NORTHWESTERN UNIVERSITY

Materials Discovery from Statistical Modeling and Atomistic Simulations

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Materials Science and Engineering

By

Abhijith M. Gopakumar

Evanston, IL
December 2022

ABSTRACT

Materials Discovery from Statistical Modeling
and Atomistic Simulations

Abhijith M. Gopakumar

Selecting the best material to deliver optimum performance in real-world applications is one of the most significant challenges in engineering. Hundreds of thousands of computationally-predicted, but experimentally unexplored materials exist today in the public inorganic material databases as candidates for consideration. This thesis discusses three projects in the domain of materials selection and discovery, and in each of them, one or more materials with a desired set of properties are identified from a large pool of candidates. The first work describes the computational discovery of three high-dielectric, high-bandgap materials within 17 selections from a set of more than 11,000 candidate materials obtained from the Open Quantum Materials Database (OQMD). We built statistical machine learning (ML) models from a sub-dataset of the Materials Project high throughput database to predict dielectric values along with the associated model uncertainty. The final material selections are made using a statistical optimization algorithm, and the final validations are done using expensive first-principles calculations to compute the dielectric properties. The second project details the identification of a new bridge material for MoS_2-based 2D electronic inks that acts as an adhesive between the 2D ink nanoparticles without interfering with the ink's electronic properties. This project uses a sequential selection workflow incorporating machine learning-aided high-throughput heuristic modeling to select the best material from a candidate set of more than 2000 materials, and subsequent estimation of the charge-transport properties from

expensive atomistic simulations. In the third project, we created a machine learning model that can identify the semiconductors and insulators which are misclassified from lower-accuracy Density Functional Theory (DFT) calculations to be metallic. The accuracy of bandgaps computed using DFT is dependent on the functionals chosen to describe the exchange-correlation energy of electron interactions. The PBE functional results in less accurate, but significantly cheaper estimations of the bandgaps compared to using the HSE hybrid functional. Our ML model predicts the bandgaps at an accuracy level of DFT-HSE at the cost of doing a cheaper DFT-PBE calculation. The reliability of ML predictions is analyzed from quantified model uncertainties and extensive literature surveys.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Prof Chris Wolverton for giving me a chance to pitch my research ideas during the APS March 2017, being my advisor for the next five years, and mentoring me on all the projects mentioned in this thesis from a scientist's point of view and also from a managerial point of view. It was great to plan, design, and execute the dielectrics project as my first doctoral research work. The side projects with OQMD were helpful in developing several personal and professional skills. There are a lot more, and thanks for everything.

Thanks to my lab mates from the Wolverton group for making life much easier at work with a strong collaborative and supportive nature all around. Especially the folks who helped with onboarding the group - Vinay, Mohan, Eric Isaacs, Xia, Shane, and Jiangang.

Thank you, Koushik, for all the collaborative work during multiple projects over the years. Bianca and Jiahong - both of you have been amazing lab mates and friends to me since the first year when we all joined the group.

Thanks to Prof. Mark Hersam for giving me an opportunity to work on the CHiMaD projects, and also to the graduate student collaborators who helped me with the projects - Lidia, Zhehao, and Sonal. Thanks to Zach Trautt and Laura Bartolo for all the help with OQMD, FAIR data, and everything related. Thanks to the folks at CECAM for collaborating to create OPTIMADE - especially Prof. Gian-Marco.

Thanks to Murat, Ph.D., Prof. Peter Voorhees, and Prof. Tobin Marks for being on my thesis and qualifier committees.

Now my folks at home:

Thanks a lot, mom and dad. None of these would have been even remotely possible if you folks hadn't prioritized your kids' education and happiness over everything else. And thanks for letting me and helping me choose my own paths in life and career all these years. Kanna, you have been a great brother all this time, and a great friend as well over the past many years.

Thanks to my amazing wife, Karen Schmitz, for helping me navigate grad school and research while also prioritizing the personal life - especially during tough times like COVID lockdowns. Also, thanks for all the help with the chores and food prep during the thesis-writing weeks.

Ali & hector - both of you are awesome friends, and also you've helped me on numerous occasions with grad school and life over the past 3-4 years. Thanks for everything.

Khyathi - I am so glad I talked to you about a small dataset in 2015. Thanks for introducing me to the world of machine learning and data science.

To Miss Archana, thanks for teaching me the fundamentals of coding in the right way. The foundation you laid in high school is still as strong as pristine monocrystalline graphene.

PUBLICATIONS LIST

- **Gopakumar, A.**, Pal, K. and Wolverton, C., 2022. Identification of high-dielectric constant compounds from statistical design. npj Computational Materials, 8(1), pp.1-10.

- **A. Gopakumar**, Pal, K., and C. Wolverton, "Identification and characterization of MoS2/GaS heterostructure as a novel 2D ink candidate", Manuscript in Preparation

- M. Liu, **A. Gopakumar**, V. Hegde, J. He, and C. Wolverton, "High-throughput Hybrid-functional DFT calculations of Bandgaps & formation energies and Multi-fidelity Learning with Uncertainty Quantification", Manuscript in Preparation

- Andersen, C.W., Armiento, R., Blokhin, E., Conduit, G.J., Dwaraknath, S., Evans, M.L., Fekete, Á., **Gopakumar, A.**, Gražulis, S., Merkys, A. and Mohamed, F., 2021. OPTIMADE, an API for exchanging materials data. Scientific data, 8(1), pp.1-10.
  (co-first author)

- Shen, J., Griesemer, S.D., **Gopakumar, A.**, Baldassarri, B., Saal, J.E., Aykol, M., Hegde, V.I. and Wolverton, C., 2022. Reflections on one million compounds in the open quantum materials database (OQMD). Journal of Physics: Materials, 5(3), p.031001.
  (co-first author)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION AND BACKGROUND

Material selection is one of the most influential subdomains of science and engineering that fuels innovation in all other fields of technology. Some of the greatest examples of material selections in the past centuries include the identification of tungsten as a good filament material for incandescent bulbs, steel as a better construction material, silicon as a suitable semiconductor, and copper as a good conductor - among numerous others. Before the 2010s, the mainstream approach to material selection had been via individualistic efforts to search through a limited set of materials that are known from literature surveys and localized databases. During the search for a new material to improve the performance of a given device, most of the computational research efforts went into finding prior data from published scientific literature, manually selecting a small set of materials to explore based on heuristic selection methods, and finally, doing material simulations to estimate the applicable material properties. This entire process has been repeated by several researchers for even the same property optimization problems, due to the limitations in reusing the resources such as data gathered by others.

In 2011, the Materials Genome Initiative (MGI)[1] triggered a new era in the scientific approach to the design and discovery of materials. MGI seeks to unify the different parts of materials research by combining the data infrastructure, computational tools, and experimental methods to accelerate materials discovery. Different parts of materials science may develop their tools and infrastructure independently but with a focus on smoothly passing the relevant knowledge to the next phase. Integrating isolated efforts in a robust infrastructure is expected to avoid delays in the materials discovery process. A good data infrastructure in place makes the old and new materials data more easily findable, accessible, interoperable, and reusable (FAIR)[2]. FAIR data enhances the manual search for individual materials and machine-actionability on large datasets. The scope and applications of MGI vary in implementation at different parts of material science, but all these related efforts are intended to contribute to the primary goal of accelerating materials discovery.

In computational materials sciences, the MGI put forward the philosophy of laying out future-proof infrastructure on large and small scales to systematically approach the materials selection with the goal of suggesting new advanced materials to experimentalists and manufacturers to consider. It includes the creation and active maintenance of large-scale material databases, automation of data access and processing, integration of statistical modeling and analysis to select materials, and validation of materials from simulation methods. Since the characterization of a material through experiments is significantly more expensive than doing the same via computational methods, the advantage of the latter comes through the capability of screening thousands or even millions of materials to select a smaller set of best candidates that the experimentalists can consider.

In this work, we focus on the discovery of inorganic crystalline materials. Some of the most prominent databases containing computational data of crystalline materials are Open Quantum Materials Database[3], [4] (OQMD), Materials Project[5], and AFLOW[6] among others[7]–[9]. These datasets hold chemical and physical property data of millions of materials generated from ab-initio atomistic simulations in a high-throughput manner. They can be used as a place to learn information on materials or to search for novel materials with enhanced properties. Most HT crystalline material databases hold information on the material's chemical composition, crystal structure, unit cell, electronic bandgap, and formation energy. Both bandgap and formation energy values are obtained from relatively cheaper-to-compute ab-initio simulations that can be automated. Even though these two properties can provide crucial knowledge into the material's electronic properties and stability, more complex and expensive simulations are often required to calculate other properties such as thermoelectric performance, photosensitivity, etc. There are smaller sub-datasets, often available from the same HT databases, containing data on more complex properties of hundreds to thousands of materials. But doing those expensive calculations on hundreds or thousands of materials becomes infeasible in most cases due to constraints on available resources. This is where statistical modeling becomes helpful in learning from the available data and then searching for better materials in the larger pool of materials with relatively much lower consumption of resources than atomistic simulations. Machine learning (ML) algorithms are among the most popular statistical modeling tools used today in materials science[10]–[15]. Several ML model architectures have been proposed over the years specifically for the domain of crystalline materials where the data is minimally available for learning most properties[11], [13], [16]. Since the ML models act like a black box when used to predict the properties with varying confidence levels in each material's prediction, other statistical methods, such as optimization algorithms, are used to aid the ML models in finalizing the material selection. The exact strategy needed in materials selection depends on how expensive are the subsequent validations. The materials selected from statistical modeling are validated further from more reliable atomistic simulations and eventually passed on to the experimentalists. Several quantum mechanical strategies exist today to simulate the atoms inside a crystal. Density functional theory[17] (DFT) is one of the most popular methods developed to do this job. DFT is used to simulate an unperturbed crystalline material at 0 K and calculate its stability, electronic band structure, optimal unit cell parameters, atomic positions, etc.

All of the projects mentioned in this thesis fit into the theme of discovering novel materials for specific applications using statistical modeling and atomistic simulations. We have used multiple HT databases as sources to train ML models and also to search for novel materials. In Chapter 3, a complete materials design workflow with multiple design cycles and knowledge feedback is implemented to identify novel high-dielectric materials that can improve the performance of batteries, memory devices, etc. In Chapter 4, a new 2D electronic ink material is identified from a single design cycle that also follows a robust materials discovery workflow involving HT databases and literature surveys, heuristic models aided by ML, and expensive characterization of material properties from ab-initio simulations. In Chapter 5, a specialized ML model is built to predict a material's bandgap with quantified prediction confidences with high accuracy at the expense of doing much cheaper, low-accuracy DFT calculations. In the end, we also discuss the outlook of this field with a focus on computational infrastructure and best practices for materials discovery.

CHAPTER 2
GENERAL METHODS AND FORMALISM

This section details the set of methods and theories used in multiple projects mentioned in this proposal.

2.1 Density Functional Theory

Density Functional Theory (DFT) is a computational modeling method used to calculate the properties of crystalline solids at 0 K. DFT computes crystal properties as a functional of the electron density, $\rho(r)$. The fundamental advantage of using DFT, as opposed to computing the full Hamiltonian for an N-electron system, is that the former reduces the dimensionality of the problem from $3N$ to 3 - making it scalable to systems with many electrons. In DFT, the Schrodinger equation for interacting many-electron system is modified to Kohn-Sham equations[17], [18] by expressing the energy as a functional of electron density in a system containing non-interacting pseudoparticles, as shown in equation 2.1.

$$E(\rho) = T_s(\rho) + \int dr\ V_{ext}(r)\rho(r) + E_H(\rho) + E_{xc}(\rho) \tag{2.1}$$

V_{ext}, T_s, and E_H denote the external potential due to the presence of positive charges in the nucleus, the kinetic energy functional, and the classical electron-electron interaction, respectively. The last part of the equation, $E_{xc}(\rho)$, represents the exchange and correlation (XC) energy functional which contains energy from all other many-electron interactions such as the interaction between electrons with opposite spins. Equation 2.1 is a complete re-formulation of the full Hamiltonian without any approximations, and solving either of them would give the exact same quantified values for material properties - including the ground state energy. But the correct functional form of $E_{xc}(\rho)$ is unknown even though it is independent of the system under consideration.

In practice, different forms of $E_{xc}(\rho)$ functional are approximated in DFT implementations with varying accuracy. The most popular XC approximations are localized density approximation (LDA) and generalized gradient approximation (GGA). In 1996, Perdew et al.[19] implemented the most used version of GGA, called Perdew-Burke-Ernzerhof functional[19] (PBE). But neither of them accounts for the discontinuities in the interaction potential; thus, they cannot find the exact ground state electron density. To further increase the accuracy of XC approximations, Heyd et al.[20] proposed a new version of hybrid functional approximations called Heyd-Scuseria-Ernzerhof (HSE) functionals. The HSE often calculates bandgap and electronic properties with higher accuracy than PBE but at about 100 times the additional computational cost. Because of that, the usage of HSE is limited to cases where the accuracy of PBE does not suffice to make a final decision. Vienna Ab-initio Simulation Package[21]–[23] (VASP) is one of the most popular software available to do DFT on materials and estimate properties. In this work, we use VASP with Projector Augmented Wave[24], [25] (PAW) potentials and PBE exchange-correlation functionals to do atomistic calculations. Additionally, we also make use of a high-throughput dataset generated by other researchers using HSE functionals to build machine learning models.

2.2 Machine Learning

Machine learning (ML) refers to the process of detecting patterns in datasets. A dataset consists of information about multiple objects - like materials. ML models learn the underlying function that maps a set of variables, called features, that are cheap to calculate and can represent a given object to another variable or a set of variables, called targets, that are often hard to estimate. This mapping between features and target(s) is initially unknown to the ML model, but it learns from the available data to best represent the underlying function with the least amount of averaged error, while also leaving room to tolerate the imperfections in data. The set of data objects whose both features and targets are initially known is called training data because this set is used to train the ML model. Once the model is trained, it is used to predict the target values of objects whose features are known, but the targets are not. The choice of features can highly influence how much information the ML model can learn. Since this is where the model searches to find new, useful data objects, it is called as the search space. As much easily-available data about the object must be encoded in the feature vector. But a large feature vector with a small number of data objects in training data can lead to underfitting of the ML model where the underlying function is too complex to be learned from the limited number of data objects. Similar to underfitting, the model can also overfit on the data where it maps a function so complex that it fits the training data perfectly without considering the possible errors during the data generation. Overfitting the data leads to large errors while predicting because of the unpredictability of the complex mapping function in unsampled areas of the feature vector space.

Different ML algorithms learn the internal mapping of features to target differently. For example, tree-based algorithms construct a set of if-else conditions, the support vector machines algorithm constructs high-dimensional hyperplanes in the feature vector space, Gaussian Process models perform bayesian inference and other probabilistic methods, and artificial neural networks create a network of linear models arranged in specialized configurations to allow approximation of complex functions. Each ML model has a different set of internal parameters to build the mapping function and a set of hyperparameters describing the model's high-level architecture. The internal parameters are learned automatically during the training process, while the hyperparameters are to be optimized either manually or from other external methods. Overall, a large number of choices can influence how well the ML model learns the mapping function, including the number of training data objects, feature vector selection, feature vector size, ML model, model hyperparameters, quality of training data, the complexity of underlying function, etc. To avoid the overfitting of data, a fraction - often 10% to 30% - of the training data is set aside and not used for training. This subset of the full training data is called the test data. Since the real target values are known in test data as well, the target values predicted by the trained ML model on the test data are compared against the real target values to assess the performance of ML models reliably predicting the target values.

Once the model is built and trained, it is used to predict target properties in the search space. But since the real target properties of search space objects are unknown, the model may produce inaccurate predictions when the search-space is much larger than the training data, even if it performed well on the test data. This happens due to the limited sampling of the feature vector space by the training data, while the objects in a larger search space may belong to an undersampled region in the feature space. One way to solve this uncertainty in accuracy is to quantify the model's prediction uncertainty. Some ML models, such as Gaussian processes, generate the quantified uncertainty inherently based on how close a new data object lies to the training data objects in the high-dimensional feature space. In other ML models, this uncertainty can be quantified by creating a statistical distribution of predictions from an ensemble of models trained with different initial parameters and possible different subsets of training data. The predicted distribution's standard deviation can be assumed as the prediction uncertainty.

In materials science, the data objects are often the materials. The features representing the materials can be generated in different ways by considering the material's composition or structure or both. It is also possible to use some of the material's properties derived from cheaper simulations as extra features, as long as they are readily available for both training data and search space materials. For example, bandgap and formation energy values are often available for millions of materials from high throughput databases. Training data in materials science can be obtained as sub-datasets of HT databases or from literature surveys and mini-throughput sets of simulations. The search space can also be defined similarly. There are ML algorithms and architectures created specifically to handle the material datasets, such as CGCNN[11] and SchNet[16]

In this work, various ML models are created and/or used in each project depending on the amount of available data and the nature of the problem at hand. We also use uncertainty quantification in all models since that will help in assessing prediction confidence for either avoiding less-reliable predictions (as in Chapter 5) or to explore undersampled feature spaces (as in chapter 3). We use data from HT databases like OQMD and Materials Project to create training data and search spaces. The ML models and their associated pre- and post-processing data pipelines are built on open-sourced libraries written in the Python programming language - such as Tensorflow[26], Keras[27], Scikit-learn[28], and OpenMDAO[29].

CHAPTER 3
DISCOVERY OF HIGH-DIELECTRIC CONSTANT COMPOUNDS IN RARE-EARTH FAMILIES FROM STATISTICAL OPTIMIZATION

3.1 Background

Dielectric materials are insulators that polarize in the presence of an external electric field. Because of their ability to store electric charges near their surfaces, dielectrics are used extensively in electronic devices such as Central Processing Units (CPUs), Dynamic Random-Access Memory (DRAM), and capacitor-based batteries. The efficiency of a dielectric material is measured in terms of the dielectric constant (ϵ). ϵ is the factor by which the electric field induced by a finite electric charge is reduced inside a dielectric material when compared to the electric field generated by the same charge in the vacuum. Hence, the materials with higher ϵ values are most often preferred for dielectric applications. Dielectric breakdown happens due to the carriers in the material's valence band crossing the electronic bandgap (E$_g$) under an external electric field, eventually resulting in leakage currents and loss of the stored charge. The maximum threshold for charge storage due to leakage currents significantly limits the usability of a given dielectric when the material has a small E$_g$. Thus, an excellent dielectric material should have a high ϵ and a large E$_g$. Dielectrics that are being used nowadays for commercial applications have their ϵ between 20 and 30 - for example, Ta$_2$O$_5$ ($\epsilon \sim$ 23-27, E_{g}=4.2 eV)[30]–[33] and TiO$_2$ (ϵ=27, E_{g}=3.5 eV)[30], [32], [34].

It is observed that the E$_g$ and ϵ are inversely proportional to each other[32], [35], making it a challenge to discover materials with large values for both E$_g$ and ϵ. Performing a high-throughput computational screening across tens of thousands of possible compounds is not feasible because the underlying simulations based on Density Functional Perturbation Theory (DFPT) to accurately calculate the total ϵ value, which is the sum of both ionic and electronic contributions, are significantly more expensive than standard DFT calculations. The primary goal of this work is to screen the large, existing high-throughput inorganic crystal structure databases and discover novel high-ϵ dielectrics with large E_g by conducting as few DFPT calculations as possible.

3.2 Results and Discussion

3.2.1 Materials design strategy

This work focuses on finding materials with optimized dielectric properties and large bandgaps. There are nine components in a material's dielectric tensor, and it will be a significantly challenging task to optimize more than a single dielectric component simultaneously. Thus, we define the target property for optimization as the largest eigenvalue of the dielectric tensor and refer to it as the ϵ. The largest eigenvalue of the total dielectric tensor is chosen here as opposed to the average eigenvalue to account for the highest dielectric nature the material will exhibit when it can be aligned perfectly in any necessary direction while making an electronic device. The total dielectric tensor is the sum of the electronic and ionic dielectric tensors obtained from DFPT calculations. As shown in Figure 3.1, each cycle in the materials design workflow implemented in this work has three steps - the collection of data, selection of materials from statistical modeling, and validation of the new materials' properties using DFPT. During the first step, the training-data and a search-space are determined. Training-data is the dataset of materials with known values for ϵ, and it is used to train the machine learning models. Search-space is the dataset of materials whose ϵ values are not known, and this is where we search for novel high-ϵ materials using machine-learning model predictions. During the second step of the workflow, reliable machine learning (ML) models are created by learning knowledge from the training-data, and then the trained models

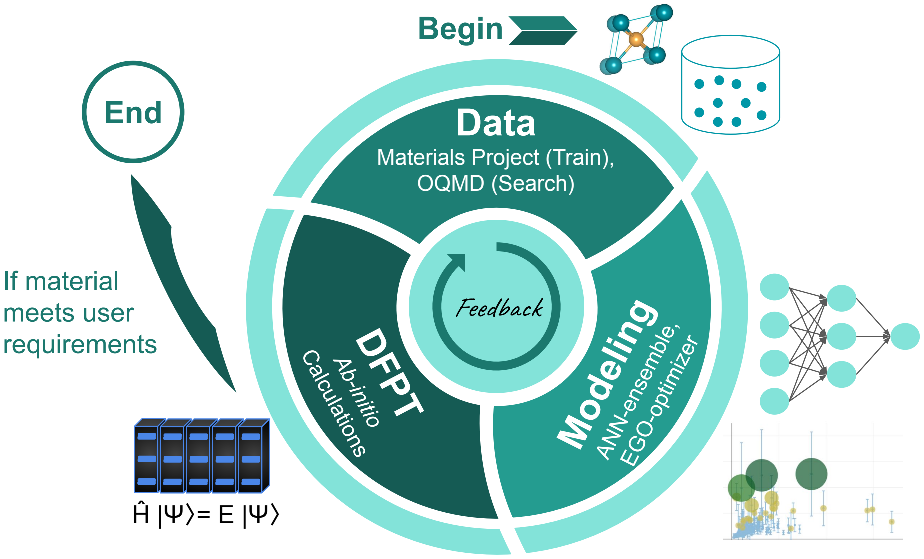


Figure 3.1: Workflow for materials design implemented in this work. A single design cycle consists of three stages - Data, Modeling, and Validation. In the first stage, the design challenge is translated into a set of datasets. Since the E_g values are readily available from high-throughput datasets, we consider only the ϵ as the unknown target property. The set of materials from the Materials Project with already known ϵ values is set as the training-data. The set of stable non-metals from OQMD whose ϵ value is not calculated yet is defined as the search-space to find the best candidates for being novel high-dielectrics. The methods to generate feature vectors representing each material in training-data and the search-space are also established in this stage. During the Modeling stage, we created an ensemble of Artificial Neural Networks (ANNs) trained on the randomly sampled subsets of training-data to predict the ϵ value of materials when their structure, chemical composition, and E_g are known. Since each ANN model predicts a single ϵ value, the ensemble with 2000 independent ANN models predicts a distribution for the ϵ value for each material in the search space. This predicted ϵ distribution is used by Efficient Global Optimization[36] (EGO) to rank the materials based on how likely the selection of that material will lead to the eventual optimization of ϵ within as few design cycles as possible. The EGO algorithm can exploit the modeling power of ANNs while also exploring the classes of materials in the search space that are not well represented in the training-data. In this work, five to seven materials are selected from EGO-based rankings in each design cycle and passed on to the next stage - Validation. During validation, the ϵ values of the selected materials are calculated using the ab-initio DFPT calculations. The design workflow ends at this stage if the design goals based on the required material property optimization are met. Otherwise, the newly generated DFPT data is fed into the next cycle by appending it to the training-data, and a new design cycle is started. We conducted three such design cycles until we identified multiple high-ϵ, high=E_g materials.

are used to predict the ϵ values of all materials in search-space with quantified model uncertainty. Further, a statistical optimization-based selection algorithm is used to select a few materials from the search-space as the best candidates for DFPT validation based on their ML-predicted ϵ data. In the workflow cycle's final step, the selected materials' dielectric properties are computed using high-accuracy DFPT methods. The newly gained knowledge of ϵ on the selected candidates is appended to the training-data for the next design cycle, and those materials are removed from the search-space. This knowledge feedback is expected to improve the ML modeling in the subsequent design cycle due to the increased sampling of materials space by the new materials in the training-data. The workflow cycles are continued sequentially until a few materials with high values for ϵ and E$_g$ are found.

3.2.2 Datasets

The initial training-data in this work consists of 1864 materials whose dielectric values were available from Materials Project[37], [38] (MP). The search-space consists of 11,102 stable non-metallic materials from OQMD. Since both MP and OQMD already contain data of DFT-estimated bandgaps, the available information for both training-data and search-space consists of the crystal structure, chemical composition, and bandgap energy values, while the training-data additionally contains ϵ as the target property. It is required to represent all the materials in vectors of the same length in order to use them in statistical modeling. We generated such material representations using Magpie[39] crystal property generator tool which inputs the crystal structure and chemical composition of a material and generates a vector of length 271. The bandgap value is appended to the vector, making the final representation vector, also called the feature vector, length to be 272. This feature vector size is further reduced to 100 by using some of the popular feature-reduction methods, such as principal component analysis and model-based feature selections - as implemented in the Scikit-Learn python library. Since we use data from separate material databases during the course of the statistical design by training the model on MP data and then using the trained model to predict on the materials from OQMD, it is crucial to consider the possible differences in the representation of the same material across databases. The main incompatibility between databases may stem from the differences in DFT calculation parameters used to relax the material's crystal structure in the databases' high throughput calculation frameworks. We investigated the similarity between Magpie-generated feature vectors for equivalent materials in OQMD and MP. In total, 1717 materials out of the 1864 materials in MP training-data had an equivalent material phase entry in OQMD as well. A material from MP is considered equivalent to a material in OQMD if both of them have the same chemical composition and crystal structure symmetry spacegroup. The results are shown in Figure 3.4. Figure 3.4a, shows that there is a negligible ($< 2\%$) difference between MP and OQMD materials in 263 out of 271 Magpie features. The rest of the 8 features also have relatively small ($< 7\%$) deviations from each other. In Figure 3.4b, the Bandgap values, which join the Magpie vector to form the final feature vector, are compared across MP and OQMD for the same 1717 equivalent materials. It also shows a negligible mean difference (0.1 eV) and a median absolute deviation (0.0 eV) between the two datasets. Hence, it is shown that the cross-database statistical design is significantly less likely to suffer from database incompatibilities in our case.

In the MP training-data, most of the ϵ values are between 0 and 25 with only less than 5% materials with ϵ above 50 - as shown in Figure 3.2. The mean, median, and standard deviation of the MP data are 20.2, 12.2, and 42.8, respectively. This bias in the training-data toward small ϵ values makes the statistical modeling prone to model fitting bias and treats the large values as rare outliers. One method to avoid the fitting bias is to reduce the numerical spacing between the target values, which avoids steep changes in the model's internal parameters such as the weights

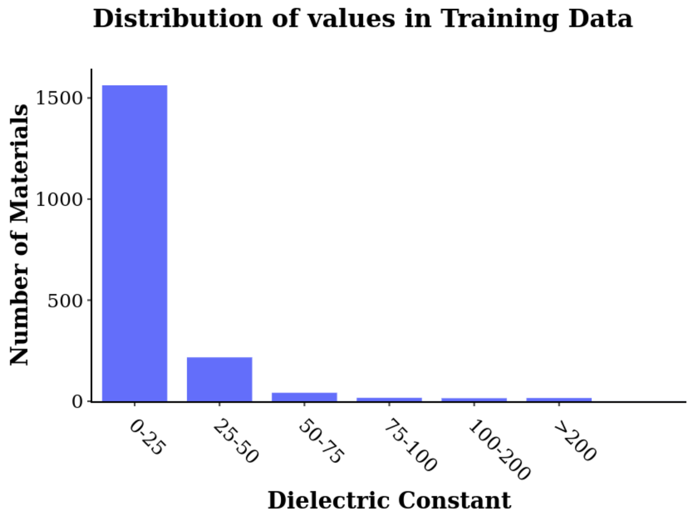
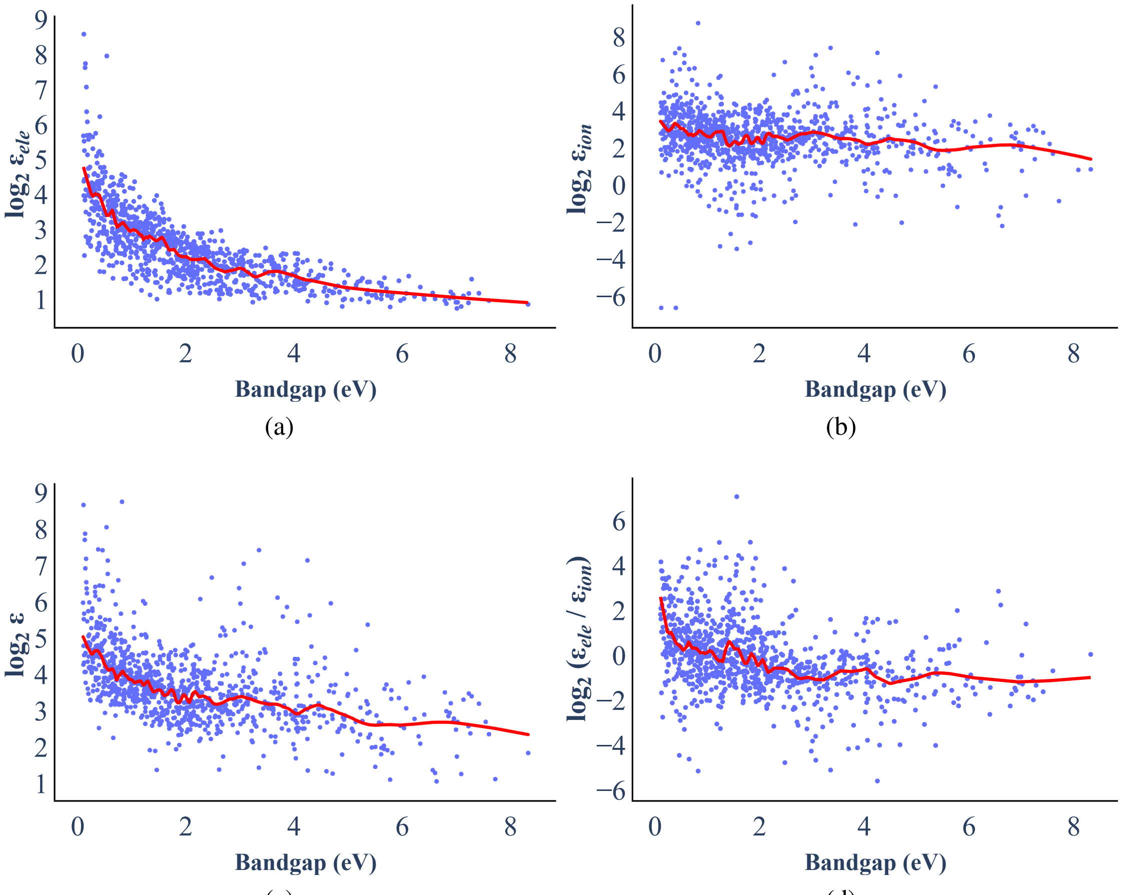


Figure 3.2: Distribution of ϵ values in the training-data. More than 95% of the values are below 50. So it is probable that most of the feature-vector space spanned by materials with large ϵ values are not well sampled within the training-data. Theoretically, exploratory materials design with knowledge feedback, as implemented in this work, is expected to perform very well even in such unevenly sampled data scenarios.

and biases in an ANN during the training. In this work, we did a log-scale transformation of the ϵ values to reduce the numerical spread. We also analyzed the trend of how ϵ and its components change with E_g in the training-data and are given in Figure 3.3. The ionic contribution to the ϵ is primarily independent of the bandgap, while the electronic contribution has a strong inverse correlation with E_g, as expected.

3.2.3 Statistical Modeling and Optimization

We created an ensemble of artificial neural network (ANN) models to learn from the data and trained each of them on a randomly sampled subset of the full training-data. The reason to choose an ensemble, instead of building a single ANN, is to quantify the model uncertainties associated with the ANN predictions of ϵ for each material in the search-space. When the training-data is



(a) (b) (c) (d)

Figure 3.3: The plots show the trends of how the electronic (ϵ_{ele}), ionic (ϵ_{ion}), ϵ (highest eigenvalue of total dielectric tensor), and $\frac{\epsilon_{ele}}{\epsilon_{ion}}$ values change with increasing bandgap value in training-data. The total dielectric tensor is obtained by tensor addition of ionic and electronic dielectric tensors. While ϵ_{ele} decays inversely with the bandgap (Subplot **a**), we find that ϵ_{ion} shows weak dependence on the bandgap (Subplot **b**). However, ϵ is dominated by the electronic contribution in the low bandgap regime, hence exhibiting an inverse relationship with the bandgap (Subplot **c**). Subplot **c** also shows that the inverse relation between ϵ and bandgap weakens at high bandgap limits because the ϵ_{ion} is the dominant factor in that range, as shown in subplot **d** that plots $\frac{\epsilon_{ele}}{\epsilon_{ion}}$. Hence, the generalization of this trend (ϵ vs. bandgap) in high-bandgap regions is less reliable due to the sparse population of data points, unlike the low-bandgap limit. The trend line is plotted using Locally Weighted Scatterplot Smoothing[40] (LOWESS) algorithm as implemented in the Plotly python package[41] with 5% of the dataset considered near each point to compute the local weights.

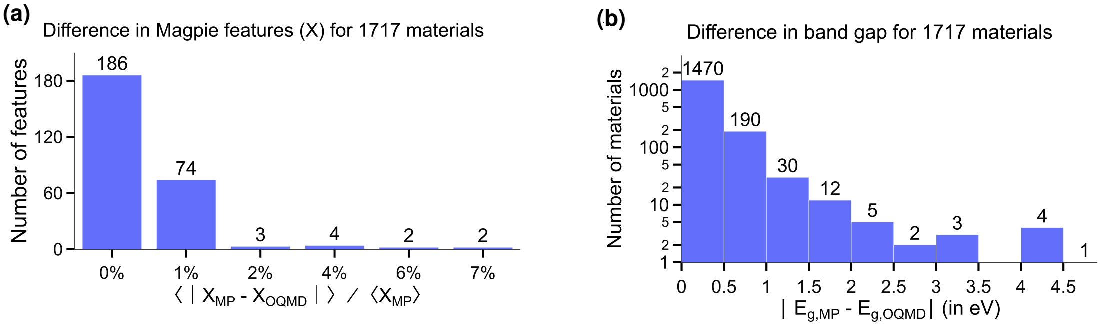


Figure 3.4: **Variance in materials data between the OQMD and MP databases.** Difference in the material representation vector components of the structures obtained from OQMD and MP for 1717 materials in the training data. **a)** Mean absolute difference in the 272 feature vectors generated from Magpie on the structures obtained from the MP and OQMD for 1717 materials in the training-data. The cross-referencing of materials across the databases was done by finding the database entries with the common ICSD Collection codes associated with their structures. We also made sure that all the 1717 materials have the same symmetry group listed in MP and OQMD, even though their lattice parameters may differ by a small amount due to differences in the DFT parameters and initial states used to generate the data. For 143 materials in the MP training-data, ICSD Collection Codes were unavailable. So they are not included in this analysis even though their counterparts in OQMD could have been found by matching composition and crystal symmetries. This strict mapping via ICSD codes helps in avoiding the materials that have changed the spacegroups after DFT relaxation. **b)** Distribution of the difference between the DFT bandgap (E_g) values listed in OQMD and MP in log scale for the same 1717 materials in the training-data. The bandgap values show very good agreement in most cases.

small compared to the candidate space, it may have under-sampled the high-dimensional vector space spanned by the search-space materials. For example, a training-dataset consisting of only organic compounds cannot reliably represent the feature-vector space that may have been spanned by inorganic compounds. In this project, a single ANN would still predict an ϵ value for even the materials that do not belong to the part of the feature space spanned by the MP training-data without any information on how confident it is about that particular prediction. Hence, quantifying the model uncertainty helps assess a prediction's reliability and explore new areas within the feature space of materials that are not sampled by the current training-data. Our model ensemble comprising 2000 independent ANNs provides a statistically relevant distribution of the predicted ϵ for each material in the search space. The standard deviation of this predicted ϵ distribution is defined as the quantified uncertainty for that material. Since the new DFPT-calculated data is added to the training-data after each design cycle, it increases the sampling range of the training-data in feature space. Because of this feedback, a new ANN ensemble is created and trained at every design cycle. Figure 3.6a shows the validation results from a randomly selected model from the second design cycle's ANN ensemble.

To assess the ANN predictions and rank the search space materials as potential candidates for DFPT calculations, a statistical optimization algorithm called Efficient Global Optimization (EGO) is used in this work. EGO takes in the mean and standard deviation of the ϵ distribution for each material and estimates a quantity called Expected Improvement, denoted as $E(I)$. The EGO algorithm does the exploitation of the available information in training-data by considering the search space materials with large mean values for their ϵ distribution, and also explores the unsampled regions of feature vector space by considering the search space materials whose ϵ distribution has a large standard deviation. The optimization that EGO looks for in this work is the maximization of the ϵ value among search space materials within as few design cycles as possible. Theoretically, the $E(I)$ value of a material is the probability that the DFPT-estimation of that particular material's ϵ would lead to the discovery of a high-ϵ material in the current or subsequent design cycles. Figure 3.6b shows the $E(I)$ values computed by EGO for the validation data in the second design cycle. An example illustration of the calculation of $E(I)$ is provided in the next section.

3.2.4 Calculation of E(I) in via an Example

Consider a set of materials M_i that belong to the search space, each with a ϵ distribution associated with it. Let y_t^{max} be the largest value for ϵ in the training-data. If we trust the ML model about its prediction capabilities completely, the best candidate for selection is the material, say M_1, with the largest means and lowest standard deviations for their ϵ distribution. That also means that if material M_2 has a mean ϵ lower than y_t^{max}, it can be discarded completely even if it has a large uncertainty associated with it. This is called the exploitation of the available knowledge by trusting both the model and the capability of training-data in reliably sampling the material feature vector space. But in reality, the training-data does not sample the feature space well, and some of the materials may have a smaller ϵ distribution mean just because the ANN models could not arrive at a confident prediction. In this case, the material M_2 is still of interest even with a smaller ϵ mean value since it has a larger uncertainty and may belong to a new class of materials that are not sampled in the training-data. This is called the exploration of the feature space. In the EGO algorithm, both exploration of the known data and the exploration of the unknown spaces are considered while ranking the materials for selection. Thus, both M_1 and M_2 are assigned a higher rank in the list of possible candidates for selection. The ranking is done by computing the $E(I)$ value as shown in Figure 3.5, and as follows:

Let Y represent the target property that is to be maximized in a given situation. In this work, Y is the dielectric constant of a material. Now, $\varphi(Y)$ is the predicted distribution of Y from the ANN ensemble. $\varphi(Y = y)$ is the probability estimated by the ANN ensemble that the real value of Y is the numerical value y. The expected improvement is calculated based on how much improvement a search space material promises above the threshold set for what is to be considered as the improvement. In the benchmarking studies of the EGO algorithm, this minimum threshold for improvement is set by the best material in the training-data. It is the largest value of the target property in the training-data, denoted as y_t^{max}. As formulated and benchmarked on surrogate models by Jones et al.[36], the EGO algorithm computes the expected improvement, $E(I)$, as

$$E(I) = \int_{y_t^{max}}^{\infty} (y - y_t^{max})\ \varphi(Y = y)\ dy \tag{3.1}$$

Balachandran *et al.* [10] benchmarked this optimization algorithm on material datasets with an approximation of the predicted distribution as a normal (i.e., Gaussian) distribution with a mean μ and a standard deviation σ. In this approximation, Equation 3.1 can be re-written as,

$$E(I) = \sigma[\phi(z) + z\Phi(z)] \tag{3.2}$$

where, $z = \frac{\mu - y_t^{max}}{\sigma}$, ϕ is the probability density function, and Φ is the cumulative distribution function[36] of the normal distribution, $\varphi(Y)$.

In the MP-dataset used in this work, the largest ϵ value belongs to TiO_2 with ϵ=988 and E_{g}=1.8 eV. But in this work, we search for high-ϵ materials with a large E_g as well. Setting the y_t^{max} value to be 988, following the theoretical strategies, will focus only on very low E_g materials, which often have very high ϵ values. For practical purposes, any new material with $epsilon$ larger than commonly used dielectrics may be considered as important as long as their bandgaps are also large enough. hence, the minimum threshold of ϵ is assigned to be 100.0 to represent our high-ϵ value preferences. With that assigned, any material whose ϵ prediction distribution extends above 100 would have a finite $E(I)$ value, even if the mean of the distribution falls below 100. The ANN ensemble is not expected to produce a Gaussian distribution for ϵ even though the EGO algorithm expects one. Instead, the mean and standard deviation of the ANN ensemble-predicted ϵ distribution is used as the mean and standard deviation of a normal distribution of ϵ assigned for each material.

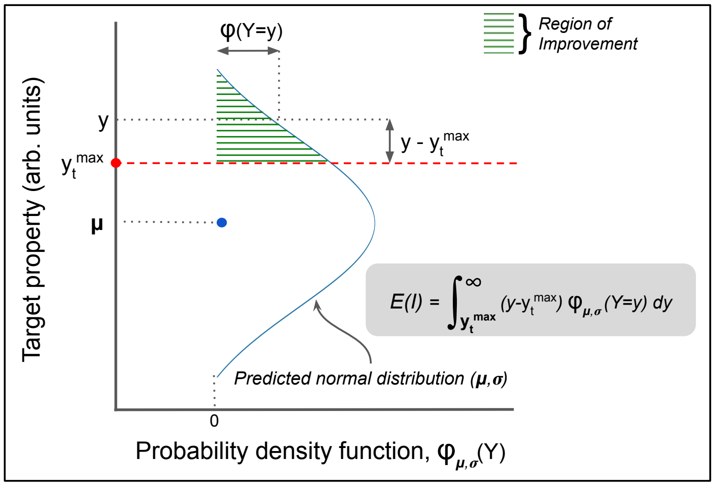

$\varphi(Y=y)$
Region of Improvement
Target property (arb. units)
y
y_t^{max}
$y - y_t^{max}$
μ
$E(I) = \int_{y_t^{max}}^{\infty} (y-y_t^{max})\, \varphi_{\mu,\sigma}(Y=y)\, dy$
Predicted normal distribution (μ,σ)
0
Probability density function, $\varphi_{\mu,\sigma}(Y)$

Figure 3.5: **The optimization algorithm.** The value y_t^{max} represents the currently available highest value of ϵ among all materials in the training-data. μ and σ represent the mean and standard deviation of the ϵ distribution for a material (blue dot) in the search-space predicted by ANN-ensemble. The predicted distribution is assumed as a normalized Gaussian function here. The region above y_t^{max} covered in green stripes represents the region of improvement. It is because, if the validation from the DFPT calculation determines that the material's ϵ lies above the minimum threshold that is y_t^{max} in this diagram, it will be considered as an improvement in property optimization among materials.

3.2.5 Design Cycles

In total, we conducted three design cycles sequentially, with 5 to 7 materials selected for DFPT calculations at the end of each of them. The training-data increased by that same amount since the new DFPT-calculated data from the previous cycle was appended to the new training-data. The feedback of data is expected to increase the confidence of ANN model predictions in the next design cycle due to the increased sampling of the feature space. In addition to this knowledge feedback mechanism between the cycles, the other main factor that influenced the selection process is the bandgap minimum cutoff imposed on the search space data. During the first design cycle, the only restriction was to have a non-zero bandgap, and it included all of the 11,102 stable non-metals from OQMD. But it resulted in the selection of very low bandgap materials since the statistical

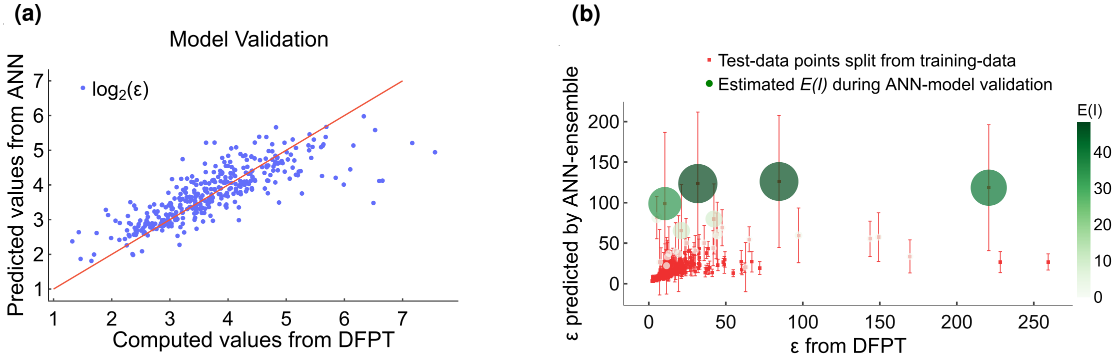


Figure 3.6: **Results from statistical modeling.** **(a)** ANN model validation on a test set of 373 materials split from the training-data. This particular model-fit plot is taken from a single ANN model that was part of the ensemble containing 2000 ANNs in the second design cycle. The ensemble was used, instead of a single ANN, to generate a distribution of ϵ for each material and quantify the uncertainty in predictions for each material in the search space. The 373 materials plotted here are part of the full MP training-data, and were not seen by this particular ANN model at any stage during the training. These predictions are made only to show this particular ANN model's learning capabilities, and they are not used anywhere during the actual selection. In the design workflow, each ANN model in the ensemble is exposed only to a unique subset of the full MP training-data, excluding 373 randomly chosen materials. The trained ANN models are used to predict the dielectric values of only the search-space materials from OQMD, not of the 373 materials split from the MP dataset, and used as test-data during ML model training. All ANNs in this work are trained to predict $\log_2(\epsilon)$ instead of ϵ, because the latter values are highly non-uniform in the training-data with most of the values below 25, making some of the very large values outliers. A log-scale transformation of ϵ reduces the numerical scale of spread among the ϵ values, making the very large values less of an outlier. The model fit shown in this plot has an R^2 score of 70%, and a Spearman's rank correlation of 85%. **(b)** This plot shows the predicted ϵ-distributions and corresponding $E(I)$ values on the same test dataset that in Subplot **(a)** consisting of 373 materials split from the training-data. The error bars represent the standard deviation in ANN-ensemble predictions. This standard deviation is quantified as the uncertainty of ANN predictions. For a clearer perspective, both the radius and color of the circles represent the same quantity - the Expected Improvement, $E(I)$. $E(I)$ value is calculated from ANN model predictions with uncertainty using the EGO algorithm. A point without an outer circle around it represents a material with a negligible ($< 10^{-3}$) value for $E(I)$. In this figure, only 25 materials have an $E(I)$ value that is greater than 10^{-3}.

modeling can get biased toward the inverse correlation of bandgap and ϵ. Thus, for the second design cycle, a bandgap minimum cutoff of 2.25 eV was set, leaving 6191 materials in the search space. During the third design cycle, this minimum cutoff was raised to 5 eV, and the search space consisted of 1046 materials with very high values for E_g. In this work, the optimization of multiple objectives - ϵ and E_g - is achieved by two different approaches. The optimization of ϵ happens via statistical modeling while the optimization of E_g is achieved via simple data filtering since all the E_g is already available. This strategy deviates from the ideal multi-objective optimization algorithms benchmarked on dummy datasets and stands as an example for situations where modifying theoretical approaches can be beneficial depending upon real-world conditions.

In total, we calculated the dielectric tensor of 17 compounds over the course of three design cycles. Table 3.1 lists all of these materials along with their ϵ eigenvalues and the bandgaps. The evolution of the Pareto-front of the known dielectric materials that include the MP training-data and the new DFPT calculations is shown in Figure 3.7. Since this is a multi-objective optimization over ϵ and E_g, modification of the Pareto-front is considered as the identification of a new promising dielectric. Five materials were selected from the search space in the first design cycle based on their $E(I)$ values. Out of the five, two of them - HoN and Bi_2SeO_2 - turned out to have very high ϵ values ($\sim$ 370) but with very low bandgap values ($\leq$0.5). This shows the reliability of the selection algorithm because it found two high-ϵ valued compounds just as it was expected to. But since the bandgaps are very small, these two materials are not fit for being good dielectrics. Another material selected in this cycle, $BaZrN_2$ (ϵ=31, E_g=1.2 eV), is, in fact, more preferred than the former two. None of the materials modified the Pareto Front as shown in Figure 3.7a. The Parento front in Figure 3.7a is fully populated by the materials in the initial MP dielectric data. Thus, we added the data of all five new dielectrics to the training-data and proceeded to the next design cycle. In the second design cycle, five more materials were selected based on their $E(I)$ values. Due to the bandgap minimum threshold filter applied on the search space, all the selected materials had their E_g $\geq$2.3 eV. All of the five materials had moderate to high dielectric values (24$\geq$ ϵ $\leq$100) as well. As shown in Figure 3.7b, one of the materials - Tl_3PbBr_5 - modified the Pareto Front due to its very high ϵ value (100.8) and moderately high E_g value (2.9 eV). Even though the other four candidate materials - Sr_2LuBiO_6, Bi_5IO_7, Bi_3ClO_4, and Bi_3BrO_4 - did not make it to the Pareto Front, they are still good contenders for regular dielectric applications due to relatively large ϵ and E_g values. During the third and the last design cycle, all the search space materials had exceptionally high E_g values ($>$ 5 eV) due to the bandgap threshold filter. Seven materials were selected in this cycle for DFPT calculations. Two of them - HoClO (ϵ=75, E_g=5.2 eV), $Eu_5SiCl_6O_4$ (ϵ=69, E_g=5.5 eV) - made it to the Pareto Front as shown in Figure 3.7c. With three new materials joining the Pareto Front, the materials design workflow ended after three design cycles.

3.2.6 New High-Dielectric Materials

The 17 materials selected and characterized using DFPT calculations during the design cycles are listed in Table 3.1. Seven materials satisfied the initial constraint ($\epsilon > 20$ and $E_g > 2.25$ eV). Out of these seven materials, three of them even improved the Pareto front of the previously known data, as seen in Figure 3.7. The most promising materials are the mixed-anion compounds monoclinic $Eu_5SiCl_6O_4$ (ϵ_{max}=69.3, E_g=5.54eV) and tetragonal HoClO (ϵ_{max}=75.1, E_g=5.19eV). The mixed-anion compounds are a class of materials with at least two different species of anions in their composition. Mixed anionic materials are a class of emerging functional materials[42], and this identification of high-dielectrics among them could boost the general research interest in them. To reiterate the advantage of statistical materials selection, the monoclinic $Eu_5SiCl_6O_4$ has 32 atoms in its primitive unit cell which often exceeds the maximum cutoff on the number of atomic sites in high throughput studies involving computationally expensive material properties[38], [43].

Statistical selection methods enable the expansion of the search during computational material selection to materials with larger unit cells as well. Another promising rare-earth halide is tetragonal Tl_3PbBr_5 (ϵ_{max}=100.8, E_g=2.86eV). In all three new high-dielectric materials, the ionic contribution to the static dielectric constant is higher than the electronic contribution, as shown in Table 3.2. The crystal structures of these three best new dielectrics are visualized in Figure 3.8.

OQMD ID	**Material**	E_g	ϵ_x	ϵ_y	ϵ_z	**Design Cycle**
681780	$CaVO_3$	0.4	4.7	4.5	4.5	1
14476	Sr_2VN_3	1.8	28.8	16.5	16.0	1
13450	$BaZrN_2$	1.2	31.2	31.2	21.7	1
1104204	HoN	0.4	376.9	373.0	372.7	1
649584	Bi_2SeO_2	0.5	377.3	371.8	118.2	1
19571	**Sr_2LuBiO_6**	2.4	24.1	19.4	18.7	2
5958	**Bi_5IO_7**	2.7	35.8	28.2	23.1	2
24994	**Bi_3ClO_4**	2.3	38.9	24.2	25.7	2
22697	**Bi_3BrO_4**	2.3	39.0	23.7	22.1	2
118234	**Tl_3PbBr_5**	2.9	100.8	36.4	36.4	2
11916	Eu_4Cl_6O	5.3	7.4	7.3	5.5	3
18953	EuClF	5.6	11.1	11.1	10.4	3
646321	Rb_2PrCl_5	5.1	12.2	11.0	8.9	3
15191	$Cs_2NaCeCl_6$	5.1	13.2	13.2	13.2	3
4063	$EuCl_2$	5.2	15.6	12.9	11.8	3
24611	**$Eu_5SiCl_6O_4$**	5.5	69.3	15.1	12.9	3
13689	**HoClO**	5.2	75.1	37.9	15.2	3

Table 3.1: DFT-calculated dielectric constants of 17 compounds selected during the three design cycles. The OQMD ID refers to the materials' unique entry ID in the OQMD database, E_g refers to the bandgap energy in eV, $\epsilon_{x,y,z}$ refers to the three eigenvalues (xx, yy, zz) of the of dielectric constant tensor, and the Design Cycle column notes the design cycle when the material was selected for the calculations of dielectric constant using DFPT. The values $\epsilon_{x,y,z}$ are ordered in such a way that $\epsilon_x > \epsilon_y > \epsilon_z$. The best materials identified in this work are highlighted in bold letters.

An important factor to consider while computing dielectric properties via DFPT is the presence of imaginary phonon modes in the calculation that will cause dynamic instability of the material[44]. All of the imaginary phonon modes observed during the DFPT calculations of the best

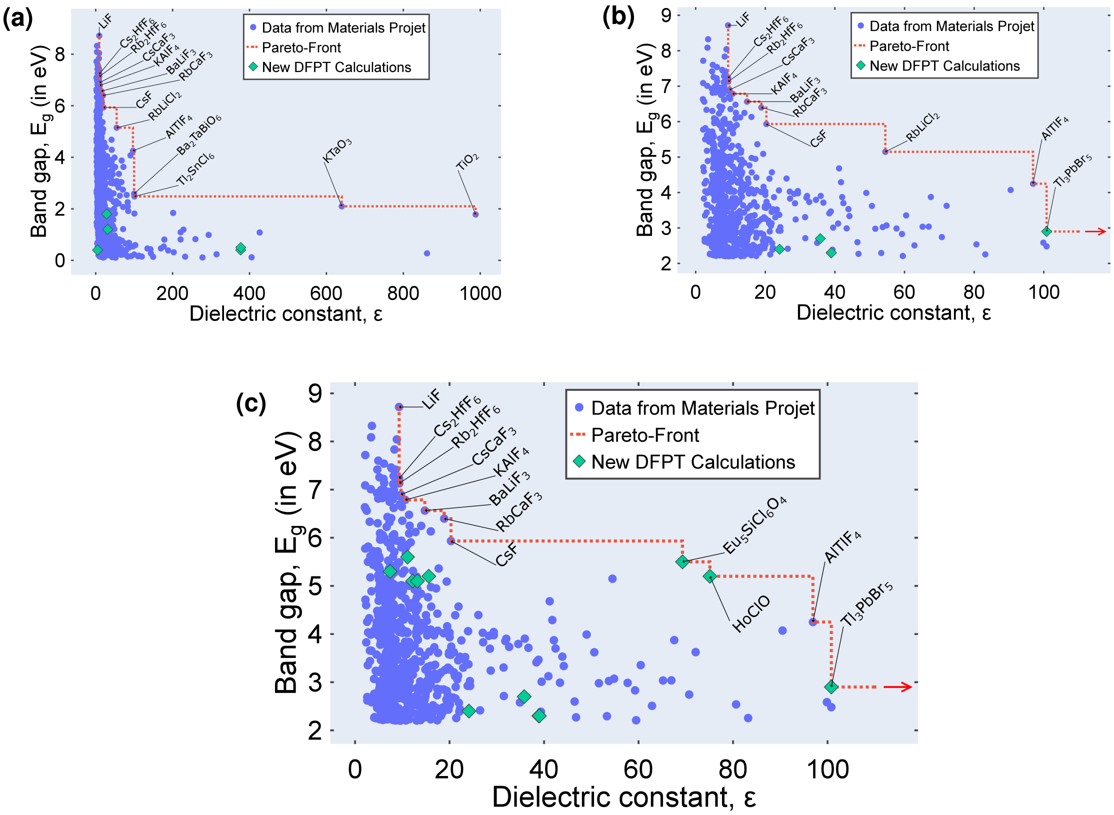


Figure 3.7: **Modification of the Pareto-front after each design cycle** The Pareto-front is the set of the most optimized group of materials in a multi-objective dataset. If a material M belongs to the Pareto-front of the known dataset in this work, that implies that there are no other materials in the known dataset that has a higher value for both ϵ and E_{g} than the material M. Subplots **(a)**, **(b)**, and **(c)** show the Pareto-front of the known data after design cycles 1, 2, and 3 respectively. The known data is the union set of the initial MP training-data and the newly characterized materials from DFPT in this work. None of the materials selected and characterized in design cycle 1 made it to the Pareto-front due to their very low bandgap values, and thus, the Pareto-front in Subplot **(a)** is the same as the Pareto-front of the initial MP data. Only the materials with E_{g} >2.0 eV are plotted in Subplots **(b)** and **(c)** to highlight the area where some of the newly discovered dielectrics in their corresponding cycles joined the Pareto-front. Two materials from the MP-dataset with very high ϵ values - tetragonal TiO_2 (ϵ=988, E_{g}=1.8 eV) and cubic $KTaO_3$ (ϵ=640, E_{g}=2.1 eV) are in the Pareto-front in plots **(b)** and **(c)**, but that part of the Pareto-front is cropped out for better visibility of the section of interest.

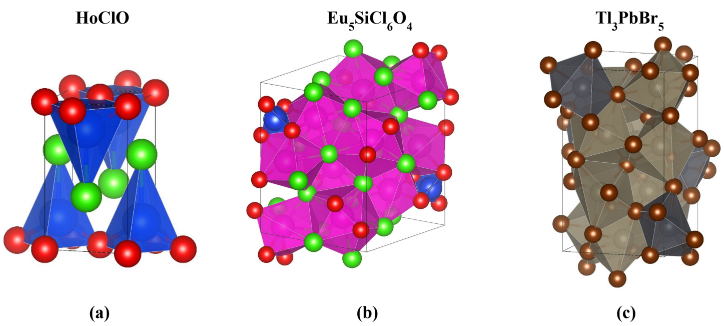


Figure 3.8: Crystal structures of **(a)** HoClO, **(b)** $Eu_5SiCl_6O_4$, and **(c)** Tl_3PbBr_5

Table 3.2: Fraction of ionic contribution $\frac{\epsilon_{ion}}{\epsilon_{ion}+\epsilon_{ele}}$ to the total static dielectric constants ($\epsilon = \epsilon_{ionic} + \epsilon_{electronic}$) for the best three high-dielectrics identified in this work. The ionic contribution is most significant in the case of Tl_3PbBr_5 and $Eu_5SiCl_6O_4$ on all three diagonal dielectric tensor components. In HoClO, both ionic and electronic contributions are similar in magnitude.

Material	$\epsilon_{\mathbf{ionic}}/\epsilon$		
	xx	yy	zz
Tl_3PbBr_5	0.9	0.9	0.9
HoClO	0.6	0.6	0.6
$Eu_5SiCl_6O_4$	0.9	0.7	0.7

three dielectrics are provided in Table 3.3. These phonon frequencies at the Γ-point are small and well within the expected range of numerical error, thus, indicating the lack of any related structural instabilities. The two rare earth oxychlorides with very large bandgaps (> 5 eV)-$Eu_5SiCl_6O_4$ and HoClO- are reported to have been experimentally synthesized [45]–[48], but their dielectric properties remain unstudied to the extent of our knowledge.

The thermodynamic stability of a dielectric is of concern when they are used in electronic circuits. A dielectric may be used in contact with other common electronic materials, such as Si, Ge, GaAs, GaN, and SiC, in addition to the chemicals in the environment, such as oxygen and nitrogen. The dielectric must remain non-reactive with all of these compounds during its deployment in real

Table 3.3: Imaginary phonon modes in high-dielectric materials. The phonon frequencies of the three acoustic phonon modes at Γ-point for the high-dielectric materials identified in this work. These small imaginary frequencies reported here fall within the numerical error of the calculations.

Material	Imaginary Phonon modes
Tl_3PbBr_5	0.005650 THz 0.021434 THz 0.022159 THz
HoClO	0.015989 THz 0.034859 THz 0.052248 THz
$Eu_5SiCl_6O_4$	0.013365 THz 0.021796 THz 0.048696 THz

devices[49]. Many high-dielectrics previously reported in the literature, such as Ta_2O_3[50]–[52], TiO_2[53], [54], $BaTiO_3$[55] and $SrTiO_3$[56], [57] suffer from reacting with Si when used in the circuits and eventually decomposes into other compounds such as SiO_x. This instability makes these dielectrics unusable in practice. The thermodynamic stability of two compounds in contact can be assessed from DFT by constructing a convex hull[58] of the phase space occupied by their constituent elements. The convex hull method uses formation energies of materials that are readily available from OQMD and other such HT databases. Every material that makes it to the convex hull not only has the lowest formation energy at its composition but also has lower energy than any linear combination of other materials in that phase space. The numerical difference between the formation energy of a compound and energy at the convex hull for the same composition is called hull distance (E_{hd}). By definition, each material that has a zero hull distance (E_{hd} = 0) is considered to be stable, while every material that has a small, but finite E_{hd} is considered as metastable

($0 < E_{\text{hd}} \leq 50$ meV per atom). Materials with larger E_{hd} are considered to be unstable ($E_{\text{hd}} > 50$ meV per atom). The cutoff between metastability and complete instability is decided based on the magnitude of E_{hd} according to the conventions practiced in literature [59]–[63]. In a convex hull phase diagram, the presence of a tie line between two compounds indicates their thermodynamic stability when in contact with each other.

In Figure 3.9, we show the convex hull built for some of the previously reported dielectrics along with common electronic materials, based on the formation energy data obtained from OQMD. The absence of tie-lines from Ta_2O_3, TiO_2, $BaTiO_3$, and $SrTiO_3$ to Si suggests that they are unstable when in contact with Si. To confirm the reliability of convex hulls, we included another material in the same phase diagram in Figure 3.9 - Gd_2O_3 ($\epsilon \sim 20$[64]). Gd_2O_3 is also experimentally proven to be stable when in contact with Si[65]. In its phase diagram, Gd_2O_3 has a tie-line to Si, making this DFT-based stability assessment agree very well with experimentally observed results in the case of both stable and unstable materials that we considered to benchmark. We constructed a new phase diagram in Figure 3.10 to assess the DFT-predicted stability of the two new high-ϵ, high E_g dielectrics - HoClO and $Eu_5SiCl_6O_4$ in a similar manner. Both of these materials are observed to have tie-lines to Si, Ge, GaAs, GaN, SiC, N, and O. This indicates that they are expected to be fit for being used in electronic devices. The other promising material that made to the Pareto-front with a larger ϵ (101) but relatively smaller bandgap (2.9 eV) - the tetragonal Tl_3PbBr_5, is observed to be metastable in convex-hull analysis with E_{hd} = 16 meV per atom. Tl_3PbBr_5 is also reported in the peer-reviewed literature as a material that had been experimentally synthesized[66]–[68], without any mention of its dielectric properties to the best of our knowledge.

We did a further computational analysis of the top three dielectrics found in this work - HoClO, $Eu_5SiCl_6O_4$, and Tl_3PbBr_5 - by computing their electronic bandstructures as shown in Figure 3.11. The bandstructures show that the conduction band maxima in each of them are occupied by lighter

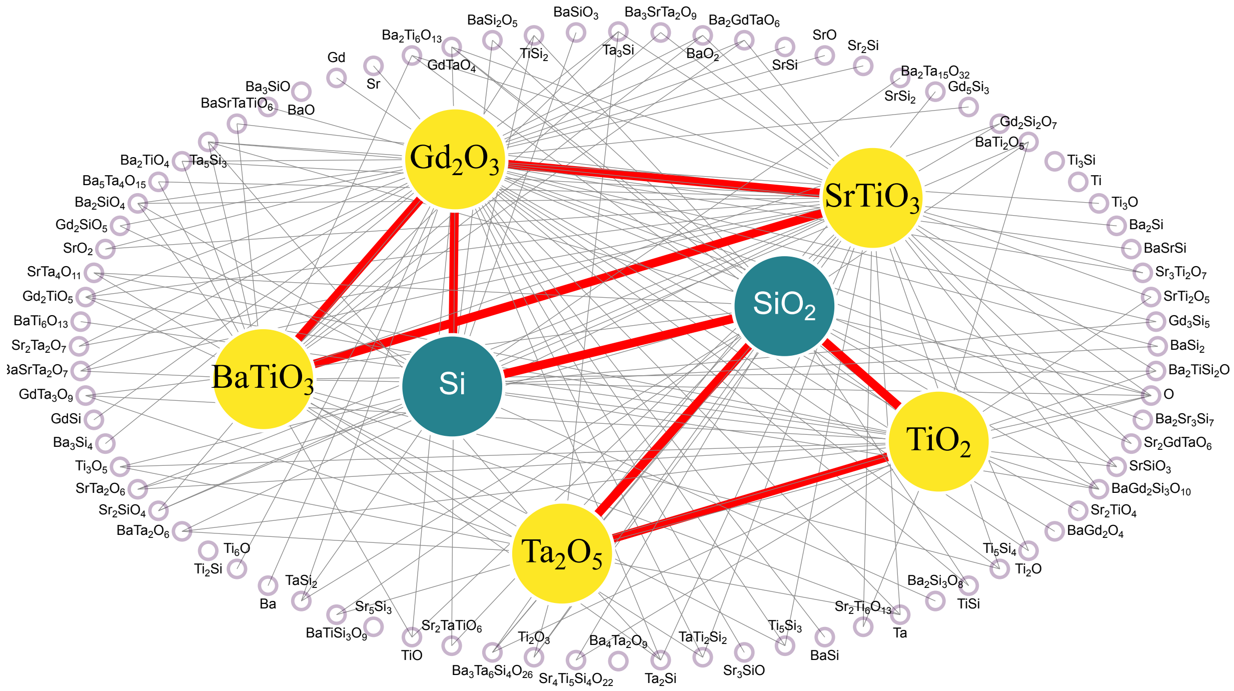


Figure 3.9: Phase diagram of Ba-Ti-O-Sr-Ta-Si phase space in OQMD database. The most relevant materials are shown in large circles in the inner shells. The tie lines between the inner-shell materials are plotted in thick red lines, while the thinner gray lines are the tie lines that connect the inner-shell materials to the materials on the outermost shell. A tie line exists between Si and Gd_2O_3 indicating the relative stability of these two materials when in contact with each other. No tie-lines originate from Si to $BaTiO_3$, $SrTiO_3$, TiO_2 or Ta_2O_5 indicating their energetic instability.

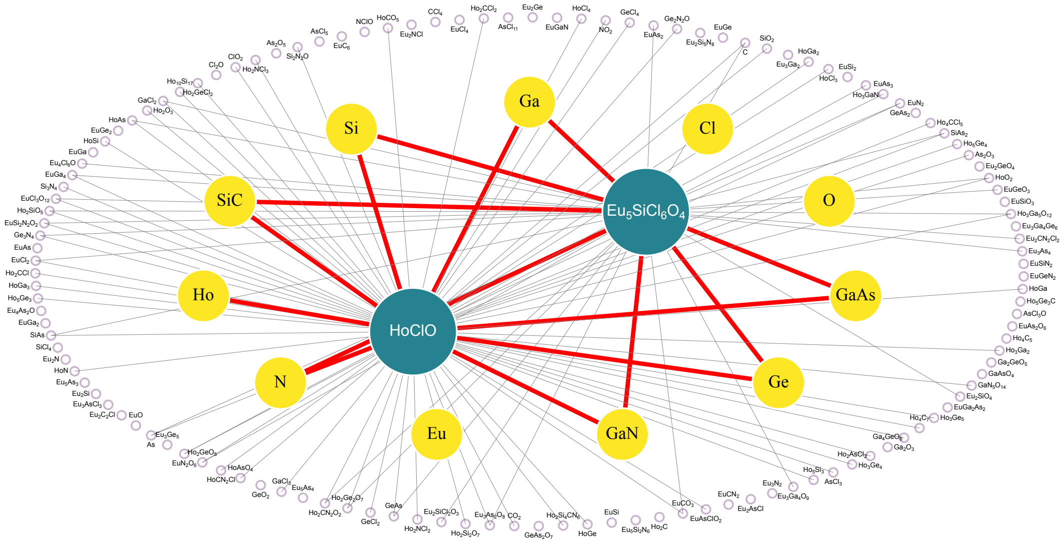


Figure 3.10: **The convex-hull phase diagram of all stable compounds in Ho-Cl-O-Eu-Si-Ge-Ga-As-C-N phase-space from OQMD (as of January 2022).** The two most promising dielectrics identified in this work - HoClO and $Eu_5SiCl_6O_4$ are plotted in large green circles in the center. The elements (Ho, Eu, Si, Cl, Ge, Ga, As, C, N, and O) and semiconductors of interest (Si, Ge, GaAs, SiC, and GaN) as suggested by Robertson[49] are plotted in the middle layer in medium-sized yellow circles. All other stable compounds in the phase diagram are plotted in small dark circles in the outermost layer. Tie-lines between the new dielectrics and the semiconductors or elements are shown as thick red lines. Other tie-lines from the dielectrics to the rest of the stable materials in the outer layer are drawn as narrow gray lines. Another 2326 tie-lines exist in this phase diagram that does not include either of the dielectrics. Those lines are not shown in this network plot for better visibility of the information relevant to the new dielectrics. The elements and compounds without any visible tie-lines in the outermost layer are still part of this phase diagram since they have tie-lines with some of the other materials in the outer layer even though they lack tie-lines to HoClO or $Eu_5SiCl_6O_4$. We observe that there exists a tie-line from each dielectric material to each semiconductor that is considered here for comparison, indicting that HoClO and $Eu_5SiCl_6O_4$ are in thermodynamic equilibrium with Si, Ge, GaAs, GaN, and SiC at 0K.

anions (Cl, Br) and, thus, much lower in energy than the conduction band minima occupied by contributions from cations (Ho, Eu, Tl). This indicates a tendency to form larger bandgaps, which is a desired property in this work. Calculations using PBE in DFT often underestimate the bandgap of materials. Because of that, the real bandgaps of HoClO and $Eu_5SiCl_6O_4$ could be higher than even the values mentioned here, making them even more resistant to leakage currents. The results of this work hint that the mixed-anion compounds formed by rare-earth elements are a class to be investigated in detail for their dielectric properties. The presence of rare-earth elements such as Ho and Eu in the new dielectrics can raise concerns over their availability in manufacturing at an industrial scale. However, this is an actively researched topic for other similar materials as well these days, and some of the most practical recommendations involve better recycling of rare-earth materials[69], [70], which may result in a sufficient supply of the rare elements for mass production of small electronic components. In fact, Ho is more abundant in the earth's crust than other widely mined elements such as Mo, Bi, and precious metals[71] but still remains an underutilized element in the industry[72]. Eu is more abundant on earth's crust than Ho and some of the heavily mined elements such as W and As[71] even though the processing methods to extract them from their ores may still be limited. Tl_3PbBr_5 is a good candidate for dielectric applications in controlled environments, but the presence of toxic elements such as Pb and Tl in it can make it less probable to be used as a dielectric in consumer electronics.

3.3 Discussion

In this work, we report three new high-ϵ, high-E_g materials found via an iterative computational materials design approach consisting of ab-initio density functional perturbation theory (DFPT) calculations, high-throughput data analysis, and statistical optimization. We also demonstrate a successful mixing of two different high-throughput databases (Materials Project and the Open

Quantum Materials Database (OQMD)) to discover new high dielectric constant materials with large bandgaps in a machine-learning-aided materials design framework. The thermodynamic stabilities of the two best newly discovered dielectrics, HoClO (ϵ=75, E_g=5.2 eV) and $Eu_5SiCl_6O_4$ (ϵ=69, E_g=5.5 eV) when in contact with other common electronic component materials are evaluated using the convex-hull construction as implemented in the OQMD. Both of these materials are found to be thermodynamically stable against common substrate materials such as Si, Ge, GaAs, GaN, and SiC. Our screening strategy also uncovers another high-ϵ material - Tl_3PbBr_5 (ϵ=101, E_g=2.9 eV), and four other dielectric materials with large E_g and relatively large ϵ - Sr_2LuBiO_6 (ϵ=24, E_g=2.4 eV), Bi_5IO_7 (ϵ=36, E_g=2.7 eV), Bi_3ClO_4(ϵ=39, E_g=2.3 eV), and Bi_3BrO_4(ϵ=39, E_g=2.3 eV).

We computed the electronic bandstructure of HoClO, $Eu_5SiCl_6O_4$, and Tl_3PbBr_5 and report the composition of valence bands predominantly by lighter anion orbitals and the domination of conduction band edges by cations, which may have contributed to the larger bandgaps in these materials. The industrial availability concerns of the constituent elements, specifically the rare earth metals, are discussed, and methods to solve this issue based on recycling are reported from previously published literature. Investing in research toward cheaper and easier extraction methods for rare earth elements may make it feasible to include them in mass-produced electronics in the near future.

The new dielectrics were discovered after conducting three materials design cycles. Each design cycle consists of an artificial neural network (ANN) model ensemble to learn from known data and predict dielectric value distributions for candidate materials, a statistical optimization model to quantify exploration-exploitation potential of predictions, a set of high-accuracy DFPT calculations on a selected subset of compounds and finally, feedback to the ANN modeling for the subsequent design cycle. Overall, this work also shows an example of how a robust materials design workflow containing high-throughput data, statistical modeling, and expensive validation methods can discover novel materials for high-performance applications within a few selections when the resources are too limited to search through all the materials in search space.

3.4 Methods

The iterative design workflow implemented in this work is described in detail in Figure 3.1. The design involves the sequential usage of an optimization algorithm and an ab-initio simulation framework. Both of them are described below.

3.4.1 Efficient Global Optimization

Efficient Global Optimization[10], [36] (EGO) is a statistical optimization algorithm, particularly applicable when the search space is significantly larger than the training-data. In this work, the predicted ϵ value distribution of search space materials is fed into the EGO algorithm. EGO assigns an expected improvement value, $E(I)$, for each material. Here, the $E(I)$ of a material in the search space is the quantified probability with which the calculation of ϵ_{max} using DFPT for that material will lead to a discovery of high-ϵ_{max} material in the design workflow within as few design cycles as possible. That means a material with large ϵ_{max} would either be a material with a large mean for its ANN ensemble prediction distribution, or a large standard deviation in ANN ensemble prediction distribution.

3.4.2 Density Functional Perturbation Theory

Density Functional Perturbation Theory (DFPT) is an ab-initio method based on DFT to calculate properties that are dependent upon some kind of perturbation in the system. High-accuracy DFPT calculations are much more expensive to do than regular DFT calculations. Here, the DFPT algorithm, with PBE functional approximation for exchange-correlation energy, as implemented in the Vienna Ab-initio Simulation Package[22], [23] (VASP), was used to compute total dielectric constant values (eigenvalues of the sum of electronic and ionic tensors) for crystals. The method that VASP follows to compute ϵ matrix is explained below, and derived in detail in Gajdoš et. al.[73] First, a polarization vector, $|\beta_{n\boldsymbol{k}}\rangle$, is defined within PAW[25] formalism as,

$$\vec{\beta}_{n\boldsymbol{k}} = \big(1 + \sum_{i} \langle \tilde{p}_{i\boldsymbol{k}} | \, Q_{ij} \, | \tilde{p}_{j\boldsymbol{k}} \rangle \big) \, | \nabla_{\boldsymbol{k}} \tilde{u}_{n\boldsymbol{k}} \rangle + i \big(\sum_{ij} \langle \tilde{p}_{i\boldsymbol{k}} | \, Q_{ij} \, | \tilde{p}_{j\boldsymbol{k}} \rangle \, (\boldsymbol{r} - \boldsymbol{R_i}) \big) \, | \tilde{u}_{n\boldsymbol{k}} \rangle \\ - i \big(\sum_{i} \langle \tilde{p}_{i\boldsymbol{k}} | \, \tau_{ij} \, | \tilde{p}_{j\boldsymbol{k}} \rangle \big) \quad (3.3)$$

where the indices i and j refer to atoms while $\boldsymbol{R_i}$ is the position of atom i. Other variables are, $\boldsymbol{k}$ = Bloch wave vector[25], n = band index, $\tilde{p}$ = projector function[24], [73], $\tilde{u}$ = cell-periodic part of PAW pseudo wave-function[24], [25], [73], Q = norm of PAW one-center charge density[24], [73], and τ = dipole moment of PAW one-center charge density[73].

The external electric field causes a modification to the one-electron wave functions, Ψ_{nk}. The first order response of $\Psi_{n\boldsymbol{k}}$ is denoted as $\xi_{n\boldsymbol{k}}$ and is computed via solving the following linear equation[73], [74],

$$\big[\boldsymbol{H}(\boldsymbol{k}) - E_{n\boldsymbol{k}}\boldsymbol{S}(\boldsymbol{k})\big] \, |\xi_{n\boldsymbol{k}}\rangle = -\Delta \boldsymbol{H_{SCF}}(\boldsymbol{k}) \, |\tilde{u}_{n\boldsymbol{k}}\rangle - \left| \hat{\boldsymbol{q}} \vec{\beta}_{n\boldsymbol{k}} \right\rangle \quad (3.4)$$

where, $\boldsymbol{H}(\boldsymbol{k})$ = Hamiltonian for cell-periodic wave-functions, $\boldsymbol{S}(\boldsymbol{k})$ = Overlap operator[73], [75], $\hat{\boldsymbol{q}}$ = direction vector in reciprocal space, E = reference Eigen energy, and the $\Delta \boldsymbol{H_{SCF}}(\boldsymbol{k})$ refers to the first-order microscopic cell-periodic change in $H(k)$ due to the external field induced modifications in Ψ[73], [75]. Presence of $\Delta \boldsymbol{H_{SCF}}(\boldsymbol{k})$ term includes the local field effects in this method.

The $\vec{\beta}$ and ξ are used while computing the electronic dielectric tensor, ε_∞, as,

$$\varepsilon_\infty(\hat{\boldsymbol{q}}) = 1 - \frac{8\pi e^2}{\Omega} \sum_{v,\boldsymbol{k}} 2w_{\boldsymbol{k}} \left\langle \hat{\boldsymbol{q}}\vec{\beta_{v\boldsymbol{k}}} \middle| \xi_{v\boldsymbol{k}} \right\rangle \tag{3.5}$$

where v refers to valence band states, $\varepsilon_\infty(\hat{\boldsymbol{q}})$ = macroscopic electronic dielectric constant in $\hat{\boldsymbol{q}}$ direction, w_k = k-point weights, Ω = primitive cell volume and the factor of 2 inside summation corresponds to a spin degenerate system.

The gradient of $\tilde{u}$ in reciprocal space, $\nabla_{\boldsymbol{k}}\tilde{u}_{n\boldsymbol{k}}$ which appears in Equation 3.3 is computed via solving the linear Sternheimer equation[73], [76],

$$\big[\boldsymbol{H}(\boldsymbol{k}) - E_{n\boldsymbol{k}}\boldsymbol{S}(\boldsymbol{k})\big] \left|\nabla_{\boldsymbol{k}}\tilde{u}_{n\boldsymbol{k}}\right\rangle = -\frac{\partial\big[\boldsymbol{H}(\boldsymbol{k}) - E_{n\boldsymbol{k}}\boldsymbol{S}(\boldsymbol{k})\big]}{\partial k} \left|\tilde{u}_{nk}\right\rangle \tag{3.6}$$

Further, the total dielectric tensor, ϵ_{total} is calculated as the sum of electronic and ionic contributions:

$$\epsilon_{total} = \epsilon_\infty + \sum_\mu \epsilon_\mu \tag{3.7}$$

where, the ϵ_μ denotes the oscillator strength of phonon mode μ and is derived as[77]–[79]:

$$\epsilon_\mu = \sum_{\omega_\mu^2 \neq 0} \frac{\left[\sum_{i\gamma} Z^*_{i\alpha\gamma}\, a_{\mu i\gamma}\, m_i^{1/2}\right]\left[\sum_{i\gamma} Z^*_{i\beta\gamma}\, a_{\mu i\gamma}\, m_i^{1/2}\right]}{\Omega \in_0 \omega_\mu^2} \tag{3.8}$$

where α and β denote Cartesian directions whose combinations create a 3x3 dielectric tensor matrix for bulk crystals. The Born effective charge and normalized eigenmode of ion i with mass m_i in the direction γ due to phonon mode μ are denoted by $Z^*_{i\alpha\gamma}$ and $a_{\mu i\gamma}$ respectively[77]. The

Table 3.4: Dielectric constants and bandgaps of common dielectric materials. We computed the dielectric properties of the commonly known dielectric materials to assess the reliability of our DFPT calculation framework. The results are listed here alongside the reference values obtained from peer-reviewed computational literature

Material	Structure	OQMD Entry ID	ϵ_x	ϵ_y	ϵ_z	Ref ϵ (DFPT)	$E_{\mathrm{g,OQMD}}$	$\mathrm{E}_{g,Exp}$
HfO_2	Cubic	647078	27	27	27	29 [80]	4.1 eV	5.7 eV [81]
ZrO_2	Tetragonal	646648	50	50	21	47 [80]	4.0 eV	5.0 eV [82]
Anatase	Tetragonal	2575	67	67	33	46-24 [83]	2.0 eV	3.2 eV [84]
Rutile	Tetragonal	2475	144	144	139	165-117 [85]	1.7 eV	3 eV [86]
EuF_2	Cubic	5660	6	6	6	8 [87]	7.9 eV	-
EuO	Tetragonal	1443633	26	26	25	24 [87]	2.9 eV	1.5 eV [88]
Ho_2O_3	Cubic	5389	13	13	13	13.1 [87]	4.1 eV	5.31 eV [89]

origin of ionic contribution to dielectric tensor is the polarization that arises when the ionic crystal structure is slightly deformed due to phonons.

We calculated the ϵ values of some of the common dielectric materials to benchmark our DFPT calculations, as listed in Table 3.4. The DFPT calculations are done using VASP and a tightly-relaxed crystal structure of the crystal as input. A high k-point density (KPPRA=8000) and tight energy convergence criteria (EDIFF=10^{-8}) are used to get accurate results.

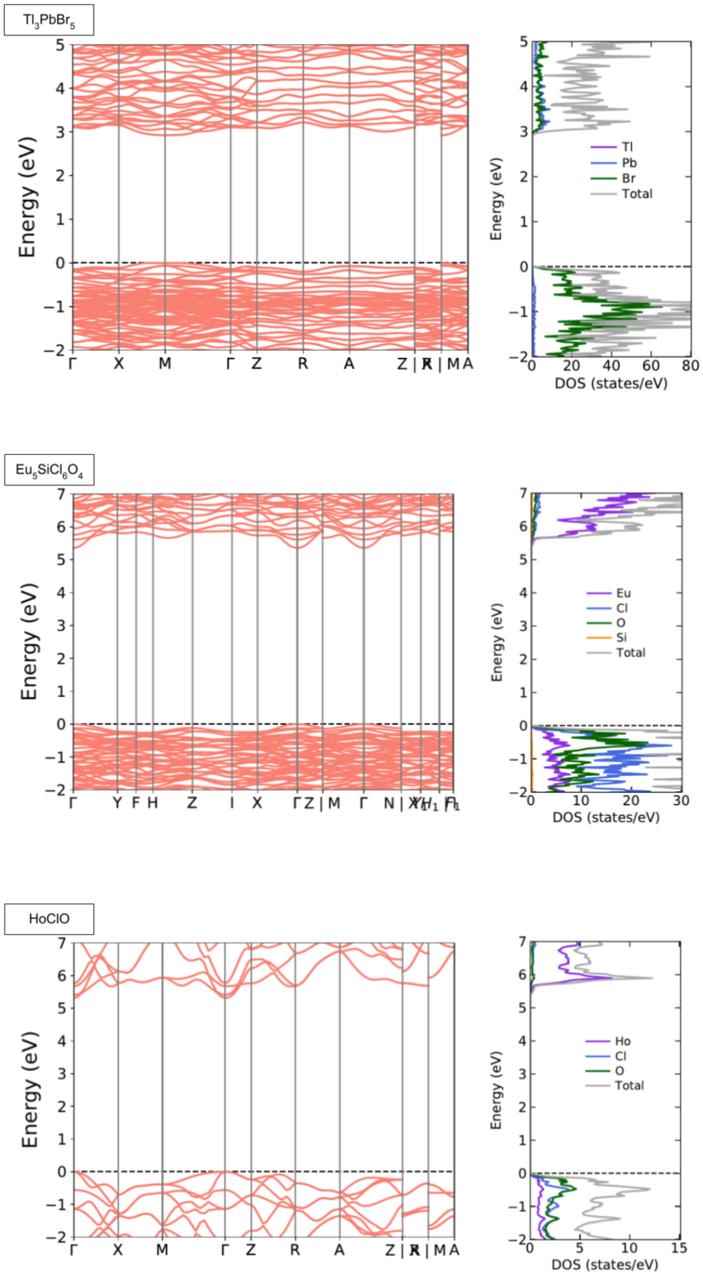


Figure 3.11: Electronic bandstructures (left-side) and partial density of states (right-side) of Tl_3PbBr_5 (top), $Eu_5SiCl_6O_4$ (middle), and HoClO (bottom). From this analysis, we find that the top of the valence band found is dominated by the orbitals of the anions, and the bottom of the conduction band primarily comes from the orbitals of the rare-earth elements.

CHAPTER 4
IDENTIFICATION OF HEXAGONAL GAS AS A BRIDGE MATERIAL FOR MOLYBDENUM DISULFIDE-BASED 2D INKS

4.1 Background

Most electronic circuits today are comprised of materials like silicon which have varied sets of electronic properties required to build components such as transistors and capacitors. But their bulk structure also limits where these circuits can be placed to only the rigid surfaces resistant to mechanical transformations of any kind. This limitation was not of significant concern to most technologies until the relatively new interest in flexible electronics emerged. The rising popularity of wearable devices and similar technologies demand the circuits be printed onto surfaces that are frequently movable, stretchable, and exposed to external forces. A material that can be used to print circuits even over flexible surfaces with reliable, long-term adhesiveness and optimal electrical properties will be capable of reshaping the consumer electronics industry. Such a capability is expected to lower manufacturing costs and decrease electronic circuits' form factor. This opens the doors for innovations and explorations into the world of 2D electronic materials that can be printed as various electronic components onto such surfaces. Over the past decade, several materials that are stable in their 2D form have been investigated for this purpose[90]–[93]. Since the materials' electronic properties change when made into a 2D sheet due to quantum confinement effects, the current set of materials that build the circuits in their bulk crystalline forms do not necessarily do well when used in 2D forms. This opens doors for exploration into a new set of electronic materials in the 2D domain. Currently, graphene and MXenes are often used as conductors while h-BN and

MoS_2 are used as insulators and semiconductors in 2D inks[93].

One of the most practically efficient ways to create circuits using 2D materials is to use them in liquid form as inks in an inkjet printer[91], [94]. Several materials, including graphene, WS_2, MoS_2, black phosphorus, etc., are widely explored to be used as such 2D ink materials. Our work focuses on MoS_2-based 2D inks due to their desirable semiconducting properties and the ease of synthesis from solution processing methods[94]–[98]. 2D-MoS_2 is reported to have been used for several semiconducting applications, including transistors[98], capacitors[99], and photodetectors[94], [100]. But depending on their application, the MoS_2 2D ink flakes may need another material to act as an adhesive between them[100], as shown in Figure 4.1. These adhesive materials are henceforth referred to as bridge-material. An ideal bridge material should not interfere with the charge transport between MoS_2 flakes, and the heterostructure formed by MoS_2 and the bridge material should have electronic properties similar to those of MoS_2. In this work, we identify hexagonal 2D-GaS as a new bridge material for MoS_2 from a material selection workflow that involves a low-fidelity high-throughput screening of more than 2000 materials followed by extensive literature surveys to select GaS, and finally, the validation of the selection using expensive DFT calculations to estimate the MoS_2/GaS heterostructure's charge transport properties and other electronic characteristics. The new bridge material is also synthesizable via solution processing - a requirement for large-scale production and acceptance of it to be used in 2D inks.

4.2 Results and Discussion

4.2.1 Selection of GaS

The complete material selection procedure implemented in this work is shown in Figure 4.3. First, we obtained the electronic bandstructure data of 2466 semiconducting (0.5 eV $\leq E_g \leq$ 2.5 eV), thermodynamically stable inorganic crystalline bulk materials from the Materials Project (MP)

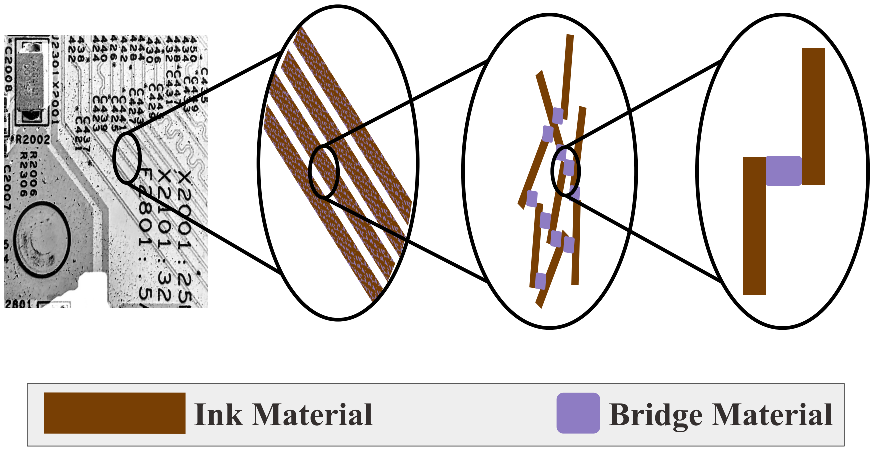


Figure 4.1: A qualitative diagram showing how the internal structure of circuits looks like when printed with 2D inks. The MoS_2 flakes are held together with smaller bridge materials that are expected to not interfere with the main ink material's electronic properties

database[5] to search for an excellent candidate to be the bridge-material in MoS_2 2D inks. The ideal bridge-material should create a 2D heterostructure with hexagonal MoS_2 without impacting the charge transport properties of MoS_2. In this work, we are looking at an extensive set of candidate materials whose heterostructure properties with MoS_2 cannot be studied in a high-throughput manner due to the substantial cost of computational resources such a study would incur. Instead, we use the electronic bandstructures of bulk materials readily available from MP generated via high throughput calculations using DFT and PBE functionals. Since the properties of the bulk structures are expected to differ from that of nanoparticles or monolayered structures due to confinement effects, the bulk bandstructure is used only as the initial screening criteria in this project, as shown in Figure 4.3. As the first step in screening for a good bridge-material, we compare the alignment of conduction band minima (CBM) between the bulk hexagonal MoS_2 and the 2466 candidate materials. The alignment of CBMs in a heterostructure with a small offset indicates with high probability that the carrier transport in the conduction band happens with the least amount of interference[101], [102]. We focus on the CBM offset, rather than of valence band maximum (VBM) offset between the materials because the MoS_2 is known to be an n-type semiconductor in most cases[103], [104]. A VBM offset is relevant to materials in which the holes are the main carriers. Even though the CBM and VBM energy levels are readily available in any bandstructure data, they are not cross-comparable between materials due to the lack of a common universal baseline. Surface slab calculations are commonly accepted as the method to get a high-fidelity estimation of a common reference energy point by treating the vacuum energy level as the zero-energy baseline. But since the surface calculations are computationally expensive, it is not viable in this project to do them for several materials. Hence, we adopt the lower-fidelity estimation of a cross-comparable energy reference baseline called branch-point energy[105], [106], denoted as E_{BP}, which can be obtained from any bandstructure data - including that from bulk structure calculations. E_{BP} is the energy level in the bandstructure at which the conduction band (CB)-like behavior and the valence band (VB)-like behavior cancel each other out. As per Tersoff[105], when two materials create a heterostructure, E_{BP} is the energy level where their electronic bands align perfectly after initial charge transfers and balances the interface dipole. Hence E_{BP} is also called the charge neutrality level. More details on the calculation of E_{BP} from a given bandstructure are provided in the Methods section (Section 4.4). We calculated the E_{BP} levels of all the 2466 materials with bandstructures. Since the E_{BP} is the same energy level across materials, the numerical values of energy levels are then shifted in each bandstructure to make E_{BP}=0 eV. With the bandstructure energy values readjusted in reference to the same baseline, their CBM values are directly comparable with each other. But another factor to consider here is the low accuracy in electronic bandgap (E_g) estimation in DFT-PBE calculations. Since the CBM values are obtained as the sum of VBM and bandgap, the accuracy of bandgap has a significant impact on the final CBM value. To correct the bandgap present in the DFT-PBE bandstructures, a machine learning

(ML) model that can predict the E$_g$ at a higher accuracy than that from DFT-PBE calculations is used to predict the better estimation of E$_g$ for each of the 2466 materials. This ML mode is built as part of another project that is discussed in Chapter 5, and also briefly described in the Methods section of this chapter (Section 4.4). Using the higher accuracy prediction of the bandgap from the ML model, we corrected the position of CBM in the band structure relative to the VBM and subsequently obtained the cross-comparable CBM values for all of the 2466 materials.

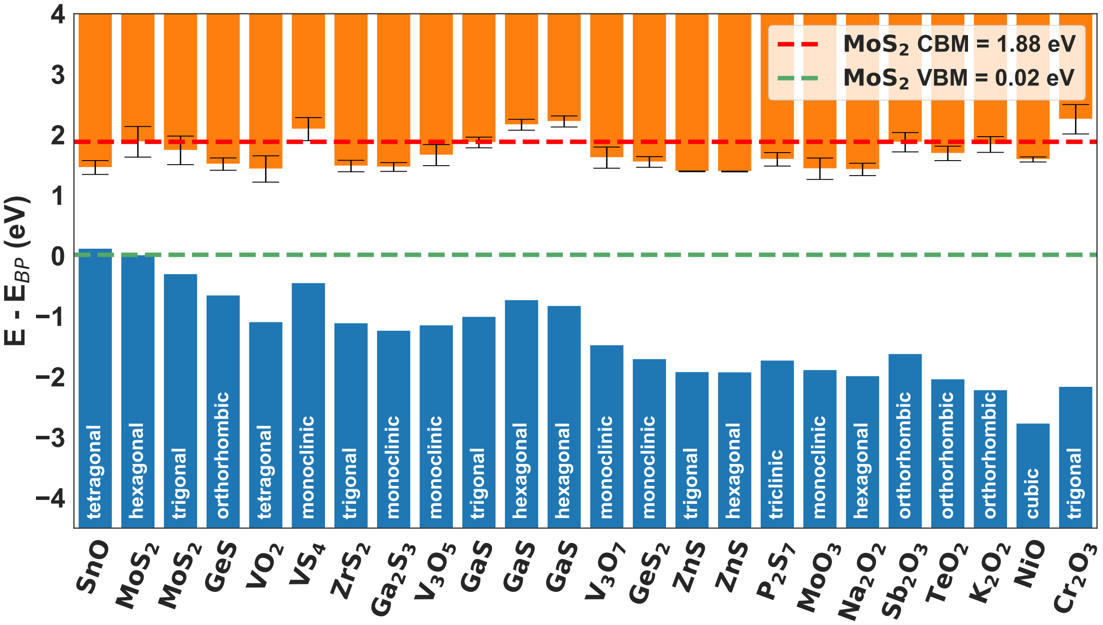


Figure 4.2: Band alignments of binary oxides and binary sulfides with their branch-point energies as the common reference energy point. Only those binary materials whose CBM offset with hexagonal MoS_2 is less than 0.2 eV are shown in this plot. The error bar values are obtained from the uncertainty quantification of the co-kriging model bandgap predictions. A similar plot containing ternary oxides and ternary sulfides is given in the appendix Figure A.1

The top 50 materials were selected for the next step based on the following criteria - *(i)* small CBM alignment offset with bulk hexagonal MoS_2 ($< 0.2 eV$), *(ii)* reported to be experimentally synthesized at ambient temperatures (T=300 ± 40 K) from Inorganic Crystal Structure Database (ICSD), and *(iii)* do not contain any elements that are rare, radioactive, expensive (Sc, Rh, Pd, Rb, Cs, Os, Ir, Pt, Au, Tl, Hf, Ru, Re), or toxic (Cd, Hg, Tl, Pb, As). This set contains 24 binary oxides and sulfides (as shown in Figure 4.2) and 26 ternary oxides and sulfides (shown in Appendix Figure A.1). Further, we conducted a thorough literature survey to find previously reported information on other relevant properties such as (i) the dominance of n-type carriers for charge transport at conduction band similar to that in MoS_2, (ii) support for solution-processing growth techniques which is a requirement for large-scale production of 2D-inks in industry, and (iii) stability and bandgap of the corresponding 2D structure. Based on the findings in all three aspects and further consideration of smaller unit cells, we narrowed the candidate set for doing DFT slab calculations to five: trigonal ZrS_2, monoclinic MoO_3, hexagonal GaS, monoclinic $MnMoO_4$, and orthorhombic $CaTiO_3$. he trigonal ZrS_2 is found to have a polar surface from DFT calculations, while the calculations of monoclinic MoO_3, monoclinic $MnMoO_4$, and orthorhombic $CaTiO_3$ did not converge within our resource limits. The DFT slab calculation converged reliably in the vase of hexagonal GaS. In DFT slab calculations, a large vacuum space is added in the z-axis of the crystal to simulate the surface exposed to the vacuum, and the material's CBM value can be calculated reliably with the common reference point of vacuum whose energy can be set to be 0 eV. Since the slab calculations converged for hexagonal GaS, we chose hexagonal GaS as the best candidate to further study using even more expensive DFT charge-transport calculations. Hexagonal GaS is reported to have been synthesized from solution processing methods[107], [108] and is also reported as a stable n-type semiconductor[109], [110]. Monolayer and few-layered phases of GaS are also reported to be stable experimentally[111].

4.2.2 2D Ink Heterostructure

The initial bulk hexagonal GaS and hexagonal MoS_2 structures for DFT calculations are obtained from the OQMD crystal structure database. A vacuum of height 15 Åis added along the z-direction

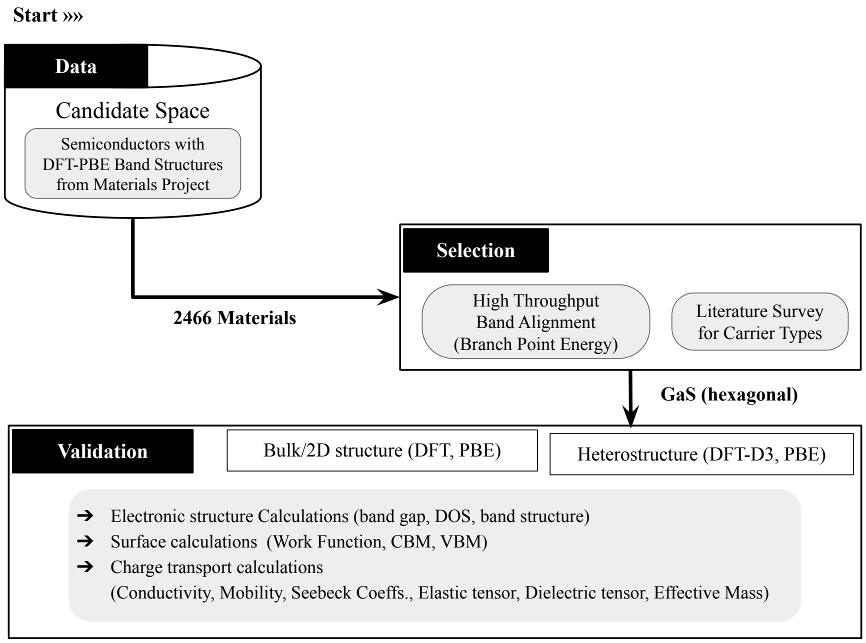

Start »»
Data
Candidate Space
Semiconductors with DFT-PBE Band Structures from Materials Project
2466 Materials
Selection
High Throughput Band Alignment (Branch Point Energy)
Literature Survey for Carrier Types
GaS (hexagonal)
Validation
Bulk/2D structure (DFT, PBE)
Heterostructure (DFT-D3, PBE)
➔ Electronic structure Calculations (band gap, DOS, band structure)
➔ Surface calculations (Work Function, CBM, VBM)
➔ Charge transport calculations
(Conductivity, Mobility, Seebeck Coeffs., Elastic tensor, Dielectric tensor, Effective Mass)

Figure 4.3: The workflow of 2D ink bridge material design implemented in this work. DFT-PBE refers to DFT calculations conducted with PBE functional approximation for exchange-correlation energy. Similarly, DFT-HSE refers to DFT calculations with HSE functional approximation for exchange-correlation energy. ML refers to machine learning. The set of 2466 materials downloaded in the first step forms the initial candidate set. After each step of the design, the number of candidate materials is reduced as the materials are picked from the candidate set based on how they meet the requirements evaluated at a given step. The preference for synthesis method is given s lower priority while filtering even though it is an important factor in manufacturing because a lack of reporting on solution processing in literature does not mean that the material cannot be synthesized using that method.

to create the 2D GaS and MoS_2 single-phase structures. We used a custom-modified version of the Supercell-core python package[112] to create a vertically stacked van der Waals 2D heterostructure from the two single phases. The Supercell-core package searches through different stacking orientations (rotations around the z-axis) between the two lattices and supercell sizes of the van der Waals heterostructure, and finds the optimum supercell with minimal strain on the individual lattices. The customization done on Supercell-core for this project are only to automate of preprocessing and post-processing of the 2D structures while keeping the main functionality intact. A search was conducted for heterostructures with maximum strain across all dimensions being less than 5% and the number of atoms per unit cell less than 100. Only one heterostructure met these criteria, and it is shown in Figure 4.4. The initial separation between the MoS_2 and GaS layers was set to be 3.6 Åbut it changed to 3.3 Åupon tight structural relaxation in DFT to make the forces on atoms less than 10^{-3} eV/Å.

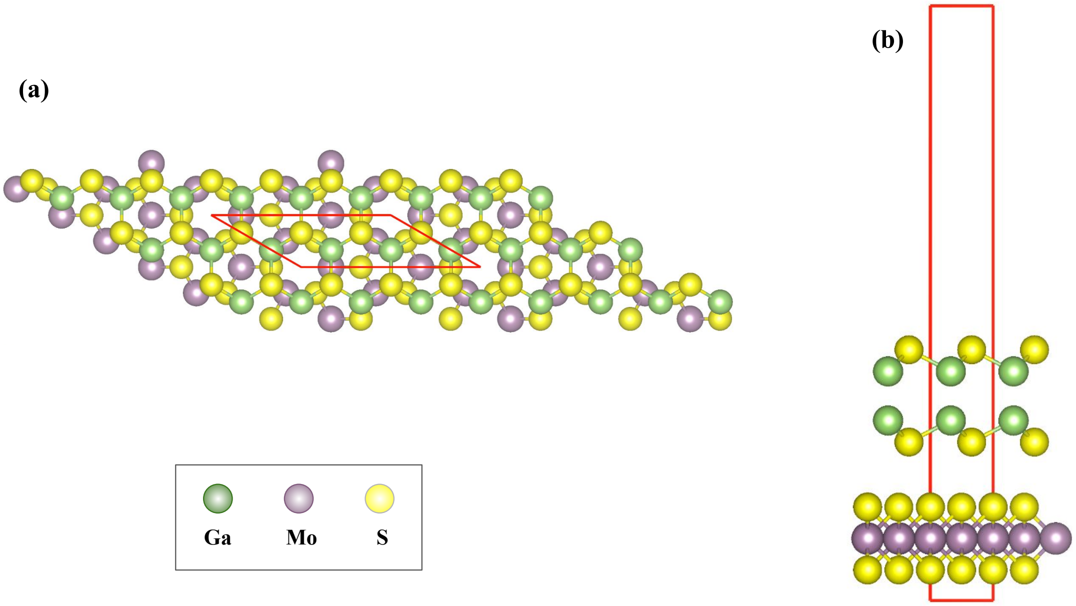


Figure 4.4: Crystal structure of the MoS_2/GaS heterostructure. This heterostructure is obtained as the structure with the lowest strain among all configurations when the 2D-MoS_2 and 2D-GaS planes on the x-y axes are vertically stacked and rotated on the z-axis. We initially set the distance between 2D-MoS_2 and 2D-GaS planes to be 3.6 Å, but it was reduced to 3.3 Åafter tight structural relaxation using DFT.

4.2.3 Heterostructure Properties

We conducted several DFT calculations to calculate and analyze the charge transport properties of the heterostructure when compared to that in the MoS_2. The orbital-projected density of electronic

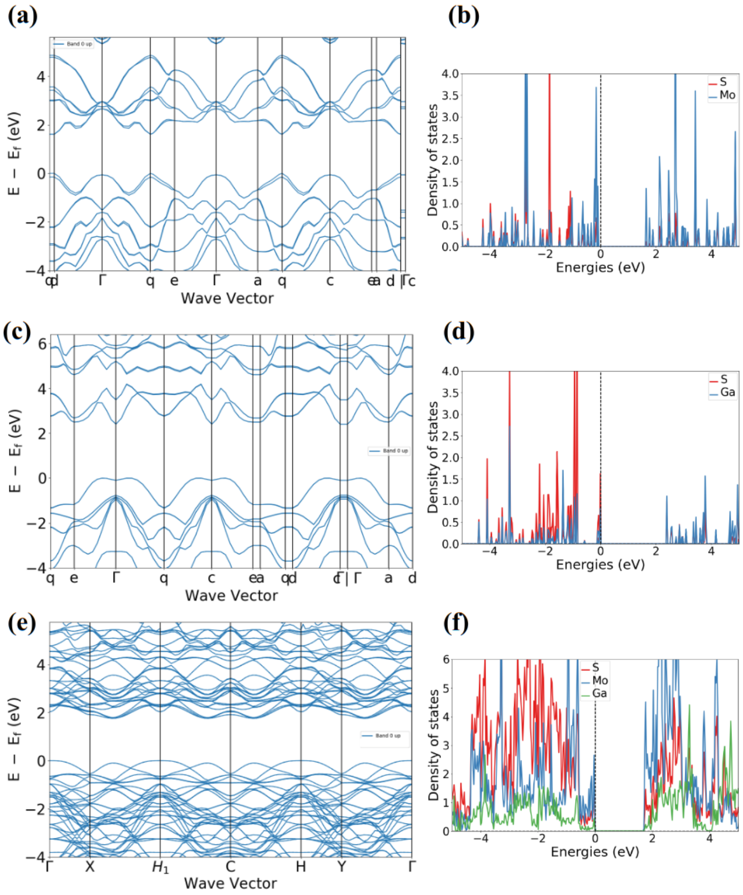


Figure 4.5: Electronic bandstructure and density of states (DOS) of **(a)** 2D-MoS_2, **(b)** 2D-GaS, and **(c)** the heterostructure from DFT calculations. The DOS in plot **(c)** shows that the orbitals of S and Mo dominate the VBM and CBM orbitals of the heterostructure

states(DOS) and the bandstructure of the 2D single phases and the heterostructure are estimated from DFT-PBE calculations and plotted in Figure 4.5. The orbital-projected band structure of the heterostructure is shown in Figure 4.6. The bandgap values obtained from the DOS for 2D-GaS, 2D-MoS_2, and the heterostructure are 2.41 eV, 1.61 eV, and 1.77 eV respectively. The 2D-MoS_2 bandgap is in very good agreement with the experimentally reported values[113]. A table of computed E_g values for 2D and bulk phases of GaS, MoS_2 alongside their reported values in the literature for reference are provided in Table 4.1. The heterostructure bandgap is close to that of the

2D-MoS$_2$, signaling a lesser deviation in the bandgap-dependent electronic properties between 2D-MoS$_2$ and the 2D-MoS$_2$/GaS heterostructure. It is also clear from the DOS and the band structure that the conduction and the valence bands are dominated by the d-orbitals of Mo and p-orbitals of S. The lack of dominance by the orbitals of Ga in the bandgap region is preferred to keep the charge transport happening in Mo and S bands similar to that in MoS$_2$.

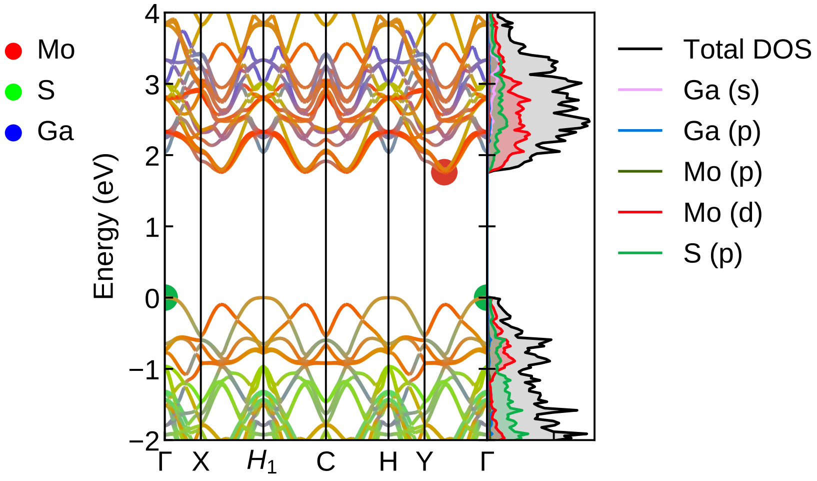


Figure 4.6: Orbital-projected band structure of the 2D MoS$_2$/GaS heterostructure. The conduction band of the heterostructure predominantly consists of the Mo d-orbital, suggesting an n-type carrier conductivity similar to that of pure 2D MoS$_2$. The heterostructure has an indirect bandgap of 1.8 eV. The direct bandgap of the heterostructure is also computed from this band structure as 1.9 eV.

We conducted slab surface calculations on the heterostructure and the 2D-MoS$_2$ to compute the work functions, CBMs, and VBMs. The CBM and VBM values computed from surface calculations are considered more reliable relative to the E$_{BP}$-based method, even though the latter is much cheaper to estimate in terms of computational resource requirements. The slab calculations enable considering the vacuum as the shared reference energy. The numerical value of energy assigned to the vacuum is subtracted from the rest of the energy levels in the band structure, effectively setting the energy of the vacuum to be 0 eV. The work function, which is calculated as the difference

Table 4.1: Bandgap values of bulk phases of MoS_2 and GaS, and 2D phases of MoS_2, GaS, and the heterostructure. The DFT-HSE bandgaps listed are obtained from the references cited, except in the case of bulk GaS whose DFT-HSE bandgap was calculated as a part of this work. The acronyms ”dir” and ”indir” refer to direct and indirect bandgaps, respectively

Material	Bandgaps (in eV)		
	DFT-PBE [This Work]	Experiment	DFT-HSE
MoS_2 (bulk)	1.4 (dir), 1.7 (indir)	1.3 (indir), 1.9 (dir)[114]	1.3 (indir), 1.8 (dir)[114]
GaS (bulk)	1.7 (dir)	2.3 (dir)[115]	2.1 (dir) (This work)
MoS_2 (2D)	1.6 (dir)	1.6 (indir), 1.9 (dir)[114]	1.5 (indir), 1.8 (dir)[114]
GaS (2D)	2.4 (indir), 2.5 (dir)	3.0 (dir)[116]	3.2 (dir)[117]
Heterostructure	1.8 (indir), 1.9 (dir)	-	-

between the fermi energy and the vacuum energy, becomes the absolute value of the Fermi energy when the vacuum energy is at 0 eV. Figure 4.7 shows the local potentials on 2D-MoS_2, 2D-GS, and the heterostructure in the direction perpendicular to the 2D plane. In all three cases, the potential surface flattens at the vacuum and shows no polarization or instability. The CBMs, VBMs, work functions, and bandgaps computed from the slab calculations are listed and compared against previously reported values in Figure 4.8. The work function values estimated for 2D-MoS_2, 2D-GaS, and the heterostructure are 5.6 eV, 5.7 eV, and 5.7 eV, respectively. Since the work functions of MoS_2 and the heterostructure differ only by 0.1 eV, the heterostructure creation is not expected to have a pronounced impact on MoS_2-based 2D inks in applications such as photoelectric devices, thermionic applications, electrocatalysis, etc. The CBM values of 2D-MoS_2 (-4.0 eV) and the heterostructure (-3.9 eV) are also observed to be close to each other, which is also expected from the dominance of conduction band states of the heterostructure by Mo d-orbitals as seen in the DOS and bandstructure (Figure 4.5e, 4.5f).

We also calculated the static dielectric and the elastic tensors, as given in Table 4.2. The dielectric constants along all three directions (xx, yy, and zz) are larger for the heterostructure than those in MoS_2. This difference is expected to alter the behavior of some of the devices made

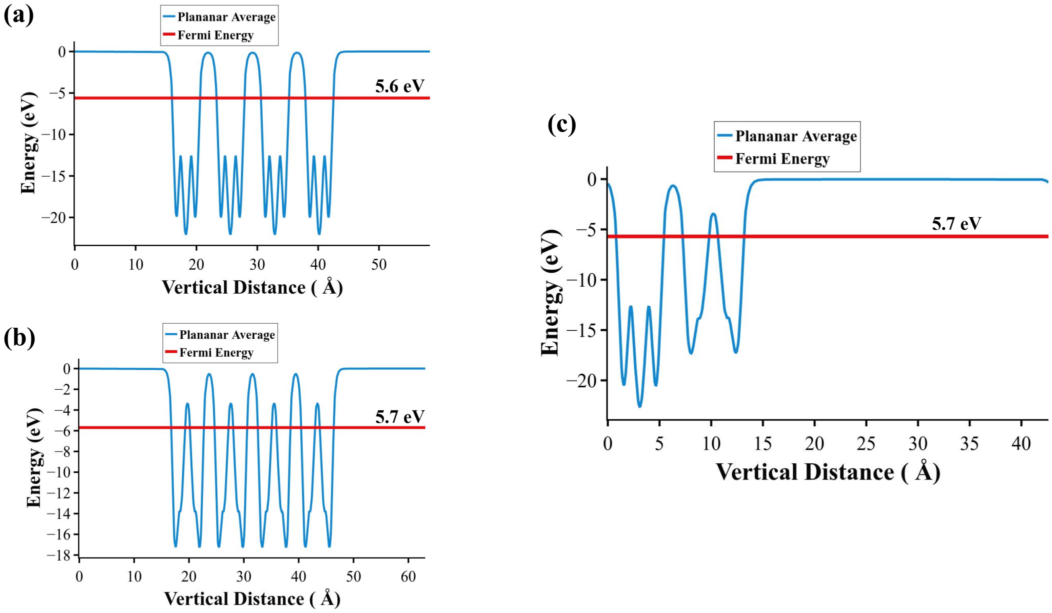


Figure 4.7: Local potential energies computed from the slab calculations. The plots **(a)**, **(b)**, and **(c)** show the energy distribution on the axis perpendicular to the 2D plane of 2D-MoS_2, 2D-GaS, and the heterostructure slabs respectively. In the case of both 2D-MoS_2 and 2D-GaS, the 2D plane is positioned in the middle of the unit cell with a 15 Å-long vacuum on top and bottom. In the case of the heterostructure, the slab is positioned at the bottom of the unit cell with a large vacuum of length 30 Åpresent only on top. Thus, the plots **(a)** and **(b)** have constant energy levels at low and high values of vertical distance while plot **(c)** has constant energy portions only at the higher values of vertical distance.

with MoS_2 2D inks such as capacitors. Even though it is not an ideal scenario for this particular work since we aim to keep the properties similar between the 2D-MoS_2 and the heterostructure. But since the dielectric constant values are higher in the heterostructure, it can be a more preferred material than even 2D-MoS_2 in applications where the higher dielectric constants are desired, for example, in charge storage applications[119]. The heterostructure's elastic constants are also computed to be higher than those in the 2D-MoS_2. Higher elastic constants are expected to benefit the 2D ink circuit components exposed to the external surface of a device and possibly subjected to external mechanical forces.

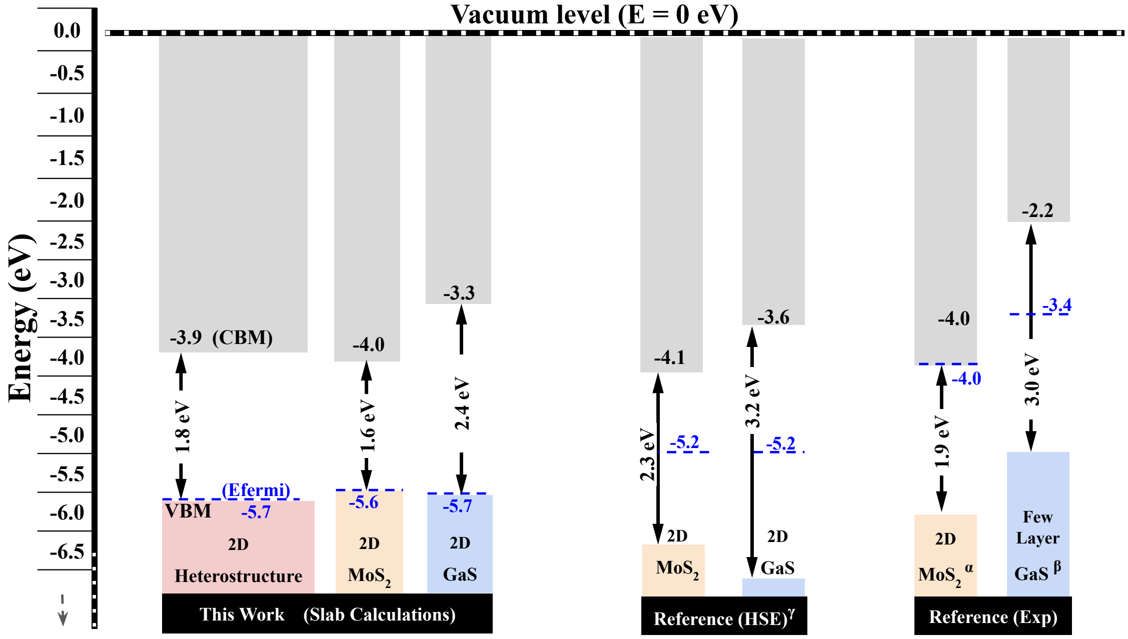
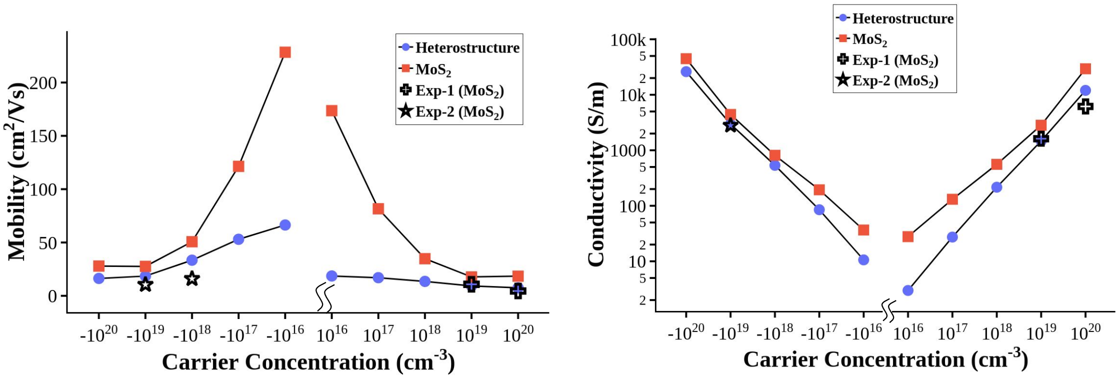


Figure 4.8: The positions of CBM (grey line), VBM (dark line), and the fermi energy (dotted line) are plotted for 2D-MoS$_2$, 2D-GaS, and the heterostructure as obtained from DFT-PBE calculations in this work, reported experimental works, and also from reported DFT-HSE works in the literature. The VBM, CBM, and fermi energy levels are marked with text on the left-most energy levels diagram (for the heterostructure). The work function is obtained as the absolute magnitude of the Fermi energy since we assume the energy of the vacuum to be 0 eV. References are as follows - α: Hu et al[118], β: Carey et al.[116], γ: Zhung et al[117]. The values of CBM and bandgap computed in this work for hexagonal 2D MoS$_2$ from slab calculations are in good agreement with the corresponding values reported from experiments. The CBM, bandgap, and work function values of 2D MoS$_2$ and the heterostructure are close to each other, which suggests a similarity in their electronic properties.

Table 4.2: The mechanical and dielectric properties of 2D-MoS$_2$ and the 2D-heterostructure. E, G, and v refer to Young's modulus, Shear Modulus, and Poisson's ratio, respectively

Material	Dielectric Constants			Mechanical Properties		
	ϵ_{xx}	ϵ_{yy}	ϵ_{zz}	E (N/m)	G (N/m)	v
MoS$_2$ (2D)	4.8	4.8	1.3	126	51	0.24
Heterostructure (2D)	12.1	12.1	5.6	217	88	0.23

Finally, the electrical conductivity, mobilities, and the Seebeck coefficients are calculated for 2D-MoS_2 and the heterostructure from Boltzman transport equations using the AMSET python package[120]. These properties are calculated at 300 K to reflect the material properties under ambient temperature conditions when used as a 2D ink. We report the electric charge transport properties computed at carrier concentration levels (in cm^3) of -10^{20}, -10^{19}, -10^{18}, -10^{17}, -10^{16}, 10^{16}, 10^{17}, 10^{18}, 10^{19}, 10^{20}. A negative carrier concentration indicates n-type carriers, and a positive carrier concentration indicates p-type carriers. Even though MoS_2 is reported as an n-type semiconductor and we focus on the conduction band charge transport behavior, the values in p-type doping are also reported here for comparison. Some of the recent studies have reported MoS_2-based devices having stable p-type carrier conduction[121]. Figure 4.9b shows a comparable conductivity trend for both 2D-MoS_2 and the heterostructure across a large range of doping concentrations. The experimental data reported in the literature also have a similar order of conductivity values at high carrier concentrations and are plotted alongside the computed values for comparison. The estimated conductivity values increase significantly from the order of 10^1 to 10^4 when the carrier concentration increases from 10^16 to 10^{20} in both n-type and p-type cases, indicating a linear trend between the carrier concentration and the conductivity. But the numerical values of the conductivity are larger in n-type doping than in p-type doping at the same carrier concentration for both MoS_2 and the heterostructure. Figure 4.9a shows the combined contribution of different scattering mechanisms to the electron and hole mobilities at different doping levels. The scattering mechanisms that are considered in this work are ionized impurity scattering (IMP), acoustic deformation potential scattering (ADP), and polar optical phonon scattering (POP) - all as implemented in the AMSET package[120]. The mobility is significantly lower in the heterostructure compared to MoS_2 at low doping levels. As expected, the carrier mobilities decrease with an increase in doping for both materials because of increased scattering from the extra carriers.



(a) **(b)**

Figure 4.9: Results from Boltzmann calculations of 2D-MoS_2 and heterostructures at 300 K and different n-type (negative carrier concentration values) and p-type (positive carrier concentration values) doping concentrations. **(a)** Mobility of carriers alongside experimentally reported values in published literature for comparison. The mobilities of 2D-MoS_2 and the heterostructure are close to each other at high carrier concentrations, while the heterostructure has significantly lower values at lower carrier concentrations. The computationally obtained values for 2D-MoS_2 agree with the experimental values at higher carrier concentrations. **(b)** The conductivity values of 2D-MoS_2 and the heterostructure are close to each other in a log scale at all carrier concentrations. The experimentally reported values of 2D-MoS_2 from published literature agree with the calculated values

The convex hull phase diagram of Mo-S-Ga composition is obtained from the OQMD database and shown in Figure 4.10. The presence of a tie-line between GaS and MoS_2 indicates that DFT predicts them to be thermodynamically stable when in contact with each other[122]. Other advantages of using GaS as a bridge-material are that neither Gallium nor Sulfur is among the generally known toxic elements, and their abundance in the earth's crust is sufficient[123] to be considered for large-scale industrial manufacturing.

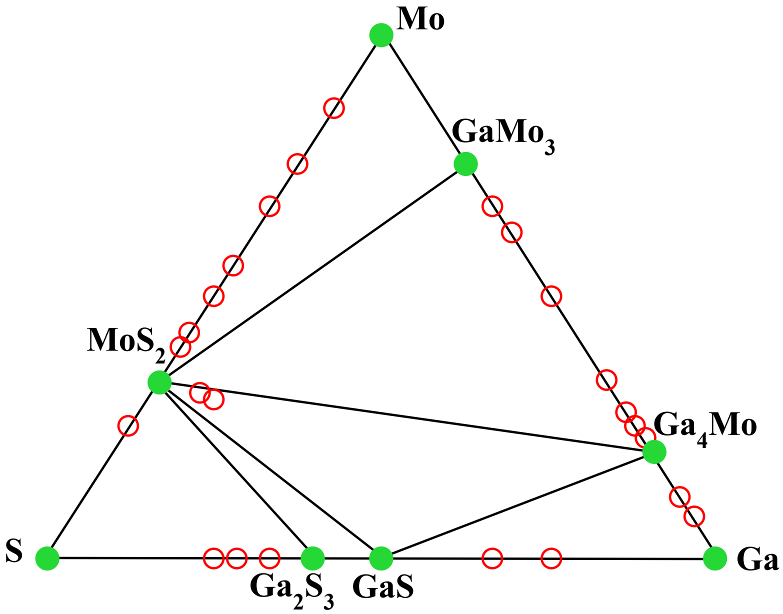


Figure 4.10: Phase diagram of Mo-Ga-S elemental phase space obtained from OQMD (as of October 14, 2022). The green-filled circles and the unfilled, red circles represent the stable and unstable compounds present in this phase-space. The presence of a tie-line between two compositions indicates the thermodynamic stability at 0 K when these two compounds are in contact with each other. Based on this phase diagram at 0K, using GaS as a bridge material in MoS_2 inks is not expected to have thermodynamic instabilities.

4.3 Discussion

In this work, we have identified hexagonal GaS as an excellent candidate to be used as a bridge-material in MoS_2-based 2D inks. The heterostructure is created with MoS_2 and GaS lattices stacked vertically with less than 5% strain on lattice vectors. The Boltzmann transport calculations based on data obtained from DFT indicate that the electronic conductivity values of the 2D-MoS_2 remain similar in many doping concentrations upon forming a heterostructure with GaS. The mobility is reduced upon heterostructure formation at lower doping concentrations, but they are similar at higher carrier concentrations ($> 10^{17}$). The conduction bands of the 2D-MoS_2 and MoS_2/GaS

2D-heterostructure are calculated to be aligned well within the required limits. The work function value of the heterostructure (5.7 eV) is also estimated to be close to that of 2D-MoS2 (5.6 eV). We also report the thermodynamic stability of GaS and MoS_2 when in contact with each other by plotting a convex hull phase diagram of the Mo-Ga-S phase space derived from the formation energy values listed in the OQMD database.

This work also shows the successful application of a machine learning-aided heuristic method in materials selection when a high-throughput computational screening of thousands of materials is infeasible. We used a multi-fidelity Co-Kriging model, which was built as a part of another work (Chapter 5) to predict bandgaps of candidate materials at DFT-HSE accuracy. The bandgap in electronic bandstructures is corrected using this model's predictions before calculating conduction band alignment.

4.4 Methods

4.4.1 Branch point Energy Calculations

Branch Point Energy (E_{BP}) calculations are considered a fast and computationally cheap method to quantify the band alignments. This method relies on finding the intrinsic charge neutrality level, E_{BP}, of a given semiconductor where the valence band-like behavior and conduction band-like behavior cancel each other inside the bandgap[124]. It is qualitatively shown in Figure 4.11, and the related quantitative estimation of E_{BP} is provided in Equation 4.1. At a heterojunction of two semiconductors, a charge transfer occurs due to the offset in the Fermi-level alignment of the semiconductors. This, in turn, causes the formation of an interfacial dipole. But the dipole causes a certain alignment of electronic bands along materials' E_{BP} energy levels, eliminating the dipole itself[105], [106]. Hence, the E_{BP} acts as a common reference energy level for the interface and makes it possible to quantify band alignments between materials. E_{BP}-based band alignment is less accurate than the results of 2D crystal slab calculations in DFT with vacuum as the reference energy point since the former method does not consider the surface properties. But E_{BP}-based calculations are significantly cheaper and quicker to perform on thousands of materials when their bulk bandstructure is already known.

$$E_{\text{BP}} = \frac{1}{2L_{\text{path}}} \sum_{j}^{N_{\text{seg}}} L_j \sum_{\mathbf{k}\in j} \frac{1}{N_{\mathbf{k},j}} \left(\frac{1}{N_c} \sum_{c_i}^{N_c} \epsilon_{c_i}^{\text{QP}}(\mathbf{k}) + \frac{1}{N_v} \sum_{v_i}^{N_v} \epsilon_{v_i}^{\text{QP}}(\mathbf{k}) \right) \tag{4.1}$$

Here, c and v represent conduction and valence bands, while N_c and N_v represent the number of electrons in those bands, respectively. L_j is the length of each segment j in the computed bandstructure that contains N_{seg} segments in total. Each such segment, j, has $N_{\mathbf{k},j}$ number of k-points. ϵ is the energy of a band and L_{path} is the total length of the path. In this work, we have taken $N_c = \frac{N_e}{8}$ and $N_v = \frac{N_e}{4}$ where N_e is the sum of number of valence electrons in s and p orbitals.

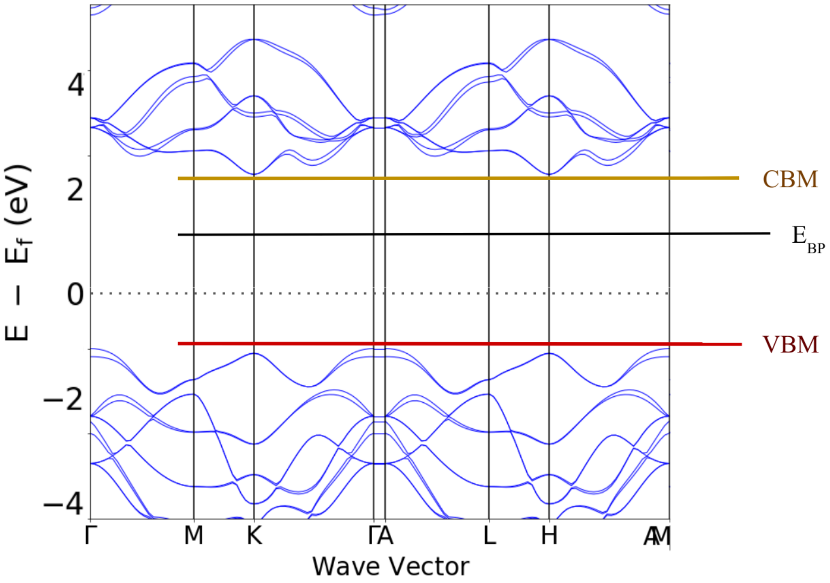


Figure 4.11: A qualitative diagram showing the VBM, CBM, and E$_{BP}$ within an electronic bandstructure.

Once the E_{BP} value is computed in a bandstructure, all the energies are shifted by the -E_{BP}

eV to make E_{BP}=0 eV. After shifting the bandstructures in this manner for two materials, their new valence band minima (VBM) can be compared by their numerical values in eV. To obtain the conduction band minimum (CBM), the bandgap value is added to the VBM. Thus, the CBM values are sensitive to the accuracy of the bandgap values. Since the PBE functionals in DFT tend to underestimate the bandgaps, we used an ML model to predict a more reliable bandgap for all the materials under consideration and used the predicted bandgaps to calculate CBM from VBM. The ML model we used in this work is obtained from a different work described in Chapter 5. This ML model is created and validated to reliably predict the bandgaps of materials at the level of accuracy expected from DFT calculations with HSE hybrid functionals.

4.4.2 DFT Calculations

All calculations in this work are done using the VASP package, and all of them were done with tight convergence settings. The K-points per reciprocal atom (KPPRA) is set to be 8000, and the EDIFF is set to be 10^{-8} for both relaxation and static calculations. The PBE functionals are used in all the calculations to approximate the exchange-correlation energy. The structural relaxation is conducted with high accuracy and tight convergence criteria to minimize the forces of atoms by setting EDIFFG=-10^{-3}. For the heterostructure calculations, the van der Waals dispersion corrections to the DFT energies are included using DFT-D3[125], [126], as implemented in VASP.

A widely accepted, high-accuracy method to quantify band alignment from DFT is the 2D-slab calculation of the material placed in the vacuum. In this method, the vacuum with an energy of 0 eV acts as a common reference point for both bandstructures. Surface calculations are of high accuracy but are computationally expensive and time-consuming. Thus, using limited computational resources, it is not viable to do surface calculations for thousands of materials. In this work, surface calculations are conducted for small unit cells of 2D-GaS, 2D-MoS_2, and also for the large unit cell of the 2D heterostructure.

4.4.3 Conductivity Calculations

The conductivity and mobility values reported in this work are calculated using Boltzmann transport equations on VASP calculation results, as implemented in the AMSET package[120]. We conducted dense uniform bandstructure calculations, finite difference calculations to estimate the elastic constants, deformation potential calculations, and the density functional perturbation theory (DFPT) calculations to estimate dielectric response and the polar phonon frequency with the tight convergence parameters. The polar phonon frequencies of 2D-MoS_2 and the heterostructure were computed to be 9.09 THz and 11.30 THz, respectively.

AMSET calculates Seeback coefficients (S), in addition to conductivity values and mobilities even though estimation of the thermoelectric properties of MoS_2 2D inks are not among the goals of this work. The S values of MoS_2 and the heterostructure at 300 K are shown in Figure 4.12 and are close to each other. The absolute magnitude of S decreases with an increase in the carrier concentrations.

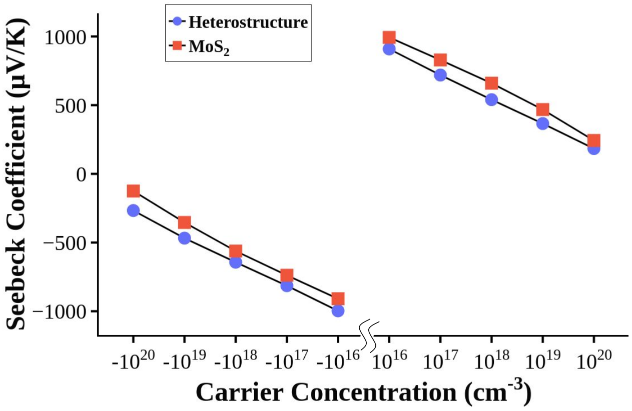


Figure 4.12: Results from Boltzmann calculations of 2D-MoS_2 and heterostructures at 300 K and different n-type (negative carrier concentration values) and p-type (positive carrier concentration values) doping concentrations. **(a)** Seebeck Coefficients of 2D-MoS_2 and the heterostructure show the values close to each other for the two materials at all carrier concentrations. Small absolute values are calculated for higher carrier concentrations.

CHAPTER 5
ASSESSING THE ACCURACY OF DFT CALCULATIONS FROM MULTIFIDELITY MODELING OF BANDGAPS

5.1 Background

High throughput databases containing computational materials data generated using DFT have changed the domain of materials selection in the past decade[3], [7]–[9], [127], [128]. An important property data available from most of these inorganic materials databases is the electronic bandgap value of materials obtained from DFT using PBE[19] version of GGA functional to approximate the exchange-correlation energy. The bandgap values play vital roles in material selection for devices such as transistors (semiconducting behavior), capacitors (insulating behavior), etc. The DFT-PBE calculations are cheaper to compute, easier to automate, and generally more accurate than gradient-less approximations such as Local Density Approximation (LDA) to the exchange-correlation functional. But it is well known that DFT-PBE tends to underestimate the bandgaps often and results in incorrect classification of electronic materials into metals, semiconductors, and high-bandgap insulators[129], [130]. More accurate approximations to exchange-correlation than PBE are available today, and one of them is HSE[20] hybrid functional. In HSE, the bandgap values are calculated more accurately than PBE when compared with the experimental values[129], [130], but they are also significantly more expensive to compute than PBE as shown in Figure 5.1.

There are more than a million materials in the OQMD database, out of which nearly 50,000 of them are thermodynamically stable at 0K, and within that, about half of them are initially obtained from the ICSD crystal structure database of experimentally synthesized compounds. Material selection strategies for novel electronic materials may look through this set containing nearly 25,000 experimentally synthesized, stable compounds and make decisions explicitly or implicitly based on the value of DFT-PBE bandgaps (E_g^{PBE}). An underestimation of bandgap from DFT-PBE on a real semiconductor can incorrectly label it as metallic with no bandgap. This would unnecessarily filter the real semiconductor out from the search-space of semiconductors in a design workflow. Because of that, it is crucial to correct the bandgap values of materials in HT databases. The best computational solution to achieve this is by re-calculating the bandgaps of all such materials using hybrid functionals, but it would become too expensive in terms of computational resources and time consumption. Mohan et al[131] computed the DFT-HSE bandgaps (E_g^{HSE}) of 1117 materials in the OQMD database and recorded them alongside their DFT-PBE bandgaps. In this work, we use this dataset of E_g^{PBE} and E_g^{HSE} to create a machine learning model that can predict the bandgaps at the accuracy level of DFT-HSE ($E_g^{PredHSE}$) when the E_g^{PBE} value and crystal structure information of a material is known. We also analyze the uncertainty quantification from the ML model to reduce the chances of selecting false-positives while searching for misclassified semiconductors among materials with E_g^{PBE}=0 eV.

Multifidelity modeling[132]–[134] is a specialized statistical learning algorithm designed to work on datasets with more than one estimation for the target property available with varying levels of accuracy (fidelity) and cost requirements. A method that provides higher accuracy in the estimation of the target property most often also requires a higher computational cost. In this work, E_g^{PBE} (low fidelity) and E_g^{HSE} (high fidelity) are estimations of material bandgaps at two different fidelities. Multifidelity modeling methods are already reported to have been successfully applied to material datasets over the past few years[132], [135]. Pilania et al[132] applied multi-fidelity modeling to a similar problem where they predicted the E_g^{HSE} values of a specific class of

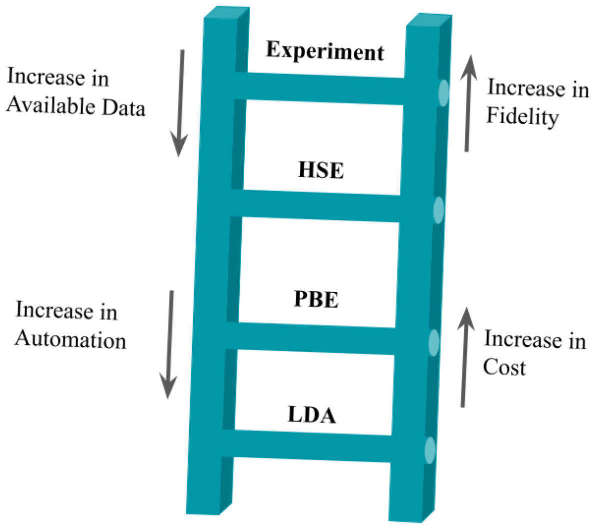


Figure 5.1: A partial diagram of Jacobs ladder in DFT[136]. This section of the ladder includes only those XC-functional approximations which are relevant to this work.

materials with high accuracy. Here, we use the same algorithm, called co-kriging, to a generalized set of materials belonging to different structural and composition classes and focus on practically integrating the model into a big HT database such as OQMD. Our goal in this work is to correct the materials data in the database and strengthen the general material selection processes, instead of focusing on the model benchmarking results alone.

5.2 Results

5.2.1 Data and Modeling

Mohan et al.[131] generated a database containing E_g^{HSE} and E_g^{PBE} data of 1117 materials in OQMD using a High Throughput (HT) DFT framework. We use this dataset, also referred to as the training-data, to train a multi-fideity co-kriging machine learning (ML) model that inputs the E_g^{PBE} and several structural and chemical attributes of a given material and predicts its E_g^{HSE}. The trained model is used to predict the E_g^{HSE} of 24,967 stable materials from OQMD whose E_g^{PBE} are already available but the E$_g^{HSE}$ values are not yet calculated from DFT. This dataset on which the E$_g^{HSE}$ is predicted by the ML model is also called the search-space. All of the materials included in the training-data are excluded from the search-space. The distribution of E$_g^{HSE}$ in the training-data, as plotted in Figure 5.2, shows a significant imbalance in the numbers of metals and non-metals. 62% of the materials in training-data are metallic with E$_g^{HSE}$=0 eV.

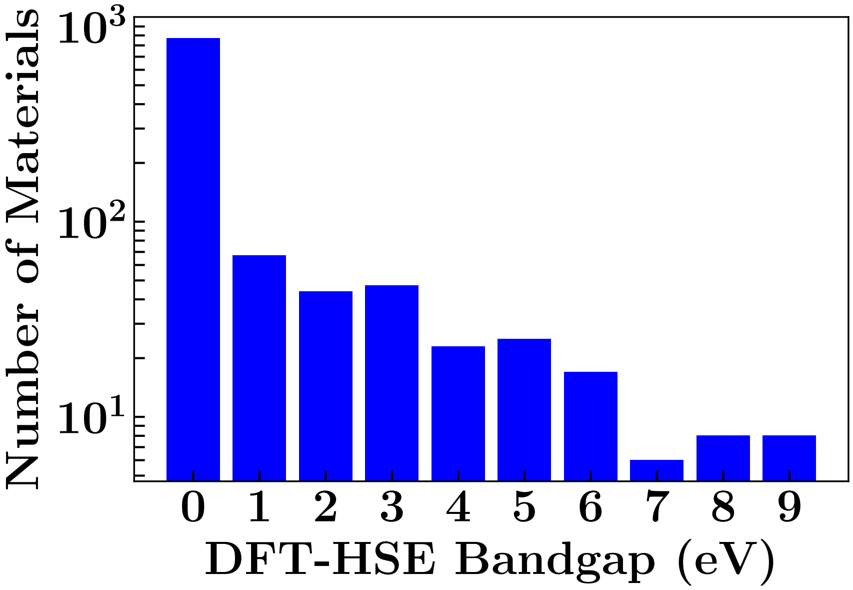


Figure 5.2: Distribution of DFT-derived E_g^{HSE} values in the training-data. The E_g^{HSE} is zero in 62% of the materials in training-data.

The structural and chemical features that are required to represent a material, in addition to the E$_g^{PBE}$, inside the co-kriging model are generated using Magpie[39] package. Examples of the Magpie features include the average number of neighbors inside the unit cell, the average electronegativity of atoms, etc. The Magpie feature vector is of size 272, and appending the E$_g^{PBE}$ makes the final input feature size to be 273 for every material. General feature dimensionality reduction methods such as principal component analysis and model-based selections are used to reduce the feature vector size to 100 before starting the co-kriging model training.

co-kriging is an extension of ordinary Kriging, a type of Gaussian process regression, designed to learn from a dataset that has data on the same property estimated at two different levels of accuracy (fidelity). The co-kriging method is originally implemented to learn knowledge even when the low-fidelity data is unavailable for some of the training-data. But in this work, we use the low-fidelity data (DFT-PBE bandgap) as one of the input features as well - making it mandatory that all materials in training-data and search-space should have their DFT-PBE already calculated. This requirement is not expected to pose an issue in the deployment of the model because most of the large HT databases already have the DFT-PBE bandgap calculated. The trained co-kriging model predicts the high-accuracy bandgap (at the HSE level) at the expense of doing a significantly cheaper-to-compute DFT-PBE bandgap calculation on a new material. A short description of the co-kriging is described below in Equation 5.1. A longer description of the algorithm is provided in the Methods section toward the end of this chapter.

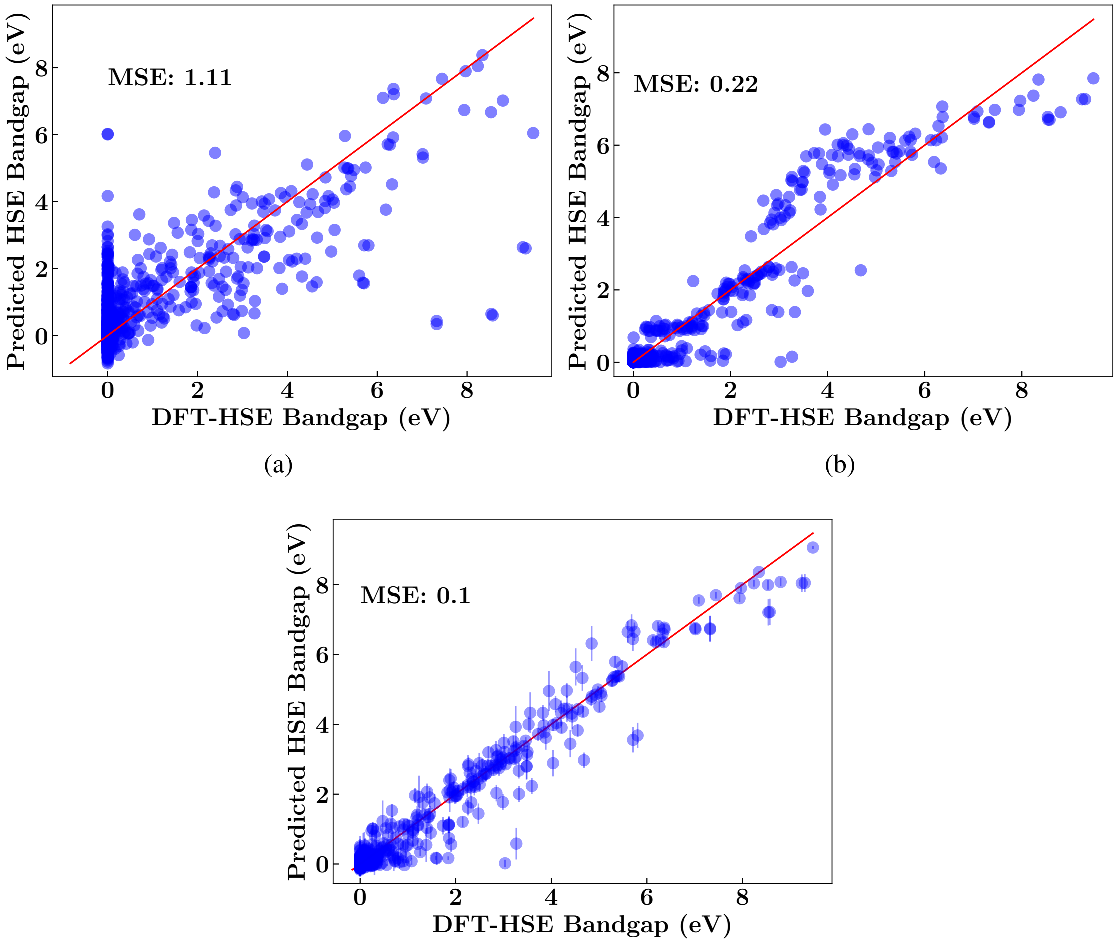
$$Z_{HSE}(x) = \rho Z_{PBE}(x) + Z_d(x) \tag{5.1}$$

Here, $Z_{HSE}(x)$ and $Z_{PBE}(x)$ are Gaussian processes which represent $E_{g,HSE}$ and $E_{g,PBE}$ respectively. The term Z_{PBE} is multiplied by a scaling parameter ρ whose value is optimized during the model training via Maximum Likelihood Estimation (MLE). The third term, Z_d, is the Gaussian function representing the difference between the high-fidelity and scaled low-fidelity processes. Since the $E_{g,PBE}$ is already known for all materials in training-data and search-space in this work, the evaluation of $Z_{PBE}(x)$ is less significant here than in the original co-kriging framework. The primary advantage of doing co-kriging over other non-Gaussian ML models is that it is less biased toward the training-data imbalance. We conducted a set of initial benchmark modeling to compare the performance of different models in the same test data - which is a split from the training-data - as shown in Figure 5.3. The random forest model and the support vector regression model suffer from low accuracy fits in this situation due to a large number of materials being predicted to be metals, which in fact, is due to the larger representation of metals in the training-data split. We have considered a data-pessimistic situation in Figure 5.3 by splitting the full training-data of 1117 materials into benchmark training:test data ratio of 2:8. Such a split of data with significantly more materials in the test data represents the realistic data scenario of this work where the full training-data (1117 materials) is significantly less than the search-space (24,967 materials). The relatively lower sampling of feature vector space by the training-data compared to a larger area spanned by the search-space can lead to a higher prediction error. In the ML model benchmarking results, the underfit of ML models due to training-data imbalance is observed to be the least influential in the case of co-kriging.

The exact model prediction errors on search-space are unverifiable until after the expensive DFT-HSE calculations are performed. Thus, it is necessary to determine the uncertainty of the model in predicting the HSE bandgap of each search-space material. Model uncertainty quantification helps avoid less-confident material selections by discarding the predictions with high uncertainty. This can benefit situations where only a limited amount of resources are available for subsequent DFT-HSE validations on the selected materials. Since the $Z_{PBE}(x)$ is already known, the advantage of using co-kriging over using regular Gaussian regression arises from the much higher dimensional covariance matrix involved in the co-kriging that explicitly considers the distribution of data points in both low- and high-fidelity domains of the training-data. A detailed discussion of the co-kriging model is provided in the Methods section.

5.2.2 Predictions and Uncertainty Quantification

We trained the co-kriging model on a train:test split of 8:2 with intense optimization of model parameters and feature-selection algorithms to use it as the final model for predictions on search-



(a)

(b)

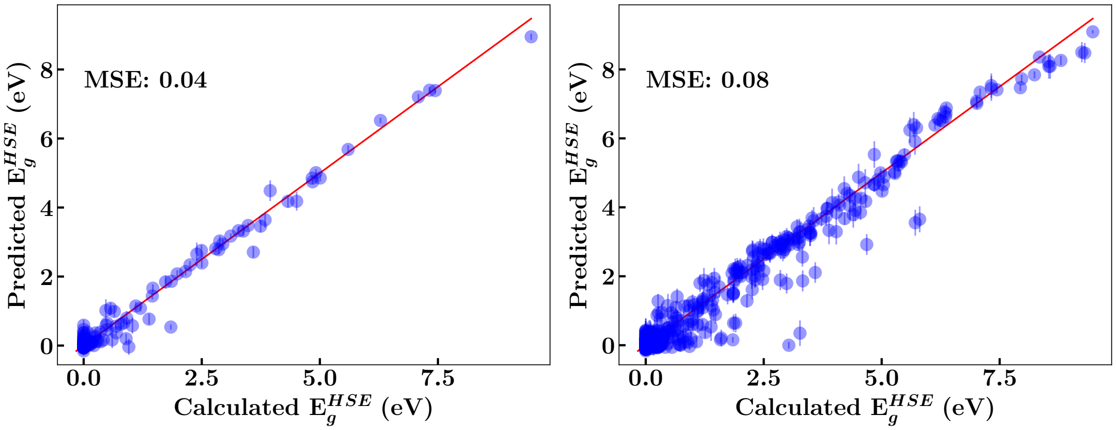
(c)

Figure 5.3: Performance of different machine learning models in the same test data sampled from the DFT-HSE dataset containing E_g^{HSE} values. The train:test split is 2:8, representing a scenario when the training-data may not have sufficiently sampled the feature space spanned by the search-space. All the models were trained on the same training-data. Both random forests regression (RFR), shown in subplot (a), and support vector regression (SVR), shown in subplot (b), suffer from the unbalanced training-dataset biased toward E_g^{HSE}=0. Since the majority of the materials in the training-data have E_g^{HSE}=0, this behavior is expected from RFR and SVR models. But as shown in subplot (c), co-kriging is significantly less biased from the training-data imbalance, and thus, more accurate at a lower mean squared error (MSE) than the other two models.

space materials. The test-data prediction accuracy of this model is plotted in Figure 5.4a. In Figure 5.4b, we use the same model architecture and the same feature-selection pipeline to build a model on a pessimistic train:test split ratio of 1:9. The difference between the models built on pessimistic data in Figure 5.3c and in the Figure 5.4b is that the former is created with modest parameter optimization and without optimizing feature-selection pipeline specifically for co-kriging to achieve an unbiased comparison with other ML models, while the latter co-kriging model is built on a model architecture and feature-selection pipeline that is fully optimized to deliver the best prediction accuracy. Henceforth, the HSE-level bandgap predictions on search-space materials used in this work are generated by the optimistic co-kriging model shown in Figure 5.4a. More details on the final model architecture and the feature-selection pipeline are provided in the Methods section.

The reliability of quantified model uncertainty in the final co-kriging model on test data is shown in Figure 5.5c. In the case of the final model with an optimistic data split of 8:2, the model estimates a high uncertainty value in 60% of the cases where there is an observable error (>0.1 eV) in the model predictions. Thus, filtering out co-kriging predictions with high uncertainty during material selection can significantly improve the probability that the selected material has the expected behavior. This is equivalent to avoiding false-positive predictions in categorical classification ML modeling. Only in 12% of all cases, the materials that are estimated to have low uncertainty in predictions also had an inaccurate HSE bandgap prediction. The focus on increasing the precision of ML model predictions is necessary when the resources to experimentally or computationally validate the ML model-based selections are limited. A similar analysis from the pessimistic data model is shown to the right of the final, optimistic model's data in Figure 5.5c to compare the performance in such conditions.

The predictions of HSE bandgap on search-space data by the final co-kriging model (Figure 5.4a) are analyzed across different classes of materials in Figure 5.6a with a focus on accurately

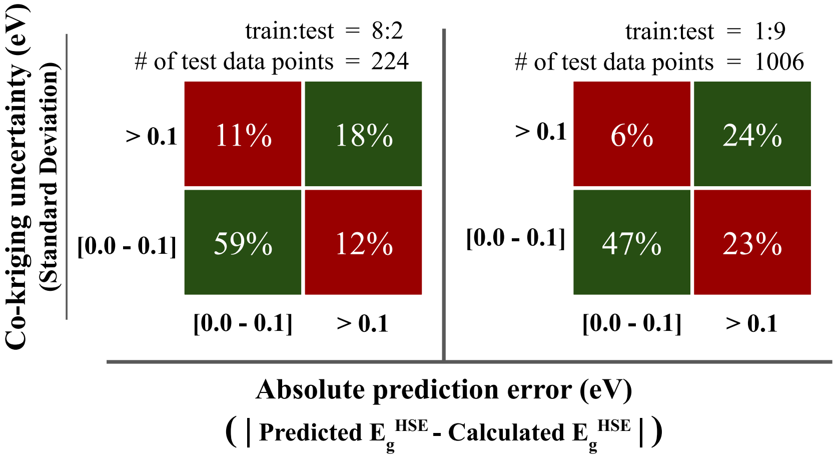


(a) Optimistic Model

(b) Pessimistic Model

Figure 5.4: Multifideity co-kriging model benchmarking results on HSE bandgap dataset of 1117 materials. A part of the full training-data is split and set aside as test-data during model training. The trained model's predictions on the test-data are plotted in these figures. The plot **(a)** has an optimistic train:test split of 8:2, which imitates a situation where training-data is large enough to reliably learn the correlation between input features and the target property, E_g^{HSE}. In such a situation, the vector space spanned by input features is sufficiently sampled by the training-data. The plot **(b)** shows the predictions from a different model trained and tested on a pessimistic train:test data split of 1:9. The pessimistic benchmarking was done to examine the prediction capability of co-kriging model in situations where training-data is not large enough to fully represent the relatively larger portion of the feature space spanned by the candidate materials. The uncertainty quantification analysis from these two models is shown in Figure 5.5

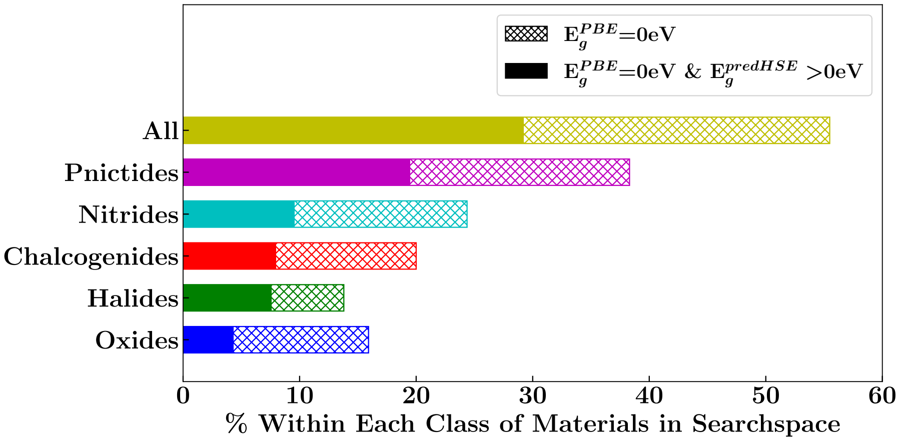
classifying a material as metallic or insulating. The co-kriging predicts that about half of the materials predicted as metals by DFT-PBE, may actually be semiconductors or insulators. This amounts to about 30% of all materials in search-space where the DFT-PBE and co-kriging model disagree on the material's conducting nature. The DFT-PBE calculations and the co-kriging predictions are in most agreement for oxides even though more than 80% of the oxides are already estimated to be insulating by DFT-PBE. The uncertainty prediction statistics of search-space materials with a focus on their constituent elements are shown in Figure 5.6b. In 12% of all hydrogen-containing compounds, a high uncertainty is associated with the model predictions.



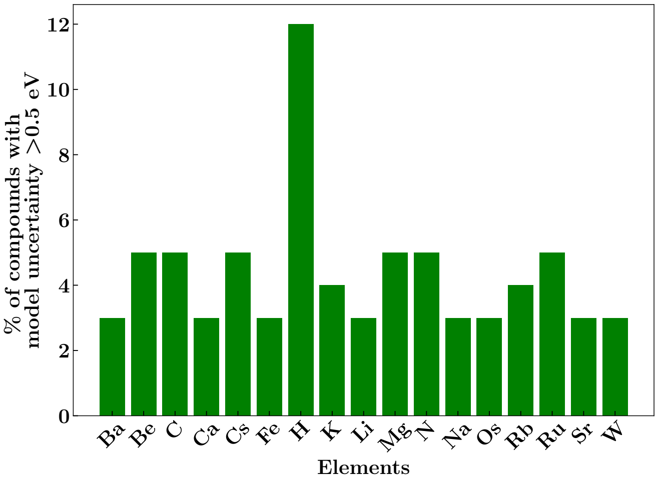
(a) co-kriging Model

Figure 5.5: The multi-fideity co-kriging model's quantified uncertainty predictions on the test-data are plotted as a part of model benchmarking. The corresponding ϵ prediction plots from these same models are shown in Figure 5.4. The train-test split is done on the full training data containing HSE bandgap values of 1117 materials. The uncertainty quantified for each test data point by the co-kriging model during the prediction is analyzed in two data conditions. In both optimistic (left) and pessimistic (right) cases, the materials in test data with a large prediction error, which can be considered as a "negative" prediction, also had a large uncertainty predicted by the co-kriging model. This shows a higher true-negative rate and a smaller false-positive rate. Considering the accurate prediction of HSE bandgaps from co-kriging as a "positive" prediction, it is shown to be possible to largely avoid false-positive predictions if the materials that have a large predicted uncertainty are excluded from the candidate list. In the optimistic case, the precision and recall based on true-positive (small error, small uncertainty), false-positive (large error, small uncertainty), and false-negative rates (small error, large uncertainty) are 83% and 84%, respectively

The general trend of the variation between E_g^{PBE} and $E_g^{predHSE}$ is analyzed in Figure 5.7. The well-known systematic underestimation of bandgaps in DFT-PBE framework is captured by the co-kriging model in the low-medium bandgap range. But at very-high bandgap limits, the co-kriging predicts a lower bandgap than the DFT-PBE. The LOWESS smoothing curve[137] fit on the E_g^{PBE} vs $E_g^{predHSE}$ plot in Figure 5.7a shows the suggested crossing of the DFT-PBE underestimation to overestimation near 7 eV. That said, this result is speculated based on the co-kriging model predictions, and not from true DFT-HSE calculations. Because of that, a reliable argument on such



(a)

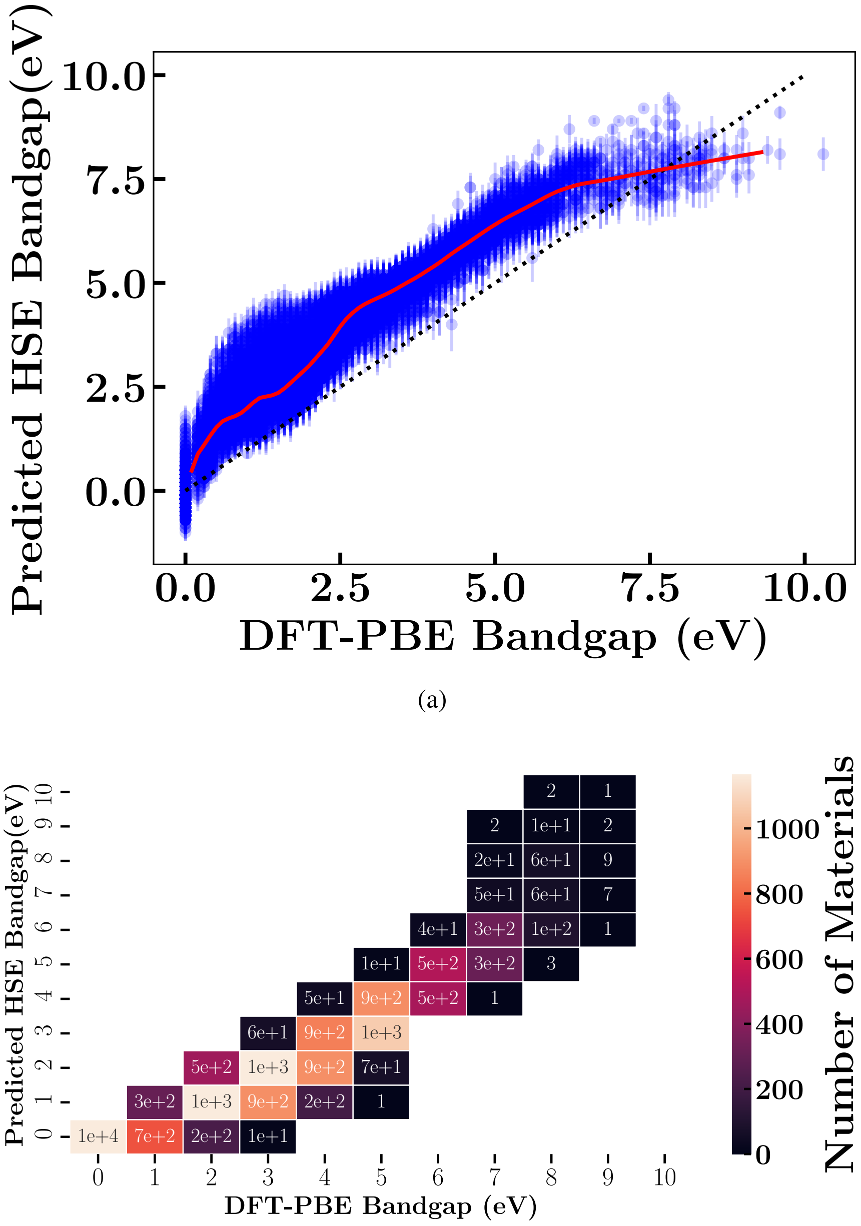


(b)

Figure 5.6: **(a)** Bandgap openings predicted by the multi-fideity model. $E_g^{predHSE}$ refers to the HSE-fidelity bandgap value predicted from co-kriging. In about 30% of all materials in the search-space, the PBE results and co-kriging model predictions disagree on whether a material is metallic or not. **(b)** Elemental distributions among the materials in search-space with the highest values for model uncertainty. The ordinate of the bar plot represents the percentage of the compounds with a high uncertainty prediction among all the compounds in the search-space that contains the element specified in the abscissa of the plot. The bars of only those elements are shown which have an ordinate value of more than 2%. Such a cutoff is kept to make the relevant information stand out and skip other elements, such as O, F, etc., that have less than 2% of the compounds predicted to have a high uncertainty value. The model uncertainty is quantified as the standard deviation of the predicted co-kriging Gaussian distribution of $E_g^{predHSE}$.

a change in trend from underestimation to overestimation will require a high-throughput study of high-bandgap materials that is out of the scope of this work.

We conducted a literature survey on 50 low-E_g^{PBE} ($<$3 eV) materials with the highest disagreement between E_g^{PBE} and $E_g^{predHSE}$. Experimental or HSE bandgap data reported on peer-reviewed articles were available for 28 of them, and are provided in Figure 5.6 for quick comparison of the reliability of co-kriging in predicting materials that are incorrectly marked as metals by DFT-PBE. The same data is also provided in Table 5.1 with external references for numerical comparison. The co-kriging model correctly predicts the bandgap openings in the 13 materials that we considered. Among the other 15 materials shown in Figure 5.6, co-kriging estimates a more reliable value in 13 materials compared to DFT-PBE. Even though the co-kriging predictions agree better with the published experimental and DFT-HSE results, in some situations, like in the case of Fe_3O_3 and $LaVO_3$, the co-kriging significantly overestimates the bandgap. But overall, within the scope of this small set of materials surveyed in literature, the co-kriging has proved effective in identifying non-metals that are incorrectly classified as metals in DFT-PBE calculations.



(a)

(b)

Figure 5.7: (**a**)Predicted HSE bandgaps of search-space materials from the co-kriging Model are plotted against the corresponding DFT-PBE bandgaps. A LOWESS smoothing function[137] is also plotted to show the trend of how the co-kriging predictions change with the DFT-PBE bandgap. The underestimation of bandgaps in DFT-PBE compared to DFT-HSE is well known, and that same behavior is also seen between DFT-PBE and co-kriging model predictions. A reversal of that trend is seen near 7 eV. (**b**) Distribution of the number of materials in a heatmap between DFT-PBE bandgaps and the co-kriging predictions.

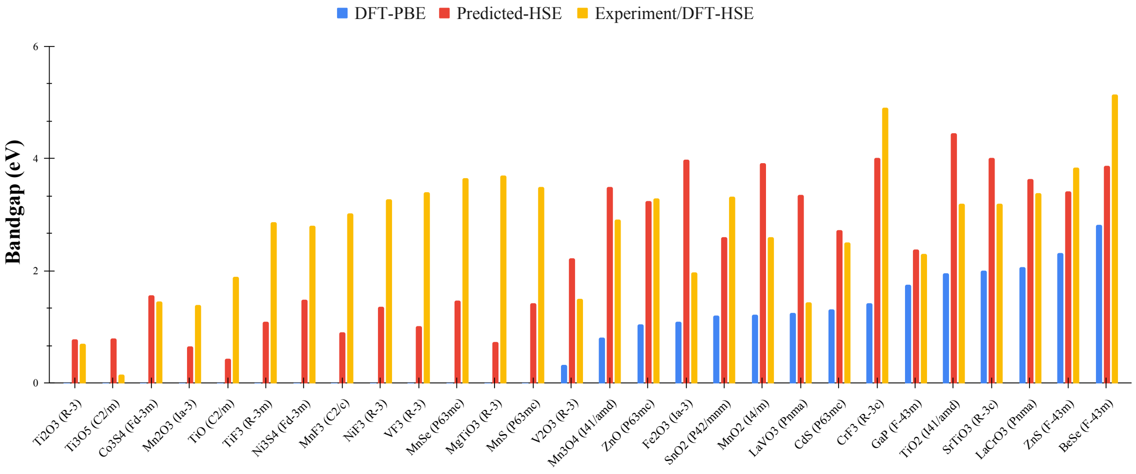


Figure 5.8: Verifying the disagreements of DFT-PBE and co-kriging when compared against the published experimental or DFT-HSE results in scientific literature. None of the DFT-HSE values in this chart are from the training-space data used in this work. The crystal structure of each material is provided in brackets beside its chemical formula. The full set of references for Experimental/DFT-HSE data is provided in the manuscript currently under author review[131]

Table 5.1: Literature comparison of co-kriging model predictions on search-space. The $E_g^{PredHSE}$ column refers to the HSE bandgap value predicted by the co-kriging model. These materials are selected based on the difference in their $E_g^{PredHSE}$ and DFT-derived E_g^{PBE}. Information about many other materials which were filtered out from search-space based on their bandgap value differences did not have any reported values on DFT-derived E_g^{HSE} or experimental bandgap in scientific literature within the scope of our search. Further details about the listed materials, including crystal structure and DFT (with PBE XC-functionals) calculation details, can be found on the material's web page identified by OQMD ID at oqmd.org

Material	OQMD ID	Spacegroup	Bandgap (eV)		
			PBE	Predicted	Expmt/HSE
Ti_2O_3	678225	R-3 (148)	0.0	0.8 ± 0.3	0.03-0.14 (EXP)[138]
Ti_3O_5	66123	C2/m (12)	0.0	0.8 ± 0.3	0.14 (EXP)[139]
Co_3S_4	4563	Fd-3m (227)	0.0	1.6 ± 0.2	1.45 (EXP)[140]
Mn_2O_3	33709	Ia-3 (206)	0.0	0.7 ± 0.3	1.4 (EXP)[141]
TiO	10207	C2/m (12)	0.0	0.4 ± 0.3	1.9 (EXP)[142]
TiF_3	5608	R-3m (166)	0.0	1.1 ± 0.3	2.87 (HSE)[143]
Ni_3S_4	6716	Fd-3m (227)	0.0	1.5 ± 0.2	2.8 (EXP)[144]
MnF_3	3777	C2/c (15)	0.0	0.9 ± 0.3	3.03 (HSE)[143]
NiF_3	15556	R-3 (148)	0.0	1.4 ± 0.3	3.28 (HSE)[143]
VF_3	5882	R-3 (148)	0.0	1.0 ± 0.3	3.40 (HSE)[143]
MnSe	30752	P63mc (186)	0.0	1.5 ± 0.2	3.5-3.8 (EXP)[145]
$MgTiO_3$	692959	R-3 (148)	0.0	0.7 ± 0.3	3.7 (EXP)[146]
MnS	646143	P63mc (186)	0.0	1.4 ± 0.2	3.7, 3 (EXP)[147], [148]

V_2O_3	678210	R-3 (148)	0.3	2.2 ± 0.5	1.51 (EXP)[142]
Mn_3O_4	5975	I41/amd (141)	0.8	3.5 ± 0.4	2.91 (EXP)[149]
ZnO	4908	P63mc (186)	1.0	3.2 ± 0.3	3.29, 3.44 (EXP)[149], [150]
Fe_2O_3	92501	Ia-3 (206)	1.1	4.0 ± 0.4	1.97 (EXP)[151]
SnO_2	2477	P42/mnm (136)	1.2	2.6 ± 0.3	3.32 (EXP)[149]
MnO_2	677684	I4/m (87)	1.2	3.9 ± 0.3	2.5, 2.7(HSE)[152], [153]
$LaVO_3$	682189	Pnma (62)	1.2	3.4 ± 0.4	1.44 (EXP)[142]
CdS	5970	P63mc (186)	1.3	2.7 ± 0.2	2.58, 2.48 (EXP)[150]
CrF_3	4854	R-3c (167)	1.4	4.0 ± 0.3	4.91 (HSE)[143]
GaP	7553	F-43m (216)	1.8	2.4 ± 0.1	2.26, 2.33 (EXP)[154], [155]
TiO_2	2575	I41/amd (141)	2.0	4.5 ± 0.3	3.2 (EXP)[156]
$SrTiO_3$	827052	R-3c (167)	2.0	4.0 ± 0.3	3.2 (EXP)[146]
$LaCrO_3$	682305	Pnma (62)	2.1	3.6 ± 0.4	3.39 (EXP)[157]
ZnS	7652	F-43m (216)	2.3	3.4 ± 0.1	3.84 (EXP)[150]
BeSe	647324	F-43m (216)	2.8	3.9 ± 0.2	5.15 (EXP)[150]

5.3 Discussion

We created and trained an ML model based on the co-kriging algorithm to do multi-fideity learning on the electronic bandgap data of 1117 materials generated from DFT-PBE and DFT-HSE calculations. This model is capable of predicting the bandgaps at the accuracy level of expensive DFT-HSE calculations (high fidelity) at the expense of doing a much cheaper bandgap estimation from DFT-PBE calculations (low fidelity). The model is benchmarked first on a test data that is split from the full training-data in data-optimistic and data-pessimistic situations. In a data-pessimistic situation during the benchmarking where the train:test split is 2:8, the co-kriging is found to be a better algorithm compared to random forest regression and support vector regression in terms of having the lowest prediction errors with the least bias toward the imbalanced data. The data imbalance comes from having 62% of the training-data being metallic with E_g=0 while the others have $0<E_g<10$. co-kriging predicts the $E_g^{PredHSE}$ and estimates an uncertainty associated with the prediction. If the material being predicted belongs to a region in the feature space that is not well represented in the training-data, the model does not learn sufficient knowledge about that part of the feature space and, thus, estimates a large uncertainty during the prediction. The usefulness of this inherent model uncertainty quantification is also assessed by analyzing how often an incorrect prediction in test data had a high uncertainty already estimated by the co-kriging model. The model is shown to estimate a high uncertainty in more than half the cases where the $E_g^{PredHSE}$ happened to be different from the real value (E_g^{HSE}).

We analyzed the co-kriging model predictions on the search-space containing 24,967 materials whose E_g^{HSE} is unknown while the E_g^{PBE} is already known from the OQMD database. The number of materials in different classes of materials, such as oxides, halides, etc., where the DFT-PBE calculated a zero bandgap while the co-kriging predicts a non-zero bandgap, is reported. The distribution of uncertainty in the search-space compositions with respect to the constituent elements is also analyzed. Finally, the trend of how the $E_g^{PredHSE}$ value from the co-kriging model changes with a change in E_g^{PBE} is also plotted and analyzed.

To validate the model predictions in the search-space, we conducted an extensive literature survey on the 50 materials whose $E_g^{PredHSE}$ differs from E_g^{PBE} especially when the E_g^{PBE}=0 eV. The literature survey found data on 28 materials and successfully validated the reliability of all 13 bandgap opening ($E_g^{PredHSE}$ >0 eV, E_g^{HSE}=0 eV) predictions that we explored. Among the other 15 materials, the $E_g^{PredHSE}$ value is seen to be closer to the literature-reported HSE or experimental bandgaps in 13 materials. The training-data with E$_g^{HSE}$ and E$_g^{PBE}$ values and the search-space with E$_g^{PredHSE}$ and E$_g^{PBE}$ values are available from the web portal at hse.oqmd.org. We created this web portal to deploy the results from this study as a part of correcting the bandgap data in OQMD.

5.4 Methods

5.4.1 co-kriging

The co-kriging method is an extension of the traditional Kriging method, with a base mathematical framework proposed by Kennedy and O'Hagan in the early 2000s[133], [158]. Forrester et al[134] successfully applied it as a demonstration of a multi-fidelity co-kriging modeling problem with two levels of fidelities. A work by Le Gratiet and Garnier[159], [160] reported the method to decouple different levels of fidelity estimations in co-kriging effectively and thus, making the co-kriging implementations scalable to real-world multi-fidelity datasets. This method has been benchmarked in materials datasets recently[132], [135]. In this work, we apply co-kriging to our two-fidelity problem (HSE vs PBE) utilizing the implementation of this algorithm in the OpenMDAO package[29]. Following the conventions and equations provided in Forrester et al[134] and Pilania et al[132], a short description of the formulation of co-kriging in this work is as follows:

The set of material representation feature vectors for n materials is denoted by X with $X = x_1, x_2, ..., x_i, ..., x_n$ with each vector x_i of length m representing the feature vector of material i. The target properties are E_g^{PBE} and E_g^{HSE} denoted by y_c and y_e, respectively, with the subscript c referring to the cheaper low-fidelity value and the subscript e referring to the expensive, high-fidelity value. Similar to Kriging and other Gaussian process regressions, any estimated value at in X is assumed to be a Gaussian distribution Z. The Gaussian random variables Z_c and Z_e represent the low-fidelity estimation (E_g^{PBE}) and the high-fidelity estimation (E_g^{HSE}). Z_e and Z_c

are connected by the scaling factor ρ and another independent Gaussian variable as:

$$Z_e(x_i) = \rho Z_c(x_i) + Z_d(x_i) \tag{5.2}$$

The formulation of Equation 5.2 assumes that the high-fidelity prediction for a given material is dependent on the low-fidelity value of the same material only, and independent of the low-fidelity value of any other material. This assumption reduces the size of the covariance matrix to consider in this work by setting

$$Cov(Z_e(x_i), Z_c(x_j))|Z_c(x_i) = 0; i \neq j \tag{5.3}$$

The covariance matrices in this work use a squared exponential kernel to compute the correlation between Z_c and Z_d. The kernel, $k(x_i, x_j)$ is of the form:

$$k(x_i, x_j) = \sigma^2 \exp\left(-\sum_{p=1}^{m} \theta_p ||x_i^p - x_j^p||^2\right) \tag{5.4}$$

The full covariance matrix for this two-fidelity scenario with X being the same for both low- and high-fidelities is:

$$K = \begin{pmatrix} Cov[Z_c(X), Z_c(X)] & Cov[Z_e(X), Z_c(X)] \\ Cov[Z_e(X), Z_c(X)] & Cov[Z_e(X), Z_e(X)] \end{pmatrix} \tag{5.5}$$

The full covariance matrix in the above equation comprises two different correlation kernel functions k_c and k_d within independent sets of σ and hyperparameters belonging to Z_c and Z_d respectively. Hence, this entire model has a large number of internal parameters to optimize via maximum likelihood estimation (MLE). Eventually, the final prediction of the high-fidelity value, in terms of mean μ_e^* and variance σ_e^*, are obtained for a material in the search-space by,

$$\mu_e^* = \hat{\mu} + \mathbf{k}^T K^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}) \tag{5.6}$$

and

$$\sigma_e^{2*} = \hat{\rho}\hat{\sigma}_c^2 + \hat{\sigma}_d^2 - \mathbf{k}^T K^{-1}\mathbf{k} \tag{5.7}$$

with,

$$\mathbf{y} = \begin{pmatrix} y_c \\ y_e \end{pmatrix},$$

$$\hat{\mu} = \mathbf{1}^T K^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}),$$

$$\mathbf{k} = \begin{pmatrix} \hat{\rho}\hat{k}_c(X, x^*) \\ \hat{\rho}^2\hat{k}_c(X, x^*) + \hat{k}_d(X, x^*) \end{pmatrix},$$

while the $\wedge$ symbol over a parameter denotes the MLE values of that parameter or those of its constituent parameters and hyperparameters (in case of $\hat{k}$). **1** is a vector consisting of 1 as its every component.

5.4.2 HSE Bandgap Dataset

The training-dataset consisting of DFT-HSE and DFT-PBE bandgap values used in this work is generated by Mohan et al, via high throughput (HT) computations. The DFT calculations in the HT workflow used Projector Augmented Wave (PAW) method as implemented in the Vienna Ab-initio Simulation Package (VASP). The k-point mesh densities were set at 8000 for DFT-PBE and 2000 for DFT-HSE. The bandgap is calculated from the electronic density of states data computed via static DFT calculations on crystal structures relaxed using PBE functionals. The full training-data that we used in this work is available from the web portal hse.oqmd.org hosted as a sub-dataset of the OQMD database.

CHAPTER 6
SUMMARY AND OUTLOOK

6.1 Summary

In this thesis, I have described three main projects that fit into the field of computational materials discovery from statistical selection and atomistic simulations. All these projects use or contribute to the general materials design workflow described in Figure 6.1.

The work on dielectrics detailed in Chapter 3 makes use of the complete design workflow for three design cycles sequentially. Most of the time and resources in this work were spent while conducting the first design cycle. Many of the resources, such as ML model pipelines, training data, search space, DFPT simulation files, etc, were reused with minimal modifications from the second design cycle onward. Because this project was built with a robust design infrastructure, its components are well connected while also being portable to other projects when required. We discovered three very high-dielectric materials and four other moderately high-dielectric materials within just 17 material selections. This high success rate may be attributed to the extra layer of statistical optimization algorithm added after the ML model, and also to the filters on search space with bandgap minimum cutoffs.

The work on 2D inks in Chapter 4 describes a different version of the workflow in Figure 6.1 where the knowledge-feedback and multiple design cycles could not be done due to a significantly higher computational cost associated with the validation of material selection. Here, the workflow was adapted to the project based on the specific constraints of the problem, but the main parts of the workflow remained the same - Data, Modeling(Selection), and Validation. We used machine

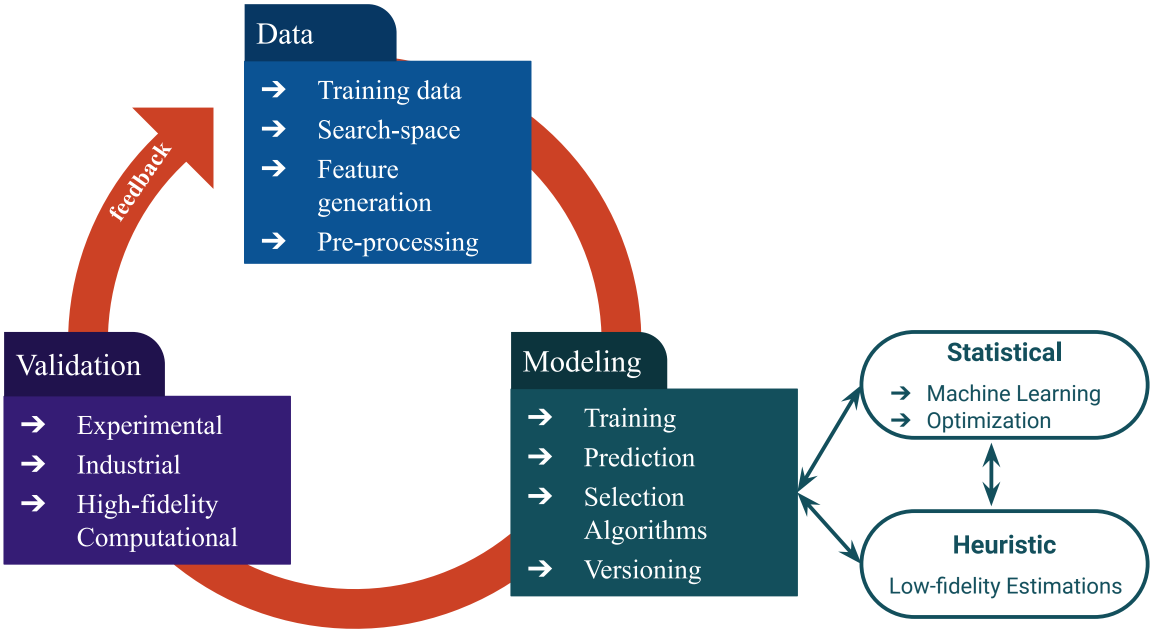


Figure 6.1: General designflow

learning in this work to aid a heuristic model to calculate the band alignment values more accurately. The limited availability of resources and time for validation required the selection process to be rigorous in confirming that the final candidate is, in fact, a good suggestion based on literature surveys as well. Eventually, the hexagonal GaS is identified and investigated for being a bridge-materials for MoS_2-based 2D inks.

In the third project in Chapter 5, we describe the works on solidifying parts of the materials design infrastructure. The materials selection strongly depends upon the quality of the data and the reliability of the modeling methods. This work is intended to improve both parts. A multi-fidelity predictive model is built to predict the material bandgap values at higher accuracy than what is estimated in most of the large HT databases from lower-accuracy first-principles calculations. The predicted bandgaps for OQMD materials are deployed via a new web portal associated with the main OQMD database. The availability of bandgap values with higher accuracy than DFT-PBE helps in identifying the real semiconductors which are misclassified as metals in DFT-PBE calculations. In addition to improving the search space for materials design, this work also shows the importance of uncertainty quantification in machine learning models. The quantified uncertainty values are shown to aid in avoiding incorrect predictions by pointing out which predictions are made by the model with less confidence. The lack of uncertainty quantification in ML modeling can lead to the unnecessary depletion of resources by focusing on false-positive selections.

Overall, we have achieved successful materials design in three different situations where the material selection from large pools of candidate materials was necessary. In one of the projects, data from two separate material repositories were combined within a statistical design - demonstrating the usefulness of interoperability between high throughput databases. We also utilized statistical model uncertainties in two different ways. In the case of dielectric selection, the quantified model uncertainty guided the search for novel materials to previously unexplored classes of materials. A prediction with high uncertainty was considered to be desired value in this case. In contrast, we showed in the multi-fidelity modeling work that the quantified uncertainty can be used to filter out unreliable predictions. This is desired when reliable bandgap predictions are preferred over looking into unexplored material classes.

We also demonstrated two different kinds of uncertainty quantifications. In the case of dielectric selection, the uncertainty was estimated by building a large number of statistical models over different subsets of the training data and using this set of models to predict a distribution of target values for each candidate material. The predicted distribution's standard deviation was considered to be the quantified uncertainty. This method is popularly known as bagging. But in the case of multi-fidelity modeling, we quantified the uncertainty from a different method. The multi-fidelity modeling is done by fitting a co-Kriging Gaussian model to the training data. The uncertainties are quantified internally inside the co-Kriging models by constructing covariance matrices to explicitly compute how close a candidate material is to the training data materials in the feature vector space.

The ML model predictions are used directly to compute material properties in the dielectric selection, and the bandgap multi-fidelity modeling works. But in the 2D inks design project, the ML model aids in improving a heuristic model instead of predicting the final property. These situations show two of the many different possible approaches to utilizing the capabilities of ML models within the material design.

6.2 Other Works

In addition to the three major projects mentioned in this thesis, there are several other relatively smaller works completed over the years that fit into one or more sections of the materials design and discovery. Since mid-2019, I have worked with the OQMD database as the primary developer of the OQMD API and the maintainer of the public server at oqmd.org. During the course of three years, three separate projects were completed that improved the public materials data infrastructure. Each of them makes the data more FAIR[2] - Findable, Accessible, Interoperable, and/or Reusable.

The first one is the development of the OPTIMADE REST API specification for universal materials data transfer in collaboration with other major material database providers and maintainers. I co-first authored the OPTIMADE specification[161] and maintained the corresponding API at OQMD servers. The primary achievement of OPTIMADE is the implementation of a fast data access system across multiple databases with a unified query syntax. The OPTIMADE-related works are intended to make the data more findable, accessible, and interoperable.

In the second work, I implemented the persistent identifiers for OQMD data entries, which was done in collaboration with the National Institute of Standards and Technology to enhance the availability of materials data in OQMD in the long term. The presence of persistent identifiers makes data less vulnerable to link-rot issues that happen when the databases are moved to a different domain URL, and any citations of data involving the previous URL lose the relevant information. This work is intended to increase the reusability and findability of the data in the long term.

In the third work, I transitioned the OQMD.org server from on-premises hosting to a fully-fledged cloud infrastructure, increasing the accessibility of the database for external researchers. The improved server infrastructure (as shown in the appendix) has resulted in faster data access and lesser downtime.

6.3 Outlook

This design process diagram in Figure 6.1 is a general suggestion for approaching most of the materials discovery challenges today. Once the real-world application requirements are translated into specific material property constraints, the workflow serves as a general guideline to follow. Individual sections of the workflow can be developed independently as long as they can be plugged into a workflow implemented for a specific materials discovery challenge in the future. There is plenty of room to improve materials data infrastructure. We still need better interoperability between databases. A system to connect the experimentally generated data to the computational databases to validate the HT simulation frameworks is still unavailable. The materials data hosting needs to be made more mainstream and easy so that more data will be available across the public domains. To achieve all the above, there need to be more initiatives like CHiMaD that connect computational researchers with experimentalists. But for now, it looks like the community is on the right track toward the goal set in 2011 by the Materials Genome Initiative. The usage of machine learning has exploded in materials science over the past half-decade. I believe that it will continue to grow as long as the data infrastructure surrounding it can support that growth.

Another area of interest is in creating better statistical prediction models and statistical selection algorithms. The size of available data in materials science can vary drastically from very small (< 10) to very large ($> 10^6$). And often, trusting the prediction from a statistical model leads to conducting expensive computations and experiments. Hence, it is crucial to avoid false positives during material selection in most projects. Material feature vector generation methods that can reliably represent materials within a small set of vector components to support small-data ML are also in demand. Overall, the statistical modeling research is yet to be fully customized to fit the materials data modeling, and that leads to plenty of opportunities for future work.

REFERENCES

[1] N. Science and T. C. (US), *Materials genome initiative for global competitiveness*. Executive Office of the President, National Science and Technology Council, 2011.

[2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, “The fair guiding principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.

[3] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, “Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd),” *Jom*, vol. 65, no. 11, pp. 1501–1509, 2013.

[4] S. Kirklin, J. E. Saal, B. Meredig, *et al.*, “The open quantum materials database (oqmd): Assessing the accuracy of dft formation energies,” *npj Computational Materials*, vol. 1, no. 1, pp. 1–15, 2015.

[5] A. Jain, S. P. Ong, G. Hautier, *et al.*, “Commentary: The materials project: A materials genome approach to accelerating materials innovation,” *APL materials*, vol. 1, no. 1, p. 011 002, 2013.

[6] S. Curtarolo, W. Setyawan, G. L. Hart, *et al.*, “Aflow: An automatic framework for high-throughput materials discovery,” *Computational Materials Science*, vol. 58, pp. 218–226, 2012.

[7] L. Talirz, S. Kumbhar, E. Passaro, *et al.*, “Materials cloud, a platform for open computational science,” *Scientific data*, vol. 7, no. 1, pp. 1–12, 2020.

[8] C. Draxl and M. Scheffler, “The nomad laboratory: From data sharing to artificial intelligence,” *Journal of Physics: Materials*, vol. 2, no. 3, p. 036 001, 2019.

[9] K. Choudhary, K. F. Garrity, A. C. Reid, *et al.*, “The joint automated repository for various integrated simulations (jarvis) for data-driven materials design,” *npj Computational Materials*, vol. 6, no. 1, pp. 1–13, 2020.

[10] P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, and T. Lookman, “Adaptive strategies for materials design using uncertainties,” *Scientific reports*, vol. 6, no. 1, pp. 1–9, 2016.

[11] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," *Physical review letters*, vol. 120, no. 14, p. 145 301, 2018.

[12] J. H. Montoya, K. T. Winther, R. A. Flores, T. Bligaard, J. S. Hummelshøj, and M. Aykol, "Autonomous intelligent agents for accelerated materials discovery," *Chemical Science*, vol. 11, no. 32, pp. 8517–8532, 2020.

[13] B. Meredig, *Five high-impact research areas in machine learning for materials science*, 2019.

[14] T. Mueller, A. G. Kusne, and R. Ramprasad, "Machine learning in materials science: Recent progress and emerging applications," *Reviews in Computational Chemistry*, vol. 29, pp. 186–273, 2016.

[15] A. M. Gopakumar, P. V. Balachandran, D. Xue, J. E. Gubernatis, and T. Lookman, "Multi-objective optimization for materials discovery via adaptive design," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[16] K. T. Schutt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Muller, "Schnet–a deep learning architecture for molecules and materials," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241 722, 2018.

[17] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Physical review*, vol. 140, no. 4A, A1133, 1965.

[18] P. Hohenberg and W. Kohn, "Phys rev 136: B864," *Kohn W, Sham LJ (1965) Phys Rev*, vol. 140, A1133, 1964.

[19] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Physical review letters*, vol. 77, no. 18, p. 3865, 1996.

[20] J. Heyd, G. E. Scuseria, and M. Ernzerhof, "Hybrid functionals based on a screened coulomb potential," *The Journal of chemical physics*, vol. 118, no. 18, pp. 8207–8215, 2003.

[21] G. Kresse and J. Hafner, "Ab initio molecular dynamics for liquid metals," *Physical Review B*, vol. 47, no. 1, p. 558, 1993.

[22] G. Kresse and J. Furthmüller, “Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set,” *Computational materials science*, vol. 6, no. 1, pp. 15–50, 1996.

[23] G. Kresse and J. Furthmüller, “Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set,” *Physical review B*, vol. 54, no. 16, p. 11 169, 1996.

[24] G. Kresse and D. Joubert, “From ultrasoft pseudopotentials to the projector augmented-wave method,” *Physical review b*, vol. 59, no. 3, p. 1758, 1999.

[25] P. E. Blöchl, “Projector augmented-wave method,” *Physical review B*, vol. 50, no. 24, p. 17 953, 1994.

[26] M. Abadi, A. Agarwal, P. Barham, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015.

[27] F. Chollet *et al.*, *Keras*, `https://keras.io`, 2015.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res*, vol. 12, pp. 2825–2830, 2011.

[29] J. S. Gray, J. T. Hwang, J. R. R. A. Martins, K. T. Moore, and B. A. Naylor, “OpenMDAO: An open-source framework for multidisciplinary design, analysis, and optimization,” *Structural and Multidisciplinary Optimization*, vol. 59, no. 4, pp. 1075–1104, Apr. 2019.

[30] R. P. Ortiz, A. Facchetti, and T. J. Marks, “High-k organic, inorganic, and hybrid dielectrics for low-voltage organic field-effect transistors,” *Chem. Rev.*, vol. 110, no. 1, pp. 205–239, 2009.

[31] Y. Iino, Y. Inoue, Y. Fujisaki, *et al.*, “Organic thin-film transistors on a plastic substrate with anodically oxidized high-dielectric-constant insulators,” *Jpn. J. Appl. Phys.*, vol. 42, no. 1R, p. 299, 2003.

[32] B. Wang, W. Huang, L. Chi, M. Al-Hashimi, T. J. Marks, and A. Facchetti, “High-k gate dielectrics for emerging flexible and stretchable electronics,” *Chemical reviews*, vol. 118, no. 11, pp. 5690–5754, 2018.

[33] K. Kukli, J. Aarik, A. Aidla, O. Kohan, T. Uustare, and V. Sammelselg, “Properties of tantalum oxide thin films grown by atomic layer deposition,” *Thin Solid Films*, vol. 260, no. 2, pp. 135–142, 1995.

[34] J. Ramajothi, S. Ochiai, K. Kojima, and T. Mizutani, "Performance of organic field-effect transistor based on poly (3-hexylthiophene) as a semiconductor and titanium dioxide gate dielectrics by the solution process," *Jpn. J. Appl. Phys.*, vol. 47, no. 11R, p. 8279, 2008.

[35] M. Lee, Y. Youn, K. Yim, and S. Han, "High-throughput ab initio calculations on dielectric constant and band gap of non-oxide dielectrics," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[36] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[37] I. Petousis, D. Mrdjenovich, E. Ballouz, *et al.*, "High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials," *Scientific data*, vol. 4, no. 1, pp. 1–12, 2017.

[38] G. Petretto, S. Dwaraknath, H. P. Miranda, *et al.*, "High-throughput density-functional perturbation theory phonons for inorganic materials," *Scientific data*, vol. 5, no. 1, pp. 1–12, 2018.

[39] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," *npj Computational Materials*, vol. 2, no. 1, pp. 1–7, 2016.

[40] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American statistical association*, vol. 83, no. 403, pp. 596–610, 1988.

[41] Plotly Technologies Inc., *Collaborative data science*, `https://plot.ly`, Accessed: 2022-04-06, 2015.

[42] H. Kageyama, K. Hayashi, K. Maeda, *et al.*, "Expanding frontiers in materials chemistry and physics with multiple anions," *Nat. Commun.*, vol. 9, no. 1, pp. 1–15, 2018.

[43] K. Choudhary, K. F. Garrity, V. Sharma, A. J. Biacchi, A. R. H. Walker, and F. Tavazza, "High-throughput density functional perturbation theory and machine learning predictions of infrared, piezoelectric, and dielectric responses," *npj Comput. Mater.*, vol. 6, no. 1, pp. 1–13, 2020.

[44] I. Pallikara, P. Kayastha, J. M. Skelton, and L. D. Whalley, "The physical significance of imaginary phonon modes in crystals," *Electronic Structure*, 2022.

[45] D. Templeton and C. H. Dauben, “Crystal structures of rare earth oxychlorides,” *J. Am. Chem. Soc.*, vol. 75, no. 23, pp. 6069–6070, 1953.

[46] J. Hölsä, M. Lahtinen, M. Lastusaari, J. Valkonen, and J. Viljanen, “Stability of rare-earth oxychloride phases: Bond valence study,” *J. Solid State Chem.*, vol. 165, no. 1, pp. 48–55, 2002.

[47] T. Basiev, N. Batyrev, V. Voronov, *et al.*, “Hydration of strontium chloride and rare-earth element oxychlorides,” *Russ. J. Appl. Chem.*, vol. 78, no. 7, pp. 1035–1037, 2005.

[48] H. Jacobsen, G. Meyer, W. Schipper, and G. Blasse, “Synthesis, structures and luminescence of two new Europium (ii) Silicate-Chlorides, $Eu_2SiO_3Cl_2$ and $Eu_5SiO_4Cl_6$,” *Z. Anorg. Allg. Chem.*, vol. 620, no. 3, pp. 451–456, 1994.

[49] J. Robertson, “High dielectric constant gate oxides for metal oxide si transistors,” *Reports on progress in Physics*, vol. 69, no. 2, p. 327, 2005.

[50] E. Atanassova and D. Spassov, “X-ray photoelectron spectroscopy of thermal thin ta2o5 films on si,” *Appl. Surf. Sci.*, vol. 135, no. 1-4, pp. 71–82, 1998.

[51] D. G. Schlom and J. H. Haeni, “A thermodynamic approach to selecting alternative gate dielectrics,” *MRS Bull.*, vol. 27, no. 3, pp. 198–204, 2002.

[52] G. Alers, D. Werder, Y. Chabal, *et al.*, “Intermixing at the tantalum oxide/silicon interface in gate dielectric structures,” *Appl. Phys. Lett.*, vol. 73, no. 11, pp. 1517–1519, 1998.

[53] M. Perego, G. Seguini, G. Scarel, M. Fanciulli, and F. Wallrapp, “Energy band alignment at ti o 2/ si interface with various interlayers,” *J. Appl. Phys.*, vol. 103, no. 4, p. 043 509, 2008.

[54] P. R. McCurdy, L. J. Sturgess, S. Kohli, and E. R. Fisher, “Investigation of the pecvd tio2–si (1 0 0) interface,” *Appl. Surf. Sci.*, vol. 233, no. 1-4, pp. 69–79, 2004.

[55] J. P. George, J. Beeckman, W. Woestenborghs, P. F. Smet, W. Bogaerts, and K. Neyts, “Preferentially oriented batio 3 thin films deposited on silicon with thin intermediate buffer layers,” *Nanoscale Res. Lett.*, vol. 8, no. 1, pp. 1–7, 2013.

[56] X. Hu, H. Li, Y. Liang, *et al.*, “The interface of epitaxial srtio 3 on silicon: In situ and ex situ studies,” *Appl. Phys. Lett.*, vol. 82, no. 2, pp. 203–205, 2003.

[57] L. Goncharova, D. Starodub, E. Garfunkel, *et al.*, “Interface structure and thermal stability of epitaxial sr ti o 3 thin films on si (001),” *J. Appl. Phys.*, vol. 100, no. 1, p. 014 912, 2006.

[58] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, “The quickhull algorithm for convex hulls,” *ACM Trans. Math. Software*, vol. 22, no. 4, pp. 469–483, 1996.

[59] W. Sun, S. T. Dacek, S. P. Ong, *et al.*, “The thermodynamic scale of inorganic crystalline metastability,” *Sci. Adv.*, vol. 2, no. 11, e1600225, 2016.

[60] P. V. Balachandran, A. A. Emery, J. E. Gubernatis, T. Lookman, C. Wolverton, and A. Zunger, “Predictions of new ABO_3 perovskite compounds by combining machine learning and density functional theory,” *Phys. Rev. Mater.*, vol. 2, no. 4, p. 043 802, 2018.

[61] Y. Wu, P. Lazic, G. Hautier, K. Persson, and G. Ceder, “First principles high throughput screening of oxynitrides for water-splitting photocatalysts,” *Energy Environ. Sci.*, vol. 6, no. 1, pp. 157–168, 2013.

[62] A. Zakutayev, X. Zhang, A. Nagaraja, *et al.*, “Theoretical prediction and experimental realization of new stable inorganic materials using the inverse design approach,” *J. Am. Chem. Soc.*, vol. 135, no. 27, pp. 10 048–10 054, 2013.

[63] K. Pal, Y. Xia, J. Shen, *et al.*, “Accelerated discovery of a large family of quaternary chalcogenides with very low lattice thermal conductivity,” *npj Comput. Mater.*, vol. 7, no. 1, pp. 1–13, 2021.

[64] J.-P. Zhou, C.-L. Chai, S.-Y. Yang, *et al.*, “Properties of high k gate dielectric gadolinium oxide deposited on si (1 0 0) by dual ion beam deposition (dibd),” *J. Cryst. Growth*, vol. 270, no. 1-2, pp. 21–29, 2004.

[65] J. Kwo, M. Hong, A. Kortan, *et al.*, “Properties of high κ gate dielectrics gd 2 o 3 and y 2 o 3 for si,” *J. Appl. Phys.*, vol. 89, no. 7, pp. 3920–3927, 2001.

[66] H.-L. Keller, “Darstellung und kristallstruktur von hoch-Tl_3PbBr_5,” *J. Less-Common Met.*, vol. 78, no. 2, pp. 281–286, 1981.

[67] N. Denysyuk, V. Bekenev, M. Karpets, O. Parasyuk, S. Danylchuk, and O. Khyzhun, “Electronic structure of the high-temperature tetragonal Tl_3PbBr_5 phase,” *J. Alloys Compd.*, vol. 576, pp. 271–278, 2013.

[68] A. Ferrier, M. Velázquez, X. Portier, J.-L. Doualan, and R. Moncorgé, “Tl_3PbBr_5: A possible crystal candidate for middle infrared nonlinear optics,” *J. Cryst. Growth*, vol. 289, no. 1, pp. 357–365, 2006.

[69] Y. Qiu and S. Suh, “Economic feasibility of recycling rare earth oxides from end-of-life lighting technologies,” *Resour. Conserv. Recycl.*, vol. 150, p. 104 432, 2019.

[70] A. Amato, A. Becci, I. Birloaga, *et al.*, “Sustainability analysis of innovative technologies for the rare earth elements recovery,” *Renewable Sustainable Energy Rev.*, vol. 106, pp. 41–53, 2019.

[71] A. Yaroshevsky, “Abundances of chemical elements in the earth’s crust,” *Geochem. Int.*, vol. 44, no. 1, pp. 48–55, 2006.

[72] B. F. Thornton and S. C. Burdette, “Homely holmium,” *Nat. Chem.*, vol. 7, no. 6, pp. 532–532, 2015.

[73] M. Gajdoš, K. Hummer, G. Kresse, J. Furthmüller, and F. Bechstedt, “Linear optical properties in the projector-augmented wave methodology,” *Physical Review B*, vol. 73, no. 4, p. 045 112, 2006.

[74] P. Giannozzi, S. De Gironcoli, P. Pavone, and S. Baroni, “Ab initio calculation of phonon dispersions in semiconductors,” *Physical Review B*, vol. 43, no. 9, p. 7231, 1991.

[75] S. Baroni, S. De Gironcoli, A. Dal Corso, and P. Giannozzi, “Phonons and related crystal properties from density-functional perturbation theory,” *Reviews of Modern Physics*, vol. 73, no. 2, p. 515, 2001.

[76] R. Sternheimer, “Electronic polarizabilities of ions from the hartree-fock wave functions,” *Physical Review*, vol. 96, no. 4, p. 951, 1954.

[77] E. Cockayne and B. P. Burton, “Phonons and static dielectric constant in catio 3 from first principles,” *Physical Review B*, vol. 62, no. 6, p. 3735, 2000.

[78] I. Petousis, W. Chen, G. Hautier, *et al.*, “Benchmarking density functional perturbation theory to enable high-throughput screening of materials for dielectric constant and refractive index,” *Physical Review B*, vol. 93, no. 11, p. 115 151, 2016.

[79] X. Gonze and C. Lee, “Dynamical matrices, born effective charges, dielectric permittivity tensors, and interatomic force constants from density-functional perturbation theory,” *Physical Review B*, vol. 55, no. 16, p. 10 355, 1997.

[80] D. Vanderbilt, X. Zhao, and D. Ceresoli, “Structural and dielectric properties of crystalline and amorphous zro2,” *Thin Solid Films*, vol. 486, no. 1-2, pp. 125–128, 2005.

[81] G. He, L. Zhu, M. Liu, Q. Fang, and L. Zhang, “Optical and electrical properties of plasma-oxidation derived hfo2 gate dielectric films,” *Applied surface science*, vol. 253, no. 7, pp. 3413–3418, 2007.

[82] C.-K. Kwok and C. R. Aita, “Indirect band gap in α-zro2,” *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, vol. 8, no. 4, pp. 3345–3346, 1990.

[83] M. Mikami, S. Nakamura, O. Kitao, and H. Arakawa, “Lattice dynamics and dielectric properties of tio 2 anatase: A first-principles study,” *Physical Review B*, vol. 66, no. 15, p. 155 213, 2002.

[84] C. Dette, M. A. Pérez-Osorio, C. S. Kley, *et al.*, “Tio2 anatase with a bandgap in the visible region,” *Nano letters*, vol. 14, no. 11, pp. 6533–6538, 2014.

[85] C. Lee, P. Ghosez, and X. Gonze, “Lattice dynamics and dielectric properties of incipient ferroelectric tio 2 rutile,” *Physical Review B*, vol. 50, no. 18, p. 13 379, 1994.

[86] T. Luttrell, S. Halpegamage, J. Tao, A. Kramer, E. Sutter, and M. Batzill, “Why is anatase a better photocatalyst than rutile?-model studies on epitaxial tio2 films,” *Scientific reports*, vol. 4, no. 1, pp. 1–8, 2014.

[87] R. D. Shannon, “Dielectric polarizabilities of ions in oxides and fluorides,” *Journal of Applied physics*, vol. 73, no. 1, pp. 348–366, 1993.

[88] J. A. C. Santana, J. M. An, N. Wu, *et al.*, “Effect of gadolinium doping on the electronic band structure of europium oxide,” *Physical Review B*, vol. 85, no. 1, p. 014 406, 2012.

[89] B. Lu, H. Cheng, X. Xu, and H. Chen, “Preparation and characterization of transparent magneto-optical ho2o3 ceramics,” *Journal of the American Ceramic Society*, vol. 102, no. 1, pp. 118–122, 2019.

[90] F. Torrisi and J. N. Coleman, “Electrifying inks with 2d materials,” *Nature nanotechnology*, vol. 9, no. 10, pp. 738–739, 2014.

[91] F. Bonaccorso, A. Bartolotta, J. N. Coleman, and C. Backes, “2d-crystal-based functional inks,” *Advanced Materials*, vol. 28, no. 29, pp. 6136–6166, 2016.

[92] M. M. Alsaif, A. F. Chrimes, T. Daeneke, *et al.*, “High-performance field effect transistors using electronic inks of 2d molybdenum oxide nanoflakes,” *Advanced Functional Materials*, vol. 26, no. 1, pp. 91–100, 2016.

[93] O. A. Moses, L. Gao, H. Zhao, *et al.*, “2d materials inks toward smart flexible electronics,” *Materials Today*, vol. 50, pp. 116–148, 2021.

[94] J.-W. T. Seo, J. Zhu, V. K. Sangwan, E. B. Secor, S. G. Wallace, and M. C. Hersam, “Fully inkjet-printed, mechanically flexible mos2 nanosheet photodetectors,” *ACS applied materials & interfaces*, vol. 11, no. 6, pp. 5675–5681, 2019.

[95] E. Carroll, D. Buckley, D. McNulty, and C. O’Dwyer, “Communication—conductive paintable 2d layered mos2 inks,” *ECS Journal of Solid State Science and Technology*, vol. 9, no. 9, p. 093 015, 2020.

[96] D. McManus, S. Vranic, F. Withers, *et al.*, “Water-based and biocompatible 2d crystal inks for all-inkjet-printed heterostructures,” *Nature nanotechnology*, vol. 12, no. 4, pp. 343–350, 2017.

[97] G. Hu, J. Kang, L. W. Ng, *et al.*, “Functional inks and printing of two-dimensional materials,” *Chemical Society Reviews*, vol. 47, no. 9, pp. 3265–3300, 2018.

[98] T. Carey, A. Arbab, L. Anzi, *et al.*, “Inkjet printed circuits with 2d semiconductor inks for high-performance electronics,” *Advanced Electronic Materials*, vol. 7, no. 7, p. 2 100 112, 2021.

[99] K. Thiyagarajan, W.-J. Song, H. Park, *et al.*, “Electroactive 1t-mos2 fluoroelastomer ink for intrinsically stretchable solid-state in-plane supercapacitors,” *ACS Applied Materials & Interfaces*, vol. 13, no. 23, pp. 26 870–26 878, 2021.

[100] L. Kuo, V. K. Sangwan, S. V. Rangnekar, *et al.*, “All-printed ultrahigh-responsivity mos2 nanosheet photodetectors enabled by megasonic exfoliation,” *Advanced Materials*, p. 2 203 772, 2022.

[101] R. Liu, F. Wang, L. Liu, *et al.*, “Band alignment engineering in two-dimensional transition metal dichalcogenide-based heterostructures for photodetectors,” *Small Structures*, vol. 2, no. 3, p. 2 000 136, 2021.

[102] K. Kim, S. Larentis, B. Fallahazad, *et al.*, "Band alignment in wse2–graphene heterostructures," *ACS nano*, vol. 9, no. 4, pp. 4527–4532, 2015.

[103] R. Ganatra and Q. Zhang, "Few-layer mos2: A promising layered semiconductor," *ACS nano*, vol. 8, no. 5, pp. 4074–4099, 2014.

[104] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, and A. Kis, "Single-layer mos2 transistors," *Nature nanotechnology*, vol. 6, no. 3, pp. 147–150, 2011.

[105] J. Tersoff, "Theory of semiconductor heterojunctions: The role of quantum dipoles," *Physical Review B*, vol. 30, no. 8, p. 4874, 1984.

[106] A. Schleife, F. Fuchs, C. Rödl, J. Furthmüller, and F. Bechstedt, "Branch-point energies and band discontinuities of iii-nitrides and iii-/ii-oxides from quasiparticle band-structure calculations," *Applied Physics Letters*, vol. 94, no. 1, p. 012 104, 2009.

[107] M. I. Zappia, G. Bianca, S. Bellani, *et al.*, "Two-dimensional gallium sulfide nanoflakes for uv-selective photoelectrochemical-type photodetectors," *The Journal of Physical Chemistry C*, vol. 125, no. 22, pp. 11 857–11 866, 2021.

[108] A. Harvey, C. Backes, Z. Gholamvand, *et al.*, "Preparation of gallium sulfide nanosheets by liquid exfoliation and their application as hydrogen evolution catalysts," *Chemistry of Materials*, vol. 27, no. 9, pp. 3483–3493, 2015.

[109] H. Chen, Y. Li, L. Huang, and J. Li, "Intrinsic defects in gallium sulfide monolayer: A first-principles study," *Rsc Advances*, vol. 5, no. 63, pp. 50 883–50 889, 2015.

[110] G. Micocci, R. Rella, P. Siciliano, and A. Tepore, "Investigation of electronic properties of gallium sulfide single crystals grown by iodine chemical transport," *Journal of applied physics*, vol. 68, no. 1, pp. 138–142, 1990.

[111] S. Yang, Y. Li, X. Wang, *et al.*, "High performance few-layer gas photodetector and its unique photo-response in different gas environments," *Nanoscale*, vol. 6, no. 5, pp. 2582–2587, 2014.

[112] T. Necio and M. Birowska, "Supercell-core software: A useful tool to generate an optimal supercell for vertically stacked nanomaterials," *AIP Advances*, vol. 10, no. 10, p. 105 105, 2020.

[113] K. F. Mak, C. Lee, J. Hone, J. Shan, and T. F. Heinz, “Atomically thin mos 2: A new direct-gap semiconductor,” *Physical review letters*, vol. 105, no. 13, p. 136 805, 2010.

[114] W. Huang, X. Luo, C. K. Gan, S. Y. Quek, and G. Liang, “Theoretical study of thermoelectric properties of few-layer mos 2 and wse 2,” *Physical Chemistry Chemical Physics*, vol. 16, no. 22, pp. 10 866–10 874, 2014.

[115] Y. Gutierrez, M. M. Giangregorio, S. Dicorato, F. Palumbo, and M. Losurdo, “Exploring the thickness-dependence of the properties of layered gallium sulfide,” *Frontiers in Chemistry*, vol. 9, 2021.

[116] B. J. Carey, J. Z. Ou, R. M. Clark, *et al.*, “Wafer-scale two-dimensional semiconductors from printed oxide skin of liquid metals,” *Nature communications*, vol. 8, no. 1, pp. 1–10, 2017.

[117] H. L. Zhuang and R. G. Hennig, “Single-layer group-iii monochalcogenide photocatalysts for water splitting,” *Chemistry of Materials*, vol. 25, no. 15, pp. 3232–3238, 2013.

[118] C. Hu, C. Yuan, A. Hong, M. Guo, T. Yu, and X. Luo, “Work function variation of monolayer mos2 by nitrogen-doping,” *Applied Physics Letters*, vol. 113, no. 4, p. 041 602, 2018.

[119] E. J. Santos and E. Kaxiras, “Electrically driven tuning of the dielectric constant in mos2 layers,” *ACS nano*, vol. 7, no. 12, pp. 10 741–10 746, 2013.

[120] A. M. Ganose, J. Park, A. Faghaninia, R. Woods-Robinson, K. A. Persson, and A. Jain, “Efficient calculation of carrier scattering rates from first principles,” *Nature communications*, vol. 12, no. 1, pp. 1–9, 2021.

[121] M. Li, J. Yao, X. Wu, *et al.*, “P-type doping in large-area monolayer mos2 by chemical vapor deposition,” *ACS applied materials & interfaces*, vol. 12, no. 5, pp. 6276–6282, 2020.

[122] V. I. Hegde, M. Aykol, S. Kirklin, and C. Wolverton, “The phase stability network of all inorganic materials,” *Science advances*, vol. 6, no. 9, eaay5606, 2020.

[123] S. I. Dutch, “Periodic tables of elemental abundance,” *Journal of chemical education*, vol. 76, no. 3, p. 356, 1999.

[124] W. Mönch, “Elementary calculation of the branch-point energy in the continuum of interface-induced gap states,” *Applied surface science*, vol. 117, pp. 380–387, 1997.

[125] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu," *The Journal of chemical physics*, vol. 132, no. 15, p. 154 104, 2010.

[126] S. Grimme, S. Ehrlich, and L. Goerigk, "Effect of the damping function in dispersion corrected density functional theory," *Journal of computational chemistry*, vol. 32, no. 7, pp. 1456–1465, 2011.

[127] C. Stefano, W. Setyawan, S. Wang, *et al.*, "Aflowlib. org: A distributed materials properties repository from high-throughput ab initio calculations," *Computational Materials Science*, vol. 58, pp. 227–235, 2012.

[128] A. Jain, G. Hautier, C. J. Moore, *et al.*, "A high-throughput infrastructure for density functional theory calculations," *Computational Materials Science*, vol. 50, no. 8, pp. 2295–2310, 2011.

[129] F. Tran and P. Blaha, "Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential," *Physical review letters*, vol. 102, no. 22, p. 226 401, 2009.

[130] M. Chan and G. Ceder, "Efficient band gap prediction for solids," *Physical review letters*, vol. 105, no. 19, p. 196 403, 2010.

[131] M. Liu, A. Gopakumar, V. Hegde, J. He, and C. Wolverton, "High-throughput hybrid-functional dft calculations of bandgaps & formation energies and multi-fidelity learning with uncertainty quantification," *Manuscript in Preparation*,

[132] G. Pilania, J. E. Gubernatis, and T. Lookman, "Multi-fidelity machine learning models for accurate bandgap predictions of solids," *Computational Materials Science*, vol. 129, pp. 156–163, 2017.

[133] M. C. Kennedy and A. O'Hagan, "Predicting the output from a complex computer code when fast approximations are available," *Biometrika*, vol. 87, no. 1, pp. 1–13, 2000.

[134] A. I. Forrester, A. Sóbester, and A. J. Keane, "Multi-fidelity optimization via surrogate modelling," *Proceedings of the royal society a: mathematical, physical and engineering sciences*, vol. 463, no. 2088, pp. 3251–3269, 2007.

[135] R. Batra, G. Pilania, B. P. Uberuaga, and R. Ramprasad, "Multifidelity information fusion with machine learning: A case study of dopant formation energies in hafnia," *ACS applied materials & interfaces*, vol. 11, no. 28, pp. 24 906–24 918, 2019.

[136] J. P. Perdew, A. Ruzsinszky, J. Tao, V. N. Staroverov, G. E. Scuseria, and G. I. Csonka, "Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits," *The Journal of chemical physics*, vol. 123, no. 6, p. 062 201, 2005.

[137] G. W. Moran, "Locally-weighted-regression scatter-plot smoothing (lowess): A graphical exploratory data analysis technique," NAVAL POSTGRADUATE SCHOOL MONTEREY CA, Tech. Rep., 1984.

[138] J. Honig and T. Reed, "Electrical properties of ti 2 o 3 single crystals," *Physical Review*, vol. 174, no. 3, p. 1020, 1968.

[139] S.-i. Ohkoshi, Y. Tsunobuchi, T. Matsuda, *et al.*, "Synthesis of a metal oxide with a room-temperature photoreversible phase transition," *Nature chemistry*, vol. 2, no. 7, p. 539, 2010.

[140] J. Qiu, W. Zheng, R. Yuan, *et al.*, "A novel 3d nanofibrous aerogel-based mos2@ co3s4 heterojunction photocatalyst for water remediation and hydrogen evolution under simulated solar irradiation," *Applied Catalysis B: Environmental*, vol. 264, p. 118 514, 2020.

[141] A. Ginsburg, D. A. Keller, H.-N. Barad, *et al.*, "One-step synthesis of crystalline mn2o3 thin film by ultrasonic spray pyrolysis," *Thin Solid Films*, vol. 615, pp. 261–264, 2016.

[142] A. Bocquet, T. Mizokawa, K. Morikawa, *et al.*, "Electronic structure of early 3d-transition-metal oxides by analysis of the 2p core-level photoemission spectra," *Physical Review B*, vol. 53, no. 3, p. 1161, 1996.

[143] S. Mattsson and B. Paulus, "Density functional theory calculations of structural, electronic, and magnetic properties of the 3d metal trifluorides mf3 (m= ti-ni) in the solid state," *Journal of computational chemistry*, vol. 40, no. 11, pp. 1190–1197, 2019.

[144] O. O. Balayeva, A. A. Azizov, M. B. Muradov, *et al.*, "β-nis and ni3s4 nanostructures: Fabrication and characterization," *Materials Research Bulletin*, vol. 75, pp. 155–161, 2016.

[145] I. T. Sines, R. Misra, P. Schiffer, and R. E. Schaak, "Colloidal synthesis of non-equilibrium wurtzite-type mnse," *Angewandte Chemie International Edition*, vol. 49, no. 27, pp. 4638–4640, 2010.

[146] L. De Haart, A. De Vries, and G. Blasse, “Photoelectrochemical properties of mgtio3 and other titanates with the ilmenite structure,” *Materials research bulletin*, vol. 19, no. 7, pp. 817–824, 1984.

[147] C. Lokhande, A. Ennaoui, P. Patil, *et al.*, “Process and characterisation of chemical bath deposited manganese sulphide (mns) thin films,” *Thin Solid Films*, vol. 330, no. 2, pp. 70–75, 1998.

[148] M. Ikeda, K. Itoh, and H. Sato, “Electrical and optical properties of cds-mns single crystals,” *Journal of the Physical Society of Japan*, vol. 25, no. 2, pp. 455–460, 1968.

[149] L. Gnanasekaran, R. Hemamalini, R. Saravanan, *et al.*, “Synthesis and characterization of metal oxides (ceo2, cuo, nio, mn3o4, sno2 and zno) nanoparticles as photo catalysts for degradation of textile dyes,” *Journal of Photochemistry and Photobiology B: Biology*, vol. 173, pp. 43–49, 2017.

[150] H. Kalt, “General properties: Datasheet from landolt-börnstein - group iii condensed matter · volume 34c2: “optical properties. part 2” in springermaterials (https://doi.org/10.1007/10860224_1),” C. Klingshirn, Ed., Copyright 2004 Springer-Verlag Berlin Heidelberg, Part of SpringerMaterials, accessed 2020-08-26.

[151] C.-W. Lee, K.-W. Lee, and J.-S. Lee, “Optoelectronic properties of β-fe2o3 hollow nanoparticles,” *Materials Letters*, vol. 62, no. 17-18, pp. 2664–2666, 2008.

[152] J. Kang, A. Hirata, L. Kang, *et al.*, “Enhanced supercapacitor performance of mno2 by atomic doping,” *Angewandte Chemie International Edition*, vol. 52, no. 6, pp. 1664–1667, 2013.

[153] M. J. Young, A. M. Holder, S. M. George, and C. B. Musgrave, “Charge storage in cation incorporated α-mno2,” *Chemistry of Materials*, vol. 27, no. 4, pp. 1172–1180, 2015.

[154] M. Lorenz, G. Pettit, and R. Taylor, “Band gap of gallium phosphide from 0 to 900 k and light emission from diodes at high temperatures,” *Physical Review*, vol. 171, no. 3, p. 876, 1968.

[155] I. Catalano, A. Cingolani, and A. Minafra, “Multiphoton transitions at the direct and indirect band gaps of gallium phosphide,” *Solid State Communications*, vol. 16, no. 4, pp. 417–420, 1975.

[156] R. Asahi, T. Morikawa, T. Ohwaki, K. Aoki, and Y. Taga, "Visible-light photocatalysis in nitrogen-doped titanium oxides," *science*, vol. 293, no. 5528, pp. 269–271, 2001.

[157] O. Polat, Z. Durmus, F. Coskun, M. Coskun, and A. Turut, "Engineering the band gap of lacro 3 doping with transition metals (co, pd, and ir)," *Journal of Materials Science*, vol. 53, no. 5, pp. 3544–3556, 2018.

[158] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001.

[159] L. Le Gratiet, "Bayesian analysis of hierarchical multifidelity codes," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 1, no. 1, pp. 244–269, 2013.

[160] L. Le Gratiet and J. Garnier, "Recursive co-kriging model for design of computer experiments with multiple levels of fidelity," *International Journal for Uncertainty Quantification*, vol. 4, no. 5, 2014.

[161] C. W. Andersen, R. Armiento, E. Blokhin, *et al.*, "Optimade, an api for exchanging materials data," *Scientific data*, vol. 8, no. 1, pp. 1–10, 2021.

[162] J. Shen, S. D. Griesemer, A. Gopakumar, *et al.*, "Reflections on one million compounds in the open quantum materials database (oqmd)," *Journal of Physics: Materials*, vol. 5, no. 3, p. 031 001, 2022.

[163] S. Sun, L. Lannom, and B. Boesch, "Handle system overview," Tech. Rep., 2003.

[164] S. Lyons, "Persistent identification of electronic documents and the future of footnotes," *Law Libr. J.*, vol. 97, p. 681, 2005.

[165] R. V. Guha, D. Brickley, and S. Macbeth, "Schema. org: Evolution of structured data on the web," *Communications of the ACM*, vol. 59, no. 2, pp. 44–51, 2016.

APPENDIX A

MAIN PROJECTS

A.1 2D Inks: Band Alignment

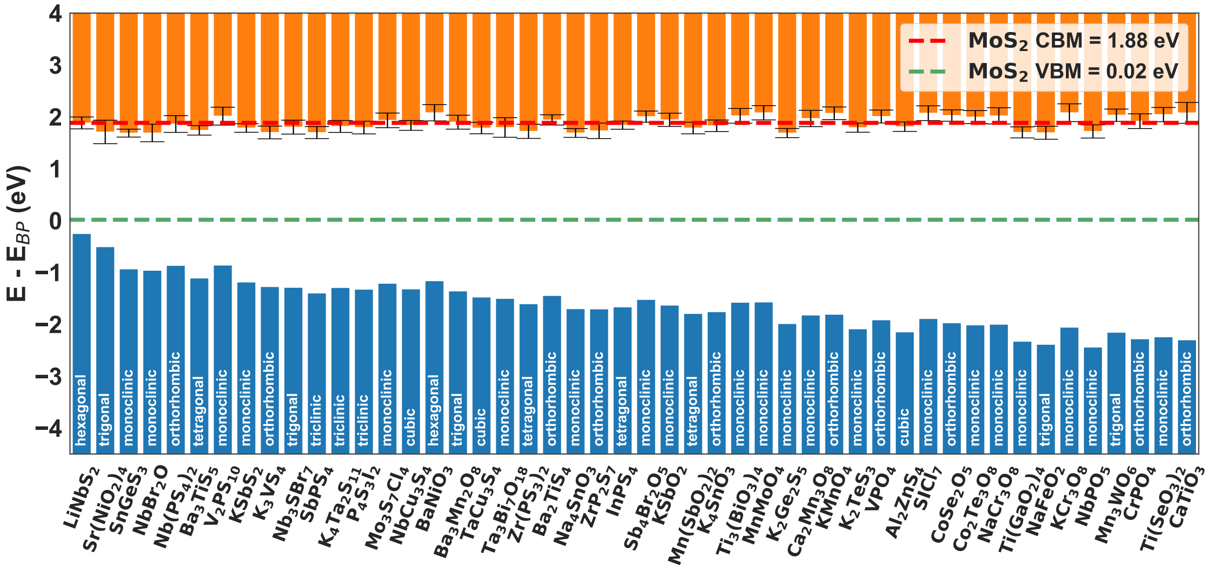
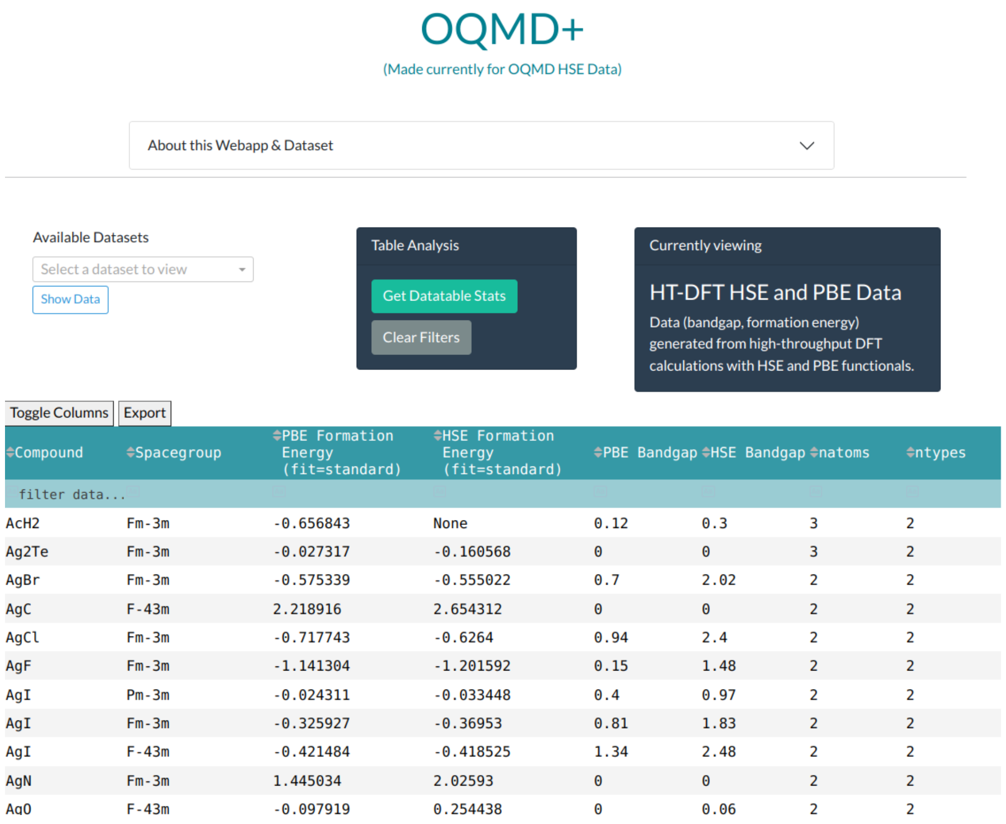


Figure A.1: The band alignment of ternary oxides and ternary sulfides based on their DFT-PBE bulk bandstructure, cross-referenced based on the common branch-point energy value. Most of the ternary materials were filtered out in the subsequent step due to the lack of experimentally reported data on their 2D phase stability, nature of carriers, synthesis methods, etc.

A.2 Multi-fidelity Modeling: Deployment of Predictions

A.2.1 hse.oqmd.org

We created a new web portal to distribute the datasets used and generated during the course of the multi-fidelity modeling project and is hosted at the URL https://hse.oqmd.org since September 2022. This web portal, as shown in Figure A.2, is created using the Plotly Dash framework and deployed on the Google Cloud Run serverless deployment platform. The data analysis operations, such as filtering, sorting, etc, are performed locally on the user's browser to avoid overload at the servers. The main achievement of this web portal in materials science is providing a scalable server solution to quickly deploy small to medium-sized material datasets in common CSV or JSON formats to the public domain via a user-friendly interface. This work contributes to the building of a robust material data infrastructure. As of November 2022, the web portal at hse.oqmd.org serves the training data and the predictions from the co-kriging model.



HT-DFT HSE and PBE Data

Data (bandgap, formation energy) generated from high-throughput DFT calculations with HSE and PBE functionals.

Compound	Spacegroup	PBE Formation Energy (fit=standard)	HSE Formation Energy (fit=standard)	PBE Bandgap	HSE Bandgap	natoms	ntypes
AcH2	Fm-3m	-0.656843	None	0.12	0.3	3	2
Ag2Te	Fm-3m	-0.027317	-0.160568	0	0	3	2
AgBr	Fm-3m	-0.575339	-0.555022	0.7	2.02	2	2
AgC	F-43m	2.218916	2.654312	0	0	2	2
AgCl	Fm-3m	-0.717743	-0.6264	0.94	2.4	2	2
AgF	Fm-3m	-1.141304	-1.201592	0.15	1.48	2	2
AgI	Pm-3m	-0.024311	-0.033448	0.4	0.97	2	2
AgI	Fm-3m	-0.325927	-0.36953	0.81	1.83	2	2
AgI	F-43m	-0.421484	-0.418525	1.34	2.48	2	2
AgN	Fm-3m	1.445034	2.02593	0	0	2	2
AgO	F-43m	-0.097919	0.254438	0	0.06	2	2

Figure A.2: The web portal created and deployed at the URL *hse.oqmd.org* to serve the datasets used and generated in the multi-fidelity co-kriging modeling project

APPENDIX B
OQMD

B.1 OQMD and FAIR Data

Open Quantum Materials Database[3], [4], [162] (OQMD) is one of the largest computational materials databases in the world as of 2022. It stands as one of the pillars of public material data infrastructure and is used by researchers around the world in their work[162]. We worked on three different projects within the OQMD toward enhancing its findability, accessibility, interoperability, and reusability (FAIR[2]).

In the first project, we co-created a universal standard for materials data transfer across most of the major high-throughput database providers, called the OPTIMADE REST API standard[161]. The OPTIMADE specification is a set of rules to follow while deploying a corresponding REST API data transfer system in materials databases. Every database that implements an OPTIMADE REST API can be queried using the same syntax by changing only the base domain URL. This work highly enhances the interoperability and accessibility of the materials databases. The OPTIMADE standard formulation work was funded by Centre Européen de Calcul Atomique et Moléculaire (CECAM), and the work on implementing the same API in OQMD was funded by grants from Toyota Research Institute (TRI).

In the second project, we created and hosted a Handle system[163] to assign persistent identifiers to the materials data in OQMD that prevents link-rot issues[164]. The persistent identifiers system implemented in OQMD are similar to Digital Object Identifier (DOI) system, and it increases the accessibility and reusability of data in OQMD. We also created and deployed structured data based on the schema.org[165] standards for all material data pages in the OQMD database to increase their findability and machine-actionability[2] in public web search engines. Both of the works completed in this project are funded by grants from the National Institute of Standards and Technology (NIST).

During the third project, we transferred the OQMD server from on-premises hosting to a scalable, container-orchestrated cloud hosting system, significantly reducing the latency and increasing the bandwidth. This transition of the hosting environment enables a more robust data infrastructure designed for long-term deployment. The full cloud architecture of OQMD as of October 2022, is given in Figure B.1.

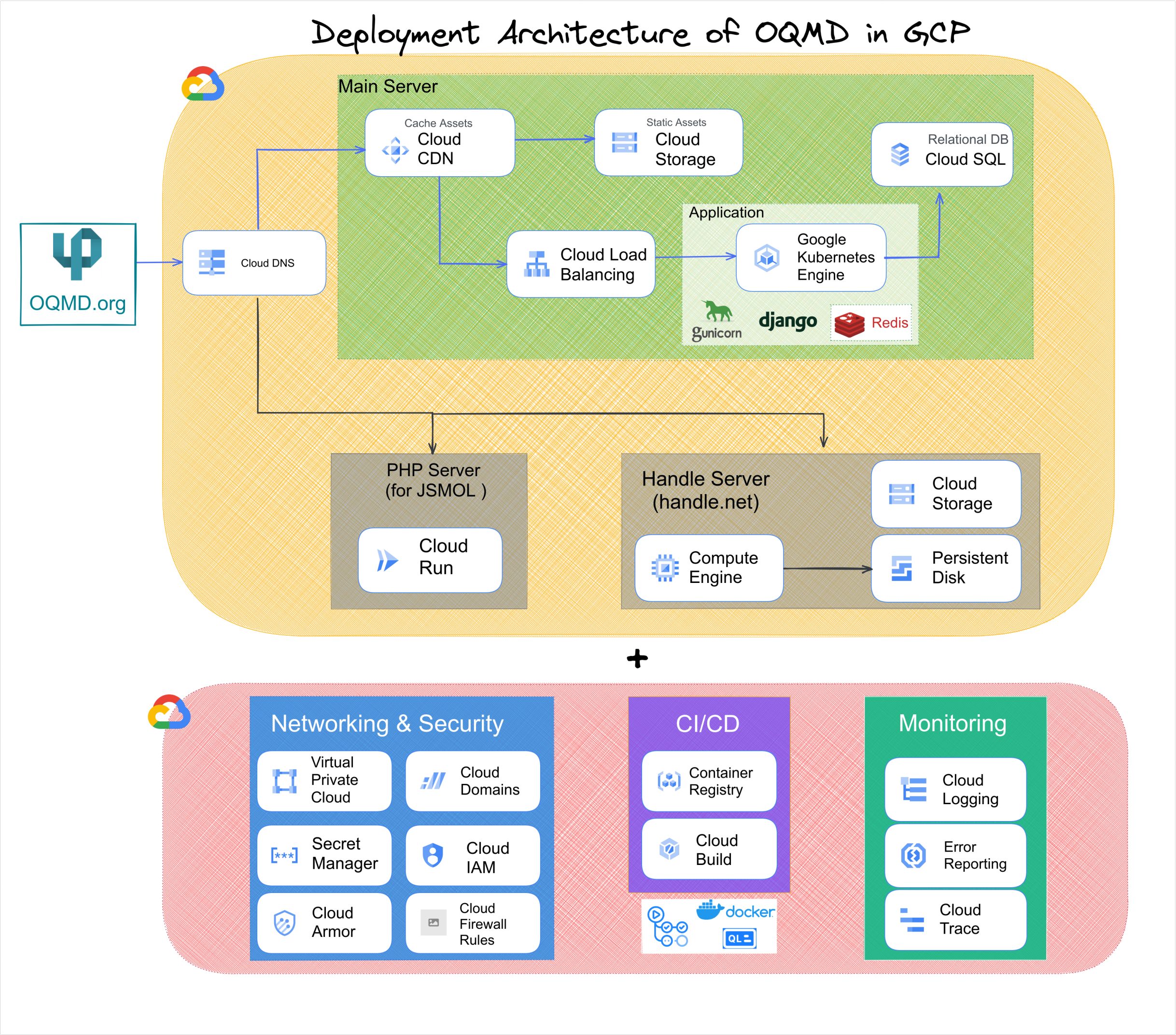


Figure B.1: The Cloud infrastructure of OQMD - as of October 2022