is categorical rather than continuous, often representing distinct groups or types. The model learns to detect patterns and correlations within the feature space that can help separate one class from another. Techniques commonly used include Logistic Regression, Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), and k-Nearest Neighbors (KNN). Performance in classification is often evaluated using metrics such as

Accuracy, Precision, Recall, F1-Score, and ROC-AUC curves, depending on whether the dataset is balanced or imbalanced. Stratified Sampling is recommended during model training to ensure that the class distribution in training and test sets matches the overall distribution, especially for imbalanced datasets.

4.2.2. Train and Test Data

In machine learning (ML), the training dataset serves as the fundamental basis for developing, training, and evaluating predictive models. It comprises input features and corresponding output labels, enabling the model to reveal complex patterns, infer relationships, and learn the mapping between inputs and outputs. The quality, size, and diversity of the training dataset are potential factors that influence model performance, as a sufficiently varied training set promotes better generalization and mitigates the risk of overfitting, wherein the model captures spurious correlations specific to the training data rather than learning generalizable trends.

Conversely, the test dataset, which typically contains only input features without associated outputs provided during model use, is employed to deliver an unbiased assessment of the models predictive capabilities on previously unseen data. Proper separation of training and test datasets is essential to ensure that performance evaluations genuinely reflect the models ability to generalize rather than its proficiency in memorizing training examples. Without separation, performance metrics may be falsely elevated, weakening the models practical adaptability.

4.2.3. Feature Representations

In the application of machine learning (ML) to materials science and molecular systems, a critical step is the development of appropriate feature representations. This process involves translating each atomic structure or molecular environment into a numerical format, known as a feature vector. Feature vectors allow ML algorithms to interpret, process, and analyze

complex data meaningfully. Since machine learning models cannot operate directly on raw material structures, effective numerical encoding is essential. A well-designed feature representation must capture essential physical properties and differences among materials, which must align with fundamental scientific principles. The quality of feature representations has a direct impact on the performance and reliability of machine learning models. Poorly designed features obscure insightful relationships or introduce noise, while well-constructed representations facilitate better learning, generalization, and interpretation.

Recognizing this, Ghiringhelli et al. (2015) outlined three key principles for constructing effective materials descriptors: (1) Uniqueness: each atomic or molecular structure must map to a distinct feature vector, (2) Proportionality: feature space should reflect gradual physical changes, (3) Dimensionality Balance: optimizing feature dimensionality to enhance computational efficiency and avoid processing bottlenecks [91].

Selecting an appropriate feature representation is particularly vital for crucial domains like catalysis and materials discovery, where both local and global environments play significant roles. For instance, in catalysis, features must encode not only the geometric arrangement of atoms but also the electronic environment, surface energies, and catalytic site properties to allow accurate prediction of reactivity and selectivity.

4.2.4. Performance Evaluation of ML Models

Performance evaluation helps determine how effectively a trained model can predict unseen data and provides insights into areas where the model may require further refinement [92]. The scikit-learn library, a widely used toolkit in Python for ML, offers an extensive range of performance metrics suitable for different tasks. Selecting the appropriate evaluation metric is essential because different ML problems, such as classification, regression, or clustering, require different criteria to assess success. For classification

tasks, commonly used performance metrics include accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC score. For regression tasks, performance is often evaluated using mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score (Coefficient of Determination) [93-95]. In this work, since the focus is primarily on regression problems, we emphasize key metrics such as MAE, RMSE, and R^2 score. These metrics collectively help to evaluate not only the accuracy but also the robustness and generalization performance of the developed regression models.

4.2.4.1. Root Mean Square Error (RMSE)

Root Mean Squared Error (RMSE) is a popular evaluation metric in machine learning, especially for regression problems [93]. RMSE measures the average magnitude of the error between predicted values and actual values. It gives higher weight to significant errors because errors are squared before averaging.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}} \quad (4.1)$$

Where, N stands for data points, y_i represents the actual points, \hat{y}_i indicates predicted value. Lower RMSE results in better fitting, higher RMSE results in worse predictions. It has the same unit as the target variable. It is easier to interpret and penalize large errors more than smaller ones. Sensitive to outliers and not normalized.

4.2.4.2. Mean Absolute Error (MAE)

MAE stands for Mean Absolute Error, and it is a common evaluation metric used in regression to measure how close predictions are to the actual outcomes [94].

$$MAE = \frac{\sum_{i=1}^N |y_i - \hat{y}_i|}{N} \quad (4.2)$$

Where, y_i indicates the actual value, \hat{y}_i represents the predicted value, and N is the number of data points. It's measured in the same units as the output variable. Additionally, the average magnitude of the errors without considering their direction, whether overprediction or underprediction, is easy to interpret. Robust to outliers (compared to MSE) because it doesn't square the error. Not differentiable at zero, which can be a problem for some optimization algorithms (like gradient descent).

4.2.4.3. R-Squared (R^2)

R^2 (R-squared), also known as the coefficient of determination, is a popular regression performance metric in machine learning. It tells you how well your model's predictions match the actual data [95].

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(\bar{y}_i - y_i)^2} \quad (4.3)$$

where y_i actual values, \hat{y}_i predicted values. \bar{y}_i is the mean of actual values and output. Additionally, R^2 value of 1 indicates a perfect model, meaning that the model explains 100% of the variance in the target variable. R^2 value of 0 suggests that the model performs no better than simply predicting the mean of the target variable. If R^2 is less than 0, the model performs worse than the baseline mean prediction, implying poor model fit.

Additionally, it is possible to achieve a high R^2 value while still having poor performance in metrics like MAE or RMSE. Therefore, R^2 should never be relied upon as the sole measure of model performance.

4.2.5. Hyperparameter Tuning

Hyperparameter tuning is a crucial step of the overall model development pipeline, as it significantly influences model performance and generalization ability. Unlike model parameters learned automatically during training, hyperparameters are external configurations set before the learning process, governing how the model learns. Common

hyperparameters include the learning rate in gradient-based optimization, the maximum depth of a decision tree, and the number of hidden layers or neurons in a neural network architecture. The objective of hyperparameter tuning is to find the optimal set of values that maximizes the model's predictive power while minimizing overfitting and underfitting. Effective hyperparameter optimization can substantially improve model accuracy, robustness, and efficiency. Different hyperparameter settings can lead to vastly different model behaviors and outcomes, even when applied to the same algorithm and dataset. Therefore, careful tuning is essential to strike the right balance between model complexity and generalization ability, ensuring that the model learns meaningful patterns from the training data without becoming overly specialized. In modern practice, advanced techniques such as Bayesian Optimization, Hyperband, and Automated Machine Learning (AutoML) frameworks are increasingly used to further automate and optimize the tuning process, especially in large-scale or high-dimensional search spaces. This section presents a detailed depiction of two commonly employed hyperparameter tuning techniques: GridSearchCV and RandomSearchCV [96].

4.2.5.1. GridSearchCV

GridSearchCV is a systematic and exhaustive method for hyperparameter optimization that aims to identify the best set of hyperparameters for a machine learning (ML) model. It combines two important concepts: Grid Search and Cross-Validation (CV). In Grid Search, all possible combinations of specified hyperparameter values are systematically explored, while Cross-Validation involves evaluating each hyperparameter combination using k-fold cross-validation to obtain a reliable and unbiased estimate of the model's performance. Hyperparameter tuning is critical because many ML models have hyperparameters that significantly influence their predictive performance. Examples include `max_depth` in decision trees, `n_estimators` in ensemble methods such as Random Forest

or XGBoost, and C and gamma parameters in Support Vector Machines (SVM). Manually selecting hyperparameters is generally ineffective and unreliable, as it often fails to cover the complexity of the hyperparameter space. Instead, GridSearchCV provides an automated and systematic approach to explore a wide range of hyperparameter combinations, ensuring a more thorough and optimized model tuning process [96].

The standard procedure involves defining a grid of hyperparameter values, training and evaluating the model on each combination using cross-validation, and selecting the configuration that yields the best performance based on a chosen evaluation metric (such as accuracy, RMSE, or F1-score). This exhaustive search ensures no potential hyperparameter setting is overlooked, maximizing the likelihood of finding an optimal model configuration. Furthermore, the key parameters in GridSearchCV that has been extensively employed for hyperparameter tuning in ML, are as follows: (1) *estimator*: specifies the machine learning model to be tuned, (2) *param_grid*: a dictionary where the keys represent hyperparameter names and the values are lists of settings to be tested during the search, (3) *cv*: defines the number of cross-validation fold used to evaluate each hyperparameter combination, (3) *scoring*: determines the evaluation metric to assess model performance, such as 'accuracy' for classification tasks or 'neg_mean_squared_error' for regression tasks, (4) *verbose*: controls the amount of logging output during the search process, with higher values producing more detailed messages, (5) *n_jobs*: sets the number of parallel jobs to run.

4.2.5.2. RandomizedSearchCV

RandomizedSearchCV offers an alternative strategy for hyperparameter optimization by randomly sampling hyperparameter combinations from predefined distributions, rather than exhaustively evaluating all possible configurations as in GridSearchCV. This random sampling approach enables the method to explore a broader and more diverse range of

hyperparameter settings while requiring fewer evaluations, thereby substantially improving computational efficiency.

A significant advantage of `RandomizedSearchCV` is its ability to effectively handle large and high-dimensional hyperparameter spaces, where evaluating every combination would be computationally prohibitive. By focusing computational resources on a random subset of potential configurations, `RandomizedSearchCV` can rapidly identify promising regions of the search space, often discovering near-optimal solutions at a significantly reduced computational cost.

The principal parameters of `GridSearchCV`, which has been widely adopted for hyperparameter optimization in ML workflows, are outlined as follows: (1) *param_distributions*: a dictionary specifying the range of hyperparameter values to sample from, which can be given as lists or probability distributions, (2) *n_iter*: defines the number of random hyperparameter combinations to test during the search process, (3) *scoring*: specifies the evaluation metric to be used for model assessment, (4) *cv*: sets the number of cross-validation folds used to evaluate each sampled hyperparameter combination, (5) *random_state*: controls random number generation for reproducibility, ensuring that the sampling process yields consistent results across runs, (6) *n_jobs*: determines the number of parallel jobs to run during hyperparameter search [97].

Due to its efficiency and flexibility, `RandomSearchCV` is particularly well-suited for time-sensitive applications or scenarios where the hyperparameter search space is very large or complex. Its ability to balance thoroughness with computational practicality makes it a preferred choice in many modern machine learning workflows

4.2.6. Cross-Validation (CV) Methods

Cross-validation (CV) is a fundamental technique for evaluating and validating machine learning models performance and generalization ability.

It assesses how well a model can maintain its predictive capabilities when exposed to new, unseen data. Instead of relying on a single train-test split, CV systematically partitions the dataset into multiple subsets, cycling through different training and testing configurations across iterations. This approach not only maximizes the use of available data but also provides a more robust and comprehensive assessment of model performance. By averaging results over multiple iterations, cross-validation reduces the influence of data variability and ensures more reliable performance estimation [98].

Traditional machine learning workflows often involve splitting the dataset once into a training set and a testing set. However, this single-split strategy can be problematic, as it may lead to biased performance estimates, either overly optimistic or overly pessimistic, depending on how representative the split is. Cross-validation focuses on this constraint by repeatedly partitioning the dataset into multiple training and testing sets, allowing models to be evaluated across different data subsets. This iterative evaluation process results in more accurate, stable, and generalizable performance metrics, significantly reducing the risk of misleading estimates that might arise from random data variations or unrepresentative splits.

4.2.6.1. K-Fold CV

In K-Fold Cross-Validation, the dataset is divided into 'k' equally sized subsets, known as folds. The machine learning model is trained on 'k-1' of these folds and validated on the remaining fold. This process is repeated 'k' times, ensuring that each fold serves once as a validation set while the others form the training set. The overall performance metric is obtained by averaging the results across all 'k' iterations, providing a more robust and statistically reliable estimate of the models actual performance [99,100]. This method helps reduce variance from a single random train-test split. K-Fold CV strikes an optimal balance between bias and variance by offering multiple evaluations, leading to a more comprehensive and dependable

assessment of a model's predictive ability. It mitigates the risks of overfitting (where a model performs well on training data but poorly on unseen data) and underfitting (where the model fails to capture underlying patterns).

Moreover, K-Fold CV ensures efficient utilization of the available data, which is especially crucial when working with limited datasets. Additionally, stratified versions of K-Fold CV, such as Stratified K-Fold, are often preferred for classification tasks, as they preserve the class distribution across folds, ensuring a fairer and more representative evaluation when dealing with imbalanced datasets.

4.2.7. ML Algorithms

4.2.7.1. K-Nearest Neighbor (KRR)

Kernel ridge regression (KRR) is a supervised ML algorithm that integrates Ridge Regression (a linear regression model with L2 regularization) using the kernel trick, enabling it to capture nonlinear relationships between features and target variables to mitigate complexity in regression analysis [101]. It belongs to the family of kernel methods, like Support Vector Machines (SVMs), and it's closely related to Gaussian Processes and Reproducing Kernel Hilbert Spaces (RKHS).

The objective of the Ridge Regression is to minimize the following loss function (**Equation 4.4**):

$$\min_w \|Xw - y\|^2 + \lambda \|w\|^2 \quad (4.4)$$

Where X indicates the matrix of input features, y denotes the target variable, w and λ the vector of coefficients and the regularization parameter, respectively.

KRR extends Ridge Regression by applying the kernel trick, which implicitly maps the input features into a higher-dimensional space without

explicitly computing the transformation. The optimization objective for KRR is then reformulated as (**Equation 4.5**):

$$\min_{\alpha} \|K_{\alpha} - y\|^2 + \lambda \alpha^T K \alpha \quad (4.5)$$

Where K is the kernel matrix, with elements $K_{ij} = k(x_i, x_j)$, where k is a kernel function, and α is the vector of dual coefficients. The kernel function $k(x_i, x_j)$ computes the inner product between the transformed features in the high-dimensional feature space. Solving this optimization problem yields the optimal values of α , which are then used to make predictions. Once α is known, predictions for a new input can be made using **Equation 4.6**.

$$y_{\text{pred}} = \sum_{i=1}^N a_i k(x_{\text{test}}, x_i) \quad (4.6)$$

Where x_{test} is the feature vector of the test instance, and N is the number of training samples. KRR is especially useful for modelling complex patterns in data, particularly in cases involving high-dimensional or nonlinear relationships between features and targets.

4.2.7.2. Random Forest Regression (RFR)

Random Forest Regressor is an ensemble learning ML algorithm that uses multiple decision trees and averages their prediction output to enhance the prediction accuracy and robustness of decision trees. It's part of the Random Forest family introduced by Leo Breiman in 2001. Random Forest is an ensemble learning method combining Bagging (Bootstrap Aggregating) and Decision Trees (DT) [101].

The working mechanism of a Random Forest Regressor can be summarized in several key steps. First, it performs bootstrap sampling, where it randomly samples the original dataset with replacement to create multiple different training datasets. Next, on each bootstrapped sample, a regression DT is trained independently. During the tree construction process, random

feature selection is applied at every split. Finally, to make a prediction, the model integrates the outputs from all individual decision trees by averaging their predictions.

Several hyperparameters are particularly important for tuning the Random Forest Regressor's performance. The hyperparameter "*n_estimators*" defines the number of trees in the forest. The "*max_depth*" hyperparameter controls the maximum depth allowed for each decision tree. The "*min_samples_leaf*" parameter specifies the minimum number of samples required to be present at a leaf node. Additionally, "*max_features*" determines how many features are considered when searching for the best split at each node.

The Random Forest Regressor offers several significant advantages, such as efficiently handling datasets with high dimensionality and being capable of capturing complex non-linear relationships between variables. Moreover, it typically does not require standardization or normalization of the input features. However, the algorithm has disadvantages, including being computationally expensive, especially when many trees are involved. Furthermore, compared to a single decision tree, Random Forest models are generally less interpretable, making it more challenging to understand the underlying decision-making process.

4.2.7.3. Gradient Boosting Regression (GBR)

Gradient Boosting Regression (GBR) is an ensemble learning technique designed specifically for regression tasks. It improves predictive performance by combining multiple weak models, typically decision trees, into a strong composite model. The fundamental approach involves sequentially building an ensemble of weak learners, with each one focusing on correcting the errors made by the ensemble of learners before it [101].

The workflow of GBR begins by initializing the ensemble with a simple model, such as a constant value, which serves as the initial prediction for all data instances. Following this, the algorithm calculates the residuals, which are the differences between the actual target values and the current predictions of the ensemble model. These residuals indicate the error that the next weak learner should attempt to correct.

In the next step, a new weak learner, typically a shallow decision tree, is trained to predict these residuals. Once trained, its predictions are added to the ensemble's overall prediction, but they are weighted according to the learning rate, which controls the influence of each tree. After this update, the residuals are recalculated by subtracting the predictions of the newly added weak learner from the previous residuals.

To ensure the model generalizes well and avoids overfitting, regularization techniques such as depth constraints, subsampling, and penalization parameters are applied. The process of adding weak learners continues until a stopping criterion is met. This criterion could be a predefined number of iterations or the point at which further improvements in the loss function become negligible.

4.2.7.4. eXtreme Gradient Boosting (XGBR)

The XGBoost Regression (XGBR) is a cutting-edge evolution of the gradient boosting algorithm, designed to maximize both computational efficiency and predictive performance [101]. It achieves this by employing optimized algorithms and advanced data structures, resulting in significant speed gains and scalability compared to traditional gradient boosting methods. The workflow begins with an initialization step, where a simple model, such as one that predicts a constant value, is used to generate the initial predictions for all instances. A differentiable loss function such as squared error or absolute error is then defined to quantify the difference between actual and predicted values. Using this loss function, the algorithm

computes the negative gradient for each instance, representing the direction in which predictions should be adjusted to reduce error.

Next, shallow decision trees are trained as weak learners to predict these negative gradients. These trained weak learners are then added to the ensemble model, and their contributions are scaled using a learning rate to prevent large updates and ensure gradual improvement. The ensemble predictions are iteratively updated by incorporating these new predictions, progressively refining the model's accuracy over time. This iterative boosting process continues until a predefined stopping criterion is met, such as a maximum number of iterations or negligible improvement in the loss function.

XGBR combines the flexibility of gradient boosting with the power of regularization and algorithmic optimization, enabling it to produce highly accurate predictions while maintaining computational efficiency and scalability. As a result, it demonstrates exceptional performance across a wide variety of regression tasks, especially when handling large, complex datasets.

4.2.7.5. Extra Trees Regression (ETR)

Extra Trees Regressor (ETR) is a machine learning algorithm based on the principles of ensemble learning. It belongs to the family of tree-based models and is closely related to the Random Forest algorithm, although it introduces several key differences that set it apart. Extra Trees is an ensemble method that builds multiple decision trees by incorporating additional randomization during the tree-building process. For regression tasks, it averages the predictions of all the individual trees, while for classification tasks, it aggregates votes.

Several hyperparameters are crucial for tuning the performance of the Extra Trees Regressor. The “*n_estimators*” parameter specifies the number of trees to build in the ensemble, while “*max_depth*” controls the maximum

depth allowed for each individual tree. The *“min_samples_split”* hyperparameter determines the minimum number of samples required to split an internal node, and *“min_samples_leaf”* specifies the minimum number of samples required to be at a leaf node. The *“max_features”* parameter governs the number of features considered when looking for the best split at each node, balancing model diversity and strength.

4.2.7.6. Adaptive Boosting (AdaBoost)

AdaBoost (Adaptive Boosting) is one of the earliest and most influential boosting algorithms, invented by Yoav Freund and Robert Schapire in 1996. The fundamental idea behind AdaBoost is to combine multiple weak learners, typically simple models such as decision stumps, to create a strong, highly accurate learner.

AdaBoost has various important hyperparameters, such as the *“base_estimator”* parameter, which specifies the type of weak learner to use, with `DecisionTreeClassifier` being the default. The *“n_estimators”* parameter defines the number of weak learners to combine, effectively controlling the size of the ensemble. The *“learning_rate”* parameter shrinks the contribution of each classifier, offering a trade-off between the number of estimators and the model’s robustness against overfitting.

4.2.7.7. Categorical Boosting (CatBoost)

CatBoost is a gradient boosting library developed by Yandex in 2017. CatBoost stands for "Categorical Boosting." CatBoost is designed to work efficiently for classification and regression, offering high performance with minimal data preprocessing. Unlike traditional gradient boosting libraries like XGBoost or LightGBM, which require manual encoding of categorical variables, CatBoost can process these features internally without needing label encoding or one-hot encoding.

CatBoost offers several important hyperparameters for tuning the model's performance. The “*iterations*” parameter defines the number of trees built during training, while the “*learning_rate*” controls the step size for updating predictions. The “*depth*” hyperparameter determines the maximum depth of the trees, and “*l2_leaf_reg*” controls the regularization strength to prevent overfitting. The “*loss_function*” can be set according to the task, such as RMSE for regression or Logloss for classification.

4.2.8. Correlation Matrices

A correlation matrix is a table that displays the correlation coefficients between pairs of selected variables or descriptors. Each cell in the matrix shows the strength and direction of the linear relationship between two variables. The correlation matrix provides an intuitive and compact way to visualize the interdependencies among multiple features, identify potential multicollinearity, and detect highly correlated pairs that might influence modeling or analysis.

4.2.8.1. Pearson Correlation Coefficient (PCC)

The Pearson correlation coefficient (PCC) quantifies the strength and direction of the linear relationship between two continuous variables, assuming values in the range $[-1, +1]$. A value of +1 denotes a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 signifies no linear association. PCC is extensively applied in fields such as statistics, machine learning, and dimensionality reduction to evaluate the fidelity of relational structures, particularly for assessing the preservation of pairwise distances between original and embedded spaces. Formally, the PCC between two variables x and y is defined as (**Equation 4.7**):

$$\text{PCC} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4.7)$$

where x_i and y_i are individual data points, and \bar{x} and \bar{y} denote their respective means. The numerator captures the covariance between x and y , while the denominator scales by their standard deviations, rendering PCC dimensionless. PCC assumes a linear relationship and is sensitive to outliers, which may bias the estimate. In cases where data exhibit nonlinearity or heavy-tailed distributions, rank-based alternatives such as Spearman's rho or Kendall's tau are often employed.

Chapter 5

Results and Discussion

5.1. Structural Design of Graphene-Supported Subnano clusters

In this section, we are exploring the putative global minimum (GM) configurations (the lowest energy configuration) and numerous low-energy metastable ensembles (LEME) of Pt_n ($n = 7-13$) subnanoclusters (SNCs) supported on graphene (Grs), in the cutoff energy of 0.4 eV relative to GM, extracted using global optimization (GO). It is to be noted that it may be difficult to explore an energy window too large in our study. Additionally, for many practical cases, there is only a limited number of thermodynamically accessible isomers with a minimum energy gap relative to the GM. Herein, we are constructing an ensemble of 6 isomers in addition to GM, attributed to the minimum energy gap relative to GM. Their isomeric distributions of all LEMEs as a function of energy and relative to 0.4 eV energy are represented in **Figure 1a** and **b**, respectively. The side and top views most stable GM configurations are represented in **Figure 1c**. The total number of LEME and the configurations with relative energy are given in **Figures A1-A7**.

LEME Structures of Pt_7/G : For Pt_7 cluster over Graphene (Pt_7/G) found to have a total of 10 isomers having different geometries. The first metastable isomer (MI) is 0.13 eV above the GM. The GM of Pt_7/G interestingly comes up with a non-planar three-dimensional cluster-like geometry with the highest symmetry, differed from the planar geometry of the GM of bare Pt_7 [102], and the prismatic GM of Pt_7/Al_2O_3 [103]. Additionally, the GM of Pt_7/G demonstrated C_{3v} point group symmetry, with a stable singlet spin state over graphene [104].

LEME Structures of Pt_8/G : By increasing one atom, the cluster geometric landscape converts into quite complex geometries. Pt_8/G exhibits a non-

rigid, naive, mostly bilayer, and a few-trilayered three-dimensional configurational distribution of Pt atoms over the graphene surface. Cluster

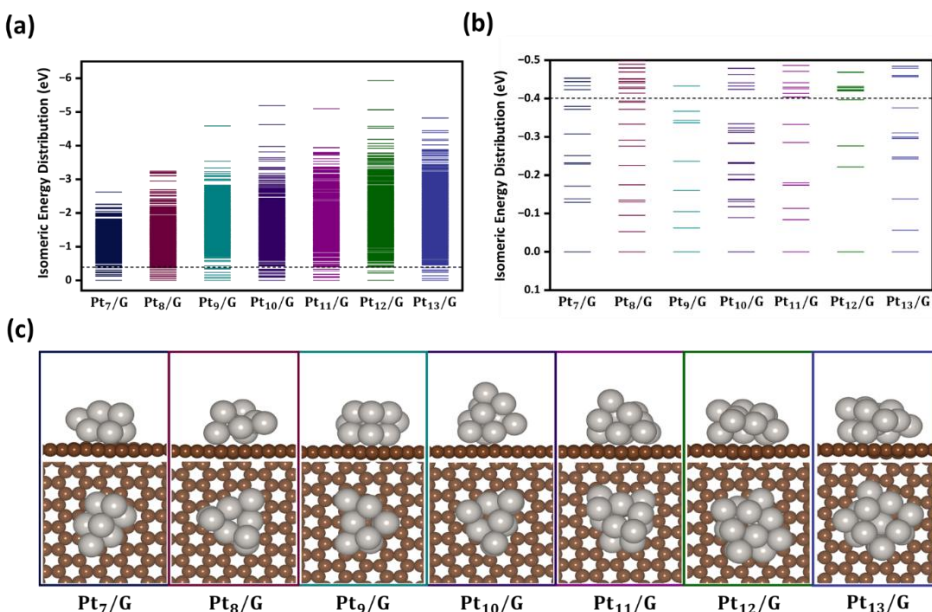


Figure 5.1: Isomeric distribution of Pt_n/G ($n = 7 - 13$) (a) as a function of energy over the entire energy pool and (b) within 0.4 eV relative to the GM energy (taken GM energy as a reference). (c) Side and top views for the energetically most stable global minimum structures of each Pt_n/G SNCs found via GO.

of Pt₈ exhibits a total of 13 isomers of diverse geometrical distributions within the same energy cutoff. The LM1 is 0.05 eV above the GM, in conjunction with a highly symmetric GM resulting in geometric similarity in low-lying MI above the GM corresponding to higher CN, resides in its quintet state [104].

LEME Structures of Pt₉/G: Pt clusters prefer planar structures of up to 9 atoms, which can be related directly to relativistic effects [105], whereas the GM of Pt₉/G exhibits a non-planar three-dimensional cluster-like geometry. Pt₉ found to have a total of 8 isomers having different geometries. The first LM is 0.06 eV above the GM, with a GM that is quite less symmetric compared to Pt₇ and Pt₈, corresponding to lower CN and significant configurational variations.

LEME Structures of Pt_{10}/G : Increasing Pt atoms from Pt_9 , a clear structural transition occurs. Furthermore, experimentally, it has been suggested that Pt_{10} is a magic cluster [105,106]. Pt_{10} found to have a total of 16 isomers having different geometries on graphene. The first low-energy MI is 0.09 eV above the GM. The GM of Pt_{10}/G demonstrating with a non-planar three-dimensional cluster-like geometry, differed from the planar geometry of the GM of bare Pt_{10}/G .

LEME Structures of Pt_{11}/G : Higher Pt clusters from Pt_{10} tetrahedral and octahedral growth motifs with highly disordered geometry are mainly observed [105,107]. Additionally, we did not find any evidence that the Pt_{11} considered as a magic cluster. Supported Pt_{11} found to have a total of 7 LEMEs having different geometries. The LM1 is 0.08 eV above the global minimum. The putative GM of Pt_{11}/G exhibits a non-planar three-dimensional cluster-like geometry, which differs from the open cluster geometry of the GM of bare Pt_{11} .

LEME Structures of Pt_{12}/G : As the number of Pt atoms in a cluster increase Pt_{12} found to have a total of 4 isomers having different geometries. Moreover, in harmony with Pt_{11} , a cluster of Pt_{12} sizes did not contain magic cluster properties. The LM1 is 0.22 eV above the global minimum, with less symmetry, resulting in a diverse geometry of low-lying MI near the GM. The putative GM of Pt_{12}/G comes up with a non-planar three-dimensional cluster-like geometry, which differed from the GM Pt_{12} of bare non-planar open cluster geometry, based on prisms [105]. Including spin-orbit coupling was found to have a minor impact (~ 0.01 eV) on the relative energy between the low-energy structures as well as isomeric configurations [108].

LEME Structures of Pt_{13}/G : Initial geometries of Pt_{13}/G obtained using GO have 9 different MI, having different geometries within 0.4 eV cut-off range relative to the GM. The first metastable isomer is 0.06 eV above the

GM. The icosahedral geometries of Pt₁₃ clusters are significantly deformed due to the strong interaction between Pt atoms and carbon atoms of support [109]. The GM of supported Pt₁₃ achieved of distorted octahedral compared to the GM of bare Pt₁₃, exhibits a low symmetry structure designated as tricapped pentagonal prismatic structure by Zhai et al or distorted structure (DIS) by Bunau et al [110,111].

5.2. Adsorption Energy Landscape on Subnano Clusters

In our investigations, the first objective is to study the adsorption energetics of the reaction intermediates for the four-electron ORR mechanism, *O, *OH, *OOH. With these three principle intermediates, we are performing a molecular adsorption simulation to identify the most stable intermediate-cluster configurations of size-selected Pt_n/G (n = 7-13) SNCs through spin-polarized DFT methods. Herein, we are exploring all non-equivalent adsorption sites including on top and bridging sites to identify the most stable adsorption configurations. The total number of configurational spaces sampled to identify the most stable configurations, including different top and bridge sites, is listed in **Table A1**. The distribution of adsorption energy corresponding to each intermediate, the adsorption energy of the most stable supported clusters-adsorbate configurations over GM and LMs of Pt_n/G (n = 7-13) are given in **Figure 2** and **A7-A14**, respectively. The adsorption energies of the ORR intermediates, hereon represented as E_{*O}, E_{*OH} and E_{*OOH}, at site *S* (top/bridge) of the cluster is computed as (**Equation 5.1**):

$$E_{\text{Pt}_n-\text{*O,*OH,*OOH}}^S = E_{\text{Pt}_n(\text{*O,*OH,*OOH})} - (E_{\text{Pt}_n} + E_{\text{*O,*OH,*OOH}}) \quad (5.1)$$

where $E_{\text{Pt}_n(\text{*O,*OH,*OOH})}$, E_{Pt_n} , and $E_{\text{*O,*OH,*OOH}}$ represents the electronic energies of the cluster with *O, *OH, and *OOH at site *S*, bare cluster and the corresponding intermediates in the gas phase, respectively.

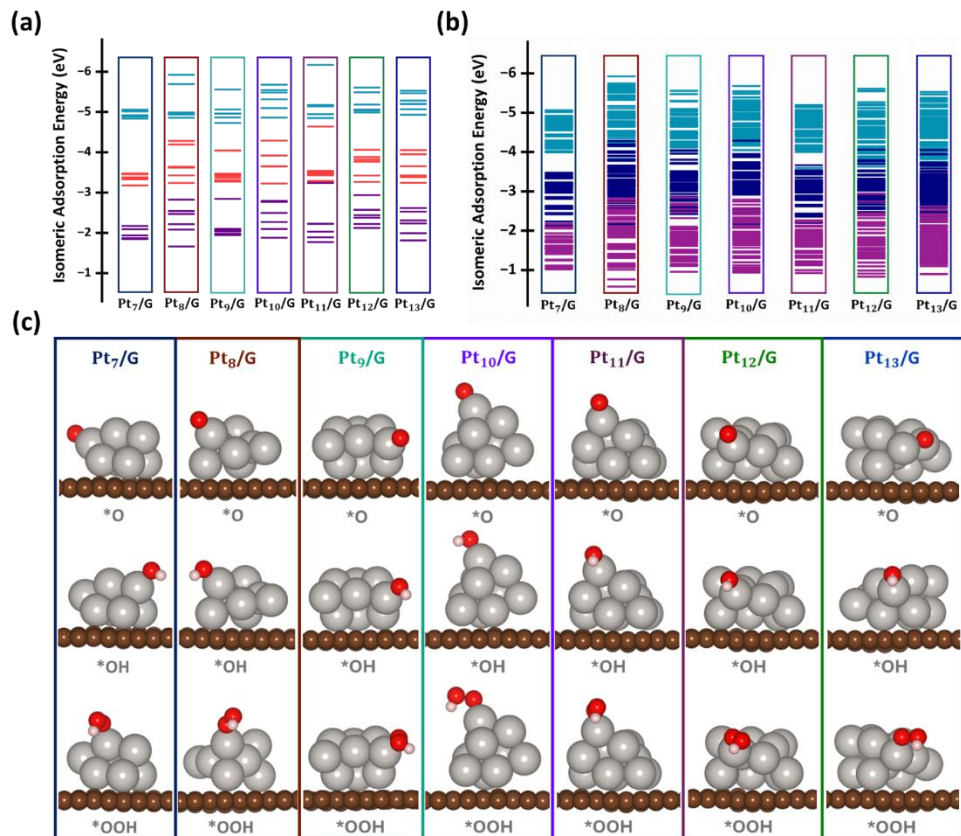


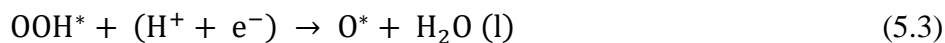
Figure 5.2: Distribution of adsorption energies of (a) the entire optimized intermediate-cluster adsorbed configurations and (b) the most stable intermediate-cluster adsorbed configurations. (c) Side view of DFT optimized most stable adsorption configurations of ORR intermediates on the GM of graphene-supported Pt_n (n = 7-13) SNCs. The gray, red, and pink indicate platinum, oxygen, and hydrogen atoms, respectively. Here, the sign * represents active sites of the catalyst.

Among all the intermediates, the *O intermediate displays the highest adsorption energy, with an almost equal preference for both top and bridging sites **Figure 5.2**. Additionally, LM3 of Pt₇/G and Pt₁₂/G, LM2 of Pt₈/G, and Pt₁₁/G, LM1 of Pt₉/G, GM of Pt₁₀/G, and LM4 of Pt₁₃/G demonstrating the highest adsorption energies for *O intermediates at the top and bridge sites, respectively (**Table A2**). Compared to the bulk Pt(111) surfaces, where *O primarily binds to the face-centered cubic (fcc) sites, *O intermediate tends to preferentially adsorb on the top and bridge sites of the Pt_n/G isomers, where n ranges from 7 to 13 [109]. Furthermore, from **Table A2**, it is evident that *OH and *OOH intermediates on different MI

both prefer to adsorb on top and bridge sites, where Pt(111) and Pt₇₉ nanoclusters adsorb on fcc sites [109]. Note that none of the intermediates prefer to bind directly attached Pt atoms bonded to the graphene sheet, but only when the surface of the Pt atoms is exposed predominantly.

5.3. Gibbs Free Energy Analysis

To investigate the thermodynamics of ORR pathways of size-selected Pt SNCs at reaction conditions, first-principles-based thermodynamics is exploited to study the Gibbs free energy of the ORR mechanistic pathways. According to Bell-Evans-Polanyi (BEP), the most stable adsorption energy is associated with a lower activation energy barrier into gain insights of reaction pathways. Note that among multiple ORR pathways, including electrochemical OOH dissociation, chemical OOH dissociation, and O₂ dissociation, have been proposed or H₂O formation under acidic conditions, along with the associative (mononuclear) mechanism involving *OOH, *OH, and *O intermediates. Furthermore, the dynamic nature of the catalytic interface and the presence of multiple non-equivalent active sites in the subnanometer regime, we are giving utmost importance solely to the associative pathway investigating multiple pathways would be quite challenging. The free energy diagrams of the Pt SNCs for the associative pathway at 0 V and 1.23 V are represented in **Figures A15** and **A16**, respectively. The change in Gibbs free energy (ΔG) associated with each step of the different metastable isomers is tabulated in **Tables A3–A9**. Therefore, utilizing the computational hydrogen electrode (CHE) model proposed by Nørskov and co-workers, we calculated the reaction energy at 0 and 1.23 V following the equation below:





In the associative pathway, the formation of H_2O from $^*\text{OH}$ was associated with the highest endothermic potential at 1.23 V. Consequently, hydrogenation of $^*\text{OH}$ resulted in the thermodynamically rate-determining step (RDS), corresponding the overpotential values are tabulated in **Table A10**. Notably, the least overpotential values (η) for the isomer in the associative pathway corresponds to the most thermodynamically active isomers. GM of Pt_7/G possessed comparable η values, LM3 of Pt_8/G , than the bulk Pt (111) surfaces ($\eta = 0.78 \text{ V}$). Also, compared to the Pt_{79} nanocluster ($\eta = 1.00 \text{ V}$) LM2, LM3 of Pt_7/G and LM3, LM4 of Pt_8/G possessed smaller η values in the gas phase, indicating enhanced thermodynamic activity and non-negligible contribution from metastable isomers.

3.5 Ab Initio Thermodynamics Phase Diagrams

Owing to structural changes of Pt/G clusters through altering the intermediates coverage under the electrochemical reaction condition to understand SNCs behavior in real-world catalytic processes, simulating realistic reaction conditions in cluster catalysis is important [114]. We are integrating DFT and ab initio thermodynamics, to investigate the stability and geometry of clusters, tendency to develop high intermediate coverage under ambient ORR condition ($\text{P}_{\text{O}_2} = 1 \text{ bar}$, $T = 300 \text{ K}$) [115]. During electrocatalytic reaction, ORR intermediates are adsorbed onto the surface of the SNCs, occupying numerous active sites. In ORR $^*\text{O}$, the intermediate displays the strongest binding tendency to clusters of each different size range. To identify the most stable oxidized phase of Pt_nO_x for the metastable isomers, we generated numerous configurations corresponding to each level coverage by sequentially adsorbing $^*\text{O}$ at various active sites. It is to be noted that we are specifically focusing on the LEME exhibiting the least overpotential value, as the least overpotential value exhibits the

highest activity among all LEMEs. Furthermore, each configuration were optimized using DFT, and the most stable configuration is selected for further analysis. The chemical potential of oxygen, μ_O is a function of temperature (T) and pressure (P). **Equation 5.6** provides a mathematical relationship to calculate μ_O (T, P), and is expressed as

$$\mu_O(T, P) = \frac{1}{2} \left[E_{O_2}(T = 0K, P^0) + \mu_{O_2}(T, P^0) + k_B \ln \left(\frac{P}{P^0} \right) \right] \quad (5.6)$$

Where T is the temperature, P is the partial pressure of oxygen, P^0 is the standard atmospheric pressure, and k_B is the Boltzmann constant. Vibrational degree of freedom of Pt and O atoms contributes to the free energy calculation. Consequently, the phase diagrams were constructed by using Python Multiscale Thermochemistry Toolbox (pMuTT) utilizing vibrational degree of freedom [116]. The phase diagram with their corresponding the most stable configuration of multiple adsorbates are shown in **Figure A17**. Remaining oxidized configurations under the ORR conditions for each isomer are represented in **Figures A18–A23**. Under the ambient ORR conditions ($T = 300$ K, $P_{O_2} = 1$ bar), the GM of Pt_7/G , LM5 of Pt_{10}/G , LM1 of Pt_{11}/G favored complete one monolayer formation, whereas the most active isomer LM3 of Pt_8/G and Pt_9/G , LM1 of Pt_{12}/G , GM of Pt_{13}/G favored Pt_8O_5 , Pt_9O_7 , $Pt_{12}O_{11}$, and $Pt_{13}O_{12}$ respectively.

3.5. Machine Learning Framework

3.5.1. Rational Design of Optimal Adsorption Energy Descriptors for ML Models

Following the acquisition of the initial adsorption energy Eads dataset, we proceeded to identify the optimal adsorption energies for each ORR intermediate. According to the Sabatier principle, the interaction between intermediates and catalyst surfaces should be optimal, neither excessively strong nor overly weak. To systematically filter inactive catalysts from Dataset 1, we applied three screening thresholds based on the Eads values of *O, *OH, and *OOH, as shown in **Figure 5.3**.

The upper and lower bounds for each intermediate were derived from established Eads values reported for benchmark systems, including Pt(111) surfaces, Pt₇₉ nanoclusters, and Pt₁₋₁₀ subnano clusters [105,117]. To account for metastability-induced reactivity and ensure statistical robustness, these thresholds were expanded by ± 0.4 eV, with Boltzmann-weighted sampling. The resulting refined dataset 2 includes 421, 422, and 416 entries for *O, *OH, and *OOH, respectively (**Figure 5.3**).

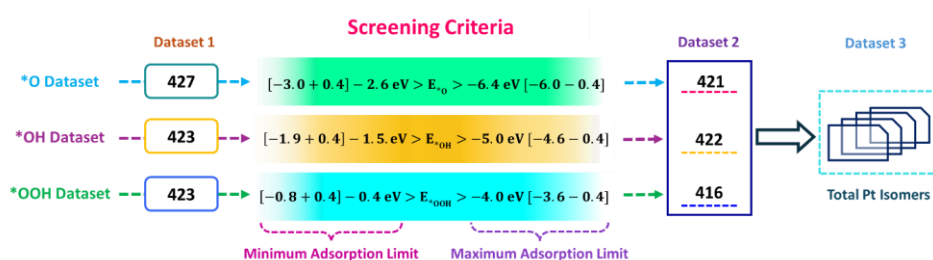


Figure 5.3: Schematic representation of the screening workflow used to identify active ORR electrocatalysts from the adsorption energy dataset, based on the Sabatier principle.

3.5.2. Descriptor Space Design for ML Prediction

The predictive performance of ML models is significantly influenced by the quality, and discriminative power of the input features. Inspired by previous works [118,119,120], 12 informative descriptors were extracted based on their accessibility and their effectiveness in characterizing the local atomic topology and accurately reflecting the structure-dependent activity trends of SNCs. The extracted features were systematically classified into three distinct groups: (1) elemental descriptors, (2) electronic descriptors, (3) geometric descriptors, tabulated in **Table 5.1**.

Table 5.1: List of Descriptors Spanning Elemental, Electronic, and Geometric Properties

Groups	Descriptors	Symbol
Elemental Descriptors	Number of Pt atoms	N
	Number of the most active Pt atoms	M
	Total number of d electrons	Σd_n

	Total atomic weight of cluster	ΣAW_{cls}
	Total electronegativity of the cluster	$\Sigma \chi_{cls}$
	Total electron affinity of cluster	ΣEA_{cls}
Geometric Descriptors	Number of sites (top, bridge)	S
	Coordination number	CN
	Distance between the adsorbate and the surface	D_{AS}
	Distance between Pt and the surface	Z
	Pt adsorbate bond length	L
	Number of Pt atoms that bind to the surface	A

The first two elemental descriptors (N, M) pertain to the physical properties of the catalysts. The latter four descriptors ($\Sigma d_n, \Sigma \chi_{cls}, \Sigma AW_{cls}, \Sigma EA_{cls}$) are connected to the electronic characteristics of the catalysts and indicate the active site's ability to donate or accept electrons during adsorption phenomena. Meanwhile, the six geometric features (S, CN, D_{AS}, Z, L, A) illustrate geometric aspects of the systems and are utilized for demonstrating the influence of the local atomic environment of SNCs.

The Pearson correlation coefficient (PCC) matrix was computed between descriptors, indicating that no single input descriptor can capture ORR adsorption energetics. As depicted in **Figure 5.4**, most feature pairs of SNCs exhibit low correlation ($|PCC| < 0.8$) [120], are allowed to coexist. Furthermore, we identified and removed the linearly strongly dependent and minimal contributions toward the output descriptor. After data mining, the final set of descriptors, as tabulated in **Table A11**, was used for machine learning analysis.

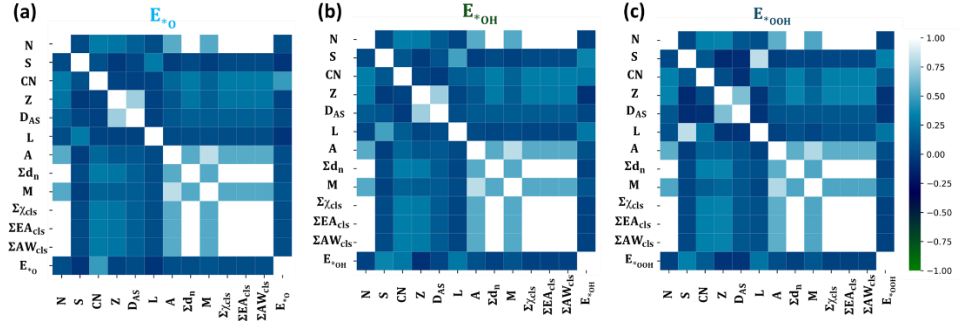


Figure 5.4: Pearson's correlation coefficient (PCC) matrices showing feature–feature and feature–target relationships for the (a) E_{*O} , (b) E_{*OH} , and (c) E_{*OOH} datasets, based on the initial 15 input features. The color bar on the right indicates the strength and direction of the correlation: white indicates a strong positive correlation, dark blue denotes negligible or no correlation, and green corresponds to a strong negative correlation.

The E_{ads} of each intermediate $*O$, $*OH$, and $*OOH$ on various Pt_n clusters were modeled as a function of a comprehensive set of features, as expressed in **Equation 5.7**:

$$E_{ads} = f(\Sigma d_n, S, CN, D_{AS}, Z, L, A) \quad (5.7)$$

This final dataset encapsulates the intrinsic complexity of adsorption behavior across diverse adsorption sites, cluster sizes, in the subnanometer regime. The high dimensionality and variability inherent to these systems necessitate a data-driven approach to effectively navigate this combinatorial landscape.

3.5.3. Machine Learning Model Training and Evaluation

Furthermore, a total of seven supervised regression algorithms have been implemented, including Kernel Ridge Regression (KRR), eXtreme Gradient Boosting Regressor (XGBR), Random Forest Regressor (RFR), AdaBoost Regression (ABR), Extra Trees Regression (ETR), Gradient Boosting Regression (GBR), and CatBoost Regressor (CR). The performance of each regression model was quantitatively assessed using standard evaluation metrics: The Coefficient of Determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). All ML

algorithms are implemented using the open-source Scikit-learn library within the Python 3 environment [121].

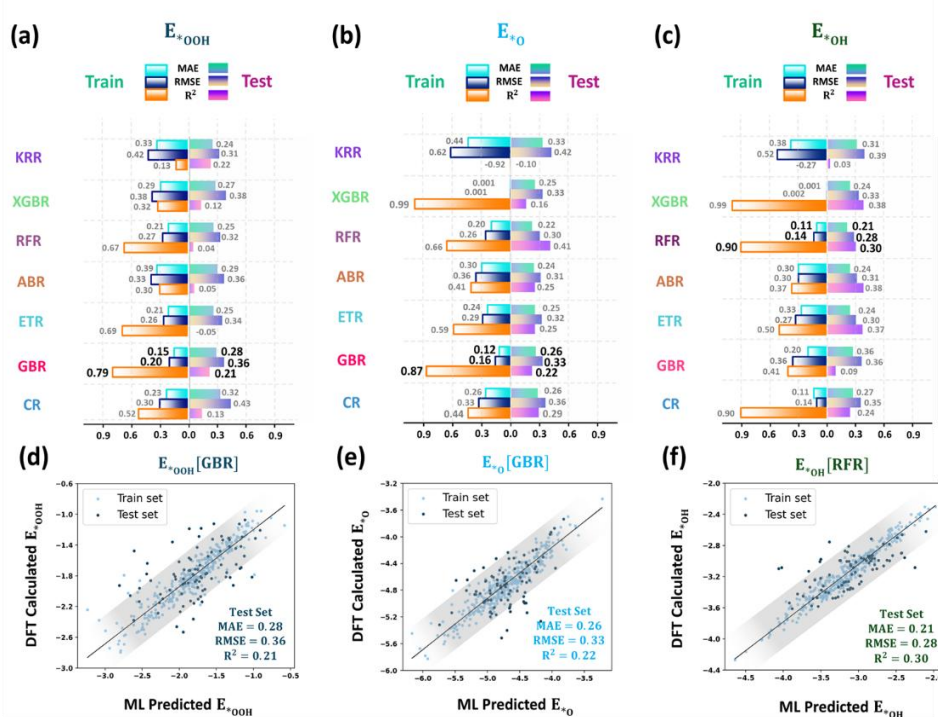


Figure 5.5: The predictive performance of five machine learning models: KRR, XGBR, RFR, ABR, ETR, GBR, CR was assessed using MAE, RMSE, and R^2 for the (a) E_{*O} , (b) E_{*OH} , and (c) E_{*OOH} datasets, reported in units of eV. For each intermediate, the model yielding the best performance metrics is highlighted with a pink rectangular box. Additionally, parity plots are presented to compare DFT-calculated versus ML-predicted values of (d) E_{*O} , (e) E_{*OH} , and (f) E_{*OOH} , as obtained from the respective best-performing models.

To enable reliable prediction of E_{*O} , E_{*OH} , and E_{*OOH} across a broad chemical space of subnanometer clusters, it is essential to employ robust ML algorithms capable of capturing complex nonlinear patterns inherent in such datasets. In this study, we selected a diverse set of ML models encompassing both ensemble-based and kernel-based approaches to ensure a balanced and comprehensive exploration of the descriptor space.

CR was chosen for its effective handling of categorical features with minimal preprocessing. GBR incrementally improves performance by

sequentially correcting the residuals of previous learners, while ABR emphasizes learning from hard-to-predict samples. RFR, through ensemble averaging of decision trees, mitigates overfitting and offers interpretable feature importance rankings. KRR, which combines ridge regression with kernel functions, enables the modeling of highly nonlinear relationships. To ensure optimal model performance, all algorithms were tuned using the *RandomizedSearchCV* method. Model evaluation was carried out using standard regression metrics, including mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2).

Figure 5.5a-c displays the final test performance for each ML model across the three ORR intermediates. Due to variations in the distribution and complexity of Eads values for $\ast\text{O}$, $\ast\text{OH}$, and $\ast\text{OOH}$, different algorithms were found to perform best for each case. Specifically, GBR demonstrated superior predictive capability for $E_{\ast\text{O}}$ and $E_{\ast\text{OOH}}$, achieving MAE/RMSE values of 0.26/0.33 eV and 0.28/0.36 eV, respectively, along with moderate R^2 scores of 0.22 and 0.21 on test datasets. While models such as ABR and RFR showed reasonable performance for $E_{\ast\text{OH}}$, elevated MAE/RMSE values in the training sets indicated underfitting. Similarly, RFR underperformed for $E_{\ast\text{OOH}}$ due to underfitting effects.

In contrast, RFR yielded the best predictive performance for $E_{\ast\text{OH}}$, achieving low MAE/RMSE values of 0.21/0.28 eV and a low R^2 score of 0.30. The accuracy of the selected models was further validated through parity plots (**Figure 5.5d-f**), which illustrate the agreement between DFT-calculated and ML-predicted adsorption energies for each intermediate across both training and test datasets. Based on these results, GBR was selected as the optimal model for $E_{\ast\text{O}}$ and $E_{\ast\text{OH}}$ while RFR was employed for $E_{\ast\text{OH}}$, and these models were subsequently used to predict adsorption energies across Dataset 2 to identify promising active electrocatalyst candidates.

To assess the reliability of our best-performing machine learning model, we predicted the adsorption energies of O, OH, and OOH intermediates for catalysts within the optimal Sabatier range. For validation and screening of electrocatalytically active candidates, we identified 6 isomers of each subnano clusters from both DFT-computed and ML-predicted datasets that exhibited valid adsorption energies for all three intermediates. These were compiled into a third dataset (Data Set 3), representing a subset suitable for mechanistic analysis. To elucidate the ORR mechanism, we considered an associative pathway involving sequential adsorption of *OOH, *O, and *OH. Although multiple ORR mechanisms have been proposed under acidic conditions, including electrochemical and chemical OOH dissociation, as well as O₂ dissociation, we restricted our focus to the associative route. This choice is due to the complexity introduced by the presence of multiple non-equivalent active sites in subnano clusters, which makes exploration of all possible pathways particularly challenging.

Figure 5.6: Construction of volcano plots for ORR activity using (a) DFT-calculated, and (b) ML-predicted η values in data set 3. Heat map corresponds to (c) DFT predicted η values, (d) ML predicted η values. Catalysts positioned at the apex outperform the Pt(111) surface (marked horizontally in brown color) at the subnanometer regime with the lower η values.

Chapter 6

Conclusions and Scope for Future Work

In this study, we presented an integrated DFT and ML framework to systematically screen size-selected Pt SNCs supported on graphene for the ORR activity. Our findings highlight the significance of considering both the GM and LM configurations to accurately capture the structural and electronic diversity of subnanometer catalysts. Diversity of adsorption energy and Gibbs free energy analyses reveal that *OH hydrogenation is the thermodynamic rate-determining step across most isomers, with several LEMEs exhibiting lower overpotentials than the bulk Pt(111) benchmark, indicating enhanced ORR performance. Furthermore, ab initio thermodynamic phase diagrams show coverage-dependent structural transformations under ORR-relevant conditions, highlighting the crucial role of surface oxidation in stabilizing active configurations.

Through the development and application of a multi-descriptor ML model, we achieved accurate predictions of adsorption energies, enabling rapid screening across the vast isomeric space. The ML framework captures non-linear relationships between the local environment and performance, while significantly reducing computational cost. Our work demonstrates the power of combining first-principles calculations with data-driven models for the rational design of subnano cluster catalysts. Additionally, the fundamental insights into the origin of activity trends in Pt-based SNCs and establish a scalable platform for discovering next-generation electrocatalysts for sustainable energy technologies.

APPENDIX-A

1. Structural extraction of LEME of graphene-supported SNCs

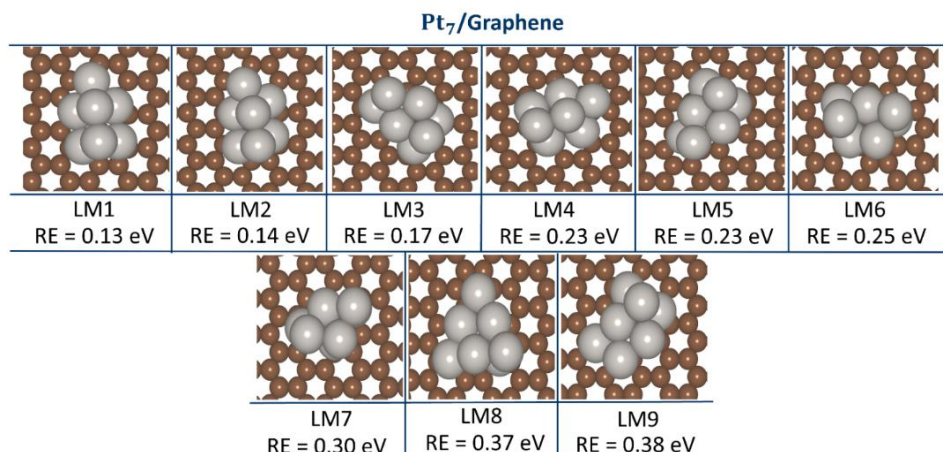


Figure A1. LEME structures of Pt₇/G SNCs: LM1-LM9 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM.

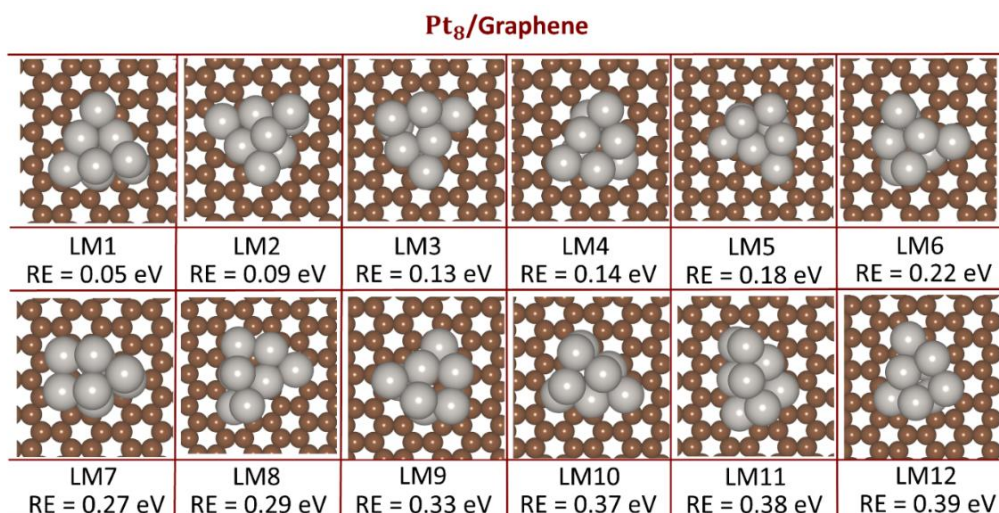


Figure A2. LEME structures of Pt₈/G SNCs: LM1-LM12 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM.

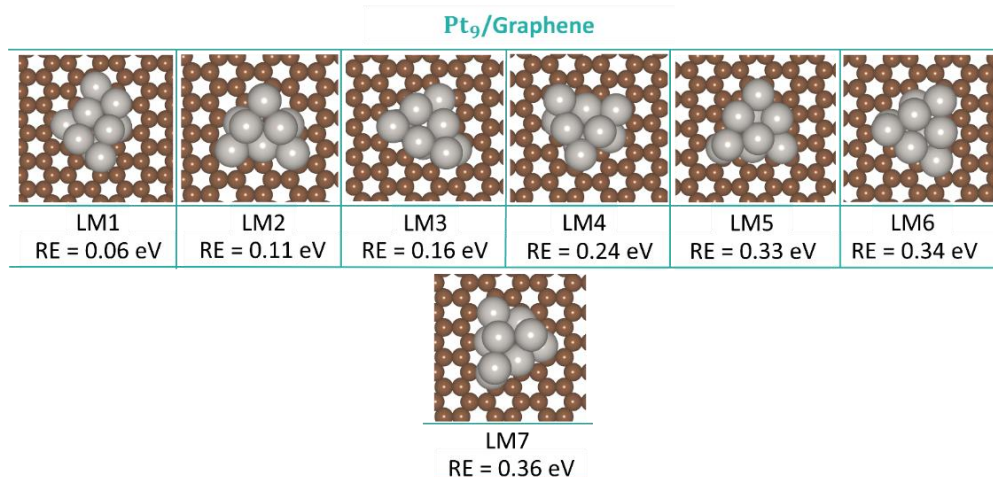


Figure A3. LEME structures of Pt₁₀/G SNCs: LM1-LM15 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM.

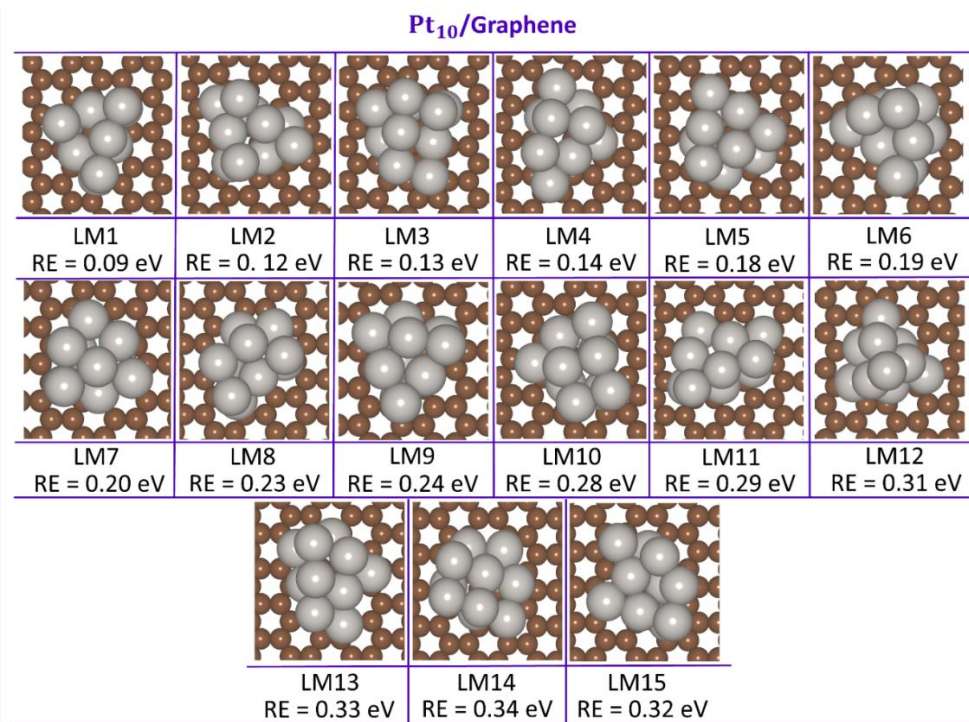


Figure A4. LEME structures of Pt₁₀/G SNCs: LM1-LM15 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM.

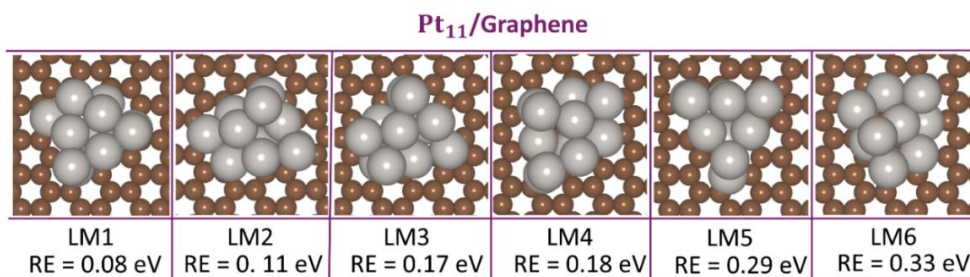


Figure A5. LEME structures of Pt₁₁/G SNCs: LM1-LM6 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM.

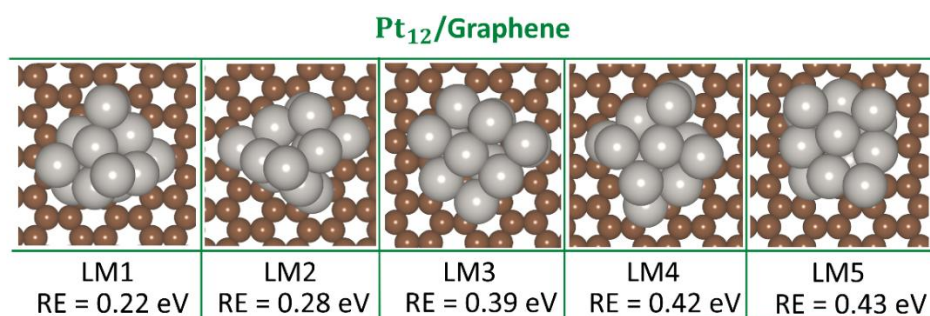


Figure A6. LEME structures of Pt₁₂/G SNCs: LM1-LM5 depict the low-energy configurations along with their relative energies (in eV) with respect to the global minimum (GM). Using GO, three structures were identified within 0.4 eV of the GM and two additional structures slightly above this threshold.

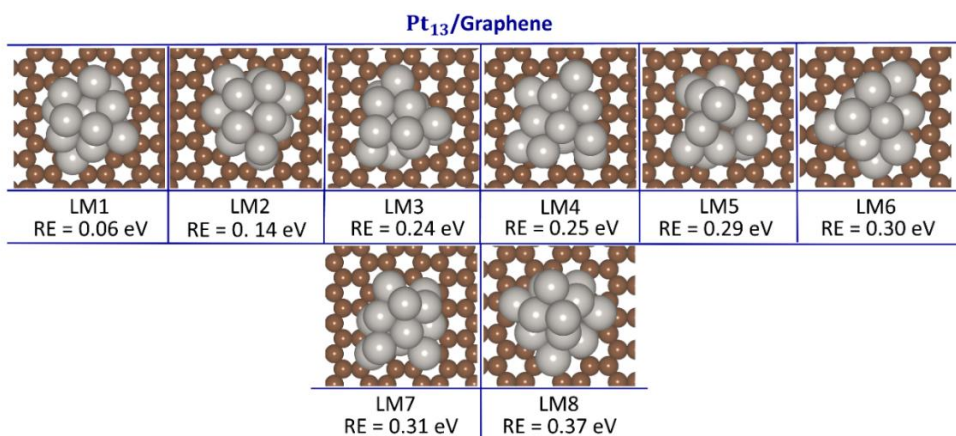


Figure A7. LEME structures of Pt₁₃/G SNCs: LM1-LM8 represent the distinct low-energy configurations identified within 0.4 eV of the GM energy, as extracted via GO. The relative energies (RE) of each structure are reported with respect to the GM.

2. Adsorption Behavior of Supported Subnanoclusters

Table A1: The total number of structural configurations sampled for each size-selected Pt_n/G ($n = 7-13$) isomer to identify the most stable adsorption geometries of ORR intermediates.

Clusters	Sites	GM	LM1	LM2	LM3	LM4	LM5
Pt_7/G	top	9	9	9	12	9	12
	bridge	6	6	6	12	6	12
Pt_8/G	top	15	18	18	15	12	18
	bridge	12	21	18	12	9	18
Pt_9/G	top	15	12	15	12	12	15
	bridge	12	9	15	9	12	12
Pt_{10}/G	top	18	9	15	18	15	9
	bridge	15	18	12	15	12	12
Pt_{11}/G	top	12	15	15	18	18	12
	bridge	9	12	12	12	15	12
Pt_{12}/G	top	15	15	18	21	18	21
	bridge	12	18	30	21	12	21
Pt_{13}/G	top	24	24	21	21	24	18
	bridge	27	33	24	21	33	15

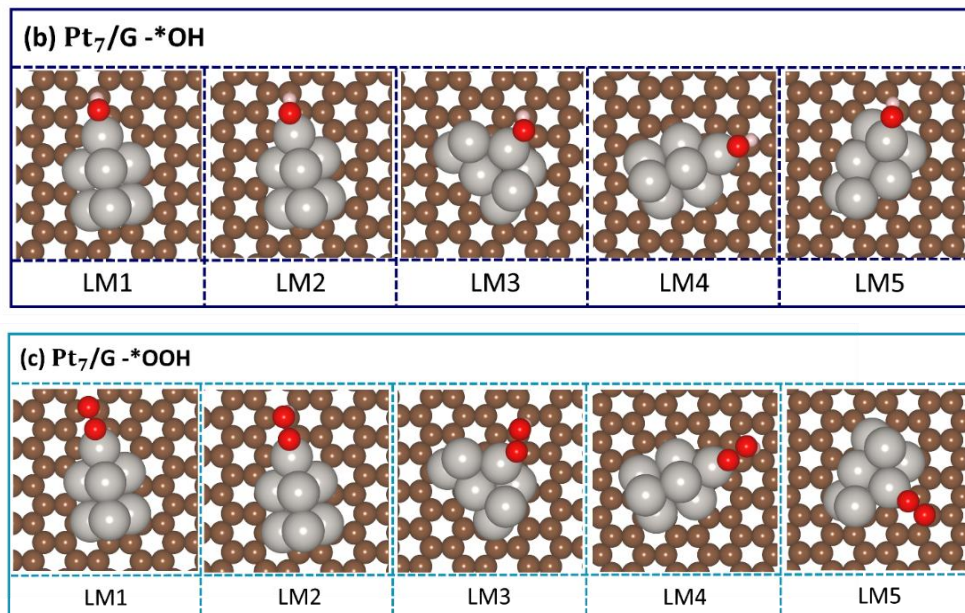


Figure A8. Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt₇ supported on graphene.

Pt₈/Graphene

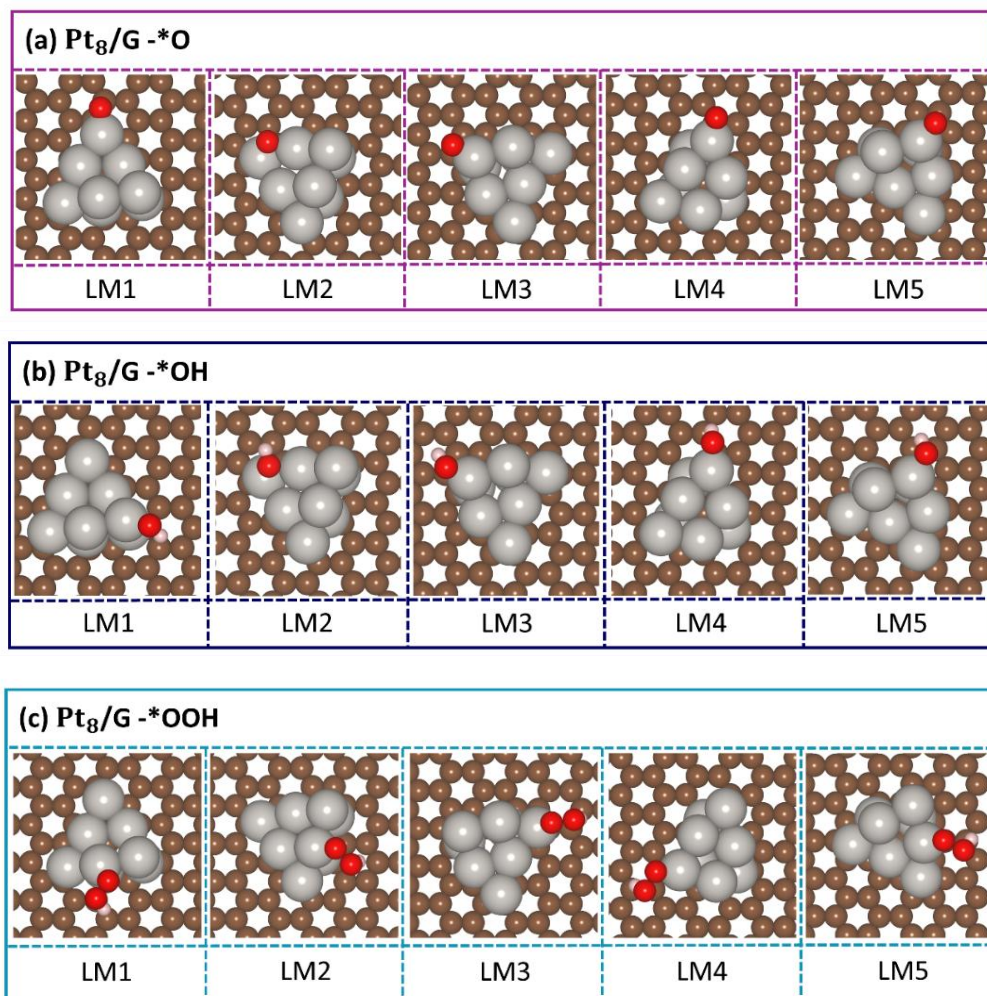


Figure A9. Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt₈ supported on graphene.

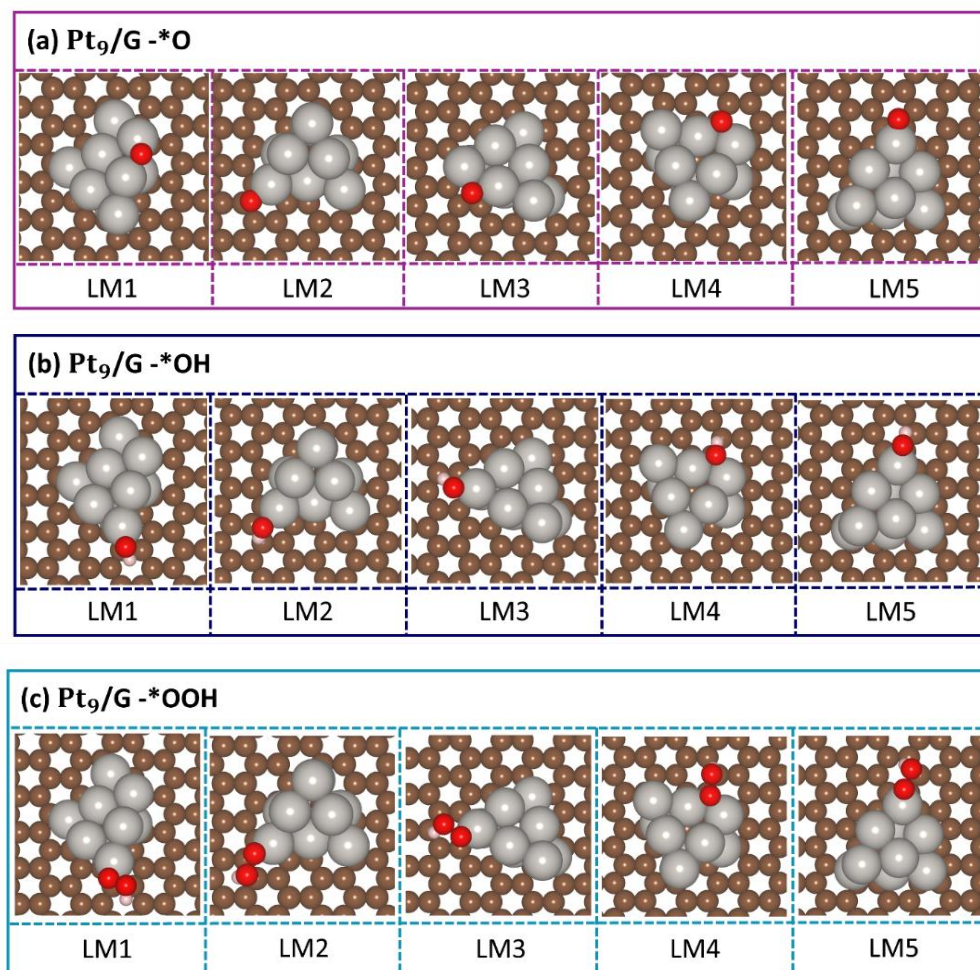


Figure A10. Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt₉ supported on graphene.

Pt₁₀/Graphene

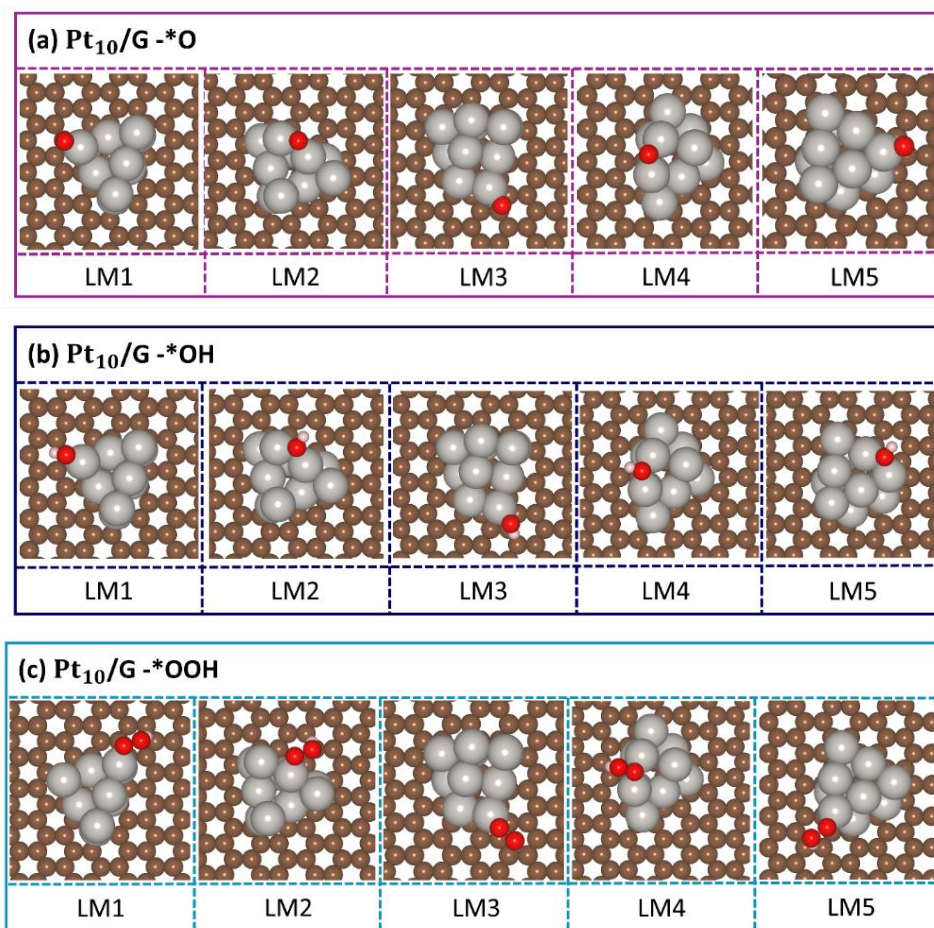


Figure A11. Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt₁₀ supported on graphene.

Pt₁₁/Graphene

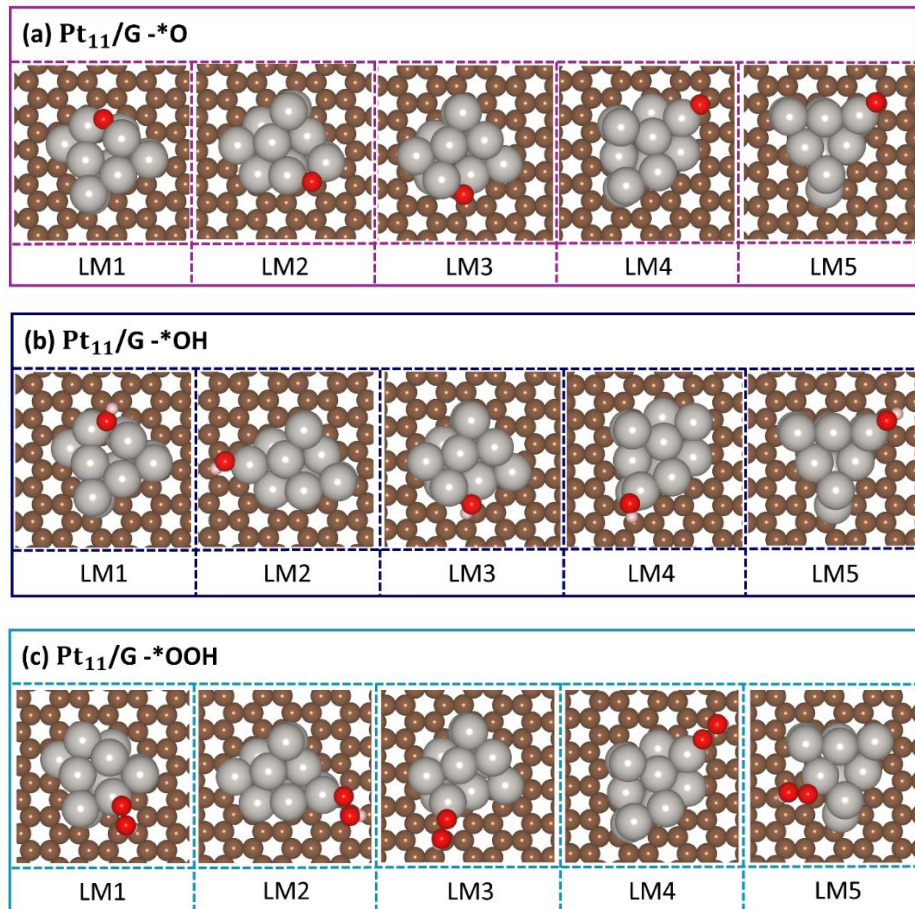


Figure A12. Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt₁₁ supported on graphene.

Pt₁₂/Graphene

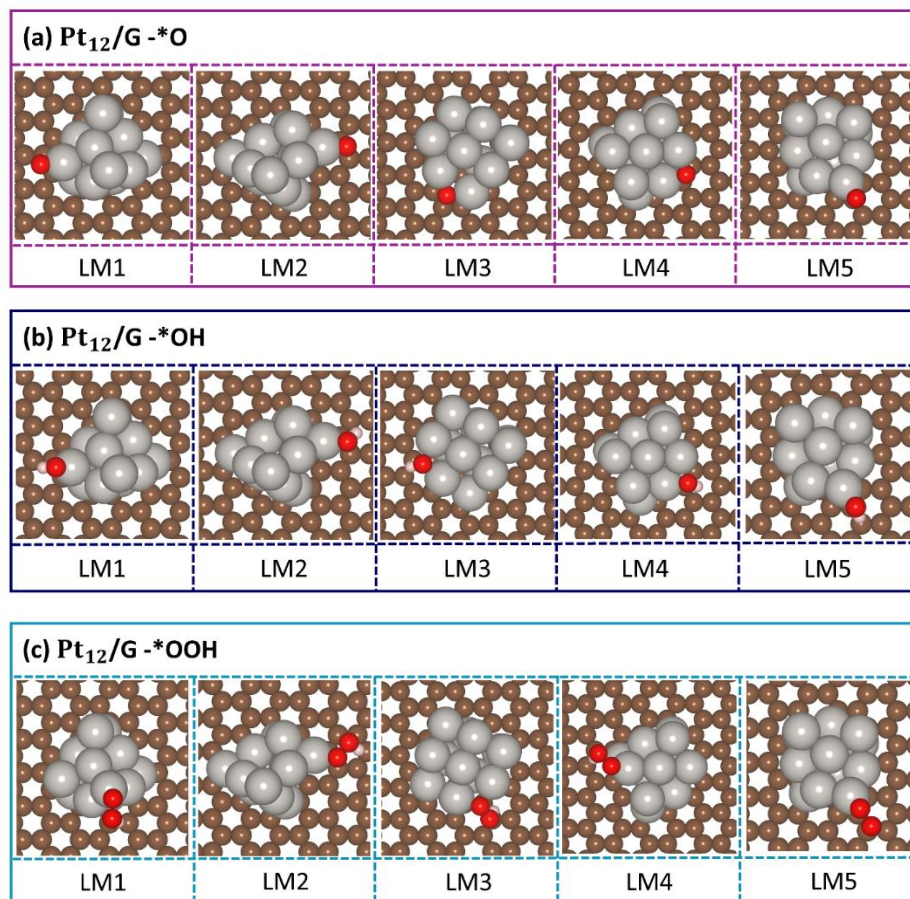


Figure A13. Depiction of the most stable binding geometries of ORR intermediates on various LM isomers of Pt₁₂ supported on graphene.

Table A2. Adsorption energies of ORR intermediates in the gas phase: adsorption energies (in eV) of ORR intermediates corresponding to their most stable intermediate–cluster adsorption configurations on Pt_n/G (n = 7-13), evaluated in the gas phase.

Clusters	Intermediates	GM	LM1	LM2	LM3	LM4	LM5
Pt ₇ /G	*O	-4.83(t)	4.91(b)	-5.01(t)	-5.06(b)	-4.89(t)	-4.89(b)
	*OH	-3.17(t)	-3.47(t)	-3.37(t)	-3.46(t)	-3.33(t)	-3.47(t)
	*OOH	-1.84(t)	-1.86(t)	-1.94(t)	-2.17(t)	-1.85(t)	-2.09(t)
Pt ₈ /G	*O	-4.94(b)	-4.96(t)	-5.92(b)	-4.83(t)	-4.99(t)	-5.69(t)
	*OH	-3.61(t)	-3.43(t)	-4.19(b)	-3.24(t)	-3.64(t)	-4.28(t)
	*OOH	-2.22(t)	-2.08(t)	-2.82(t)	-1.66(t)	-2.47(t)	-2.54(t)
Pt ₉ /G	*O	-4.86(t)	-5.56(b)	-4.95(t)	-4.72(b)	-5.06(b)	-4.97(t)
	*OH	-3.31(t)	-4.04(t)	-3.41(t)	-3.27(t)	-3.47(b)	-3.37(t)
	*OOH	-2.10(t)	-2.84(t)	-2.03(t)	-1.94(t)	-1.97(b)	-2.07(t)
Pt ₁₀ /G	*O	-5.67(t)	-5.09(t)	-5.54(b)	-5.48(t)	-5.31(b)	-4.86(t)
	*OH	-4.29(t)	-3.65(t)	-3.92(b)	-3.92(t)	-3.65(b)	-3.23(b)
	*OOH	-2.79(t)	-2.09(t)	-2.77(t)	-2.49(t)	-2.27(b)	-1.87(t)
Pt ₁₁ /G	*O	-4.94(t)	-5.14(b)	-6.16(b)	-5.14(b)	-5.17(t)	-4.85(t)
	*OH	-3.47(t)	-3.28(b)	-4.64(t)	-3.51(b)	-3.53(t)	-3.42(t)
	*OOH	-2.22(t)	-1.89(t)	-3.23(t)	-2.02(t)	-2.23(t)	-1.77(t)
Pt ₁₂ /G	*O	-4.98(t)	-5.03(t)	-5.18(t)	-5.60(b)	-5.49(b)	-5.06(t)
	*OH	-3.42(t)	-3.26(t)	-3.76(t)	-4.06(b)	-3.81(b)	-3.88(t)
	*OOH	-2.21(t)	-2.12(t)	-2.44(t)	-2.93(b)	-2.57(t)	-2.37(t)
Pt ₁₃ /G	*O	-5.06(b)	-5.19(b)	-5.46(t)	-5.28(b)	-5.53(t)	-4.93(t)
	*OH	-3.24(t)	-3.38(b)	-3.95(t)	-3.65(t)	-4.05(t)	-3.43(t)
	*OOH	-1.81(t)	-1.99(b)	-2.52(t)	-2.31(b)	-2.62(b)	-2.24(t)
Bare Pt ₁₃ SNC	*O	-4.92(b)	-4.82(t)	-4.60(t)	-5.03(t)	-5.12(b)	-4.81(t)
	*OH	3.26(t)	-3.44(t)	-3.30(t)	-3.31(t)	-3.43(t)	-3.29(t)
	*OOH	-1.87(b)	-2.05(t)	-1.91(t)	-2.11(t)	-1.96(t)	-1.97(t)

3. Gibbs Free Energy

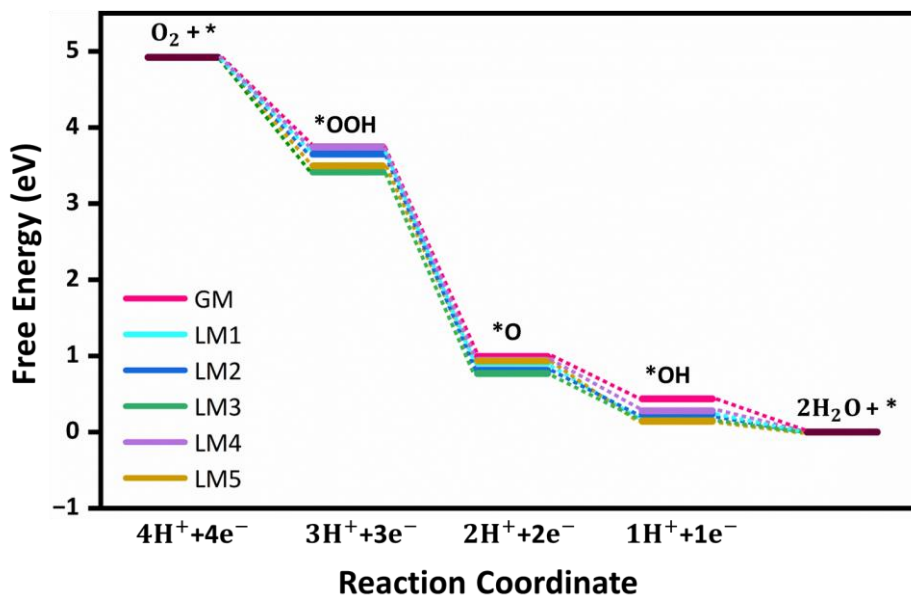


Figure A14: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt_7/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

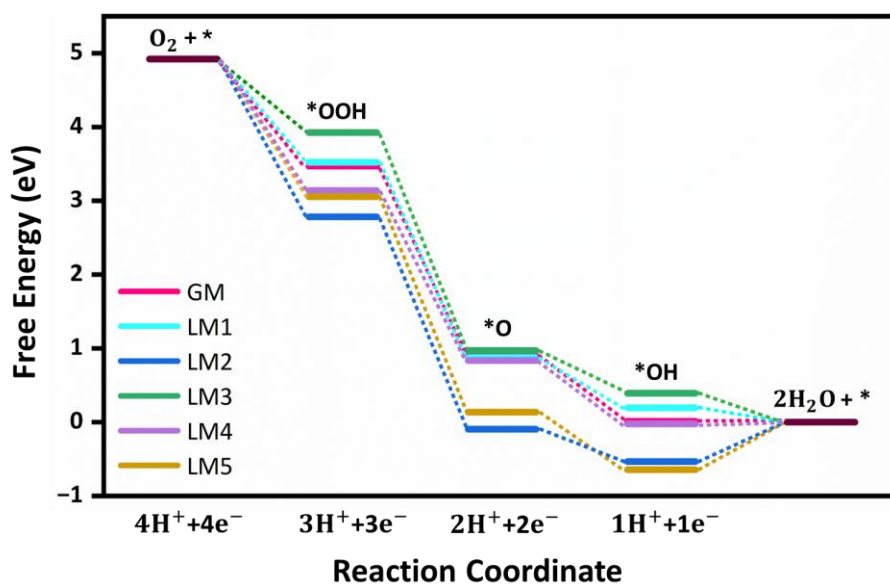


Figure A15: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway

for different isomers of Pt_8/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

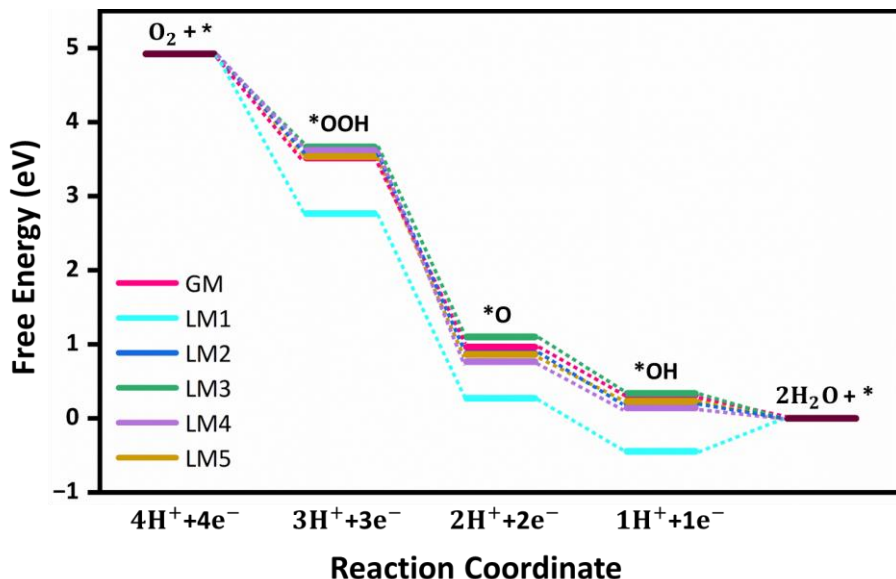


Figure A16: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt_9/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

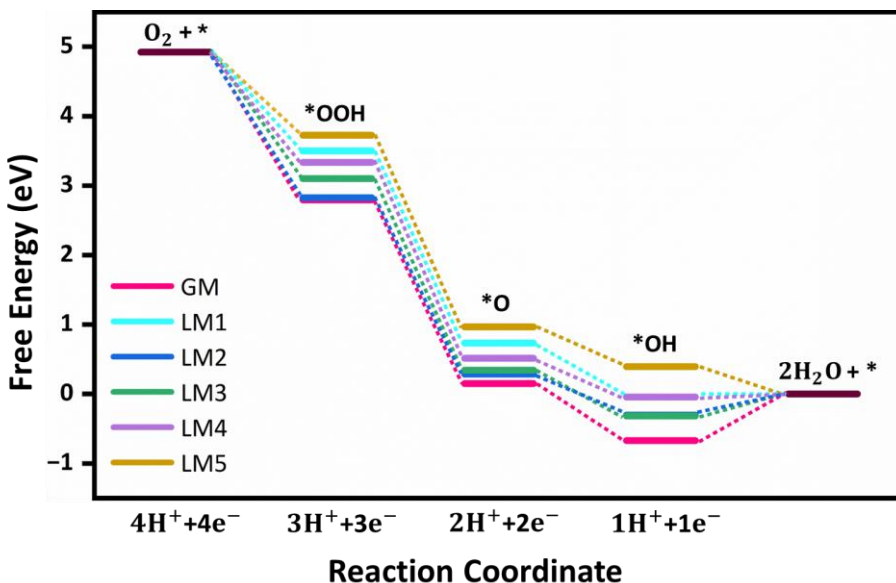


Figure A17: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway

for different isomers of Pt_{10}/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

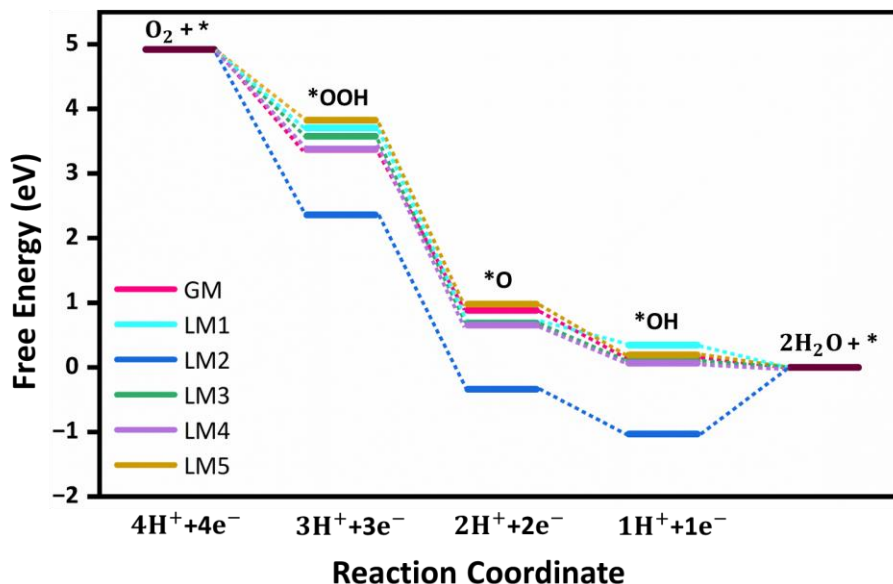


Figure A18: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt_{11}/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

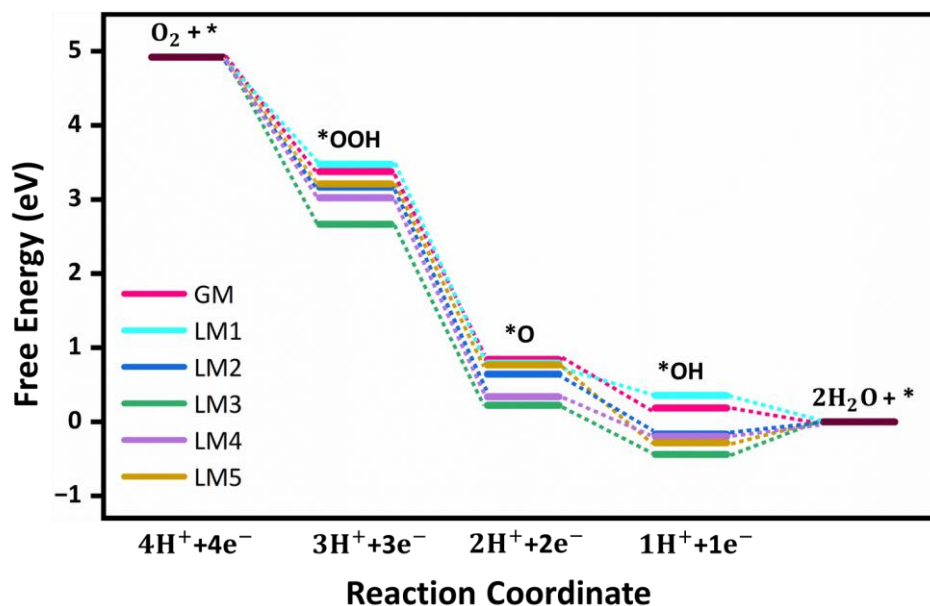


Figure A19: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt_{12}/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

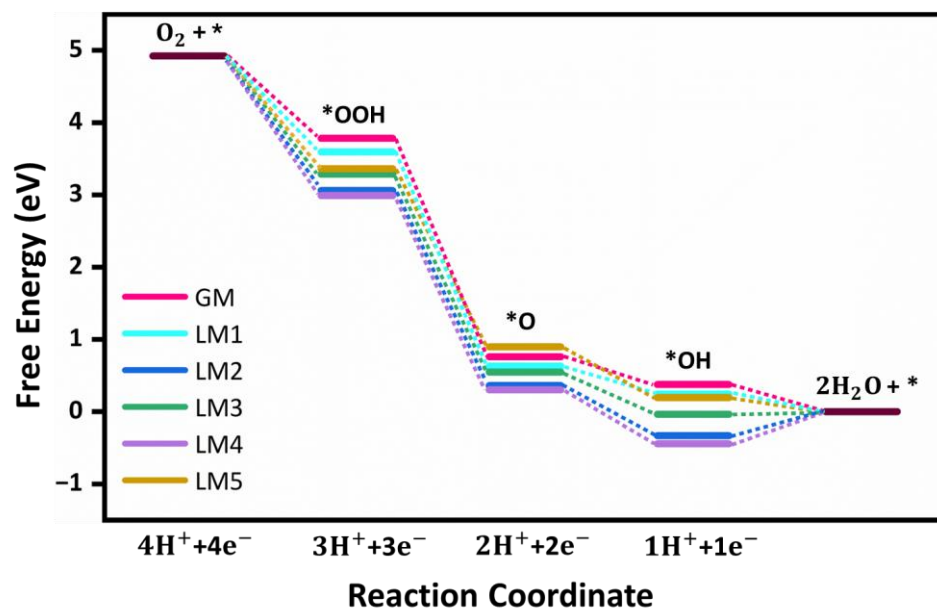


Figure A20: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt_{13}/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

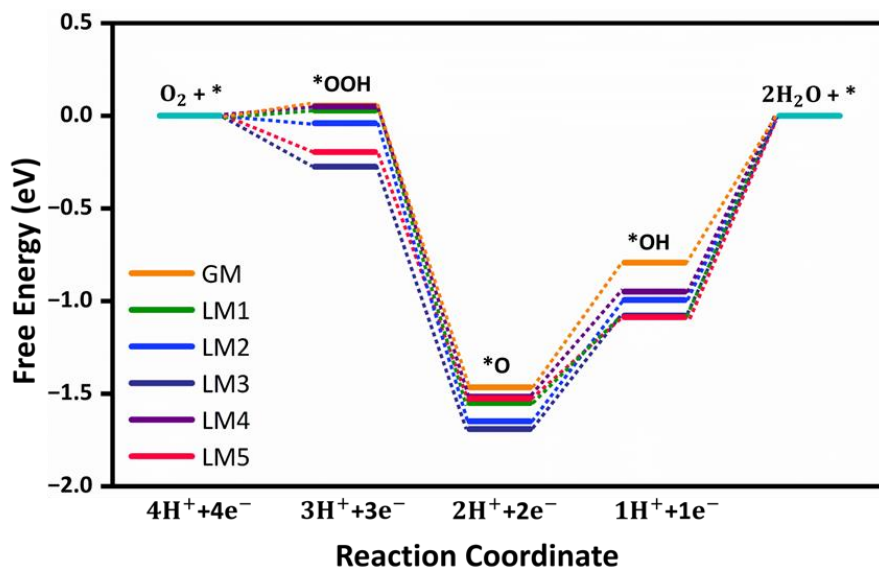


Figure A21: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt_7/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

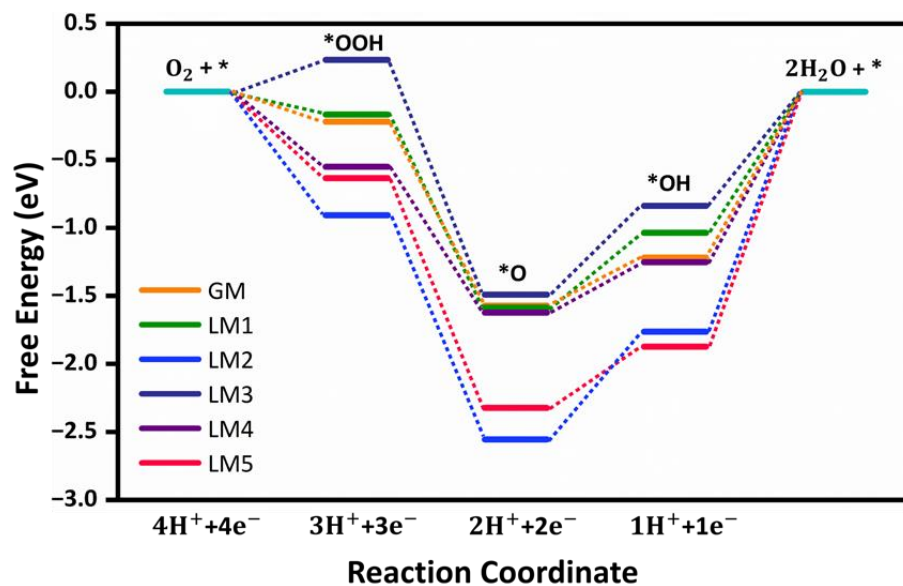


Figure A22: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt_8/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

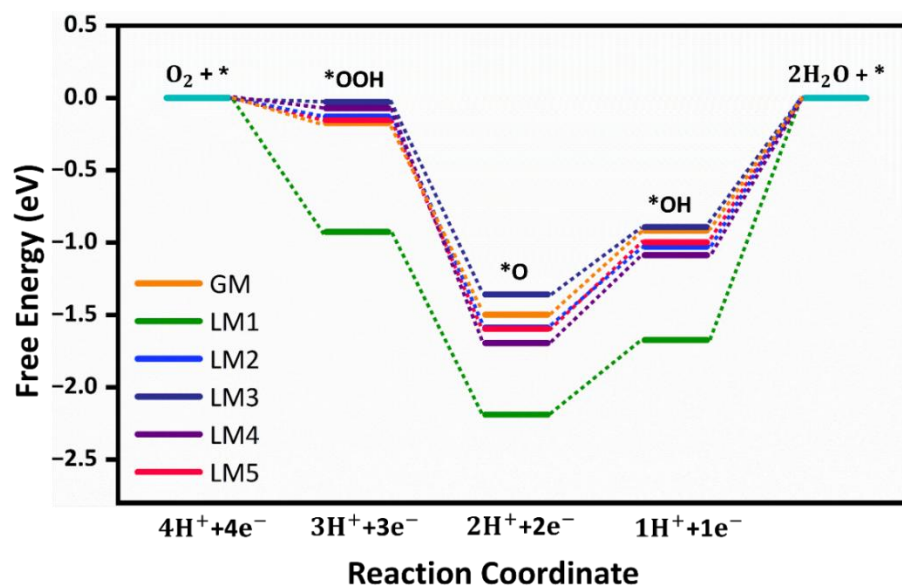


Figure A23: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt_9/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

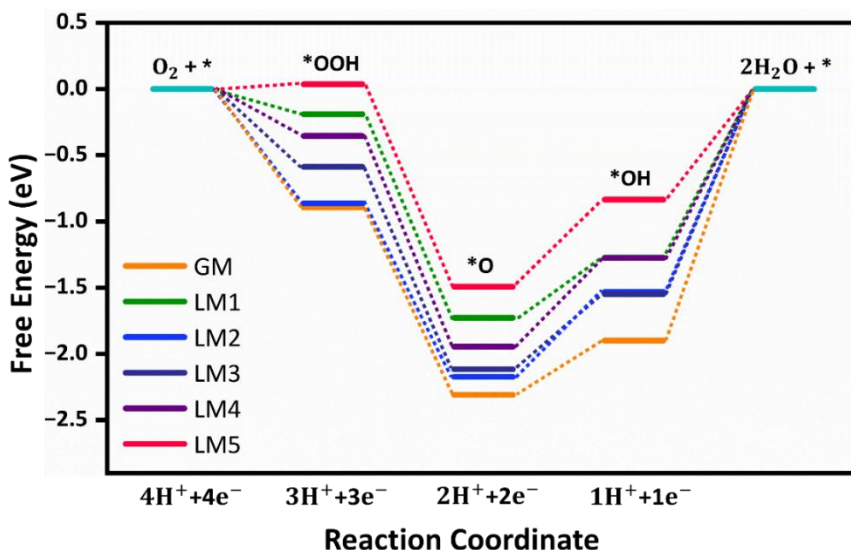


Figure A24: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt_{10}/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

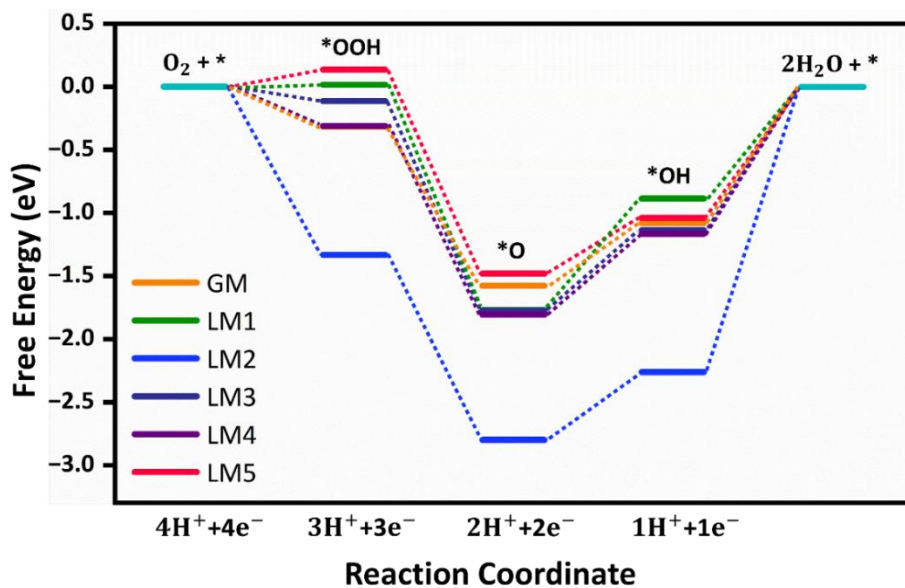


Figure A25: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt_{11}/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

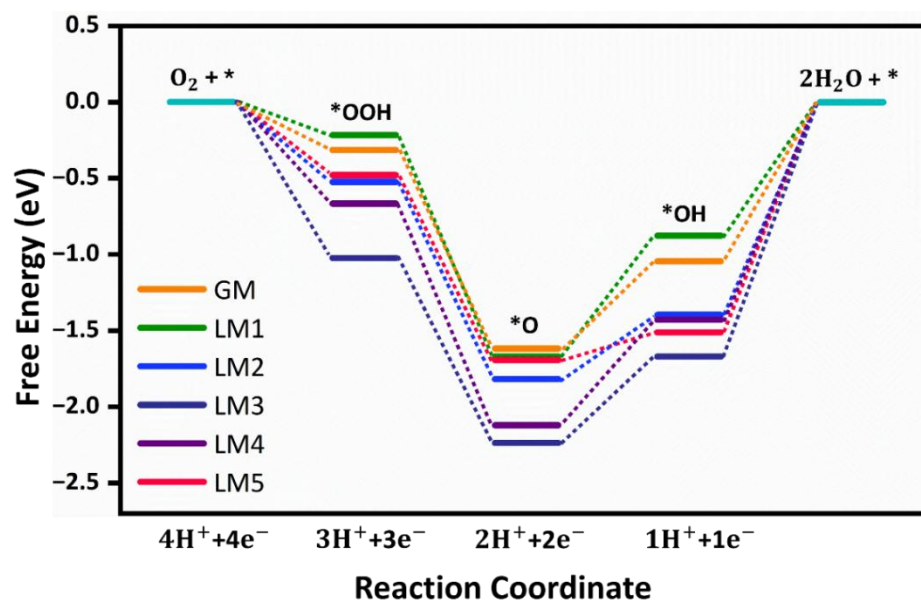


Figure A26: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt₁₂/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

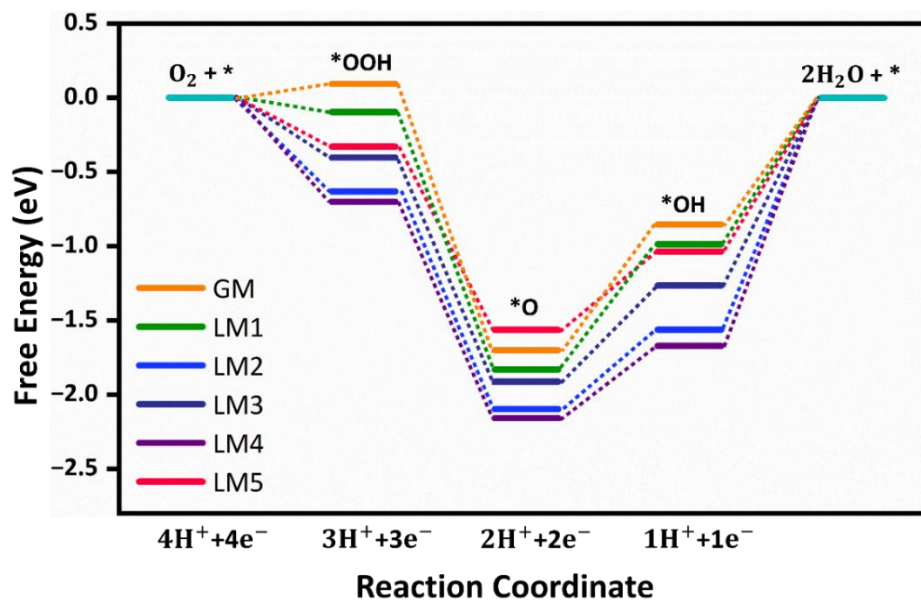


Figure A27: Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt₁₃/G within the LEME. Here, the asterisk * represents active sites of the catalyst.

Table A3. Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt₇/G within the LEME.

Isomers	0 V					1.23 V				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
GM	4.92	3.75	0.99	0.44	0.00	0.00	0.06	-1.47	-0.79	0.00
LM1	4.92	3.72	0.91	0.15	0.00	0.00	0.03	-1.55	-1.08	0.00
LM2	4.92	3.65	0.81	0.24	0.00	0.00	-0.04	-1.65	-0.99	0.00
LM3	4.92	3.42	0.77	0.15	0.00	0.00	-0.27	-1.69	-1.08	0.00
LM4	4.92	3.74	0.94	0.28	0.00	0.00	0.04	-1.52	-0.95	0.00
LM5	4.92	3.5	0.93	0.14	0.00	0.00	-0.20	-1.53	-1.09	0.00

Table A4. Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt₈/G within the LEME.

Isomers	0 V					1.23 V				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
GM	4.92	3.47	0.89	0.01	0.00	0.00	-0.22	-1.57	-1.22	0.00
LM1	4.92	3.52	0.87	0.19	0.00	0.00	-0.17	-1.59	-1.04	0.00
LM2	4.92	2.78	-0.10	-0.53	0.00	0.00	-0.91	-2.56	-1.76	0.00
LM3	4.92	3.93	0.97	0.39	0.00	0.00	0.24	-1.49	-0.84	0.00
LM4	4.92	3.14	0.84	-0.02	0.00	0.00	-0.55	-1.62	-1.25	0.00
LM5	4.92	3.05	0.14	-0.64	0.00	0.00	-0.63	-2.32	-1.87	0.00

Table A5. Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt₉/G within the LEME.

Isomers	0 V					1.23 V				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
GM	4.92	3.52	0.96	0.31	0.00	0.00	-0.17	-1.50	-0.92	0.00
LM1	4.92	2.76	0.27	-0.44	0.00	0.00	-0.93	-2.19	-1.67	0.00
LM2	4.92	3.56	0.87	0.20	0.00	0.00	-0.13	-1.59	-1.03	0.00
LM3	4.92	3.66	1.10	0.33	0.00	0.00	-0.03	-1.36	-0.89	0.00
LM4	4.92	3.62	0.77	0.14	0.00	0.00	-0.07	-1.69	-1.09	0.00
LM5	4.92	3.54	0.86	0.23	0.00	0.00	-0.15	-1.60	-1.00	0.00

Table A6. Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt₁₀/G within the LEME.

Isomers	0 V					1.23 V				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
GM	4.92	2.80	0.15	-0.67	0.00	0.00	-0.89	-2.30	-1.90	0.00
LM1	4.92	3.50	0.73	-0.04	0.00	0.00	-0.19	-1.73	-1.27	0.00
LM2	4.92	2.83	0.29	-0.30	0.00	0.00	-0.86	-2.17	-1.53	0.00
LM3	4.92	3.10	0.34	-0.32	0.00	0.00	-0.59	-2.12	-1.55	0.00
LM4	4.92	3.34	0.51	-0.04	0.00	0.00	-0.35	-1.95	-1.28	0.00
LM5	4.92	3.73	0.97	0.39	0.00	0.00	0.04	-1.49	-0.84	0.00

Table A7. Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 0 V via the associative pathway for different isomers of Pt₁₁/G within the LEME.

Isomers	0 V					1.23 V				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
GM	4.92	3.37	0.88	0.15	0.00	0.00	-0.32	-1.58	-1.08	0.00
LM1	4.92	3.71	0.69	0.34	0.00	0.00	0.02	-1.77	-0.89	0.00
LM2	4.92	2.36	-0.34	-1.03	0.00	0.00	-1.33	-2.80	-2.26	0.00
LM3	4.92	3.58	0.69	0.09	0.00	0.00	-0.11	-1.77	-1.14	0.00
LM4	4.92	3.34	0.66	0.07	0.00	0.00	-0.31	-1.80	-1.16	0.00
LM5	4.92	3.83	0.98	0.19	0.00	0.00	0.13	-1.48	-1.04	0.00

Table A8. Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR at 1.23 V via the associative pathway for different isomers of Pt₁₂/G within the LEME.

Isomers	0 V					1.23 V				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
GM	4.92	3.38	0.84	0.19	0.00	0.00	-0.31	-1.62	-1.04	0.00
LM1	4.92	3.47	0.79	0.35	0.00	0.00	-0.22	-1.67	-0.88	0.00
LM2	4.92	3.17	0.64	-0.17	0.00	0.00	-0.52	-1.82	-1.40	0.00
LM3	4.92	2.67	0.22	-0.44	0.00	0.00	-1.02	-2.24	-1.67	0.00
LM4	4.92	3.02	0.34	-0.20	0.00	0.00	-0.67	-2.12	-1.43	0.00
LM5	4.92	3.21	0.77	-0.28	0.00	0.00	-0.48	-1.69	-1.51	0.00

Table A9. Gibbs free energy changes are associated with the elementary steps (R1-R5) of the ORR via the associative pathway for different isomers of Pt_{13}/G within the LEME.

Isomers	0 V					1.23 V				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
GM	4.92	3.78	0.76	0.38	0.00	0.00	0.09	-1.70	-0.85	0.00
LM1	4.92	3.59	0.63	0.24	0.00	0.00	-0.10	-1.83	-0.99	0.00
LM2	4.92	3.05	0.36	-0.33	0.00	0.00	-0.63	-2.10	-1.56	0.00
LM3	4.92	3.29	0.55	-0.03	0.00	0.00	-0.40	-1.91	-1.26	0.00
LM4	4.92	2.99	0.30	-0.44	0.00	0.00	-0.70	-2.16	-1.67	0.00
LM5	4.92	3.36	0.90	0.19	0.00	0.00	-0.33	-1.56	-1.04	0.00

Table A10. Identified rate-determining steps (RDS) for various metastable isomers of each Pt_n/G ($n = 7-13$) SNCs, evaluated under gas-phase conditions following the associative mechanism of the ORR.

Isomers	GM	LM1	LM2	LM3	LM4	LM5
Pt_7/G	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$
Pt_8/G	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$
Pt_9/G	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$
Pt_{10}/G	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$
Pt_{11}/G	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$
Pt_{12}/G	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$
Pt_{13}/G	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$	$\text{OH}^* \rightarrow \text{H}_2\text{O}(\text{l})$

4. Ab Initio Phase Behavior

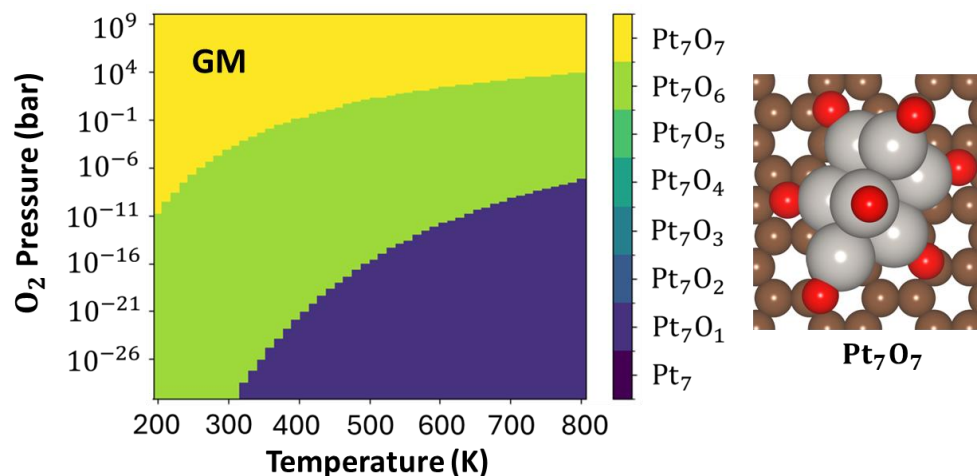


Figure A28. Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt₇O_x (x=1-7).

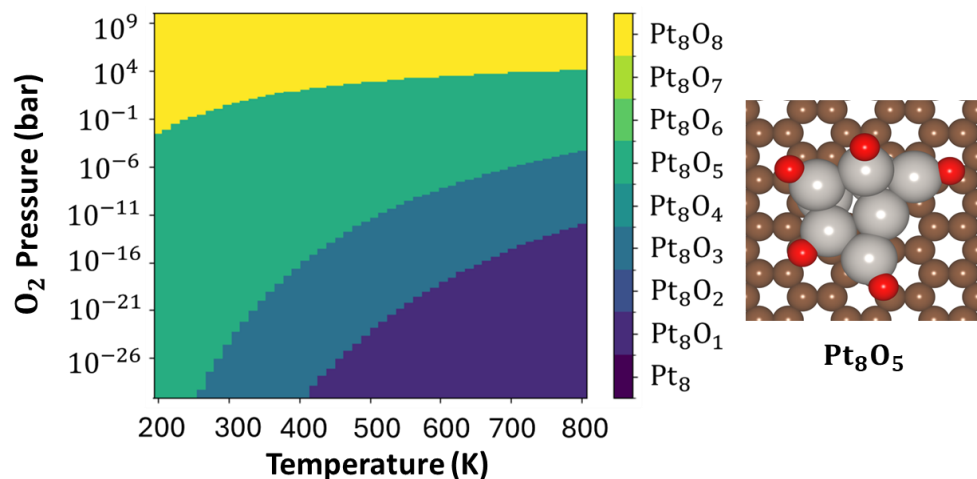


Figure A29. Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt₈O_x (x=1-8).

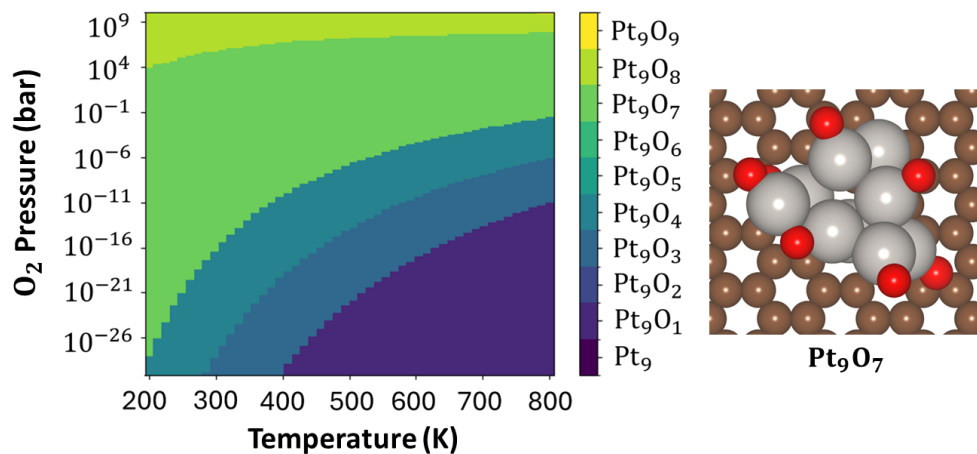


Figure A30. Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt₉O_x (x=1–9).

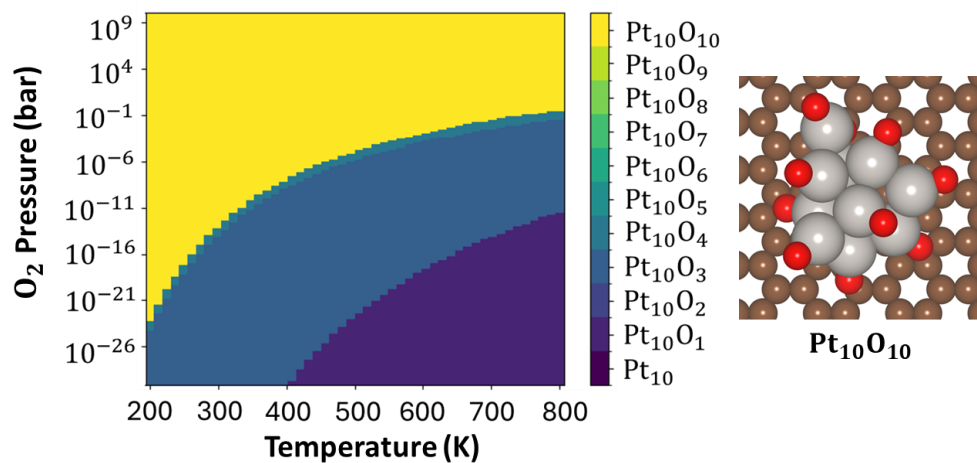


Figure A31. Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt₁₀O_x (x =1–10).

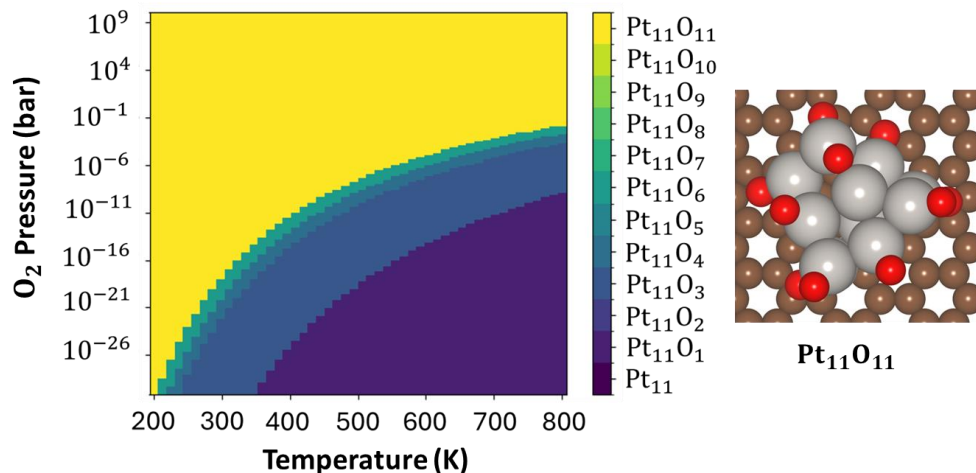


Figure A32. Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt₁₁O_x (x = 1–11).

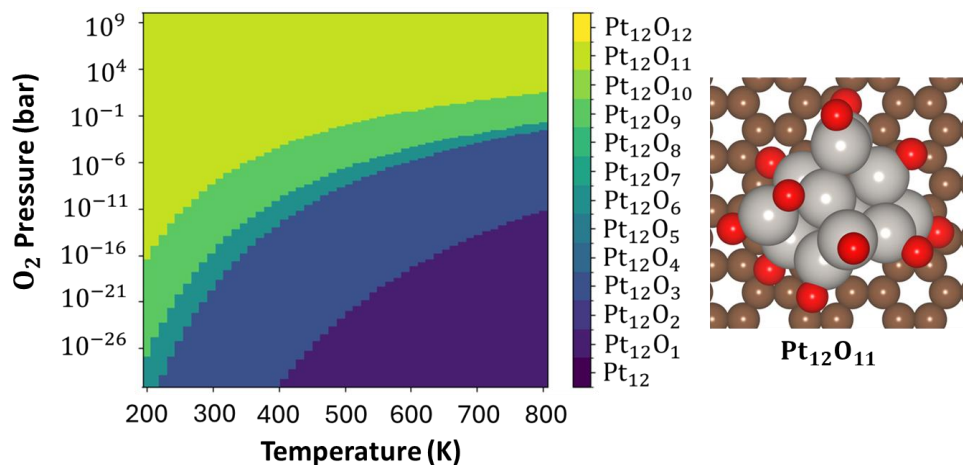


Figure A33. Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt₁₂O_x (x = 1–12).

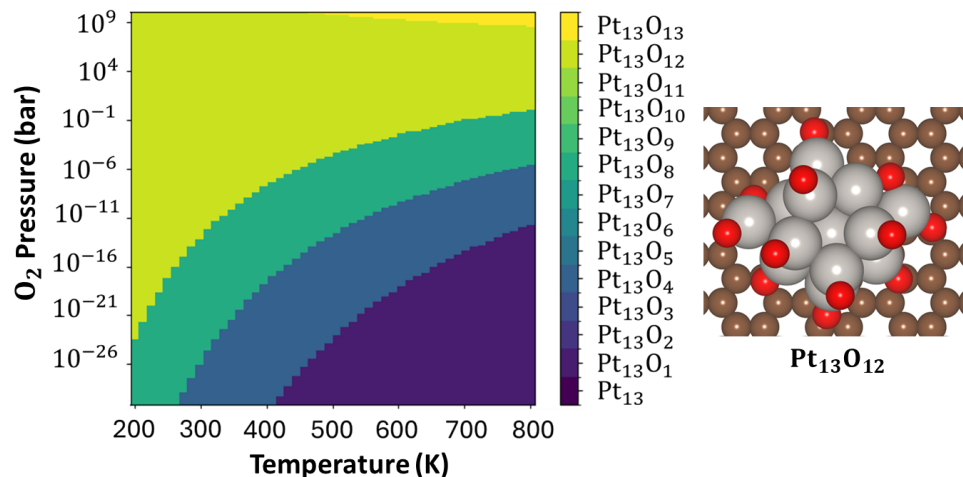


Figure A34. Second-order phase diagrams and most oxidized configurations of thermodynamically active metastable isomers: oxygen coverage (number of O atoms) as a function of oxygen chemical potential for Pt_{13}O_x ($x = 1-13$) Pt_n/G SNCs.

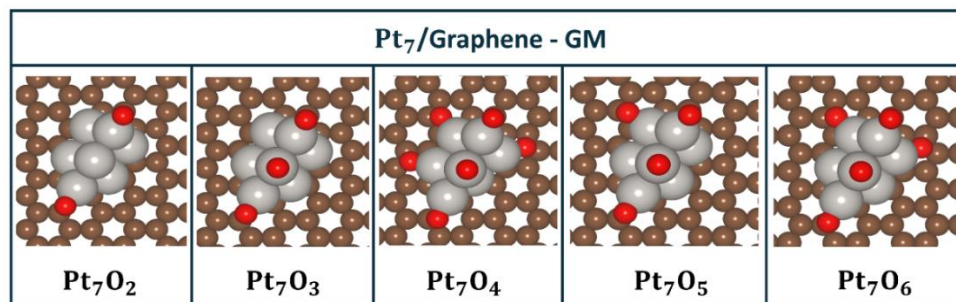


Figure A35: DFT optimized oxidized structures of the most active GM isomer of $\text{Pt}_7\text{O}_x/\text{G}$ ($x=2-6$) at varying oxygen coverage.

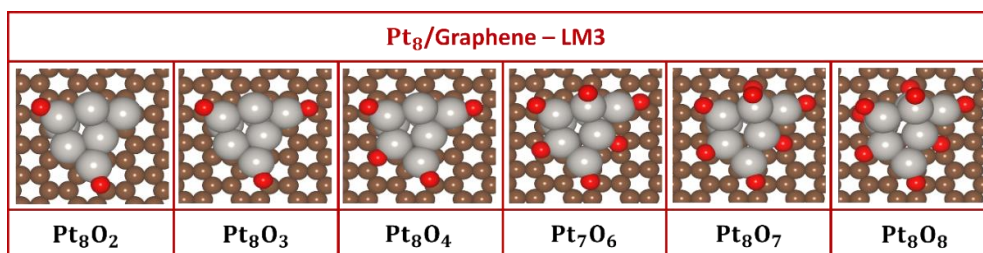


Figure A36. DFT optimized oxidized structure of the most active LM3 isomer of $\text{Pt}_8\text{O}_x/\text{G}$ ($x=2-8$, $x \neq 5$) at varying oxygen coverage.

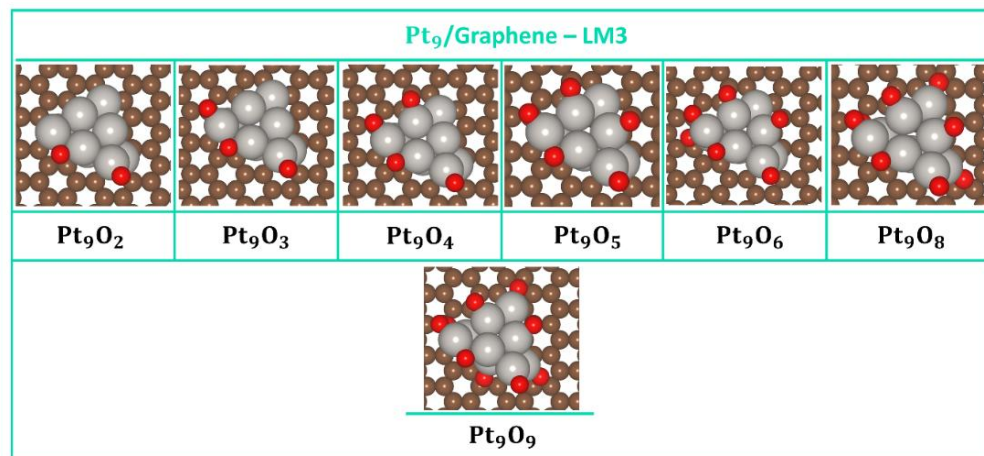


Figure A37. DFT optimized oxidized structures of the most active LM3 isomer of Pt₉O_x/G (x=2-9, x≠7) at varying oxygen coverage.

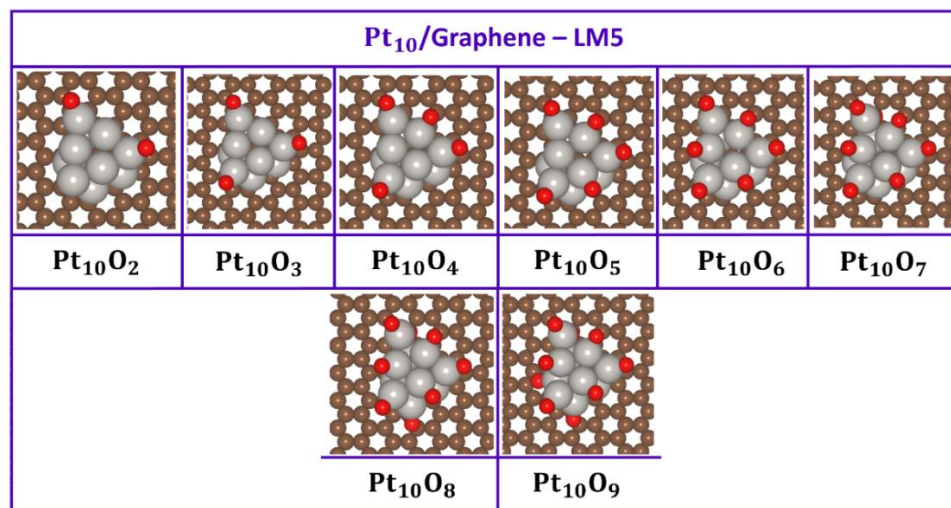


Figure A38. DFT optimized oxidized structures of the most active LM5 isomer of Pt₁₀O_x/G (x=2-9) at varying oxygen coverage..

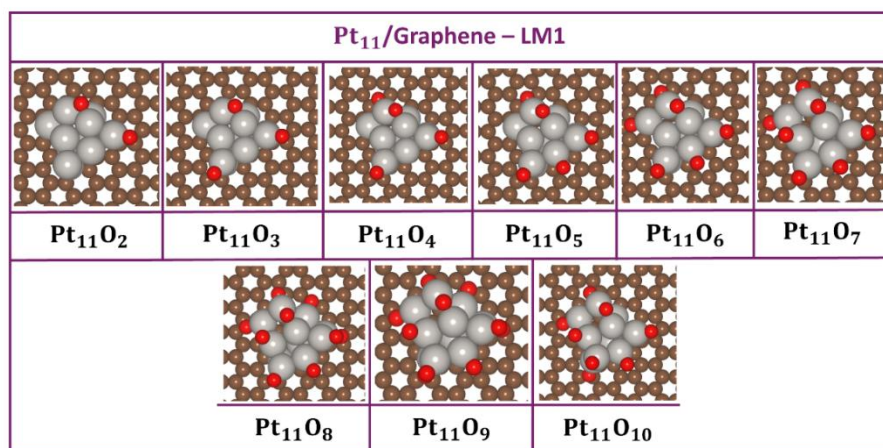


Figure A39: DFT optimized oxidized structures of the most active LM1 isomer of Pt₁₁O_x/G (x=2-11, x≠11) at varying oxygen coverage.

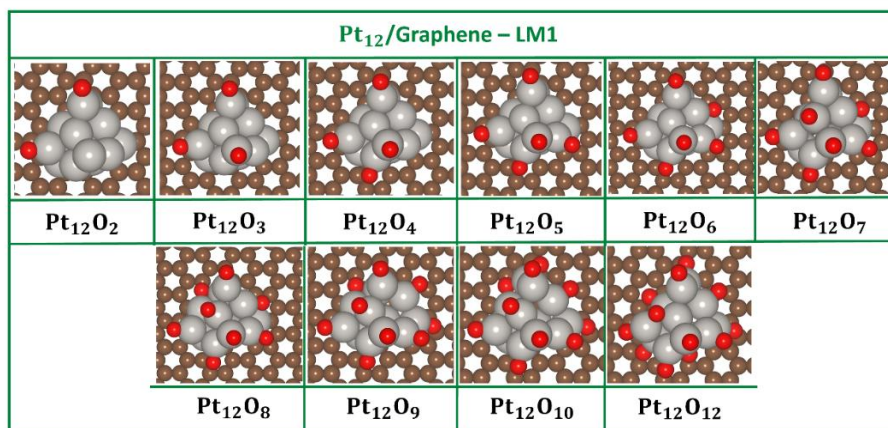


Figure A41: DFT optimized oxidized structures of the most active LM1 isomer of Pt₁₂O_x/G (x=2-12, x≠11) at varying oxygen coverage.

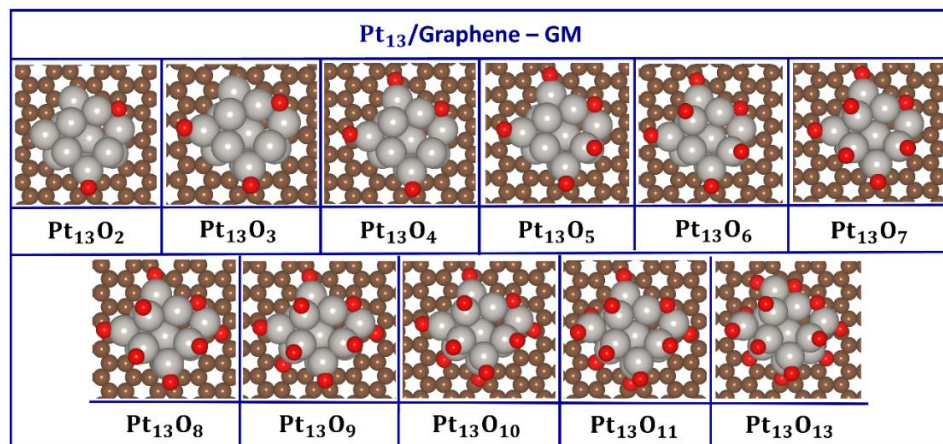


Figure A42: DFT optimized oxidized structures of the most active GM isomer of Pt₁₃O_x/G (x=2-13, x≠12) at varying oxygen coverage.

6. Feature Description

Table A11: Refined set of descriptors retained for machine learning analysis following the elimination of highly correlated features.

Groups	Descriptors	Symbol
Electronic Descriptors	Total number of d electrons	Σd_n
Geometric Descriptors	Number of sites (top, bridge)	S
	Coordination number	CN
	Distance between the adsorbate and the surface	D_{AS}
	Distance between Pt and the surface	Z
	Pt adsorbate bond length	D_{AP}
	Number of Pt atoms that bind to the surface	A

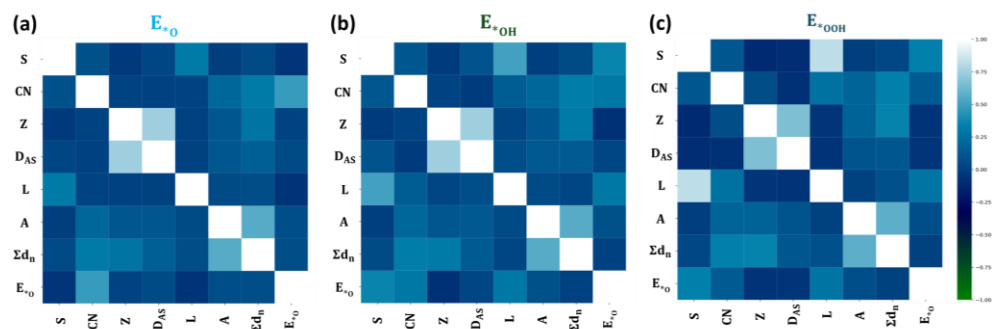


Figure A43. Pearson's correlation coefficient (PCC) matrices for adsorption energy datasets: Correlation matrices illustrating feature-feature and feature-target relationships for (a) E^*_o , (b) E^*_{OH} , and (c) E^*_{OOH} datasets, computed using the extracted 7 features.

REFERENCES

1. Jin, S. (2020). COVID-19, Climate Change, and Renewable Energy Research: We Are All in This Together, and the Time to Act Is Now. *ACS Energy Lett.*, 5, 1709–1711 (DOI: 10.1021/acsenergylett.0c00910)
2. Zhu, J., Hu, L., Zhao, P., Lee, L. Y. S., Wong, K.-Y. (2020). Recent Advances in Electrocatalytic Hydrogen Evolution Using Nanoparticles. *Chem. Rev.*, 120, 851–918 (DOI: 10.1021/acs.chemrev.9b00248)
3. Giannouli, M. (2020). Current Status of Emerging PV Technologies: A Comparative Study of Dye-Sensitized, Organic, and Perovskite Solar Cells. *Int. J. Photoenergy*, 2021, 1–19 (DOI: 10.1155/2021/6692858)
4. Butler, K. T., Sai Gautam, G., Canepa, P. (2019). Designing Interfaces in Energy Materials Applications with First-Principles Calculations. *npj Comput. Mater.*, 5, 19 (DOI: 10.1038/s41524-019-0160-9)
5. Lubitz, W., Tumas, W. (2007). Hydrogen: An Overview. *Chem. Rev.*, 107, 3900–3903 (DOI: 10.1021/cr050200z)
6. Albarbar, A., Alrweq, M. (2018). Proton Exchange Membrane Fuel Cells: Design, Modelling, and Performance Assessment Techniques. Springer International Publishing, pp. 4–5 (ISBN: 978-3-319-70726-6)
7. O'Hare, R., Cha, S. W., Collella, W. G., Prinz, F. B. (2016). Fuel Cell Fundamentals, 3rd ed. Wiley, USA, pp. 6–9 (ISBN: 978-1-119-11380-5)
8. Qin, C., Wang, J., Yang, D., Li, B., Zhang, C. (2016). Proton Exchange Membrane Fuel Cell Reversal: A Review. *Catalysts*, 6, 197 (DOI: 10.3390/catal6120197)
9. Mahata, A., Nair, A. S., Pathak, B. (2019). Recent Advancements in Pt-Nanostructure-Based Electrocatalysts for the Oxygen Reduction Reaction. *Catal. Sci. Technol.*, 9, 4835–4863 (DOI: 10.1039/C9CY00895K)
10. Zhang, J., Zhao, Z.; Xia, Z.; Dai, L. A metal-free bifunctional electrocatalyst for oxygen reduction and oxygen evolution

- reactions. *Nature Nanotechnol.* 2015, 10, 444 (DOI: 10.1038/nnano.2015.48)
11. Ekspong, J., Gracia-Espino, E., Wagberg, T. Hydrogen Evolution Reaction Activity of Heterogeneous Materials: A Theoretical Model. *J. Phys. Chem. C* 2020, 124 (38), 20911– 20921 (DOI: 10.1021/acs.jpcc.0c05243)
 12. Chen, Z., Gariepy, Z., Chen, L., Yao, X., Anand, A., Liu, S., Giresse, C., Feugmo, T., Tamblyn, I., Singh, C. V. Machine-learning-driven high-entropy alloy catalyst discovery to circumvent the scaling relation for CO₂ reduction reaction. *ACS Catal.* 2022, 12, 14864– 14871 (DOI: 10.1021/acscatal.2c03675)
 13. Xue, Z., Zhang, X., Qin, J., Liu, R. High-throughput identifications of high activity and selectivity transition metal single-atom catalysts for nitrogen reduction. *Nano Energy* 2021, 80, 105527 (DOI: 10.1016/j.nanoen.2020.105527)
 14. Batchelor, T. A. A., Pedersen, J. K., Winther, S. H.; Castelli, I. E., Jacobsen, K. W., Rossmeisl, J. High-Entropy Alloys as a Discovery Platform for Electrocatalysis. *Joule* 2019, 3 (3), 834– 845 (DOI: 10.1016/j.joule.2018.12.015)
 15. Qian, Y., Liu, Y., Zhao, Y., Zhang, X., Yu, G. Single vs Double Atom Catalyst for N₂ Activation in Nitrogen Reduction Reaction: A DFT Perspective. *EcoMat* 2020, 2 (1), e12014 (DOI: 10.1002/eom2.12014)
 16. Chittibabu, D. K. D.; Chen, H.-T. Computational Design of Two-Dimensional Transition Metal Supported Biphenylene as Efficient Electrocatalysts Toward Nitrogen Reduction Reaction. *Electrochim. Acta* 2024, 497, 144578 (DOI: j.electacta.2024.144578)
 17. Ali, S., Fu Liu, T., Lian, Z., Li, B., Sheng Su, D. The Effect of Defects on the Catalytic Activity of Single Au Atom Supported Carbon Nanotubes and Reaction Mechanism for Co Oxidation. *Phys. Chem. Chem. Phys.* 2017, 19, 22344– 22354 (DOI: 10.1039/C7CP03793G)

18. Khanna S. N., Jena P. (1995), Atomic clusters: Building blocks for a class of solids, *Phys. Rev. B*, 51, 13705–16 (DOI: 10.1103/PhysRevB.51.13705)
19. Somorjai G. A., Contreras A. M., Montano M., Rioux R. M. (2006), Clusters, surfaces, and catalysis. *Proc. Natl. Acad. Sci. U S A*, 103, 10577–10583 (DOI: 10.1073/pnas.0507691103)
20. Ferrando R., Jellinek J., Johnston R. L. (2008), Nanoalloys: From theory to applications of alloy clusters and nanoparticles. *Chem. Rev.*, 108, 845–910 (DOI: 10.1021/cr040090g)
21. Sanchez A., Abbet S., Heiz U., Schneider, W.-D., Häkkinen, H. N., Barnett R, et al., (1999), When Gold is not noble: Nanoscale gold catalysts, *J. Phys. Chem. A*, 103, 9573–9578 (DOI: 0.1021/jp9935992)
22. Mao, X., Wang, L., Xu, Y. et al. (2021), Computational high-throughput screening of alloy nanoclusters for electrocatalytic hydrogen evolution, *npj Comput. Mater.*, 7, 46 (DOI: 10.1038/s41524-021-00514-8)
23. Mahata, A., Rawat, K. S., Choudhuri, I., Pathak, B. (2016), Cuboctahedral: vs. octahedral platinum nanoclusters: Insights into the shape-dependent catalytic activity for fuel cell applications, *Catal. Sci. Technol.*, 6, 7913–7923 (DOI: 10.1039/C6CY01709F)
24. Liu, L., Corma, A. (2018), Metal Catalysts for Heterogeneous Catalysis: From Single Atoms to Nanoclusters and Nanoparticles, *Chem. Rev.*, 118, 4981–5079 (DOI: 10.1021/acs.chemrev.7b00776)
25. Jin, R. (2015), Atomically precise metal nanoclusters: Stable sizes and optical properties. *Nanoscale*, 7, 1549-1565 (DOI: 10.1039/C4NR05794E)
26. Chu, S., Majumdar, A. Opportunities and challenges for a sustainable energy future. *Nature* 2012, 488, 294-303 (DOI: 10.1038/nature11475)
27. Joya, K. S., Joya, Y. F., Ocakoglu, K., van de Krol, R. Water-splitting catalysis and solar fuel devices: artificial leaves on the move. *Angew. Chem. Int. Ed* 2013, 52, 10426– 10437 (DOI: 10.1002/anie.201300136)

28. Lewis, N. S., Nocera, D. G. Powering the planet: Chemical challenges in solar energy utilization. *Proc. Natl. Acad. Sci. U. S. A* 2006, 103, 15729– 15735 (DOI: 10.1073/pnas.0603395103)
29. Mekhilef, S., Saidur, R., Safari, A. Comparative study of different fuel cell technologies. *Renew. Sustain. Energy Rev.* 2012, 16, 981– 989 (DOI: 10.1016/j.rser.2011.09.020)
30. Watanabe, M., Tryk, D. A.; Wakisaka, M., Yano, H., Uchida, H. Overview of Recent Developments in Oxygen Reduction Electrocatalysis. *Electrochim. Acta* 2012, 84, 187– 201 (DOI: 10.1016/j.electacta.2012.04.035)
31. Shao, M., Chang, Q., Dodelet, J.-P., Chenitz, R. Recent Advances in Electrocatalysts for Oxygen Reduction Reaction. *Chem. Rev.* 2016, 116 (6), 3594– 3657 (DOI: 10.1021/acs.chemrev.5b00462)
32. Kulkarni, A.; Siahrostami, S., Patel, A., Nørskov, J. K. Understanding Catalytic Activity Trends in the Oxygen Reduction Reaction. *Chem. Rev.* 2018, 118 (5), 2302– 2312 (DOI: 10.1021/acs.chemrev.7b00488)
33. Yoo, E., Okata, T., Akita, T., Kohyama, M., Nakamura, J., Honma, I. Enhanced Electrocatalytic Activity of Pt Subnanoclusters on Graphene Nanosheet Surface. *Nano Lett.* 2009, 9, 2255– 2259 (DOI: 10.1021/nl900397t)
34. Felix, J. P. C. S., Batista, K. E. A., Morais, W. O., Nagurniak, G. R., Orenha, R. P., Rêgo, C. R. C., Guedes-Sobrinho, D., Parreira, R. L. T., Ferrer, M. M., Piotrowski, M. J. Molecular adsorption on coinage metal subnanoclusters: A DFT+D3 investigation. *J. Comput. Chem.* 2023, 44, 1040– 1051 (DOI: 10.1002/jcc.27063)
35. Zandkarimi, B., Alexandrova, A. N. Can Fluxionality of Subnanometer Cluster Catalysts Solely Cause Non-Arrhenius Behavior in Catalysis?. *J. Phys. Chem. C* 2020, 124, 19556– 19562 DOI: 10.1021/acs.jpcc.0c04136
36. Zhang, Z., Zandkarimi, B., Munarriz, J., Dickerson, C. E., Alexandrova, A. N. Fluxionality of Subnano Clusters Reshapes the

- Activity Volcano of
Electrocatalysis. *ChemCatChem* 2022, 14, e202200345 (DOI: 10.1002/cctc.202200345)
37. Zhang, Z., Zandkarimi, B., Alexandrova, A. N. Ensembles of Metastable States Govern Heterogeneous Catalysis on Dynamic Interfaces. *Acc. Chem. Res.* 2020, 53 (2), 447– 458 (DOI: 10.1021/acs.accounts.9b00531)
 38. Li, G., Jin, R. Atomically precise gold nanoclusters as new model catalysts. *Acc. Chem. Res.* 2013, 46 (8), 1749– 1758 (DOI: 10.1021/ar300213z)
 39. Lim, D.-H., Wilcox, J. DFT-Based Study on Oxygen Adsorption on Defective Graphene-Supported Pt Nanoparticles. *J. Phys. Chem. C* 2011, 115, 22742– 22747 (DOI: 10.1021/jp205244m)
 40. Kumari, S., Masubuchi, T., White, H. S., Alexandrova, A., Anderson, S. L., Sautet, P. Electrocatalytic Hydrogen Evolution at Full Atomic Utilization over ITO-Supported Sub-nano-Pt_n Clusters: High, Size-Dependent Activity Controlled by Fluxional Pt Hydride Species. *J. Am. Chem. Soc.* 2023, 145 (10), 5834– 5845 (DOI: 10.1021/jacs.2c13063)
 41. Zandkarimi, B., Alexandrova, A. N. Dynamics of Subnanometer Pt Clusters Can Break the Scaling Relationships in Catalysis. *J. Phys. Chem. Lett.* 2019, 10, 460– 467 (DOI: 10.1021/acs.jpcllett.8b03680)
 42. Zhang, Z.; Zandkarimi, B., Munarriz, J., Dickerson, C. E., Alexandrova, A. N. Fluxionality of Subnano Clusters Reshapes the Activity Volcano of
Electrocatalysis. *ChemCatChem* 2022, 14, e202200345 (DOI: 10.1002/cctc.202200345)
 43. Zhang, Z., Cui, Z. H., Jimenez-Izal, E., Sautet, P., Alexandrova, A. N. Hydrogen Evolution on Restructured B-Rich WB: Metastable Surface States and Isolated Active Sites. *ACS Catal.* 2020, 10, 13867– 13877 (DOI: 10.1021/acscatal.0c03410)

44. Yoon, A., Poon, J., Grosse, P., Chee, S. W., Cuenya, B. R. Iodide-mediated Cu Catalyst Restructuring during CO₂ Electroreduction. *J. Mater. Chem. A* 2022, 10, 14041– 14050 (DOI: 10.1039/D1TA11089F)
45. Wu, L. P., Guo, T., Li, T. Rational Design of Transition Metal Single-Atom Electrocatalysts: A Simulation-Based, Machine Learning-Accelerated Study. *J. Mater. Chem. A* 2020, 8, 19290– 19299 (DOI: 10.1039/D0TA06207C)
46. Wan, X. H., Zhang, Z. F., Niu, H., Yin, Y. H., Kuai, C. G., Wang, J., Shao, C., Guo, Y. Z. Machine-Learning-Accelerated Catalytic Activity Predictions of Transition Metal Phthalocyanine Dual-Metal-Site Catalysts for CO₂ Reduction. *J. Phys. Chem. Lett.* 2021, 12, 6111– 6118 (DOI: 10.1021/acs.jpcclett.1c01526)
47. Chen, A., Zhang, X., Zhou, Z. Machine Learning: Accelerating Materials Development for Energy Storage and Conversion. *InfoMat* 2020, 2, 553– 576 (DOI: 10.1002/inf2.12094)
48. Zhong, M., Tran, K., Min, Y. M., Wang, C. H., Wang, Z. Y., Dinh, C. T., De Luna, P., Yu, Z. Q., Rasouli, A. S., Brodersen, P., Sun, S.; Voznyy, O., Tan, C. S., Askerka, M.; Che, F. L., Liu, M., Seifitokaldani, A., Pang, Y. J., Lo, S. C., Ip, A., Ulissi, Z., Sargent, E. H. Accelerated Discovery of CO₂ Electrocatalysts Using Active Machine Learning. *Nature* 2020, 581, 178– 183 (DOI: 10.1038/s41586-020-2242-8)
49. Wang, X. M., Xiao, B., Li, Y. H., Tang, Y. C., Liu, F., Chen, J. H., Liu, Y. First-Principles Based Machine Learning Study of Oxygen Evolution Reactions of Perovskite Oxides Using a Surface Center-Environment Feature Model. *Appl. Surf. Sci.* 2020, 531, 147323 (DOI: 10.1016/j.apsusc.2020.147323)
50. Secor, M., Soudackov, A. V., Hammes-Schiffer, S. Density Matrix-Based Features as Descriptors for Oxygen Reduction and Evolution

- Catalysts. J. Phys. Chem. C 2023, 127, 15246– 15256 (DOI: 10.1021/acs.jpcc.3c03392)
51. Xing, M. J., Zhang, Y. J., Li, S. Y., He, H., Sun, S. R. Prediction of Carbon Dioxide Reduction Catalyst Using Machine Learning with a Few-Feature Model: WLEDZ. J. Phys. Chem. C 2022, 126, 17025– 17035 (DOI: 10.1021/acs.jpcc.2c0211)
 52. Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M., Fazzio, A. From DFT to Machine Learning: Recent Approaches to Materials Science—a Review. J. Phys. Mater. 2019, 2, 032001 (DOI: 10.1088/2515-7639/ab084b)
 53. Wei, J.; Chu, X.; Sun, X. Y.; Xu, K.; Deng, H. X.; Chen, J. G.; Wei, Z. M.; Lei, M. Machine Learning in Materials Science. InfoMat 2019, 1, 338– 358 (DOI: 10.1002/inf2.12028)
 54. Jordan, M. I., Mitchell, T. M. Machine Learning: Trends, Perspectives, and Prospects. Science 2015, 349, 255– 260 (DOI: 10.1126/science.aaa8415)
 55. Born M., Oppenheimer, J. R. (1928). Zur quanten theorie der molekeln. Ann. Phys., 84, 457–484 (DOI: 10.1002/andp.19273892002)
 56. Hohenberg, P., Kohn, W. (1964). Inhomogeneous electron gas. Phys. Rev., 136, B864–B871 (DOI: 10.1103/PhysRev.136.B864)
 57. Kohn, W., Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. Phys. Rev., 140, A1133–A1138 (DOI: 10.1103/PhysRev.140.A1133)
 58. Hohenberg P. C., Kohn W., Sham L. J. (1990), The Beginnings and Some Thoughts on the Future, Advances in Quantum Chemistry, 21, 7– 26 (DOI: 10.1016/S0065-3276(08)60589-4)
 59. Martin R. M. (2004). Electronic structure: basic theory and practical methods. Cambridge University Press.
 60. Ceperley D. M., Alder B. J. (1980). Ground State of the Electron Gas by a Stochastic Method. Phys. Rev. Lett., 45, 566 (DOI: 10.1103/PhysRevLett.45.566)

61. Perdew J. P., Burke K., Ernzerhof M. (1996). Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77, 3865–3868 (DOI: 10.1103/PhysRevLett.77.3865)
62. Perdew J. P., Wang Y. (1992). Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B*, 45, 13244 (DOI: 10.1103/PhysRevB.45.13244)
63. Perdew J. P., Ruzsinszky A., Csonka G. I., Vydrov O. A., Scuseria G. E., Constantin L. A., Zhou X., Burke K. (2008). Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.*, 100, 136406 (DOI: 10.1103/PhysRevLett.100.136406)
64. Vanderbilt D. (1990). Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Phys. Rev. B*, 41, 7892–7895 (DOI: 10.1103/PhysRevB.41.7892)
65. Blöchl, P. E. (1994). Projector augmented-wave method. *Phys. Rev. B*, 50, 17953–17979 (DOI: 10.1103/PhysRevB.50.17953)
66. Andersen, O. K. (1975). Linear methods in band theory. *Phys. Rev. B*, 12, 3060–3083 (DOI: 10.1103/PhysRevB.12.3060)
67. Hamann, D. R. (1979). Norm-conserving pseudopotentials. *Phys. Rev. Lett.*, 43, 1494–1497 (DOI: 10.1103/PhysRevLett.43.1494)
68. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. (2010). A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, 132, 154104 (DOI: 10.1063/1.3382344)
69. Barth, U. von., Hedin, L. (1972). A local exchange-correlation potential for the spin-polarized case. *J. Phys. C: Solid State Phys.*, 5, 1629–1642 (DOI: 10.1088/0022-3719/5/13/012)
70. Pant, M. M., Rajagopal, A. K. (1972). Theory of inhomogeneous magnetic electron gases. *Solid State Commun.*, 10, 1157–1160 (DOI: 10.1016/0038-1098(72)90949-4)

71. Terakura, K., Williams, A. R., Kugler, J., Kübler, J. (1984). Transition-metal monoxides: band or Mott insulators. *Phys. Rev. Lett.*, 52, 1830 (DOI: 10.1103/PhysRevLett.52.1830)
72. Oliver, G. L., Perdew, J. P. (1979). Spin-density gradient expansion for the kinetic energy. *Phys. Rev. A*, 20, 397 (DOI: 10.1103/PhysRevA.20.397)
73. Okabayashi, J., Okabayashi, J., Rader, O., Mizokawa, T., Fujimori, A., Hayashi, T., Tanaka, M. (1998). Core-level photoemission study of $\text{Ga}_{1-x}\text{Mn}_x\text{As}$. *Phys. Rev. B*, 58, R4211 (DOI: 10.1103/PhysRevB.58.R4211)
74. Sato, K., Bergqvist, L., Kudrnovsky, J., Dederichs, P. H., Eriksson, O., Turek, I., Sanyal, B., Bouzerar, G., Katayama-Yoshida, H., Dinh, V. A., Fukushima, T., Kizaki, H., Zeller, R. (2010), First-principles theory of dilute magnetic semiconductors. *Rev. Mod. Phys.*, 82, 1633 (DOI: 10.1103/RevModPhys.82.1633)
75. Sato K., Dederichs P. H., Katayama-Yoshida H., Kudrnovsky J., (2004), Exchange interactions in diluted magnetic semiconductors. *J. Phys.: Condens. Matter*. 16, 5491 (DOI: 0953-8984/16/48/003)
76. Panchmatia P. M., Sanyal B., (2008), GGA + U modeling of structural, electronic, and magnetic properties of iron porphyrin-type molecules. *Chem. Phys.* 343, 47 (DOI: 10.1016/j.chemphys.2007.10.030)
77. Bernien M., Miguel J., Weis C., Ali Md. E., Kurde J., Krumme B., Panchmatia P. M., Sanyal B., Piantek M., Srivastava P., Baberschke K., Oppeneer P. M., Eriksson O., Kuch W., Wende H., (2009), Tailoring the nature of magnetic coupling of Fe-porphyrin molecules to ferromagnetic substrates. *Phys. Rev. Lett.* 102, 047202 (DOI: 10.1103/PhysRevLett.102.047202)
78. Shick A. B., Kudrnovsky J., Drchal V., *Phys. Rev. B* (2004), Coulomb correlation effects on the electronic structure of III-V diluted magnetic semiconductors. 69, 125207 (DOI: 10.1103/PhysRevB.69.125207)

79. Anisimov, V. I., Zaanen, J., Andersen, O. K. (1991). Band theory and Mott insulators: Hubbard U instead of Stoner I. *Phys. Rev. B*, 44, 943 (DOI: 10.1103/PhysRevB.44.943)
80. Martin R. L., Illas R., (1997), Antiferromagnetic exchange interactions from hybrid functional theory. *Phys. Rev. Lett.*, 79, 1539 (DOI: 10.1103/PhysRevLett.79.1539)
81. Becke, A. D. (1993). Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98, 5648 (DOI: 10.1063/1.464913)
82. Lee, C., Yang, W., Parr, R. G. (1988). Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37, 785 (DOI: 10.1103/PhysRevB.37.785)
83. Becke, A. D. (1993). Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98, 5648 (DOI: 10.1063/1.464913)
84. Lee, C., Yang, W., Parr, R. G. (1988). Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys.* 37, 785 (DOI: 10.1103/PhysRevB.37.785)
85. Lee C. T., Yang W. T., Parr R. G. (1988), *Phys. Rev. B: Condens. Matter Mater.* Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys.* 37, 785–789 (DOI: 10.1103/PhysRevB.37.785)
86. Hobbs D., Kresse G., Hafner J. (2000), Fully unconstrained noncollinear magnetism within the projector augmented-wave method. *Phys. Rev. B*, 62, 11556–11570. (DOI: 10.1103/PhysRevB.62.11556)
87. Ihm J., Zunger A., Cohen M. L. (1979), Momentum-space formalism for the total energy of solids, *J. Phys. C: Solid State Physics*, 12, 4409 (DOI: 10.1088/0022-3719/12/21/009)
88. Boys S. F., Bernardi F. (1970), The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors, *Mol. Phys.*, 19, 553–566 (DOI: 10.1080/00268977000101561)

89. Nørskov, J. K.; Rossmeisl, J.; Logadottir, A.; Lindqvist, L. Origin of the Overpotential for Oxygen Reduction at A Fuel-Cell Cathode. *J. Phys. Chem. B* 2004, 108, 17886– 17892 (DOI: 10.1021/jp047349j)
90. Nasteski V. (2017), An overview of the supervised machine learning methods. *Horizons. B* 4, 51–62 (DOI: 10.20544/HORIZONS.B.04.1.17.P05)
91. Ghiringhelli L. M., Vybiral J., Levchenko S. V., Draxl C., Scheffler M. (2015), Big Data of Materials Science: Critical Role of the Descriptor, *Phys. Rev. Lett.* 114, 105503 (DOI: 10.1103/PhysRevLett.114.105503)
- Pedregosa, F. et al. (2011), Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.* 12, 2825–2830
92. Hyndman R. J., Koehler A. B. (2006), Another look at measures of forecast accuracy, *Int. J. Forecast.* 22, 679–688 (DOI: 10.1016/j.ijforecast.2006.03.001)
93. Willmott C. J., Matsuura K. (2005), Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, *Climate research* 30, 79–82 (DOI: 10.3354/cr030079)
94. Ash A., Shwartz M. (1999), R^2 : a useful measure of model performance when predicting a dichotomous outcome, *Stat. Med.* 18, 375–384 (DOI: 10.1002/(sici)1097-0258(19990228)18:4<375::aid-sim20>3.0.co;2-j)
95. Schratz P., Muenchow J., Iturritxa E., Richter J., Brenning A. (2019), Hyperparameter Tuning and Performance Assessment of Statistical and Machine-Learning Algorithms Using Spatial Data, *Ecol Modell* 406, 109–120 (DOI: 10.1016/J.ECOLMODEL.2019.06.002)
96. Arafath Y., Roy A. C., Shamim K., M., Arefin M. S. (2022), Developing a Framework for Credit Card Fraud Detection, *Lecture Notes on Data Engineering and Communications Technologies* 95, 637–651 (DOI:10.1007/978-981-16-6636-0_48/COVER)
97. Browne M.W. (2000), Cross-Validation Methods, *J. Math. Psychol.* 44, 108–132 (DOI: :10.1006/jmps.1999.1279)

98. Wong T.T., Yeh P.Y. (2019), Reliable accuracy estimates from k-fold cross validation, *IEEE Trans. Knowl. Data Eng.* 32, 1586–1594 (DOI: 10.1109/TKDE.2019.2912815)
99. Berrar D. (2019), Cross-validation, *Encyclopedia of Bioinformatics and Computational Biology* 1, 542–545
Suthaharan S. (2016), Decision tree learning. In *Machine Learning Models and Algorithms for Big Data Classification*, Springer: Berlin, Germany, 237–269.
100. Segal M. R. (2004), *Machine Learning Benchmarks and Random Forest Regression*; Center for Bioinformatics & Molecular Biostatistics, San Francisco, USA.
101. Chen, T., Guestrin C. (2016), XGBoost: A Scalable Tree Boosting System, In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
102. Nair, A. S.; Anoop, A.; Ahuja, R.; Pathak, B. Relativistic Effects in Platinum Nanocluster Catalysis: A Statistical Ensemble-Based Analysis. *J. Phys. Chem. A* 2022, 126, 1345– 1359 (DOI: 10.1021/acs.jpca.1c09981)
103. Baxter, E. T., Ha, M.-A., Cass, A. C., Alexandrova, A. N., Anderson, S. L. Ethylene Dehydrogenation on Pt_{4,7,8} Clusters on Al₂O₃: Strong Cluster Size Dependence Linked to Preferred Catalyst Morphologies. *ACS Catal.* 2017, 7, 3322– 3335 (DOI: 10.1021/acscatal.7b00409)
104. Sharma, R. K., Minhas, H., Pathak, B. Investigating the Metastability-Triggered Reactivity of Pt_{7,8} Clusters on Graphene: Unraveling Statistical Ensemble Representation for ORR in Gas and Implicit Solvent Phases. *J. Phys. Chem. C* 2024, 128 (18), 7504– 7517 (DOI: 10.1021/acs.jpcc.4c00376)
105. Chaves, A. S., Piotrowski, M. J., Da Silva, J. L. F. Evolution of the Structural, Energetic, and Electronic Properties of the 3d, 4d, and 5d Transition-Metal Clusters (30 TM_n Systems for n = 2 - 15): a Density

- Functional Theory Investigation. *Phys. Chem. Chem. Phys.* 2017, 19, 15484– 15502 (DOI: 10.1039/c7cp02240a)
106. Chaves, A. S., Rondina, G. G., Piotrowski, M. J., Tereshchuk, P., Da Silva, J. L. F. The Role of Charge States in the Atomic Structure of Cu_n and Pt_n ($n = 2\text{--}14$ Atoms) Clusters: A DFT Investigation. *J. Phys. Chem. A* 2014, 118, 10813– 10821 (DOI: 10.1021/jp508220h)
 107. Zhai, H., Alexandrova, A. N. Ensemble-Average Representation of Pt Clusters in Conditions of Catalysis Accessed through GPU Accelerated Deep Neural Network Fitting Global Optimization. *J. Chem. Theory Comput.* 2016, 12 (12), 6213– 6226 (DOI: 10.1021/acs.jctc.6b00994)
 108. Bunău, O., Bartolomé, J., Bartolomé, F., Garcia, L.-M. Large Orbital Magnetic Moment in Pt_{13} Clusters. *J. Phys. Condens. Matter Inst. Phys. J.* 2014, 26 (19), 196006 (DOI: 10.1088/0953-8984/26/19/196006)
 109. Mahata, A., Rawat, K. S., Choudhuri, I., Pathak, B. Cuboctahedral vs. Octahedral Platinum Nanoclusters: Insights into the Shape-Dependent Catalytic Activity for Fuel Cell Applications. *Catal. Sci. Technol.* 2016, 6 (21), 7913– 7923 (DOI: 10.1039/C6CY01709F)
 110. Duan, Z., Wang, G. A First Principles Study of Oxygen Reduction Reaction on a Pt(111) Surface Modified by a Subsurface Transition Metal M ($M = \text{Ni}, \text{Co}, \text{or Fe}$). *Phys. Chem. Chem. Phys.* 2011, 13 (45), 20178– 20187 (DOI: 10.1039/c1cp21687b)
 111. Kulkarni, A., Siahrostami, S., Patel, A., Nørskov, J. K. Understanding Catalytic Activity Trends in the Oxygen Reduction Reaction. *Chem. Rev.* 2018, 118 (5), 2302– 2312 (DOI: 10.1021/acs.chemrev.7b00488)
 112. Nair, A. S., Pathak, B. Computational Screening for ORR Activity of 3d Transition Metal Based M@Pt Core-Shell Clusters. *J. Phys. Chem. C* 2019, 123, 3634– 3644 (DOI: 10.1021/acs.jpcc.8b11483)

113. Kumari, S., Masubuchi, T., White, H. S., Alexandrova, A., Anderson, S. L., Sautet, P. Electrocatalytic Hydrogen Evolution at Full Atomic Utilization over ITO-Supported Sub-nano-Pt_n Clusters: High, Size-Dependent Activity Controlled by Fluxional Pt Hydride Species. *J. Am. Chem. Soc.* 2023, 145 (10), 5834– 5845 (DOI: 10.1021/jacs.2c13063)
114. Taleblou, M., Camellone, M. F., Fabris, S., Piccinin, S. Oxidation of Gas-Phase and Supported Pt Nanoclusters: An Ab Initio Investigation. *J. Phys. Chem. C* 2022, 126, 10880– 10888 (DOI: 10.1021/acs.jpcc.2c02176)
115. Lym, J., Wittreich, G. R., Vlachos, D. G. A Python Multiscale Thermochemistry Toolbox (PMuTT) for Thermochemical and Kinetic Parameter Estimation. *Comput. Phys. Commun.* 2020, 247, 106864 (DOI: 10.1016/j.cpc.2019.106864)
116. Sharma, R. K., Minhas, H., Pathak, B. Investigating the Metastability-Triggered Reactivity of Pt_{7,8} Clusters on Graphene: Unraveling Statistical Ensemble Representation for ORR in Gas and Implicit Solvent Phases. *J. Phys. Chem. C* 2024, 128 (18), 7504– 7517 (DOI: 10.1021/acs.jpcc.4c00376)
117. Sedgwick, P. Pearson's Correlation Coefficient. *Bmj* 2012, 345, e4483 (DOI: 10.1136/bmj.e4483)
118. Lundberg, S. M., Lee, S. I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U. Von, Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; Vol. 30, pp 4768– 4777.
119. Mai, H., Le, T. C., Chen, D., Winkler, D. A., Caruso, R. A. Machine Learning for Electrocatalyst and Photocatalyst Design and Discovery. *Chem. Rev.* 2022, 122, 13478– 13515 (DOI: 10.1021/acs.chemrev.2c00061)

120. Fang, Z., Li, S., Zhang, Y., Wang, Y., Meng, K., Huang, C., Sun, S. The DFT and Machine Learning Method Accelerated the Discovery of DMSCs with High ORR and OER Catalytic Activities. *J. Phys. Chem. Lett.* 2024, 15, 281– 289 (DOI: 10.1021/acs.jpclett.3c02938)
 121. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011, 12, 2825– 283
-