Computational Materials Science 243 (2024) 113108

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Full length article

A guide to discovering next-generation semiconductor materials using atomistic simulations and machine learning

Arun Mannodi-Kanakkithodi

School of Materials Engineering, Purdue University, West Lafayette IN 47904, USA

ARTICLE INFO

Keywords:
Semiconductors
Density functional theory
Machine learning
Materials informatics
Computational screening

ABSTRACT

With massive influx of new funding and emergence of modern facilities and centers, the area of semiconductor manufacturing and processing has attained national and global momentum like never before. To meet the rising industry demands via accelerated materials design, fundamental atomistic-level insights are crucial. Today, there are a number of published works using some combination of first principles-based density functional theory (DFT) simulations and machine learning (ML) for accelerating the prediction of semiconductor properties over large chemical spaces. Materials discovery based on "DFT-ML" significantly shortens the time required to select promising chemistries for laboratory synthesis and characterization, and are thus frequently combined with rational experimentation for a variety of semiconductor classes and properties, for applications such as solar absorption, light emission, power electronics, photocatalysis, and quantum technology. In this review article, I discuss some of the key concepts behind accelerating the prediction of fundamental semiconductor properties, highlighting some available datasets and tools. In the context of published literature as well as my own past and ongoing research, I discuss some major studies involving the discovery of new materials for solar cells, water splitting, and wide band gap semiconductors, focusing on the application of DFT and/or ML for prediction and understanding of semiconductor alloys, bulk stability, electronic, optical, and dielectric properties, and defect behavior. This article aims to highlight established computational recipes, large databases, and ML-based prediction and optimization tools that may guide researchers in their own semiconductor design and discovery endeavors.

1. Introduction

One of the clearest markers of human progress is the signing of major scientific legislation from the governments of the world. The CHIPS and Science Act, signed into law in August 2022, provided a historically large amount of money for semiconductor research and manufacturing within the United states [1]. As a result, a number of large semiconductor manufacturing facilities, fabrication plants, and research centers have emerged across the country. Many international partnerships have been fostered with leading semiconductor companies and thousands of new jobs have either already been created or are on the way. Furthermore, universities have begun developing semiconductor degree programs, hiring more semiconductor-related faculty, and generally incorporating concepts of semiconductor research and development in curricula [2]. We now find ourselves in a "semiconductor moment", which promises to lead to the development of next-generation transistors, photodiodes, solar cells, electronics, and other technologies. Fig. 1(a) shows the projected growth of the semiconductor market size, expected to cross 800 billion dollars by 2030 [3].

The immediate impact of the CHIPS Act has echos of the Materials Genome Initiative (MGI) [4], announced by the US government in 2011 with the aim of creating policy, resources, and infrastructure for "discovering, manufacturing, and deploying advanced materials twice as fast and at a fraction of the cost compared to traditional methods". The MGI has been a resounding success to date, with great quantity and quality of computational and experimental research resulting in the discovery of new and improved battery electrodes and electrolytes [5], absorbers for solar cells [6], capacitors for dielectrics [7], thermoelectrics for refrigeration [8], and so on. The application of methods rooted in data science, machine learning (ML), and artificial intelligence (AI) is key to accelerated data-driven materials discovery [9], even giving birth to the field of *materials informatics*, which is now sufficiently mainstream that there are graduate courses, PhD projects, and degree programs in materials science and engineering departments with a focus on informatics. The 2021 MGI Strategic Plan additionally identifies the goals of unifying the materials innovation infrastructure, harnessing the power of materials data, and educating the next generation of materials research workforce; this is pictorially represented in the graphic in Fig. 1(b). These goals connect rather well with the goals of the CHIPS Act. The use of AI/ML and ma-

E-mail address: amannodi@purdue.edu.

https://doi.org/10.1016/j.commatsci.2024.113108
Received 31 July 2023; Received in revised form 3 April 2024; Accepted 15 May 2024
Available online 27 May 2024
0927-0256/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

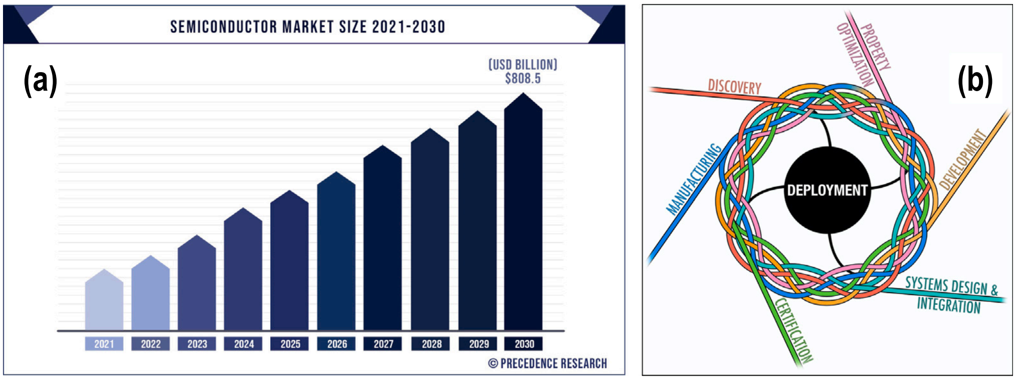

Fig. 1. (a) Projected growth of the semiconductor industry [3]. (b) A glimpse of MGI's 2021 Strategic Plan [4].

terials data, and training in materials informatics, will be invaluable for designing next-generation semiconductor materials and obtaining increased understanding of structure–chemistry–property relationships that contribute to improved performance.

Computational materials scientists have a significant role to play in the future of semiconductor research, especially within the context of the MGI. Today, materials discovery happens on the computer long before any experimental confirmation or deployment in real-world application. Computational methods, including first principles simulations, classical mechanics, and ML predictions, are vital cogs in the materials discovery machinery. In particular, high-throughput (HT) density functional theory (DFT) computations, often coupled with ML models, are instrumental in accurate atomistic-level property estimates and virtual screening of promising materials [10,11]. DFT provides a convenient way to simulate crystalline structures of inorganic and organic materials, determine relative stabilities, explain measured diffraction or luminescence peaks, and generally guide experimental synthesis and characterization [12], while ML models trained upon DFT data ensure that computations do not need to be performed endlessly, and property predictions over hundreds of thousands of materials with DFT-level accuracy can be made in a fraction of the time [11,13]. In the space of semiconductors, DFT and ML have been extensively applied for successful discovery of new compositions, structures, dopants, etc. that lead to desired electronic band gaps, optical absorption coefficients, dielectric breakdown strength, and defect formation energies [14,15]. A typical DFT-driven semiconductor materials discovery endeavor can be formulated as: identifying a chemical space of interest, benchmarking the DFT level of theory and generating a dataset, training ML models for predictions over much larger spaces, and finally, performing experimental validation [11,13]. Such an approach, or at least certain elements of it, has been applied plenty of times over the last couple of decades, and promises to continue being the way forward.

Novel semiconductors are sought after for applications that include solar absorption, catalysis, transistors, light emission, power devices, and quantum information sciences. Experimentally, it is often challenging to determine multiple properties of interest simultaneously, which may include characterizing the crystal structure and determining long-term stability in addition to electronic, optical, dielectric, surface, and defect properties; ab initio simulations provide the way out here and can reasonably be applied for multi-objective estimates and optimization across a large number of candidates. When performing computational materials design, it is important to first determine the thermodynamic stability and "synthesizability" of any candidate, often in the presence of a solvent or atmospheric conditions of moisture and temperature. One may also be interested in understanding the temporal structural robustness and resistance to degradation of the material (e.g., halide perovskites, as exciting as they are as absorbers in solar cells, are notoriously unstable in air or water [16,17]), as well as in unraveling relevant decomposition pathways. DFT provides a fundamental understanding of bulk stability of a material in different phases and its decomposition likelihood, even though synthesizability predictions (recently being explored in-depth using ML [18,19]) and long-term stability may be harder targets.

By accessing all possible low energy configurations for given chemistries—something DFT is reliable at—many properties of interest can then be calculated using appropriate methods and approximations. DFT-computed electronic band structures reveal band gaps, carrier effective masses, and mobilities; perturbation theory calculations help estimate dielectric constants, vibrational frequencies, and phonon properties; large supercell defect calculations lead to formation energies and energy levels of point defects and dopants; and surface slab and heterostructure calculations reveal surface and interface properties. Given the wide variety of approaches within DFT (in conjunction with ML) currently in use for calculating stability and properties of interest, I believe it is vital to review these methods and discuss their utility, in the context of how they have been applied in the literature. Such an overview would be of particular benefit to experimental researchers without in-depth knowledge of ab initio methods interested in understanding the strengths and limitations of DFT-ML insights, as well as computational researchers aiming to build expertise in certain types of property predictions.

The plethora of studies on computational prediction of key semiconductor properties provides both challenges and opportunities. Almost always, there exists some published dataset, computational recipe, tool, or model, that can be applied by any researcher for predicting properties of interest for new materials, and performing design/optimization for a given application; however, it can also be confusing to determine the ideal DFT level of theory, computational parameters, or ML models to use for specific chemistries, structures, and properties. The semiconductor and materials informatics communities will benefit from a summary of relevant computational recipes and approximations, valuable datasets and tools, and successful examples of how such methods have been applied in the recent past. With this in mind, in this article, I aim to review the broad topic of using DFT and ML for accelerated discovery and understanding of semiconductors, with a focus on solid-state inorganic materials and their applications typically requiring some combination of desired electronic, optical, dielectric, and defect properties. Ultimately, this article serves to act as a guide for researchers in their semiconductor design endeavors, by covering three major things:

1. Methods to computationally predict structure, stability, band gap, dielectric constant and breakdown, optical absorption and related efficiencies, surface and interface energies, and defect formation energies and other defect-related properties.
2. Major datasets, tools, and ML models from the literature that can be readily utilized by researchers in their own work, towards predicting the aforementioned properties and driving materials discovery for a variety of electronic applications.

3. A few specific case studies from the literature on how HT-DFT, computational screening, and ML have been used to discover new solar cell materials, WBG semiconductors, and photocatalysts.

While ML models are also trained on experimental data and are actively being applied to perform autonomous synthesis and testing and to accelerate materials design right in the laboratory [20,21], this article will primarily cover ML applied on DFT data to aid prediction and screening with first principles accuracy prior to experimental consideration. In the following sections, the essential concepts of a DFT-ML methodology are discussed, followed by how it is applied to predict crystal structures, bulk and alloy stability, electronic, optical, and dielectric properties, surface and interface properties, and defect behavior. A few major examples from the literature are then highlighted, including most useful datasets and models, the discovery of novel semiconductor absorbers for solar absorption, new candidates for photocatalysis, and wide band gap materials for power electronics. The article will finish with a summary and some words on the promise of DFT-ML for future semiconductor R&D.

2. DFT-ML for materials discovery

Improving the performance of semiconductor-based devices is fundamentally an exercise in mathematical optimization performed within the atom-composition-structure (ACS) space [22] of semiconductors to achieve the desired combination of electronic, optical, dielectric, thermoelectric, and defect properties. When considering all possible cations and anions, phases and polymorphs, alloying and doping, etc., the chemical space of well-known semiconductor classes becomes so prohibitively large that it is impossible to perform brute-force experiments or even computations to evaluate the structure and properties of all materials [11,23]. However, this also provides massive opportunities for composition- and structure-engineering in order to tune the properties. One of the most interesting ways of semiconductor engineering is via chemical substitution in “prototype phases” within different chemical sub-spaces. Materials represented by a general chemical formula, such as “AB” binaries, “ABC_2” ternaries, “A_2BX_4” ternaries, etc., are typically found to occur in one of few well-defined crystalline arrangements or symmetries, such as cubic or orthorhombic, zincblende or Wurtzite, vacancy-ordered, layered, etc. For a complete understanding of structure–property relationships, semiconductors must be studied in different possible phases, leading to phase stability diagrams, and encouraging the simulation of completely novel compositions within the same phases.

Some classic examples of semiconductor classes and the types of phases or crystal structures they adopt are listed below:

- Canonical group IV, III–V and II–VI binaries, ternaries, quaternaries, and beyond (via ion-mixing or alloying), in multiple prototype phases such as Wurtzite (WZ), zincblende (ZB), and rocksalt (RS) [24].
- ZB-derived materials such as I–III–VI_2 (e.g., $CuInSe_2$) and I_2-II-IV-VI_4 (e.g., Cu_2ZnSnS_4 or CZTS) compounds, which adopt phases such as Kesterite and Stannite [25,26].
- Perovskites that manifest as standard ABX_3 crystalline structures with corner-shared octahedra, AA’BB’X_6 double perovskites, AA’BX_6 vacancy-ordered double perovskites, or layered 2D perovskites. Perovskites can be halide, oxide, chalcogenide, or nitride, and either purely inorganic or hybrid organic–inorganic [11,27,28].
- A host of binary, ternary, or quaternary oxides (e.g., Ga_2O_3), chalcogenides (e.g., SnS, Sb_2Se_3), and nitrides (e.g., $ZnSnN_2$), many of which can be found in online databases [29,30].

All the above types of semiconductors and more have been extensively studied using DFT. Although theory always has its limitations in terms of accuracy and computational time, DFT-level prediction of structure, stability, and properties is a widely accepted “initial step” towards eventual materials discovery. With modern computing power and many available tools such as atomate [31] and pylada [32], high-throughput computations and subsequent analysis and screening can be readily performed for most problems, involving tens to hundreds to thousands of computations [28,33,34]. However, because the ACS space is practically infinite in every material class—since any number of competing phases and alloy compositions could be generated, in theory—DFT only takes us so far without being coupled with state-of-the-art regression, classification, or optimization algorithms that help make accurate predictions over massive spaces or directly generate new ACS combinations that satisfy multiple objectives [11,35,36]. Thus, “DFT-ML” is the crucial marker for virtual materials design and guiding experimental synthesis and testing for validation and discovery. Fig. 2 shows a typical workflow of a DFT-ML process: from defining the semiconductor chemical space, to benchmarking the level of theory being used in DFT against known experimental values, to generating a reliable DFT dataset, to training a variety of ML models, to performing inverse design as well as high-throughput prediction and screening, to finally closing the loop on materials discovery via experiments. It should be noted that error bars or uncertainties will typically be different from experiments and from DFT—depending on the synthesis and characterization conditions, the same property values may not be reproducible, which often results in different studies reporting slightly different properties for the same material, such as the band gap and photovoltaic efficiency of a given perovskite composition [37]. DFT is more uniform since the crystal structure and computational parameters can be kept fixed from one study to another, but will show differences if the level of theory or DFT functional is changed. This difference in uncertainties should be kept in mind when comparing experiment to different flavors of DFT.

DFT involves solving many one-electron quantum mechanical (QM) equations to go from an input structure to the outputs (optimized structure, energies, band gap, etc.), via intermediate calculations of electron densities, forces and energies, and ultimately, the desired properties [12]. Published literature contains many examples of training ML force-fields or interatomic potentials by fitting functions to DFT data [38,39], such that atomic forces and energies may be predicted instantly without using QM for any new configurations. Similarly, ML is used for predicting charge densities to accelerate the DFT computation [40]. The type of DFT-ML this article focuses on is where the property of interest is directly predicted from a given semiconductor ACS input [41,42]. This involves training predictive models based on random forests, Gaussian processes, neural networks, or other state-of-the-art techniques, using the *input* → *output* formatted DFT dataset, where all materials are converted into a set of unique but easily attainable inputs [43,44]. If a predefined chemical space has $> 10^5$ possible compounds, such models could be trained on a representative dataset containing $< 10^3$ points and used to make predictions across the entire space enumerated in terms of their input vectors, based on which screening is performed for promising materials. This process is often made more efficient by using “inverse design” strategies [35,36], where the ML models may act as surrogates for predicting properties of materials generated using evolutionary algorithms, Bayesian optimization, or generative methods [45–48], with the aim of minimizing a loss function that includes multiple target properties. Zunger defined inverse design as declaring a desired functionality (properties of interest) of a new material and performing theoretical calculations (such as using DFT) to determine which candidate(s) might show such properties [22]. The direct prediction of suitable semiconductor materials that are stable and show desired electronic, optical, and defect properties can be achieved via inverse design strategies that include high-throughput computational screening, using correlations and chemical design rules

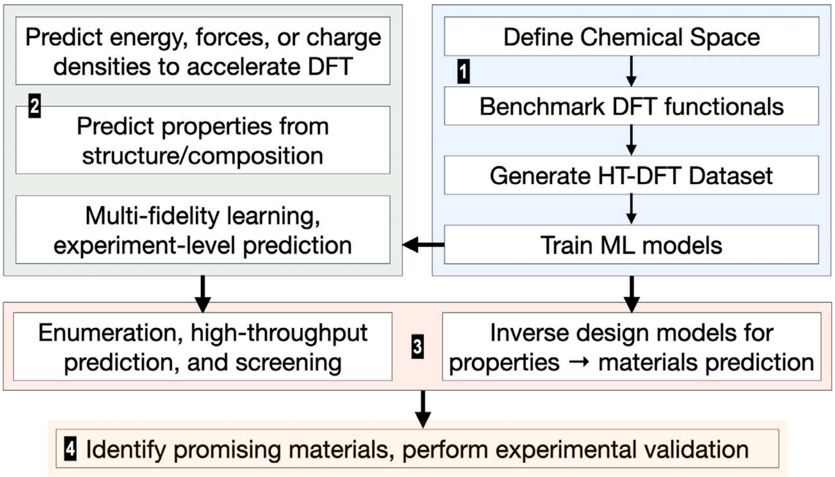
Predict energy, forces, or charge densities to accelerate DFT
2
Predict properties from structure/composition
Multi-fidelity learning, experiment-level prediction
Define Chemical Space
1
Benchmark DFT functionals
Generate HT-DFT Dataset
Train ML models
Enumeration, high-throughput prediction, and screening
3
Inverse design models for properties → materials prediction
4
Identify promising materials, perform experimental validation

Fig. 2. A typical DFT-ML framework for materials discovery.

to narrow down the search, and using a variety of global optimization or generative algorithms [49].

For successful development of a DFT-ML prediction and design framework which can be coupled with experiments, there are many important things to keep in mind. First and foremost is the quality and quantity of computational data. It is now fairly routine to perform geometry optimization and electronic structure computations on crystal structures with up to ~100 atoms: the accuracy of the calculated properties heavily depends on the nature of the DFT functional used [50–52], as will be discussed more in the next sections. If DFT has a large error compared with measurements, the ML models will only propagate this error further, leading to unreliable prediction and screening [13]. Thus, DFT-experiment benchmarking is crucial. Once reliable DFT data is generated for compounds that sufficiently represent the ACS diversity of the chemical space, suitable input descriptors or feature vectors must be defined: typical inputs include one-hot encoded composition vectors, elemental properties of constituent atoms, cheaper surrogate or proxy properties, Coulomb matrices, atomic overlap functions, and crystal graph-based representations [43,44,53]. Next, the data must be split into training, validation, and test sets, and a reasonable ML algorithm must be chosen and rigorously optimized. The choice of the specific algorithm is less important than the data and descriptors, as it has been well demonstrated that almost any nonlinear regression method will return roughly similar predictions on the test data points once it is optimized completely, in terms of training data size, input dimensions, and regression hyperparameters [9,23].

Standard ML methods include random forests, ridge regression, Gaussian processes, decision trees, gradient boosting, and simple neural networks [9,54], whereas massive datasets and complex predictive tasks are typically handled using “deep learning” (DL), which refers to a sophisticated neural network model with many hidden layers [55]. Each of these methods have their merits for different tasks, but DL approaches are becoming increasingly prominent for predicting atomic forces and crystal energies [56,57], and for generative design of novel inorganic materials [58,59]. Furthermore, a key component of training and optimizing an ML model is “feature engineering”, wherein a high-dimensional feature set perceived to contain all the necessary information for accurate prediction may be trimmed down to the essential dimensions or converted into a set of new features that better correlate with the outputs. Typical feature engineering methods include eliminating dimensions based on Pearson correlation coefficients, principal component analysis, 2-point statistics, and methods such as LASSO and SISSO where millions of compound features are generated using primary features and the best ones are selected for eventual model training [60–62]. Performance of the DFT-ML models is evaluated using metrics such as root mean square error (RMSE), R^2 value, or mean absolute error (MAE), following which the best models are deployed for prediction, screening, and inverse design. In the following sections, the accuracy of DFT and ML for predicting the structure, stability, and various relevant properties of semiconductors is discussed, in an attempt to shed light on how such predictions contribute to materials discovery for various important applications.

3. Recipes for prediction of semiconductor properties

3.1. Crystal structure and stability prediction

There is a vast body of work on predicting the thermodynamic stability of bulk crystalline materials from first principles, which helps in determining ground state and competing structures/phases for any given material composition [63–65]. Integral to such approaches is the efficient navigation of the potential energy landscape via evolutionary, minima or basin hopping, or ML-based algorithms [65–67], and the accurate estimation of a DFT-level energy for structures that are metastable or lie on local or global minima. For predicting the structure and stability of semiconductors using DFT-based geometry optimization, it is important to simulate them in multiple prototype phases and explore polymorphs via intentional atomic or polyhedral distortions as well as re-optimization in larger supercells [68]. Naturally, efficient and accurate crystal structure prediction is a major field of study in itself [66,69–71], necessitated especially for completely new materials that may not have analogues in structure databases, but a detailed look at structure prediction is beyond the scope of this article.

While determining ground state structures is important, ***polymorphs*** are almost equally important, as the best property estimates may arise from an ensemble of competing phases, using for instance a Boltzmann factor-based weighted property average [72] rather than properties from a single structure. Eq. (1) shows how such a statistically averaged property value could be determined using properties from multiple low energy polymorphs or competing phases.

$$P = (\Sigma_i P_i exp(-E_i/k_B T))/(\Sigma_i exp(-E_i/k_B T)) \qquad (1)$$

Here, E_i is the difference in the formation energy or decomposition energy (or energy above hull [73]) of any polymorph i compared to the ground state, P_i is the computed property (e.g., band gap) of the polymorph i, k_B is the Boltzmann constant, and T is the temperature of interest [72]. Polymorphs or metastable structures may also manifest in the form of ***prototype phases*** which maintain certain symmetries and are applicable to specific material classes. Fig. 3(a) shows four different phases shown by compounds belonging to the III–V or II–VI semiconductor family, as found in the Materials Project (MP) database [29]: ZB,

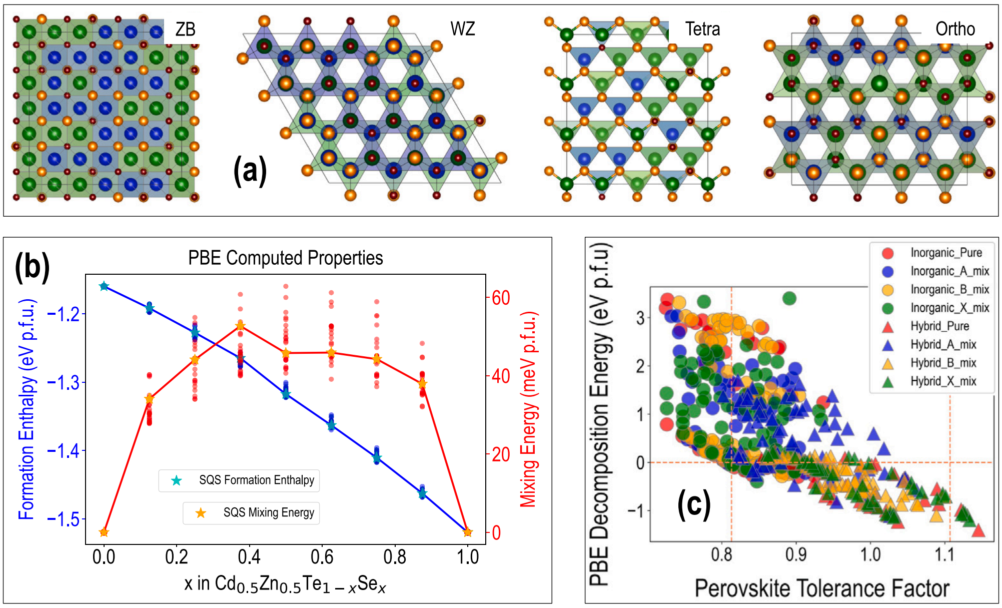

Fig. 3. (a) Four different phases found in the Materials Project [29] for binary III–V and II–VI semiconductors. (b) DFT-predicted formation enthalpy and mixing energy for $Cd_{0.5}Zn_{0.5}Te_{1-x}Se_x$ alloys, as a function of x [75]. (c) DFT-computed decomposition energy for inorganic and hybrid ABX_3 halide perovskite alloys plotted against the tolerance factor [76,77].

Source: Permission to reuse figures (b) and (c) have been obtained from IOPscience and RSC.

WZ, tetragonal, and orthorhombic phases. Similarly, ABX_3 perovskites and their derivatives adopt cubic, tetragonal, orthorhombic, hexagonal, needle-like, or other distorted phases [11,74]. For any material in a given semiconductor class, stability and properties of any compound should be determined in different possible phases before making any conclusions on the most likely structure and properties.

The semi-local GGA-PBE functional [50] is often sufficiently accurate for determining lattice parameters and relative stability. As shown in Fig. 5(a), the cubic lattice constants of ZB compounds computed from PBE match remarkably well with experimental values, and the use of the more computationally expensive hybrid HSE06 functional [51] is unnecessary for geometry optimization [75]. While GGA-PBE is the standard DFT functional of choice for most studies, it is often replaced by PBEsol [78] (a PBE functional better parameterized for solids) or PBE-D3 [79] (additional terms to take weak dispersion interactions into account, especially when organic species are present) to better reproduce lattice parameters. The stability of a crystalline semiconductor is defined in terms of formation energy (decomposition of the compound to its constituent atomic phases, such as AB → A + B), decomposition energy (compound decomposing to alternative phases, such as $ABX_3 \rightarrow AX + BX_2$), or mixing energy (for alloys, $ABB'X_3 \rightarrow ABX_3 + AB'X_3$) [23,75]. Sufficiently low or negative values indicate robust stability. For alloys, the mixing entropy is very important and often has a major stabilization effect on top of the DFT energies. For a hypothetical series of alloys $AB_xB'_{1-x}$, the mixing energy with entropic contribution can be given by Eq. (2) [23,75,76].

$$\Delta H(AB_xB'_{1-x}) = E(AB_xB'_{1-x}) - xE(AB) - (1-x)E(AB') - k_BT(xln(x) + (1-x)ln(1-x)) \qquad (2)$$

Fig. 3(b) shows the formation and mixing energy of $Cd_{0.5}Zn_{0.5}Te_{1-x}Se_x$ quaternary alloys in the ZB phase, plotted as a function of Se fraction [75]: it is seen that the former is negative throughout, whereas the latter reaches a peak of ∼50 meV per formula unit (p.f.u.) close to the middle of the series. At room temperature, the mixing entropy contribution for the 50–50 composition will be around −18 meV p.f.u., which stabilizes the alloys to some extent but keeps the total mixing energy high enough that compositions closer to the end points will be more stable. It can also be seen from Fig. 3(b) that every alloy composition has a number of configurations, 20 to be precise, and one of them corresponds to the special quasi-random structure (SQS) which represents the most statistically random alloy arrangement possible [80].

In addition to the semiconductor phase, it is important to evaluate the dependence of ***ionic ordering*** on the stability and properties, although the SQS structure is often reliably utilized as the most representative configuration. Furthermore, certain semiconductor classes have established figures of merit for determining stability and formability—such as the Goldschmidt tolerance (t) and octahedral (o) factors for ABX_3 perovskites [77], which take into account the ionic radii of A (r_A), B (r_B), and X (r_X) species. Eq. (3) to 6 show formulas for t and o, as well as modified perovskite stability factors discovered using ML by Bartel et al. [77] (t_{Bartel}) and Sun et al. [81] (t_{Sun}). Fig. 3(c) shows the perovskite decomposition energy (ΔH) plotted against the tolerance factor [76]: it can be seen that while nearly all the materials that satisfy the $\Delta H < 0$ eV condition (indicating resistance to decomposition to AX and BX_2 phases) also occur in the desired range of t values (between 0.8 and 1.1), vice-versa is not true, meaning that t alone will not reveal everything about the perovskite stability, and DFT-level energy estimates are vital. The same observations hold for t_{Bartel} plotted against ΔH as well.

Octahedral factor:

$$o = \frac{r_B}{r_X} \qquad (3)$$

Tolerance factor:

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)} \qquad (4)$$

Modified perovskite tolerance factor from Bartel et al. [77]:

$$t_{Bartel} = \frac{r_X}{r_B} - [1 - \frac{\frac{r_A}{r_B}}{ln(\frac{r_A}{r_B})}] \qquad (5)$$

Modified perovskite stability metric from Sun et al. [81] (η is the crystal volume fraction):

$$t_{Sun} = (\mu + t)^{\eta} \qquad (6)$$

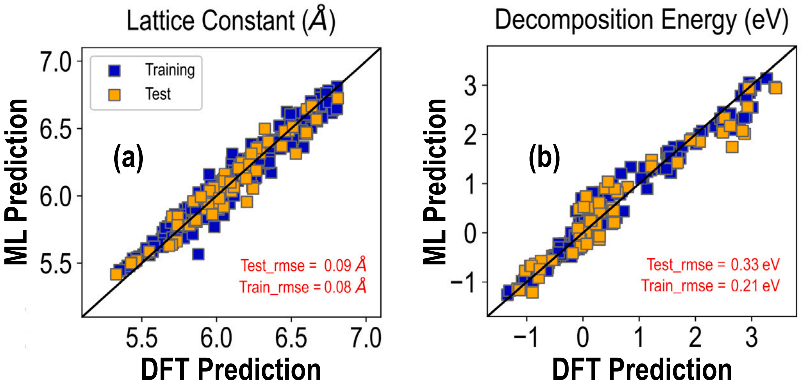

Fig. 4. Gaussian process regression (GPR) models trained on a dataset of ABX_3 halide perovskite alloys for PBE-level prediction of (a) pseudo-cubic lattice constant, and (b) decomposition energy [23].
Source: Permission to reuse the figures have been obtained from RSC.

In recent work from my research group [23], we showed that descriptors based purely on composition and well-known physical properties of constituent species (such as ionic radius and electronegativity) are sufficient to train accurate regression models for a dataset of 229 ABX_3 halide perovskite alloys with B-site mixing allowed, to predict the pseudo-cubic lattice constant (a) and ΔH. Such models trained for PBE-computed properties using Gaussian process regression (GPR) are presented in Fig. 4, showing test RMSE values of 0.09 Åfor a and 0.33 eV for ΔH, which are less than 10% errors considering the range of values across the dataset. Recently, we expanded this work to include a variety of A-site and X-site mixing as well, and allowing each compound to adopt multiple possible non-cubic phases, with the same types of descriptors as before but adding a few extra dimensions to represent the perovskite phase [76,82,83]. We find that fully-optimized random forest regression (RFR) models trained on this extended dataset of ~800 compounds show test RMSE of <0.20 eV for both PBE and HSE ΔH [83], which translates to ~20 meV per atom, highly competitive with the state-of-the-art for formation energy predictions [84,85].

While composition-based predictions typically only provide assumed ground state energies, more general structure-based models are possible using crystal Graph-based Neural Networks (GNNs), some well-demonstrated examples of which include Crystal Graph Convolutional Neural Network (CGCNN) [86], Materials Graph Network (MEGNET) [87], and Atomistic Line Graph Neural Network (ALIGNN) [88]. In their excellent paper from 2020 [84], Bartel et al. provide a comprehensive review of different types of energies computed as surrogates for stability of inorganic compounds, highlighting the most useful ML models in the literature and comparing them, and observing that structure-based models for formation/decomposition energy prediction are indeed far more useful than composition-based models, especially in chemical spaces with a sparsity of actually stable materials. Our research group recently started applying crystal graph-based methods to our perovskite dataset, by including thousands of metastable structures for every composition, leading to ALIGNN-based ΔH predictions with test RMSE ~0.08 eV [89]. Thus, a DFT-ML framework developed using an adequate DFT structure-energy dataset will yield optimized lattice parameters and stability metrics for any semiconductor, using appropriate material descriptors and ML algorithms. It should be noted that DFT-ML approaches used for semiconductor alloys mirror those used for metallic alloys [90], as simulating atomic/ionic mixing at specific lattice sites using SQS or other approaches are broadly applicable to both, and the alloy stability can be determined the same way by incorporating mixing entropy terms in formation enthalpy equations. A key difference of course would be the consideration of electronic and optical properties (such as the need for a non-zero band gap and visible range absorption) for semiconductors whereas mechanical or thermal properties may be more interesting for metallic alloys.

3.2. Predicting electronic, optical, and dielectric properties

Perhaps the single most important semiconductor property is the electronic band gap (E_g). A narrow E_g is suitable for IR sensors [93], E_g in the lower visible range is desired for single-junction solar absorption [94], a somewhat wider E_g is appropriate for tandem, multi-junction, or bifacial solar cells [95,96], and a range of E_g values may be suitable for light emission. Ultra-wide E_g semiconductors with large dielectric breakdown fields are attractive for power electronics [97–99]. Suitably aligned valence band (VB) and conduction band (CB) edges that straddle the redox potentials are important for semiconductors to be used as photocatalysts [100,101], such as for CO_2 reduction or H_2O splitting, as well as for junctions with other semiconductors or insulators in optoelectronic devices. E_g is typically computed from DFT using a highly accurate electronic structure calculation that samples a dense mesh of reciprocal space k-points, typically along a high-symmetry path [23,76,99]. A high-level electronic band structure yields the E_g as a difference between the highest occupied (VB) and lowest unoccupied (CB) states, and also yields other properties such as hole and electronic effective masses and mobilities based on the curvature of the electronic bands [99,102]. Equations for calculating electron and hole effective masses (m_e or m_h, using E-k relationships) and mobilities (using hole/electron concentrations p/n and scattering times τ_e / τ_h) are shown below.

$$m_e(m_h) = \hbar^2(d^2E/dk^2) \tag{7}$$

$$\mu_e = n\tau_e/m_e, \mu_h = p\tau_h/m_h \tag{8}$$

The single-electron energy levels from GGA-PBE are not sufficient to accurately predict E_g, as a result of which the hybrid HSE06 functional (which combines exact exchange energy from Hartree–Fock theory with the exchange–correlation energy from PBE [51]) or beyond-DFT GW approximation (which expands the self-energy in terms of the single-particle Green's function G and the screened Coulomb interaction W [52]) are necessary to obtain accurate band gaps. However, the computational expense steeply rises from PBE to HSE to GW, making it prohibitive to apply advanced theories across many materials. Whereas PBE E_g values are found to be under-predicted by 1 eV or more for most materials [103], for certain semiconductor classes such as hybrid organic–inorganic perovskites, PBE E_g is often accidentally accurate if spin–orbit coupling (SOC) or other corrections are neglected, which means it can be used as a surrogate for initial screening [23,76,104]. In HSE, the mixing parameter α must be tuned for different materials for matching band gaps with measured values [52], as the default of $\alpha = 0.25$ is not always sufficient: this is highly sensitive to material composition and adds another layer of complexity in accurate computational prediction and screening of electronic properties. For instance,

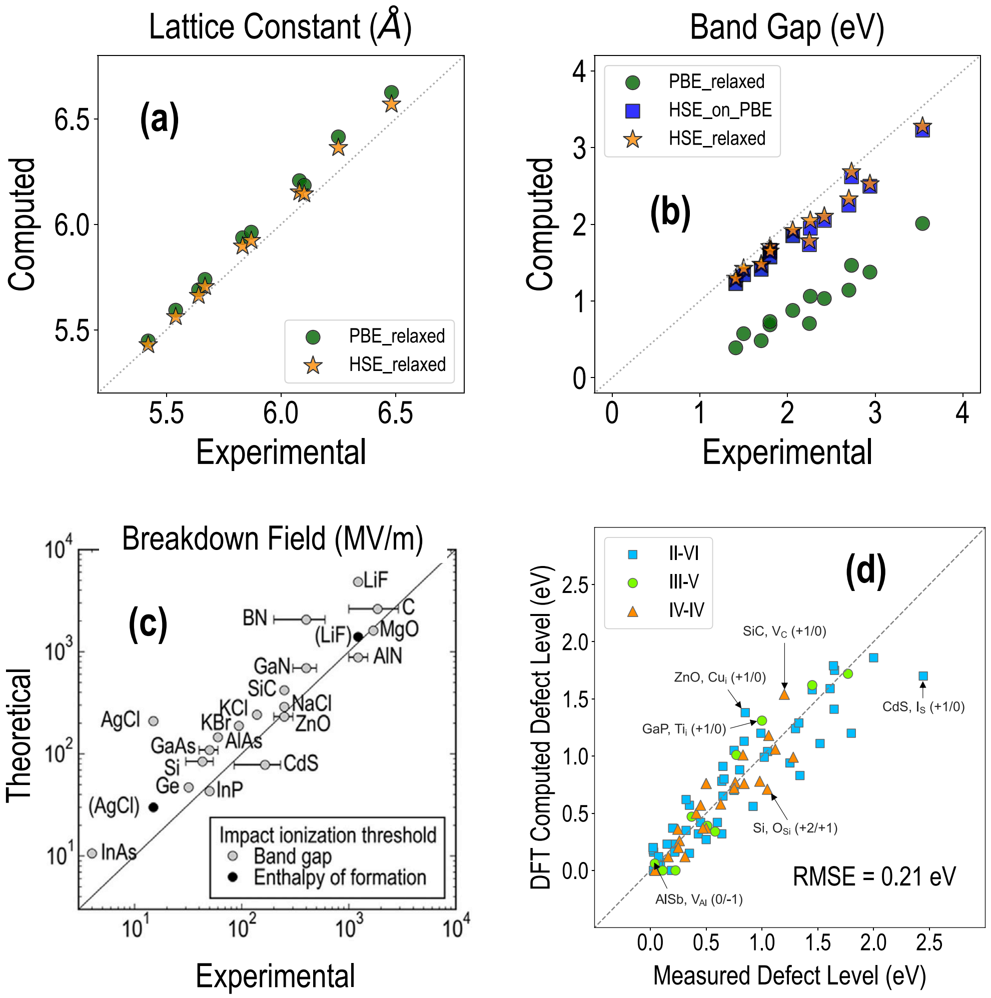

Fig. 5. A comparison of DFT-computed semiconductor properties with experimental values: (a) cubic lattice constants of ZB binaries [75], (b) band gaps of ZB binaries [75], (c) intrinsic breakdown field of inorganic insulators [91], and (d) defect energy levels of ZB binaries [92].
Source: Permission to reuse the figures have been obtained from IOPscience, ACS, and Elsevier.

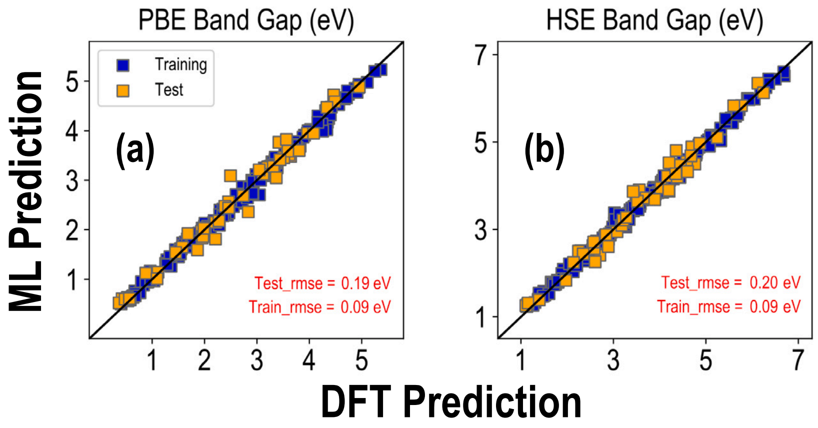

Fig. 6. RFR models trained for predicting the band gap of ABX_3 halide perovskite alloys [23], at the (a) PBE-level, and (b) HSE-level.
Source: Permission to reuse the figures have been obtained from RSC.

Pan et al. determined that $\alpha = 0.33$ reproduces the band gap of CdTe correctly [105], while in recent work, we showed that $\alpha = 0.48$ works best for $CsPbI_3$ and $CsPbBr_3$ [82].

Fig. 5(b) shows that while PBE E_g is heavily under-predicted for ZB binary compounds, static HSE on the PBE-optimized structure and full HSE optimization both predict E_g equally well with a far superior accuracy [75]. In recently published work [23], we trained RFR models for PBE and HSE E_g on the dataset of 229 ABX_3 perovskite alloys, using descriptors based on composition and elemental properties, resulting in test RMSE of ∼0.2 eV, which are also highly competitive with

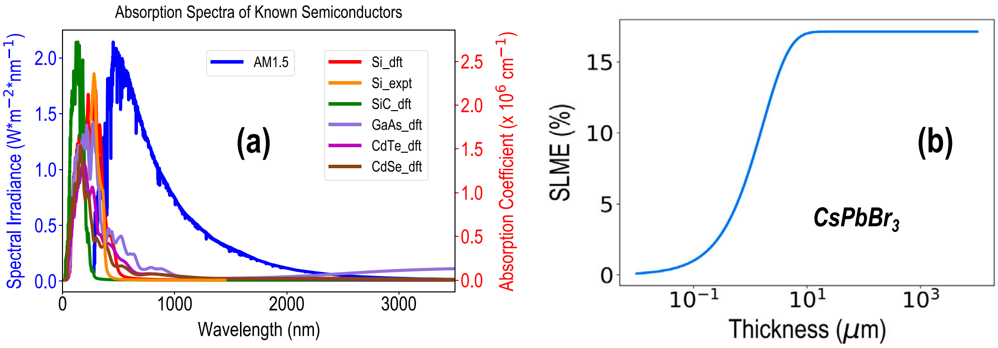

Fig. 7. (a) DFT-computed optical absorption spectra for a variety of well-known semiconductors [75]. (b) Optical absorption-derived SLME in % plotted against sample thickness for CsPbBr$_3$ [76].
Source: Permission to reuse the figures have been obtained from IOPscience and RSC.

the state-of-the-art in ML-based band gap prediction [106,107]. These models are pictured in Fig. 6, and now being extended by training over 800 compounds with A, B, or X-site mixing, in multiple phases, with E$_g$ computed from both PBE and HSE (including SOC [108]). By combining these data with around 100 data points collected from the literature for experimental band gaps of known ABX$_3$ compounds [37], we trained multi-fidelity regression models [109] to simultaneously predict PBE, HSE, and experimental E$_g$ with high accuracy: this shows an RMSE of ∼0.2 eV for the DFT points and ∼0.1 eV for the experimental points [83]. The latter is in a more narrow range than the DFT data which ranges from E$_g$ of <1 eV to >5 eV, but the distinct advantage of the multi-fidelity learning approach is that (a) accurate experiment-fidelity predictions can be made even with a small dataset, and (b) experimental predictions can be made even in regions of the chemical space not covered by the experimental dataset, by systematically learning the complex inter-dependencies between the composition and the PBE, HSE, and experimental estimates. Furthermore, our ongoing work includes training GNN models for PBE E$_g$ from a dataset of ∼ 1000 perovskite structures, which also shows RMSE < 0.2 eV, showing great promise for accurate structure to band gap prediction [89].

Optical properties of semiconductors are important for evaluating their suitability for optoelectronic applications such as photodiodes and solar cells. Optical absorption is typically simulated in DFT from complex dielectric functions using the allowed optical transitions between the energy bands in k-space, yielding as output the optical absorption coefficients (α) as a function of incident photon energy [76,110–112]. The α value remains zero until the E$_g$ when it starts rising in a linear or quadratic manner, depending on whether the band gap is direct or indirect [112]. Fig. 7(a) shows the DFT-computed absorption plots for well-known semiconductors such as SiC and CdTe [75], with the measured spectrum for Si and the AM1.5 solar irradiance spectrum [113] shown for comparison. Using the computed α values, the AM1.5 spectrum, and the nature of the band gap, the spectroscopic limited maximum efficiency (SLME) can be computed as a function of sample thickness [114,115]; as an example, the SLME vs thickness plot for the perovskite CsPbBr$_3$ in the cubic phase is shown in Fig. 7(b) [76]. Equations for calculating the absorptivity *a(E)*, the current density J, and the efficiency η as a function of sample thickness L, calculated using the solar spectral intensity I_{sun}, are shown below. The SLME is conveniently calculated vs L using an open-source package [116].

$$a(E) = 1 - e^{-2\alpha(E)L} \tag{9}$$

$$J = e\int_0^\infty a(E)I_{sun}(E)dE - J_0(1 - e^{\frac{eV}{k_BT}}) \tag{10}$$

$$\eta = \frac{P_m}{P_{in}} = \frac{max(J \times V)}{P_{in}} \tag{11}$$

The band gap inaccuracy of GGA-PBE translates to α and SLME as well, although the absorption spectrum is expected to qualitatively remain the same across different functionals. As a workaround, we showed that we could shift the PBE-computed spectrum by the difference between the PBE and HSE E$_g$ to compute an effective HSE-level SLME [76,82]. Recently, we combined a dataset of PBE and HSE SLME values for ABX$_3$ perovskites with experimental PCE values collected from the literature and trained a multi-fidelity RFR model using composition-based descriptors [83], but found errors larger than equivalent models for ΔH and E$_g$. It should be noted that excitons should be taken into account for truly reproducing the optical absorption features; this can be achieved by using GW approaches along with the Bethe–Salpeter equation (BSE) [110], which yields the optical absorption and emission spectra as well as exciton binding energies.

Finally, dielectric properties are important to evaluate the suitability of semiconductors in various electronic applications. For instance, wide band gap semiconductors (WBGSs) used in power electronics must have a static component of the dielectric constant > 10 and a dielectric breakdown field as high as possible [99]. Density functional perturbation theory (DFPT) [117] can be reliably used to calculate the static (electronic) and ionic components (ϵ_{elec} and ϵ_{ionic}, respectively) of the dielectric tensor, based on the optimized structure as input. Obtaining the intrinsic breakdown field (E$_{bd}$) from DFT involves a very complex calculation of phonon properties and electron–electron and electron–phonon interactions. Such calculations were performed by Kim et al. [91] for a variety of elemental and binary semiconductors and insulators, and a comparison of computed values with known experiments is provided in Fig. 5(c). The accuracy is more than reasonable, but these expensive computations do not provide a viable path for screening across massive chemical spaces, which is why Kim et al. developed an interpretable ML predictive model which yielded an actual function to predict the intrinsic breakdown field E$_{bd}$ using the PBE band gap (E$_g$) and the highest phonon frequency obtained from DFPT (ω_{max}), as shown in Eq. (12). While traditional DFT-ML models can be used to accelerate the prediction of ϵ_{elec} and ϵ_{ionic}, the empirical model can be reliably applied to calculate E$_{bd}$, and high-throughput screening can be performed for hundreds of thousands of semiconductors based on their dielectric properties.

$$E_{bd} = 24.442 * exp[0.315 * sqrt(E_g * \omega_{max})] \tag{12}$$

3.3. Predicting surface and interface properties

Determining the stability of free surfaces in semiconductors is crucial for understanding their performance in electronic devices, passivation effects, segregation of defects and dopants, catalytic properties, and preferred crystal growth orientations. DFT provides an ideal avenue for simulating surface slab structures by cleaving a large supercell along the plane of interest (such as (100), (110), (111), etc.) and

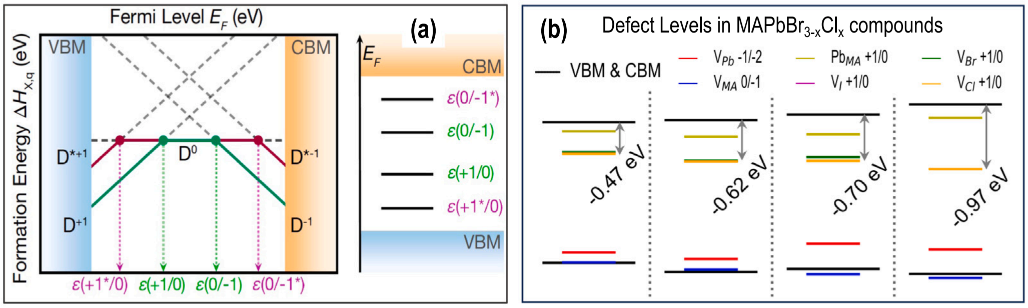

Fig. 8. (a) A typical q-dependent E_{def} vs E_F plot, and associated CTLs. (b) CTLs computed for defects in $MAPb(Br,Cl)_3$ perovskites, becoming deeper as the E_g increases with increase in Cl-content [129].
Source: Permission to reuse the figures have been obtained from RSC and ACS.

incorporating a sufficient amount of vacuum on top to prevent spurious interactions of two consecutive surfaces [118,119]. The DFT energy difference between the optimized slab and crystalline bulk along with necessary chemical potentials of species involved in creating the surface, divided by the surface area, gives the surface energy value which can be compared with experimental values in J/m^2 and used to identify the lowest energy surfaces for the material. Using a suitably thick bulk region within the slab also helps align electronic band edges w.r.t. vacuum level (far away from the surfaces), by comparing the averaged electrostatic potential in the middle to the same in a bulk crystalline supercell [119].

Additionally, studying the adsorption of chemical species on free surfaces is crucial for catalysis. Investigating suitable catalysts for CO_2 reduction or H_2O splitting would involve adsorbing atomic and molecules species such as CO, CO_2, H, H_2, and O_2, on various top sites, bridge sites, and hollow sites of the low energy surfaces identified for the given semiconductor [120]. The adsorption energy can then be computed using the total DFT energy of the optimized slab+adsorbate structure, the slab alone, and the isolated adsorbate. Zero-point energies and entropic contributions can further be calculated for the chemical process based on its likely mechanism, and added to the adsorption energy to yield a free energy for the catalytic reaction [121]. An ideal catalyst has a free energy close to 0 eV. Desirable surface and adsorption behavior, in addition to large optical absorption in the visible range and band edges that straddle the redox potentials, helps identify suitable candidates for photocatalytic reduction of CO_2, pollutant degradation, H_2 or O_2 generation, etc., providing an efficient and green avenue for performing these important chemical processes [122,123].

Stabilizing the free surfaces of multiple semiconductors also helps in creating heterostructures and studying interfaces between different semiconductors or semiconductors and metals [124]. This is achieved by combining two optimized slabs into one structure with a small amount of lattice strain, and the DFT-optimized energy of the heterostructure along with the individual slab energies yields the stability of the interface [119]. Such calculations also help in aligning the band edges of one semiconductor with the other, revealing whether they create Type I or Type II junctions [125], and determining their suitability for pn-junction devices. It has been shown that reasonably stable superlattices created in this fashion often have superior electronic and optical properties compared to the individual bulk materials, adding another dimension of tunability in semiconductors along with composition and structure [119,126]. Emerging quantum phenomena such as the Moire effect can also be explored by creating twisted superlattices [127,128], where the interplay between atomic structure and electronic correlations lead to exciton trapping, magnetism, and superconductivity.

3.4. Predicting defect properties

Point defects such as vacancies, self-interstitials, and substitutions, can form spontaneously in crystalline semiconductors and affect carrier recombination, conductivity, energy efficiency, and electronic performance [92,130–132]. As the band gap increases, defects are likely to create “deep” energy levels [129,131], which are generally harmful in terms of behaving as traps for non-radiative recombination of charge carriers, but potentially advantageous in terms of forming intermediate bands for enhanced optical absorption or being used as qubits for quantum computing [104,129]. Experimentally, defect levels are measured using techniques such as cathodoluminescence (CL) [133] or deep level transient/optical spectroscopy (DLTS or DLOS) [134]. However, sample preparation with these techniques is not trivial, and it is often impossible to assign measured levels to specific defects without utilizing formation energies and electronic signatures of defects from DFT computations. DFT is extensively applied to calculate the formation energy (E^f) of point defects as a function of defect charge (q), chemical potential (μ), and the Fermi level (E_F) as it traverses the band gap region from the VB to the CB [130]. However, DFT becomes expensive when using large supercells and advanced levels of theory, which can be addressed by coupling DFT data with ML models to accelerate predictions and screen across a combinatorial space of semiconductors and defects [92,132]. In addition to native and extrinsic point defects, it is essential to exhaustively consider different types of defect complexes which are combinations of different single defects. This adds substantially to the computational expense; e.g., V_{Ga}–O_N–2H was shown to form a stable triple complex in GaN [135], and it was recently suggested that than an O_i-As_{Te}-O_i complex creates a very low energy complex in n-type CdTe during PV application [136].

$$E^f\left(D^q, E_F\right) = E\left(D^q\right) - E\left(bulk\right) + \Sigma_i n_i \mu_i + q\left(E_F + E_{VBM}\right) + E_{corr} \quad (13)$$

$$\varepsilon\left(q_1/q_2\right) = \frac{E^f\left(q_1, E_F = 0\right) - E^f\left(q_2, E_F = 0\right)}{q_2 - q_1} \quad (14)$$

Eqs. (13) and (14) are typically used to calculate E^f for any defect as a function of q, E_F (which goes from p-type to intrinsic to n-type), and μ (such as cation-rich or anion-rich), and any defect charge transition level (CTL), $\epsilon(q_1/q_2)$. Fig. 8(a) shows a typical E^f vs E_F plot with positively charged (donors) and negative charged (acceptors) defects; line slopes are equal to q (+1, 0, or −1, in this case), and points where a defect transitions from one stable charge state to another (e.g., (+1/0)) are CTLs. Changing the μ values, such as going from Ga-rich to N-rich conditions in GaN, will move the lines up or down without changing CTLs. Typically, the same DFT functional requirements that apply for electronic and optical properties would apply for defect calculations as well, since the correct placement of the VB edge and accurate estimate of E_g is essential for plotting the E^f vs E_F. However, total energy differences are generally reliable from PBE, which means that E^f values are often reliable, at least for a first level of screening. Fig. 5(d) shows the

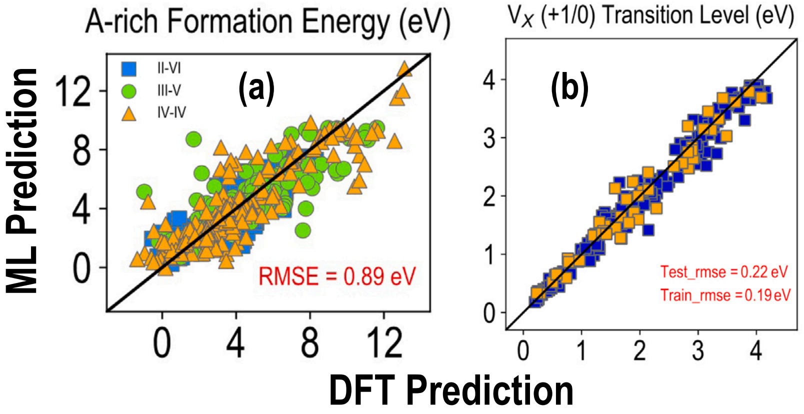

Fig. 9. (a) GPR model for E_{def} of binary compounds denoted as "AB" (where A is the cation and B is the anion) at A-rich chemical potential conditions [92]. (b) NN regression model trained for the +1/0 CTL of X-site vacancies in ABX_3 halide perovskite alloys [23].
Source: Permission to reuse the figures have been obtained from Elsevier and RSC.

PBE-computed CTLs of various native point defects and impurities in ZB binaries [92], plotted against experimentally measured defect levels collected from the literature: it is seen that by artificially extending the CB edge to satisfy the known experimental E_g, the PBE-computed CTLs match remarkably well with experiments, showing an RMSE of only 0.21 eV. This may not be a general rule for all semiconductors, but the impressive correspondence is encouraging. Defect properties may initially be computed from PBE and used for evaluating defect tolerance or potential deep defects, before additional corrections are made as necessary. Some possible corrections include using the modified band alignment approach based on PBE and HSE E_g values [137], shifting both band edge positions using GW quasiparticle energies [138], and performing full high-fidelity HSE06 or GW calculations for possible low energy defects in different charged states.

Fig. 8(b) shows CTLs computed for low energy native defects in the halide perovskite series $MAPbBr_{3-x}Cl_x$ (MA = methylammonium, on organic cation) [129]: it is seen that the +1/0 CTL of the halogen vacancy defect V_{Cl} becomes deeper in the band gap as x increases and the perovskite becomes more Cl-rich. It is thus important to evaluate the same type of defect across related compositions to see how their behavior changes. Defect formation energy plots can reveal a lot of crucial information about the semiconductor, including:

- The lowest energy donor and acceptor type defects under various chemical potential conditions.
- The equilibrium E_F as pinned by the lowest energy defects, based on charge neutrality conditions: this shows the equilibrium conductivity expected in the material (p-type, intrinsic, or n-type).
- The identity of deep and shallow level native defects, and how far from the band edges defect levels may be expected.
- Whether a compound is defect tolerant or not. Defect tolerance can be defined as the absence of any deep defect levels and E^f values not becoming negative for a majority of the band gap.
- The identity of possible impurities that may form and create harmful levels, or possible dopants that may be incorporated in the lattice to induce a p-type or n-type shift in conductivity.
- Suitable growth conditions (such as cation-rich, anion-rich, or intermediate) for stabilizing or de-stabilizing certain defects and tuning the equilibrium conductivity.
- The nature of compensating defects in any compound; i.e., which specific interstitials, vacancies, or substitutions may occur so as to neutralize any low energy defect or impurity. This may be studied by simulating different point defects individually or together in the lattice as defect complexes.

Predicting E^f and CTLs using ML is still in its infancy, owing to the lack of large and diverse defect datasets and the difficulty of correctly describing charge-dependent defect chemistries for prediction. Some recent efforts, including my work, have shown that regression models trained on a DFT dataset using information about the defect coordination environment and elemental properties of defect atoms can yield accurate prediction and screening over tens of thousands of defects and impurities [92,132,139,140]. Regression models trained for neutral-state E^f of native point defects and impurities in ZB binaries [92] and for the +1/0 CTL of X-site vacancies in a dataset of 229 ABX_3 halide perovskite alloys [23] are pictured in Fig. 9**(a)** and **(b)**, respectively. Reasonable RMSE < 10% of the total range of values are seen for both, meaning that many new predictions can be made, but models could also be improved with more data and better descriptors. Using such models, we screened all "dominating" impurities in ZB compounds which will create lower energy defects than native defects and may or may not create mid-gap defect levels [92]: examples include Ti_i as an n-type dopant in ZnS and Ni_{Ga} as a p-type dopant in GaP. Recently, we extended this work to train GNN models for direct E^f prediction using a defective crystal structure [136]. We collected thousands of partially optimized structures from defect calculations to generate a massive dataset of ∼15,000 structures and their corresponding E^f values at E_F = 0 eV (predicting this quantity for different q values helps produce the entire E^f vs E_F plot using lines of slope = q), and trained ALIGNN models to obtain test RMSE of ∼0.3 eV, far superior to the errors around 0.9 eV observed for the composition-based (or neighboring atom identity-based) model shown in Fig. 9(a).

The DFT-GNN approach is very promising for defects, as the same framework allows for the consideration of semiconductor alloys and defect complexes as well, and the prediction of defect energetics by applying a series of distortions on the local defect configuration of any hypothetical defect structure built upon ground state structures [141,142], thus meaningfully sampling all defect polymorphs and reaching the defect ground state without getting trapped in symmetric local minima. ML-predicted E^f values at E_F = 0 eV can further be used to determine all possible CTLs in the band gap. A semiconductor used in electronics or optical absorption must demonstrate intrinsic defect tolerance and/or the ability to be tuned by doping (dopability), which can be conveniently and quickly evaluated using DFT-ML predictions of defect properties. As noted before, these on-demand defect prediction models can also be extended to DFT data from higher levels of theory, either by retraining them with high-fidelity training data or by applying band edge corrections based on semi-local predictions [137,138]

4. Literature containing useful datasets, models, and tools

In the era of Findable, Accessible, Interoperable, Reusable (FAIR) [143] data, it is paramount that major materials informatics studies release their datasets (crystal structures, computed properties, etc.) as well as all code or scripts associated with data analysis, ML model development, and making new ML predictions. FAIR data prevents unnecessary duplication of work and provides researchers with a solid foundation to build their own work on. With this in mind, a number of prominent semiconductor-related computational datasets, ML models, and tools/repositories from the recent literature are highlighted below.

1. W.H. Strehlow et al. 2009 (old but gold) [144]: Experimental band gaps are reported in the paper for a dataset of 723 compounds, along with the method of measurement used.
2. K. Yim et al. 2015 [145]: HSE06 band gaps (on GGA-optimized structures) and dielectric constants using the LDA functional are computed for ~ 1800 binary and ternary oxides selected from ICSD. Selected data are tabulated in the paper, while the rest are available from the authors.
3. M. Lee et al. 2018 [146]: HSE06 band gaps (on GGA-optimized structures) and dielectric constants using the LDA functional are computed for 869 non-oxide inorganic compounds selected from ICSD. Tabulated data are available in the SI.
4. S. Kim et al. 2020 [147]: HSE06 band gaps are computed for 10,481 inorganic materials, showing an RMSE compared to experiments = 0.36 eV. All data available are online on Figshare and all code and tools are on Github.
5. G.S. Na et al. 2020 [148]: GNN models are developed to predict band gaps at different levels of theory, using different published datasets. GW band gaps are predicted for a massive set of 45,835 materials, and all predictions are available on Github.
6. Z. Wan et al. 2021 [149]: A dataset of 150 PBE and experimental band gaps is reported, as well as a neural network model that bridges the gap between PBE and experiment. All data are available in the SI, along with selected HSE06 and GW computations.
7. X. Li et al. 2021 [150]: A multi-fidelity GNN model is trained on a dataset of >800,000 points containing band gaps from multiple levels of theory as well as experimental data, showing a mean absolute error of 0.23 eV. All data are available through the Materials Data Facility (MDF) [151] and the models are on Github.
8. M. Gao et al. 2023 [152]: DFT-ML models are developed to predict the band gaps of over 2000 quaternary semiconductors. All data are available in the SI.
9. Xu et al. 2022 [153]: Multiple regression models are developed to train the band gap of ABO_3 perovskites, with mixing allowed at cation sites. The best models are available as part of a user-friendly online tool.
10. W. Li et al. 2022 [154]: Regression models were trained to bridge the differences between HSE06 and PBEsol band gaps of ABO_3 compounds, by determining necessary shifts that must be applied to either band edge. All data and models are available on Github.
11. G. Petretto et al. 2018 [155]: A DFT dataset of phonon and dielectric properties of 1521 semiconductor materials is presented. All documentation and data can be accessed through Github.
12. I. Petousis et al. 2017 [156]: A large DFPT dataset of the dielectric constant and refractive index of 1056 compounds is presented. All the data are available as part of the Dryad Digital Repository and also incorporated within the Materials Project.
13. A. Talapatra et al. 2021 [157]: ML models are trained on a large dataset containing 1500 experimental points and nearly 3500 DFT calculations, to predict the stability and formability of oxide single and double perovskites. All training data along with hundreds of thousands of novel hypothetical compounds predicted to be stable/formable are made available through the SI.
14. H. Wang et al. 2021 [158]: A chemical similarity-based high-throughput screening approach is applied to discover ~18,000 novel stable crystalline compounds, structures for which are made available through the SI. Thousands of these compounds are predicted to be semiconductors or insulators.
15. D. Broberg et al. 2023 [159]: A dataset of GGA-computed point defect properties (245 data points) in a variety of semiconductors is presented and benchmarked against higher-fidelity hybrid functional computations. All data are available as a contribution to the Materials Project.
16. D. Dahliah et al. 2021 [160]: Defect formation energies and transition levels, carrier lifetimes, and theoretical photovoltaic (PV) efficiencies are presented for dozens of semiconductors, resulting in the screening and discovery of novel Cu-based defect-tolerant solar absorbers. All data are available as part of the SI.

The datasets listed above are simply a subset of DFT-ML studies within the semiconductor design space. The intention with highlighting these datasets is to emphasize that any new project involving screening and design of semiconductors with targeted properties must not start from scratch, but from one or more of the above datasets and models. Even if considering completely novel classes of materials that have not been studied before, existing ML models may provide reasonable initial predictions to determine the potentially most rewarding candidates that must be studied such that the ML models can be updated and made suitable for the new materials. Similarly, existing datasets can be merged with new computations to train newer ML models and obtain better physics-based correlations and understanding of structure–property relationships. A natural question may arise at this point: how reliable is every dataset and model presented in the literature? This is a difficult question to answer, and made especially complicated by the fact that the same materials might often have a range of different property values reported in different studies (e.g., band gaps of the hybrid perovskite, $MAPbI_3$ [37]), with the differences arising either from different experimental/computational setups or unresolved errors. This article is not intended to validate the quality and reliability of any dataset or model, because that is a task best left to individual researchers who may have their own opinions on the relative merits of DFT functionals, structural approximations, ML descriptors, and algorithms being used. Regardless, it is extremely valuable to have open-source peer-reviewed data and code available to serve as a foundation for all ongoing and future research, and the concise list presented here is aimed at streamlining the process of searching for suitable datasets and models to assist with desired semiconductor design objectives.

5. Semiconductor design case studies

This section extends the previous section by covering three major applications of semiconductor materials, namely solar absorption, power electronics, and photocatalysis, through the lens of HT-DFT and ML. A number of examples from the published literature are discussed in terms of the methods applied, properties investigated, and best materials discovered.

5.1. Semiconductor absorbers for solar cells

There are numerous success stories of applying DFT and/or ML to discover new semiconductors that may be used as absorbers in single-junction or tandem solar cells to achieve high power conversion efficiency [11,114]. Suitable materials for solar absorption must have robust bulk stability, resistance to photo and thermal degradation, tolerance to the formation of harmful defects, band gap in the visible range, high PV efficiency, long carrier diffusion lengths and radiative lifetimes, and large carrier mobilities, among many other requirements [23,76]. DFT-ML can be used to calculate many of these metrics and perform HT-screening. Khmaissia et al. [161] trained a variety of

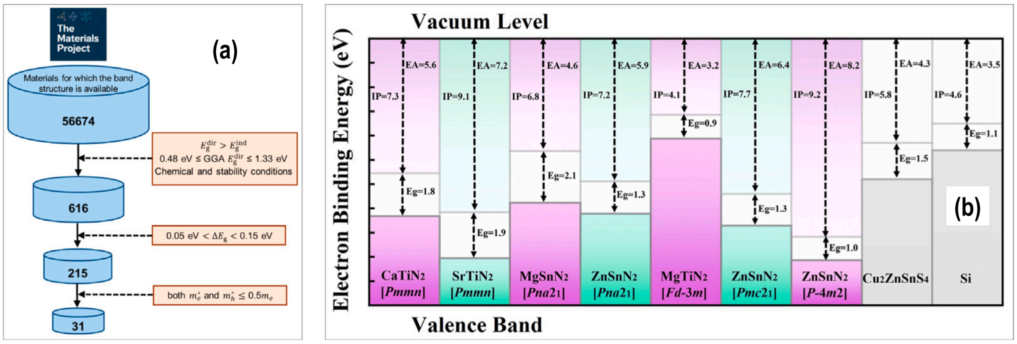

Fig. 10. (a) High-throughput screening approach applied by Kang et al. to discover novel indirect-gap PV materials [162]. (b) Band edges computed for nitrides with promising PV properties, by Xue et al. [33].
Source: Permission to reuse the figures have been obtained from ACS.

linear regression models to predict the E_g of ABC_2 ternary chalcopyrite compounds using a dataset containing both DFT and experimental gaps. The authors found the best subsets of input features that help improve E_g prediction, of which the bond dissociation energies of B-C and C–C bonds are two of the most important. Ultimately, E_g predictions were made for 156 new chalcopyrites, many of which are suitable for solar absorption. The authors have made all the data from this paper available on a google drive link. Kang et al. [162] utilized the Materials Project database [29] to perform screening for indirect-gap materials that exhibit long photocarrier lifetimes and are suitable for PV absorption, using the procedure pictured in Fig. 10(a). They found 31 indirect-gap compounds with small difference in the direct vs indirect E_g, as well as low hole and electronic effective masses, and by applying the constraint of element earth-abundance, discovered three promising compounds: $GeAs_2$, $SiAs_2$, and $NaSbS_2$. The authors further performed accurate optical absorption and defect formation energy calculations to confirm the screened compounds' PV utility and obtain information on how to optimize growth conditions. Tabulated values of computed band gaps, effective masses, defect properties, and PV efficiencies are available in the paper.

Feng et al. [163] used a combination of DFT and ML to discover four promising inorganic PV materials, namely $Ba_4Te_{12}Ge_4$, $Ba_8P_8Ge_4$, $Sr_8P_8Sn_4$, and $Y_4Te_4Se_2$. The dataset used for this purpose contained around 2400 materials obtained from the New Light Harvesting Materials project [34]. By training a variety of regression models for E_g and making new predictions, the authors identified the four compounds and studied them further using ab-initio molecular dynamics (AIMD), band structure, and optical absorption calculations, concluding that the compounds are likely to show SLME greater than 26%. Feng et al. make all data for band gaps, heat of formation, and PV efficiencies available in the form of spreadsheets in the SI. Xue et al. [33] used an HT-DFT framework to discover novel ternary nitrides for PV absorption. Their workflow includes evolutionary algorithm-based crystal structure prediction of several nitrides followed by evaluation of the energy above hull and calculation of E_g and SLME from HSE06, leading to identification of attractive candidates such as $MgTiN_2$ and $ZnSnN_2$, for which additional band structure, optical absorption, and band alignment calculations were performed. Their computed band edge values w.r.t. vacuum level are pictured in Fig. 10(b), for several selected nitrides. All computed properties are available as tabulated data in the SI.

In my research over the last 3 years or so, I studied halide perovskites as well as canonical group IV, III–V, and II–VI semiconductors as possible PV materials. Based on a dataset of 229 ABX_3 halide perovskite alloys, we trained ML models to predict a variety of properties and via predictions made across ∼18,000 hypothetical compositions, identified 392 compounds with bulk stability, E_g in the PV range, defect tolerance based on A-site and X-site vacancy formation, and high absorption efficiencies [23]. This work has been successfully extended to critically examine how factors such as the DFT functional (PBE vs HSE, with or without SOC, with or without full HSE geometry optimization), perovskite phase, and other types of polymorphism influence the computed properties [76,82], and trained all new multi-fidelity regression models on a fusion PBE-HSE-experiment dataset to ultimately identify >2500 stable ABX_3 alloys in 4 different prototype phases with suitable E_g and SLME > 15% [83]. All the raw DFT data from this project are available on MDF [164,165], whereas the ML models and associated code/scripts for analyzing DFT data and making new ML predictions are added to the SI and available on Github [166]. Importantly, this project was used as a suitable case study in several materials informatics workshops organized on behalf of the Materials Research Society [167], as well as for exercises in my modeling and informatics graduate course at Purdue.

5.2. Wide band gap semiconductors for power electronics

Power electronics—the control and conversion of electric power—is of vital importance for energy harvesting, transmission, and storage technologies. While semiconductor power devices primarily use Si, a variety of WBGSs such as SiC, Ga_2O_3, and GaN have been suggested as replacements, owing to their large breakdown fields and high thermal resistance [97,98]. In addition to being cheap to synthesize and process, a suitable semiconductor for power electronics must display a large E_g and dielectric constant, high dielectric breakdown field, high carrier mobilities and concentrations, high thermal conductivity, tolerance to formation of potentially harmful native point defects or impurities, and desirable dopability depending on the application. In their comprehensive review, Kioupakis et al. [168] discuss how different flavors of DFT computations are applied to study the properties of technologically-relevant ultra-WBG compounds such as diamond, cubic BN, β-Ga_2O_3, and AlGaN alloys. The authors describe computations for studying alloy formation energies and ionic ordering, band structures and band edges, thermal conductivities, defect formation energies, carrier mobilities, polarons, dielectric and optical properties, and carrier recombination coefficients. Green et al. [169] and Oshima et al. [170] published detailed reviews of one of the most important UWBG semiconductors, Ga_2O_3, showing how DFT is used to calculate their many electronic and dielectric properties, how it connects to experiments and device-level performance, and guidelines for further improvement. A comparison of the breakdown field and band gap of α-phase Ga_2O_3 with other well-known semiconductors is presented in Fig. 11(a) [170].

Gorai et al. [99] applied a HT-DFT screening procedure to identify several novel inorganic compounds likely to be desirable for power electronics. The authors sampled >10,000 oxides, sulphides, nitrides, carbides, silicides, and borides from the Inorganic Crystal Structure Database (ICSD) [171] and used a series of highly involved DFT computations as well as semi-empirical models to estimate multiple properties

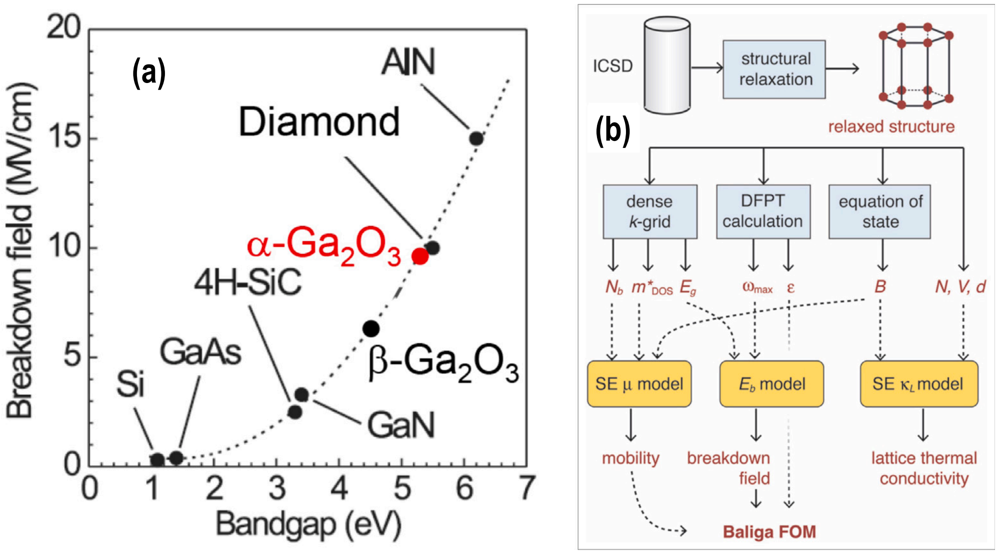

Fig. 11. (a) Breakdown field values of several known semiconductors plotted against their band gap [170]. (b) The computational screening procedure applied by Gorai et al. [99] to discover hundreds of novel WBG semiconductors for power electronics.
Source: Permission to reuse the figures have been obtained from AIP and RSC.

of 863 compounds with finite band gaps. The computational procedure in pictured in Fig. 11(b). Highly accurate electronic structure calculations were performed to determine carrier effective mass and E_g, and a semi-empirical model was used the subsequently estimate the carrier mobility (μ) using the band degeneracy (N_b) and the band effective mass (m^*) from the electronic density of states. DFPT computations [117] yielded the dielectric constant (ϵ) and phonon frequencies (ω), which along with the PBE E_g was used to semi-empirically determine the breakdown field (E_{bd}) based on the phenomenological model developed by Kim et al. [91], which is the same as Eq. (12) discussed in Section 3.2. Finally, the thermal conductivity (κ_L) was also calculated semi-empirically using the bulk modulus (B), number of atoms in the primitive cell (N), volume (V), and density (D). The Baliga figure of merit (BFOM) for application in power devices was calculated using Eq. (15), and applied along with κ_L for screening of suitable candidates. Based on this process, the authors discovered several oxides, carbides, and nitrides with promise for power devices, including known materials SiC, GaN and Ga_2O_3, and related compounds such as ZnO and BN. As part of the SI, Gorai et al. provide a comprehensive tabulated dataset of each of the above properties estimated for 863 inorganic compounds, which may be readily adapted and applied by any researchers.

$$BFOM = \epsilon * \mu * (E_{bd})^3 \tag{15}$$

5.3. Semiconductors for photocatalysis

Semiconductor n-type photoanodes and p-type photocathodes are desired for application in photoelectrochemical cells (PECs), in conjunction with electrolytes and metal electrodes [100,101]. This is generally used for processes such as the photo-driven electrolysis of water to produce H_2 and O_2, photo-oxidation of CN^- by O_2, photo-reduction of CO_2 to useful fuels, and the formation of methane (CH_4) from CH_3COOH. Arunachalam et al. [101] and many other publications report the use of a variety of inorganic oxides as suitable electrodes in PECs, such as ZnO, TiO_2, $SrTiO_3$, $BiVO_4$, and Fe_2O_3. Materials used for photocatalysis must have an ideal semiconductor band structure, high PV absorption and charge carrier concentrations, appropriate VB and CB edges that straddle the redox potentials of the desired chemical process, as well as suitable surface and interface properties. While accurate estimates of E_g and optical absorption are achieved from higher levels of theory including HSE06 and GW, calculating band edges is a far more complicated prospect that typically entails bulk, surface, and heterostructure calculations in order to align mean electrostatic potentials and subsequently the band edges with respect to the vacuum level. Simulating surface slab structures of superlattices is quite expensive due to the large number of atoms, and certainly very prohibitive when using advanced functionals. These calculations become even more expensive if band edges need to be computed in aqueous environments [172]. Many empirical methods are thus used for approximating the band edges and aligning them w.r.t. vacuum and other semiconductors, with the most common approach involving the use of Mulliken electronegativity values of cation and anion species along with the computed E_g and normal hydrogen electrode potential [173].

Jin et al. [175] performed multi-objective HT-screening over nearly 84,000 compounds in the Materials Project [29] to discover 22 materials that show promise for photocatalytic water splitting. By applying the conditions that the crystal structure must be ordered and present in the ICSD, and the computed energy above hull must be low enough, the authors reduced this large number to around 5000. Using the MP E_g values and empirically determined band edges, this number further came down to 22 compounds that have E_g expected to be between 1.6 eV and 3.0 eV, CB edge below 0 eV, and VB edge above 1.23 eV, as desired for water splitting. Finally, electronic band structure calculations were performed using HSE06 to accurately determine E_g and band edges, and optical absorption calculations were performed, revealing ZnSe, Ga_2Se_3, and $Na_2Zn_2O_3$ as the most promising candidates. Tabulated data are made available within the main text of the paper. Similarly, Antoniuk et al. [174] performed two types of screening across 75,000 compounds and applied criteria of low intrinsic emittance score, desired stability and synthesizability, robust work function, commercial availability and finally, low intrinsic emittance from DFT, leading to 11 candidates for low intrinsic emittance materials, and additional conditions of visible light emission and air stability to find strong candidates that may be used as photocathodes. These compounds include Na_2O, K_2O, and Rb_2O, for which the authors performed HSE06 computations to accurately determine important electronic properties. Their screening procedure is pictured in Fig. 12(a), and the authors made all tabulated data available in the main text and the SI of the paper. Finally, in my past collaborative work [173,176], we trained DFT-ML models for accurate E_g prediction for $AA'BB'O_6$ double perovskite oxides, at the GLLB-SC level of theory [177] which has a better accuracy than semi-local functionals, and

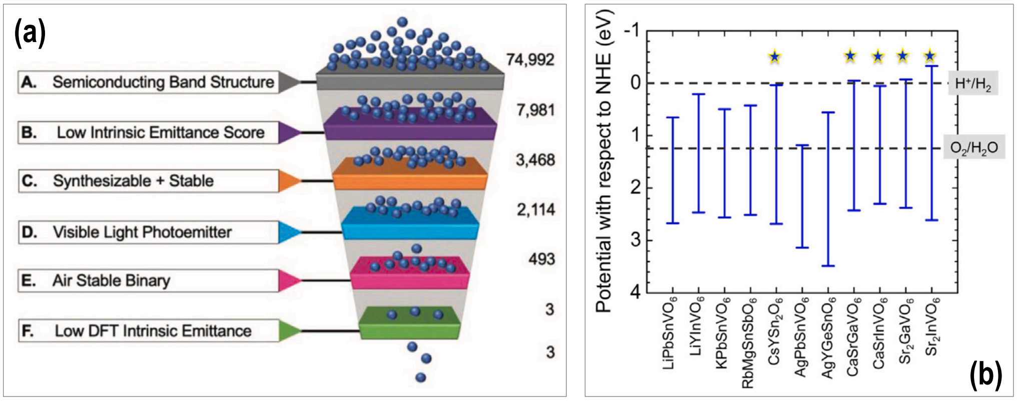

Fig. 12. (a) Screening procedure applied by Antoniuk et al. [174] to discover promising low emittance compounds. (b) VB and CB edges of various oxide double perovskites, aligned against H_2 and O_2 redox potentials [173].
Source: Permission to reuse the figures have been obtained from John Wiley and Sons and Springer Nature.

identified several compounds with suitable E_g and empirical VB and CB edges that straddle the H^+/H_2 and O_2/H_2O levels for water splitting, as shown in Fig. 12(b). We performed in-depth crystal structure prediction and electronic structure calculations to confirm the promise of four compounds, namely $CaSrGaVO_6$, $CaSrInVO_6$, Sr_2GaVO_6, and Sr_2InVO_6. Naturally, all DFT data are made available as part of the SI.

Summary and future outlook

This article presents an overview of the many technologically-relevant properties of semiconductors that can be computed using a combination of high-throughput density functional theory (DFT) computations and machine learning (ML), leading to screening and identification of suitable candidates for a variety of applications including solar absorption, power electronics, and photocatalysis. “DFT-ML” is now a routine approach applied towards virtual materials design to guide experimental discovery. The fundamental atomistic insights and ability to perform composition- and structure-engineering using DFT-ML is vital to fulfilling the purpose of the CHIPS Act as well as accomplishing the updated goals of the MGI. When it comes to designing semiconductors for devices with improved performance, DFT and ML are invaluable for performing crystal structure prediction, estimating bulk stability, calculating band gap and band edges, determining optical absorption efficiency, evaluating defect tolerance and dopability using defect formation energies, and studying surfaces and interfaces, among other things. This article covers some important examples from the literature of how existing databases as well as new property predictions are used to create lists of promising semiconductors that are worthy of experimental investigation. The computational recipes, specific datasets and models, and case studies presented here are intended to serve as a guide for researchers as they embark on their own semiconductor design efforts, hopefully eliminating guesswork or duplication of research, helping further accelerate materials discovery.

The key takeaways from this review are as follows:

- Discovering novel semiconductor compositions for a variety of applications is an exercise in performing multi-objective design over massive chemical spaces. This task is made easier by the use of DFT-ML, wherein large datasets of first principles-based properties lead to predictive models that enable high-throughput screening of promising new candidates. Such models can be trained in a “multi-fidelity” manner by combining several levels of theory and experiments, to ensure eventual quality and reliability of screened compounds.
- DFT is invaluable in evaluating the stability of materials in multiple possible “phases”, which can be defined as the possible structural configurations that may be adopted by the given set of atoms within a crystalline framework. Examples include the cubic, tetragonal, or orthorhombic phases of 3D ABX_3 perovskites. Conveniently, most materials can be considered in “prototype” phases given their composition and types of cations/anions in the system; such prototype structures can be obtained from available databases such as Materials Project [29] and used as starting configurations for novel materials. By investigating several such material phases as well as other competing polymorphs arising from distortions or strains, DFT can comprehensively reveal formation energies of crystals and the “convex hull” [73,178] for a given set of materials, which can be used to understand the likelihood that any semiconductor crystal may decompose to alternative phases.
- Navigating through a sea of DFT-ML papers on semiconductor properties, this article presents a few important methods that may be used to determine bulk stability, formability, electronic band gaps, carrier mobilities, photovoltaic efficiencies, dielectric constants and breakdown fields, surface and interface energies, and defect formation energies, for any new materials of interest. The DFT recipes and/or ML models discussed in this work enable such predictions, with some caveats and limitations from the levels of theory or types of materials being studied.
- This article lists several important computational datasets and ML models from the literature that are pertinent to semiconductors for electronic and energy-relevant applications. Many such contributions are often overlooked which leads to unnecessary new calculations; by providing a list of useful datasets and brief descriptions of the associated properties, models, and data-hosting locations, I hope to encourage researchers to extensively utilize these past works to guide their ongoing work, in terms of the specific chemistries that must be studied, how to calculate and where to find properties of interest, and how best to replicate DFT-ML for their own research problems.
- Finally, an in-depth look at some examples in the literature sheds light on the procedures used for computer-aided design of new solar cell absorbers, photocatalysts, and WBG compounds for power electronics—three important applications where tailoring the properties of semiconductors is paramount. These studies reveal specific methods that best enable multi-objective design across wide chemical spaces, such as empirical calculations of breakdown fields and thermal conductivity, the use of random forests or neural networks for training regression models, and useful property targets and ranges that account for possible uncertainties in prediction while screening.

For continued discovery of next-generation semiconductor materials, it is crucial that materials scientists be trained in data science and ML, in addition to obtaining an appreciation and understanding for first principles simulations. This can be achieved by incorporating materials informatics in graduate and undergraduate education, as well as conducting regular hands-on workshops [167]. The importance of generating large amounts of high-quality data cannot be emphasized enough. There must also be an increased focus on making all compiled data, simulation workflows, and tools for training models and predicting properties, openly available in a FAIR [143] format. With open-science policies released by the government and the emergence of journals and conferences that strongly adhere to FAIR principles, the future of materials data is in good hands. In the coming years, DFT-ML in synergy with accelerated experiments will be invaluable in leading semiconductor discovery and reducing the time and costs involved in technological advancements.

CRediT authorship contribution statement

Arun Mannodi-Kanakkithodi: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

A.M.K. acknowledges support from the School of Materials Engineering at Purdue University under account number F.10023800.05.002.

References

[1] FACT SHEET: CHIPS and science act will lower costs, create jobs, strengthen supply chains, and counter China, 2022, The White House.

[2] SEMICONDUCTORS @ PURDUE: Educating the next generation of workforce leaders in semiconductors and microelectronics, 2023, Purdue University. Updated.

[3] Semiconductor market revenue to reach US$ 772 billion by 2030, 2021, Market Stats News.

[4] Materials genome initiative strategic plan: A report by the subcommittee on the materials genome initiative committee on technology of the national science and technology council, 2021, National Science and Technology Council.

[5] Chen Ling, A review of the recent progress in battery informatics, npj Comput. Mater. 8 (2022) 33.

[6] Jorge Felipe Gaviria, Gabriel Narváez, Camilo Guillen, Luis Felipe Giraldo, Michael Bressan, Machine learning in photovoltaic systems: A review, Renew. Energy 196 (2022) 298–318.

[7] Arun Mannodi-Kanakkithodi, Anand Chandrasekaran, Chiho Kim, Tran Doan Huan, Ghanshyam Pilania, Venkatesh Botu, Rampi Ramprasad, Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond, Mater. Today 21 (7) (2018) 785–796.

[8] Xiangdong Wang, Ye Sheng, Jinyan Ning, Jinyang Xi, Lili Xi, Di Qiu, Jiong Yang, Xuezhi Ke, A critical review of machine learning techniques on thermoelectric materials, J. Phys. Chem. Lett. 14 (7) (2023) 1808–1822.

[9] Jonathan Schmidt, Mário R.G. Marques, Silvana Botti, Miguel A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, npj Comput. Mater. 5 (1) (2019) 83.

[10] Deepak Jain, Suryanaman Chaube, Prerna Khullar, Sriram Goverapet Srinivasan, Beena Rai, Bulk and surface DFT investigations of inorganic halide perovskites screened using machine learning and materials property databases, Phys. Chem. Chem. Phys. 21 (2019) 19423–19436.

[11] Jiaqi Yang, Arun Mannodi-Kanakkithodi, High-throughput computations and machine learning for halide perovskite discovery, MRS Bull. (2022).

[12] Giulia Galli, The long and winding road: Predicting materials properties through theory and computation, in: Wanda Andreoni, Sidney Yip (Eds.), Handbook of Materials Modeling: Methods: Theory and Modeling, Springer International Publishing, Cham, 2020, pp. 37–48.

[13] Arun Mannodi-Kanakkithodi, Maria K.Y. Chan, Computational data-driven materials discovery, Trends Chem. 3 (2) (2021) 79–82, Special Issue: Machine Learning for Molecules and Materials.

[14] Jared C. Stanley, Felix Mayr, Alessio Gagliardi, Machine learning stability and bandgaps of lead-free perovskites for photovoltaics, Adv. Theory Simul. 3 (1) (2020) 1900178.

[15] Diego Milardovich, Markus Jech, Dominic Waldhoer, Michael Waltl, Tibor Grasser, Machine learning prediction of defect formation energies in a-SiO_2, in: 2020 International Conference on Simulation of Semiconductor Processes and Devices, SISPAD, 2020, pp. 339–342.

[16] Manala Tabu Mbumba, Davy Maurice Malouangou, Jadel Matondo Tsiba, Luyun Bai, Yifan Yang, Mina Guli, Degradation mechanism and addressing techniques of thermal instability in halide perovskite solar cells, Sol. Energy 230 (2021) 954–978.

[17] Juanita Hidalgo, Waldemar Kaiser, Yu An, Ruipeng Li, Zion Oh, Andrés-Felipe Castro-Méndez, Diana K. LaFollette, Sanggyun Kim, Barry Lai, Joachim Breternitz, Susan Schorr, Carlo A.R. Perini, Edoardo Mosconi, Filippo De Angelis, Juan-Pablo Correa-Baena, Synergistic role of water and oxygen leads to degradation in formamidinium-based halide perovskites, J. Am. Chem. Soc. 145 (45) (2023) 24549–24557.

[18] Ruiming Zhu, Siyu Isaac Parker Tian, Zekun Ren, Jiali Li, Tonio Buonassisi, Kedar Hippalgaonkar, Predicting synthesizability using machine learning on databases of existing inorganic materials, ACS Omega 8 (9) (2023) 8210–8218.

[19] Geun Ho Gu, Jidon Jang, Juhwan Noh, Aron Walsh, Yousung Jung, Perovskite synthesizability using graph neural networks, npj Comput. Mater. 8 (1) (2022) 71.

[20] A. Gilad Kusne, Austin McDannald, Scalable multi-agent lab framework for lab optimization, Matter 6 (6) (2023) 1880–1893.

[21] Rishi E. Kumar, Armi Tiihonen, Shijing Sun, David P. Fenning, Zhe Liu, Tonio Buonassisi, Opportunities for machine learning to accelerate halide-perovskite commercialization and scale-up, Matter 5 (5) (2022) 1353–1366.

[22] Alex Zunger, Inverse design in search of materials with target functionalities, Nat. Rev. Chem. 2 (4) (2018) 0121.

[23] Arun Mannodi-Kanakkithodi, Maria K.Y. Chan, Data-driven design of novel halide perovskite alloys, Energy Environ. Sci. 15 (2022) 1930–1949.

[24] Structural properties, in: Properties of Group-IV, III-V and II-VI Semiconductors, John Wiley & Sons, Ltd, 2005, pp. 1–21.

[25] Mirjana Dimitrievska, Federica Boero, Alexander P. Litvinchuk, Simona Delsante, Gabriella Borzone, Alejandro Perez-Rodriguez, Victor Izquierdo-Roca, Structural polymorphism in "kesterite" Cu2ZnSnS4: Raman spectroscopy and first-principles calculations analysis, Inorg. Chem. 56 (6) (2017) 3467–3474.

[26] Gopalakrishnan Sai Gautam, Thomas P. Senftle, Emily A. Carter, Understanding the effects of Cd and Ag doping in Cu2ZnSnS4 solar cells, Chem. Mater. 30 (14) (2018) 4543–4555.

[27] Joseph S. Manser, Jeffrey A. Christians, Prashant V. Kamat, Intriguing optoelectronic properties of metal halide perovskites, Chem. Rev. 116 (21) (2016) 12956–13008.

[28] Tao Zhang, Zenghua Cai, Shiyou Chen, Chemical trends in the thermodynamic stability and band gaps of 980 halide double perovskites: A high-throughput first-principles study, ACS Appl. Mater. Interfaces 12 (18) (2020) 20680–20690.

[29] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, Kristin A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013) 011002.

[30] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD), JOM 65 (11) (2013) 1501–1509.

[31] K. Mathew, J.H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. Tang, I. Chu, T. Smidt, B. Bocklund, M. Horton, J. Dagdelen, B. Wood, Z.-K. Liu, J. Neaton, S.P. Ong, K. Persson, A. Jain, Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows, Comput. Mater. Sci. 139 (2017) 140–152.

[32] A. Goyal, P. Gorai, H. Peng, S. Lany, V. Stevanović, A computational framework for automation of point defect calculations, Comput. Mater. Sci. 130 (2017) 1–9.

[33] Pengyan Xue, Dongdong Chu, Congwei Xie, Evgenii Tikhonov, Keith T. Butler, Design of new ternary nitrides for photovoltaic applications via high-throughput calculations, J. Phys. Chem. C 126 (40) (2022) 17398–17405.

[34] Ivano E. Castelli, Falco Hüser, Mohnish Pandey, Hong Li, Kristian S. Thygesen, Brian Seger, Anubhav Jain, Kristin A. Persson, Gerbrand Ceder, Karsten W. Jacobsen, New light-harvesting materials using accurate and efficient bandgap calculations, Adv. Energy Mater. 5 (2) (2015) 1400915.

[35] Milica Todorović, Michael U. Gutmann, Jukka Corander, Patrick Rinke, Bayesian inference of atomistic structure in functional materials, npj Comput. Mater. 5 (1) (2019) 35.

[36] Yashaswi Pathak, Karandeep Singh Juneja, Girish Varma, Masahiro Ehara, U. Deva Priyakumar, Deep learning enabled inorganic material generator, Phys. Chem. Chem. Phys. 22 (2020) 26935–26943.

[37] Osbel Almora, Derya Baran, Guillermo C. Bazan, Christian Berger, Carlos I. Cabrera, Kylie R. Catchpole, Sule Erten-Ela, Fei Guo, Jens Hauch, Anita W.Y. Ho-Baillie, T. Jesper Jacobsson, Rene A.J. Janssen, Thomas Kirchartz, Nikos Kopidakis, Yongfang Li, Maria A. Loi, Richard R. Lunt, Xavier Mathew, Michael D. McGehee, Jie Min, David B. Mitzi, Mohammad K. Nazeeruddin, Jenny Nelson, Ana F. Nogueira, Ulrich W. Paetzold, Nam-Gyu Park, Barry P. Rand, Uwe Rau, Henry J. Snaith, Eva Unger, Lídice Vaillant-Roca, Hin-Lap Yip, Christoph J. Brabec, Device performance of emerging photovoltaic materials (Version 1), Adv. Energy Mater. 11 (11) (2021) 2002774.

[38] Y. Mishin, Machine-learning interatomic potentials for materials science, Acta Mater. 214 (2021) 116980.

[39] Dylan M. Anstine, Olexandr Isayev, Machine learning interatomic potentials and long-range physics, J. Phys. Chem. A 127 (11) (2023) 2417–2431.

[40] Anand Chandrasekaran, Deepak Kamal, Rohit Batra, Chiho Kim, Lihua Chen, Rampi Ramprasad, Solving the electronic structure problem with machine learning, npj Comput. Mater. 5 (1) (2019) 22.

[41] Heesoo Park, Raghvendra Mall, Fahhad H. Alharbi, Stefano Sanvito, Nouar Tabet, Halima Bensmail, Fedwa El-Mellouhi, Exploring new approaches towards the formability of mixed-ion perovskites by dft and machine learning, Phys. Chem. Chem. Phys. 21 (3) (2019) 1078–1088.

[42] Deepak Jain, Suryanaman Chaube, Prerna Khullar, Sriram Goverapet Srinivasan, Beena Rai, Bulk and surface DFT investigations of inorganic halide perovskites screened using machine learning and materials property databases, Phys. Chem. Chem. Phys. 21 (2019) 19423–19436.

[43] Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, Adam S. Foster, DScribe: Library of descriptors for machine learning in materials science, Comput. Phys. Comm. 247 (2020) 106949.

[44] Atsuto Seko, Atsushi Togo, Isao Tanaka, Descriptors for machine learning of materials data, in: Isao Tanaka (Ed.), Nanoinformatics, Springer Singapore, Singapore, 2018, pp. 3–23.

[45] Yashaswi Pathak, Karandeep Singh Juneja, Girish Varma, Masahiro Ehara, U. Deva Priyakumar, Deep learning enabled inorganic material generator, Phys. Chem. Chem. Phys. 22 (2020) 26935–26943.

[46] Hitarth Choubisa, Mikhail Askerka, Kevin Ryczko, Oleksandr Voznyy, Kyle Mills, Isaac Tamblyn, Edward H. Sargent, Crystal site feature embedding enables exploration of large chemical spaces, Matter 3 (2) (2020) 433–448.

[47] Juhwan Noh, Jaehoon Kim, Helge S. Stein, Benjamin Sanchez-Lengeling, John M. Gregoire, Alan Aspuru-Guzik, Yousung Jung, Inverse design of solid-state materials via a continuous representation, Matter 1 (5) (2019) 1370–1384.

[48] Paul C. Jennings, Steen Lysgaard, Jens Strabo Hummelshoj, Tejs Vegge, Thomas Bligaard, Genetic algorithms for computational materials discovery accelerated by machine learning, Nat. Comput. Sci. 5 (46) (2019).

[49] Jia Wang, Yingxue Wang, Yanan Chen, Inverse design of materials by machine learning, Materials 15 (5) (2022).

[50] John P. Perdew, Kieron Burke, Matthias Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865–3868.

[51] Jochen Heyd, Gustavo E. Scuseria, Matthias Ernzerhof, Hybrid functionals based on a screened Coulomb potential, J. Chem. Phys. 118 (18) (2003) 8207–8215.

[52] Cecilia Vona, Dmitrii Nabok, Claudia Draxl, Electronic structure of (organic-)inorganic metal halide perovskites: The dilemma of choosing the right functional, Adv. Theory Simul. 5 (1) (2022) 2100496.

[53] Olaf Delgado-Friedrichs, Michael O'Keeffe, Crystal nets as graphs: Terminology and definitions, J. Solid State Chem. 178 (8) (2005) 2480–2485.

[54] Anthony Yu-Tung Wang, Ryan J. Murdock, Steven K. Kauwe, Anton O. Oliynyk, Aleksander Gurlo, Jakoah Brgoch, Kristin A. Persson, Taylor D. Sparks, Machine learning for materials scientists: An introductory guide toward best practices, Chem. Mater. 32 (12) (2020) 4954–4965.

[55] K. Choudhary, B. DeCost, C. Chen, et al., Recent advances and applications of deep learning methods in materials science, npj Comput. Mater. 8 (59) (2022).

[56] K.T. Schütt, P.-J. Kindermans, H.E. Sauceda, S. Chmiela, A. Tkatchenko, K.-R. Müller, SchNet: A continuous-filter convolutional neural network for modeling quantum interactions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17, Curran Associates Inc., 2017, pp. 992–1002.

[57] O.T. Unke, S. Chmiela, M. Gastegger, et al., SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects, Nature Commun. 12 (7273) (2021).

[58] Yashaswi Pathak, Karandeep Singh Juneja, Girish Varma, Masahiro Ehara, U. Deva Priyakumar, Deep learning enabled inorganic material generator, Phys. Chem. Chem. Phys. 22 (2020) 26935–26943.

[59] Addis S. Fuhr, Bobby G. Sumpter, Deep generative models for materials discovery and machine learning-accelerated innovation, Front. Mater. 9 (2022).

[60] Surya R. Kalidindi, Feature engineering of material structure for AI-based materials knowledge systems, J. Appl. Phys. 128 (4) (2020) 041103.

[61] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, Chiho Kim, Machine learning in materials informatics: recent applications and prospects, npj Comput. Mater. 3 (2017) 54.

[62] Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, Luca M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Phys. Rev. Mater. 2 (2018) 083802.

[63] Félix Therrien, Eric B. Jones, Vladan Stevanović, Metastable materials discovery in the age of large-scale computation, Appl. Phys. Rev. 8 (3) (2021) 031310.

[64] Shubham Pandey, Jiaxing Qu, Vladan Stevanović, Peter St. John, Prashun Gorai, Predicting energy and stability of known and hypothetical crystals using graph neural network, Patterns 2 (11) (2021) 100361.

[65] Jeffrey N. Law, Shubham Pandey, Prashun Gorai, Peter C. St. John, Upper-bound energy minimization to search for stable functional materials with graph neural networks, JACS 3 (1) (2023) 113–123.

[66] Andriy O. Lyakhov, Artem R. Oganov, Harold T. Stokes, Qiang Zhu, New developments in evolutionary structure prediction algorithm USPEX, Comput. Phys. Comm. 184 (4) (2013) 1172–1182.

[67] Stefan Goedecker, Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems, J. Chem. Phys. 120 (21) (2004) 9911–9917.

[68] Xin-Gang Zhao, Gustavo M. Dalpian, Zhi Wang, Alex Zunger, Polymorphous nature of cubic halide perovskites, Phys. Rev. B 101 (2020) 155137.

[69] Lai Wei, Nihang Fu, Edirisuriya M.D. Siriwardane, Wenhui Yang, Sadman Sadeed Omee, Rongzhi Dong, Rui Xin, Jianjun Hu, TCSP: a template-based crystal structure prediction algorithm for materials discovery, Inorg. Chem. 61 (22) (2022) 8431–8439.

[70] Tomoki Yamashita, Hiori Kino amd Koji Tsuda, Takashi Miyake, Tamio Oguchi, Hybrid algorithm of Bayesian optimization and evolutionary algorithm in crystal structure prediction, Sci. Technol. Adv. Mater.: Methods 2 (1) (2022) 67–74.

[71] B.C. Revard, W.W. Tipton, R.G. Hennig, Genetic algorithm for structure and phase prediction, 2018, GitHub repository. https://github.com/henniggroup/GASP-python.

[72] Dirk Gillespie, Computing the partition function, ensemble averages, and density of states for lattice spin systems by sampling the mean, J. Comput. Phys. 250 (2013) 1–12.

[73] Christopher R. Weinberger, Xiao-Xiang Yu, Hang Yu, Gregory B. Thompson, Ab initio investigations of the phase stability in group IVB and VB transition metal nitrides, Comput. Mater. Sci. 138 (2017) 333–345.

[74] Yi-Yang Sun, Michael L. Agiorgousis, Peihong Zhang, Shengbai Zhang, Chalcogenide perovskites for photovoltaics, Nano Lett. 15 (1) (2015) 581–585.

[75] Arun Mannodi-Kanakkithodi, A first principles investigation of ternary and quaternary II–VI zincblende semiconductor alloys, Modelling Simul. Mater. Sci. Eng. 30 (4) (2022) 044001.

[76] Jiaqi Yang, Panayotis Manganaris, Arun Mannodi-Kanakkithodi, A high-throughput computational dataset of halide perovskite alloys, Digit. Discov. 2 (2023) 856–870.

[77] J. Bartel Christopher, Christopher Sutton, R. Goldsmith Bryan, Runhai Ouyang, B. Musgrave Charles, M. Ghiringhelli Luca, Matthias Scheffler, New tolerance factor to predict the stability of perovskite oxides and halides, Sci. Adv. 5 (2) (2019) eaav0693.

[78] Gábor I. Csonka, John P. Perdew, Adrienn Ruzsinszky, Pier H.T. Philipsen, Sébastien Lebègue, Joachim Paier, Oleg A. Vydrov, János G. Ángyán, Assessing the performance of recent density functionals for bulk solids, Phys. Rev. B 79 (2009) 155107.

[79] Stefan Grimme, Jens Antony, Stephan Ehrlich, Helge Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, J. Chem. Phys. 132 (15) (2010) 154104.

[80] Zhijun Jiang, Yousra Nahas, Bin Xu, Sergey Prosandeev, Dawei Wang, Laurent Bellaiche, Special quasirandom structures for perovskite solid solutions, J. Phys.: Condens. Matter. 28 (47) (2016) 475901.

[81] Qingde Sun, Wan-Jian Yin, Thermodynamic stability trend of cubic perovskites, J. Am. Chem. Soc. 139 (42) (2017) 14905–14908.

[82] J. Yang, A. Mannodi-Kanakkithodi, First principles investigation of polymorphism in halide perovskites, 2023 (submitted for publication), preprint: https://arxiv.org/abs/2309.16095.

[83] Jiaqi Yang, Panayotis Manganaris, Arun Mannodi-Kanakkithodi, Discovering novel halide perovskite alloys using multi-fidelity machine learning and genetic algorithm, J. Chem. Phys. 160 (6) (2024) 064114.

[84] Christopher J. Bartel, Amalie Trewartha, Qi Wang, Alexander Dunn, Anubhav Jain, Gerbrand Ceder, A critical examination of compound stability predictions from machine-learned formation energies, npj Comput. Mater. 6 (2020) 97.

[85] Gordon G.C. Peterson, Jakoah Brgoch, Materials discovery through machine learning formation energy, J. Phys.: Energy 3 (2) (2021) 022002.

[86] Tian Xie, Jeffrey C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (2018) 145301.

[87] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, Shyue Ping Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chem. Mater. 31 (9) (2019) 3564–3572.

[88] Kamal Choudhary, Brian DeCost, Atomistic line graph neural network for improved materials property predictions, npj Comput. Mater. 7 (185) (2021).

[89] M. Biswas, G. Bidna, M.H. Rahman, R. Edlabadkar, R. Desai, A. Mannodi-Kanakkithodi, Discovering novel halide perovskites using graph-based neural networks, 2024, (submitted for publication).

[90] Sangqi Xiong, Xin Li, Xiangwei Wu, Jin Yu, Oleg I. Gorbatov, Igor Di Marco, Paul R.C. Kent, Weiwei Sun, A combined machine learning and density functional theory study of binary Ti-Nb and Ti-Zr alloys: Stability and Young's modulus, Comput. Mater. Sci. 184 (2020) 109830.

[91] Chiho Kim, Ghanshyam Pilania, Ramamurthy Ramprasad, From organized high-throughput data to phenomenological theory using machine learning: The example of dielectric breakdown, Chem. Mater. 28 (5) (2016) 1304–1311.

[92] Arun Mannodi-Kanakkithodi, Xiaofeng Xiang, Laura Jacoby, Robert Biegaj, Scott T. Dunham, Daniel R. Gamelin, Maria K.Y. Chan, Universal machine learning framework for defect predictions in zinc blende semiconductors, Patterns 3 (3) (2022) 100450.

[93] Hong-Jhang Syu, Yu-Chieh Huang, Zih-Chun Su, Ruei-Lien Sun, Ching-Fuh Lin, An alternative to compound semiconductors using a Si-Based IR detector, IEEE Trans. Electron Devices 69 (1) (2022) 205–211.

[94] Hanul Min, Do Yoon Lee, Junu Kim, Gwisu Kim, Kyoung Su Lee, Jongbeom Kim, Min Jae Paik, Young Ki Kim, Kwang S. Kim, Min Gyu Kim, Tae Joo Shin, Sang Il Seok, Perovskite solar cells with atomically coherent interlayers on SnO2 electrodes, Nature 598 (7881) (2021) 444–450.

[95] Hui Li, Wei Zhang, Perovskite tandem solar cells: From fundamentals to commercial deployment, Chem. Rev. 120 (18) (2020) 9835–9950.

[96] R. Guerrero-Lemus, R. Vega, Taehyeon Kim, Amy Kimm, L.E. Shephard, Bifacial solar photovoltaics – A technology review, Renew. Sustain. Energy Rev. 60 (2016) 1533–1549.

[97] H. Amano, Y. Baines, E. Beam, Matteo Borga, T. Bouchet, Paul R. Chalker, M. Charles, Kevin J. Chen, Nadim Chowdhury, Rongming Chu, Carlo De Santi, Maria Merlyne De Souza, Stefaan Decoutere, L. Di Cioccio, Bernd Eckardt, Takashi Egawa, P. Fay, Joseph J. Freedsman, L. Guido, Oliver Häberlen, Geoff Haynes, Thomas Heckel, Dilini Hemakumara, Peter Houston, Jie Hu, Mengyuan Hua, Qingyun Huang, Alex Huang, Sheng Jiang, H. Kawai, Dan Kinzer, Martin Kuball, Ashwani Kumar, Kean Boon Lee, Xu Li, Denis Marcon, Martin März, R. McCarthy, Gaudenzio Meneghesso, Matteo Meneghini, E. Morvan, A. Nakajima, E.M.S. Narayanan, Stephen Oliver, Tomás Palacios, Daniel Piedra, M. Plissonnier, R. Reddy, Min Sun, Iain Thayne, A. Torres, Nicola Trivellin, V. Unni, Michael J. Uren, Marleen Van Hove, David J. Wallis, J. Wang, J. Xie, S. Yagi, Shu Yang, C. Youtsey, Ruiyang Yu, Enrico Zanoni, Stefan Zeltner, Yuhao Zhang, The 2018 GaN power electronics roadmap, J. Phys. D: Appl. Phys. 51 (16) (2018) 163001.

[98] K. Shenai, R.S. Scott, B.J. Baliga, Optimum semiconductors for high-power electronics, IEEE Trans. Electron Devices 36 (9) (1989) 1811–1823.

[99] Prashun Gorai, Robert W. McKinney, Nancy M. Haegel, Andriy Zakutayev, Vladan Stevanovic, A computational survey of semiconductors for power electronics, Energy Environ. Sci. 12 (2019) 3338–3347.

[100] N. Serpone, A.V. Emeline, Semiconductor photocatalysis — Past, present, and future outlook, J. Phys. Chem. Lett. 3 (5) (2012) 673–677.

[101] Prabhakarn Arunachalam, Abdullah M. Al Mayouf, Chapter 28 - Photoelectrochemical water splitting, in: Satyabrata Mohapatra, Tuan Anh Nguyen, Phuong Nguyen-Tri (Eds.), Noble Metal-Metal Oxide Hybrid Nanoparticles, in: Micro and Nano Technologies, Woodhead Publishing, 2019, pp. 585–606.

[102] Georg K.H. Madsen, David J. Singh, BoltzTraP. a code for calculating band-structure dependent quantities, Comput. Phys. Comm. 175 (1) (2006) 67–71.

[103] M.K.Y. Chan, G. Ceder, Efficient band gap prediction for solids, Phys. Rev. Lett. 105 (2010) 196403.

[104] Arun Mannodi-Kanakkithodi, Ji-Sang Park, Nari Jeon, Duyen H. Cao, David J. Gosztola, Alex B.F. Martinson, Maria K.Y. Chan, Comprehensive computational study of partial lead substitution in methylammonium lead bromide, Chem. Mater. 31 (10) (2019) 3599–3612.

[105] Jie Pan, Wyatt K. Metzger, Stephan Lany, Spin-orbit coupling effects on predicting defect properties with hybrid functionals: A case study in cdte, Phys. Rev. B 98 (2018) 054108.

[106] Ya Zhuo, Aria Mansouri Tehrani, Jakoah Brgoch, Predicting the band gaps of inorganic solids by machine learning, J. Phys. Chem. Lett. 9 (7) (2018) 1668–1673.

[107] Chan Gao, Xiaoyong Yang, Ming Jiang, Lixin Chen, Zhiwen Chen, Chandra Veer Singh, Machine learning-enabled band gap prediction of monolayer transition metal chalcogenide alloys, Phys. Chem. Chem. Phys. 24 (2022) 4653–4665.

[108] Tilak Das, Giovanni Di Liberto, Gianfranco Pacchioni, Density functional theory estimate of halide perovskite band gap: When spin orbit coupling helps, J. Phys. Chem. C 126 (4) (2022) 2184–2198.

[109] G. Pilania, J.E. Gubernatis, T. Lookman, Multi-fidelity machine learning models for accurate bandgap predictions of solids, Comput. Mater. Sci. 129 (2017) 156–163.

[110] Zhun Liu, Guangren Na, Fuyu Tian, Liping Yu, Jingbo Li, Lijun Zhang, Computational functionality-driven design of semiconductors for optoelectronic applications, InfoMat 2 (5) (2020) 879–904.

[111] Fumiyasu Oba, Yu Kumagai, Design and exploration of semiconductors from first principles: A review of recent advances, Appl. Phys. Express 11 (6) (2018) 060101.

[112] M. Gajdoš, K. Hummer, G. Kresse, J. Furthmüller, F. Bechstedt, Linear optical properties in the projector-augmented wave methodology, Phys. Rev. B 73 (2006) 045112.

[113] C.A. Gueymard, D. Myers, K. Emery, Proposed reference irradiance spectra for solar energy systems testing, Sol. Energy 73 (6) (2002) 443–467.

[114] Marnik Bercx, Nasrin Sarmadian, Rolando Saniz, Bart Partoens, Dirk Lamoen, First-principles analysis of the spectroscopic limited maximum efficiency of photovoltaic absorber layers for CuAu-like chalcogenides and silicon, Phys. Chem. Chem. Phys. 18 (2016) 20542–20549.

[115] Kamal Choudhary, Marnik Bercx, Jie Jiang, Ruth Pachter, Dirk Lamoen, Francesca Tavazza, Accelerated discovery of efficient solar cell materials using quantum and machine-learning methods, Chem. Mater. 31 (15) (2019) 5900–5908.

[116] Logan Williams, Sl3me – a Python3 implementation of the spectroscopic limited maximum efficiency (SLME) analysis of solar absorbers, 2022.

[117] Stefano Baroni, Paolo Giannozzi, Eyvaz Isaev, Density-functional perturbation theory for quasi-harmonic calculations, Rev. Mineral. Geochem. 71 (1) (2010) 39–57.

[118] Azimatu Seidu, Marc Dvorak, Jari Järvi, Patrick Rinke, Jingrui Li, Surface reconstruction of tetragonal methylammonium lead triiodide, APL Mater. 9 (11) (2021) 111102.

[119] Duyen H. Cao, Peijun Guo, Arun Mannodi-Kanakkithodi, Gary P. Wiederrecht, David J. Gosztola, Nari Jeon, Richard D. Schaller, Maria K.Y. Chan, Alex B.F. Martinson, Charge Transfer Dynamics of Phase-Segregated Halide Perovskites: $CH_3NH_3PbCl_3$ and $CH_3NH_3PbI_3$ or $(C_4H_9NH_3)_2(CH_3NH_3)_{n-1}Pb_nI_{3n+1}$ Mixtures, ACS Appl. Mater. Interfaces 11 (9) (2019) 9583–9593.

[120] Oxana Andriuc, Martin Siron, Joseph H. Montoya, Matthew Horton, Kristin A. Persson, Automated adsorption workflow for semiconductor surfaces and the application to zinc telluride, J. Chem. Inf. Model. 61 (8) (2021) 3908–3916.

[121] Md. Habibur Rahman, Jiaqi Yang, Yujie Sun, Arun Mannodi-Kanakkithodi, Defect engineering in ZnIn2X4 (X=S, Se, Te) semiconductors for improved photocatalysis, Surf. Interfaces 39 (2023) 102960.

[122] Jia Yuan, Hongli Liu, Shirong Wang, Xianggao Li, How to apply metal halide perovskites to photocatalysis: challenges and development, Nanoscale 13 (2021) 10281–10304.

[123] Kuankuan Ren, Shizhong Yue, Chunhe Li, Zebo Fang, Khaled A.M. Gasem, Jerzy Leszczynski, Shengchun Qu, Zhijie Wang, Maohong Fan, Metal halide perovskites for photocatalysis applications, J. Mater. Chem. A 10 (2022) 407–429.

[124] Ivan A. Aleksandrov, Timur V. Malin, Konstantin S. Zhuravlev, Svetlana V. Trubina, Simon B. Erenburg, Bela Pecz, Yahor V. Lebiadok, Diffusion in GaN/AlN superlattices: DFT and EXAFS study, Appl. Surf. Sci. 515 (2020) 146001.

[125] Yashaswi Nandan, Mohan Singh Mehata, Wavefunction engineering of type-I/type-II excitons of CdSe/CdS core-shell quantum dots, Sci. Rep. 9 (2019) 2.

[126] Liping Cheng, Baoen Xu, Yanli Zeng, Lingpeng Meng, Intrinsic defects at the interface of the FAPbI3/MAPbI3 superlattice: insight from first-principles calculations, Phys. Chem. Chem. Phys. 25 (2023) 6369–6379.

[127] Leo J. McGilly, Alexander Kerelsky, Nathan R. Finney, Konstantin Shapovalov, En-Min Shih, Augusto Ghiotto, Yihang Zeng, Samuel L. Moore, Wenjing Wu, Yusong Bai, Kenji Watanabe, Takashi Taniguchi, Massimiliano Stengel, Lin Zhou, James Hone, Xiaoyang Zhu, Dmitri N. Basov, Cory Dean, Cyrus E. Dreyer, Abhay N. Pasupathy, Visualization of moiré superlattices, Nature Nanotechnol. 15 (2020) 580–584.

[128] Matthew J. Hamer, Alessio Giampietri, Viktor Kandyba, Francesca Genuzio, Tevfik O. Menteş, Andrea Locatelli, Roman V. Gorbachev, Alexei Barinov, Marcin Mucha-Kruczyński, Moiré superlattice effects and band structure evolution in near-30-degree twisted bilayer graphene, ACS Nano 16 (2) (2022) 1954–1962.

[129] Arun Mannodi-Kanakkithodi, Ji-Sang Park, Alex B.F. Martinson, Maria K.Y. Chan, Defect energetics in pseudo-cubic mixed halide lead perovskites from first-principles, J. Phys. Chem. C 124 (31) (2020) 16729–16738.

[130] Christoph Freysoldt, Blazej Grabowski, Tilmann Hickel, Jörg Neugebauer, Georg Kresse, Anderson Janotti, Chris G. Van de Walle, First-principles calculations for point defects in solids, Rev. Mod. Phys. 86 (2014) 253–305.

[131] Ji Sang Park, Sunghyun Kim, Zijuan Xie, Aron Walsh, Point defect engineering in thin-film solar cells, Nat. Rev. Mater. 3 (2018) 194–210.

[132] Arun Mannodi-Kanakkithodi, Michael Y. Toriyama, Fatih G. Sen, Michael J. Davis, Robert F. Klie, Maria K.Y. Chan, Machine-learned impurity level prediction for semiconductors: the example of cd-based chalcogenides, npj Comput. Mater. 6 (1) (2020).

[133] J. Schäfer, A.P. Young, T.M. Levin, L.J. Brillson, J.J. Paggel, L. Vanzetti, A. Franciosi, Cathodoluminescence spectroscopy of deep defect levels at the ZnSe/Gaas interface with a composition-control interface layer, J. Electron. Mater. 28 (1999) 881–886.

[134] Aurangzeb Khan, Yamaguchi Masafumi, Deep level transient spectroscopy: A powerful experimental technique for understanding the physics and engineering of photo-carrier generation, escape, loss and collection processes in photovoltaic materials, in: Leonid A. Kosyachenko (Ed.), Solar Cells, IntechOpen, Rijeka, 2015.

[135] John L. Lyons, Darshana Wickramaratne, Chris G. Van de Walle, A first-principles understanding of point defects and impurities in GaN, J. Appl. Phys. 129 (11) (2021) 111101.

[136] Md Habibur Rahman, Prince Gollapalli, Panayotis Manganaris, Satyesh Kumar Yadav, Ghanshyam Pilania, Brian DeCost, Kamal Choudhary, Arun Mannodi-Kanakkithodi, Accelerating defect predictions in semiconductors using graph neural networks, APL Mach. Learn. 2 (1) (2024) 016122.

[137] Maciej P. Polak, Ryan Jacobs, Arun Mannodi-Kanakkithodi, Maria K.Y. Chan, Dane Morgan, Machine learning for impurity charge-state transition levels in semiconductors from elemental properties using multi-fidelity datasets, J. Chem. Phys. 156 (11) (2022) 114110.

[138] Michael Y. Toriyama, Jiaxing Qu, G. Jeffrey Snyder, Prashun Gorai, Defect chemistry and doping of BiCuSeO, J. Mater. Chem. A 9 (2021) 20685–20694.

[139] Nathan C. Frey, Deji Akinwande, Deep Jariwala, Vivek B. Shenoy, Machine learning-enabled design of point defects in 2D materials for quantum and neuromorphic information processing, ACS Nano 14 (10) (2020) 13406–13417.

[140] M. Witman, A. Goyal, T. Ogitsu, A. McDaniel, S. Lany, Materials discovery for high-temperature, clean-energy applications using graph neural network models of vacancy defects and free-energy calculations, 2023, ChemRxiv. Cambridge: Cambridge Open Engage.

[141] Irea Mosquera-Lois, Seán R. Kavanagh, Aron Walsh, David O. Scanlon, Identifying the ground state structures of point defects in solids, npj Comput. Mater. 9 (1) (2023) 25.

[142] Irea Mosquera-Lois, Seán R. Kavanagh, Aron Walsh, David O. Scanlon, ShakeNBreak: Navigating the defect configurational landscape, J. Open Source Softw. 7 (80) (2022) 4817.

[143] L. Catherine Brinson, Laura M. Bartolo, Ben Blaiszik, David Elbert, Ian Foster, Alejandro Strachan, Peter W. Voorhees, Community action on FAIR data will fuel a revolution in materials research, MRS Bull. (2023).

[144] W.H. Strehlow, E.L. Cook, Compilation of energy band gaps in elemental and binary compound semiconductors and insulators, J. Phys. Chem. Ref. Data 2 (1) (2009) 163–200.

[145] Kanghoon Yim, Youn Yong, Joohee Lee, Kyuhyun Lee, Ho-Hyun Nahm, Jiho Yoo, Chanhee Lee, Cheol Seong Hwang, Seungwu Han, Novel high-k dielectrics for next-generation electronic devices screened by automated ab initio calculations, NPG Asia Mater. 7 (6) (2015).

[146] Miso Lee, Yong Youn, Kanghoon Yim, Seungwu Han, High-throughput ab initio calculations on dielectric constant and band gap of non-oxide dielectrics, Sci. Rep. 8 (1) (2018).

[147] Sangtae Kim, Miso Lee, Changho Hong, Youngchae Yoon, Hyungmin An, Dongheon Lee, Wonseok Jeong, Dongsun Yoo, Youngho Kang, Yong Youn, Seungwu Han, A band-gap database for semiconducting inorganic materials calculated with hybrid functional, Sci. Data 7 (1) (2020).

[148] Gyoung S. Na, Seunghun Jang, Yea-Lee Lee, Hyunju Chang, Tuplewise material representation based machine learning for accurate band gap prediction, J. Phys. Chem. A 124 (50) (2020) 10616–10623.

[149] Zhongyu Wan, Quan-De Wang, Dongchang Liu, Jinhu Liang, Effectively improving the accuracy of PBE functional in calculating the solid band gap via machine learning, Comput. Mater. Sci. 198 (2021) 110699.

[150] Xiang-Guo Li, Ben Blaiszik, Marcus Emory Schwarting, Ryan Jacobs, Aristana Scourtas, K.J. Schmidt, Paul M. Voyles, Dane Morgan, Graph network based deep learning of bandgaps, J. Chem. Phys. 155 (15) (2021) 154702.

[151] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, I. Foster, The materials data facility: Data services to advance materials science research, JOM 68 (2016) 2045–2052.

[152] Mengwei Gao, Bo Cai, Gaoyu Liu, Lili Xu, Shengli Zhang, Haibo Zeng, Machine learning and density functional theory simulation of the electronic structural properties for novel quaternary semiconductors, Phys. Chem. Chem. Phys. 25 (2023) 9123–9130.

[153] Pengcheng Xu, Dongping Chang, Tian Lu, Long Li, Minjie Li, Wencong Lu, Search for ABO3 type ferroelectric perovskites with targeted multi-properties by machine learning strategies, J. Chem. Inf. Model. 62 (21) (2022) 5038–5049.

[154] Wei Li, Zigeng Wang, Xia Xiao, Zhiqiang Zhang, Anderson Janotti, Sanguthevar Rajasekaran, Bharat Medasani, Predicting band gaps and band-edge positions of oxide perovskites using density functional theory and machine learning, Phys. Rev. B 106 (2022) 155156.

[155] Guido Petretto, Shyam Dwaraknath, Henrique P.C. Miranda, Donald Winston, Matteo Giantomassi, Michiel J. van Setten, Xavier Gonze, Kristin A. Persson, Geoffroy Hautier, Gian-Marco Rignanese, High-throughput density-functional perturbation theory phonons for inorganic materials, Sci. Data 5 (1) (2018) 180065.

[156] Ioannis Petousis, David Mrdjenovich, Eric Ballouz, Miao Liu, Donald Winston, Wei Chen, Tanja Graf, Thomas D. Schladt, Kristin A. Persson, Fritz B. Prinz, High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials, Sci. Data 4 (1) (2017) 160134.

[157] Anjana Talapatra, Blas P. Uberuaga, Christopher R. Stanek, Ghanshyam Pilania, A machine learning approach for the prediction of formability and thermodynamic stability of single and double perovskite oxides, Chem. Mater. 33 (3) (2021) 845–858.

[158] Hai-Chen Wang, Silvana Botti, Miguel A.L. Marques, Predicting stable crystalline compounds using chemical similarity, npj Comput. Mater. 7 (1) (2021) 12.

[159] Danny Broberg, Kyle Bystrom, Shivani Srivastava, Diana Dahliah, Benjamin A.D. Williamson, Leigh Weston, David O. Scanlon, Gian-Marco Rignanese, Shyam Dwaraknath, Joel Varley, Kristin A. Persson, Mark Asta, Geoffroy Hautier, High-throughput calculations of charged point defect properties with semi-local density functional theory—performance benchmarks for materials screening applications, npj Comput. Mater. 9 (1) (2023) 72.

[160] Diana Dahliah, Guillaume Brunin, Janine George, Viet-Anh Ha, Gian-Marco Rignanese, Geoffroy Hautier, High-throughput computational search for high carrier lifetime, defect-tolerant solar absorbers, Energy Environ. Sci. 14 (2021) 5057–5073.

[161] Fadoua Khmaissia, Hichem Frigui, Mahendra Sunkara, Jacek Jasinski, Alejandro Martinez Garcia, Tom Pace, Madhu Menon, Accelerating band gap prediction for solar materials using feature selection and regression techniques, Comput. Mater. Sci. 147 (2018) 304–315.

[162] Youngho Kang, Yong Youn, Seungwu Han, Jiwon Park, Chang-Seok Oh, Computational screening of indirect-gap semiconductors for potential photovoltaic absorbers, Chem. Mater. 31 (11) (2019) 4072–4080.

[163] Hong-Jian Feng, Kan Wu, Zun-Yi Deng, Predicting inorganic photovoltaic materials with efficiencies >26 structure-relevant machine learning and density functional calculations, Cell Rep. Phys. Sci. 1 (9) (2020) 100179.

[164] A. Mannodi-Kanakkithodi, M.K.Y. Chan, High-throughput density functional theory dataset of pb-site impurities in hybrid perovskites, Mater. Data Facil. (2022).

[165] A. Mannodi-Kanakkithodi, M.K.Y. Chan, J. Yang, P. Manganaris, High-throughput DFT dataset of halide perovskite alloys, Mater. Data Facil. (2022).

[166] A. Mannodi-Kanakkithodi, et al., Github, 2023, https://github.com/mannodiarun/perovs_mfml_ga.

[167] Arun Mannodi-Kanakkithodi, Austin McDannald, Shijing Sun, Saaketh Desai, Keith A. Brown, A. Gilad Kusne, A framework for materials informatics education through workshops, MRS Bull. 48 (2023) 560–569.

[168] Emmanouil Kioupakis, Sieun Chae, Kyle Bushick, Nick Pant, Xiao Zhang, Woncheol Lee, Theoretical characterization and computational discovery of ultra-wide-band-gap semiconductors with predictive atomistic calculations, J. Mater. Res. 36 (23) (2021) 4616–4637.

[169] Andrew J. Green, James Speck, Grace Xing, Peter Moens, Fredrik Allerstam, Krister Gumaelius, Thomas Neyer, Andrea Arias-Purdue, Vivek Mehrotra, Akito Kuramata, Kohei Sasaki, Shinya Watanabe, Kimiyoshi Koshi, John Blevins, Oliver Bierwagen, Sriram Krishnamoorthy, Kevin Leedy, Aaron R. Arehart, Adam T. Neal, Shin Mou, Steven A. Ringel, Avinash Kumar, Ankit Sharma, Krishnendu Ghosh, Uttam Singisetti, Wenshen Li, Kelson Chabak, Kyle Liddy, Ahmad Islam, Siddharth Rajan, Samuel Graham, Sukwon Choi, Zhe Cheng, Masataka Higashiwaki, β-Gallium oxide power electronics, APL Mater. 10 (2) (2022) 029201.

[170] Yuichi Oshima, Elaheh Ahmadi, Progress and challenges in the development of ultra-wide bandgap semiconductor α-Ga_2O_3 toward realizing power device applications, Appl. Phys. Lett. 121 (26) (2022) 260501.

[171] Mariette Hellenbrandt, The inorganic crystal structure database (ICSD)—Present and future, Crystallogr. Rev. 10 (1) (2004) 17–22.

[172] Yabi Wu, M.K.Y. Chan, G. Ceder, Prediction of semiconductor band edge positions in aqueous environments from first principles, Phys. Rev. B 83 (2011) 235301.

[173] G. Pilania, Mannodi-Kanakkithodi, First-principles identification of novel double perovskites for water-splitting applications, J. Mater. Sci. 52 (2017) 8518–8525.

[174] Evan R. Antoniuk, Peter Schindler, W. Andreas Schroeder, Bruce Dunham, Piero Pianetta, Theodore Vecchione, Evan J. Reed, Novel ultrabright and air-stable photocathodes discovered from machine learning and density functional theory driven screening, Adv. Mater. 33 (44) (2021) 2104081.

[175] Hao Jin, Huijun Zhang, Jianwei Li, Tao Wang, Langhui Wan, Hong Guo, Yadong Wei, Data-driven systematic search of promising photocatalysts for water splitting under visible light, J. Phys. Chem. Lett. 10 (17) (2019) 5211–5218.

[176] G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Sci. Rep. 6 (1) (2016) 19375.

[177] Fabien Tran, Sohaib Ehsan, Peter Blaha, Assessment of the GLLB-SC potential for solid-state properties and attempts for improvement, Phys. Rev. Mater. 2 (2018) 023802.

[178] Andrea Anelli, Edgar A. Engel, Chris J. Pickard, Michele Ceriotti, Generalized convex hull construction for materials discovery, Phys. Rev. Mater. 2 (2018) 103804.