Prediction of 2D van der Waals Heterostructures for Photocatalysis using Machine Learning and DFT

Arooba Kanwal

Reg. No: 0000294443

Department of Physics

Allama Iqbal Open University, Islamabad

(2024)

Prediction of 2D van der Waals Heterostructures for Photocatalysis using Machine Learning and DFT

Arooba Kanwal

Reg. No: 0000294443

Submitted in partial fulfillment of the requirements for the Master of Philosophy degree in discipline of Physics at Faculty of Science.

Supervisor

Dr. Abdul Jalil

Department of Physics

Allama Iqbal Open University

Department of Physics

Allama Iqbal Open University, Islamabad

(2024)

DECLARATION BY THE SUPERVISOR AT THE TIME OF FORWARDING THE RESEARCH THESIS TO THE CHAIRMAN OF THE DEPARTMENT FOR EXTERNAL EVALUATION

I Prof./Dr. Abdul Jalil Supervisor of an AlOU research student Mr./Ms. Arooba Kanwal do hereby solemnly declare that the thesis entitled Prediction of 2D van der Waals Heterostructures for Photocatalysis using Machine Learning and DFT being submitted as partial fulfillment of MS/M.Phil/MSc. (Hons) degree in the discipline of Physics has been completed under my guidance and supervision and is an original work of the student except where otherwise acknowledged in the text. It has not been submitted or published earlier for obtaining any degree from this or any other University or Institution.

The thesis is complete in all respects and I am fully satisfied with the quality of the student's research work. Now it is ready to be evaluated by external subject experts.

Date: _____

Signature _____

Name in full Abdul Jalil

Address Department of Physics

Allama Iqbal Open University

Phone _____

Email abdul.jaleel@aiou.edu.pk

CHECK LIST

Title of the Thesis Prediction of 2D van der Waals Heterostructures for Photocatalysis using Machine Learning and DFT

Name of the Student Arooba Kanwal

Name of Supervisor Abdul Jalil

Level/ Program M.Phil. Physics

	Yes	No	N/A
1. Student has observed format of the faculty.	✓		
2. Student has observed APA/MLA style	✓		
3. This thesis is an original piece of work.	✓		
4. Content given is relevant to the statement of the problem.	✓		
5. Scientific style of presentation has been adopted	✓		
6. Proper methodology has been adopted for research.	✓		
7. Population and samples are consistent throughout the thesis.			✓
8. Related researches have been included in relevant chapters.	✓		
9. Data is presented in tabular form of the faculty.	✓		
10. Discussion on tables has been mode adequately.	✓		
11. Findings are based on analysis.	✓		
12. Conclusions are based on findings.	✓		
13. Appropriate statistical methods have been applied in support of reliability & validity of the data collected.			✓
14. Recommendations are based on conclusions.	✓		
15. Annexures are placed properly.	✓		
16. All the authors given inside are included in the reference list.	✓		
17. Spellings and names of authors given in the reference list are same as are in the text.	✓		
18. The thesis is free of language errors.	✓		
19. Proper editing has been made.	✓		
20. Every part of the thesis is free of plagiarism.	✓		
21. This thesis adds to the existing stock of knowledge.	✓		

Dated: _____

Supervisor: Abdul Jalil

Signature: _____

DECLARATION BY THE STUDENT AT THE TIME OF SUBMISSION OF THESIS TO THE SUPERVISOR FOR EXTERNAL EVALUATION

(To be retained by the Controller of Examinations)

I Arooba Kanwal Son/Daughter of Muhammad Zahid Registration No. 0000294443 student of M.Phil. at Allama Iqbal Open University do hereby solemnly declare that the thesis entitled Prediction of 2D van der Waals Heterostructures for Photocatalysis using Machine Learning and DFT submitted by me in partial fulfillment of M.Phil. degree in discipline of Physics is my original work, except where otherwise acknowledged in the text, and has not been submitted or published earlier and shall not, in future, be submitted by me for obtaining any degree from this or any other university or institution.

Signature: _____

Name in Full: Arooba Kanwal

ALLAMA IQBAL OPEN UNIVERSITY ISLAMABAD

Department of Physics

(Acceptance by the Viva Voce Committee)

Title of thesis: Prediction of 2D van der Waals Heterostructures for Photocatalysis using Machine Learning and DFT

Name of Student: Arooba Kanwal

Accepted by the Faculty of Science Allama Iqbal Open University in partial fulfillment of the requirements for the Master of Philosophy Degree in the discipline of Physics.

_____

Supervisor

Viva Voce Committee

_____

External Examiner

_____

Chairperson/Director

_____

Dean Faculty of Sciences

(Day, Month, Year)

DEDICATION

To my beloved Mother,

whose memory continues to inspire and guide me

ACKNOWLEDGEMENT

"In the name of ALLAH, the most Merciful, the most Beneficent"

Firstly, thanks to **Almighty Allah** for His limitless blessings and the courage He granted to complete this task successfully.

Heartfelt appreciation goes to **Dr. Abdul Jalil**, my supervisor, for his consistent support and the ideas he imparted to me. Indeed, to him, I owe this thesis. Further, I want to express my sincere gratitude to the Head of the Physics Department, **Dr. Raza Ali Raza**, for the inspiration and constructive comments. He was indeed a great source of motivation for me throughout this journey. Finally, this work would not have been possible without the endless support of my senior, **Noor ul Ain**, who was more than a family to me.

Arooba Kanwal

ABSTRACT

The van der Waals (vdW) heterostructures, formed by vertical stacking of two distinct 2D materials, exhibit unique electronic properties. The weak vdW interactions tend to provide the benefit of efficient carrier separation, thus making them promising candidates for photocatalysis. However, analyzing all possible combinations of 2D materials is impractical through traditional approaches, necessitating the development of predictive models to automate and accelerate the quest. Herein, a hybrid approach using machine learning (ML) in conjugation with first-principles calculations is proposed to predict the properties of hexagonal vdW bilayers for application in photocatalysis. Our ML workflow comprises of following major steps: (1) constructing a vast material space of bilayers and their descriptors using a 2D material database, (2) labeling a diverse set of bilayers using Density Functional Theory (DFT) calculations, (3) training the supervised ML models on a labeled dataset for binding energy, interlayer distance, bandgap, work function and band edges of heterostructures, (4) evaluating the performance of models on the validation set, and (5) predicting the properties of the unlabeled dataset and screening the bilayers feasible for overall water-splitting photocatalysis. Various linear, boosting and kernel-based data-driven models were implemented. Ensemble techniques such as Random Forest and Gradient Boosting Trees outperformed other approaches in predicting the desired attributes of 2D vdW heterostructures. The computational framework presented here tends to establish the relationship between 2D monolayers and vdW bilayers. Our findings highlight the potential of this approach in accelerating the search for novel photocatalysts by efficiently and accurately predicting their properties, thereby contributing to the broader goal of sustainable energy production.

TABLE OF CONTENTS

DEDICATION vii

ACKNOWLEDGEMENT viii

ABSTRACT ix

LIST OF FIGURES xii

LIST OF TABLES xiv

CHAPTER 1 INTRODUCTION 1

1.1 Background and Significance 2

1.1.1 Evolution of Two-Dimensional Materials 2

1.1.2 van der Waals Heterostructures 2

1.1.3 The Role of Photocatalysis 4

1.1.4 Challenges in Material Discovery 4

1.1.5 Significance of Integrating Machine Learning in DFT 4

1.2 Machine Learning in Materials Science 5

1.2.1 Revolutionizing Material Discovery 5

1.2.2 Techniques & Applications 5

1.2.3 Integration of Computational Methods 6

1.2.4 Impact on Material Development 6

1.3 Context of the problem 6

1.4 Research Objectives 7

1.4.1 Primary Objective 7

1.4.2 Specific Objectives 7

CHAPTER 2 LITERATURE REVIEW 8

2.1 DFT Applications in Photocatalysis 9

2.2 Integration of ML and DFT 9

CHAPTER 3 METHODOLOGY 14

3.1 Research Design 15

3.2 Data Collection and Preprocessing 15

3.3 Feature Selection and Engineering 17
3.4 High-throughput Density Functional Theory (DFT) Calculations 17
3.5 Machine Learning Algorithms 18
3.5.1 Least absolute shrinkage and selection operation (LASSO) 18
3.5.2 Ridge 18
3.5.3 Support Vector Machines (SVM) 18
3.5.4 AdaBoost (AB) 18
3.5.5 Random Forest (RF) 19
3.5.6 Gradient Boosting Trees (GBT) 19
3.5.7 Elastic Net (EN) 19
3.5.8 Kernel Ridge Regression (KRR) 19
3.5.9 Stacked Ensemble Meta-learner (SEM) 19
3.6 Performance Evaluation Criteria 20
CHAPTER 4 RESULTS AND DISCUSSION 21
4.1 Machine Learning-Assisted Screening of Hexagonal 2D van der Waals Bilayers for Photocatalysis 22
4.1.1 Introduction 22
4.1.2 Methodology 26
4.1.3 Results and discussion 33
4.1.4 Prediction for Water-Splitting Photocatalysis using ML Models 41
CHAPTER 5 CONCLUSIONS 43
5.1 Conclusions 44
REFERENCES 46
Appendix A: Data Processing and Bilayer Features 58
Appendix B: Labeled vdW Bilayers 72
Appendix C: Predictions on Unlabeled vdW Bilayers 79

LIST OF FIGURES

Figure 1.1 Band edge alignment of (a) Type-I, (b) Type-II, and (c) Type-II heterostructures 3
Figure 3.1. The systematic workflow of DFT-ML hybrid approach. The steps involve (i) data collection using database and high-throughput first-principles calculations and processing, (ii) monolayer feature selection, (iii) bilayer feature engineering from the monolayer descriptors, (iv) labeling bilayer target properties by conducting DFT calculations, (v) model selection and training, and (vi) property prediction, i.e., filtering the vdW bilayers feasible for hydrogen and oxygen evolution reactions (HER and OER) from labeled and unlabeled datasets. 16
Figure 4.1. The systematic mechanism of interfacial charge transfer in (a) type-I and (b) type-II heterostructures. The vacuum (Fermi) levels of the monolayers are aligned before (after) making the junction. 23
Figure 4.2. The phenomenon of water-splitting photocatalysis for (a) type-I and (b) type-II heterostructures. 24
Figure 4.3. The selected monolayer prototypes from the C2DB database. 27
Figure 4.4. (a) The resulting bilayer prototypes by stacking two different monolayers vertically, (b) bilayer prototypes and percentage of three types of band alignments for each class, (c) labeling the target properties for bilayers using DFT calculations, and (d) the indication of interlayer distance in Å for bilayers. 28
Figure 4.5. Comparison of DFT and ML predicted (a) IE (b) EA (c) Φ (d) bandgap for validation set using SEM and (e) DFT and Anderson's bandgap. The evaluation metrics are provided for each target. 36
Figure 4.6. Comparison of DFT and ML predicted (a) E_{CBM} (b) E_{VBM} (c) E_b and (d) d_o for validation set using SEM. The evaluation metrics are provided for each target. ... 40
Figure 4.7. The overall water-splitting photocatalysts obtained from (a) the labeled DFT calculated set and (b) the unlabeled ML predicted dataset. Water reduction and oxidation levels are marked by black dotted lines. 42
Figure C.1. The LASSO coefficients for LASSO-selected set of descriptors for IE. . 84
Figure C.2. Correlation map for LASSO-selected set of descriptors for IE. 85
Figure C.3. The LASSO coefficients for LASSO-selected set of descriptors for EA. 86
Figure C.4. Correlation map for LASSO-selected set of descriptors for EA. 87

Figure C.5. The LASSO coefficients for LASSO-selected set of descriptors for E_g. . 88
Figure C.6. Correlation map for LASSO-selected set of descriptors for E_g. 89
Figure C.7. The LASSO coefficients for LASSO-selected set of descriptors for E_{CBM}. 90
Figure C.8. Correlation map for LASSO-selected set of descriptors for E_{CBM}. 91
Figure C.9. The LASSO coefficients for LASSO-selected set of descriptors for E_{VBM}. 92
Figure C.10. Correlation map for LASSO-selected set of descriptors for E_{VBM}. 93
Figure C.11. The LASSO coefficients for LASSO-selected set of descriptors for Φ. 94
Figure C.12. Correlation map for LASSO-selected set of descriptors for Φ. 95
Figure C.13. The LASSO coefficients for LASSO-selected set of descriptors for E_b. 96
Figure C.14. Correlation map for LASSO-selected set of descriptors for E_b. 97
Figure C.15. The LASSO coefficients for LASSO-selected set of descriptors for d_o. 98
Figure C.16. Correlation map for LASSO-selected set of descriptors for d_o. 99
Figure C.17. Comparison of DFT and ML predicted IE for validation set employing multiple models. The evaluation metrics are provided for each model. 100
Figure C.18. Comparison of DFT and ML predicted EA for validation set employing multiple models. The evaluation metrics are provided for each model. 101
Figure C.19. Comparison of DFT and ML predicted E_g for validation set employing multiple models. The evaluation metrics are provided for each model. 102
Figure C.20. Comparison of DFT and ML predicted Φ for validation set employing multiple models. The evaluation metrics are provided for each model. 103
Figure C.21. Comparison of DFT and ML predicted E_{CBM} for validation set employing multiple models. The evaluation metrics are provided for each model. 104
Figure C.22. Comparison of DFT and ML predicted E_{VBM} for validation set employing multiple models. The evaluation metrics are provided for each model. 105
Figure C.23. Comparison of DFT and ML predicted E_b for validation set employing multiple models. The evaluation metrics are provided for each model. 106
Figure C.24. Comparison of DFT and ML predicted d_o for validation set employing multiple models. The evaluation metrics are provided for each model. 107

LIST OF TABLES

Table 4.1. The number of monolayers belonging to six selected prototypes. 27
Table 4.2. MAE ± STD of IE, EA and E_g pipelines for training (validation) set determined by five-fold cross-validation. 35
Table 4.3. MAE ± STD of Φ, E_{CBM}, E_{VBM} pipelines for training (validation) set determined by five-fold cross-validation. 38
Table 4.4. R^2 score of data-driven ML models on unseen dataset indicating prediction accuracy of electronic properties. 39
Table 4.5. MAE ± STD for training (validation) set determined by five-fold cross validation and R^2 score of models on unseen dataset indicating predicting accuracy of data-driven ML models for interlayer distance and interlayer binding energy. 40
Table A.1. List of selected 197 monolayers from C2DB. Monolayer UID is same as provided by the database (version of 30-11-2022). 58
Table A.2. List of 97 selected bilayer descriptors for set of vdW bilayers. 67
Table B.1. Formula, mismatch (%), bilayer prototype, binding energy (E_b) in meVÅ^{-2}, interlayer distance (d_o) in Å, bandgap (E_g), ionization energy (IE), electron affinity (EA), work function (Φ), energies of conduction band minimum (E_{CBM}) and valence band maximum (E_{VBM}) in eV for 47 DFT-predicted direct bandgap vdW bilayer heterostructures. 72
Table B.2. Formula, mismatch (%), bilayer prototype, binding energy (BE) in meVÅ^{-2}, interlayer distance (d_o) in Å, bandgap (Eg), ionization energy (IE), electron affinity (EA), work function (Φ), energies of conduction band minimum (E_{CBM}) and valence band maximum (E_{VBM}) in eV for 63 DFT-predicted vdW bilayer photocatalysts feasible for overall water splitting photocatalysis along with their bandgap type and band alignment. 75
Table C.1. Formula, mismatch (%), bilayer prototype, binding energy (B.E) in meVÅ^{-2}, interlayer distance (d_o) in Å, bandgap (Eg), ionization energy (IE), electron affinity (EA), work function (Φ), energies of conduction band minimum (E_{CBM}) and valence band maximum (E_{VBM}) in eV for 93 ML-predicted vdW bilayer photocatalysts feasible for overall water splitting photocatalysis. 79

CHAPTER 1
INTRODUCTION

1.1 Background and Significance

1.1.1 Evolution of Two-Dimensional Materials

Two-dimensional (2D) materials have drawn a lot of interest owing to their distinct features that set them apart from their bulk counterparts. These materials have atomic-scale thicknesses and exceptional stability. In 2004, the remarkable isolation of graphene, a single sheet of carbon atoms organized in a honeycomb lattice, revealed a variety of outstanding attributes, including excellent electrical and thermal conductivity and remarkable mechanical strength. This discovery encouraged the investigation of other 2D materials beyond graphene (Peng & De, 2013; W. Zhu, Perebeinos, Freitag, & Avouris, 2009), such as boron nitride (BN) (Wickramaratne, Weston, & Van De Walle, 2018), transition metal dichalcogenides (TMDs) (B. Liu et al., 2014; Opoku, Govender, Van Sittert, & Govender, 2017; Tang & Jiang, 2015), gallium sulphide (GaS) and gallium selenide (GaSe) (Yagmurcukardes, Senger, Peeters, & Sahin, 2016), which exhibit noteworthy electronic, optical, and mechanical properties, making them ideal across multiple applications, including electronics, optoelectronics, and energy conversion.

1.1.2 van der Waals Heterostructures

2D van der Waals (vdW) heterostructures have entirely transformed the field of materials research by providing a flexible framework for creating sophisticated materials with unique features (Geim & Grigorieva, 2013; Pierucci et al., 2016). They are assembled by vertically stacking the different 2D materials without necessitating direct chemical bonds between them. Moreover, these heterostructures use weak vdW forces to integrate distinct features of 2D materials, resulting in unique functionality. Each layer retains its unique properties, allowing the combination of desired attributes from several materials. They have attracted paramount interest in photocatalysis because of their remarkable capacity to promote photocatalytic processes through effective charge separation, improved light absorption, and active surface sites (Bian et al., 2022; L. Song et al., 2022; X. Zhou et al., 2018).

1.1.2.1 Type-I (Straddling) Heterostructures

The conduction and valence band edges of one semiconductor layer (SC-I) align with the conduction band minimum (CBM) and valence band maximum (VBM) of the other layer (SC-II), respectively, in type-I heterostructures, as presented in **Figure 1.1 (a)**. Because of this alignment, electron-hole recombination is efficient, making type-I heterostructures excellent for optoelectronic devices such as lasers and light-emitting diodes (LEDs).

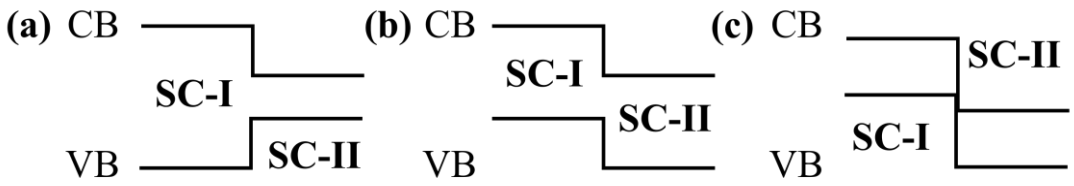

Figure 1.1 Band edge alignment of (a) Type-I, (b) Type-II, and (c) Type-II heterostructures

1.1.2.2 Type-II (Staggered) Heterostructures

Figure 1.1 (b) illustrates that type-II heterostructures feature mismatched CBMs and VBMs between two semiconductor layers. Electrons and holes are spatially isolated in various layers in these devices, which reduces recombination rates. This spatial separation can result in unique electrical and optical features. Type-II heterostructures are frequently employed in light detectors and solar cells.

1.1.2.3 Type-III (Broken gap) Heterostructures

In type-III heterostructure, the VBM of the first semiconductor (SC-I) is higher than that of the second (SC-II). In contrast, the CBM of the first is lower than that of the second semiconductor, as given in **Figure 1.1 (c)**. Despite being spatially confined, wavefunctions of electrons and holes slightly overlap, which results in intermediate radiative recombination rates. Because of this alignment, type-III heterostructures can have fascinating electrical and optical characteristics, making them flexible for several applications. In these heterostructures, device performance is enhanced by effective carrier control and recombination. Type-III heterostructures are employed in LEDs, tunnel diodes, and transistors.

1.1.3 The Role of Photocatalysis

Photocatalysis is a process that uses light to induce chemical processes, which has important implications for environmental remediation and sustainable technological developments. Ideal photocatalysts must absorb light efficiently, produce electron-hole pairs, and promote their separation and migration to reactive sites. The structural and electrical adaptability of 2D vdW heterostructures makes them ideal candidates for photocatalytic applications, where fine regulation of band alignment and charge dynamics are essential to optimize catalytic performance (Paquin, Rivnay, Salleo, Stingelin, & Silva, 2015).

The importance of exploring vdW bilayer photocatalysts stems from their potential to address major environmental and energy concerns. Researchers may develop photocatalysts with higher efficiency, selectivity, and stability for diverse light-driven processes, such as pollutant degradation, water splitting, and CO_2 reduction, by leveraging the synergistic effects of unique monolayers within a heterostructure. The control over the band alignment, electrical structure, and surface chemistry paves the way to develop eco-friendly energy sources and the degradation of environmental pollutants (Jalil, Zhao, Kanwal, & Ahmed, 2023).

1.1.4 Challenges in Material Discovery

The extensive combinatorial space of potential material combinations makes the identification and development of 2D vdW heterostructures for photocatalysis challenging. Traditional approaches are frequently time-consuming and resource-intensive, emphasizing the need for computational tools to speed up the discovery process and direct experimental efforts.

1.1.5 Significance of Integrating Machine Learning in DFT

Combining ML with DFT provides a synergistic approach that leverages the benefits of both methodologies. While machine learning may provide quick predictions, DFT calculations give a strong foundation in quantum mechanics and precise predictions of material properties (Wan, Zhang, Yu, & Guo, 2021). This hybrid technique tries to capitalize on the accuracy of DFT whilst benefiting from the productivity and versatility of ML (Wei et al., 2019). The DFT-ML hybrid approach includes training the data-

driven models on target properties obtained by standard computational methods, determining their performance on training and validation datasets and then using those trained models to predict the properties of unseen and unlabeled datasets (Chibani & Coudert, 2020). This approach can efficiently predict the properties of materials, minimizing the computational cost without compromising the accuracy of predictions.

1.2 Machine Learning in Materials Science

1.2.1 Revolutionizing Material Discovery

ML is transforming materials science by providing new tools for material discovery, identification, and optimization. Unlike conventional trial-and-error approaches, machine learning uses data and algorithms to determine patterns and generate predictions, which speeds up the design and discovery processes and lowers the computational cost.

1.2.2 Techniques & Applications

Material science typically employs the following machine learning techniques:

1.2.2.1 Supervised Learning

Supervised learning involves defined input and target variables. The ML models are trained on the labeled data (Chibani & Coudert, 2020). The models then learn to relate input and target features. This learning can interpret relationships driving predictions.

1.2.2.2 Unsupervised Learning

On the other hand, unsupervised learning involves defined input variables only. The ML models are trained on the unlabeled data. The models then seek patterns in the input data and make predictions. This type of learning is implemented for data exploration, feature engineering, detecting and treating outliers in data and dimensionality reduction.

1.2.2.3 Reinforcement Learning

Reinforcement learning is a feedback-based learning. It improves the decision-making processes of ML models through feedback, which is necessary in synthesis optimization.

1.2.2.4 Deep Learning

Deep learning is a subset of machine learning which employs neural networks to handle complicated datasets, allowing for the prediction of complex material characteristics.

1.2.3 Integration of Computational Methods

Combining machine learning with computational approaches such as DFT improves predictive capabilities. DFT generates high-quality data to educate machine learning models. Meanwhile, ML models can forecast the properties of materials based on DFT calculations, allowing for efficient screening. Further, ML algorithms may optimize material parameters to attain desired characteristics, directing experimental synthesis.

1.2.4 Impact on Material Development

The integration of ML in material science drives multiple advancements, such as:

1. Quickly identifying the promising materials for particular applications.
2. Eliminating the need for significant experimental testing.
3. Providing deep insights into structure-property relationships.

1.3 Context of the problem

The machine learning has the benefit of being computationally efficient. Traditional computational approaches, such as ab initio calculations, provide accuracy but may be costly and time-consuming, especially when working with big datasets or complicated material design. As a result, there is an urgent need for a data-driven methodology that may alleviate the discrepancy between computational efficiency and prediction accuracy.

1.4 Research Objectives

1.4.1 Primary Objective

The primary objective of this work is to establish and implement an integrated strategy that uses the DFT-ML hybrid approach to identify, optimize and propose 2D vdW heterostructures for photocatalysis, more effectively and precisely in contrast to traditional techniques.

1.4.2 Specific Objectives

1. Create a bilayer design space from a collection of distinct 2D materials. Generate and analyze high-quality datasets from DFT computations on diverse 2D vdW heterostructures.
2. Create machine learning models that predict essential aspects of 2D vdW bilayers, such as bandgap and band edge energies.
3. Employ labeled data to validate the prediction models and test their performance on new or unknown materials.
4. Use the derived models to predict properties of unlabeled datasets and screen the potential candidates from them for overall water-splitting photocatalysis.

CHAPTER 2
LITERATURE REVIEW

2.1 DFT Applications in Photocatalysis

Single-component photocatalysts consist of a single material. As a result, having a broad light absorption spectrum and a strong redox capacity is challenging for them since photo-excited electrons in the CB can return to the VB or be trapped in the defect state before reuniting with the holes, drastically lowering the solar energy deployment efficiency. Developing bilayer photocatalysts can be an ideal way to overcome the recombination problem (Luis, 2016). It has been discovered that 2D semiconductors with large specific surface areas and numerous surface-active sites may be attractive photocatalysts, offering improved electron mobility and low diffusion distances for effective charge separation (Guo, Zhou, Zhu, & Sun, 2016; Li, Dai, Li, Wei, & Huang, 2015; Lv et al., 2017). Furthermore, the graphene-like structure of two-dimensional materials with weak vdW interactions and surface free of dangling bonds facilitates the development of vdW heterostructures, which can be merged with not only 2D but also multidimensional materials. The formation of these composites provides unique features and boosts photocatalytic activity due to the synergy of their components (Low, Cao, Yu, & Wageh, 2014; Shifa, Wang, Liu, & He, 2019; B. Song & Jin, 2017).

Tan et al (Tan et al., 2024) reviewed several MXene-based nanocomposites for applications in photocatalysis and photoelectrochemical sensors. Min et al. (Min et al., 2021) reported $SeGa_2Te/SeIn_2Te$ vdW heterostructure, possessing type-II band alignment, for water-splitting photocatalysis. The material fulfilled the criteria of the water-splitting mechanism under different pH levels, i.e., pH = 0 and pH = 7. Ye et al. (Ye et al., 2023) proposed GaN/WSe_2 for visible-light-driven photocatalysis. Wang et al. (Y. Wang, Ding, Arif, Jiang, & Zeng, 2022) reviewed several MoS_2, WS_2, graphitic carbon nitride (CN), BN and MXenes based heterostructures for photocatalytic hydrogen production.

2.2 Integration of ML and DFT

ML can analyze massive datasets to identify patterns and estimate properties of novel materials. It has evolved as an effective technique in materials research in recent years, presenting an appealing solution to the property estimation problem (Choudhary,

Garrity, Pilania, & Tavazza, 2020; Willhelm et al., 2022). ML techniques have shown amazing success in various fields, including materials science, where these have been used to solve complicated tasks, optimize material discovery, and forecast material attributes with high accuracy (Chibani & Coudert, 2020; Pederson, Kalita, & Burke, 2022). These approaches have the potential to dramatically improve the capacity to predict new materials and estimate their properties by understanding subtle patterns and correlations within data that conventional techniques may not be able to perceive. On the other hand, DFT offers a quantum mechanical context for evaluating the electronic structure of materials. Integrating ML with DFT allows for rapid material screening and evaluation, paving the way for predicting and optimizing the properties of 2D vdW heterostructures for photocatalytic applications by establishing structure-property relationships. Researchers can accelerate the screening of promising photocatalysts by exploiting the predictive power of ML algorithms and the computational accuracy of DFT computations (Duan, Liu, Nandy, & Kulik, 2021; Wan et al., 2021).

These techniques (Mueller, Kusne, & Ramprasad, 2016) have widely been implemented with high-throughput density functional theory (DFT) calculation (M. Liu, Gopakumar, Hegde, He, & Wolverton, 2024) for accurate predictions of photocatalysts (Ge, Ke, & Li, 2023; Yan et al., 2022), thermoelectric materials (T. Wang, Zhang, Snoussi, & Zhang, 2020), electrocatalysts (Chen et al., 2024), gas sensors (Q. Zhou, Luo, Xue, & Liao, 2023), photoelectrochemical sensors (Tan et al., 2024), energy storage materials (Huang, Huang, & Dong, 2024; Qian, Sun, & Bao, 2022) and piezoelectric materials (Jing, Guan, Yang, & Zhu, 2023). Fiedler et al. (Fiedler et al., 2023) created an ML framework to predict the electronic structure on any length scale. It achieved remarkable efficiency on systems where DFT is feasible and, more crucially, permits estimations on scales where DFT computations are impossible. The study revealed how machine learning overcomes a long-standing computing constraint and propels materials science to advance. Back et al. (Back, Tran, & Ulissi, 2019) implemented ML and high-throughput DFT calculations to estimate Oxygen evolution reaction (OER) overpotential. Abraham et al. (Abraham, Sinha, Halder, & Singh, 2023) developed a robust and general-purpose multistep workflow that relies on DFT-ML to assess Hydrogen evolution reaction (HER) performance using DFT calculations. They implemented boosting algorithms to precisely and efficiently determine the Gibbs free

energy of hydrogen adsorption with a small mean absolute error (MAE) of 0.358 eV. Wan et al. (Wan et al., 2021) worked on the design and performance testing of a DFT-based ML method for catalysis program (DMCP) to implement the DFT-ML technique. They introduced DMCP as a user-friendly and productive program that may fulfil the demands for carrying out ML modelling based on data provided by DFT computations or from a materials repository. To demonstrate the overall workflow of the DFT-ML hybrid strategy and DMCP system, they tested double-atom electrocatalysts enabling carbon reduction reactions.

Jiang et al. (Jiang, Wang, Jia, Qu, & Qin, 2022) proposed a data-driven strategy to forecast the overpotential for (Ni-Fe-Co) O_x catalysts. The random forest regression model performed well with a small mean error. The variations in first ionization energies and d-orbital electrons revealed a linearly decreasing relationship with overpotential. They provided a direct pathway for the design of overpotential-oriented components for (Ni-Fe-Co) O_x catalysts. Gao et al. (M. Gao et al., 2023) employed a hybrid approach to project the bandgaps of 2180 underdeveloped yet sustainable quaternary semiconductors. The evaluation coefficient (R^2) implementing a random forest approach reached 0.93. The model was subsequently employed to choose four new quaternary direct bandgap semiconductors, i.e., $AgZn_2InS_4$, Ag_2InGaS_4, Ag_2ZnSnS_4 and $AgZn_2GaS_4$. Their electrical structures and optical characteristics were then validated and analyzed using DFT calculations. Their analysis has a definite reference value in the quest for luminous materials and devices and might significantly speed up the development of innovative optoelectronic semiconductors.

Choudhary et al. (Choudhary, Garrity, Pilania, et al., 2020) featured a computational database, machine learning models and web-apps to facilitate the design and discovery of two-dimensional (2D) heterostructures. They developed 226,779 potential heterostructures using density functional theory (DFT)-based parameters for non-metallic 2D-materials. They categorized those heterostructures into three types based on Anderson criterion. The most frequent heterostructure type was type II, while the least common was type III. The chemical patterns for each heterostructure type were examined in terms of the periodic arrangement of component elements. Willhelm et al. (Willhelm et al., 2022) deployed a computational framework that integrates electronic structure calculations, a 2D material database, and trained machine learning algorithms

to predict electronic and structural features of van der Waals (vdW) heterostructures. The data-driven strategy facilitated the fast screening and development of vdW heterostructures with desirable electrical and optical characteristics for specific device applications.

Gao et al. (Y. Gao, Zhang, Hu, & Yang, 2024) constructed a computational framework to screen two-dimensional polar materials for water-splitting photocatalysis. They performed high-throughput DFT calculations to determine the band-edge alignment. Further, they tested the thermodynamic stability of potential photocatalysts in an operational aqueous environment. The reaction-free energies validate the hydrogen and oxygen evolution reactions. Moreover, they studied excited state carrier dynamics to gain deeper insight into the charge transfer mechanism. Zhou et al. (P. Zhou et al., 2024) constructed a machine-learning framework to screen promising metal-oxide catalysts. They observed the best performance for the xgboost model with a coefficient of determination (R^2) of 0.908. Following this, they identified 10 materials showing potential for photocatalysis.

Huang et al. (Huang et al., 2024) reviewed data-driven studies on energy storage materials, followed by prediction and discovery. They argued that deep learning methods reduce the tediousness of the descriptor-construction process by automating the workflow using a crystal graph convolution neural network (CGCNN). This framework tends to increase the efficiency and accuracy of the screening process, minimizing errors. Sinha et al. (Sinha, Jyothirmai, Abraham, & Singh, 2024) reviewed prediction catalysts for hydrogen evolution reaction (HER) using a data-driven approach. They argued that artificial neural networks (ANN) and k-nearest neighbors (KNN) are compatible with large datasets, simplifying the training process by automatic feature selection. Unlike support vector machines (SVM), these models do not require careful tuning of feature sets.

Parse et al. (Parse & Pinitsoontorn, 2023) employed machine learning to predict promising thermoelectric materials. To improve the performance of models, they incorporated experimental data and features. They achieved an RMSE of 0.156 and an R^2 of 0.859, using a regression model. They further constructed a user-friendly web application to predict ZT values of thermoelectric materials with selective doping. Wang et al. (T. Wang et al., 2020) reviewed several machine learning approaches being

employed for the hunt of thermoelectric materials. They discussed that the accuracy portrayed by the neural network (NN) may arise from its complexity. They further argued that careful descriptor selection is a crucial step for thermoelectric materials to capture the diversity in the dataset.

Liu et al. (M. Liu et al., 2024) reported that HSE06 functional show small mean absolute error for bandgap and formation energies (0.687 eV and 0.147 eV/atom) in contrast to PBE functional (1.184 eV and 0.175 eV/atom). By enabling fast screening and recognition of heterostructures with specific properties, our method has the potential to accelerate progress in a range of disciplines, including electronics, photonics, energy generation and storage, and others (Abraham et al., 2023; Chandrasekaran et al., 2019; Fiedler et al., 2023; Pederson, Kalita, & Burke, 2022; Schleder, Padilha, Acosta, Costa, & Fazzio, 2019; Sun, Cao, Li, Li, & Ao, 2023; Wu, Guo, & Li, 2021). It might also be used as a foundation model for dealing with difficult material design challenges employing a data-driven approach. The choice of machine learning as a crucial tool in this research is motivated by its known efficiency in materials science, computational benefits, and compliance with the ultimate objective of enhancing heterostructure prediction (Mueller et al., 2016).

CHAPTER 3
METHODOLOGY

3.1 Research Design

This study leverages the benefits of data-driven and quantum-mechanical approaches to construct a computational framework to accurately estimate and optimize the properties of 2D vdW bilayer heterostructures from the constituent monolayers. We further employed this framework for efficient screening of water-splitting photocatalysts among them. The detailed systematic workflow for this particular study is discussed in **Section 4.1.2**.

The workflow comprises several steps, as shown in **Figure 3.1**. Firstly, data is collected from a monolayer database, cleaned and pre-processed. The monolayer crystal structures are then obtained to construct vast bilayer material space. Bilayer features are engineered by aggregating the monolayer features. The bilayer target properties are labeled employing high-throughput first-principles calculations. Data-driven ML models are subsequently trained on target properties, and their performance is evaluated. The model exhibiting the best performance is used further to predict the properties of unlabeled dataset.

3.2 Data Collection and Preprocessing

Firstly, relevant data is extracted from a computational material database such as JARVIS-DFT (Choudhary, Garrity, Reid, et al., 2020), Computational 2D Materials Database (C2DB) (Haastrup et al., 2021), Materials Project (Jain et al., 2013), Novel Materials Discovery (NOMAD) Repository (Sbailò, Fekete, & Ghiringhelli, 2022), Materials Cloud (Talirz et al., 2020), Inorganic Crystal Structure Database (ICSD) (Belsky & Lynn, 2002), Virtual 2D Materials Database (V2DB) (Sorkun, Astruc, Koelman, & Er, 2020), Open Quantum Materials Database (OQMD) (Kirklin et al., 2015), etc. Next, the data is filtered, cleaned and preprocessed for missing values. The missing data values are treated by a suitable method, such as elimination or aggregation. The outliers in the training data are identified and eliminated.

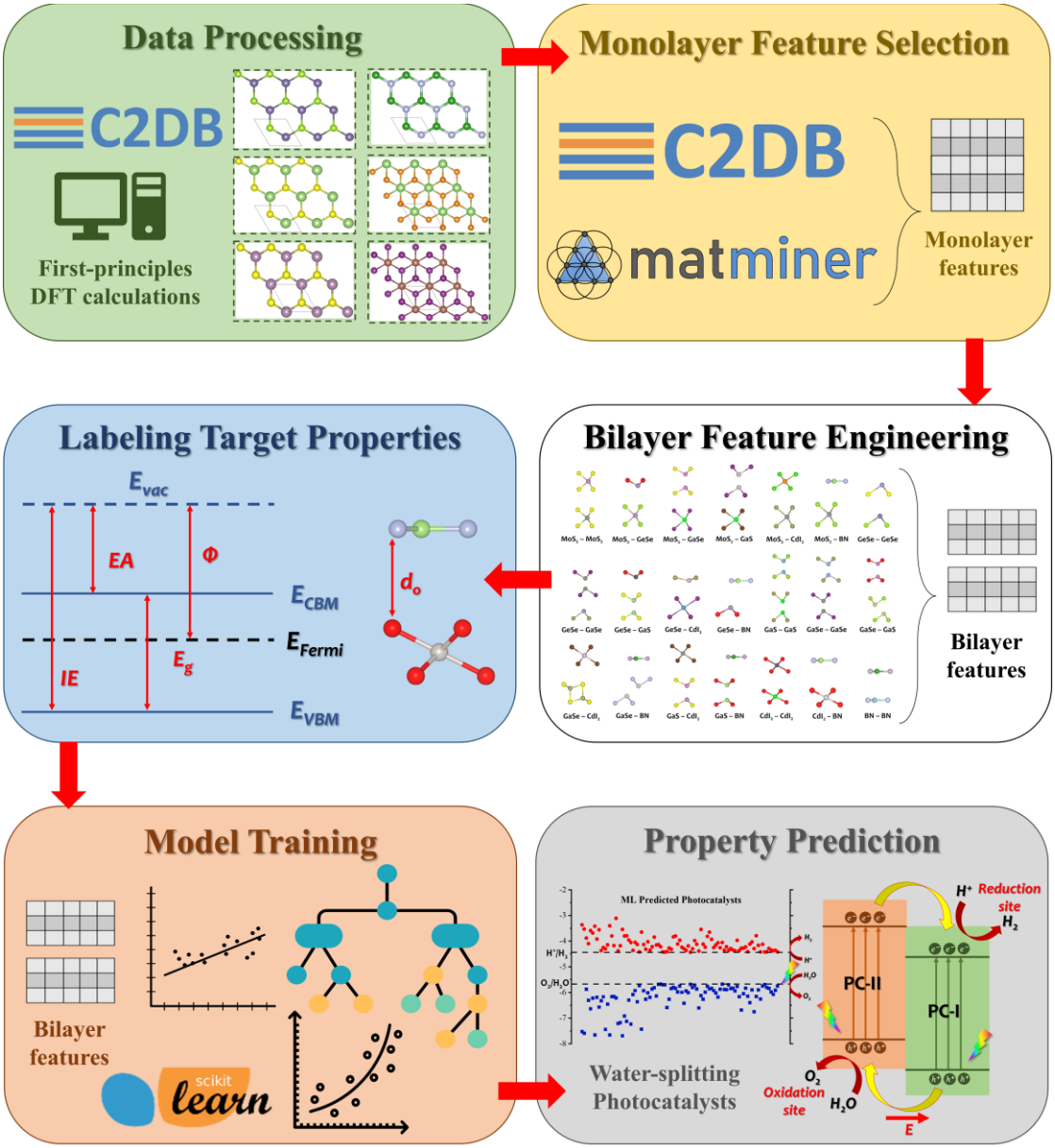

Figure 3.1. The systematic workflow of DFT-ML hybrid approach. The steps involve (i) data collection using database and high-throughput first-principles calculations and processing, (ii) monolayer feature selection, (iii) bilayer feature engineering from the monolayer descriptors, (iv) labeling bilayer target properties by conducting DFT calculations, (v) model selection and training, and (vi) property prediction, i.e., filtering the vdW bilayers feasible for hydrogen and oxygen evolution reactions (HER and OER) from labeled and unlabeled datasets.

Herein, we constructed vdW heterostructures by extracting 2D materials from C2DB and stacking them using the Atomic Simulation Environment (ASE) library. The generated bilayer data is then divided into two subsets: labeled and unlabeled datasets. The labeled dataset is the one for which target properties are calculated by DFT calculation. On the contrary, the unlabeled dataset is the one which has unknown target properties. For unlabeled datasets, target properties are determined by applying the best data-driven model.

3.3 Feature Selection and Engineering

Feature extraction and engineering rely on the target properties. Some features are extracted from the database; others are constructed using featurizers provided by matminer (Ward et al., 2018), while some can be engineered if they are significant for a better understanding of target properties. Feature construction and engineering is carried out for both labeled and unlabeled datasets. The features selected for this work are discussed in detail in **Section 4.1.2.4**. The labeled dataset is subdivided into training and testing datasets. The ML models are subsequently trained on the training dataset and their performance (i.e., accuracy of their predictions) is evaluated on the validation dataset.

3.4 High-throughput Density Functional Theory (DFT) Calculations

This study relies on Density Functional Theory (DFT) (Sholl & Steckel, 2009), which provides a quantum mechanical approach for describing the electronic structure and vast range of properties. DFT computations are employed to determine parameters such as binding energy, electronic bandgap, and band edges, which are essential for understanding the photocatalytic activity of vdW bilayers. The DFT workflow involves atomic structure relaxation to minimize total energy and electronic structure calculations to determine the target properties. High-throughput DFT calculations are employed to label the dataset by calculating the target features. The dataset is subsequently filtered based on target properties.

3.5 Machine Learning Algorithms

After determining the target properties from DFT calculations, suitable ML algorithms are trained for those properties on the labeled dataset, using scikit-learn python library (Pedregosa et al., 2011). The supervised ML models are categorized into classification and regression models. The classification models predict discrete values or labels of the classes. The regression models construct patterns between the dependent variable (target property) and the independent variables (input parameters) to predict continuous values (Cherkassky & Ma, 2003). Herein, we implemented multiple ML regression algorithms for predicting selected target properties, as discussed below.

3.5.1 Least absolute shrinkage and selection operation (LASSO)

LASSO uses L1 regularization on linear regression to penalize the absolute values of the coefficients. It yields sparse solutions by typically choosing a subset of features by setting specific coefficients to zero, thus providing feature selection and regularization (Kim, Koh, Lustig, Boyd, & Gorinevsky, 2007).

3.5.2 Ridge

Ridge Regression uses L2 regularization in linear regression by penalizing the squared magnitude of the coefficients. It avoids overfitting by penalizing significant coefficients and enhancing model generalization while maintaining all descriptors (McCullagh & Nelder, 2017).

3.5.3 Support Vector Machines (SVM)

The support vector machine (Platt, 1999) algorithm uses kernel functions to create hyperplanes in high-dimensional space to perform tasks such as classification and regression. It excels in high-dimensional domains, particularly if the number of dimensions exceeds the number of samples. This approach is commonly utilized in materials science research.

3.5.4 AdaBoost (AB)

AdaBoost (J. Zhu, Zou, Rosset, & Hastie, 2009), also known as Adaptive Boosting, combines several weak learners into a strong learner by repeatedly modifying the

weights of training instances. It focuses on misclassified instances and constructs a final model based on the weighted votes of the weak learners.

3.5.5 Random Forest (RF)

Random Forest (Louppe, 2014) builds a community of decision trees by training each tree on a randomly selected subset of the data and features. It enhances predictive accuracy and robustness by aggregating individual tree estimations, preventing overfitting.

3.5.6 Gradient Boosting Trees (GBT)

Gradient Boosting Trees construct an ensemble of decision trees systematically (Friedman, 2001). Each tree corrects the mistakes of its predecessor by fitting the residuals of the prior tree and optimizing the loss function via gradient descent, resulting in a robust predictive model.

3.5.7 Elastic Net (EN)

Elastic Net uses L1 and L2 regularization constraints in linear regression. It combines the feature selection traits of LASSO with the regularization power of Ridge Regression, resulting in more effective handling of correlated features and complicated datasets.

3.5.8 Kernel Ridge Regression (KRR)

Kernel Ridge Regression (Kevin P. Murphy, 2012) builds on Ridge Regression by introducing kernel functions. It allows for nonlinear interactions between features and target variables, allowing the method to capture complicated patterns in the data while still using L2 regularization.

3.5.9 Stacked Ensemble Meta-learner (SEM)

A Stacked Ensemble Meta-learner (Wolpert, 1992) integrates predictions from numerous base models to improve overall performance. It uses a meta-model to determine the optimum way to combine the predictions of the base models, thus boosting accuracy and robustness by exploiting their complementary capabilities.

3.6 Performance Evaluation Criteria

The performance of ML models is determined using suitable evaluation metrics. The metrics include mean absolute error (MAE), coefficient of determination (R^2), root mean square error (RMSE) and mean squared error (MSE). Generally, cross-validation (CV) or repeated k-Fold CV (Kohavi & Edu, 1993) is employed to determine the evaluation metrics. Compared to random train-test splits, k-Fold cross-validation offers more accurate and robust estimations of model performance. The process involves random splitting of the dataset into k groups, with each group serving as the test dataset and the remaining ones acting as the training dataset. The model is fitted to the training set and assessed on the test set. The cross-validation performance is calculated by aggregating the performance of all trained models.

CHAPTER 4
RESULTS AND DISCUSSION

4.1 Machine Learning-Assisted Screening of Hexagonal 2D van der Waals Bilayers for Photocatalysis

4.1.1 Introduction

Since the discovery of semiconducting TiO_2 as a photocatalyst in 1972 (Fujishima & Honda, 1972), water-splitting photocatalysis has emerged as an effective technique for environmental remediation and sustainable energy production (Jalil et al., 2023). The design of van der Waals (vdW) bilayer heterostructures may offer surface-active platforms for photocatalytic processes. The close interfacial contact of two-component monolayers can increase the transmission of charge and dissociation of photogenerated electron-hole pairs (Choudhary, Garrity, Pilania, et al., 2020; Q. Zheng et al., 2018). Two-dimensional (2D) vdW heterostructures are compound materials formed by vertically stacking two or more distinct 2D materials, where the layers are held by weak vdW interactions (Geim & Grigorieva, 2013). These heterostructures are recognized for their unique chemical and physical characteristics that emerge from the combinations of different materials, offering use in electronics, optoelectronics, catalysis, and other fields (Jin et al., 2018). The bilayer heterostructures are classified into type-I, type-II and type-III, depending on the alignment of band edges of constituent monolayers (Lee, Hwang, & Chung, 2015). The electron-hole recombination is efficient in type-I band alignment. Type-II heterostructures, on the other hand, offer spatially separated electrons and holes, which reduces their recombination rate. Such bilayers are applicable in light detectors and solar cells. In type-III band alignment, the wavefunctions of electrons and holes slightly overlap, resulting in intermediate radiative recombination rates.

Before making the junction, the difference in work functions of the two photocatalysts develops an electric field at the interface. The monolayer with small Φ develops a positive charge by losing electrons more rapidly. While the monolayer having large Φ develops a negative charge. This electric field is directed from the layer with a lower work function to the one with a higher work function, promoting band-edge bending and facilitating interfacial charge transfer. **Figure 4.1** depicts the complete charge transfer mechanism for type-I and type-II heterostructures. The more the difference in work functions, the greater the interfacial electric field and the more efficient the charge

transfer at the interface. The interfacial charge transfer is efficient in inferior band edges of type-II bilayers, while superior band edges provide spatial charge separation. In contrast, charge transfer at the interface in type-I bilayers facilitates carrier recombination.

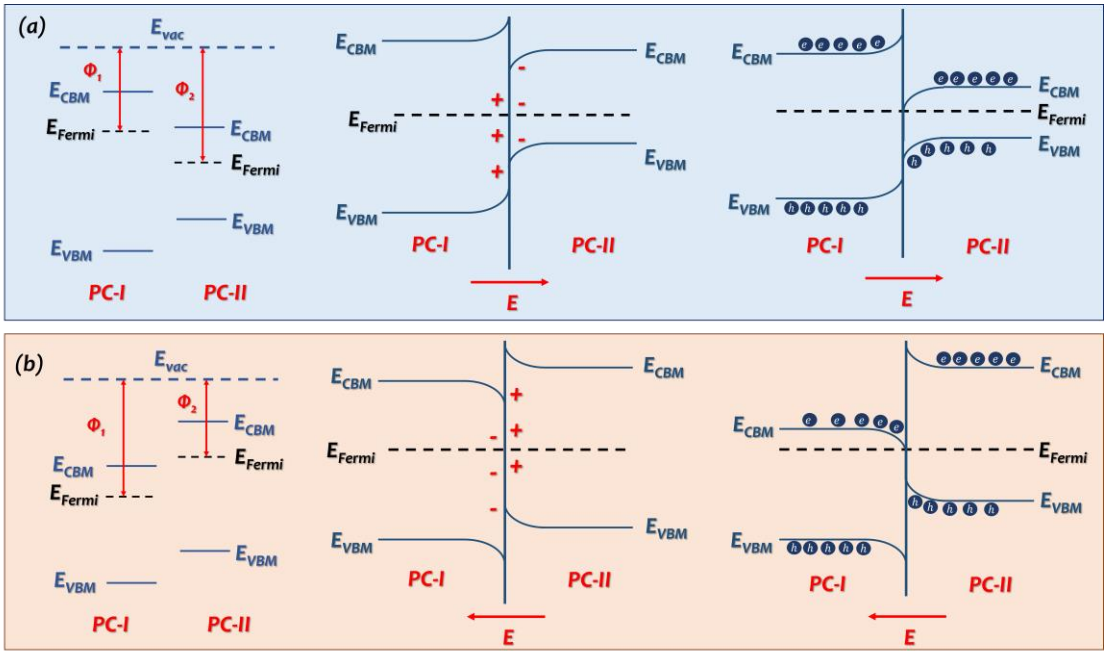

Figure 4.1. The systematic mechanism of interfacial charge transfer in (a) type-I and (b) type-II heterostructures. The vacuum (Fermi) levels of the monolayers are aligned before (after) making the junction.

In the context of type-II heterostructures, the photoexcited electrons travel from the higher conduction band (CB) of photocatalyst-II (PC-II) to the lower CB of photocatalyst-I (PC-I). The holes, in contrast, migrate from the lower valence band (VB) of PC-I to the higher VB of PC-II, increasing charge carrier separation and promoting photocatalytic oxidation and reduction reactions. PC-I becomes electron-rich while PC-II becomes electron-deficient. CB of PC-I acts as a reduction site facilitating hydrogen evolution reaction (HER), whereas VB of PC-II acts as an oxidation site facilitating oxygen evolution reaction (OER). Therefore, PC-I functions as a hydrogen-producing photocatalyst (HPP), while PC-II acts as an oxygen-producing photocatalyst (OPP). The corresponding reduction and oxidation reactions are as follows.

$$2H^+ + 2e^- \rightarrow H_2 \qquad \textbf{4.1}$$

$$4OH^- + 4h^+ \rightarrow O_2 + 2H_2O \qquad \textbf{4.2}$$

The light irradiation on type-I heterostructures also results in the excitation of electrons in the CB of both photocatalyst monolayers. However, due to the aligned band edges, these excited electrons and holes tend to recombine rapidly, causing minimum charge separation. PC-II acts as HPP and OPP since both reduction and oxidation sites lie in PC-II. Consequently, type-I alignment is less effective for photocatalysis. **Figure 4.2** illustrates the mechanism of photocatalysis for type-I and type-II bilayers.

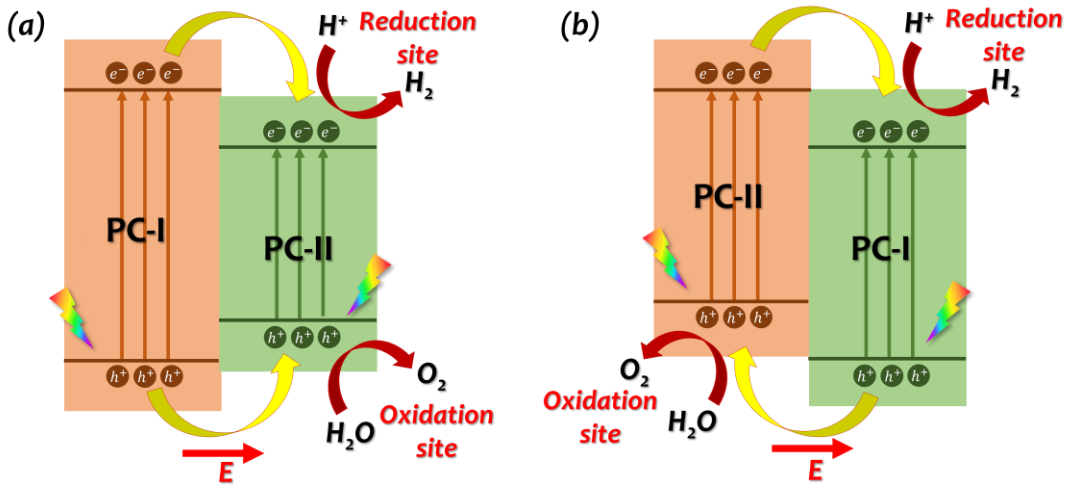

Figure 4.2. The phenomenon of water-splitting photocatalysis for (a) type-I and (b) type-II heterostructures.

Gao et al. (M. Gao et al., 2023) employed the DFT-ML approach to project the bandgaps of 2180 underdeveloped yet sustainable quaternary semiconductors. The evaluation coefficient (R^2) of the model employing a random forest approach reached 0.93. Dong et al. (Dong, Jacob, Bourdais, & Sanvito, 2021) employed a band structure folding approach on the 2H and 1T polymorphs of monolayers to determine the electronic structure of vdW hetero-bilayers. From the MXenes class, Zheng et al. (J. Zheng et al., 2020) screened 299 MXene hydrogen evolution reaction (HER) catalysts employing the DFT-ML approach. Mishra et al. (Mishra et al., 2019) devised a machine-learning model to locate band edges and accelerate band identification of MXene materials. Tawfik et al. (Tawfik et al., 2019) predicted the interlayer distance and electronic bandgap of 1500 bilayer vdW heterostructures based on 267 DFT-labeled bilayers using

a hybrid approach. They achieved an R^2 score of 0.83, employing the support vector machines algorithm.

Abraham et al. (Abraham, Sinha, Halder, & Singh, 2023) employed multiple ML algorithms to filter 2D MXene-based HER catalysts. Their results revealed that boosting models show better performance than the rest. They reported best testing MAE and R^2 of 0.424 (0.399) eV and 0.771 (0.788) for Gibbs energy with (without) 10-fold cross-validation, employing the gradient boosting trees. Choudhary et al. (Choudhary, Garrity, Pilania, et al., 2020) featured a computational database, machine learning models and web apps to facilitate the design and discovery of 2D heterostructures. They developed a large set of 226,779 potential heterostructures using DFT-based parameters for non-metallic 2D materials. They categorized those heterostructures into three types based on Anderson's criterion. The most frequent heterostructure type was type II, while the least common was type III. The chemical patterns for each heterostructure type were examined in terms of the periodic arrangement of component elements.

ML-based data-driven models are attracting attention since they are valuable and advantageous in predicting properties of materials that are difficult to measure through traditional methods because they are either costly or time-consuming. Herein, we develop a computational framework integrating a comprehensive 2D material database, electronic structure calculations, and supervised ML techniques. This framework aims to construct data-driven models capable of accurately and efficiently predicting properties of vdW bilayers from their constituent monolayers. The bilayer descriptors were constructed by aggregating the monolayer descriptors, and the ML models were subsequently trained for selected target properties. For the validation dataset, the ML-predicted properties are verified against the DFT-calculated properties. Flaws and inconsistencies, such as bias, overfitting, and under-fitting, are examined to assess the predictive potential and accuracy of each model. To address these flaws, we pre-processed the data by eliminating biased data. We further used regularization, hyper-parameter optimization, cross-validation and ensemble techniques. Afterwards, the trained data-driven models are employed to screen the potential bilayer photocatalysts from an unseen and unlabeled dataset. This approach proposes a large set of novel vdW bilayer photocatalysts efficiently.

4.1.2 Methodology

4.1.2.1 Hybrid Workflow

At first, the monolayer data and ground-state crystal structures are fetched from a 2D material database. These crystal structures are later stacked vertically, resulting in an extensive material space of unique vdW bilayer heterostructures. The bilayer material space is sampled, and a diverse set of bilayers is chosen for training the machine learning models. The selected bilayers are optimized using high-throughput DFT calculations (Bonacci et al., 2023), and their electronic properties are subsequently calculated. We determined the following eight target properties from our DFT computations: interlayer distance (d_o), binding energy (E_b), ionization energy (IE), electron affinity (EA), bandgap (E_g), work function (Φ), band edges, i.e., conduction band minimum (E_{CBM}) and valence band maximum (E_{VBM}) with respect to vacuum level. The monolayer data and descriptors are defined for constituent monolayers using pymatgen (Ong et al., 2013) and matminer (Ward et al., 2018) and then aggregated to construct a set of unique bilayer descriptors. The descriptors include elemental features, lattice parameters, band edges, van der Waals radii, atomic compositions, and electronic features. The labeled dataset is divided into training and validation sets. Next, the supervised ML models are trained for the mentioned eight target properties. The ML-predicted properties are authenticated against the DFT-calculated properties on the validation set to evaluate the performance of ML models. The trained data-driven models are then used to predict properties of unseen and unlabeled vdW bilayer heterostructures.

4.1.2.2 Monolayer Data Mining and Bilayer Configuration Space

The ground-state crystal structures for 2D materials are extracted from the Computational 2D Materials Database (C2DB) (Gjerding et al., 2021; Haastrup et al., 2021, 2018), which is an extensive and open-source database consisting of over 15,000 materials (as of 30-11-2022 version), their crystal structures and high-throughput DFT-calculated properties. The monolayers are selected based on six 2D prototypes: boron nitride (BN), molybdenum disulfide (MoS_2), gallium sulfide (GaS), gallium selenide

(GaSe), germanium selenide (GeSe) and cadmium iodide (CdI_2), as illustrated in **Figure 4.3**.

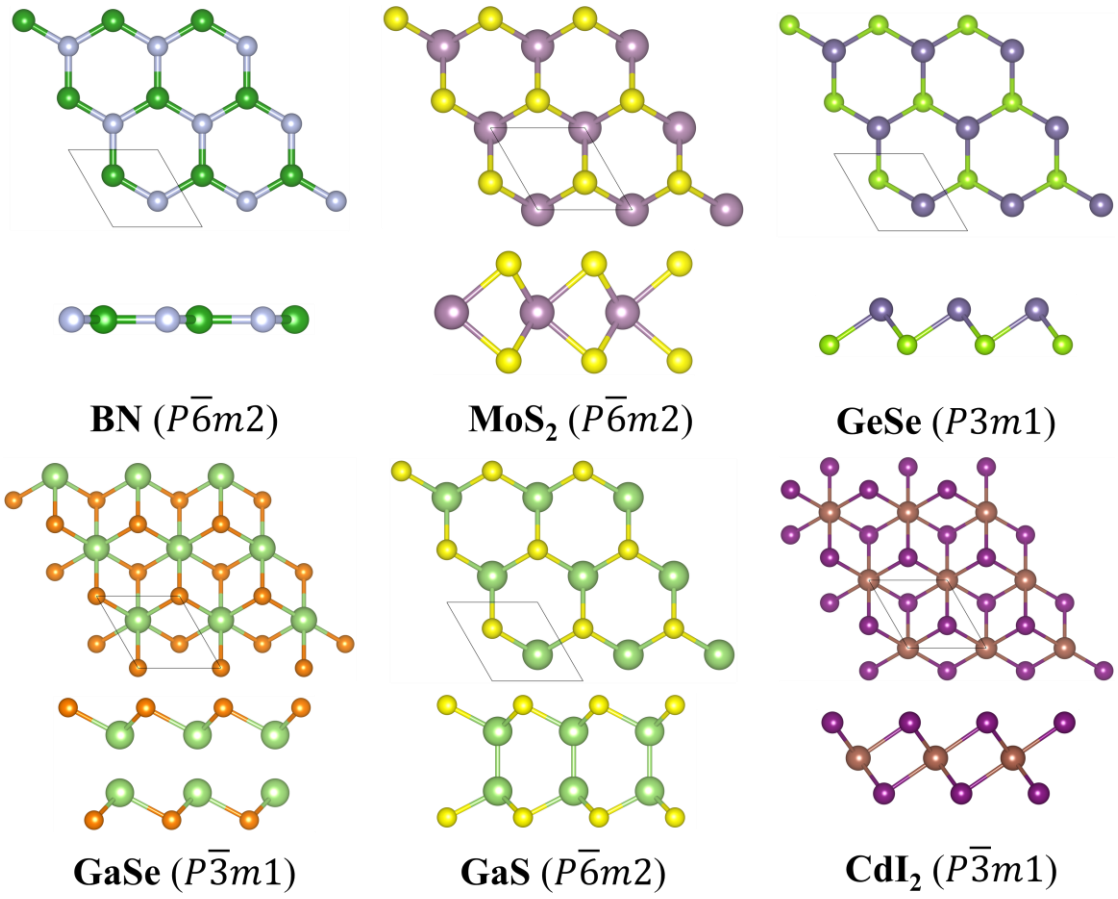

Figure 4.3. The selected monolayer prototypes from the C2DB database.

Table 4.1. The number of monolayers belonging to six selected prototypes.

S. No.	2D Prototype	General formula	Space group	No. of monolayers
1.	GaSe	A_2B_2	P-3m1	58
2.	CdI_2	AB_2	P-3m1	57
3.	MoS_2	AB_2	P-6m2	44
4.	GaS	A_2B_2	P-6m2	19
5.	GeSe	AB	P3m1	14
6.	BN	AB	P-6m2	5

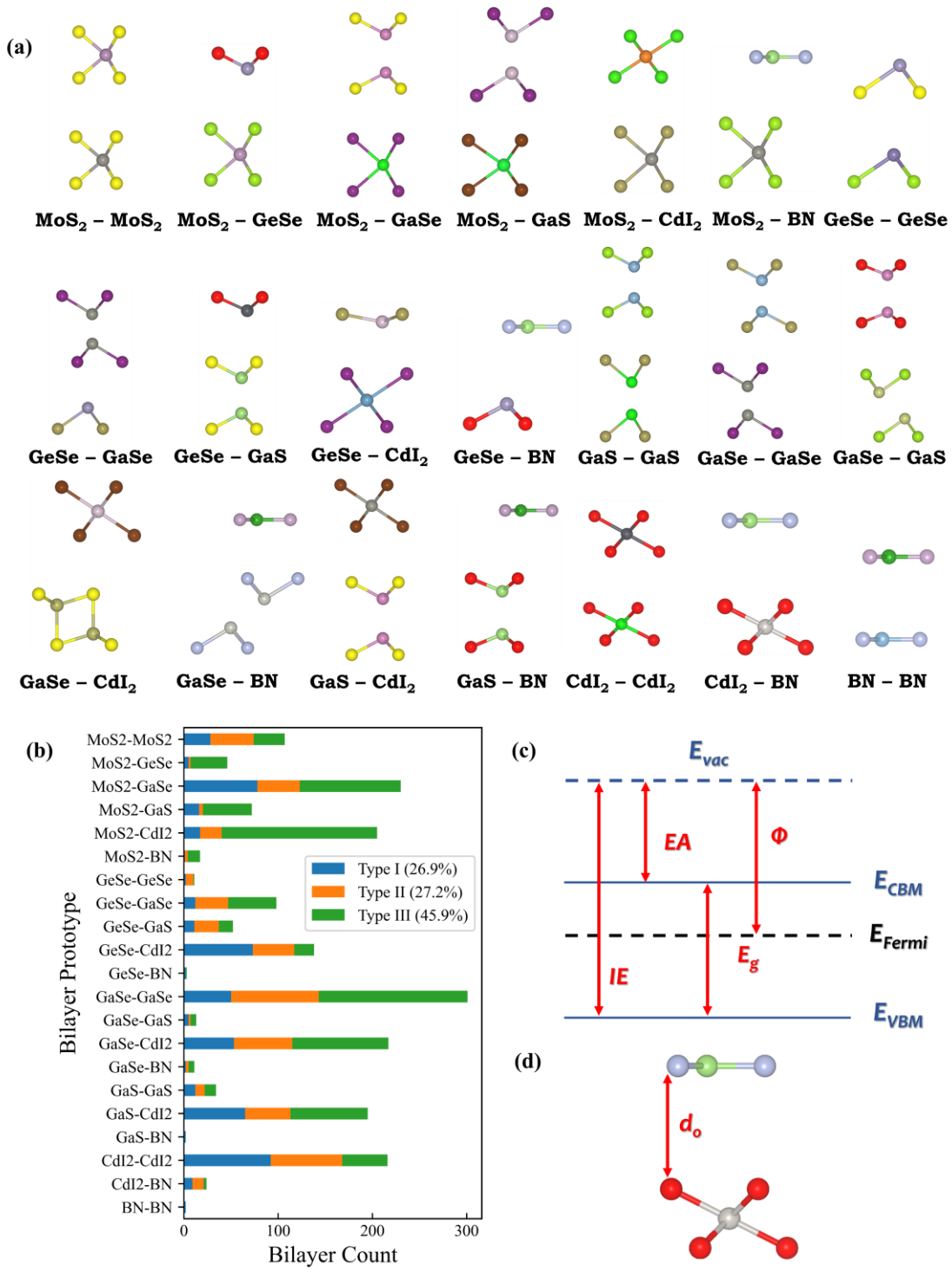

Figure 4.4. (a) The resulting bilayer prototypes by stacking two different monolayers vertically, (b) bilayer prototypes and percentage of three types of band alignments for each class, (c) labeling the target properties for bilayers using DFT calculations, and (d) the indication of interlayer distance in Å for bilayers.

To minimize the computational cost, we refine this subset by including only the monolayers possessing hexagonal lattice symmetry and are non-metallic and non-magnetic. Further, we only select the monolayers which exhibit high thermodynamic and dynamic stability, as determined by the C2DB database. As a result, 197 unique monolayers are obtained, which are used to construct vdW bilayers. The selected monolayers are binary compounds with a maximum of four atoms per unit cell and have general formulae of the form AB, AB_2, and A_2B_2, where A is either a metal or a semiconductor species. The count for each monolayer prototype is listed in **Table 4.1**. The lattice constant and bandgap lie in the range of 2.51 – 5.52 Å and 0.01 – 5.94 eV, respectively, as summarized in **Table A.1 (Appendix A)**.

The crystallographic information of selected monolayers is utilized to construct a diverse set of bilayers from 19306 possible combinations by stacking two monolayers vertically, employing the Atomic Simulation Environment (ASE) library (Larsen, Mortsen, & Blomqvist, 2019). Initially, an interlayer distance of 4 Å is specified. The pairs of monolayers are selected such that the lattice mismatch is 4% or less, and the pairs with zero mismatch are eliminated, resulting in 1994 bilayer heterostructures. The lattice mismatch (Δa) is determined using the relation given in **Equation 4.3**, where a_1 and a_2 are the lattice constants of constituent monolayers and a_1 is the smaller one among them (Ye et al., 2023).

$$\Delta a = \frac{|a_1 - a_2|}{a_1} \qquad \textbf{4.3}$$

A vacuum space of about 38 Å is kept along the z-axis to avoid interaction between adjacent layers, thus ensuring a space of at least 15 Å on each side. The combinations of six monolayer prototypes result in 21 bilayer prototypes, illustrated in **Figure 4.4 (a)**.

Since most of the monolayers belong to GaSe, CdI_2 and MoS_2 classes, bilayers from these classes dominate the configuration space. Classifying vdW bilayers in type-I, type-II and type-III based on the band offsets of their constituent monolayers, we determined the fraction of each type of band alignment from each bilayer class as represented in **Figure 4.4 (b)**. The two monolayers in a bilayer heterostructure must have their vacuum levels aligned at the same energy level to comply with Anderson's rule. We accomplish this by aligning the two vacuum levels at 0 eV and relocating band

edges of monolayers, E_{VBM} and E_{CBM}, relative to the vacuum level. Once the vacuum levels of monolayers are aligned, the difference between the higher VBM and the lower CBM can lead to a rough estimation of the bilayer bandgap as determined by Anderson's rule **(Equation 4.4)**.

$$\text{Anderson's gap} = \text{Lower } E_{CBM} - \text{Higher } E_{VBM} \quad \textbf{4.4}$$

This technique works best when there is a significant vacuum space between the constituents. Besides these estimations, our approach goes beyond this crude bandgap estimation since the ML models rely on several other descriptors besides these band edge features.

4.1.2.3 Labeling bilayers using High-Throughput DFT Workflow

We sampled 412 vdW bilayer heterostructures (~20% of bilayer space) such that the sampled data is a representative of entire bilayer design space. We labeled this training set by conducting first-principles DFT calculations (Sholl & Steckel, 2009) using Vienna ab initio simulation package (VASP) (Hafner & Kresse, 1997). The electron-ion interactions were considered using Projector augmented wave (PAW) method (Aewfc, 2022; Mortensen, Hansen, & Jacobsen, 2005). The electron exchange and correlation were estimated using Perdew Burke Ernzerhof (PBE) functional within generalized gradient approximation (GGA) (Perdew, Burke, & Ernzerhof, 1996). The weak vdW interactions were incorporated using Grimme's DFT-D3 approach (Grimme, Antony, Ehrlich, & Krieg, 2010). A plane-wave cut-off energy of 500 eV was considered. Monkhorst-Pack (Pack & Monkhorst, 1976) k-mesh of 9×9×1 was used for Brillouin zone sampling. Furthermore, in vdW heterostructures, dispersion effects can be observed at large distances. For this reason, a vacuum of 38 Å was maintained in the z-direction to avoid contact with nearby replicas. For structural relaxation, the convergence threshold for atomic force and electronic energy was set as 0.02 $eVÅ^{-1}$ and 1×10^{-6} eV, respectively.

The eight target properties were determined from these DFT calculations. Firstly, the interlayer binding energy (E_b), in $meVÅ^{-2}$, was estimated using the relation provided in **Equation 4.5**.

$$E_b = \frac{E_1 + E_2 - E_{hetero}}{S} \quad \textbf{4.5}$$

E_1 and E_2 represent total energy of the constituent monolayers, while E_{hetero} and S indicate the total energy of the heterostructure and its surface area, respectively. Next, the interlayer distance is taken as the vertical distance between topmost atom of the bottom monolayer and the lowest atom of the top monolayer as shown in **Figure 4.4 (d)**. The IE, EA, bandgap and work function, displayed in **Figure 4.4 (c)**, were determined using the relations specified in **Equations 4.6 – 4.9**.

$$IE = E_{vacuum} - E_{VBM} \quad \textbf{4.6}$$

$$EA = E_{vacuum} - E_{CBM} \quad \textbf{4.7}$$

$$E_g = IE - EA \quad \textbf{4.8}$$

$$\Phi = E_{vacuum} - E_F \quad \textbf{4.9}$$

4.1.2.4 Machine Learning Descriptors for van der Waals Heterostructures

The feature extraction for selected monolayers is performed using the C2DB (Haastrup et al., 2021) database and featurizers provided by the matminer (Ward et al., 2018) library. The structure-based features were obtained from 'StrToComposition' and 'ElementProperty' featurizers. On the other hand, the density and hybridization-based descriptors were obtained from the 'DOSFeaturizer' and 'Hybridization' featurizers.

DOSFeaturizer takes partial or complete density of states along with structure as input. It was employed on the pymatgen CompleteDos object, as obtained from the vasprun.xml file. The resulting features include information on species present in CBM (cbm_specie) and VBM (vbm_specie), the contribution of each of their orbitals (cbm_character and vbm_character), the amount of hybridization at band edges (cbm_hybridization and vbm_hybridization) and the fractional contribution of the first orbital (cbm_score and vbm_score). Whereas, the Hybridization featurizer quantifies the orbital and hybridizing orbital characters.

The orbitals were further encoded by introducing a cardinal descriptor which assigns the values of 0, 1 and 2 for s, p and d orbitals. The major orbital contributions were

encoded by One Hot Encoder that assigns a dummy binary value (0=False or 1=True). The vdW interactions for the species contributing to band edges were encoded by ElementProperty featurizer to include element-based properties such as melting point, Mendeleev number, atomic volume, vdW radius, and atomic radius (covalent radius). Consequently, 152 monolayer descriptors are obtained. In the cleaning and pre-processing, the data is analyzed for missing values and those missing data values are treated by taking the mean. The data points that deviate considerably from the usual patterns are filtered out, leaving us with 130 descriptors.

Feature engineering for bilayers was carried out by taking the mean of the component monolayer descriptors. Some additional bilayer features were constructed to include significant electronic properties, i.e., anderson_gap, wf_diff, min_cbm, max_cbm, min_vbm, max_vbm determining the Anderson's bandgap (anderson_gap = min_cbm - max_vbm), difference of monolayer work functions, and lower and higher band edges among the component monolayers. The categorical features for band alignment were constructed by determining the type-I, type-II and type-III heterostructures. As a result, 141 bilayer descriptors were obtained. Next, the descriptors with high skewness and low variance were filtered out, resulting in 97 bilayer descriptors, as listed in **Table A.2 (Appendix A)**.

Descriptor selection for each target property uses the L1 regularization technique offered by the Least absolute shrinkage and selection operation (LASSO) estimator (Kim et al., 2007). Optionally, it takes a penalty term to select descriptors by shrinking the LASSO coefficients of descriptors to zero. LASSO-selected set of bilayer descriptors is then supplied to multiple machine learning regression models (Cherkassky & Ma, 2003) supplied by Scikit-learn (Pedregosa et al., 2011) library. The ML models used here to construct ML pipelines include AdaBoost (AB) (J. Zhu et al., 2009), Elastic Net (EN), Ridge Regression (Ridge) (McCullagh & Nelder, 2017), Gradient Boosting Trees (GBT) (Friedman, 2001), Random Forest (RF) (Louppe, 2014), LASSO, Kernel Ridge Regression (KRR) (Kevin P. Murphy, 2012), Support Vector Machines (SVM) (Hastie, Rosset, Tibshirani, & Zhu, 2004; Platt, 1999) and Stacked Ensemble Meta-learner (SEM) (Wolpert, 1992).

We used repeated k-fold cross-validation (CV) (Kohavi & Edu, 1993) approach to estimate the generalization error. We applied five-fold CV with 20 repetitions and took

the mean score along with the standard deviation (STD). The performance of models on unseen data is assessed by plotting the predicted values against the actual values. Here, we used mean absolute error with standard deviation (MAE ± STD) and the coefficient of determination (R^2) as evaluation metrics to measure the prediction accuracy of each model. MAE is the average absolute difference between predicted and true values. A low MAE value indicates that the predictions made by the model are close to the true values, signifying good performance. Besides this, the absolute in MAE suggests that it is robust to outliers. R^2 represents the proportion of variance in the dependent variable that can be predicted by the independent variable. A high R^2 value indicates that the model is capable of explaining the diversity and spread of values in the target variable. MAE and R^2 metrics are determined by relations given in **Equation 4.10** and **4.11**, where $y_i, \overline{y}$ and $\dot{y}_i$ represent the predicted, average and true value of the target, respectively. The value of MAE (R^2) close to 0 (1) indicate the best performing model.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}(y_i - \dot{y}_i)}{n} \qquad \textbf{4.10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \overline{y})}{\sum_{i=1}^{n}(\dot{y}_i - \overline{y})} \qquad \textbf{4.11}$$

4.1.3 Results and discussion

4.1.3.1 High-Throughput DFT Results

The lattice mismatch for the labeled training set of 412 bilayers ranges between 0.02 and 3.99 %, while the interlayer distance falls in the range of 1.98 – 5.51 Å, with an average distance of 3.67 Å. **Equation 4.5** suggests that positive binding energy values indicate stable structures. Therefore, we eliminated the bilayers having interlayer distance less than 2.50 Å and negative binding energies. We eliminated the bilayers by selecting only those having binding energies ≤ 500 meVÅ^{-2}, leaving behind 348 bilayers for training the models. In the resulted structural dataset, binding energy and interlayer distance lie in the ranges of 1.42 – 390 meVÅ^{-2} and 2.61 – 4.62 Å. The bandgap values of labeled bilayers range between ~0 eV and 3.66 eV. For training the models for

electronic properties, we removed the metallic and semi-metallic bilayers since they may cause complications for ML models. Further, we eliminated bilayers with type-III band alignment, which leaves us with 230 bilayers. After filtering the labeled datasets, training data was scaled to have zero mean and unit standard deviation (STD). Among them, 47 bilayers are identified as direct bandgap vdW heterostructures as listed in **Table B.1 (Appendix B)**.

4.1.3.2 ML-based Prediction of Electronic properties

After filtering, we used 230 bilayers to train the ML models. The bilayer descriptors were first passed to LASSO to select a unique set of features for each target property. LASSO-selected bilayer features are then employed by other ML models to predict the target property. LASSO selected 33 descriptors for IE and EA pipelines while 64 descriptors for E_g pipeline. The descriptor space is reduced from 97 to 50 for E_{CBM} and E_{VBM} pipelines and further reduced to 34 for the work function. The corresponding lasso coefficients and correlation map for the selected set of descriptors are shown in **Figure C1 – C12 (Appendix C)**. The magnitude of Lasso coefficients assesses the importance of features for a particular target property. The higher the coefficient value for a feature, the more important a feature is. Meanwhile, the positive and negative correlations in the correlation map represent the direct and inverse relation between descriptors, respectively.

For all electronic properties, average lattice parameter (avg_lattice_param) acts as an important feature. Further, it can be observed that the maximum value of VBM among the component monolayers (max_vbm) is a strong predictor of bilayer IE. Similarly, the minimum value of CBM among two monolayers (min_cbm) is a strong predictor of bilayer EA. These findings align well with the relations of IE and EA with band edges. Anderson's bandgap was calculated as a rough estimation of bandgap, which is determined by the difference between the lower E_{CBM} and the higher E_{VBM} value among the constituent monolayers (**Equation 4.4**).

For the bilayer bandgap energy, the mean of the bandgaps of the component monolayers (avg_gap_nosoc) acts as a strong predictor. Most importantly, the features based on vdW interactions at the band edges are crucial in determining E_{CBM} and E_{VBM}. In contrast to the band edge-based features, the mean work function (avg_workfunction)

and mean bandgap (avg_gap_nosoc, avg_gap_dir) of the two monolayers act as a strong predictor of bilayer band edge energy levels (E_{CBM} and E_{VBM}).

The mean work function (avg_workfunction) and the difference of work functions of the constituents (wf_diff) are strong predictors of the bilayer work function. The work function formula suggests that the Fermi energy level and vacuum level should play a significant role in predicting bilayer work function. However, these features are absent from the LASSO-selected list of features. It may result in poor performance of models trained on the selected set of features. The high value of wf_diff suggests a greater electric field at the interface and accelerated the interfacial charge transfer.

The LASSO-selected features are used to train several linear, non-linear and ensemble models, and the performance of models on training and validation sets is determined by average MAE with a standard deviation by five-fold cross-validation. Further, we calculated the R^2 score on the validation set to compare the predictive performance of ML models.

Table 4.2. MAE ± STD of IE, EA and E_g pipelines for training (validation) set determined by five-fold cross-validation.

S. No.	ML Model	IE (eV)	EA (eV)	E_g (eV)
1.	LASSO	0.268 ± 0.053 (0.213)	0.342 ± 0.060 (0.241)	0.286 ± 0.040 (0.213)
2.	Ridge	0.271 ± 0.052 (0.223)	0.336 ± 0.062 (0.254)	0.284 ± 0.041 (0.228)
3.	SVM (rbf)	0.290 ± 0.072 (0.209)	0.330 ± 0.078 (0.24)	0.299 ± 0.050 (0.257)
4.	SVM (linear)	0.256 ± 0.065 (0.194)	0.335 ± 0.072 (0.251)	0.288 ± 0.040 (0.236)
5.	AB	0.326 ± 0.064 (0.278)	0.390 ± 0.064 (0.257)	0.331 ± 0.045 (0.268)
6.	RF	0.310 ± 0.065 (0.26)	0.362 ± 0.071 (0.232)	0.334 ± 0.046 (0.257)
7.	GBT	n/a	n/a	0.287 ± 0.052 (0.227)
8.	EN	0.270 ± 0.052 (0.216)	0.342 ± 0.061 (0.242)	0.286 ± 0.040 (0.222)
9.	KRR (poly)	n/a	n/a	0.268 ± 0.040 (0.191)

10.	SEM	0.260 ± 0.054 (0.232)	0.323 ± 0.060 (0.238)	0.261 ± 0.043 (0.20)

For IE, EA and Eg, the ML models yield predictions with average MAEs as low as 0.19, 0.23 and 0.19 eV, respectively. **Table 4.2** shows that training MAE across these targets is as little as 0.25 eV. Further, the minor standard deviation in error, in hundredths of eV, suggests the robustness of models against data changes caused by resampling.

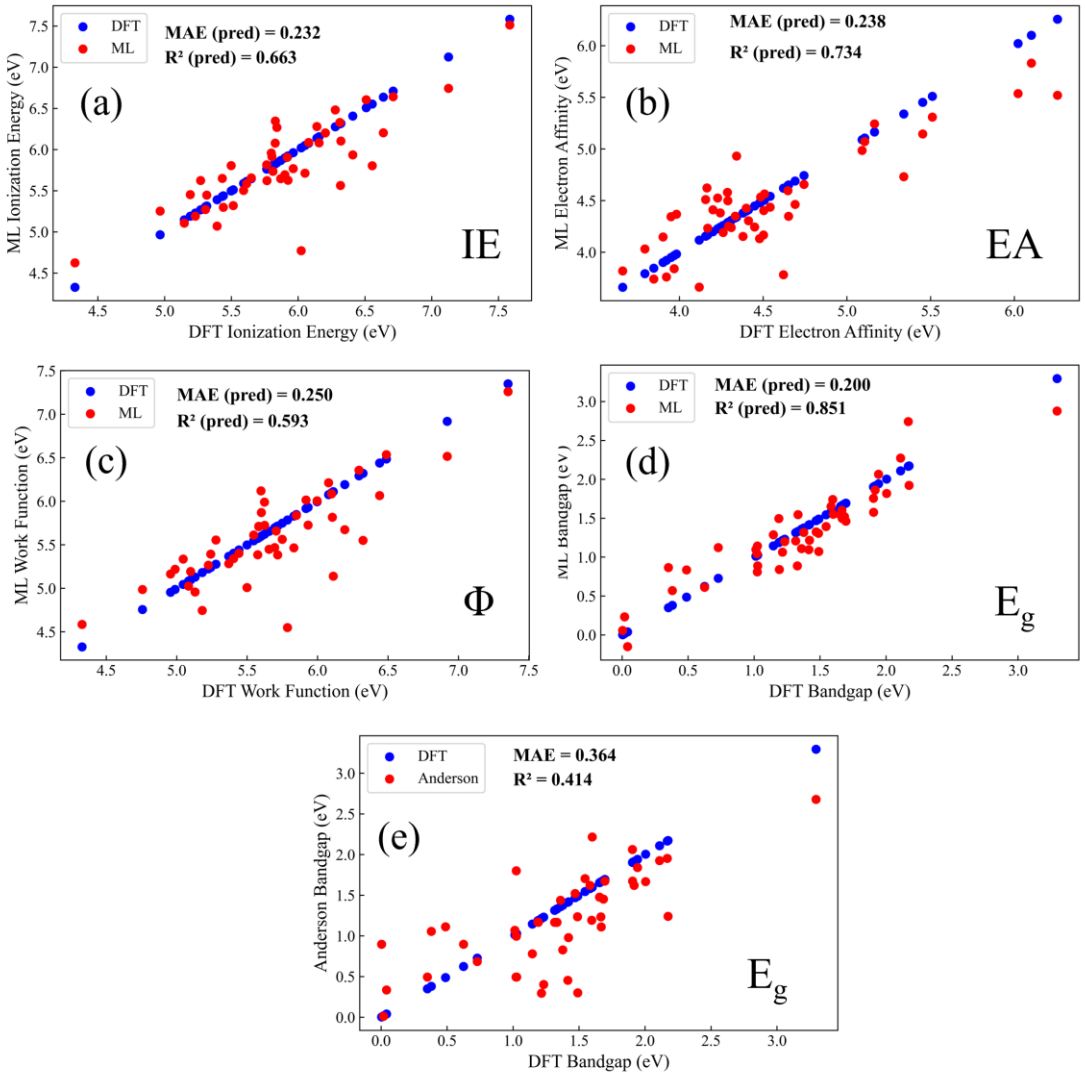

Figure 4.5. Comparison of DFT and ML predicted (a) IE (b) EA (c) Φ (d) bandgap for validation set using SEM and (e) DFT and Anderson's bandgap. The evaluation metrics are provided for each target.

Each model has its strengths and weaknesses. Therefore, selecting the correct model is necessary to minimize the prediction error. The regularized linear models such as Lasso,

Ridge, and EN give higher MAE. These models lack complexity and are convenient to implement, offering no place for non-linear relationships. In contrast, the tree-based boosting models such as GBT, RF and AB capture non-linearity in data and thus perform far better than linear models. However, these models are sensitive to noise, resulting in bias and overfitting, and may require large datasets for training to minimize noise. One way to alleviate this noise is regularization, which we applied before training the models (Kim et al., 2007). These models usually give lower errors on the validation set than on the training set, suggesting overfitting. The KRR model with a polynomial kernel also captures the non-linear relation by fitting a polynomial curve, improving the prediction accuracy.

To balance the strengths and weaknesses of different ML models, we employed a Stacked Ensemble Meta-learner (SEM), which takes multiple models as base models. SEM is known to perform far better than the individual models. For diversity, we stacked several models (i.e., tree-based, linear and kernel-based), GBT, Ridge and polynomial KRR to construct a high-level meta-learner. The predictions from these base models are then passed to the ridge regression meta-estimator. The predictions using SEM for IE, EA and Eg are shown in **Figure 4.5 (a, b, c).**

Figure 4.5 (d, e) suggests that the performance of stacked regressor (SEM) is much better than Anderson's bandgap with an average MAE of 0.2 eV, since it takes several other bilayer descriptors besides the edge-based features that are considered by Anderson's rule. LASSO estimator suggests that the bandgap target strongly relies on the vdW radius, interfacial charge transfer and average bandgap of constituent layers. Whereas, Anderson's approach considers only the VBM and CBM energy levels. Therefore, bandgap estimation using ML models suggests greater accuracy.

For Φ, E_{CBM} and E_{VBM}, the ML models yield predictions with average MAEs as low as 0.185, 0.17 and 0.166 eV, respectively, as listed in **Table 4.3**. The training MAE across these electronic target properties is as little as 0.20 eV. Additionally, the minor STD in average MAE, in hundredths of eV, suggests the robustness of models against data changes caused by resampling. The GBT and polynomial KRR models are discarded for the work function pipeline due to high generalization error.

Table 4.3. MAE ± STD of Φ, E_{CBM}, E_{VBM} pipelines for training (validation) set determined by five-fold cross-validation.

S. No.	ML Model	Φ (eV)	E_{CBM} (eV)	E_{VBM} (eV)
1.	LASSO	0.278 ± 0.052 (0.221)	0.297 ± 0.035 (0.195)	0.207 ± 0.031 (0.188)
2.	Ridge	0.283 ± 0.051 (0.227)	0.278 ± 0.036 (0.207)	0.213 ± 0.031 (0.181)
3.	SVM (rbf)	0.293 ± 0.072 (0.193)	0.291 ± 0.046 (0.234)	0.255 ± 0.042 (0.193)
4.	SVM (linear)	0.269 ± 0.066 (0.185)	0.289 ± 0.041 (0.217)	0.230 ± 0.040 (0.193)
5.	AB	0.337 ± 0.089 (0.268)	0.336 ± 0.038 (0.229)	0.267 ± 0.036 (0.256)
6.	RF	0.311 ± 0.064 (0.238)	0.330 ± 0.046 (0.224)	0.259 ± 0.038 (0.223)
7.	GBT	n/a	0.293 ± 0.044 (0.192)	0.223 ± 0.033 (0.215)
8.	EN	0.281 ± 0.050 (0.223)	0.287 ± 0.036 (0.204)	0.206 ± 0.031 (0.183)
9.	KRR (poly)	n/a	0.273 ± 0.040 (0.212)	0.229 ± 0.038 (0.166)
10.	SEM	0.272 ± 0.054 (0.25)	0.263 ± 0.036 (0.174)	0.201 ± 0.036 (0.185)

The coefficient of determination (R^2) for the validation set determines the predictive performance of trained models, where a value close to 1 suggests the best performance. The performance of employed ML models on the validation set for electronic properties is represented in **Figure C17 – C22 (Appendix C)**.

R^2 scores, listed in **Table 4.4**, suggest that the trained models predict E_{CBM} and E_{VBM} with higher accuracy in contrast to other targets. Further, SEM, SVM regressor and linear models perform better than the rest. The models employed here may perform better in the absence of cross-validation (Abraham et al., 2023), reducing the MAE and improving R^2 further. Further, employing HSE06 functional for labeling electronic properties may significantly improve the predictive performance of models (Z. Zhu, Dong, Guo, Yang, & Zhang, 2020). The predictions using SEM for band edges are shown in **Figure 4.6 (a, b)**. It can be observed that E_{CBM} and E_{VBM} pipelines perform much better than IE and EA pipelines.

Table 4.4. R^2 score of data-driven ML models on unseen dataset indicating prediction accuracy of electronic properties.

S. No.	ML Model	IE	EA	E_g	Φ	E_{CBM}	E_{VBM}
1.	LASSO	0.749	0.716	0.812	0.718	0.823	0.81
2.	Ridge	0.729	0.705	0.793	0.699	0.8	0.83
3.	SVM (rbf)	0.759	0.642	0.776	0.767	0.7	0.791
4.	SVM (linear)	0.778	0.698	0.787	0.785	0.773	0.812
5.	AB	0.581	0.699	0.69	0.57	0.759	0.646
6.	RF	0.582	0.713	0.712	0.642	0.741	0.711
7.	GBT	n/a	n/a	0.774	n/a	0.819	0.7
8.	EN	0.742	0.716	0.804	0.708	0.806	0.819
9.	KRR (poly)	n/a	n/a	0.85	n/a	0.789	0.828
10.	SEM	0.663	0.734	0.851	0.593	0.857	0.813

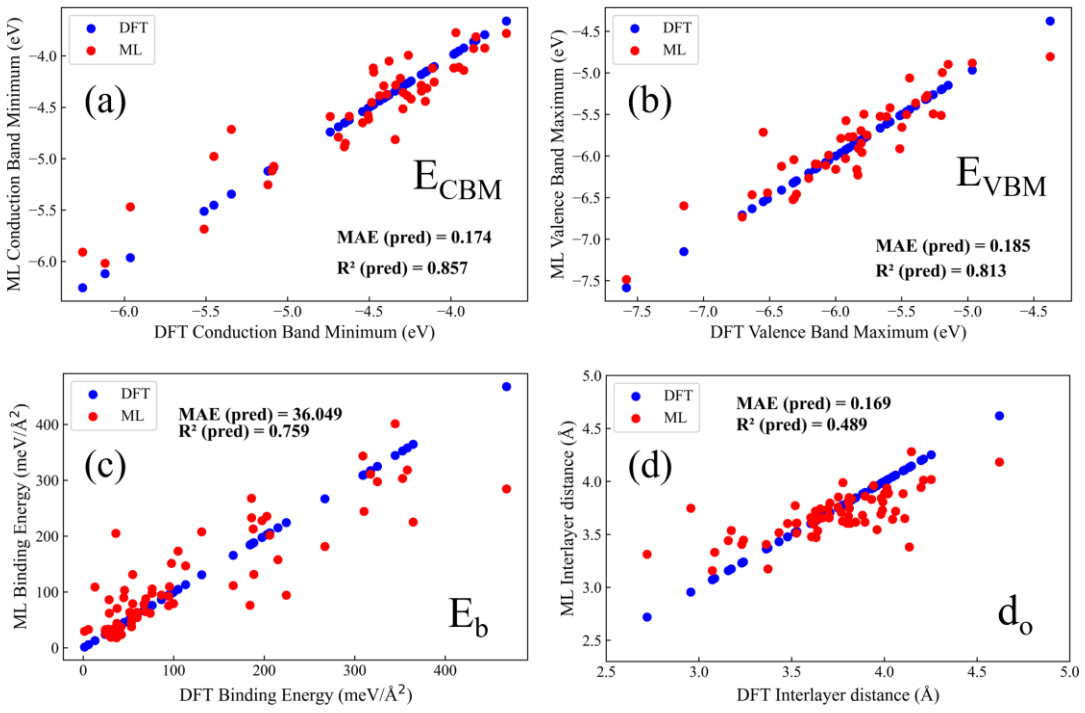

Figure 4.6. Comparison of DFT and ML predicted (a) E_{CBM} (b) E_{VBM} (c) E_b and (d) d_o for validation set using SEM. The evaluation metrics are provided for each target.

4.1.3.3 ML-based Prediction of Binding Energy and Interlayer Distance

After filtering, we obtained 348 bilayers for training the ML models for interlayer distance and binding energy. The descriptor space is reduced from 97 to 43 and 37 for interlayer E_b and d_o pipelines. The corresponding lasso coefficients and correlation map for the selected set of descriptors are shown in **Figure C13 – C16 (Appendix C)**.

Table 4.5. MAE ± STD for training (validation) set determined by five-fold cross validation and R^2 score of models on unseen dataset indicating predicting accuracy of data-driven ML models for interlayer distance and interlayer binding energy.

S. No.	ML Model	d_o		E_b	
		MAE ± STD (Å)	R^2 Score	MAE ± STD (meV/Å^{-2})	R^2 Score
1.	LASSO	0.199 ± 0.033 (0.189)	0.408	42.380 ± 5.661 (44.526)	0.685
2.	Ridge	0.192 ± 0.032 (0.187)	0.418	41.673 ± 5.244 (45.474)	0.682
3.	SVM (rbf kernel)	0.175 ± 0.040 (0.166)	0.496	n/a	
4.	SVM (linear kernel)	0.178 ± 0.037 (0.183)	0.413	41.614 ± 5.194 (45.072)	0.684
5.	AB	0.218 ± 0.044 (0.201)	0.403	38.038 ± 5.558 (42.975)	0.68
6.	RF	0.189 ± 0.039 (0.164)	0.511	38.332 ± 5.494 (37.018)	0.733
7.	GBT	0.185 ± 0.046 (0.176)	0.473	35.345 ± 5.597 (35.428)	0.76
8.	EN	0.196 ± 0.033 (0.19)	0.398	41.775 ± 5.266 (45.283)	0.683
9.	KRR (poly kernel)	0.179 ± 0.037 (0.17)	0.457	37.986 ± 5.657 (37.326)	0.729
10.	SEM	0.192 ± 0.028 (0.169)	0.489	35.090 ± 5.092 (36.049)	0.759

Besides the elemental features, the vdW radius and work function features are strong predictor of both targets. It suggests that amount of interfacial charge transfer strongly affects the interlayer distance and binding energy. Further, the score of p-orbital of band

edges is highly correlated to vdW radius feature. The cell area is a good predictor of E_b target, agreeing well with its formula. On the other hand, lattice mismatch and type of band edge alignment are weak predictors of binding energy.

We pass the LASSO-selected set of descriptors to multiple ML models and assess their performance by five-fold cross-validation. For E_b and d_o, the ML models yield predictions with average MAEs as low as 35.428 meV/Å^{-2} and 0.164 Å, respectively, as listed in **Table 4.5**. The training MAE across these target properties is as little as 35 meV/Å^{-2} and 0.175 Å. Additionally, the minor standard deviation in average MAE for d_o, in hundredths of Å, suggests the robustness of models against data changes caused by resampling. The interlayer binding energy shows high MAEs since we expressed it in units of meV/Å^{-2}. These findings suggest that E_b is predicted within tens of meV/Å^{-2} while d_o is estimated within tenths (1/10) of Å.

The R^2 score suggests the predictive performance of trained models. The performance of employed ML models on the validation set for these target properties is represented in **Figure C23 – C24 (Appendix C)**. The predictions using SEM for interlayer distance and binding energy are shown in **Figure 4.6 (c, d)**. R^2 scores, listed in **Table 4.5**, suggest that the tree-based models (i.e., GBT and RF) and SEM predict properties with greater accuracy in contrast to other models.

4.1.4 Prediction for Water-Splitting Photocatalysis using ML Models

The ability to oxidize or reduce water can be evaluated by aligning the band edges of bilayers against water redox potentials. For a water-splitting photocatalyst to facilitate hydrogen evolution reaction (HER), its CBM should be higher than the water reduction potential, i.e., $E_{CBM} > V_{H_+/H_2} = -4.44$ eV. Similarly, to promote oxygen evolution reaction (OER), its VBM should be lower than the water oxidation potential, i.e., $E_{VBM} < V_{H_2O/O_2} = -5.67$ eV (J. Liu et al., 2014).

In the labeled dataset of 230 vdW bilayers, we identified 128 (142) bilayers that incorporate only water reduction (oxidation) potential, revealing their suitability for HER (OER). Among them, we identified 63 vdW bilayers with the feasibility of overall photocatalysis by incorporating water reduction and oxidation levels. The DFT-

predicted bilayers feasible for both hydrogen and oxygen production are summarized in **Table B.2 (Appendix B)**. The trained data-driven models were employed to predict the target properties of the unlabeled dataset. Based on the target properties determined by SEM, 424, 575 and 93 vdW bilayers were determined with suitability for HER, OER and overall water-splitting photocatalysis, respectively. The ML-predicted bilayers driving both HER and OER are summarized in **Table C.1 (Appendix C)**. Further, the bandgaps of these bilayers lie in the desirable range of 0.98 – 4 eV (UV-Visible region). The alignment of band edges for these photocatalysts against water redox levels is shown in **Figure 4.7**.

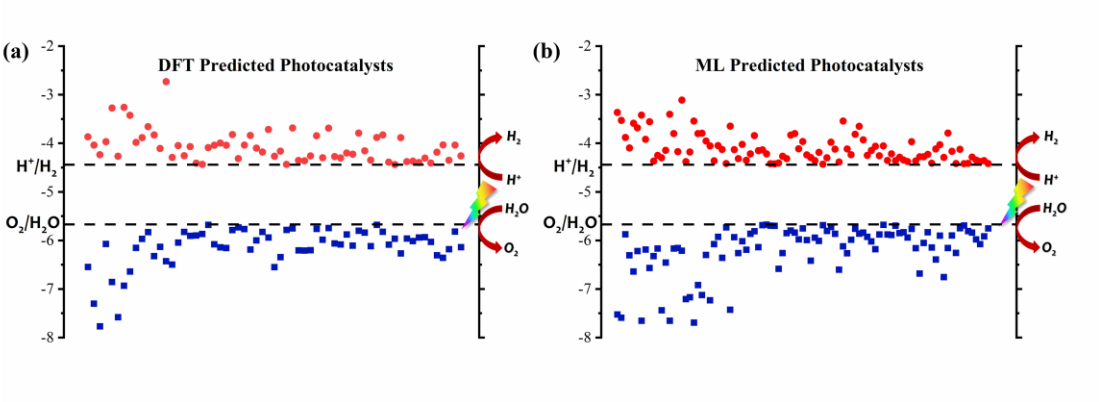

Figure 4.7. The overall water-splitting photocatalysts obtained from (a) the labeled DFT calculated set and (b) the unlabeled ML predicted dataset. Water reduction and oxidation levels are marked by black dotted lines.

CHAPTER 5
CONCLUSIONS

5.1 Conclusions

Hexagonal 2D materials have evolved as a potential class of 2D nanomaterials with extraordinary features, drawing significant research since the discovery of graphene in 2004. Tailoring the properties of hexagonal monolayers by constructing vdW heterostructures is a great way to benefit from the properties of individual monolayers. The weak vdW forces permit control over the monolayer characteristics and offer adequate spatial separation of charge carriers in contrast to the chemical bonds in the traditional heterostructures. The formation of these composites provides unique features and boosts photocatalytic activity due to the synergy of their components. We developed a machine learning framework with high-throughput DFT calculations to predict the properties of bilayer vdW heterostructures. We aggregated the monolayer features to construct the bilayer descriptors. Efficient feature selection using the LASSO estimator permitted the training of more accurate and efficient machine learning models on a smaller feature set, enhancing predictive performance while maintaining result integrity. Multiple supervised ML models were trained on those descriptors to construct ML pipelines for eight target properties: IE, EA, E_g, Φ, E_{CBM}, E_{VBM}, E_b and d_o. The models were evaluated for their performance on training and validation sets. Once trained, models were employed to predict the selected target properties of unseen and unlabeled datasets. The E_g, E_{CBM} and E_{VBM} pipelines revealed MAE (R^2) of 0.2 (0.85), 0.17 (0.86) and 0.18 (0.81) eV, respectively, employing stacked regressor with five-fold cross-validation.

The hybrid technique proved that data-driven models trained on DFT-calculated attributes could accurately predict the structural and electronic properties of novel, unexplored vdW heterostructures, allowing for rapid identification and screening of materials with optimum properties. From the labeled and unlabeled datasets, the bilayer photocatalysts feasible for HER, OER and overall photocatalysis were screened. From the DFT and ML-predicted datasets, 370 and 999 bilayers were identified as promoting only HER and OER. Among them, 156 bilayers incorporate both water oxidation and reduction potentials, highlighting their suitability for both HER and OER. Therefore, our ML framework can confidently predict a large set of 2D vdW bilayer photocatalysts efficiently and accurately. These findings demonstrate the potential of merging DFT and machine learning to speed up the development and optimization of innovative materials

for particular applications, such as photocatalysis, by offering a robust prediction strategy. Future studies may concentrate on applying this technique to diverse material systems, incorporating experimental validation, and investigating sophisticated machine-learning models to improve prediction capabilities.

REFERENCES

Abraham, B. M., Sinha, P., Halder, P., & Singh, J. K. (2023). Fusing a machine learning strategy with density functional theory to hasten the discovery of 2D MXene-based catalysts for hydrogen generation. *Journal of Materials Chemistry A*, *11*(15), 8091–8100. https://doi.org/10.1039/d3ta00344b

Aewfc, T. (2022). *PAW All-Electron Wavefunction in VASP*. (6), 1–9.

Back, S., Tran, K., & Ulissi, Z. W. (2019). Toward a Design of Active Oxygen Evolution Catalysts: Insights from Automated Density Functional Theory Calculations and Machine Learning. *ACS Catalysis*, *9*(9), 7651–7659. https://doi.org/10.1021/acscatal.9b02416

Belsky, A., & Lynn, V. (2002). New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica*, *58*, 364–369.

Bian, R., Li, C., Liu, Q., Cao, G., Fu, Q., Meng, P., … Liu, Z. (2022). Recent progress in the synthesis of novel two-dimensional van der Waals materials. *National Science Review*, *9*(5). https://doi.org/10.1093/nsr/nwab164

Bonacci, M., Qiao, J., Spallanzani, N., Marrazzo, A., Pizzi, G., Molinari, E., … Prezzi, D. (2023). Towards high-throughput many-body perturbation theory: efficient algorithms and automated workflows. *Npj Computational Materials*, *9*(1), 1–10. https://doi.org/10.1038/s41524-023-01027-2

Chandrasekaran, A., Kamal, D., Batra, R., Kim, C., Chen, L., & Ramprasad, R. (2019). Solving the electronic structure problem with machine learning. *Npj Computational Materials*, *5*(1). https://doi.org/10.1038/s41524-019-0162-7

Chen, C., Xiao, B., Li, Z., Li, W., Li, Q., & Yu, X. (2024). High throughput screening for electrocatalysts for nitrogen reduction reaction using metal-doped bilayer borophene: A combined approach of DFT and machine learning. *Molecular Catalysis*, *557*(February), 113972. https://doi.org/10.1016/j.mcat.2024.113972

Cherkassky, V., & Ma, Y. (2003). Comparison of model selection for regression. *Neural Computation*, *15*(7), 1691–1714. https://doi.org/10.1162/089976603321891864

Chibani, S., & Coudert, F. X. (2020). Machine learning approaches for the prediction of materials properties. *APL Materials*, *8*(8). https://doi.org/10.1063/5.0018384

Choudhary, K., Garrity, K. F., Pilania, G., & Tavazza, F. (2020). *Efficient Computational Design of 2D van der Waals Heterostructures: Band-Alignment, Lattice-Mismatch, Web-app Generation and Machine-learning*. 1–29. Retrieved from http://arxiv.org/abs/2004.03025

Choudhary, K., Garrity, K. F., Reid, A. C. E., Decost, B., Biacchi, A. J., Walker, A. R. H., … Kalinin, S. V. (2020). The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *Npj Computational Materials*, *6*(173). https://doi.org/10.1038/s41524-020-00440-1

Dong, R., Jacob, A., Bourdais, S., & Sanvito, S. (2021). High-throughput bandstructure simulations of van der Waals hetero-bilayers formed by 1T and 2H monolayers. *Npj 2D Materials and Applications*, *5*(1), 1–12. https://doi.org/10.1038/s41699-021-00200-9

Duan, C., Liu, F., Nandy, A., & Kulik, H. J. (2021). Putting Density Functional Theory to the Test in Machine-Learning-Accelerated Materials Discovery. *Journal of Physical Chemistry Letters*, *12*(19), 4628–4637. https://doi.org/10.1021/acs.jpclett.1c00631

Fiedler, L., Modine, N. A., Schmerler, S., Vogel, D. J., Popoola, G. A., Thompson, A. P., … Cangi, A. (2023). Predicting electronic structures at any length scale with machine learning. *Npj Computational Materials*, *9*(1), 1–10. https://doi.org/10.1038/s41524-023-01070-z

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Fujishima, A., & Honda, K. (1972). Electrochemical Photolysis of Water at a Semiconductor Electrode. *Nature*, *238*(5358), 37–38. https://doi.org/10.1038/238037a0

Gao, M., Cai, B., Liu, G., Xu, L., Zhang, S., & Zeng, H. (2023). Machine learning and density functional theory simulation of the electronic structural properties for

novel quaternary semiconductors. *Physical Chemistry Chemical Physics*, *25*(13), 9123–9130. https://doi.org/10.1039/d2cp04244d

Gao, Y., Zhang, Q., Hu, W., & Yang, J. (2024). First-Principles Computational Screening of Two-Dimensional Polar Materials for Photocatalytic Water Splitting. *ACS Nano*, *18*(29), 19381–19390. https://doi.org/10.1021/acsnano.4c06544

Ge, L., Ke, Y., & Li, X. (2023). Machine learning integrated photocatalysis: progress and challenges. *Chemical Communications*, *59*(39), 5795–5806. https://doi.org/10.1039/d3cc00989k

Geim, A. K., & Grigorieva, I. V. (2013). Van der Waals heterostructures. *Nature*, *499*(7459), 419–425. https://doi.org/10.1038/nature12385

Gjerding, M. N., Taghizadeh, A., Rasmussen, A., Ali, S., Bertoldo, F., Deilmann, T., … Thygesen, K. S. (2021). Recent progress of the computational 2D materials database (C2DB). *2D Materials*, *8*(4). https://doi.org/10.1088/2053-1583/ac1059

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. *Journal of Chemical Physics*, *132*(15). https://doi.org/10.1063/1.3382344

Guo, Z., Zhou, J., Zhu, L., & Sun, Z. (2016). MXene: A promising photocatalyst for water splitting. *Journal of Materials Chemistry A*, *4*(29), 11446–11452. https://doi.org/10.1039/c6ta04414j

Haastrup, S., Strange, M., Pandey, M., Deilmann, T., Schmidt, P. S., Hinsche, F., … Kristian, S. (2021). *Computational 2D Materials Database (C2DB)*. 1–13.

Haastrup, S., Strange, M., Pandey, M., Deilmann, T., Schmidt, P. S., Hinsche, N. F., … Thygesen, K. S. (2018). The Computational 2D Materials Database: High-throughput modeling and discovery of atomically thin crystals. *2D Materials*, *5*(4). https://doi.org/10.1088/2053-1583/aacfc1

Hafner, J., & Kresse, G. (1997). The Vienna AB-Initio Simulation Program VASP: An Efficient and Versatile Tool for Studying the Structural, Dynamic, and Electronic

Properties of Materials. In *Properties of Complex Inorganic Solids*. https://doi.org/10.1007/978-1-4615-5943-6

Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The Entire Regularization Path for the Support Vector Machine. *Journal OfMachine Learning Research*, (5), 1391–1415. https://doi.org/10.1007/11866565_30

Huang, G., Huang, F., & Dong, W. (2024). Machine learning in energy storage material discovery and performance prediction. *Chemical Engineering Journal*, 152294. https://doi.org/10.1016/j.cej.2024.152294

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., … Persson, K. A. (2013). The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, *1*(011002). https://doi.org/10.1063/1.4812323

Jalil, A., Zhao, T., Kanwal, A., & Ahmed, I. (2023). Prediction of direct Z-scheme H and Ḧ-phase of MoSi2N4/MoSX (X = S, Se) van der Waals heterostructures: A promising candidate for photocatalysis. *Chemical Engineering Journal*, *470*, 144239. https://doi.org/10.1016/j.cej.2023.144239

Jiang, X., Wang, Y., Jia, B., Qu, X., & Qin, M. (2022). Prediction of Oxygen Evolution Activity for NiCoFe Oxide Catalysts via Machine Learning. *ACS Omega*, *7*(16), 14160–14164. https://doi.org/10.1021/acsomega.2c00776

Jin, C., Ma, E. Y., Karni, O., Regan, E. C., Wang, F., & Heinz, T. F. (2018). Ultrafast dynamics in van der Waals heterostructures. *Nature Nanotechnology*, *13*(11), 994–1003. https://doi.org/10.1038/s41565-018-0298-5

Jing, H., Guan, C., Yang, Y., & Zhu, H. (2023). Machine learning-assisted design of AlN-based high-performance piezoelectric materials. *Journal of Materials Chemistry A*, *11*(27), 14840–14849. https://doi.org/10.1039/d3ta02095a

Kevin P. Murphy. (2012). *Machine learning: a probabilistic perspective*. Retrieved from https://research.google/pubs/pub38136/%0Ahttps://books.google.co.uk/books?hl=en&lr=&id=RC43AgAAQBAJ&oi=fnd&pg=PR7&dq=Machine+Learning+-+A+Probabilistic+Perspective+(2012)&ots=umlueCPx-

7&sig=Fez88u5ceM3dJjO9DGl3dugeR6Y#v=onepage&q=Machine Learning - A Prob

Kim, S. J., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2007). An interior-point method for large-scale ℓ1-regularized least squares. *IEEE Journal on Selected Topics in Signal Processing*, *1*(4), 606–617. https://doi.org/10.1109/JSTSP.2007.910971

Kirklin, S., Saal, J. E., Meredig, B., Thompson, A., Doak, J. W., Aykol, M., … Wolverton, C. (2015). The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *Npj Computational Materials*, *1*(September). https://doi.org/10.1038/npjcompumats.2015.10

Kohavi, R., & Edu, S. (1993). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, *2*, 1137–1143.

Larsen, A. H., Mortsen, J. J., & Blomqvist, J. (2019). The Atomic Simulation Environment — A Python library for working with atoms. In *Materials Today: Proceedings* (Vol. 27). Retrieved from https://doi.org/10.1016/j.jare.2020.01.010%0Ahttps://doi.org/10.1016/j.nano.2021.102426%0Ahttps://doi.org/10.1080/03008207.2019.1617280%0Ahttp://dx.doi.org/10.1038/s41598-019-38972-2%0Ahttps://doi.org/10.1016/j.matpr.2019.12.188%0Ahttps://doi.org/10.1016/

Lee, Y., Hwang, Y., & Chung, Y. C. (2015). Achieving Type I, II, and III Heterojunctions Using Functionalized MXene. *ACS Applied Materials and Interfaces*, *7*(13), 7163–7169. https://doi.org/10.1021/acsami.5b00063

Li, X., Dai, Y., Li, M., Wei, W., & Huang, B. (2015). Stable Si-based pentagonal monolayers: High carrier mobilities and applications in photocatalytic water splitting. *Journal of Materials Chemistry A*, *3*(47), 24055–24063. https://doi.org/10.1039/c5ta05770a

Liu, B., Chen, L., Liu, G., Abbas, A. N., Fathi, M., & Zhou, C. (2014). High-performance chemical sensing using Schottky-contacted chemical vapor

deposition grown monolayer MoS2 transistors. *ACS Nano*, *8*(5), 5304–5314. https://doi.org/10.1021/nn5015215

Liu, J., Li, X. B., Wang, D., Liu, H., Peng, P., & Liu, L. M. (2014). Single-layer Group-IVB nitride halides as promising photocatalysts. *Journal of Materials Chemistry A*, *2*(19), 6755–6761. https://doi.org/10.1039/c3ta15431a

Liu, M., Gopakumar, A., Hegde, V. I., He, J., & Wolverton, C. (2024). High-Throughput hybrid-functional DFT calculations of bandgaps and formation energies and multifidelity learning with uncertainty quantification. *Physical Review Materials*, *8*(4), 1–12. https://doi.org/10.1103/PhysRevMaterials.8.043803

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. (July). Retrieved from http://arxiv.org/abs/1407.7502

Low, J., Cao, S., Yu, J., & Wageh, S. (2014). Two-dimensional layered composite photocatalysts. *Chemical Communications*, *50*(74), 10768–10777. https://doi.org/10.1039/c4cc02553a

Luis, N. (2016). *Advanced Catalytic Materials* (L. E. Norena & J.-A. Wang, Eds.). Rijeka: IntechOpen. https://doi.org/10.5772/60491

Lv, X., Wei, W., Sun, Q., Li, F., Huang, B., & Dai, Y. (2017). Two-dimensional germanium monochalcogenides for photocatalytic water splitting with high carrier mobility. *Applied Catalysis B: Environmental*, *217*, 275–284. https://doi.org/10.1016/j.apcatb.2017.05.087

McCullagh, P., & Nelder, J. A. (2017). *Generalized linear models*. https://doi.org/10.1201/9780203738535

Min, J., Zhou, M., Zhang, C., Tang, C., Peng, X., & Zhong, J. (2021). Type-II vdW heterojunction SeGa2Te/SeIn2Se as a high-efficiency visible-light-driven water-splitting photocatalyst. *Physics Letters A*, *413*, 127594. https://doi.org/https://doi.org/10.1016/j.physleta.2021.127594

Mishra, A., Satsangi, S., Rajan, A. C., Mizuseki, H., Lee, K. R., & Singh, A. K. (2019). Accelerated data-driven accurate positioning of the band edges of MXenes.

Journal of Physical Chemistry Letters, *10*(4), 780–785. https://doi.org/10.1021/acs.jpclett.9b00009

Mortensen, J. J., Hansen, L. B., & Jacobsen, K. W. (2005). Real-space grid implementation of the projector augmented wave method. *Physical Review B - Condensed Matter and Materials Physics*, *71*(3), 1–11. https://doi.org/10.1103/PhysRevB.71.035109

Mueller, T., Kusne, A., & Ramprasad, R. (2016). Machine Learning in Materials Science: Recent progress and emerging applications. *Reviews in Computational Chemistry*, *29*(i), 186–273.

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., … Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, *68*(February), 314–319. https://doi.org/10.1016/j.commatsci.2012.10.028

Opoku, F., Govender, K. K., Van Sittert, C. G. C. E., & Govender, P. P. (2017). Role of MoS2 and WS2 monolayers on photocatalytic hydrogen production and the pollutant degradation of monoclinic BiVO4: A first-principles study. *New Journal of Chemistry*, *41*(20), 11701–11713. https://doi.org/10.1039/c7nj02340e

Pack, J. D., & Monkhorst, H. J. (1976). Special points for Brillouin-zone integrations. *Physical Review B*, *13*(12), 5188. https://doi.org/10.1103/PhysRevB.13.5188

Paquin, F., Rivnay, J., Salleo, A., Stingelin, N., & Silva, C. (2015). Recent Advances in 2D Materials for Photocatalysis. *J. Mater. Chem. C*, *3*, 10715–10722. https://doi.org/10.1039/b000000x

Parse, N., & Pinitsoontorn, S. (2023). Machine learning for predicting ZT values of high-performance thermoelectric materials in mid-temperature range. *APL Materials*, *11*(8). https://doi.org/10.1063/5.0160055

Pederson, R., Kalita, B., & Burke, K. (2022). Machine learning and density functional theory. *Nature Reviews Physics*, *4*(6), 357–358. https://doi.org/10.1038/s42254-022-00470-2

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, Q., & De, S. (2013). Mechanical properties and instabilities of ordered graphene oxide C 6O monolayers. *RSC Advances*, *3*(46), 24337–24344. https://doi.org/10.1039/c3ra44949a

Perdew, J. P., Burke, K., & Ernzerhof, M. (1996). Generalized Gradient Approximation Made Simple. *Physical Review Letters*, *77*(18), 3865. https://doi.org/10.1103/physrevlett.77.3865

Pierucci, D., Henck, H., Avila, J., Balan, A., Naylor, C. H., Patriarche, G., … Ouerghi, A. (2016). Band alignment and minigaps in monolayer MoS2-graphene van der Waals heterostructures. *Nano Letters*, *16*(7), 4054–4061. https://doi.org/10.1021/acs.nanolett.6b00609

Platt, J. C. (1999). Probablistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers* (Vol. 10, pp. 61–74).

Qian, C., Sun, K., & Bao, W. (2022). Recent advance on machine learning of MXenes for energy storage and conversion. *International Journal of Energy Research*, *46*(15), 21511–21522. https://doi.org/10.1002/er.7833

Sbailò, L., Fekete, Á., & Ghiringhelli, L. M. (2022). The NOMAD Artificial-Intelligence Toolkit: turning materials-science data into knowledge and understanding. *Npj Computational Materials*, *8*(250), 1–7. https://doi.org/10.1038/s41524-022-00935-z

Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M., & Fazzio, A. (2019). From DFT to machine learning: Recent approaches to materials science - A review. *JPhys Materials*, *2*(3). https://doi.org/10.1088/2515-7639/ab084b

Shifa, T. A., Wang, F., Liu, Y., & He, J. (2019). Heterostructures Based on 2D Materials: A Versatile Platform for Efficient Catalysis. *Advanced Materials*, *31*(45). https://doi.org/10.1002/adma.201804828

Sholl, D. S. ;, & Steckel, J. A. (2009). *Density Functional Theory: A Practical Introduction*. https://doi.org/10.1002/9780470447710

Sinha, P., Jyothirmai, M. V., Abraham, B. M., & Singh, J. K. (2024). Machine learning driven advancements in catalysis for predicting hydrogen evolution reaction activity. *Neural Networks*, 1–16. https://doi.org/10.1016/j.matchemphys.2024.129805

Song, B., & Jin, S. (2017). Two Are Better than One: Heterostructures Improve Hydrogen Evolution Catalysis. *Joule*, *1*(2), 220–221. https://doi.org/10.1016/j.joule.2017.09.012

Song, L., Song, M., Lu, Z., Yu, G., Liang, Z., Hou, W., … Song, Y. (2022). Recent Advances of Preparation and Application of Two-Dimension van der Waals Heterostructure. *Coatings*, *12*(8). https://doi.org/10.3390/coatings12081152

Sorkun, M. C., Astruc, S., Koelman, J. M. V. A., & Er, S. (2020). An artificial intelligence-aided virtual screening recipe for two-dimensional materials discovery. *Npj Computational Materials*, *6*(1), 1–10. https://doi.org/10.1038/s41524-020-00375-7

Sun, Y., Cao, J., Li, Q., Li, D., & Ao, Z. (2023). Identifying key factors of peroxymonosulfate activation on single-atom M–N–C catalysts: a combined density functional theory and machine learning study. *Journal of Materials Chemistry A*. https://doi.org/10.1039/d3ta02371k

Talirz, L., Kumbhar, S., Passaro, E., Yakutovich, A. V, Granata, V., Gargiulo, F., … Marzari, N. (2020). Materials Cloud, a platform for open computational science. *Scientific Data*, *7*(299), 1–12. https://doi.org/10.1038/s41597-020-00637-5

Tan, A. Y. S., Awan, H. T. A., Cheng, F., Zhang, M., Tan, M. T. T., Manickam, S., … Muthoosamy, K. (2024). Recent advances in the use of MXenes for photoelectrochemical sensors. *Chemical Engineering Journal*, *482*(January), 148774. https://doi.org/10.1016/j.cej.2024.148774

Tang, Q., & Jiang, D. E. (2015). Stabilization and band-gap tuning of the 1T-MoS2 monolayer by covalent functionalization. *Chemistry of Materials*, *27*(10), 3743–3748. https://doi.org/10.1021/acs.chemmater.5b00986

Tawfik, S. A., Isayev, O., Stampfl, C., Shapter, J., Winkler, D. A., & Ford, M. J. (2019). Efficient Prediction of Structural and Electronic Properties of Hybrid 2D Materials Using Complementary DFT and Machine Learning Approaches. *Advanced Theory and Simulations*, *2*(1). https://doi.org/10.1002/adts.201800128

Wan, X., Zhang, Z., Yu, W., & Guo, Y. (2021). A density-functional-theory-based and machine-learning-accelerated hybrid method for intricate system catalysis. *Materials Reports: Energy*, *1*(3), 100046. https://doi.org/10.1016/j.matre.2021.100046

Wang, T., Zhang, C., Snoussi, H., & Zhang, G. (2020). Machine Learning Approaches for Thermoelectric Materials Research. *Advanced Functional Materials*, *30*(5), 1–14. https://doi.org/10.1002/adfm.201906041

Wang, Y., Ding, Z., Arif, N., Jiang, W. C., & Zeng, Y. J. (2022). 2D material based heterostructures for solar light driven photocatalytic H2 production. *Materials Advances*, 3389–3417. https://doi.org/10.1039/d2ma00191h

Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N. E. R., Bajaj, S., Wang, Q., … Jain, A. (2018). Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, *152*(May), 60–69. https://doi.org/10.1016/j.commatsci.2018.05.018

Wei, J., Chu, X., Sun, X. Y., Xu, K., Deng, H. X., Chen, J., … Lei, M. (2019). Machine learning in materials science. *InfoMat*, *1*(3), 338–358. https://doi.org/10.1002/inf2.12028

Wickramaratne, D., Weston, L., & Van De Walle, C. G. (2018). Monolayer to Bulk Properties of Hexagonal Boron Nitride. *Journal of Physical Chemistry C*, *122*(44), 25524–25529. https://doi.org/10.1021/acs.jpcc.8b09087

Willhelm, D., Wilson, N., Arroyave, R., Qian, X., Cagin, T., Pachter, R., & Qian, X. (2022). Predicting Van der Waals Heterostructures by a Combined Machine Learning and Density Functional Theory Approach. *ACS Applied Materials and Interfaces*, *14*(22), 25907–25919. https://doi.org/10.1021/acsami.2c04403

Wolpert, D. (1992). Stacked Generalization. *Neural Networks*, *5*, 241–259.

Wu, L., Guo, T., & Li, T. (2021). Machine learning-accelerated prediction of overpotential of oxygen evolution reaction of single-atom catalysts. *IScience*, *24*(5), 102398. https://doi.org/10.1016/j.isci.2021.102398

Yagmurcukardes, M., Senger, R. T., Peeters, F. M., & Sahin, H. (2016). Mechanical properties of monolayer GaS and GaSe crystals. *Physical Review B*, *94*(24), 1–7. https://doi.org/10.1103/PhysRevB.94.245407

Yan, L., Zhong, S., Igou, T., Gao, H., Li, J., & Chen, Y. (2022). Development of machine learning models to enhance element-doped g-C3N4 photocatalyst for hydrogen production through splitting water. *International Journal of Hydrogen Energy*, *47*(80), 34075–34089. https://doi.org/10.1016/j.ijhydene.2022.08.013

Ye, X., Zhuang, F., Si, Y., He, J., Xue, Y., Li, H., … Zhang, R. (2023). Direct Z-scheme GaN/WSe2 heterostructure for enhanced photocatalytic water splitting under visible spectrum. *RSC Advances*, *13*(29), 20179–20186. https://doi.org/10.1039/d3ra00928a

Zheng, J., Sun, X., Qiu, C., Yan, Y., Yao, Z., Deng, S., … Wang, J. (2020). High Throughput Screening of Hydrogen Evolution Reaction Catalysts in MXenes Materials. *The Journal of Physical Chemistry C*, *124*(25), 13695–13705.

Zheng, Q., Xie, Y., Lan, Z., Prezhdo, O. V., Saidi, W. A., & Zhao, J. (2018). Phonon-coupled ultrafast interlayer charge oscillation at van der Waals heterostructure interfaces. *Physical Review B*, *97*(20). https://doi.org/10.1103/PhysRevB.97.205417

Zhou, P., Wang, M., Tang, F., Ling, L., Yu, H., & Chen, X. (2024). Machine learning accelerates the screening of efficient metal-oxide catalysts for photocatalytic water splitting. *Materials Research Bulletin*, *179*(June), 112956. https://doi.org/10.1016/j.materresbull.2024.112956

Zhou, Q., Luo, S., Xue, W., & Liao, N. (2023). Accelerated screening of sensitive and selective MoO3-based gas sensing materials by combining first-principles and machine learning approach. *Chemical Engineering Journal*, *475*(July), 146318. https://doi.org/10.1016/j.cej.2023.146318

Zhou, X., Hu, X., Yu, J., Liu, S., Shu, Z., Zhang, Q., … Zhai, T. (2018). 2D Layered Material-Based van der Waals Heterostructures for Optoelectronics. *Advanced Functional Materials*, *28*(14), 1–28. https://doi.org/10.1002/adfm.201706587

Zhu, J., Zou, H., Rosset, S., & Hastie, T. (2009). Multi-class AdaBoost ∗. *Statistics and Its Interface*, *2*, 349–360.

Zhu, W., Perebeinos, V., Freitag, M., & Avouris, P. (2009). Carrier scattering, mobilities, and electrostatic potential in monolayer, bilayer, and trilayer graphene. *Physical Review B - Condensed Matter and Materials Physics*, *80*(23), 1–8. https://doi.org/10.1103/PhysRevB.80.235402

Zhu, Z., Dong, B., Guo, H., Yang, T., & Zhang, Z. (2020). Fundamental band gap and alignment of two-dimensional semiconductors explored by machine learning. *Chinese Physics B*, *29*(4), 1–20. https://doi.org/10.1088/1674-1056/ab75d5

Appendix A: Data Processing and Bilayer Features

Table A.1. List of selected 197 monolayers from C2DB. Monolayer UID is same as provided by the database (version of 30-11-2022).

Monolayer UID	Chemical formula	Lattice constant (Å)	Prototype	Space group	Bandgap (eV)	Work function (eV)
BaBr2-1a59eff92917	BaBr2	4.69	MoS2	P-6m2	4.16	5.27
HfTe2-59c0e014651d	HfTe2	3.91	MoS2	P-6m2	0.13	5.15
CrO2-2433700165bb	CrO2	2.63	MoS2	P-6m2	0.42	7.14
MoTe2-38a53176109a	MoTe2	3.55	MoS2	P-6m2	0.93	4.3
TiBr2-57116f9a9a4e	Br2Ti	3.47	MoS2	P-6m2	0.76	3.5
CaBr2-fbb623b6f288	Br2Ca	4.12	MoS2	P-6m2	4.13	5.49
CaCl2-3ca106221b9b	CaCl2	3.95	MoS2	P-6m2	4.77	5.87
SrI2-6cfaae647808	I2Sr	4.64	MoS2	P-6m2	3.42	4.92
WS2-64090c9845f8	S2W	3.19	MoS2	P-6m2	1.53	4.73
BaCl2-54ec344f88a7	BaCl2	4.53	MoS2	P-6m2	4.75	5.47
CrTe2-c31911a1b3f9	CrTe2	3.47	MoS2	P-6m2	0.45	4.55
HfBr2-84e9162c0c53	Br2Hf	3.5	MoS2	P-6m2	0.73	2.99
PdSe2-0ae696751911	PdSe2	4	MoS2	P-6m2	0.23	3.66
ZrI2-9c024b5a2e89	I2Zr	3.83	MoS2	P-6m2	0.69	3.05
MoS2-b3b4685fb6e1	MoS2	3.18	MoS2	P-6m2	1.58	5.1
PbS2-372c217dd52f	PbS2	4.73	MoS2	P-6m2	1.39	4.67
ZrTe2-f7ad606317e6	Te2Zr	3.92	MoS2	P-6m2	0.26	5.28
TiCl2-95688ba68ca1	Cl2Ti	3.28	MoS2	P-6m2	0.9	3.79
CaI2-066f40f26c53	CaI2	4.39	MoS2	P-6m2	2.95	4.88

PbBr2-cbdc15b42a05	Br2Pb	4.31	MoS2	P-6m2	2.27	5.98
BaI2-c4707a226b8f	BaI2	4.92	MoS2	P-6m2	3.33	4.97
CrS2-c5ee5e35d2b4	CrS2	3.05	MoS2	P-6m2	0.88	5.4
HfCl2-864f8b497185	Cl2Hf	3.35	MoS2	P-6m2	0.91	3.25
PbCl2-b0b142073783	Cl2Pb	4.18	MoS2	P-6m2	2.78	6.32
PbSe2-0bc5d11454a7	PbSe2	4.88	MoS2	P-6m2	1.03	4.46
TiH2-2d25b4df71af	H2Ti	2.79	MoS2	P-6m2	0.02	4.82
WSe2-1cfbe6183886	Se2W	3.32	MoS2	P-6m2	1.24	4.23
SnI2-d9c422656482	I2Sn	4.42	MoS2	P-6m2	1.93	4.91
SrBr2-2876a0cb2478	Br2Sr	4.39	MoS2	P-6m2	4.31	5.36
ZrBr2-7897c7cc2491	Br2Zr	3.56	MoS2	P-6m2	0.83	3.24
TiI2-088e8488f895	I2Ti	3.77	MoS2	P-6m2	0.6	3.33
TiSe2-509ef368050d	Se2Ti	3.5	MoS2	P-6m2	0.52	5.83
TiF2-32023209e1ed	F2Ti	2.85	MoS2	P-6m2	1.29	4.44
HfSe2-d2d9fee03594	HfSe2	3.68	MoS2	P-6m2	0.81	5.86
MoSe2-f61b14d398c7	MoSe2	3.32	MoS2	P-6m2	1.32	4.57
TiTe2-9ebfacbfe4b5	Te2Ti	3.74	MoS2	P-6m2	0.04	5.15
WO2-94cfbb3f9284	O2W	2.83	MoS2	P-6m2	1.34	5.71
HfI2-05a69240794c	HfI2	3.77	MoS2	P-6m2	0.63	2.79
MoO2-152bd69757aa	MoO2	2.82	MoS2	P-6m2	0.91	6.21
CrSe2-9a6ff6a3c41a	CrSe2	3.21	MoS2	P-6m2	0.69	4.88
PbI2-9e6494406d07	I2Pb	4.5	MoS2	P-6m2	1.58	5.45
SrCl2-77398c835c11	Cl2Sr	4.22	MoS2	P-6m2	4.94	5.61
WTe2-3c87365bc48c	Te2W	3.55	MoS2	P-6m2	0.73	4.05
ZrCl2-dc09b7c396eb	Cl2Zr	3.41	MoS2	P-6m2	0.98	3.52
CdF2-c43f24bb6516	CdF2	3.51	CdI2	P-3m1	3.81	6.86

PdS2-02762a09ebbf	PdS2	3.55	CdI2	P-3m1	1.17	5.71
SnS2-42a44e4e7298	S2Sn	3.7	CdI2	P-3m1	1.59	6.03
BaBr2-74dc20d8f1ff	BaBr2	4.96	CdI2	P-3m1	4.9	4.78
CaBr2-819000a41305	Br2Ca	4.3	CdI2	P-3m1	4.86	5.02
CdCl2-3861891e1b05	CdCl2	3.9	CdI2	P-3m1	3.87	6.07
HgBr2-2af12bd12520	Br2Hg	4.14	CdI2	P-3m1	1.98	5.92
ZrSe2-7dc2a0a57b42	Se2Zr	3.79	CdI2	P-3m1	0.34	5.29
BaCl2-00a09165a480	BaCl2	4.81	CdI2	P-3m1	5.62	4.84
GeS2-6b1efcdcfd40	GeS2	3.44	CdI2	P-3m1	0.73	6.09
PbO2-e2a7fd3e05fc	O2Pb	3.4	CdI2	P-3m1	1.35	7.42
SnSe2-2c6d4c024ca0	Se2Sn	3.86	CdI2	P-3m1	0.76	5.62
HfO2-a555d8dafaf2	HfO2	3.25	CdI2	P-3m1	4.78	5.48
HgCl2-31f13a321b70	Cl2Hg	3.99	CdI2	P-3m1	2.45	6.4
PtSe2-d000f0288397	PtSe2	3.75	CdI2	P-3m1	1.17	4.88
SrI2-7d9241a9f3b1	I2Sr	4.84	CdI2	P-3m1	3.98	4.54
ZnBr2-88424f85cf2b	Br2Zn	3.8	CdI2	P-3m1	3.35	5.26
CaCl2-643810e6bb76	CaCl2	4.14	CdI2	P-3m1	5.78	5.26
NiO2-c444fe118e1b	NiO2	2.84	CdI2	P-3m1	1.28	7.46
PbBr2-0dc6977a707b	Br2Pb	4.49	CdI2	P-3m1	2	5.97
CdI2-b474b300c0f6	CdI2	4.32	CdI2	P-3m1	2.14	4.96
PdSe2-f7acacb69123	PdSe2	3.73	CdI2	P-3m1	0.56	5.06
ZnCl2-29e6de323923	Cl2Zn	3.6	CdI2	P-3m1	4.45	5.63
BaI2-080cc67ae6c9	BaI2	5.19	CdI2	P-3m1	3.87	4.6
HfS2-d48ddb34811e	HfS2	3.65	CdI2	P-3m1	1.22	5.73

MgBr2-702277c8c7ed	Br2Mg	3.88	CdI2	P-3m1	4.59	4.94
PbS2-b874a236aaca	PbS2	3.85	CdI2	P-3m1	0.76	6.39
PtTe2-747f8e0087c0	PtTe2	4.02	CdI2	P-3m1	0.3	4.13
ZnF2-eaf518aac509	F2Zn	3.14	CdI2	P-3m1	4.51	6.77
CaI2-ee886d522d75	CaI2	4.55	CdI2	P-3m1	3.54	4.59
MgCl2-85b63b2e4472	Cl2Mg	3.67	CdI2	P-3m1	5.94	5.28
PbCl2-ec34fbc566ef	Cl2Pb	4.39	CdI2	P-3m1	2.37	6.15
RuBr2-88699c6d39e3	Br2Ru	3.78	CdI2	P-3m1	1.05	4.19
SnI2-bf884c9076a0	I2Sn	4.58	CdI2	P-3m1	1.84	4.99
ZnI2-69f5f8718ba5	I2Zn	4.1	CdI2	P-3m1	1.75	4.95
ZrO2-f09f2879d995	O2Zr	3.27	CdI2	P-3m1	4.45	5.6
HgI2-efa4b9d311d3	HgI2	4.39	CdI2	P-3m1	1.2	5.27
NiS2-c07043374f26	NiS2	3.35	CdI2	P-3m1	0.58	5.5
PtO2-6f43cd15097d	O2Pt	3.15	CdI2	P-3m1	1.67	6.03
GeI2-24970c2cdc0f	GeI2	4.3	CdI2	P-3m1	1.95	4.88
HfSe2-7a708c5759cf	HfSe2	3.77	CdI2	P-3m1	0.43	5.16
MgI2-440644551de5	I2Mg	4.21	CdI2	P-3m1	3.27	4.44
PbSe2-d0003c75d651	PbSe2	4	CdI2	P-3m1	0.08	5.88
RuCl2-6be75763d9c1	Cl2Ru	3.59	CdI2	P-3m1	1.18	4.36
SrBr2-a17f01139eff	Br2Sr	4.61	CdI2	P-3m1	4.93	4.95
SnO2-9c24e87a3dde	O2Sn	3.22	CdI2	P-3m1	2.68	6.99
ZrS2-8913f4e54692	S2Zr	3.68	CdI2	P-3m1	1.16	5.85
HgF2-24d1e0feea03	F2Hg	3.66	CdI2	P-3m1	1.73	7.12
CdBr2-5acf0d634aa9	Br2Cd	4.06	CdI2	P-3m1	3.07	5.59
PbI2-82db29775962	I2Pb	4.66	CdI2	P-3m1	1.5	5.55
PdO2-de604d79f9c3	O2Pd	3.09	CdI2	P-3m1	1.38	6.83
SiS2-c70107ea4985	S2Si	3.3	CdI2	P-3m1	1.38	5.78
SrCl2-67727fec854b	Cl2Sr	4.46	CdI2	P-3m1	5.74	5.07
GeO2-630e9718ed1f	GeO2	2.9	CdI2	P-3m1	3.64	7.05

NiSe2-c1ffa88ee5ca	NiSe2	3.54	CdI2	P-3m1	0.06	4.85
PtS2-77caffd1d3ed	PtS2	3.57	CdI2	P-3m1	1.69	5.4
SiSe2-0dae128e0ffe	Se2Si	3.51	CdI2	P-3m1	0.47	5.41
Al2S2-f9df9f4a5c34	Al2S2	3.58	GaS	P-6m2	2.09	5.16
S2Tl2-751e767bff79	S2Tl2	4.06	GaS	P-6m2	0.67	5.97
Te2Zr2-8912432cb37b	Te2Zr2	3.75	GaS	P-6m2	0.45	4.61
Bi2O2-53ac438f321b	Bi2O2	3.98	GaS	P-6m2	0.12	4.28
Al2Se2-129a514b51ad	Al2Se2	3.78	GaS	P-6m2	2	4.83
Ga2Te2-55c23ca88a05	Ga2Te2	4.13	GaS	P-6m2	1.3	4.52
BN-4a5edc763604	BN	2.51	BN	P-6m2	4.67	3.49
BP-0a5e44762c75	BP	3.22	BN	P-6m2	0.9	4.68
In2Se2-eb204c739879	In2Se2	4.07	GaS	P-6m2	1.4	5.21
GaN-c973e283b023	GaN	3.27	BN	P-6m2	1.88	4.38
Al2Te2-e54041554385	Al2Te2	4.12	GaS	P-6m2	1.75	4.4
Ga2O2-16c96094d1a0	Ga2O2	3.13	GaS	P-6m2	1.56	6.34
Se2Tl2-625697b299d1	Se2Tl2	4.22	GaS	P-6m2	0.49	5.59
In2Te2-fcd97ff5abcd	In2Te2	4.37	GaS	P-6m2	1.23	4.65
Te2Ti2-b43c14735d8e	Te2Ti2	3.61	GaS	P-6m2	0.22	4.61
AlN-4fcd8dcb9c90	AlN	3.13	BN	P-6m2	2.89	3.88
In2O2-d14171d2ba1a	In2O2	3.44	GaS	P-6m2	0.39	6.25
AuI-a18bfea44dc0	AuI	4.79	BN	P-6m2	0.05	5.25
Al2O2-bce0ccee4eca	Al2O2	2.98	GaS	P-6m2	1.32	4.35
Se2Zr2-f89b20d72c95	Se2Zr2	3.6	GaS	P-6m2	0.06	4.81

Ga2S2-ac002f4ce724	Ga2S2	3.65	GaS	P-6m2	2.32	5.11
Ga2Se2-394e5709a3ac	Ga2Se2	3.82	GaS	P-6m2	1.76	4.91
In2S2-172ef584c4a6	In2S2	3.91	GaS	P-6m2	1.68	5.51
Te2Tl2-73117163f0e2	Te2Tl2	4.51	GaS	P-6m2	0.34	5.01
Al2S2-dce599ea3912	Al2S2	3.6	GaSe	P-3m1	2.14	5.09
Br2Sb2-1b03587c5de9	Br2Sb2	5.24	GaSe	P-3m1	0.44	4.84
In2S2-1b7899449ed6	In2S2	3.92	GaSe	P-3m1	1.6	5.52
S2Ti2-89ee439948ac	S2Ti2	3.92	GaSe	P-3m1	0.05	4.12
Te2Zr2-6f8a67e0a912	Te2Zr2	3.73	GaSe	P-3m1	0.21	4.62
I2Pt2-1e64e8d4982d	I2Pt2	3.96	GaSe	P-3m1	0.36	4.05
Cl2Pt2-93dfef2d1004	Cl2Pt2	3.63	GaSe	P-3m1	1.31	5.07
Ga2Se2-8a6a5075fdf0	Ga2Se2	3.84	GaSe	P-3m1	1.64	4.95
Ir2S2-bf9a5b4e8e88	Ir2S2	3.28	GaSe	P-3m1	0.04	4.82
S2Zn2-fdc73006aad8	S2Zn2	3.92	GaSe	P-3m1	2.56	5.02
Cl2Ge2-db2546129ff7	Cl2Ge2	4.23	GaSe	P-3m1	0.12	6.25
Cu2I2-ac7333c6ab94	Cu2I2	4.18	GaSe	P-3m1	1.79	3.65
Ga2Te2-1b0b43b113f5	Ga2Te2	4.15	GaSe	P-3m1	1.27	4.53
F2Ge2-aac9cc861c7f	F2Ge2	4.25	GaSe	P-3m1	0.12	7.13
F2Si2-aa042bf2fe06	F2Si2	3.94	GaSe	P-3m1	0.67	6.18
Hf2Se2-a2f6b7b424d2	Hf2Se2	4.27	GaSe	P-3m1	0.04	3.6

In2Se2-178396d52e5c	In2Se2	4.08	GaSe	P-3m1	1.32	5.23
Br2Cd2-2b24ece7d0b2	Br2Cd2	4.1	GaSe	P-3m1	1.84	5.29
Cd2Cl2-0089607de839	Cd2Cl2	3.93	GaSe	P-3m1	1.67	5.47
F2Pt2-bad21013bab6	F2Pt2	3.39	GaSe	P-3m1	0.81	6.11
Al2Se2-136900211245	Al2Se2	3.79	GaSe	P-3m1	2.14	4.72
F2Sn2-327ecc5b4be0	F2Sn2	4.96	GaSe	P-3m1	0.32	6.87
Ir2Se2-a4d54cf399c5	Ir2Se2	3.41	GaSe	P-3m1	0.07	4.38
S2Tl2-fb0b06c4bab6	S2Tl2	4.06	GaSe	P-3m1	0.63	5.99
F2Pd2-28a66c9b9a33	F2Pd2	3.29	GaSe	P-3m1	0.86	7.13
I2Zn2-748053f4b718	I2Zn2	4.19	GaSe	P-3m1	2.09	4.45
Cd2I2-369d58ce1d9f	Cd2I2	4.4	GaSe	P-3m1	1.94	4.78
Al2Te2-1584210aaef9	Al2Te2	4.13	GaSe	P-3m1	1.96	4.27
I2In2-9ab97b2e26fc	I2In2	4.48	GaSe	P-3m1	0.99	4.94
S2Zr2-8e309f5eccc0	S2Zr2	4.2	GaSe	P-3m1	0.13	3.54
Se2Ti2-f346fce11db3	Se2Ti2	4.04	GaSe	P-3m1	0.04	4.18
Br2Hg2-915289d04444	Br2Hg2	4.18	GaSe	P-3m1	1.52	6.03
Ag2Br2-9d25bdce5cb7	Ag2Br2	4.43	GaSe	P-3m1	1.72	5.37
Br2Cu2-1d4af93f2f8f	Br2Cu2	3.95	GaSe	P-3m1	1.52	4.5
C2F2-02ae369021e9	C2F2	2.6	GaSe	P-3m1	3.17	6.1
In2Te2-3cf9ea4cebc4	In2Te2	4.38	GaSe	P-3m1	1.15	4.68

Sb2Se2-a59fe0d07081	Sb2Se2	4.82	GaSe	P-3m1	0.44	4.53
Se2Tl2-16490f81af26	Se2Tl2	4.23	GaSe	P-3m1	0.45	5.62
Se2Zn2-4349b5bd2e79	Se2Zn2	4.08	GaSe	P-3m1	1.62	4.92
Cl2Pd2-5ae6f8403434	Cl2Pd2	3.67	GaSe	P-3m1	0.16	5.4
Bi2Se2-d13cbee86f79	Bi2Se2	4.92	GaSe	P-3m1	0.35	4.69
Cd2S2-d9ca01c61c76	Cd2S2	4.32	GaSe	P-3m1	1.79	5.22
C2H2-1b265530a869	C2H2	2.54	GaSe	P-3m1	3.46	3
Br2Ge2-282b13289e74	Br2Ge2	4.25	GaSe	P-3m1	0.17	5.76
Ag2Cl2-dd5f0964d63d	Ag2Cl2	4.4	GaSe	P-3m1	1.57	5.04
Cd2Se2-6e69c77cf8fc	Cd2Se2	4.45	GaSe	P-3m1	1.4	5.14
Te2Ti2-ecf4b958e9a1	Te2Ti2	3.58	GaSe	P-3m1	0.11	4.6
Cl2Sb2-a6ccabb6ac8a	Cl2Sb2	5.22	GaSe	P-3m1	0.44	5.06
Se2Zr2-a1b04a4b1428	Se2Zr2	4.33	GaSe	P-3m1	0.02	3.78
Te2Zn2-4e0f8a77aaf0	Te2Zn2	4.33	GaSe	P-3m1	0.54	4.52
Br2Pt2-046bb8e1f5ad	Br2Pt2	3.76	GaSe	P-3m1	1.18	4.71
Cd2Te2-3a9e0525e324	Cd2Te2	4.67	GaSe	P-3m1	0.63	4.78
Ga2S2-6f3b2e422ac9	Ga2S2	3.66	GaSe	P-3m1	2.18	5.14

Hg2I2-f7e70d2b90ad	Hg2I2	4.46	GaSe	P-3m1	1.28	5.27
Ag2I2-0c52f79478e2	Ag2I2	4.57	GaSe	P-3m1	1.77	4.53
Pb2Te2-8541bd4f0edd	Pb2Te2	5.52	GaSe	P-3m1	0.43	4.12
Te2Tl2-44ffc9d405b8	Te2Tl2	4.52	GaSe	P-3m1	0.29	5.05
Hf2S2-e7d34362869c	Hf2S2	4.15	GaSe	P-3m1	0.13	3.36
GeSe-211bcb7f05d6	GeSe	3.67	GeSe	P3m1	2.22	4.65
OPb-2a393480e273	OPb	3.59	GeSe	P3m1	1.8	4.97
HgSe-619ed885f677	HgSe	4.46	GeSe	P3m1	0.12	5.43
OSn-026ebfd86b48	OSn	3.34	GeSe	P3m1	1.71	5.07
PbTe-3bc08d486d65	PbTe	4.32	GeSe	P3m1	0.93	4.17
SnTe-e688959ea45b	SnTe	4.17	GeSe	P3m1	1.58	4.18
GeTe-eadd37f03ca5	GeTe	3.94	GeSe	P3m1	1.47	4.22
HgTe-1a3bdd1b142a	HgTe	4.73	GeSe	P3m1	0.15	5.03
SSn-f98da23471a1	SSn	3.75	GeSe	P3m1	2.29	4.94
PbS-5e4ff1f56b4a	PbS	3.95	GeSe	P3m1	1.82	4.83
GeO-a42f736f1682	GeO	3.01	GeSe	P3m1	2.13	5.49
PbSe-a0dbdc6630fa	PbSe	4.08	GeSe	P3m1	1.53	4.62
SeSn-d59c96fdfda1	SeSn	3.91	GeSe	P3m1	2.18	4.65
GeS-227b12019ade	GeS	3.49	GeSe	P3m1	2.45	4.94

Table A.2. List of 97 selected bilayer descriptors for set of vdW bilayers.

Bilayer Descriptor	Description	mean	std	min	max
avg_hform	Heat of formation	-0.792388156	0.447033268	-2.837496801	0.006623155
avg_evac	Vacuum energy level in eV	3.405025923	0.49384742	1.546130228	4.813278232
avg_efermi	Fermi energy level in eV	-1.65158858	0.812084302	-4.465950736	0.843573311
avg_gap	Bandgap in eV	1.788054614	0.996738722	0.075944654	4.641257784
avg_vbm	VBM energy by PBE functional	-2.545615887	1.123111208	-5.754269491	0.558760383
avg_cbm	CBM energy by PBE functional	-0.757561274	0.744531452	-3.543100258	1.203303177
avg_gap_dir	PBE direct bandgap in eV	1.991078473	0.970037449	0.078255023	4.808550569
avg_gap_dir_nosoc	PBE direct bandgap in eV without SOC	2.200837724	0.880377327	0.361018115	5.009862045
avg_gap_nosoc	PBE bandgap in eV without SOC	1.949897715	0.949198357	0.25249283	4.729466855
avg_workfunction	Work function in eV	5.056614503	0.574764767	3.243910798	7.138246654
avg_vbm_hse	VBM energy by HSE06 functional	-2.91499188	1.306166729	-5.999360874	0.623145747
avg_cbm_hse	CBM energy by HSE06 functional	-0.355612536	0.726458634	-2.983820246	1.554009622
avg_gap_dir_hse	HSE06 direct bandgap in eV	2.767393509	1.165705113	0.080698612	6.040127806
avg_gap_hse	HSE06 bandgap in eV	2.559379343	1.196029973	0.080698612	5.820847474
avg_c_11	Stiffness Tensor 11-component	57.98026758	35.72552946	-2.514967515	270.3055822
avg_c_12	Stiffness Tensor 12-component	19.64765498	12.27938234	3.001278616	66.3202841
avg_E_B	Exciton binding energy in eV	1.061335159	0.330484912	0.347858655	2.225000532
avg_emass_vb_dir1	Effective mass in direction 1	-10.79038741	82.04658067	-1662.837104	-0.085763008
avg_emass_vb_dir2	Effective mass in direction 2	-1.487395598	2.358645664	-26.58056625	6.677671864
avg_emass_cb_dir1	Effective mass in direction 1	0.350220736	0.515546957	-2.83223047	3.071287146
avg_emass_cb_dir2	Effective mass in direction 2	0.823694603	1.351696823	-1.252579894	12.66013929
avg_alphax	Static polarizability in x direction	254.6463151	814.6258659	1.169124935	6359.611076
avg_alphaz	Static polarizability in z direction	0.393503013	0.264159383	0	4.77017121
avg_cell_area	Area of primitive cell	13.09038362	2.766241677	5.519881632	21.13074958
avg_cbm_hybridization	Amount of hybridization at the CBM	1.329583445	0.315253116	0.301495068	2.061669974
avg_cbm_score	Fraction of orbital contributing to CBM	0.486611867	0.152366514	0.21978891	0.92939494
avg_vbm_hybridization	Amount of hybridization at the VBM	1.130460957	0.275959724	0.046175455	1.815667734

avg_vbm_score	Fraction of orbital contributing to VBM	0.475042398	0.133028776	0.262724934	0.990723765
avg_cbm_s	Fractional contribution of s orbital to CBM	0.299483366	0.21488676	0.003479536	0.901373709
avg_cbm_p	Fractional contribution of p orbital to CBM	0.387191004	0.181703541	0.056766886	0.918497662
avg_cbm_d	Fractional contribution of d orbital to CBM	0.31332563	0.26479657	0	0.92939494
avg_cbm_sp	Fractional contribution of sp hybridized orbital to CBM	0.420179594	0.271270687	0	0.957328572
avg_cbm_sd	Fractional contribution of sd hybridized orbital to CBM	0.067611439	0.050840812	0	0.452543957
avg_cbm_pd	Fractional contribution of pd hybridized orbital to CBM	0.260413159	0.196140352	0	0.808045978
avg_vbm_s	Fractional contribution of s orbital to VBM	0.054985558	0.064910695	0	0.328142458
avg_vbm_p	Fractional contribution of p orbital to VBM	0.697596746	0.22136243	0.055200692	1
avg_vbm_d	Fractional contribution of d orbital to VBM	0.247417695	0.235270714	0	0.942132821
avg_vbm_sp	Fractional contribution of sp hybridized orbital to VBM	0.157079333	0.175694418	0	0.790251678
avg_vbm_sd	Fractional contribution of sd hybridized orbital to VBM	0.015801173	0.024135148	0	0.216808275
avg_vbm_pd	Fractional contribution of pd hybridized orbital to VBM	0.296661632	0.198541813	0	0.941632613
avg_lattice_param	Mean of monolayer lattice constants	3.864606329	0.414977804	2.52459994	4.939554195
avg_PymatgenData minimum X	Minimum of Composition-based electronegativity	1.777718447	0.249563635	0.92	2.345
avg_PymatgenData maximum X	Maximum of Composition-based electronegativity	2.802524272	0.317372979	2.1	3.71
avg_PymatgenData range X	Range of Composition-based electronegativity	1.024805825	0.37460615	0.11	2.235
avg_PymatgenData mean X	Mean of Composition-based electronegativity	2.383667071	0.23644113	1.8975	3.143333333
avg_PymatgenData std_dev X	Standard deviation of Composition-based electronegativity	0.724647148	0.264886549	0.077781746	1.580383656
avg_PymatgenData minimum atomic_mass	Minimum of Composition-based atomic mass	54.6349934	25.35965636	5.90947	127.6
avg_PymatgenData maximum atomic_mass	Maximum of Composition-based atomic mass	124.6353627	38.83626808	13.0087	207.2
avg_PymatgenData range atomic_mass	Range of Composition-based atomic mass	70.00036934	37.87312253	5.0834614	183.6430984
avg_PymatgenData mean atomic_mass	Mean of Composition-based atomic mass	85.80241235	26.65821169	9.459085	163.0152597

avg_PymatgenData std_dev atomic_mass	Standard deviation of Composition-based atomic mass	49.49773585	26.78034177	3.594550028	129.8552802
avg_PymatgenData minimum atomic_radius	Minimum of Composition-based atomic radius	1.067900485	0.20326477	0.425	1.4
avg_PymatgenData maximum atomic_radius	Maximum of Composition-based atomic radius	1.489381068	0.156191497	0.775	2.075
avg_PymatgenData range atomic_radius	Range of Composition-based atomic radius	0.421480583	0.215297829	0.025	1.125
avg_PymatgenData mean atomic_radius	Mean of Composition-based atomic radius	1.242738673	0.154545649	0.6125	1.5625
avg_PymatgenData std_dev atomic_radius	Standard deviation of Composition-based atomic radius	0.298031778	0.152238555	0.01767767	0.795495129
avg_PymatgenData minimum thermal_conductivity	Minimum of Composition-based thermal conductivity	0.496406444	0.615756096	0.0089	3
avg_PymatgenData maximum thermal_conductivity	Maximum of Composition-based thermal conductivity	89.99381068	61.80309152	8.15	430
avg_PymatgenData range thermal_conductivity	Range of Composition-based thermal conductivity	89.49740424	61.80092551	6.39	429.77105
avg_PymatgenData mean thermal_conductivity	Mean of Composition-based thermal conductivity	38.99438969	30.27320429	2.809633333	215.114475
avg_PymatgenData std_dev thermal_conductivity	Standard deviation of Composition-based thermal conductivity	63.28422143	43.69985351	4.518412332	303.8940238
avg_PymatgenData minimum melting_point	Minimum of Composition-based melting point	329.8661772	132.2441236	34.405	722.66
avg_PymatgenData maximum melting_point	Maximum of Composition-based melting point	1258.379053	617.8875543	250.06	3295.5
avg_PymatgenData range melting_point	Range of Composition-based melting point	928.5128762	624.0503497	26.235	3035.97
avg_PymatgenData mean melting_point	Mean of Composition-based melting point	701.1830542	257.4870623	191.8741667	1566.395
avg_PymatgenData std_dev melting_point	Standard deviation of Composition-based melting point	656.5577512	441.2702341	18.5509464	2146.754974
avg_cbm_spd_card	Major orbital contributor in CBM as cardinal descriptor (s=0, p=1, d=2)	1.003640777	0.632156428	0	2
avg_vbm_spd_card	Major orbital contributor in VBM as cardinal descriptor (s=0, p=1, d=2)	1.247572816	0.309019219	1	2
avg_vbmsite_vdw_radius	Mean vdW radius of the species contributing to VBM	1.923106796	0.139915019	1.495	2.23
avg_cbmsite_vdw_radius	Mean vdW radius of the species contributing to CBM	2.08434466	0.130625699	1.66	2.585

avg_cbm_mend_no	Mendeleev number of the species contributing to CBM	45.6723301	15.45763889	5.5	82
avg_cbmsite_mass	Mean atomic mass of the species contributing to CBM	108.5998499	40.56275415	11.41085	207.2
avg_cbmsite_en	Mean electronegativity of the species contributing to CBM	1.881662621	0.340816916	0.92	3.01
avg_cbmsite_nvalence	Number of valence electrons of the species contributing to CBM	3.406553398	1.115203827	0	6
avg_cbmsite_mp	Melting point of the species contributing to CBM	1190.066529	651.1204472	178.855	3295.5
avg_cbmsite_atomic_vol	Atomic volume of the species contributing to CBM	13.40310064	5.009893565	1.861392075	37.57001301
avg_cbmsite_atomic_radius	Mean atomic radius of the species contributing to CBM	1.433616505	0.195046076	0.75	2.075
avg_vbm_mend_no	Mendeleev number of the species contributing to VBM	35.32281553	14.43266651	6.5	77.5
avg_vbmsite_mass	Mean atomic mass of the species contributing to VBM	82.36237218	37.03902312	13.0087	193.6505
avg_vbmsite_en	Mean electronegativity of the species contributing to VBM	2.485643204	0.398347574	1.33	3.71
avg_vbmsite_nvalence	Number of valence electrons of the species contributing to VBM	5.040048544	1.881908489	-2	7
avg_vbmsite_mp	Melting point of the species contributing to VBM	840.1643204	610.7541254	54.165	3295.5
avg_vbmsite_atomic_vol	Atomic volume of the species contributing to VBM	8.041820986	3.218305658	0.714188127	16.37789407
avg_vbmsite_atomic_radius	Mean atomic radius of the species contributing to VBM	1.189320388	0.198943436	0.55	1.575
avg_vbm_character_d	Major d-orbital character for VBM, encoded by One Hot Encoder as Boolean	0.264563107	0.318384843	0	1
avg_vbm_character_p	Major p-orbital character for VBM, encoded by One Hot Encoder as Boolean	0.726941748	0.317881317	0	1
avg_vbm_character_s	Major s-orbital character for VBM, encoded by One Hot Encoder as Boolean	0.008495146	0.064695938	0	0.5
avg_cbm_character_d	Major d-orbital character for CBM, encoded by One Hot Encoder as Boolean	0.372572816	0.354189637	0	1
avg_cbm_character_p	Major p-orbital character for CBM, encoded by One Hot Encoder as Boolean	0.252427184	0.316794978	0	1
avg_cbm_character_s	Major s-orbital character for CBM, encoded by One Hot Encoder as Boolean	0.375	0.351612552	0	1

min_cbm	The lower CBM energy among the CBMs of constituent monolayers	-1.383624564	1.015989896	-4.409745704	1.045044663
max_cbm	The higher CBM energy among the CBMs of constituent monolayers	-0.131497984	0.818863996	-2.884205782	1.843731732
min_vbm	The lower VBM energy among the VBMs of constituent monolayers	-3.469841111	1.316211713	-6.824461818	0.30812823
max_vbm	The higher VBM energy among the VBMs of constituent monolayers	-1.621390664	1.326899292	-4.92371398	1.214032039
anderson_gap	Bandgap by Anderson rule, i.e., min_cbm - max_vbm	0.237766101	1.539518932	-4.679126682	3.460110743
wf_diff	Difference of work functions of constituent monolayers	1.029136818	0.871729535	0.010483479	4.429202609
mismatch	Lattice mismatch determined by relation $\Delta a = \frac{\lvert a_1 - a_2 \rvert}{a_1}$ for $a_1 < a_2$	1.94604786	1.169593814	0.019746142	3.999309037

Appendix B: Labeled vdW Bilayers

Table B.1. Formula, mismatch (%), bilayer prototype, binding energy (E_b) in meVÅ^{-2}, interlayer distance (d_o) in Å, bandgap (E_g), ionization energy (IE), electron affinity (EA), work function (Φ), energies of conduction band minimum (E_{CBM}) and valence band maximum (E_{VBM}) in eV for 47 DFT-predicted direct bandgap vdW bilayer heterostructures.

Formula	Mismatch (%)	Bilayer Prototype	E_b	d_o	Eg	IE	EA	Φ	E_{CBM}	E_{VBM}
			meVÅ^{-2}	Å	eV	eV	eV	eV	eV	eV
PbS2-PbSe2	3.13	MoS2-MoS2	17.35	3	1.61	5.16	3.55	4.95	-3.55	-5.16
Br2Zr-Te2W	0.34	MoS2-MoS2	292.74	3.98	0.14	3.77	3.63	3.7	-3.62	-3.78
Br2Ca-Br2Hg	3.68	CdI2-CdI2	0.62	3.97	1.99	7.1	5.11	6.88	-5.11	-7.1
Br2Cd-Br2Hg	1.98	CdI2-CdI2	3.44	3.93	2.03	7	4.97	6.78	-4.97	-7
Br2Hg-I2Mg	1.54	CdI2-CdI2	8.53	4.04	1.53	6.53	5	6.32	-5	-6.53
Cd2Cl2-S2Zn2	0.24	GaSe-GaSe	58.75	3.7	1.69	6.32	4.62	6.11	-4.62	-6.32
S2Zn2-Se2Zn2	3.82	GaSe-GaSe	77.92	3.56	1.84	5.91	4.07	5.69	-4.07	-5.91
Al2S2-Cl2Pt2	0.92	GaSe-GaSe	94.61	3.52	1.49	5.9	4.41	5.7	-4.41	-5.9
In2S2-Se2Zn2	3.9	GaSe-GaSe	47.85	3.8	0.87	5.58	4.71	5.37	-4.71	-5.58
Ag2I2-Cd2I2	3.88	GaSe-GaSe	55.21	4.11	1.59	5.68	4.09	5.47	-4.09	-5.68
Cd2Te2-Sb2Se2	3.11	GaSe-GaSe	19.62	3.94	0.06	4.76	4.7	4.73	-4.7	-4.76

Ag2Cl2-Cd2Se2	1.18	GaSe-GaSe	61.36	2.83	0.58	5.82	5.25	5.61	-5.25	-5.83
Ag2Cl2-Ag2I2	3.89	GaSe-GaSe	68.49	3.16	1.41	5.92	4.51	5.72	-4.51	-5.92
BP-O2Zr	1.63	CdI2-BN	97.78	3.42	0.88	5.17	4.29	4.93	-4.29	-5.17
Cl2Ru-GeS	2.96	GeSe-CdI2	40.35	3.67	1.19	5.66	4.47	5.43	-4.47	-5.67
PbTe-Se2Tl2	2.37	GeSe-GaS	19.22	3.81	0.1	5.47	5.38	5.42	-5.38	-5.48
Br2Cd2-SnTe	1.66	GeSe-GaSe	26.3	3.95	0.71	5.41	4.7	5.2	-4.7	-5.41
AlN-CrS2	2.82	MoS2-BN	107.39	2.78	0.4	5.72	5.32	5.52	-5.31	-5.72
AlN-S2W	1.63	MoS2-BN	224.41	2.96	1.49	5.44	3.95	5.22	-3.95	-5.44
CrSe2-HfO2	1.37	MoS2-CdI2	310.2	3.48	0.73	5.23	4.5	4.99	-4.47	-5.2
MoSe2-O2Zr	1.55	MoS2-CdI2	114.36	3.39	1.52	5.32	3.79	5.11	-3.8	-5.32
Cl2Mg-Te2W	3.4	MoS2-CdI2	172.75	3.92	0.91	4.76	3.85	4.53	-3.85	-4.76
Ag2I2-BaBr2	2.53	MoS2-GaSe	50.22	4.02	2.07	5.75	3.68	5.55	-3.69	-5.76
CaCl2-Se2Zn2	3.23	MoS2-GaSe	54.61	3.9	1.92	5.76	3.85	5.55	-3.85	-5.76
Ag2Br2-BaCl2	2.22	MoS2-GaSe	46.78	3.58	1.69	5.99	4.3	5.77	-4.3	-5.99
Ag2I2-BaCl2	0.91	MoS2-GaSe	49.38	3.94	2.06	5.75	3.69	5.53	-3.69	-5.75
Br2Sr-Cd2S2	1.69	MoS2-GaSe	37.01	3.7	1.78	6.08	4.3	5.85	-4.3	-6.08
Ag2Cl2-Br2Sr	0.26	MoS2-GaSe	50.95	3.38	1.68	5.88	4.2	5.65	-4.2	-5.88
Cd2S2-Cl2Sr	2.13	MoS2-GaSe	34.03	3.55	1.88	6.11	4.23	5.89	-4.23	-6.11
Ag2I2-PbS2	3.27	MoS2-GaSe	43.42	3.81	1.65	5.48	3.83	5.27	-3.83	-5.48
Ag2Br2-CaI2	1.1	MoS2-GaSe	56.23	4.04	1.65	6.14	4.48	5.92	-4.49	-6.14

Ag2Cl2-CaI2	0.33	MoS2-GaSe	53.58	3.53	1.68	5.84	4.16	5.62	-4.16	-5.84
Br2Pb-Cd2S2	0.18	MoS2-GaSe	31.92	3.6	1.77	6.12	4.35	5.9	-4.35	-6.12
Ag2Br2-I2Pb	1.53	MoS2-GaSe	50.62	3.89	1.7	6.09	4.39	5.88	-4.39	-6.09
Cl2Sr-Cu2I2	1.13	MoS2-GaSe	73.12	3.96	1.68	4.85	3.17	4.64	-3.17	-4.85
Br2Zn-S2Zn2	3.23	GaSe-CdI2	68.69	3.78	2.38	6.27	3.89	6.06	-3.89	-6.27
I2Mg-Se2Zn2	3.19	GaSe-CdI2	59.41	4.01	1.58	5.96	4.38	5.75	-4.38	-5.96
Cd2Se2-CdI2	2.92	GaSe-CdI2	43.29	3.96	1.55	5.94	4.4	5.73	-4.4	-5.94
Cd2Se2-GeI2	3.48	GaSe-CdI2	48.18	4.04	1.63	5.94	4.31	5.72	-4.31	-5.94
I2Mg-Te2Zn2	2.86	GaSe-CdI2	50.02	4.09	1.03	4.98	3.96	4.77	-3.96	-4.98
CdCl2-S2Zn2	0.59	GaSe-CdI2	52.39	3.64	2.12	6.31	4.18	6.1	-4.18	-6.31
Br2Cd-S2Zn2	3.49	GaSe-CdI2	53.93	3.7	2.33	6.36	4.03	6.14	-4.03	-6.36
Cu2I2-GeI2	2.91	GaSe-CdI2	82.94	4.05	1.14	4.86	3.71	4.64	-3.72	-4.86
Cd2S2-I2Mg	2.54	GaSe-CdI2	39.03	3.98	1.88	6.14	4.26	5.93	-4.26	-6.14
Ag2Cl2-HgI2	0.3	GaSe-CdI2	32.88	3.24	1.18	5.83	4.65	5.62	-4.81	-5.83
AlN-BP	2.56	BN-BN	34.45	3.88	0.62	4.97	4.34	4.76	-4.34	-4.97
BP-GaN	1.62	BN-BN	48.06	3.92	0.66	5.06	4.4	4.84	-4.4	-5.06

Table B.2. Formula, mismatch (%), bilayer prototype, binding energy (BE) in meVÅ^{-2}, interlayer distance (d_o) in Å, bandgap (Eg), ionization energy (IE), electron affinity (EA), work function (Φ), energies of conduction band minimum (E_{CBM}) and valence band maximum (E_{VBM}) in eV for 63 DFT-predicted vdW bilayer photocatalysts feasible for overall water splitting photocatalysis along with their bandgap type and band alignment.

Formula	Mismatch (%)	Bilayer Prototype	BE	d_o	Eg	IE	EA	Φ	E_{CBM}	E_{VBM}	Bandgap Type	Band alignment
			meV Å^{-2}	Å	eV	eV	eV	eV	eV	eV		
BaCl2-I2Pb	0.69	MoS2-MoS2	27.46	4	2.68	6.55	3.87	6.33	-3.87	-6.55	ID	I
Br2Pb-Br2Sr	1.86	MoS2-MoS2	28.57	4	3.24	7.28	4.04	7.08	-4.04	-7.3	ID	I
Cl2Pb-Cl2Sr	1.05	MoS2-MoS2	21.25	3.72	3.54	7.77	4.23	7.54	-4.24	-7.77	ID	I
Br2Sr-I2Sn	0.7	MoS2-MoS2	32.9	4.02	2.11	6.08	3.97	5.85	-3.97	-6.08	ID	I
BaBr2-I2Sr	1.22	MoS2-MoS2	33.33	4.01	3.58	6.86	3.28	6.63	-3.28	-6.86	ID	II
Br2Ca-Cl2Pb	1.48	MoS2-MoS2	24.23	3.99	3.3	7.58	4.29	7.35	-4.27	-7.58	ID	II
BaCl2-I2Sr	2.23	MoS2-MoS2	31.36	4	3.66	6.93	3.27	6.7	-3.26	-6.94	ID	II
CaI2-Cl2Sr	3.75	MoS2-MoS2	33.35	4.01	3.2	6.64	3.44	6.43	-3.43	-6.64	ID	II
I2Pb-I2Sn	1.86	MoS2-MoS2	30.38	4.11	2.17	6.16	3.98	5.93	-3.98	-6.15	ID	II
Br2Ca-GeI2	0.04	CdI2-CdI2	46.29	4.03	2.09	5.97	3.88	5.75	-3.89	-5.97	ID	I
CaCl2-GeI2	3.86	CdI2-CdI2	42.11	4.02	2.17	5.83	3.66	5.6	-3.66	-5.83	ID	I
CdI2-I2Mg	2.74	CdI2-CdI2	46.6	4.27	2.45	6.32	3.88	6.11	-3.83	-6.32	ID	I
Cl2Sr-GeI2	3.71	CdI2-CdI2	40.84	4	2.02	6.13	4.11	5.9	-4.11	-6.13	ID	I

CaCl2-I2Mg	1.67	CdI2-CdI2	38.58	4.03	3.65	6.43	2.78	6.21	-2.73	-6.43	ID	II
CdI2-Cl2Pb	1.56	CdI2-CdI2	29.67	3.92	2.2	6.5	4.29	6.28	-4.29	-6.5	ID	II
Al2S2-Ga2S2	1.7	GaS-GaS	91.49	4	1.94	6.05	4.1	5.83	-4.05	-6.05	ID	I
Al2Se2-Ga2S2	3.57	GaS-GaS	86.27	4.02	1.57	5.83	4.26	5.61	-4.26	-5.83	ID	II
S2Zn2-Se2Zn2	3.82	GaSe-GaSe	77.92	3.56	1.84	5.91	4.07	5.69	-4.07	-5.91	D	I
Al2S2-Cl2Pt2	0.92	GaSe-GaSe	94.61	3.52	1.49	5.9	4.41	5.7	-4.41	-5.9	D	II
Cl2Pt2-Ga2S2	0.67	GaSe-GaSe	95.35	3.7	1.42	5.87	4.45	5.66	-4.44	-5.87	ID	II
Ag2I2-Cd2I2	3.88	GaSe-GaSe	55.21	4.11	1.59	5.68	4.09	5.47	-4.09	-5.68	D	II
Al2S2-Cl2Mg	2.45	GaS-CdI2	64.88	3.71	1.99	6.08	4.09	5.86	-4.04	-6.08	ID	I
Cl2Mg-Ga2S2	0.75	GaS-CdI2	65.88	3.69	2.14	6.14	4	5.93	-4	-6.15	ID	I
Al2S2-Cl2Zn	0.41	GaS-CdI2	71.98	4.01	2.06	6.15	4.09	5.93	-4.04	-6.16	ID	II
Al2Se2-Br2Zn	0.54	GaS-CdI2	74.72	4	1.94	5.79	3.85	5.57	-3.82	-5.79	ID	II
CdCl2-Ga2Se2	1.95	GaS-CdI2	57.99	3.76	1.41	5.73	4.32	5.51	-4.32	-5.73	ID	II
Br2Zn-Ga2Se2	0.69	GaS-CdI2	74.69	3.97	1.73	5.76	4.04	5.54	-4.04	-5.77	ID	II
Cl2Zn-Ga2S2	1.28	GaS-CdI2	72.74	3.67	2.34	6.17	3.84	5.97	-3.84	-6.19	ID	II
Al2S2-Al2S2	0.39	GaSe-GaS	94.4	4.62	1.9	6.02	4.12	5.79	-4.1	-6	ID	II
Br2Zn-GeSe	3.53	GeSe-CdI2	53.03	3.81	1.6	5.83	4.23	5.6	-4.18	-5.83	ID	I
Cl2Mg-GeSe	0.18	GeSe-CdI2	42.19	3.83	2.22	5.94	3.72	5.72	-3.72	-5.94	ID	I
Cl2Mg-SSn	2.13	GeSe-CdI2	23.01	3.93	2.28	6.54	4.26	6.33	-4.27	-6.55	ID	I

Cl2Zn-GeS	3.12	GeSe-CdI2	47.41	3.73	2.18	6.34	4.16	6.12	-4.16	-6.35	ID	I
Ga2S2-GeSe	0.57	GeSe-GaS	68.88	3.86	1.34	5.77	4.44	5.57	-4.44	-5.78	ID	II
Ag2I2-BaBr2	2.53	MoS2-GaSe	50.22	4.02	2.07	5.75	3.68	5.55	-3.69	-5.76	D	I
Br2Ca-Br2Cd2	0.36	MoS2-GaSe	41.47	3.8	1.85	6.21	4.35	6	-4.35	-6.21	ID	I
Br2Ca-Br2Hg2	1.46	MoS2-GaSe	97.8	3.9	1.85	6.21	4.36	6	-4.36	-6.21	ID	I
Br2Cd2-CaCl2	3.93	MoS2-GaSe	37.48	3.75	1.94	6.2	4.26	6	-4.26	-6.2	ID	I
CaCl2-Se2Zn2	3.23	MoS2-GaSe	54.61	3.9	1.92	5.76	3.85	5.55	-3.85	-5.76	D	I
Ag2Br2-BaCl2	2.22	MoS2-GaSe	46.78	3.58	1.69	5.99	4.3	5.77	-4.3	-5.99	D	I
Ag2I2-BaCl2	0.91	MoS2-GaSe	49.38	3.94	2.06	5.75	3.69	5.53	-3.69	-5.75	D	I
CaI2-Cd2S2	1.61	MoS2-GaSe	39.37	3.86	1.8	6.07	4.27	5.83	-4.28	-6.06	ID	I
Br2Sr-Cd2S2	1.69	MoS2-GaSe	37.01	3.7	1.78	6.08	4.3	5.85	-4.3	-6.08	D	I
Ag2Cl2-Br2Sr	0.26	MoS2-GaSe	50.95	3.38	1.68	5.88	4.2	5.65	-4.2	-5.88	D	I
Cd2S2-Cl2Sr	2.13	MoS2-GaSe	34.03	3.55	1.88	6.11	4.23	5.89	-4.23	-6.11	D	I
CaI2-Cd2I2	0.35	MoS2-GaSe	41.66	4.25	2	5.8	3.79	5.58	-3.79	-5.8	ID	II
Ag2Cl2-CaI2	0.33	MoS2-GaSe	53.58	3.53	1.68	5.84	4.16	5.62	-4.16	-5.84	D	II
Br2Pb-Cd2S2	0.18	MoS2-GaSe	31.92	3.6	1.77	6.12	4.35	5.9	-4.35	-6.12	D	II
Ag2I2-I2Sn	3.46	MoS2-GaSe	52.92	4.15	1.8	5.68	3.88	5.48	-3.88	-5.68	ID	II
Br2Sr-Cd2I2	0.27	MoS2-GaSe	38.96	4.06	1.98	5.81	3.83	5.59	-3.83	-5.82	ID	II
Ag2Br2-I2Pb	1.53	MoS2-GaSe	50.62	3.89	1.7	6.09	4.39	5.88	-4.39	-6.09	D	II

Cl2Mg-Cl2Pt2	1.14	GaSe-CdI2	68.46	3.56	1.51	5.97	4.45	5.76	-4.44	-5.97	ID	I
Br2Zn-S2Zn2	3.23	GaSe-CdI2	68.69	3.78	2.38	6.27	3.89	6.06	-3.89	-6.27	D	I
I2Mg-Se2Zn2	3.19	GaSe-CdI2	59.41	4.01	1.58	5.96	4.38	5.75	-4.38	-5.96	D	I
Cd2S2-HgI2	1.65	GaSe-CdI2	17.83	3.58	1.65	6.01	4.37	5.8	-4.37	-6.02	ID	I
Cd2Se2-CdI2	2.92	GaSe-CdI2	43.29	3.96	1.55	5.94	4.4	5.73	-4.4	-5.94	D	I
Cd2Se2-GeI2	3.48	GaSe-CdI2	48.18	4.04	1.63	5.94	4.31	5.72	-4.31	-5.94	D	I
Cl2Pt2-Cl2Zn	0.9	GaSe-CdI2	75.14	3.54	1.61	6.03	4.43	5.83	-4.41	-6.03	ID	II
CdCl2-S2Zn2	0.59	GaSe-CdI2	52.39	3.64	2.12	6.31	4.18	6.1	-4.18	-6.31	D	II
Br2Cd-S2Zn2	3.49	GaSe-CdI2	53.93	3.7	2.33	6.36	4.03	6.14	-4.03	-6.36	D	II
Br2Cd2-I2Zn	0.11	GaSe-CdI2	52.58	4.03	1.84	6.19	4.35	5.97	-4.35	-6.18	ID	II
Cd2I2-GeI2	2.31	GaSe-CdI2	49.02	4.24	1.79	5.83	4.04	5.6	-4.04	-5.82	ID	II
Cd2S2-I2Mg	2.54	GaSe-CdI2	39.03	3.98	1.88	6.14	4.26	5.93	-4.26	-6.14	D	II

Appendix C: Predictions on Unlabeled vdW Bilayers

Table C.1. Formula, mismatch (%), bilayer prototype, binding energy (B.E) in meVÅ^{-2}, interlayer distance (d_o) in Å, bandgap (Eg), ionization energy (IE), electron affinity (EA), work function (Φ), energies of conduction band minimum (E_{CBM}) and valence band maximum (E_{VBM}) in eV for 93 ML-predicted vdW bilayer photocatalysts feasible for overall water splitting photocatalysis.

Formula	Mismatch (%)	Bilayer Prototype	E_b	d_o	Eg	IE	EA	Φ	E_{CBM}	E_{VBM}
			meVÅ^{-2}	Å	eV	eV	eV	eV	eV	eV
BaBr2-BaCl2	3.44	MoS2-MoS2	27.67	4.01	3.97	7.08	3.32	6.88	-3.37	-7.53
Br2Ca-Cl2Sr	2.53	MoS2-MoS2	43.87	3.79	3.95	7.48	3.59	7.18	-3.53	-7.59
I2Sr-PbS2	1.96	MoS2-MoS2	27.93	3.86	1.88	6.1	3.99	5.81	-3.88	-5.88
I2Pb-I2Sr	2.91	MoS2-MoS2	27.63	3.92	2.26	6.43	4.36	6.18	-4.1	-6.31
BaCl2-CaI2	3.31	MoS2-MoS2	34.14	4.08	3.2	6.37	3.55	6.27	-3.59	-6.64
BaCl2-I2Sn	2.54	MoS2-MoS2	21.64	3.94	2.32	5.95	3.71	5.83	-3.68	-6.22
BaCl2-Br2Sr	3.24	MoS2-MoS2	45.8	3.96	4.01	7.3	3.49	7.06	-3.42	-7.66
CaI2-I2Sn	0.77	MoS2-MoS2	33.47	4	1.93	6.22	4.07	5.92	-3.92	-6.19
Br2Sr-CaI2	0.07	MoS2-MoS2	44.78	3.93	2.91	6.23	3.67	5.92	-3.56	-6.57
CaI2-I2Pb	2.63	MoS2-MoS2	33.67	3.93	2.02	6.46	4.58	6.22	-4.37	-6.32
Br2Pb-I2Sn	2.56	MoS2-MoS2	14.92	3.6	1.85	6.09	4.53	5.76	-4.25	-6.17

Br2Pb-Cl2Sr	1.96	MoS2-MoS2	25.05	3.81	3.48	7.47	4.49	7.19	-4.3	-7.44
Br2Sr-I2Pb	2.56	MoS2-MoS2	28.34	3.9	2.46	6.67	4.46	6.37	-4.15	-6.46
Br2Sr-Cl2Sr	3.82	MoS2-MoS2	45.21	3.94	4.11	7.53	3.64	7.24	-3.4	-7.66
Br2Ca-CdI2	0.6	CdI2-CdI2	19.19	4	2.55	6.24	3.8	6.02	-3.8	-6.17
Br2Ca-HgI2	2.04	CdI2-CdI2	14.52	3.74	2.17	6.42	4.37	6.18	-4.17	-6.16
Br2Ca-I2Mg	2.15	CdI2-CdI2	43.54	4.06	3.3	6.35	3.34	6.11	-3.11	-6.21
Br2Zn-CdCl2	2.64	CdI2-CdI2	29.93	3.79	2.35	7.11	4.57	6.84	-4.38	-7.21
Br2Mg-CdCl2	0.41	CdI2-CdI2	39.03	3.85	2.68	7	4.11	6.71	-4.18	-7.17
HfO2-O2Zr	0.44	CdI2-CdI2	308.32	3.65	3.66	7.25	3.72	7.15	-3.55	-7.69
Br2Mg-Br2Zn	2.22	CdI2-CdI2	39.55	3.93	2.65	6.8	4.05	6.54	-3.8	-6.92
Br2Zn-Cl2Mg	3.35	CdI2-CdI2	35.66	3.82	2.96	6.95	4.19	6.85	-3.79	-7.13
CaCl2-I2Zn	0.92	CdI2-CdI2	24.26	3.89	2.46	6.34	3.96	6.09	-3.96	-6.3
Br2Cd-CaCl2	1.84	CdI2-CdI2	16.58	3.82	3.23	7.17	4.05	6.88	-4.05	-7.23
CdI2-HgI2	1.44	CdI2-CdI2	7.57	3.84	1.75	6.27	4.66	6.01	-4.36	-6.08
CdI2-GeI2	0.56	CdI2-CdI2	28.86	3.96	1.81	5.98	4	5.76	-4.05	-5.93
CdI2-Cl2Sr	3.15	CdI2-CdI2	13.65	4	2.39	6.39	4.22	6.15	-4.13	-6.36
Cl2Mg-PdSe2	1.67	CdI2-CdI2	69.18	3.58	1.5	5.82	4.55	5.6	-4.42	-5.73
Cl2Mg-Cl2Zn	2.04	CdI2-CdI2	37.06	3.71	3.32	7.23	3.76	6.96	-3.65	-7.43
CaI2-HgI2	3.62	CdI2-CdI2	22.13	3.86	1.94	6.26	4.33	6.01	-4.13	-5.94
Cl2Mg-PtS2	2.74	CdI2-CdI2	85.86	3.62	1.81	6.27	4.39	6.12	-4.32	-6.26
I2Mg-I2Zn	2.59	CdI2-CdI2	47.4	4.13	2.02	6.11	4.26	5.89	-4.03	-6.02
Br2Cd-I2Zn	0.93	CdI2-CdI2	24.46	3.91	1.74	6.04	4.53	5.95	-4.35	-6.19
GeI2-HgI2	2	CdI2-CdI2	15.74	3.72	1.53	6.3	4.52	6.02	-4.22	-5.87
GeI2-I2Mg	2.19	CdI2-CdI2	53.93	4.04	1.97	5.98	3.97	5.77	-3.85	-5.81

Br2Cd-I2Mg	3.51	CdI2-CdI2	39.23	3.98	2.18	6.28	4.39	6.04	-4.15	-6.14
Ga2Se2-S2Zn2	2.25	GaSe-GaSe	60.57	3.82	1.6	5.51	3.84	5.28	-4.14	-5.69
Br2Cd2-I2Zn2	2.11	GaSe-GaSe	51	4.01	1.45	5.55	4.22	5.39	-4.22	-5.68
Ag2I2-Cd2Se2	2.71	GaSe-GaSe	36.76	3.74	1.25	5.61	4.37	5.36	-4.42	-5.7
Ag2I2-Hg2I2	2.65	GaSe-GaSe	20.28	4.03	1.34	5.5	4.27	5.3	-4.43	-5.7
AuI-BaBr2	3.5	CdI2-BN	36.39	4.56	1.81	6.55	6.2	6.73	-4.41	-6.59
AuI-BaCl2	0.44	CdI2-BN	37.54	3.54	1.85	6.94	4.67	6.73	-4.27	-6.26
AuI-I2Sr	0.97	CdI2-BN	65.11	3.53	1.51	6.7	4.87	6.46	-4.32	-5.85
Al2Se2-Br2Mg	2.76	GaS-CdI2	63.35	3.97	1.85	5.7	3.98	5.52	-3.83	-5.8
Al2Se2-Cl2Mg	2.81	GaS-CdI2	58.27	3.88	2.08	5.77	3.9	5.56	-3.8	-5.84
CaCl2-In2Se2	1.54	GaS-CdI2	31.21	3.8	1.8	5.92	3.96	5.69	-4.11	-5.98
Br2Mg-Ga2Se2	1.53	GaS-CdI2	51.49	3.98	1.76	5.65	3.91	5.44	-3.96	-5.75
Cl2Zn-GeSe	1.85	GeSe-CdI2	33.16	3.76	1.75	6.01	4.28	5.77	-4.24	-6
Br2Mg-SSn	3.44	GeSe-CdI2	40.06	3.91	1.94	6.38	4.38	6.18	-4.29	-6.42
Br2Zn-SeSn	2.82	GeSe-CdI2	32.17	3.72	1.39	6.1	4.51	5.88	-4.36	-5.94
Br2Mg-SeSn	0.6	GeSe-CdI2	29.85	3.82	1.71	6.05	4.26	5.85	-4.19	-6.01
Al2Se2-SeSn	3.36	GeSe-GaS	40.47	3.87	1.23	5.94	4.69	5.74	-4.43	-5.69
S2Zn2-SeSn	0.4	GeSe-GaSe	42.41	3.65	1.35	5.58	3.96	5.38	-4.3	-5.81
HfO2-MoS2	2.19	MoS2-CdI2	242.4	3.63	1.64	5.77	4.04	5.75	-3.98	-5.73
Br2Ca-Se2Zn2	1.06	MoS2-GaSe	53.22	3.65	1.77	5.75	3.92	5.52	-4.12	-5.87
Br2Ca-Br2Ge2	3.1	MoS2-GaSe	57.35	3.82	1.97	7.07	5.08	6.74	-4.39	-6.6

CaCl2-S2Zn2	0.59	MoS2-GaSe	73.84	3.56	2.13	5.95	3.17	5.7	-3.54	-6.12
CaCl2-Cd2Cl2	0.35	MoS2-GaSe	81.78	3.57	1.99	6.24	4.02	6	-4.12	-6.27
Hg2I2-I2Sr	3.96	MoS2-GaSe	24.74	4.1	1.78	5.95	4.31	5.7	-4.24	-5.99
Ag2I2-I2Sr	1.32	MoS2-GaSe	44.77	3.97	1.86	5.75	3.85	5.51	-3.81	-5.76
BaCl2-Cd2I2	2.97	MoS2-GaSe	31.43	3.81	2.09	5.75	3.76	5.57	-3.65	-5.86
BaCl2-I2In2	1.21	MoS2-GaSe	38.94	4.04	1.98	5.36	3.92	5.27	-3.93	-5.84
Ag2Cl2-BaCl2	2.98	MoS2-GaSe	51.12	3.51	1.65	6.05	4.12	5.86	-4.25	-5.93
BaCl2-Cd2Se2	1.8	MoS2-GaSe	21.69	3.71	1.87	5.88	3.98	5.66	-4.15	-6.02
BaCl2-Hg2I2	1.73	MoS2-GaSe	17.86	3.99	2.06	5.85	4.07	5.74	-4.06	-6.18
CaI2-I2In2	2.1	MoS2-GaSe	50.07	4.11	1.56	5.72	4.4	5.44	-4.26	-5.88
Cd2I2-I2Sn	0.42	MoS2-GaSe	16.28	4.04	1.7	5.68	3.92	5.45	-4.06	-5.68
Ag2Br2-I2Sn	0.33	MoS2-GaSe	41.35	3.67	1.48	5.81	4.29	5.58	-4.36	-5.88
Cd2S2-I2Sn	2.38	MoS2-GaSe	24.32	3.81	1.56	5.6	3.97	5.29	-4.22	-5.89
Br2Ge2-I2Sn	3.95	MoS2-GaSe	42.91	3.5	0.98	5.88	5.18	5.66	-4.36	-5.7
Cd2Se2-I2Sn	0.75	MoS2-GaSe	21.04	3.79	1.41	5.79	4.17	5.56	-4.29	-5.85
Hg2I2-I2Sn	0.81	MoS2-GaSe	7.38	4.03	1.47	5.59	4.02	5.36	-4.35	-5.83
Br2Sr-Cl2Ge2	3.63	MoS2-GaSe	75.76	3.95	1.55	6.35	4.79	6.09	-4.38	-5.94
Br2Sr-I2In2	2.03	MoS2-GaSe	50	3.95	1.97	5.79	4.23	5.58	-3.97	-5.77
Ag2Br2-Br2Sr	1.02	MoS2-GaSe	53.32	3.63	1.86	6.35	4.39	6.07	-4.36	-6.16
Br2Ge2-Br2Sr	3.25	MoS2-GaSe	60.15	3.86	2.08	7.13	5.13	6.81	-4.28	-6.68
Br2Sr-Cd2Se2	1.44	MoS2-GaSe	33.22	3.76	1.74	6.03	4.27	5.79	-4.28	-6.05
Ag2I2-I2Pb	1.6	MoS2-GaSe	30.78	3.91	1.69	5.88	4.43	5.64	-4.23	-5.83
Cl2Ge2-Cl2Sr	0.19	MoS2-GaSe	69.43	3.88	1.8	6.4	4.65	6.09	-4.41	-6.15

Br2Cd2-Cl2Sr	2.88	MoS2-GaSe	45.36	3.87	2.08	6.3	4.31	6.06	-4.12	-6.39
Cl2Sr-Se2Zn2	3.59	MoS2-GaSe	58.75	3.66	1.86	5.79	4.04	5.73	-4.03	-5.9
Br2Ge2-Cl2Sr	0.57	MoS2-GaSe	62.04	3.8	2.26	7.25	5.02	6.89	-4.29	-6.76
Br2Mg-S2Zn2	1	GaSe-CdI2	53.74	3.85	2.29	5.92	3.46	5.68	-3.79	-6.16
Br2Cd2-I2Mg	2.48	GaSe-CdI2	54.41	4.05	1.79	5.87	4.27	5.67	-4.17	-5.9
Br2Mg-Cd2Cl2	1.24	GaSe-CdI2	48.24	3.8	1.74	6.03	4.24	5.75	-4.42	-6.26
Cd2I2-CdI2	1.75	GaSe-CdI2	21.74	4.06	1.75	5.85	4.2	5.61	-4.12	-5.75
Cd2I2-HgI2	0.31	GaSe-CdI2	11.29	3.84	1.38	5.83	4.56	5.63	-4.42	-5.69
Ag2Br2-GeI2	3.06	GaSe-CdI2	54.67	3.66	1.43	5.95	4.45	5.7	-4.42	-5.81
I2Zn-Se2Zn2	0.6	GaSe-CdI2	47.04	3.89	1.39	5.86	4.3	5.66	-4.29	-5.84
Br2Cd-Se2Zn2	0.33	GaSe-CdI2	37.9	3.71	1.44	5.78	4.15	5.56	-4.35	-5.98
Cd2S2-CdI2	0.21	GaSe-CdI2	22.71	3.81	1.71	5.92	4.15	5.68	-4.38	-6.08
Cd2S2-GeI2	0.35	GaSe-CdI2	32.47	3.78	1.48	5.74	4.08	5.55	-4.36	-5.92
Br2Ge2-I2Mg	0.97	GaSe-CdI2	66.38	3.86	1.51	6.07	5.17	5.9	-4.42	-5.75

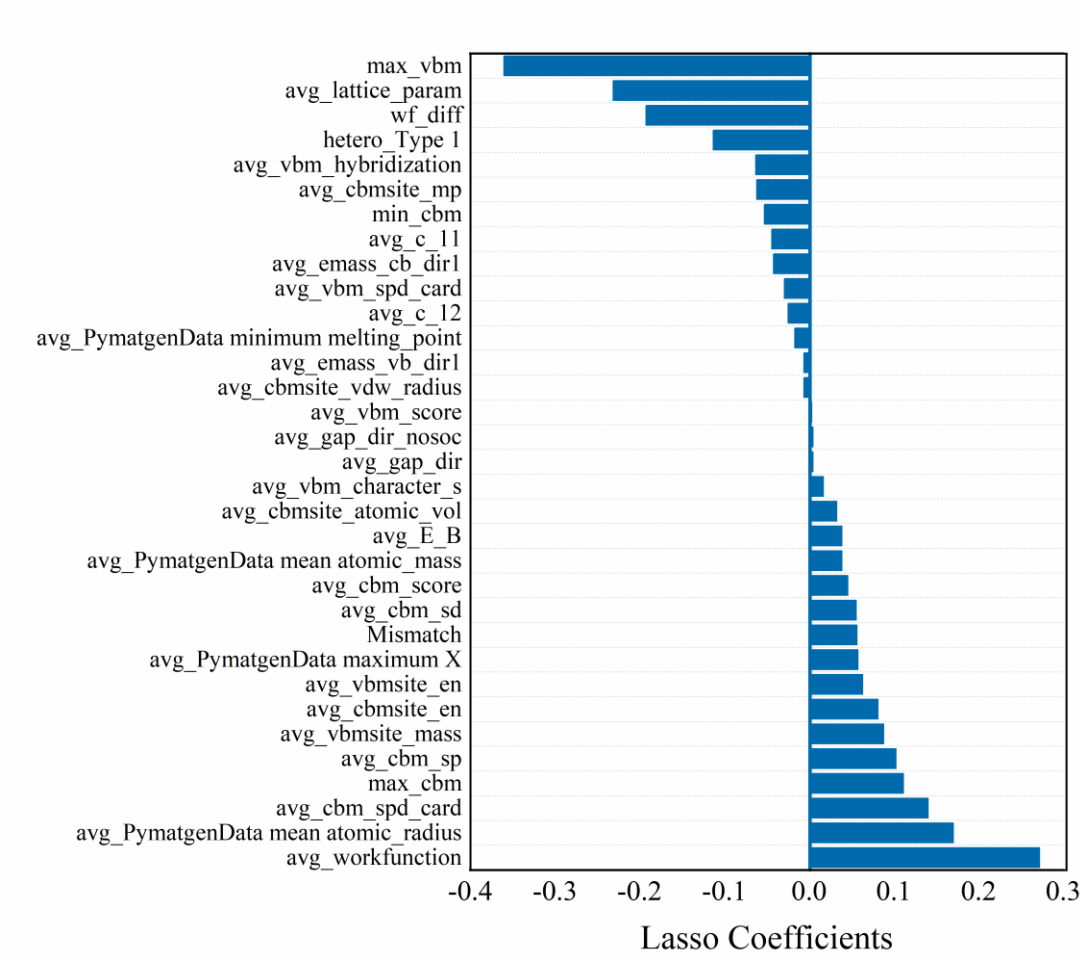

Figure C.1. The LASSO coefficients for LASSO-selected set of descriptors for IE.

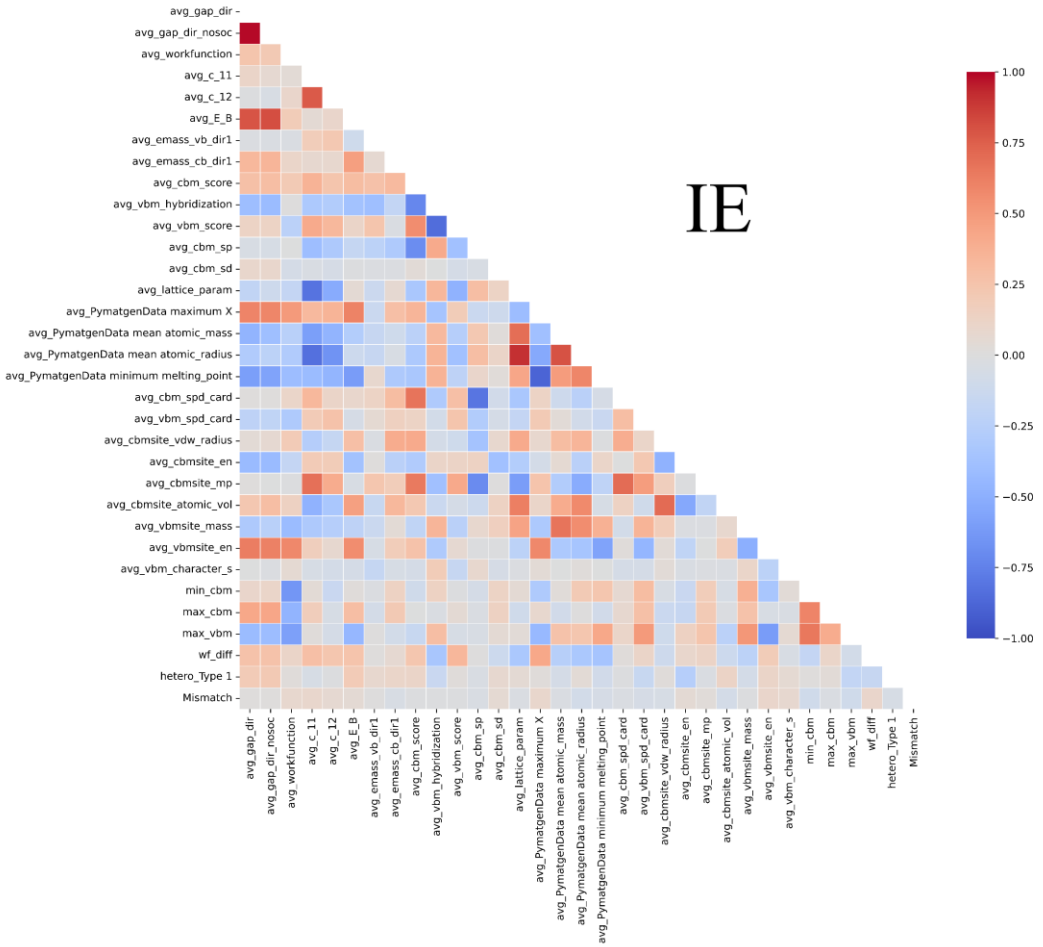

Figure C.2. Correlation map for LASSO-selected set of descriptors for IE.

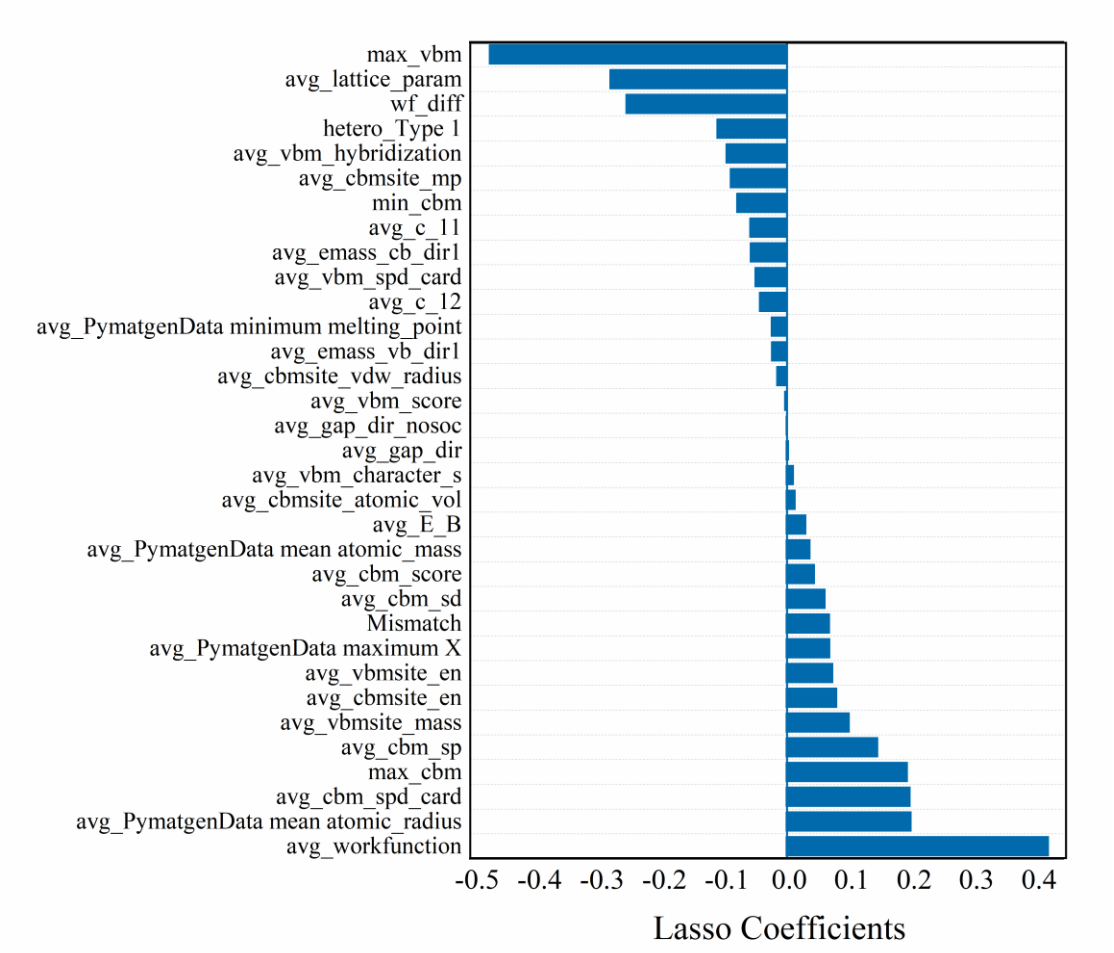

Figure C.3. The LASSO coefficients for LASSO-selected set of descriptors for EA.

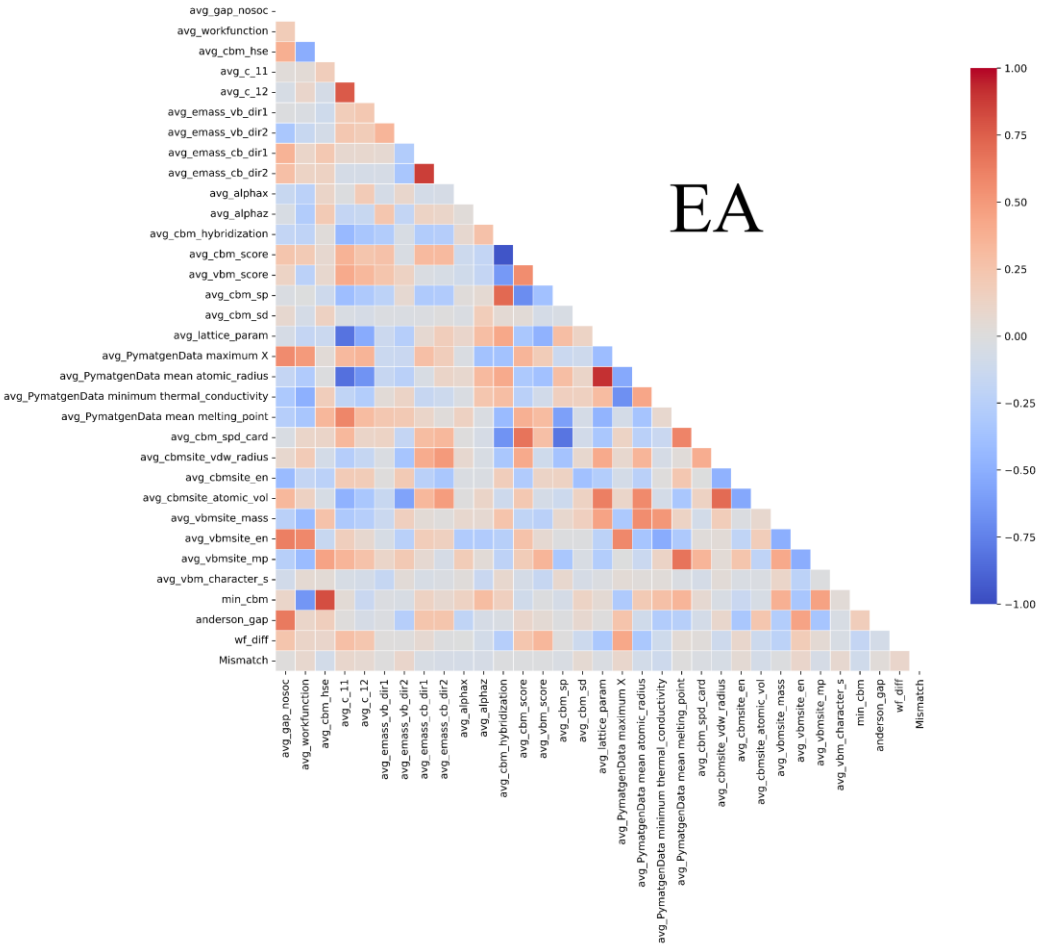

Figure C.4. Correlation map for LASSO-selected set of descriptors for EA.

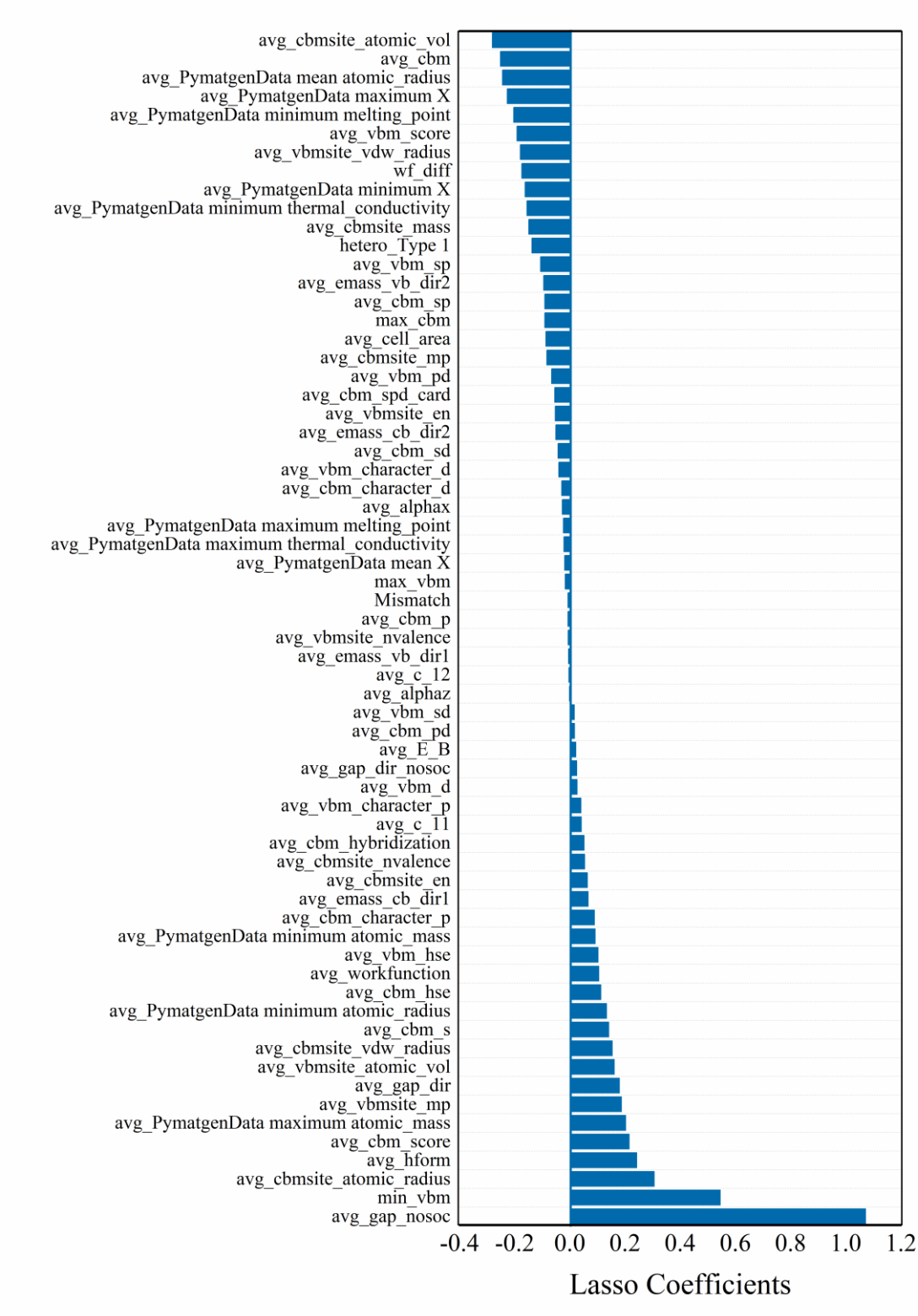

Figure C.5. The LASSO coefficients for LASSO-selected set of descriptors for E_g.

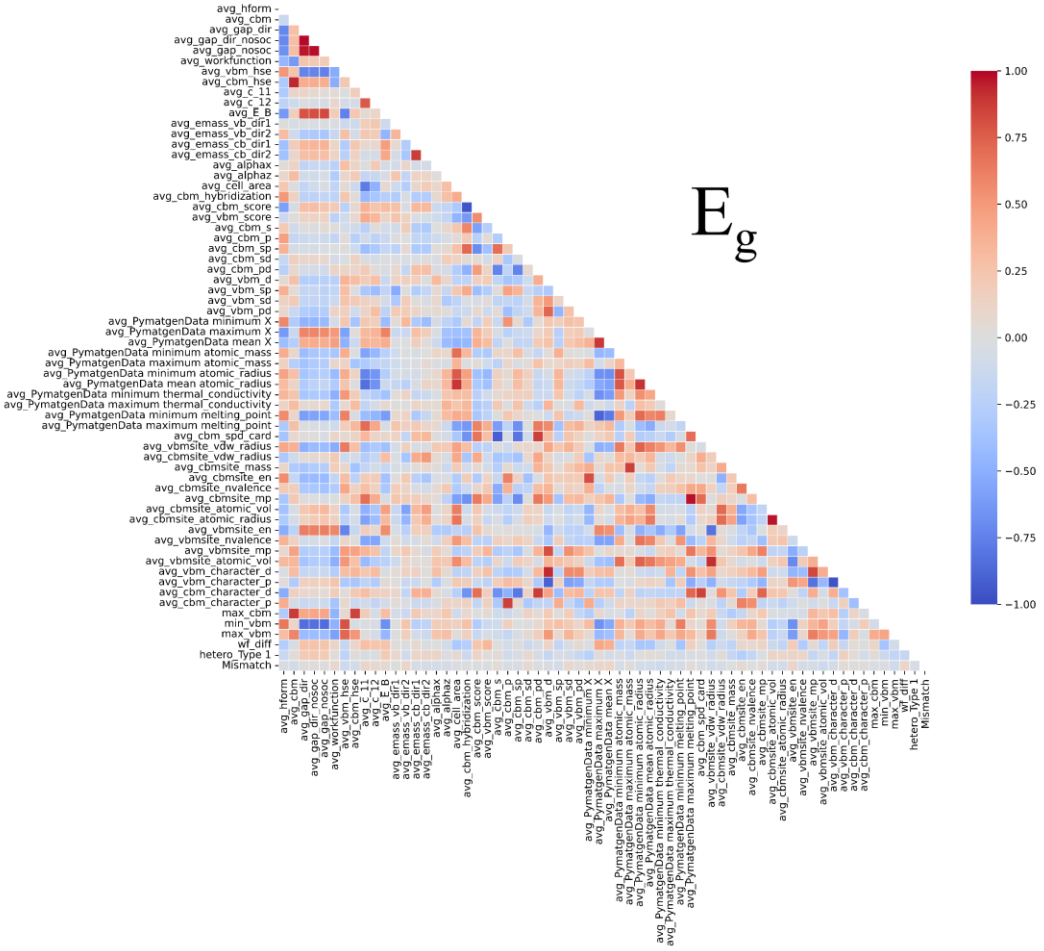

Figure C.6. Correlation map for LASSO-selected set of descriptors for E_g.

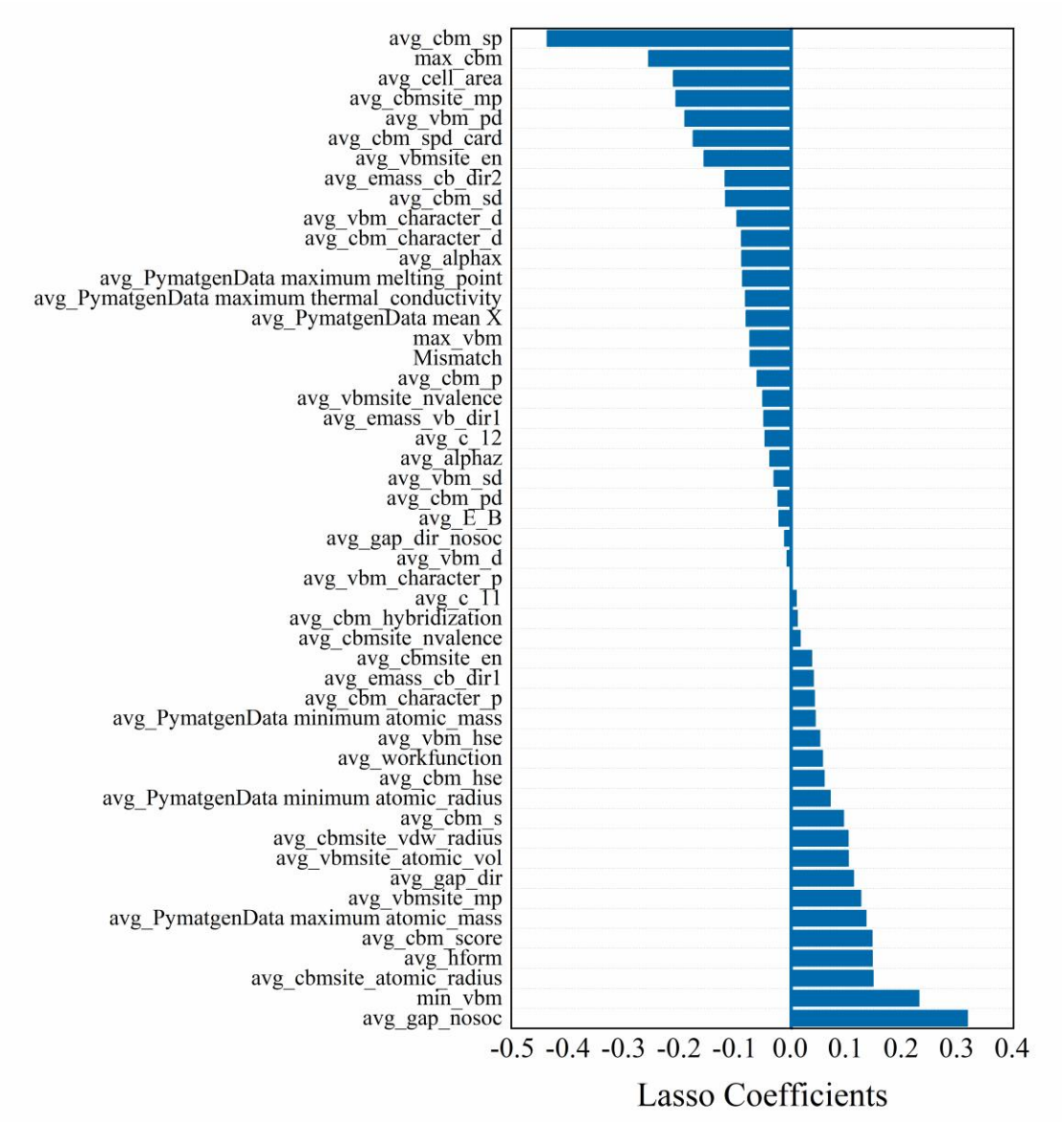

Figure C.7. The LASSO coefficients for LASSO-selected set of descriptors for E_{CBM}.

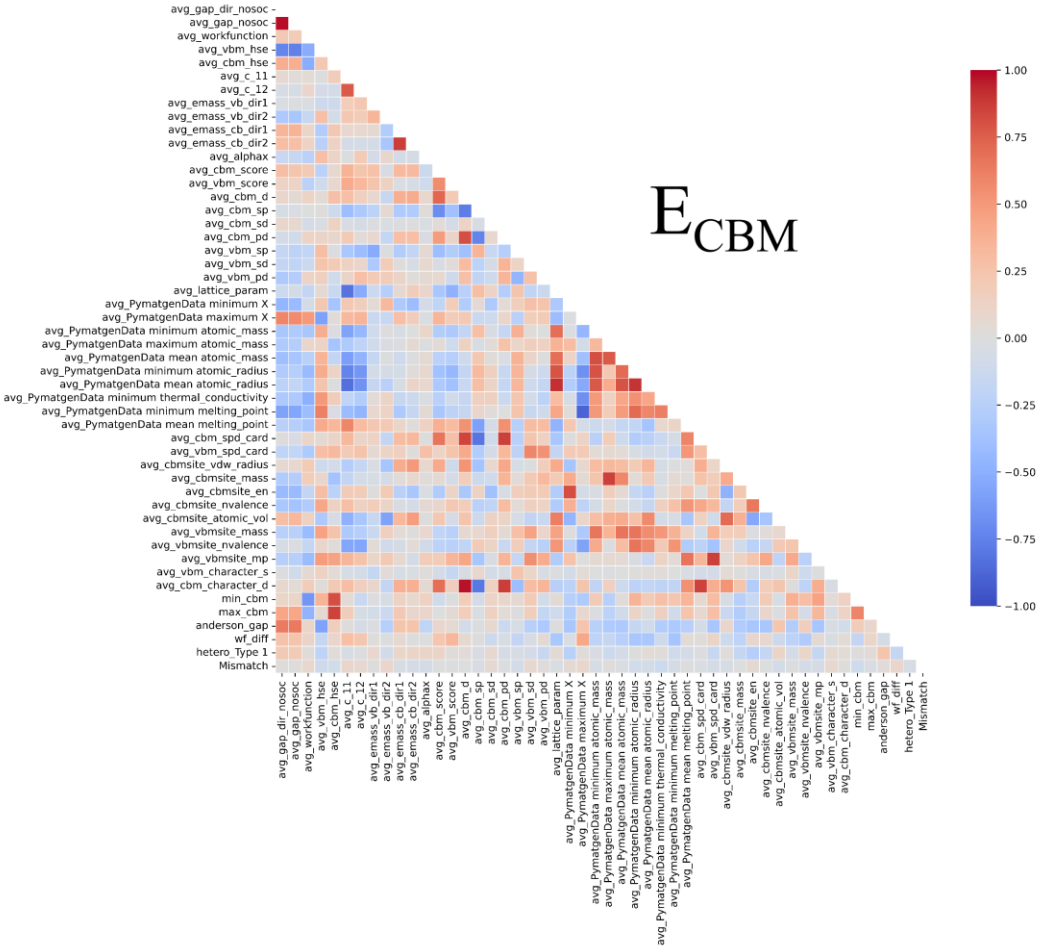

Figure C.8. Correlation map for LASSO-selected set of descriptors for E_{CBM}.

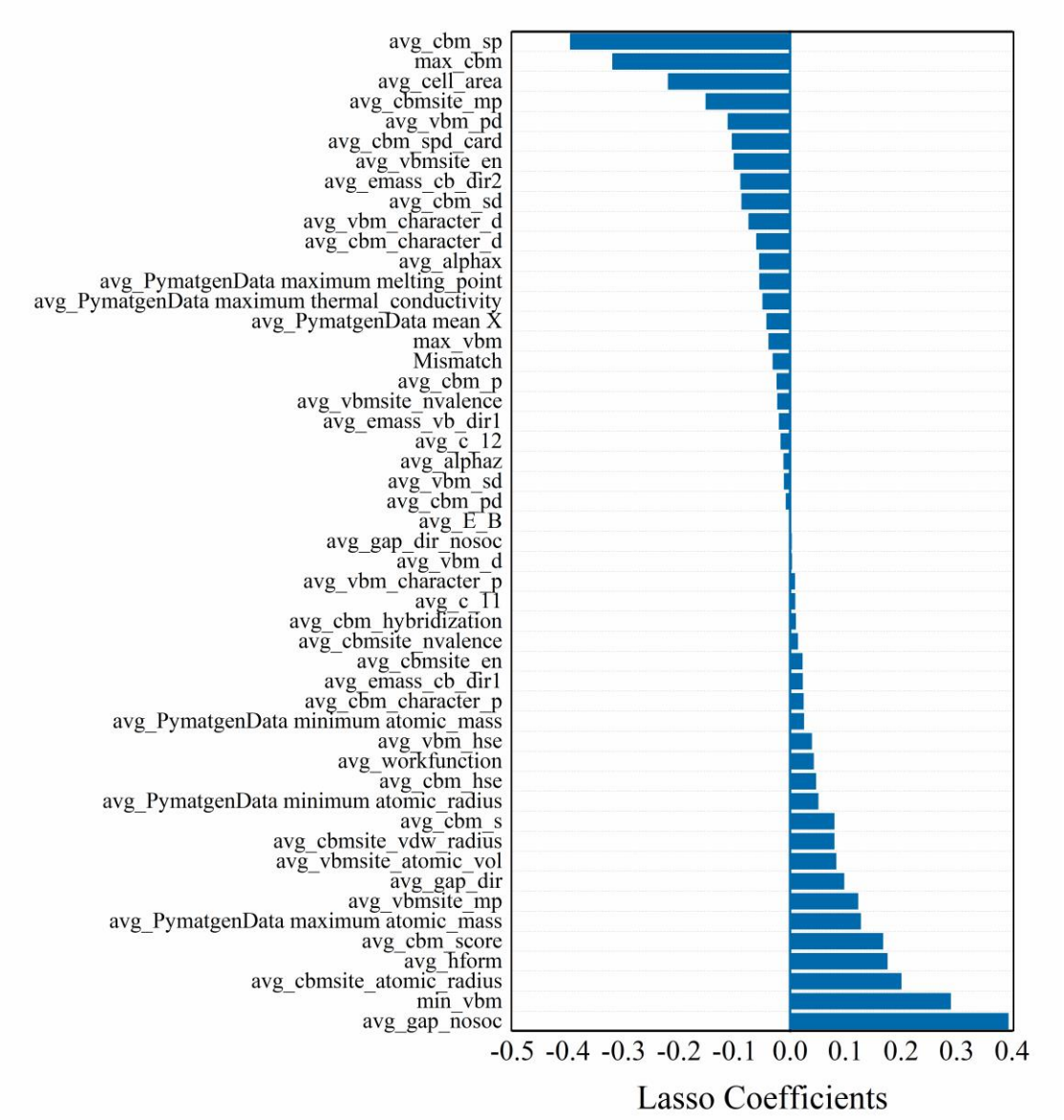

Figure C.9. The LASSO coefficients for LASSO-selected set of descriptors for E_{VBM}.

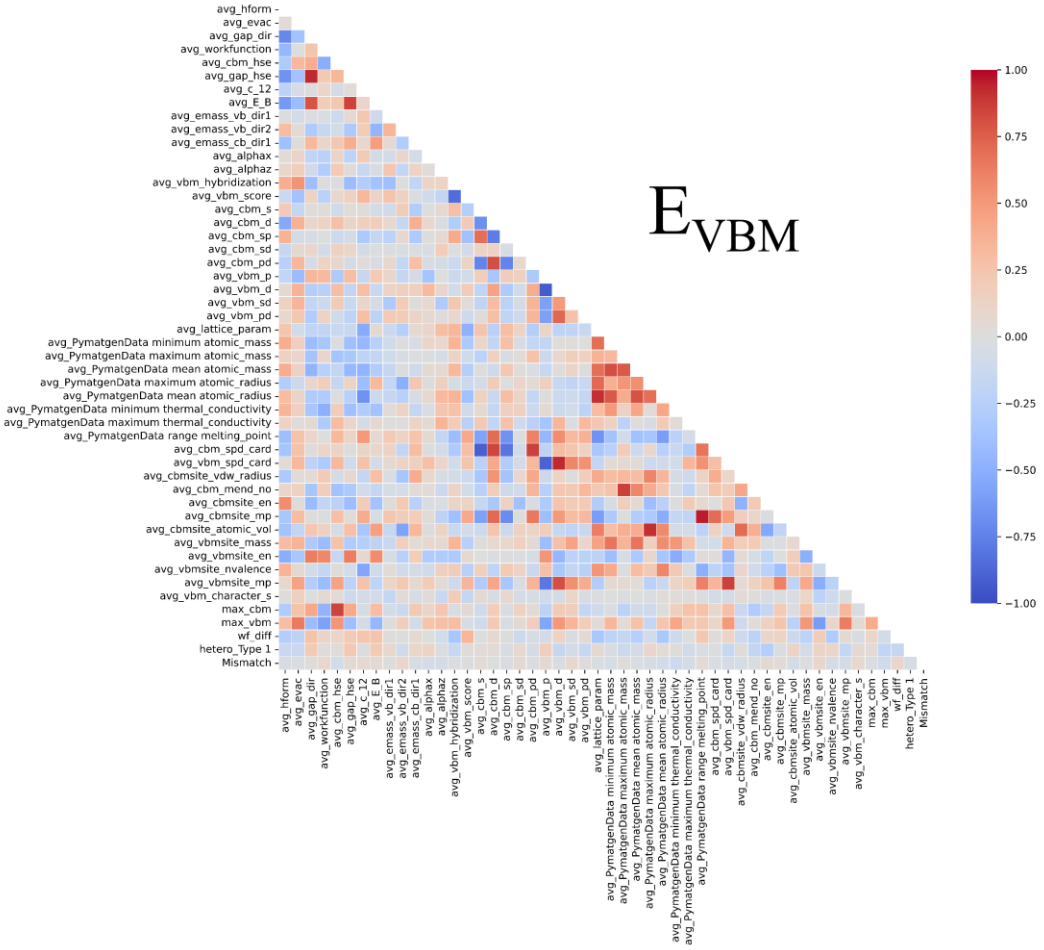

Figure C.10. Correlation map for LASSO-selected set of descriptors for E_{VBM}.

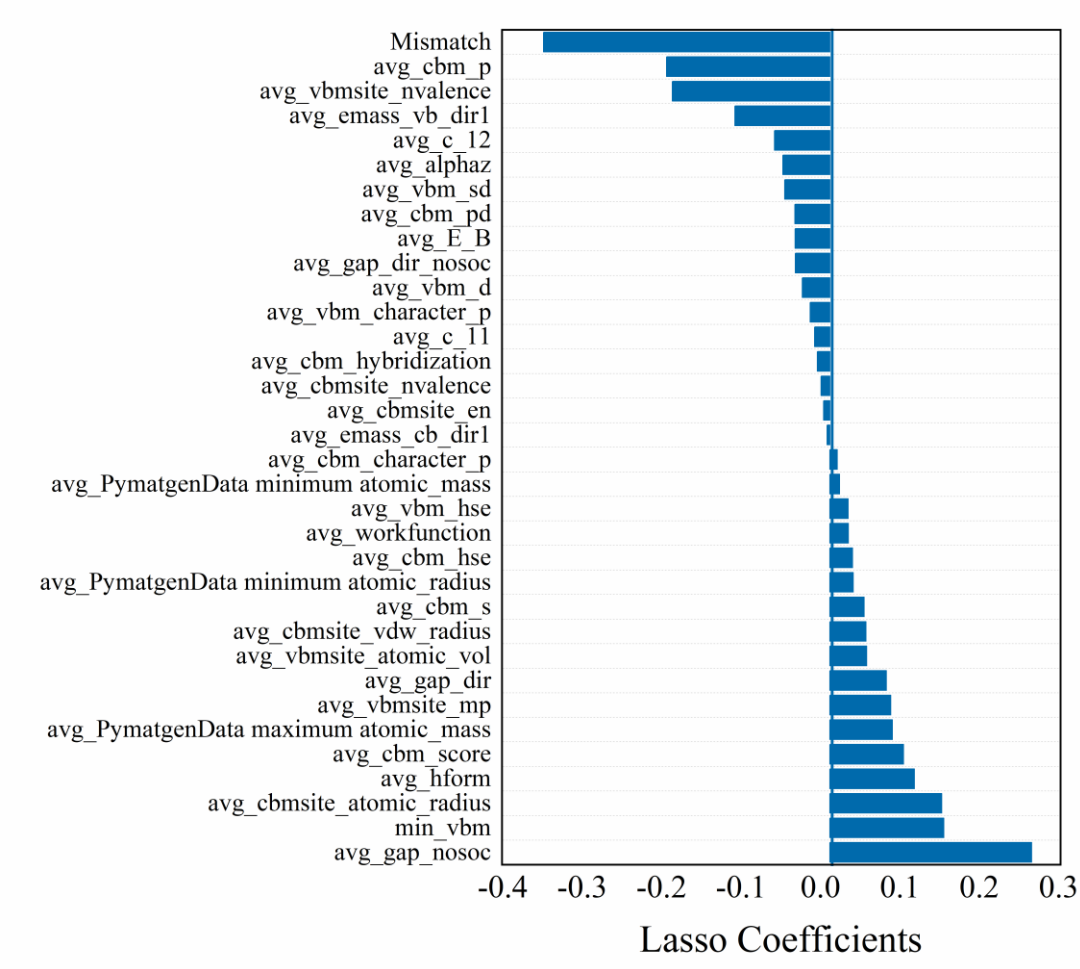

Figure C.11. The LASSO coefficients for LASSO-selected set of descriptors for Φ.

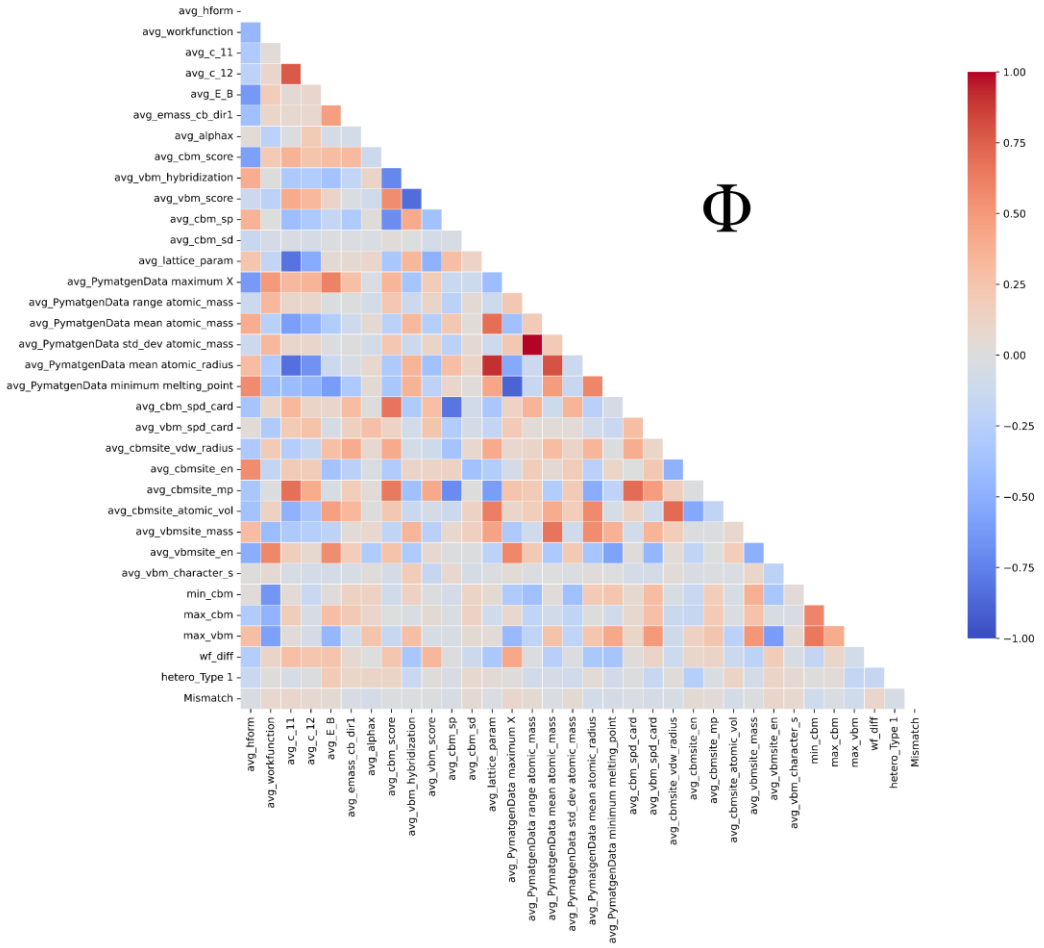

Figure C.12. Correlation map for LASSO-selected set of descriptors for Φ.

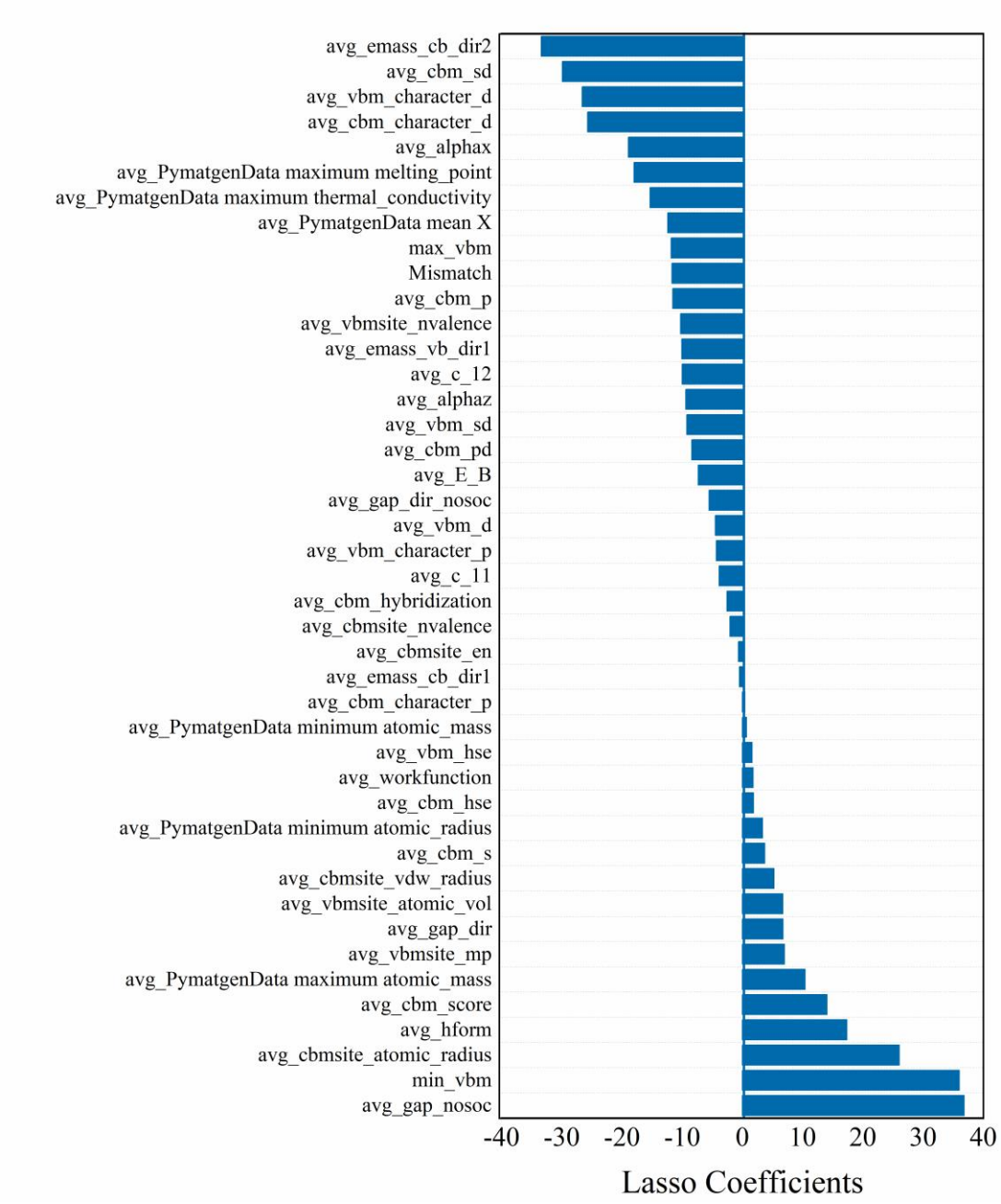

Figure C.13. The LASSO coefficients for LASSO-selected set of descriptors for E_b.

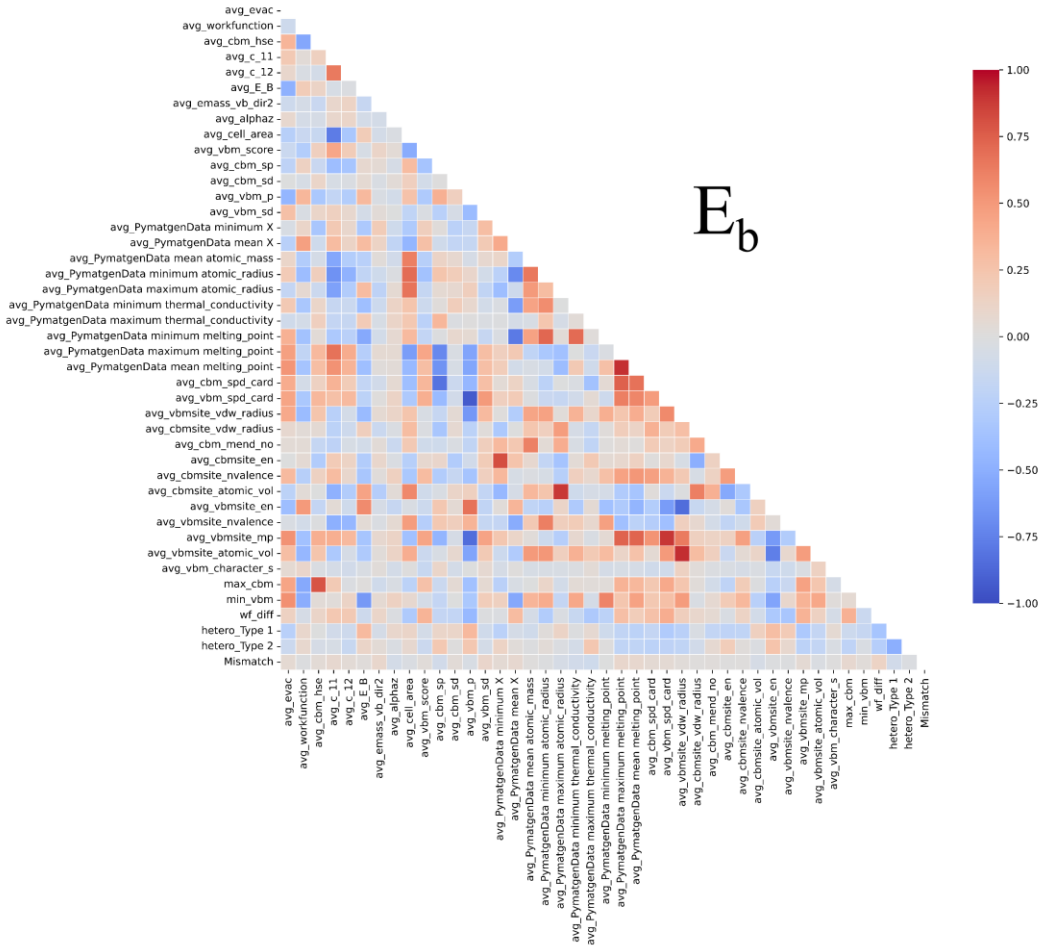

Figure C.14. Correlation map for LASSO-selected set of descriptors for E_b.

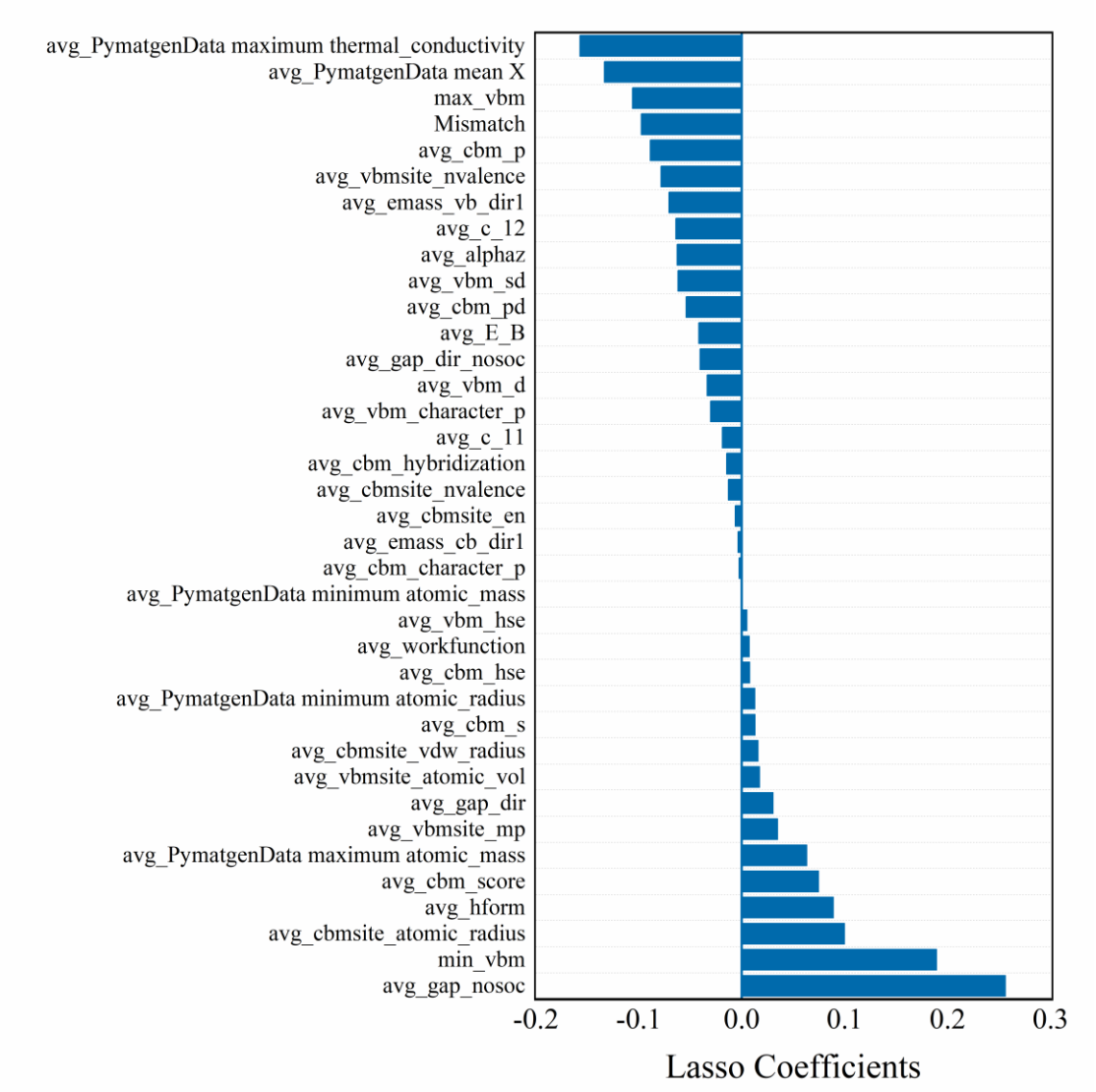

Figure C.15. The LASSO coefficients for LASSO-selected set of descriptors for d_o.

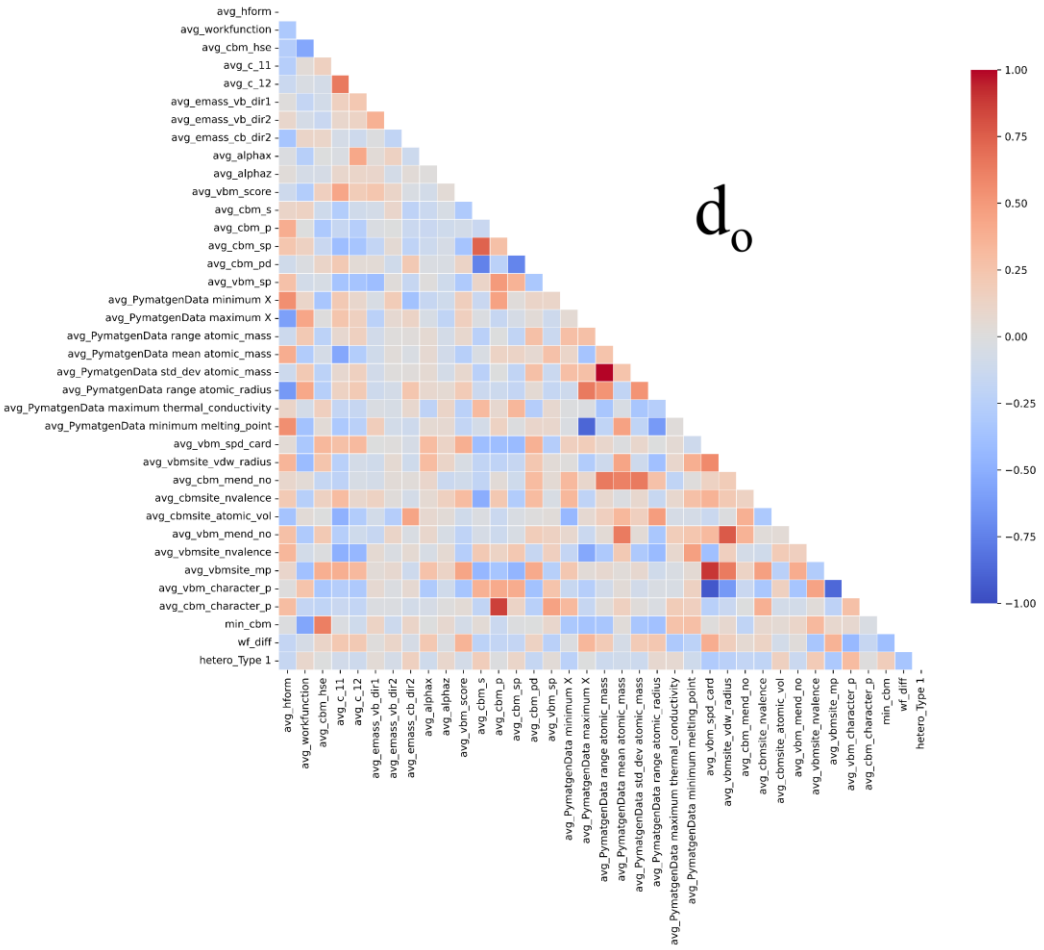

Figure C.16. Correlation map for LASSO-selected set of descriptors for d_o.

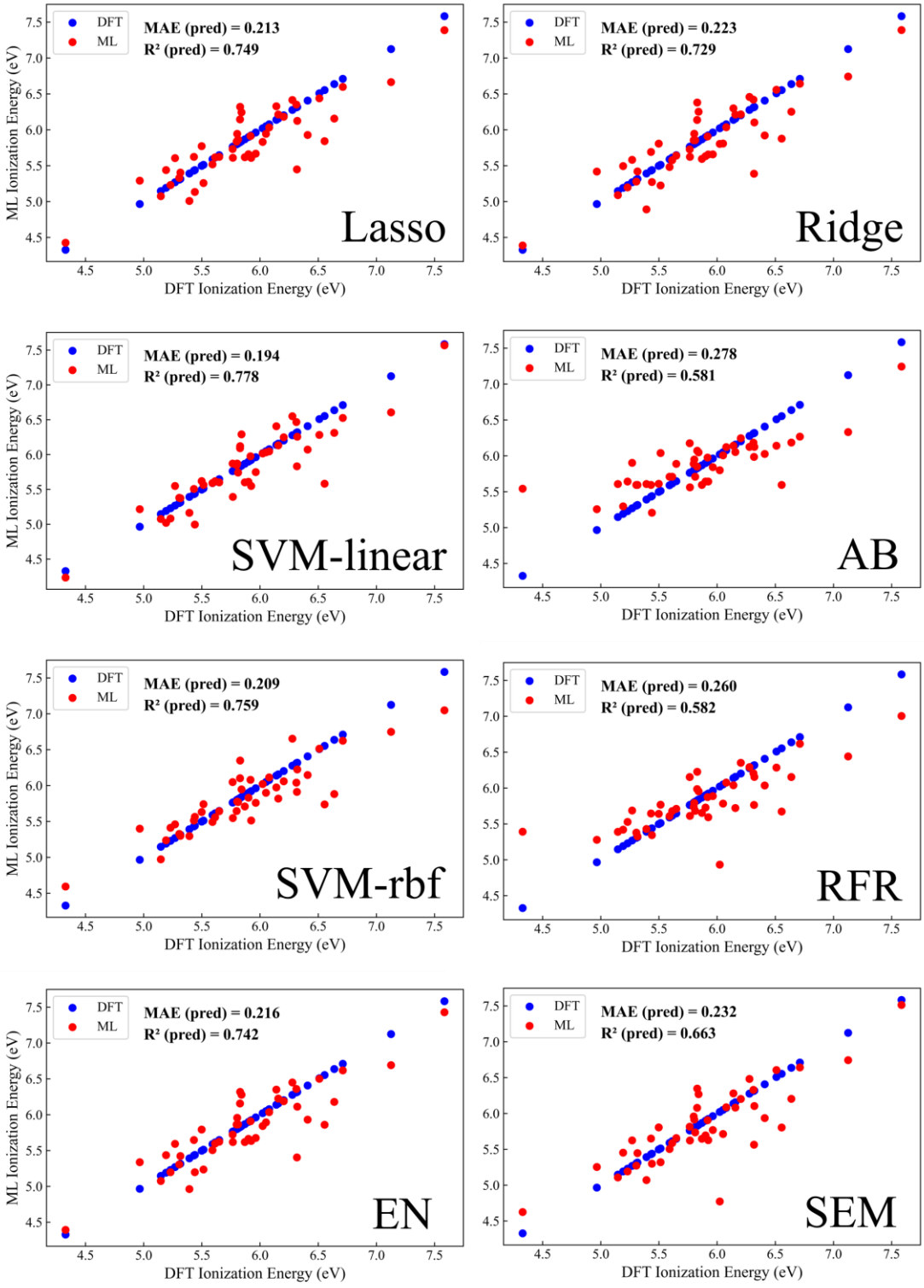

Figure C.17. Comparison of DFT and ML predicted IE for validation set employing multiple models. The evaluation metrics are provided for each model.

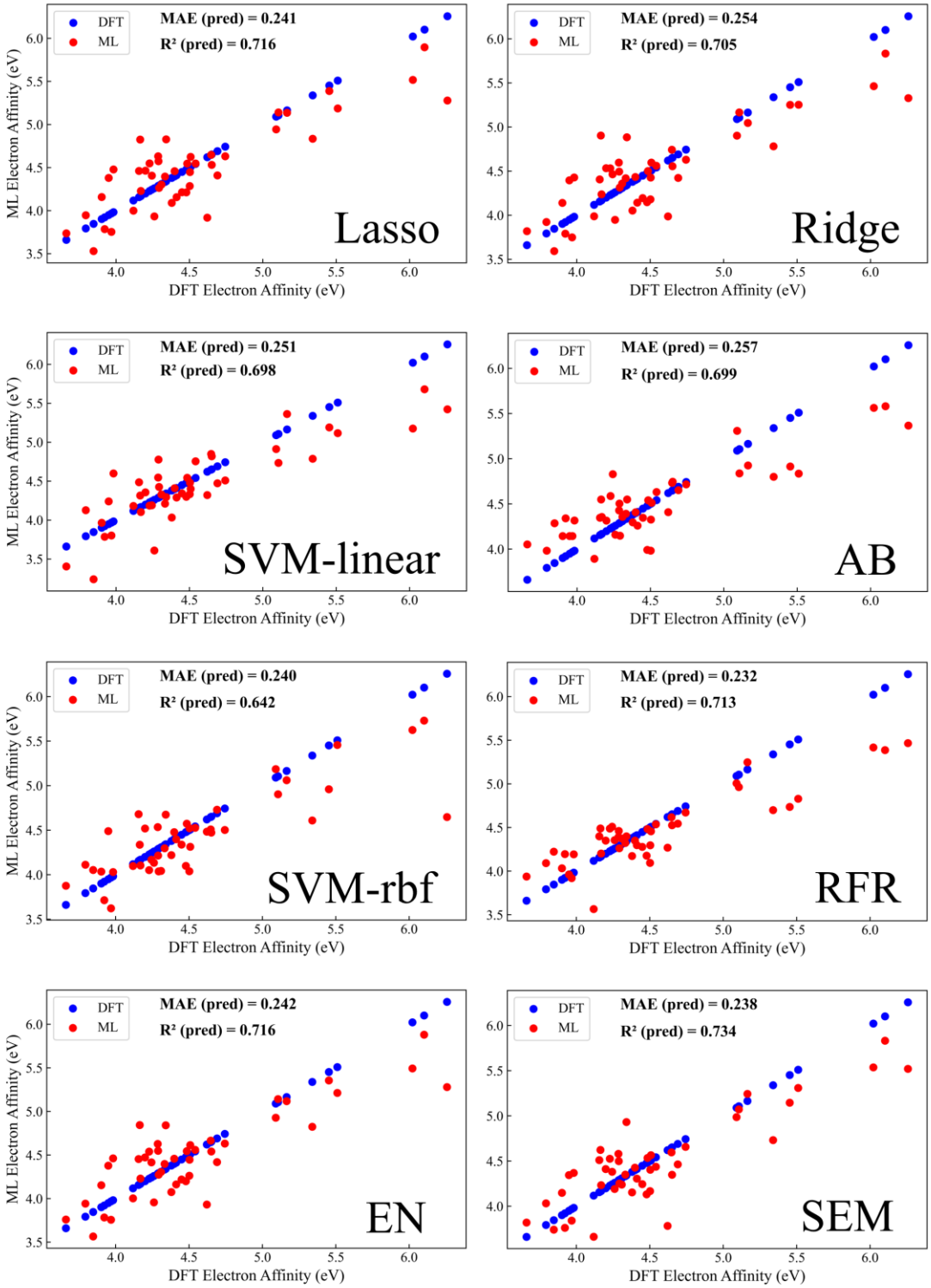

Figure C.18. Comparison of DFT and ML predicted EA for validation set employing multiple models. The evaluation metrics are provided for each model.

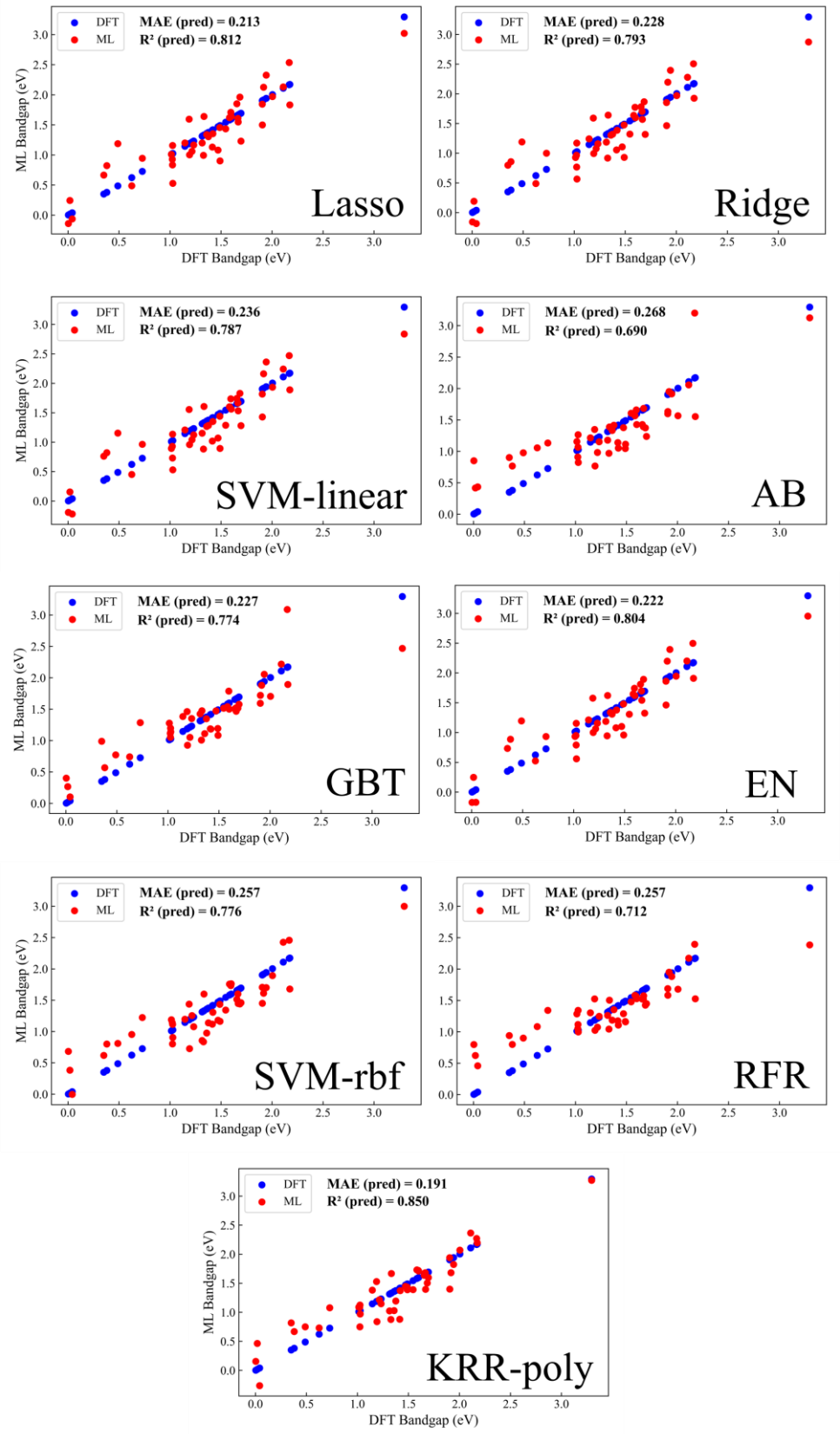

Figure C.19. Comparison of DFT and ML predicted E_g for validation set employing multiple models. The evaluation metrics are provided for each model.

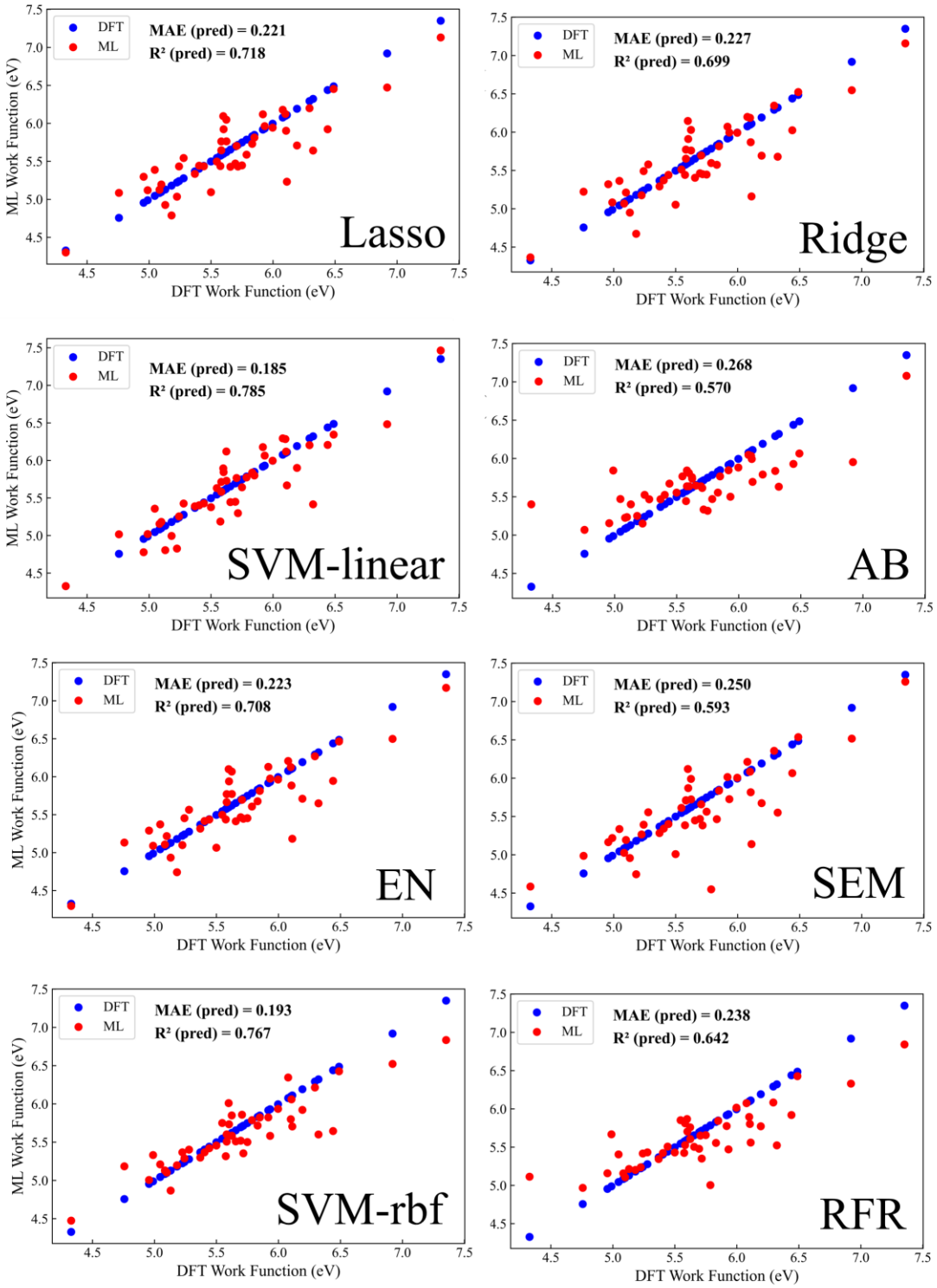

Figure C.20. Comparison of DFT and ML predicted Φ for validation set employing multiple models. The evaluation metrics are provided for each model.

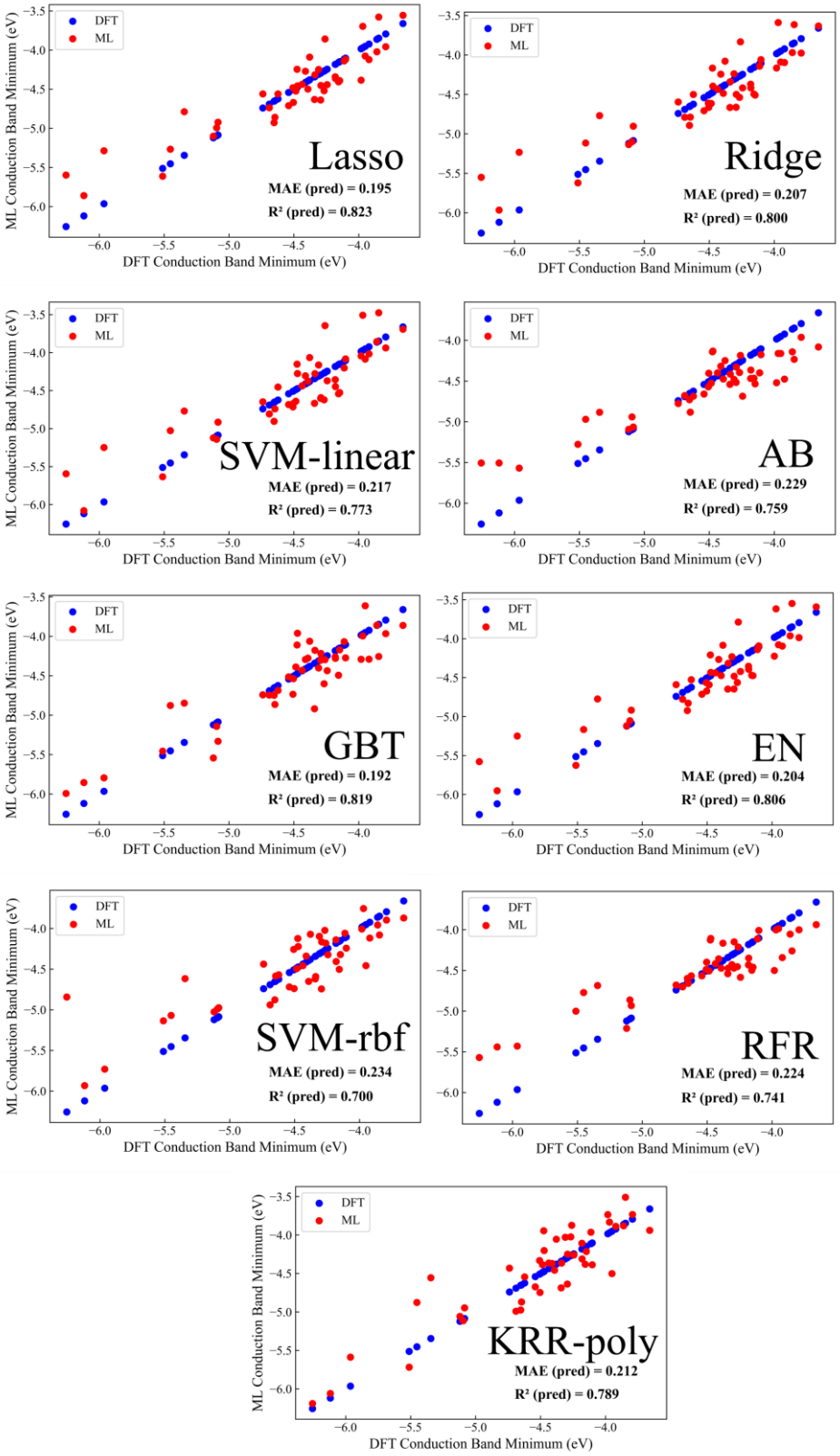

Figure C.21. Comparison of DFT and ML predicted E_{CBM} for validation set employing multiple models. The evaluation metrics are provided for each model.

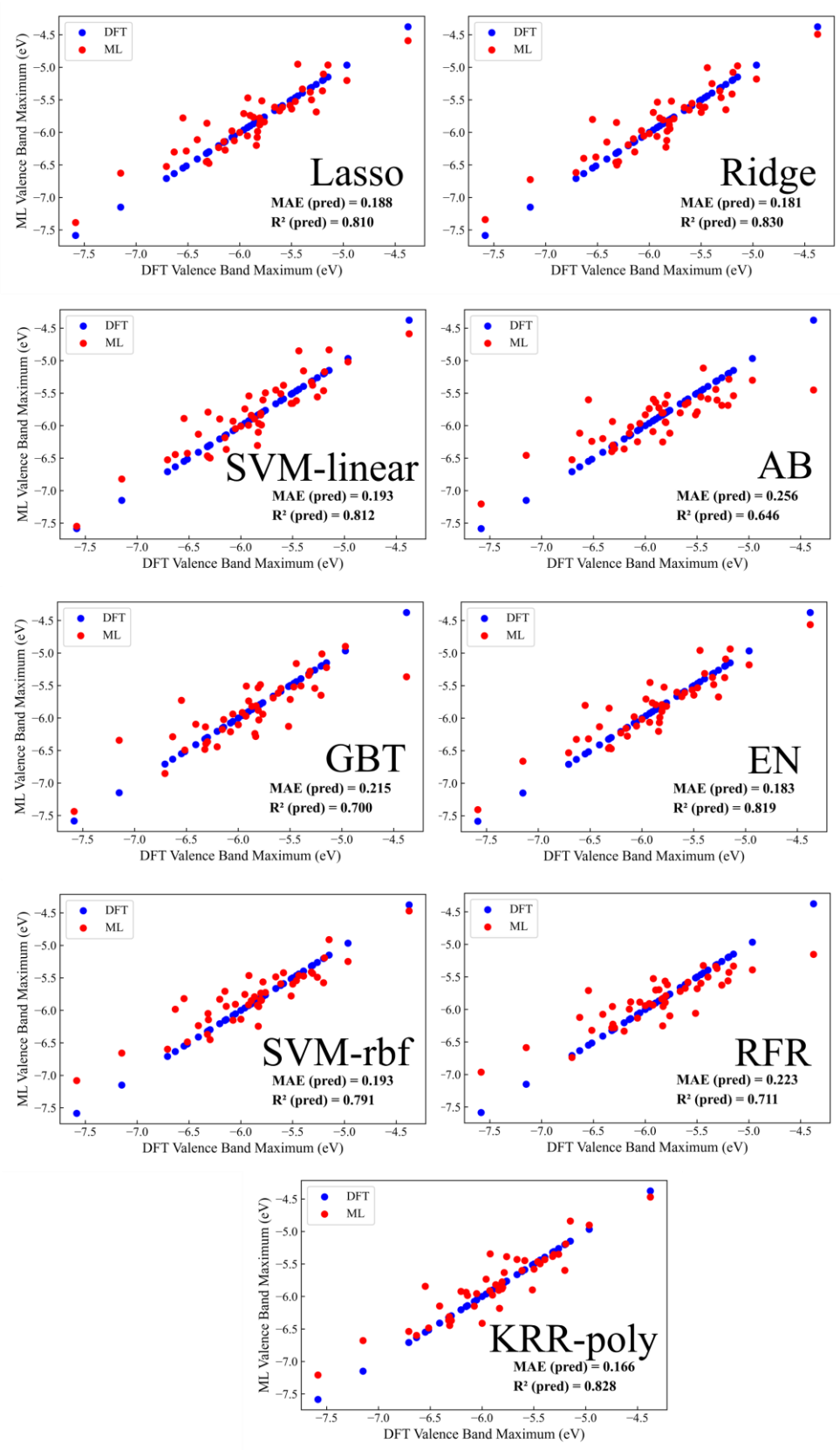

Figure C.22. Comparison of DFT and ML predicted E_{VBM} for validation set employing multiple models. The evaluation metrics are provided for each model.

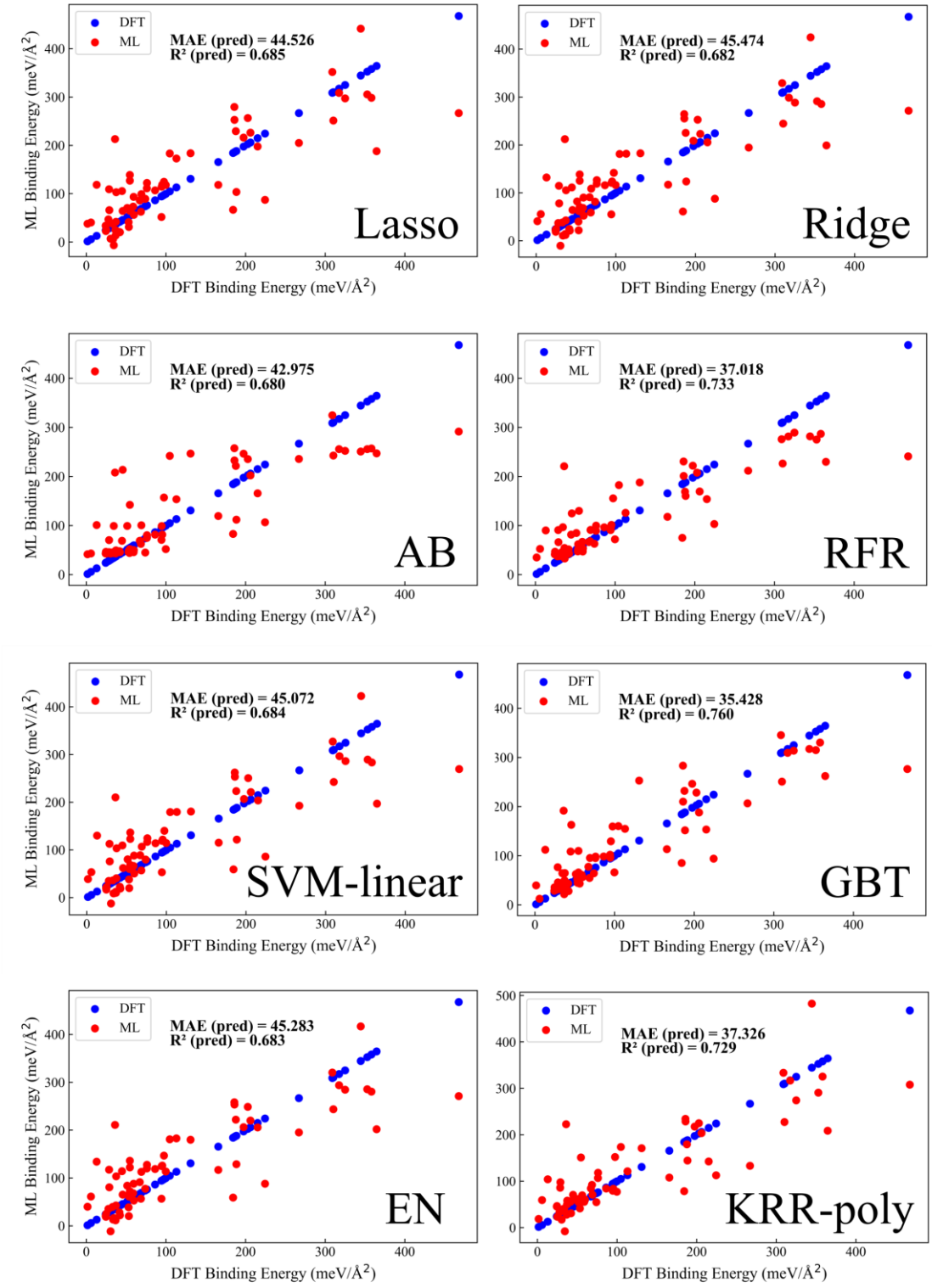

Figure C.23. Comparison of DFT and ML predicted E_b for validation set employing multiple models. The evaluation metrics are provided for each model.

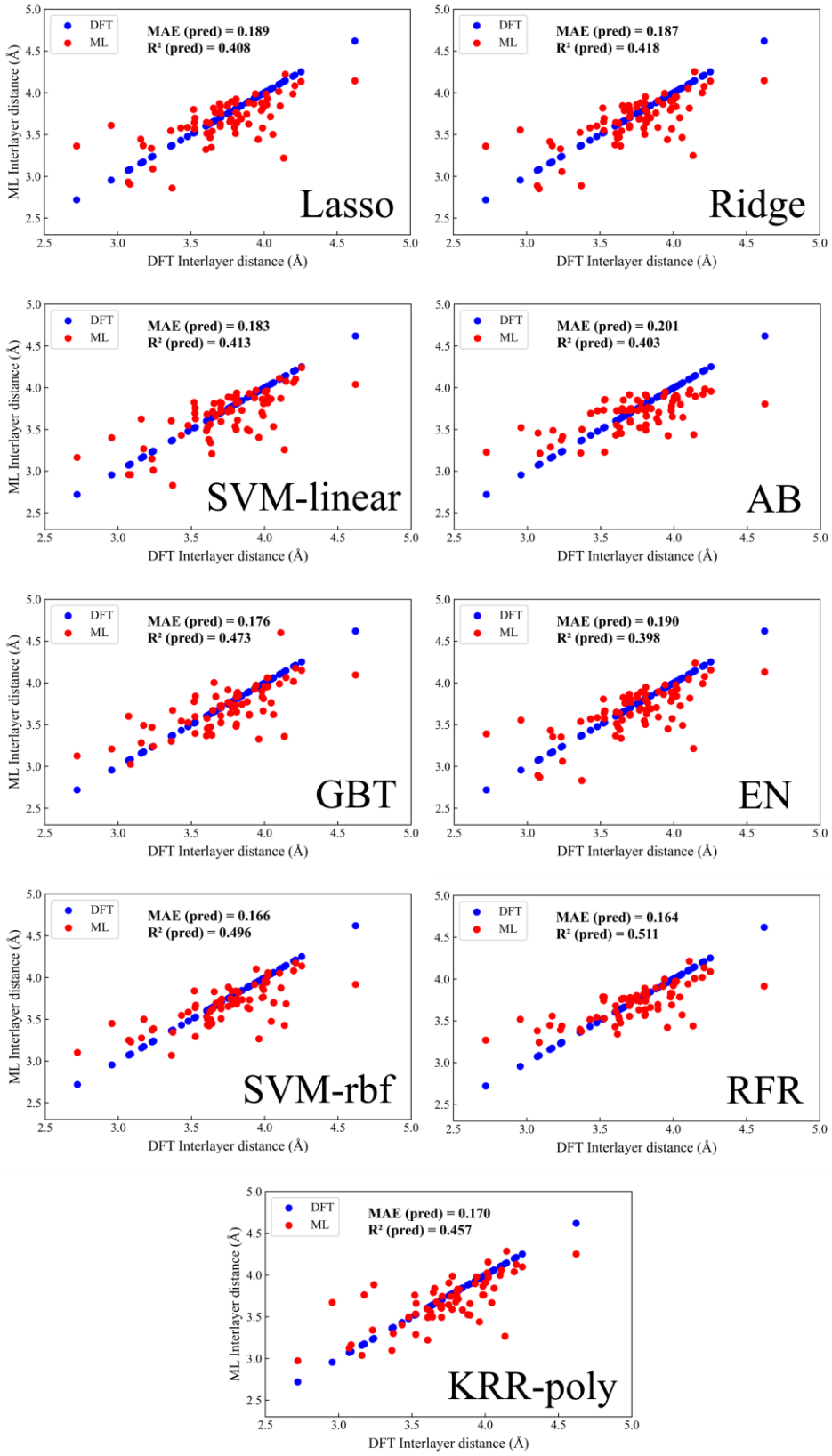

Figure C.24. Comparison of DFT and ML predicted d_o for validation set employing multiple models. The evaluation metrics are provided for each model.