

## High-throughput hybrid-functional DFT calculations of bandgaps and formation energies and multifidelity learning with uncertainty quantification

Mohan Liu , Abhijith Gopakumar , Vinay Ishwar Hegde, Jiangang He , and Chris Wolverton \*

*Department of Materials Science and Engineering, Northwestern University, Evanston, Illinois 60208, USA*



(Received 3 October 2023; accepted 8 February 2024; published 16 April 2024)

Despite the fact that first-principles methods are critical tools in the study and design of materials today, the accuracy of density functional theory (DFT) prediction is fundamentally reliant on the exchange-correlation functional chosen to approximate the interactions between electrons. Although the general improvement in accurately calculating the bandgap with the Heyd-Scuseria-Ernzerhof (HSE) hybrid-functional method over the conventional semilocal DFT is well accepted, other properties such as formation energy have not been systematically studied and have yet to be evaluated thoroughly for different classes of materials. A high-throughput hybrid-functional DFT investigation on materials bandgaps and formation energies is therefore performed in this work. By evaluating over a thousand materials, including metals, semiconductors, and insulators, we have quantitatively verified that the materials bandgaps obtained through HSE [mean absolute error (MAE) = 0.687 eV] are more accurate than those from the Perdew-Burke-Ernzerhof (PBE) functional (MAE = 1.184 eV) when compared to the experimental values. For formation energies, PBE systematically underestimates the magnitude of the formation enthalpies (MAE = 0.175 eV/atom), whereas formation enthalpies obtained from the HSE method are generally more accurate (MAE = 0.147 eV/atom). We have also found that HSE significantly increases the accuracy of formation energy prediction for insulators and strongly bound compounds. A primary application of this new dataset is achieved by building a cokriging multifidelity machine learning (ML) model to quickly predict the bandgaps with HSE-level accuracy when its PBE bandgap is available from DFT calculations. The preliminary goal of our ML model, benchmarked in this work, is to select the semiconductors and insulators which may have been mislabeled as metals from the DFT-PBE calculations in the existing Open Quantum Materials Database. The performance of the cokriging model in reliably predicting HSE bandgaps with quantified model uncertainty is analyzed by comparing the results against published experimental data from the literature.

DOI: [10.1103/PhysRevMaterials.8.043803](https://doi.org/10.1103/PhysRevMaterials.8.043803)

### I. INTRODUCTION

Since its original formulation in the 1960s, density functional theory (DFT) [1] has evolved to become a key method to discover novel materials computationally and to estimate their material properties starting with little to no experimental input [2]. More recently, the versatility of DFT coupled with the increase in computer power and decrease in CPU costs has enabled high-throughput calculations of thousands of compounds. The ability to compute efficiently and exhaustively the properties of crystalline structures has led to the creation of several materials databases, each encompassing hundreds of thousands or even millions of different compounds [3–8].

Even though databases of DFT calculations have proven to be useful for discovering materials for a variety of applications at 0 K [2,9–11], the accuracy of predictions is fundamentally limited by the exchange-correlation functional employed in the calculations. A well-known inaccuracy is the large difference between experimental bandgaps and the bandgaps predicted using generalized gradient approximation [12] (GGA) exchange-correlation (XC) functionals—the current standard for high-throughput calculations [13]. In many cases, the bandgap energy can be predicted much more accurately using techniques such as the modified

Becke-Johnson (mBJ) potential [13] or hybrid functionals [14,15]. There are several hybrid-functional approaches, such as the Becke three-parameter Lee-Yang-Parr (B3LYP), Perdew-Burke-Ernzerhof (PBE0) and Heyd-Scuseria-Ernzerhof (HSE06) functionals [14,15], while HSE06 is particularly relevant for solids. Unfortunately, HSE calculations can be 2 orders of magnitude more computationally expensive than GGA calculations, making their use on a large scale problematic. Furthermore, for an arbitrary material, it is difficult to know *a priori* whether HSE calculations will be significantly more accurate than those with GGA functionals.

The improvement in calculated properties with HSE over GGA is often not systematic and has yet to be evaluated thoroughly for different classes of materials even though the estimation of bandgaps is shown to be generally improved with HSE06 compared to PBE [13,16]. Pilania *et al.* [17] studied 250  $A_2BB'X_6$ -type elpasolite compounds and observed a roughly linear trend in PBE and HSE bandgaps. For transition-metal oxides, HSE improves not only the bandgaps but also the properties related to thermochemistry, which includes formation energies [18,19]. In an analysis by Chevrier *et al.* [20], HSE06 was shown to predict more accurate formation energies than PBE for transition-metal oxides, but surprisingly, this was not true for non-transition-metal oxides. Zhang *et al.* [21] showed that HSE can drastically improve the predicted thermodynamic stabilities (i.e., formation

\*c-wolverton@northwestern.edu

energies) in the case of the metallic Cu-Au alloy system, where PBE fails to accurately predict the formation energies for the intermetallics with an error approaching 50%. Even for the prototypical example of the CsCl structure, both PBE and HSE surprisingly predict CsCl will adopt the NaCl structure [22]. All these examples show that when (and to what extent) HSE would provide an advantage over PBE has still not been well established. Considering that HSE is at least 50–100 times more expensive than GGA when calculating a typical compound, it would be very useful for computational materials scientists to have a computationally efficient method to predict *a priori* which method would be more appropriate for predicting a certain property and when both would fail.

In this work, we utilize high-throughput computational techniques to quantify the level of accuracy expected from HSE calculations for over 1200 solid-state materials. We find that HSE performs well in computing materials bandgaps with a decrease in the mean absolute error of more than 40% to 0.687 eV relative to PBE when compared against experimental values. For formation energy calculations, we find that HSE performs remarkably well for compounds with nonzero bandgaps and for compounds that are strongly bound. For the rest of the compounds, the improvement from PBE to HSE is not that significant. Overall, HSE is a significant advance in describing both bandgaps and formation energies for certain classes of compounds. However, considering the huge computational cost of HSE, a data-driven selection strategy is necessary to wisely choose which method should be used.

Further, to take advantage of having data on both PBE and HSE bandgaps of materials, we apply multifidelity [17,23–25] statistical modeling to our dataset. Multifidelity models are statistical machine learning models employed to fit datasets where each data point has multiple values available to describe the same target property. One of these target-property values for each data point is determined via a less accurate method (low fidelity), while the other is obtained using a more accurate but relatively expensive strategy (high fidelity). The goal of a multifidelity model is to predict the target property with high-fidelity accuracy when a low-fidelity estimation of the target property is known for the same data point. In this work, fidelity is defined as the accuracy of a DFT method in calculating bandgaps. For a given material, the high-fidelity and low-fidelity data correspond to HSE and PBE bandgap values, respectively. We report the multifidelity model validation details and predictive ability for a set of materials whose HSE bandgap values were not calculated in our high-throughput DFT work. The model uncertainties are also quantified and analyzed during the training and prediction stages. A subset of model predictions is analyzed in comparison to experimental and HSE bandgap values reported in the published scientific literature. The final multifidelity model is shown to be reliable for predicting bandgap values in HSE fidelity at the cost of doing a lower-fidelity DFT calculation to estimate the PBE bandgap.

## II. METHODOLOGY

### A. High-throughput calculation framework

This work has extensively utilized the framework of the Open Quantum Materials Database (OQMD) [4]. In order to

perform hybrid-functional calculations in a high-throughput manner, we implemented new features into the automated framework within the PYTHON application programming interface (API) of the OQMD, called QMPY, to handle DFT calculations with a different functional. These new features include the support for parallelization of jobs that perform hybrid-functional DFT calculations using the Vienna *Ab initio* Simulation Package (VASP), associated error handling, and parsing of data from DFT calculation output files unique to hybrid-functional calculations. The existing version of QMPY has functionalities (1) to read (write) crystal structure data from (into) VASP POSCAR and CIF formats, (2) to generate input files for a DFT calculation using VASP with different functionals, (3) to generate batch script files to run the calculations on various clusters, (4) to monitor the progress of the jobs submitted on different clusters, (5) to handle common run-time errors associated with VASP simulations such as insufficient convergence of energy and forces, (6) to parse the output and extract relevant data once the calculations are successfully completed, (7) to store the resulting data in a MYSQL database, (8) to perform thermodynamic analysis to determine the phase stability and generate phase diagrams, and (9) to enable simple, but powerful, querying of all the parameters of the calculations and properties of the compounds stored in the database via a web API created with the open-source DJANGO PYTHON API framework.

For each individual material, we perform four individual DFT calculations in order, starting with an initial full structure relaxation with the PBE functional. All the subsequent calculations are done on this PBE-relaxed structure. In the second step, an accurate wave function for the structure (WAVECAR output file from VASP) is calculated from a single-point DFT-PBE calculation. The third and final stage involves a hybrid-functional calculation in which the wave function generated in the second step is used as input. Based on a few example test cases, we found that the single-point PBE wave function calculation before the actual HSE calculation is necessary in order to increase the efficiency of the overall hybrid-functional calculation. The workflow is illustrated in Fig. 1.

### B. Hybrid-functional DFT settings

DFT calculations are performed using the projector augmented-wave [27] (PAW) method implemented in VASP [28]. The plane-wave cutoffs were set at 520 and 400 eV for PBE and HSE, respectively. We employ the recommended VASP 5.2 PBE PAW potentials for both PBE and HSE calculations.  $k$ -point meshes are generated using generalized Monkhorst-Pack grids with a  $k$ -point density of 8000 for PBE and 2000 for HSE [29]. Spin-polarized calculations are done for magnetic compounds. The magnetic nature of a compound during the initialization of the first structural relaxation is determined based on the number of outer shell  $d$  and  $f$  electrons in the nonionic states of the constituent elements. The relaxation calculation output from VASP contains the integrated magnetic moments derived from the simulation, which are used in the input of subsequent static calculations.

Formation energies are calculated by

$$\Delta E_f = E - \sum_i x_i \mu_i, \quad (1)$$

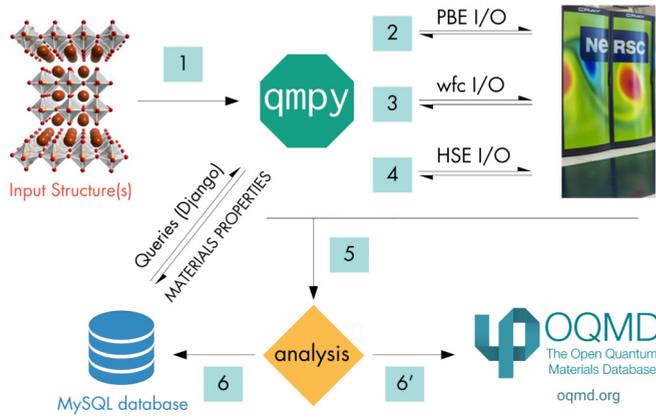


FIG. 1. The workflow for hybrid-functional calculations implemented within the QMPY package. For given input structure(s), QMPY generates the VASP input files to perform a relaxation run with the PBE functional, followed by a calculation to generate the electronic wave function (wfc) and, finally, the hybrid-functional calculation. The TASKSERVER and JOBSERVER modules communicate with High-Performance Computing Cluster (HPCC) resources, monitor running jobs, and transfer data between local database storage and HPCC servers. The final calculation data undergo thermodynamic analysis and are stored in a MySQL database. The web interface of the Open Quantum Materials Database (OQMD) [26] provides convenient database access to the research community.

where  $E$  is the total energy of a compound containing  $x_i$  atoms of element  $i$  in the formula unit and  $\mu_i$  is the elemental chemical potential. Unless specified otherwise, all formation energies mentioned in this work are calculated by taking DFT total energies of elemental ground states at 0 K as the corresponding elemental chemical potentials. As an additional step to improve the accuracy of formation energy calculations, the experimental, room-temperature formation energy data available from external databases can be used to correct the DFT-based elemental chemical potentials [30]. This method is discussed separately in detail in Sec. III C. In this work, we make the assumption that the pressure-volume (pV) contributions to the enthalpies are small for solid materials at ambient pressure and therefore  $\Delta E_f \approx \Delta H_f$ , where the latter is the formation enthalpy.

### C. Statistical modeling

The goal of statistical modeling in this work is to learn information from the calculated HSE bandgap data (also called training data) and use it to predict the same property for the set of materials whose HSE calculations have not been carried out. To achieve this, we perform multifidelity modeling based on cokriging [31–33]. Cokriging is an extension of the kriging interpolation, also known as Gaussian process regression in the machine learning literature. Each material in the training dataset has an HSE bandgap  $E_g^{\text{HSE}}$  and a PBE bandgap  $E_g^{\text{PBE}}$  derived via high-throughput DFT calculations in this work.  $E_g^{\text{HSE}}$  is the value with higher fidelity, while  $E_g^{\text{PBE}}$  is the corresponding low-fidelity data. A cokriging-based multifidelity machine learning model was created and trained over this training data with the goal of predicting a higher-fidelity  $E_g^{\text{HSE}}$

value for a candidate material when its  $E_g^{\text{PBE}}$  value is known. The set of candidate materials for prediction (called the search space) is curated here as the set of stable materials in OQMD whose PBE calculations have already been carried out but whose HSE calculations have not. The search space contains 24 967 materials available in the OQMD [26] which fall on the convex hull of ground-state-stable compositions. Each material is represented by a set of input features generated using the MAGPIE [34] package based on its crystal structure and chemical composition. A simplified representation of the cokriging model is given in the following equation, while a detailed discussion of the theory of and algorithm for cokriging modeling is provided by Pilia *et al.* [17]:

$$Z_{\text{HSE}}(x) = \rho Z_{\text{PBE}}(x) + Z_d(x). \quad (2)$$

Here,  $Z_{\text{HSE}}(x)$  and  $Z_{\text{PBE}}(x)$  are Gaussian processes representing  $E_g^{\text{HSE}}$  and  $E_g^{\text{PBE}}$ , respectively. The term  $Z_{\text{PBE}}$  is multiplied by a scaling parameter  $\rho$  whose value is optimized during the model training via maximum likelihood estimation (MLE) [35]. The MLE method finds the values of the Gaussian process parameters that maximize the joint probability of the training dataset being sampled by the associated statistical model. The third term,  $Z_d$ , is the Gaussian function representing the difference between the high-fidelity process and the scaled low-fidelity process. In the original cokriging models, the low-fidelity data are not required to be known for all the data points, and thus, they learn to predict both  $Z_{\text{PBE}}$  and  $Z_d$  for a given set of input features. But in this work, all the materials in both the training data and search space have a known value of  $E_g^{\text{PBE}}$ . So slightly diverging from the original cokriging procedure,  $E_g^{\text{PBE}}$  was added as an input feature in addition to MAGPIE features, thereby making the prediction of  $Z_{\text{PBE}}$  obsolete while placing more importance on the calculation of  $Z_d$ . The main advantage of performing cokriging in this work is attributed to the involvement of a much larger set of hyperparameters to define the Gaussian processes, quantification of model prediction uncertainties, and consideration of a relatively higher-dimensional covariance matrix [17,32] compared to other regression models. In addition, the importance of the  $E_g^{\text{PBE}}$  value is emphasized in the model because it is explicitly present in Eq. (2), while all other input features are used only while constructing the covariance matrix. In this work, the cokriging algorithm is implemented in PYTHON primarily using the OPENMDAO [36] library.

## III. RESULTS

### A. Bandgap: HSE vs PBE vs experiment

We first evaluate the performance of bandgap predictions between PBE and HSE based on 1135 compounds compared to the experimental values. In Fig. 2, we observe that almost all bandgap energies calculated by PBE are smaller than those calculated by HSE. Furthermore, HSE found more nonzero bandgap materials than PBE. For 106 compounds, PBE predicts a zero bandgap, but HSE predicts a nonzero bandgap. To conclude whether HSE outperforms PBE in bandgap prediction, a comparison with the experimental measurements is carried out below. The experimental data are taken from

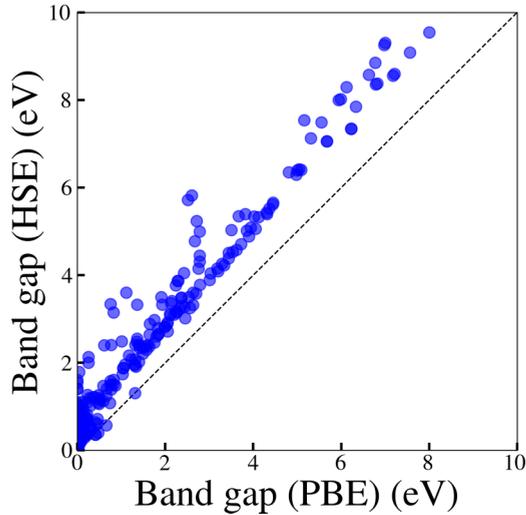


FIG. 2. Comparison of bandgaps for 1135 compounds between PBE and HSE.

the Database of Forbidden Zones of Solids (DFZS) [37]. The bandgap values listed in DFZS are accumulated from several handbooks and articles reporting results from one of several experimental bandgap characterization techniques such as Hall measurement, photoconductivity measurement, etc. The comparison between DFT-calculated and experimentally measured bandgaps for 146 materials is shown in Fig. 3. In general, PBE calculations significantly underestimate the bandgaps, but HSE calculations greatly increase the accuracy in estimating bandgaps.

Prediction errors for different groups of materials are presented in Table I. In group 1, both PBE and HSE calculate zero bandgaps for these 20 cases. Experimental data have shown that these materials have nonzero bandgaps, but the values are relatively small. The average bandgap value for these materials is only 0.45 eV. In group 2, we found 23 cases where HSE successfully identifies the nonzero bandgaps but PBE fails to. The mean absolute error using HSE is only 0.55 eV, which is smaller than that of PBE, with a mean absolute error of 0.82 eV. Furthermore, both PBE and HSE are able to predict the nonzero bandgaps for the materials in group 3. For these materials, HSE significantly decreases the prediction error

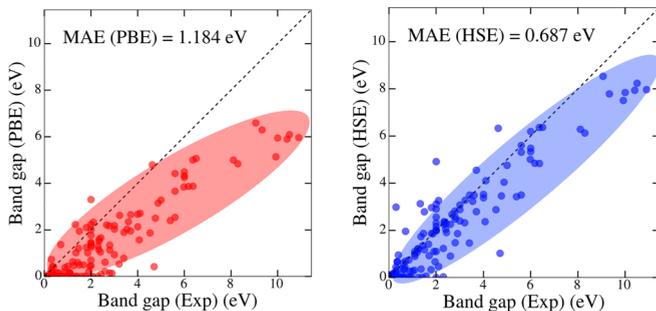


FIG. 3. Comparisons between DFT (left: PBE, right: HSE) and experimental bandgaps for 146 compounds. The experimental data are taken from the Database of Forbidden Zones of Solids (DFZS) [37].

TABLE I. Prediction errors (mean absolute error) of bandgaps calculated using PBE and HSE when compared to the experimental values. In group 1, both PBE and HSE found zero bandgaps. In group 2, PBE found zero bandgaps, but HSE found nonzero bandgaps. In group 3, both PBE and HSE found nonzero bandgaps.

Group	Count	MAE <sub>PBE</sub> (eV)	MAE <sub>HSE</sub> (eV)
1	20	0.45	0.45
2	23	0.82	0.55
3	103	1.41	0.76

from 1.41 to 0.76 eV. The numbers of materials correctly classified as metals or nonmetals by HSE and PBE compared to experimental data are given in Table II. All four materials with zero bandgap in the experimental data are predicted to have zero bandgap by the PBE and HSE calculations as well. Among all the nonmetals, 11% are predicted correctly to be nonmetals by HSE but incorrectly classified to be metals by PBE. In 19% of the cases, both PBE and HSE misclassify the nonmetals as metals.

### B. Formation energies: HSE vs PBE vs experiment

Formation energy values for all 1135 compounds were calculated using both PBE and HSE functionals (shown in Fig. 4). We observed a strong linear relationship between PBE-calculated and HSE-calculated formation energies. A linear regression fitting was performed between PBE and HSE, and the slope of the linear fitting line is 0.887. This result indicates that there are systematic differences between PBE and HSE when computing materials formation energies. To better evaluate the reliability of both methods, we compare our DFT-calculated formation energies using both methods to their corresponding experimental measurements. The experimental formation energy data come from two sources: the Scientific Group Thermodata Europe Solid Substance (SSUB) database [38] and the thermodynamic database of the Thermal Processing Technology Center at the Illinois Institute of Technology (IIT) [39], where the IIT database mostly focuses on intermetallic compounds. The absolute errors of calculated formation energies are shown in Fig. 5 (PBE on the left and HSE on the right). For PBE, most of the data are in the right section of this plot, indicating that PBE generally underestimates the magnitude of the formation energy. The overall mean absolute error (MAE) is 0.175 eV/atom, and  $\Delta H_f$  for a few metal oxides are poorly estimated. On the other hand,

TABLE II. Prediction of the existence of a nonzero bandgap from HSE and PBE among all the 142 nonmetallic materials in the experimental dataset ( $E_g^{\text{EXP}} > 0\text{eV}$ ). Only four metals are present in the experimental dataset ( $E_g^{\text{EXP}} = 0\text{eV}$ ), and they all are predicted to have zero bandgaps from HSE and PBE as well.

	$E_g^{\text{HSE}} = 0\text{eV}$	$E_g^{\text{HSE}} > 0\text{eV}$
$E_g^{\text{PBE}} = 0\text{eV}$	27 (19%)	16 (11%)
$E_g^{\text{PBE}} > 0\text{eV}$	0	99 (70%)
Total ( $E_g^{\text{PBE}} \geq 0\text{eV}$ )	27 (19%)	115 (81%)

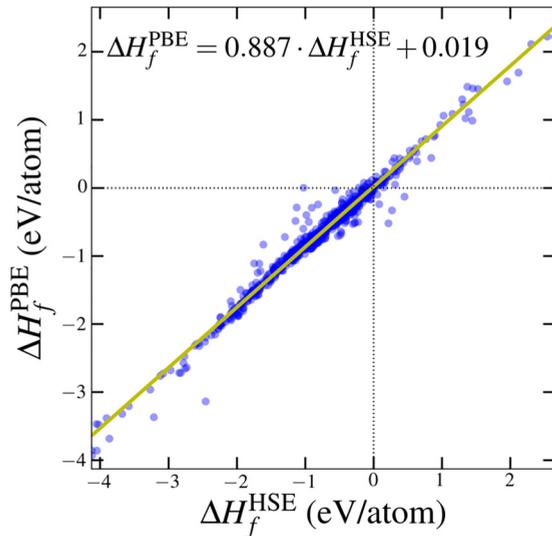


FIG. 4. Comparison of  $\Delta H_f$  between PBE and HSE.

formation energies calculated by HSE (in the right plot) appear to be more accurate, with a decreased mean absolute error of 0.147 eV/atom. By comparing the distributions of both methods, we find the data points from HSE are more symmetric. Therefore, the HSE method statistically outperforms PBE considering both variance and bias in the distribution.

However, we cannot draw the conclusion that HSE formation energies are always more accurate than PBE formation energies. There are certainly a number of compounds for which HSE either overestimates (e.g., some intermetallic compounds) or underestimates (e.g.,  $\text{Mo}_2\text{C}$ ,  $\text{CuO}$ , etc.) the magnitude of the formation energies. In order to analyze these errors separately, we partition the compounds into two groups based on PBE-calculated formation energies. We choose to classify strongly bound compounds with  $\Delta H_f(\text{PBE}) < -1$  eV/atom and the rest as weakly bound compounds, following a similar strategy in a related work [40]. It is worth noting that the choices of PBE-calculated formation energies and the threshold of  $-1$  eV/atom are somewhat arbitrary but fair enough to perform a reasonable analysis. Some slight variations of these choices will not lead to large changes

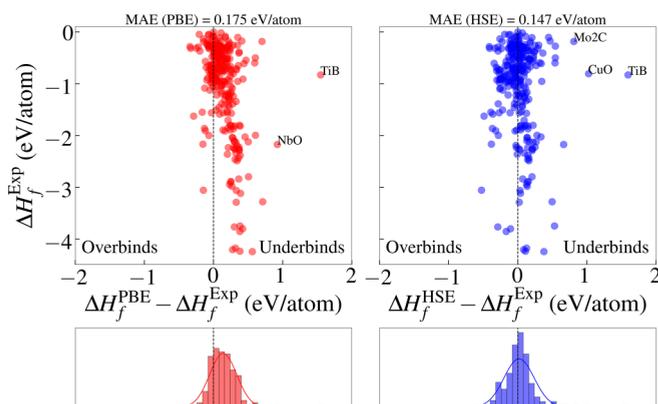


FIG. 5.  $\Delta H_f$  comparisons between DFT (left: PBE, right: HSE) and experiments.

in the conclusions discussed below. Using this convention, we consider mean error, MAE, and root-mean-square error (RMSE) to make comparisons, and the results are shown in Fig. 6. For each subset of compounds, all three error metrics qualitatively follow the same trend. In general, HSE significantly outperforms PBE for strongly bound compounds by all error metrics, whereas HSE is only slightly better than PBE for weakly bound compounds.

Similarly, we also group the compounds into metals, small-bandgap materials, and large-bandgap materials based on the PBE-computed bandgap value. We define materials with a PBE bandgap  $< 1.0$  eV as small-bandgap compounds and the rest as large-bandgap compounds. Here, we find that PBE has a slightly lower MAE than HSE for metals, which is the only case where PBE is statistically better than HSE. Within these three groups, PBE provides a sizable error for small-bandgap compounds and a slightly lower error for large-bandgap compounds. On the other hand, errors in formation energy predicted by HSE are relatively the same between metals and small-bandgap compounds but much smaller for large-bandgap compounds. In Fig. 7, we show the relation between formation energy error and calculated bandgaps. We find that several oxide compounds, including  $\text{VO}$ ,  $\text{NiO}$ ,  $\text{CuO}$ , etc., have large deviations in the PBE-calculated formation energy. There are also some outliers like  $\text{ScN}$ ,  $\text{YN}$ ,  $\text{EuO}$ ,  $\text{EuS}$ , and  $\text{CaSe}$  for which HSE significantly underestimates the magnitude of the formation energy. The box plot in the right panel of Fig. 7 implies that HSE significantly outperforms PBE in formation energy prediction for these nonzero bandgap compounds. Furthermore, we compare the errors between experimental and DFT-predicted formation energies for a variety of material classes using PBE and HSE. The MAEs in formation energy calculations for six different material classes are shown in Fig. 8. For alkali, alkaline earth, and halide compounds, HSE significantly increases the accuracy. However, for basic metal, semimetal, and transition-metal compounds, the two methods estimate formation energy values with a similar MAE.

### C. Elemental chemical potential corrections

Up to this point, the formation energies reported in this work used DFT total energies of elements at 0 K as the elemental chemical potential, as explained in Sec. II. But this strategy disregards the changes in elemental ground states when the temperature is raised from 0 K to room temperature and also the inaccuracies in DFT for the calculation of liquid and gas phase reference states. One of the available methods to work around this issue and improve the accuracy of formation energy estimations is to make small corrections to the DFT total-energy elemental chemical potentials based on the available experimental formation energies [5,30,41]. Hence, to improve the accuracy of DFT-predicted formation energies, we here perform a simultaneous least-squares fitting [41] of the elemental chemical potential  $\mu$  for all available elements. For all 290 compounds with DFT-calculated and experimentally measured formation energies, both the MAE and RMSE of  $\Delta H$  are reduced when the chemical potential corrections are made. The computed chemical potential corrections are plotted in Fig. 9. We find, consistent with

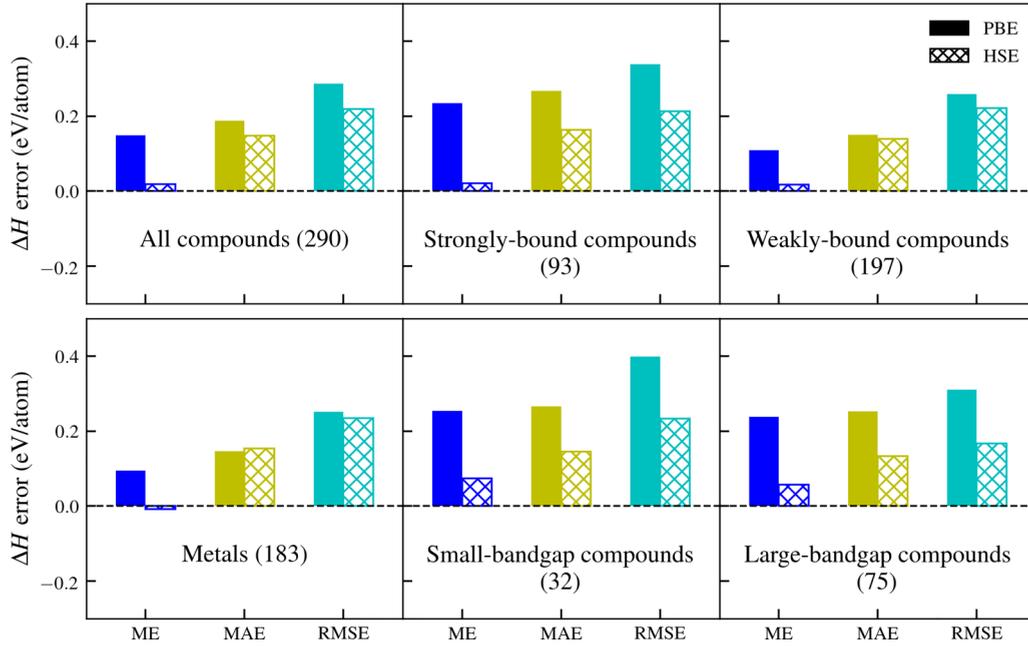


FIG. 6. Comparisons of errors in calculating the formation energies via PBE and HSE relative to experiment. Compounds are partitioned into groups: all compounds, strongly bound compounds, weakly bound compounds, all metals, small-bandgap compounds, and large-bandgap compounds. ME, MAE, and RMSE refer to mean error, mean absolute error and root-mean-square error, respectively. The numbers of compounds in each group are denoted in parentheses.

previous work [40], significant positive PBE corrections for electronegative elements such as O, S, F, Br, and Cl. These large corrections are caused by PBE tending to underestimate the magnitude of the formation energy for compounds that have these elements in them. On the contrary, the HSE chemical potential corrections for these electronegative elements are much smaller. In addition to these elements, we find that, in general, HSE chemical potential corrections are smaller than PBE corrections in magnitude, which implies chemical potential fitting will have less influence on HSE formation energies than PBE ones. However, there are elements with large HSE corrections like Ge and Ga due to the fact that HSE systematically underbinds compounds like GaP, GaAs, GeP, and GeSe compared to PBE. We also want to point out that

Ni, Fe, Se, Rb, Hf, U, and Pu have opposite signs for chemical potential corrections between HSE and PBE and the rest of the elements have the same signs.

The overall MAEs and RMSEs for corrected PBE and HSE formation energies are shown in Fig. 10. With elemental chemical potential fitting, the MAE reduces from 0.187 to 0.102 eV/atom for PBE and 0.147 to 0.100 eV/atom for HSE. Similarly, the RMSE drops from 0.287 to 0.173 eV/atom for PBE and 0.219 to 0.164 eV/atom for HSE. For both errors, we can see that HSE always outperforms PBE with or without elemental fitting, but the gap gets closer when corrections are utilized. Furthermore, we find that the  $\Delta H$  RMSEs for

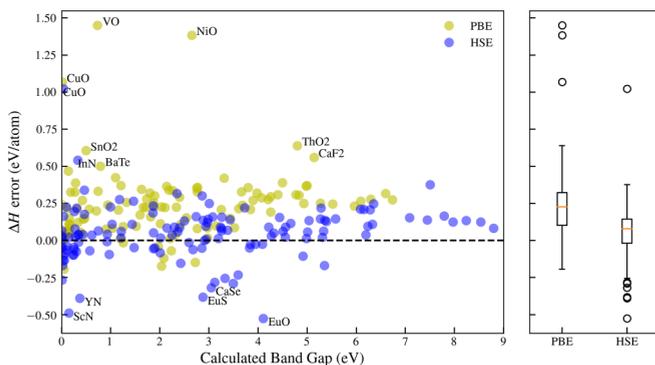


FIG. 7. Left: DFT errors in prediction  $\Delta H_f$  with respect to calculated bandgaps for nonzero bandgap compounds. Right: Box plots for DFT calculation errors in formation energy for nonzero bandgap compounds.

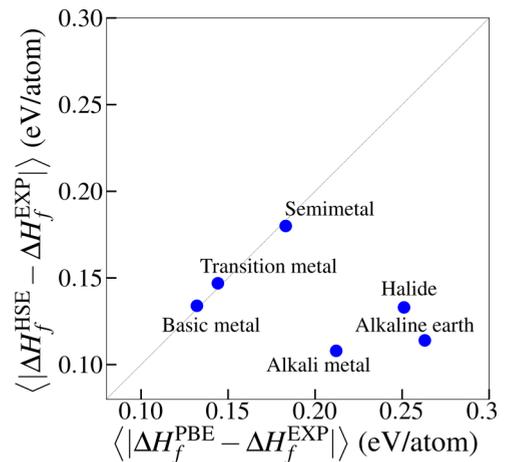


FIG. 8. Mean absolute errors in DFT estimation of  $\Delta H_f$  using PBE and HSE when compared against experimental data among different classes of binary compounds.

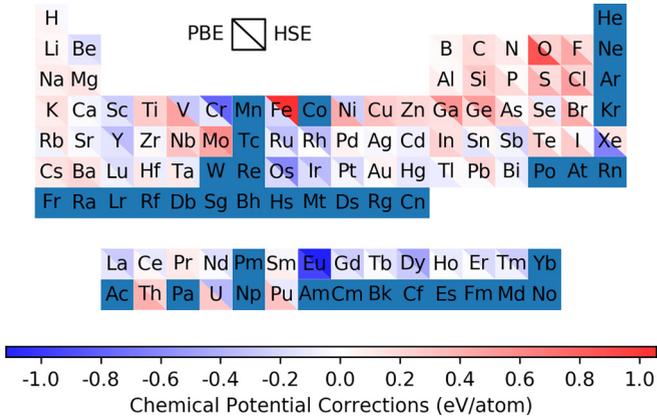


FIG. 9. Elemental chemical potential corrections  $\delta\mu$  obtained via fitting for PBE and HSE. Within each cell, the upper triangle shows HSE chemical potential corrections, and the lower triangle shows PBE corrections. Elements with positive corrections are colored red, and negative ones are colored blue. The elements for which we are currently unable to compute the chemical potentials are enclosed in solid squares without a diagonal line.

metals after fitting are very close between PBE and HSE. But HSE outperforms PBE by 20 meV/atom in predicting  $\Delta H$  for insulators (in Table III). The elemental correction values for five diatomic molecules are given in Table IV alongside the correction values reported by Grindy *et al.* [42].

#### D. Predictions from multifidelity modeling

The multifidelity cokriging model benchmarking results for different data environments are shown in Fig. 11. Figures 11(a) and 11(b) show the optimistic and pessimistic data situations, respectively. In Figs. 11(a), the data are split in such a way that most of the known data (80%) are used to train the model and the testing is done only on the rest of the data (20%) to assess the best performance of the cokriging algorithm in the HSE bandgap dataset when most of the information is provided to the model. In Fig. 11(b), the model is trained on a very small subset of known data (10%), and the test data are significantly larger (90% of all known data)

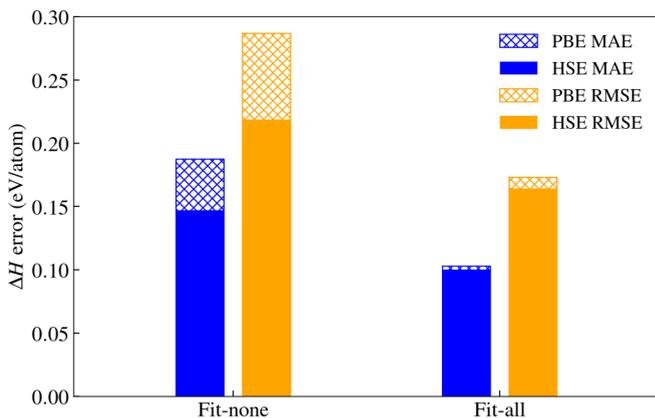


FIG. 10. MAE and RMSE for corrected PBE and HSE formation energies for the cases of fitting  $\mu$  for no elements (“Fit-none”) and all elements (“Fit-all”).

TABLE III. Formation energy RMSEs (in eV/atom) for PBE and HSE before and after elemental fitting.

	All		Metals		Insulators	
	Fit none	Fit all	Fit none	Fit all	Fit none	Fit all
PBE	0.287	0.173	0.251	0.183	0.339	0.154
HSE	0.219	0.164	0.235	0.180	0.190	0.133

than the training data. This is a situation where the training data do not well represent the feature vector space spanned by the test data. The cokriging models show a reliable fit in both data situations with a low MAE. In addition to the prediction accuracy, another important aspect of cokriging is the quantification of model uncertainty, also called epistemic uncertainty. In Fig. 11(c), we show the usefulness of uncertainty quantification. Among the materials with an absolute error of more than 0.1 eV between the model-predicted HSE bandgap and DFT-calculated HSE bandgap in both data optimistic and data pessimistic cases, more than half of them have a quantified model uncertainty greater than 0.1 eV. We set the upper threshold of 0.1 eV to analyze the model predictions in a scenario that demands very high accuracy in bandgap values. In the optimistic case with a train:test ratio of 8:2, only 12% of the materials have inaccurate predictions without a significant value for predicted uncertainty. The magnitude of the quantified uncertainty can be used to focus only on the reliable predictions from the model in a materials design workflow. In both the optimistic and pessimistic models shown in Fig. 11(c), the uncertainty quantification helps avoid selection of more than half the materials with incorrectly predicted values. Such a filter is preferable in situations where a false-negative prediction is more desirable over a false-positive prediction.

This shows that the uncertainty value which is quantified as the standard deviation of the Gaussian process is an important factor to consider while selecting design candidates based on cokriging model predictions. A final model trained on a train:test data split ratio of 9:1 was used to predict  $E_g^{\text{HSE}}$  of materials in the search space.

TABLE IV. Elemental chemical potential corrections for five diatomic molecules as reported by Grindy *et al.* [42] and this work for comparison. The chemical potential correction values of the rest of the elements calculated in this work are provided in the OQMD+ data portal [65]. Perdew-Wang 1991 generalized gradient approximation (PW91-GGA); Perdew-Zunger Local-density approximations (PZ-LDA).

Element	From Grindy <i>et al.</i> (eV/atom)			This work (eV/atom)	
	PBE-GGA	PW91-GGA	PZ-LDA	PBE-GGA	HSE
O	-0.599	-0.4445	0.127	0.8698	0.2384
H	-0.142	-0.091	0.085	0.0535	-0.0403
N	-0.446	-0.3345	0.2195	0.1128	0.0202
F	-0.442	-0.3515	-0.042	0.4378	0.1652
Cl	-0.483	-0.417	-0.185	0.4914	0.218

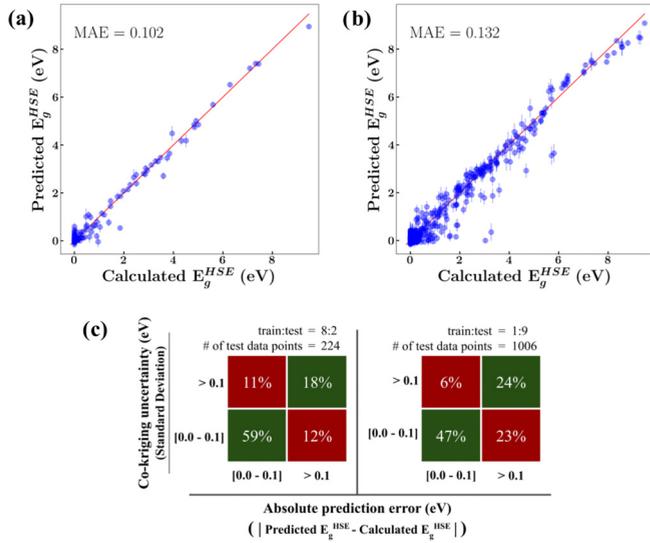


FIG. 11. Multifidelity cokriging model benchmarking results for the HSE bandgap dataset of 1135 materials. Some of the full training data are split and set aside as test data during the training of the model. The trained model's predictions on the test data are plotted. The plot in (a) has an optimistic train:test split of 8:2, which imitates a situation where training data are large enough to reliably learn the correlation between input features and the target property  $E_g^{\text{HSE}}$ . In such a situation, the vector space spanned by input features is sufficiently sampled by the training data. The plot in (b) shows the predictions from a different model trained and tested on a pessimistic train:test data split of 1:9. The pessimistic benchmarking was done to examine the prediction capability of the cokriging model in situations where training data are not large enough to fully represent the relatively larger portion of the feature space spanned by the candidate materials. In (c), the uncertainty quantified for each test data point by the cokriging model during the prediction is analyzed.

The results from multifidelity cokriging model predictions of bandgaps with an HSE fidelity, referred to as  $E_g^{\text{predHSE}}$ , for all materials in the search space (stable OQMD compounds with no available HSE calculations) are provided in Fig. 12. As shown in the plot, 55% of the materials in the search space have no bandgap according to their PBE calculations, while more than half of them are predicted by the machine learning (ML) model to have a nonzero bandgap when calculated using the HSE functional. In total, the cokriging predictions and PBE disagree upon whether a given compound is metallic or not for 30% of the materials in the search space. Among different classes of materials, oxides have the best agreement between PBE and cokriging predictions when classifying a material as metallic. In the same criterion, pnictides and halides are predicted to disagree the most based on the fractions of their total size.

We attempted to validate these multifidelity cokriging model predictions by searching for experimental and calculated (HSE) bandgaps in previously published studies. None of these experimental or HSE bandgaps were used in the ML model training, and they were collected from a literature survey that was conducted only after the final cokriging model predictions were made. A comparison of cokriging-predicted bandgap  $E_g^{\text{predHSE}}$  to the reported DFT-derived  $E_g^{\text{HSE}}$  and ex-

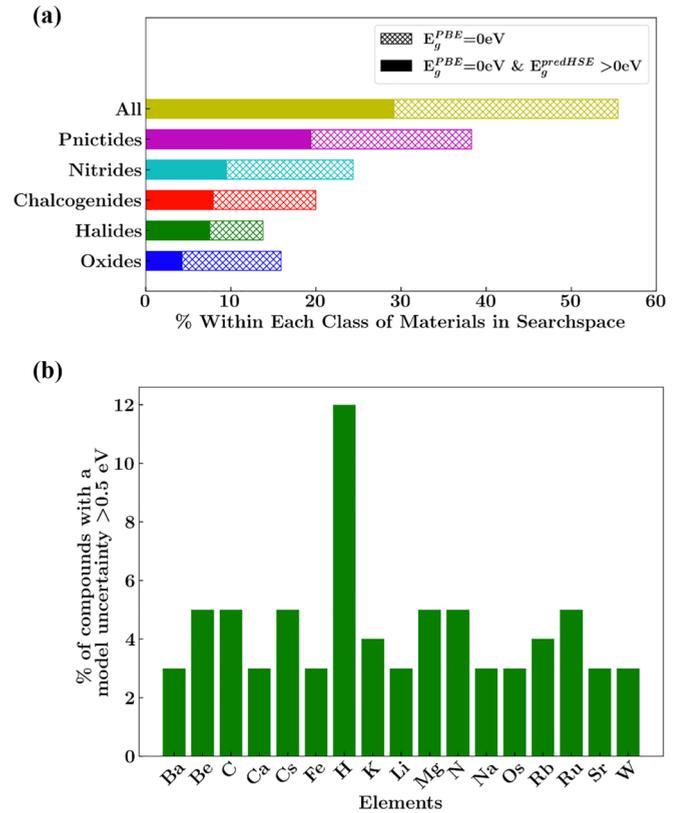


FIG. 12. (a) Bandgap openings predicted by the multifidelity model.  $E_g^{\text{predHSE}}$  refers to the HSE-fidelity bandgap value predicted from cokriging. In about 30% of all materials in the search space, the PBE results and cokriging model predictions disagree on whether a material is metallic or not. (b) Elemental distributions among the materials in the search space with the highest values for model uncertainty. The ordinate of the bar plot represents the percentage of the compounds with a high uncertainty prediction among all the compounds in the search space that contain the element specified on the abscissa of the plot. The bars of only those elements that have an ordinate value of more than 2% are shown. Such a cutoff is kept to make the relevant information stand out and skip other elements such as O, F, etc., which have less than 2% of the compounds predicted to have a high uncertainty value. The model uncertainty is quantified as the standard deviation of the predicted cokriging Gaussian distribution of  $E_g^{\text{predHSE}}$ .

perimental bandgap  $E_g^{\text{EXP}}$  for a small subset of materials from the search space is provided in Table V. We were able to find 13 materials which were predicted correctly as semiconductors by the cokriging model but classified as metallic by PBE calculations. Another 15 materials whose experimental or DFT-derived  $E_g^{\text{HSE}}$  were more accurately predicted by the cokriging model compared to their corresponding DFT-derived  $E_g^{\text{PBE}}$  values are also listed. The 28 materials listed in Table V were not included in the high-throughput HSE study, and thus, they were not included in the training data. As shown in Fig. 12(a), the best agreement between DFT-PBE calculations and cokriging predictions (with HSE fidelity) happens in the case of oxides, followed by halides and chalcogenides. The data in Table V show the capability of the cokriging model to serve as an initial screening method

TABLE V. Literature comparison of cokriging model predictions for the search space. The  $E_g^{\text{predHSE}}$  column refers to the HSE bandgap value predicted by the cokriging model. These materials are selected based on the difference in their  $E_g^{\text{predHSE}}$  and DFT-derived  $E_g^{\text{PBE}}$ . Information about many other materials which were filtered out from the search space based on their bandgap value differences did not have any reported values for DFT-derived  $E_g^{\text{HSE}}$  or the experimental bandgap in the scientific literature within the scope of our search. Further details about the listed materials, including the crystal structure and DFT (with PBE XC functionals) calculation details, can be found on the material's web page identified by OQMD ID [26].

Material	OQMD ID	Space group	$E_g^{\text{PBE}}$ (eV)	$E_g^{\text{predHSE}}$ (eV)	$E_g^{\text{EXP/HSE}}$ (eV)
Ti <sub>2</sub> O <sub>3</sub>	678225	<i>R</i> -3 (148)	0.0	0.8 ± 0.3	0.03-0.14 (EXP) [43]
Ti <sub>3</sub> O <sub>5</sub>	66123	<i>C</i> 2/ <i>m</i> (12)	0.0	0.8 ± 0.3	0.14 (EXP) [44]
Co <sub>3</sub> S <sub>4</sub>	4563	<i>Fd</i> -3 <i>m</i> (227)	0.0	1.6 ± 0.2	1.45 (EXP) [45]
Mn <sub>2</sub> O <sub>3</sub>	33709	<i>Ia</i> -3 (206)	0.0	0.7 ± 0.3	1.4 (EXP) [46]
TiO	10207	<i>C</i> 2/ <i>m</i> (12)	0.0	0.4 ± 0.3	1.9 (EXP) [47]
TiF <sub>3</sub>	5608	<i>R</i> -3 <i>m</i> (166)	0.0	1.1 ± 0.3	2.87 (HSE) [48]
Ni <sub>3</sub> S <sub>4</sub>	6716	<i>Fd</i> -3 <i>m</i> (227)	0.0	1.5 ± 0.2	2.8 (EXP) [49]
MnF <sub>3</sub>	3777	<i>C</i> 2/ <i>c</i> (15)	0.0	0.9 ± 0.3	3.03 (HSE) [48]
NiF <sub>3</sub>	15556	<i>R</i> -3 (148)	0.0	1.4 ± 0.3	3.28 (HSE) [48]
VF <sub>3</sub>	5882	<i>R</i> -3 (148)	0.0	1.0 ± 0.3	3.40 (HSE) [48]
MnSe	30752	<i>P</i> 63 <i>mc</i> (186)	0.0	1.5 ± 0.2	3.5-3.8 (EXP) [50]
MgTiO <sub>3</sub>	692959	<i>R</i> -3 (148)	0.0	0.7 ± 0.3	3.7 (EXP) [51]
MnS	646143	<i>P</i> 63 <i>mc</i> (186)	0.0	1.4 ± 0.2	3.7, 3 (EXP) [52,53]
V <sub>2</sub> O <sub>3</sub>	678210	<i>R</i> -3 (148)	0.3	2.2 ± 0.5	1.51 (EXP) [47]
Mn <sub>3</sub> O <sub>4</sub>	5975	<i>I</i> 41/ <i>amd</i> (141)	0.8	3.5 ± 0.4	2.91 (EXP) [54]
ZnO	4908	<i>P</i> 63 <i>mc</i> (186)	1.0	3.2 ± 0.3	3.29, 3.44 (EXP) [54,55]
Fe <sub>2</sub> O <sub>3</sub>	92501	<i>Ia</i> -3 (206)	1.1	4.0 ± 0.4	1.97 (EXP) [25]
SnO <sub>2</sub>	2477	<i>P</i> 42/ <i>mnm</i> (136)	1.2	2.6 ± 0.3	3.32 (EXP) [54]
MnO <sub>2</sub>	677684	<i>I</i> 4/ <i>m</i> (87)	1.2	3.9 ± 0.3	2.5, 2.7(HSE) [56,57]
LaVO <sub>3</sub>	682189	<i>Pnma</i> (62)	1.2	3.4 ± 0.4	1.44 (EXP) [47]
CdS	5970	<i>P</i> 63 <i>mc</i> (186)	1.3	2.7 ± 0.2	2.58, 2.48 (EXP) [55]
CrF <sub>3</sub>	4854	<i>R</i> -3 <i>c</i> (167)	1.4	4.0 ± 0.3	4.91 (HSE) [48]
GaP	7553	<i>F</i> -43 <i>m</i> (216)	1.8	2.4 ± 0.1	2.26, 2.33 (EXP) [58,59]
TiO <sub>2</sub>	2575	<i>I</i> 41/ <i>amd</i> (141)	2.0	4.5 ± 0.3	3.2 (EXP) [60]
SrTiO <sub>3</sub>	827052	<i>R</i> -3 <i>c</i> (167)	2.0	4.0 ± 0.3	3.2 (EXP) [51]
LaCrO <sub>3</sub>	682305	<i>Pnma</i> (62)	2.1	3.6 ± 0.4	3.39 (EXP) [61]
ZnS	7652	<i>F</i> -43 <i>m</i> (216)	2.3	3.4 ± 0.1	3.84 (EXP) [55]
BeSe	647324	<i>F</i> -43 <i>m</i> (216)	2.8	3.9 ± 0.2	5.15 (EXP) [55]

in materials design challenges to achieve bandgap predictions at HSE fidelity without having to do the computationally expensive DFT-HSE calculations. In Fig. 12(b), the cokriging model prediction uncertainties are analyzed. The uncertainty value is high important when exploring the material spaces in an optimization-based iterative materials design [62–64]. The highest quantified uncertainty was 0.6 eV, and it was assigned for 72 compounds in the search space. Out of these, 52 had hydrogen as a constituent element in their composition. Hydrogen stands out as the only element with a relatively high value on the y axis in Fig. 12(b), indicating difficulty in learning the mapping between input features and the target property of hydrides during the model training. Such uncertainty is not solely due to the limited amount of training data since there are roughly similar numbers of hydrides and oxides in the training data (~4%) but the oxides in the search space are predicted with much less uncertainty than in the case of hydrides. The quantified uncertainty can also be used to filter out material classes from the search space when the high-accuracy prediction of the bandgap with high confidence in trusting the results is significantly more important than

investigating less explored material classes. The cokriging predictions availed in our work can be used to identify the best candidates to conduct further DFT-HSE calculations to find the semiconductors or insulators that are misidentified as conductors in PBE calculations.

#### IV. CONCLUSION

To accelerate the design and discovery of novel materials using computational data, we generated large amounts of computational data on materials and used advanced strategies to efficiently mine these datasets to describe materials properties. In this work, we constructed datasets of bandgaps and formation energies using advanced exchange-correlation functionals (HSE). By analyzing over 1000 materials together, we found that the HSE method in general improves materials' bandgap and formation energy estimation accuracy. However, there are still cases where the computationally efficient PBE outperforms HSE. Therefore, machine learning models are beneficial to define decision rules for identifying when using HSE would significantly improve the theoretical predictions

over PBE. One such approach we reported here based on the cokriging model successfully filtered out a set of materials that were incorrectly predicted as being metallic by DFT calculations using PBE functionals. The multifidelity model achieved reliable accuracy in the prediction of bandgaps with HSE fidelity at the cost of conducting a much cheaper estimation of bandgap using the DFT calculation with PBE functionals. Cokriging also quantifies the epistemic uncertainty in model predictions, which is an important factor in statistical optimization-based materials design workflows.

An interactive user interface is hosted by the Open Quantum Materials Database [65] and contains the three datasets generated in this work: (1) the high-throughput DFT-HSE and DFT-PBE data of material bandgaps and formation energies, (2) elemental chemical potentials and corrections from DFT-PBE and DFT-HSE, and (3) the predicted DFT-HSE bandgap values along with quantified uncertainty from the multifidelity cokriging model for more than 20 000 materials. Alternatively, all of the above data are also available in plain CSV format [66].

## ACKNOWLEDGMENTS

We gratefully acknowledge the support for the High Throughput (HT) generation of the database (M.L., V.I.H., C.W.) by the Toyota Research Institute through the Accelerated Materials Design and Discovery program and for the data analysis work (A.G., J.H.) by the U.S. Department of Commerce, National Institute of Standards and Technology, as part of the Center for Hierarchical Materials Design (CHiMaD) Award No. 70NANB14H012. The high-throughput calculations performed in this work made use of resources in the National Energy Research Scientific Computing Center (NERSC) and ASCR Leadership Computing Challenge (ALCC) award. The statistical learning calculations mentioned in this work made use of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation, Grant No. ACI-1548562. Specifically, it used the Bridges system, which is supported by National Science Foundation Grant No. ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

- 
- [1] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Phys. Rev.* **136**, B864 (1964); W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, *ibid.* **140**, A1133 (1965).
- [2] N. Marzari, A. Ferretti, and C. Wolverton, Electronic-structure methods for materials design, *Nat. Mater.* **20**, 736 (2021).
- [3] C. Stefano, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, and M. Buongiorno-Nardelli, AFLOWLIB.ORG: A distributed materials properties repository from high-throughput *ab initio* calculations, *Comput. Mater. Sci.* **58**, 227 (2012).
- [4] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD), *JOM* **65**, 1501 (2013).
- [5] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, A high-throughput infrastructure for density functional theory calculations, *Comput. Mater. Sci.* **50**, 2295 (2011).
- [6] L. Talirz *et al.*, Materials Cloud, a platform for open computational science, *Sci. Data* **7**, 299 (2020).
- [7] C. Draxl and M. Scheffler, The NOMAD laboratory: From data sharing to artificial intelligence, *J. Phys.: Mater.* **2**, 036001 (2019).
- [8] K. Choudhary *et al.*, The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design, *npj Comput. Mater.* **6**, 1 (2020).
- [9] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, The high-throughput highway to computational materials design, *Nat. Mater.* **12**, 191 (2013).
- [10] G. Brunin, F. Ricci, V.-A. Ha, G.-M. Rignanese, and G. Hautier, Transparent conducting materials discovery using high-throughput computing, *npj Comput. Mater.* **5**, 63 (2019).
- [11] J.-P. Correa-Baena, K. Hippalgaonkar, J. van Duren, S. Jaffer, V. R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha, and T. Buonassisi, Accelerating materials development via automation, machine learning, and high-performance computing, *Joule* **2**, 1410 (2018).
- [12] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [13] F. Tran and P. Blaha, Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential, *Phys. Rev. Lett.* **102**, 226401 (2009).
- [14] J. Heyd, G. E. Scuseria, and M. Ernzerhof, Hybrid functionals based on a screened Coulomb potential, *J. Chem. Phys.* **118**, 8207 (2003).
- [15] A. V. Krukau, O. A. Vydrov, A. F. Izmaylov, and G. E. Scuseria, Influence of the exchange screening parameter on the performance of screened hybrid functionals, *J. Chem. Phys.* **125**, 224106 (2006).
- [16] M. K. Y. Chan and G. Ceder, Efficient band gap prediction for solids, *Phys. Rev. Lett.* **105**, 196403 (2010).
- [17] G. Pilania, J. E. Gubernatis, and T. Lookman, Multi-fidelity machine learning models for accurate bandgap predictions of solids, *Comput. Mater. Sci.* **129**, 156 (2017).
- [18] C. Franchini, R. Podloucky, J. Paier, M. Marsman, and G. Kresse, Ground-state properties of multivalent manganese oxides: Density functional and hybrid density functional calculations, *Phys. Rev. B* **75**, 195128 (2007).
- [19] L.-F. Huang and J. M. Rondinelli, Electrochemical phase diagrams for Ti oxides from density functional calculations, *Phys. Rev. B* **92**, 245126 (2015).
- [20] V. L. Chevrier, S. P. Ong, R. Armiento, M. K. Chan, and G. Ceder, Hybrid density functional calculations of redox potentials and formation energies of transition metal compounds, *Phys. Rev. B* **82**, 075122 (2010).
- [21] Y. Zhang, G. Kresse, and C. Wolverton, Nonlocal first-principles calculations in Cu-Au and other intermetallic alloys, *Phys. Rev. Lett.* **112**, 075502 (2014).
- [22] F. Zhang, J. Gale, B. Uberuaga, C. Stanek, and N. Marks, Importance of dispersion in density functional calculations of

- cesium chloride and its related halides, *Phys. Rev. B* **88**, 054112 (2013).
- [23] R. Batra, G. Pilania, B. P. Uberuaga, and R. Ramprasad, Multifidelity information fusion with machine learning: A case study of dopant formation energies in hafnia, *ACS Appl. Mater. Interfaces* **11**, 24906 (2019).
- [24] P. Perdikaris, D. Venturi, J. O. Royset, and G. E. Karniadakis, Multi-fidelity modelling via recursive co-kriging and Gaussian-Markov random fields, *Proc. R. Soc. A* **471**, 20150018 (2015).
- [25] C.-W. Lee, K.-W. Lee, and J.-S. Lee, Optoelectronic properties of  $\beta$ -Fe<sub>2</sub>O<sub>3</sub> hollow nanoparticles, *Mater. Lett.* **62**, 2664 (2008).
- [26] Open Quantum Materials Database, <https://oqmd.org>.
- [27] P. E. Blöchl, Projector augmented-wave method, *Phys. Rev. B* **50**, 17953 (1994).
- [28] G. Kresse and J. Furthmüller, Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set, *Comput. Mater. Sci.* **6**, 15 (1996).
- [29] P. Wisesa, K. A. McGill, and T. Mueller, Efficient generation of generalized Monkhorst-Pack grids through the use of informatics, *Phys. Rev. B* **93**, 155109 (2016).
- [30] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies, *npj Comput. Mater.* **1**, 15010 (2015).
- [31] M. C. Kennedy and A. O'Hagan, Predicting the output from a complex computer code when fast approximations are available, *Biometrika* **87**, 1 (2000).
- [32] A. I. J. Forrester, A. Söbester, and A. J. Keane, Multi-fidelity optimization via surrogate modelling, *Proc. R. Soc. A* **463**, 3251 (2007).
- [33] L. Le Gratiet and J. Garnier, Recursive co-kriging model for design of computer experiments with multiple levels of fidelity, *Int. J. Uncertainty Quantif.* **4**, 365 (2014).
- [34] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, *npj Comput. Mater.* **2**, 16028 (2016).
- [35] I. J. Myung, Tutorial on maximum likelihood estimation, *J. Math. Psychol.* **47**, 90 (2003).
- [36] J. S. Gray, J. T. Hwang, J. R. R. A. Martins, K. T. Moore, and B. A. Naylor, OpenMDAO: An open-source framework for multidisciplinary design, analysis, and optimization, *Struct. Multidiscip. Optim.* **59**, 1075 (2019).
- [37] N. N. Kiselyova, V. A. Dudarev, and M. A. Korzhuyev, Database on the bandgap of inorganic substances and materials, *Inorganic Materials: Appl. Res.* **7**, 34 (2016).
- [38] S. G. T. Europe, *Thermodynamic Properties of Inorganic Materials: Thermodynamic Properties of Inorganic Materials* (Springer, 1999).
- [39] G. Kim, S. V. Meschel, P. Nash, and W. Chen, Experimental formation enthalpies for intermetallic phases and other inorganic compounds, *Scientific Data* **4**, 170162 (2017).
- [40] E. B. Isaacs and C. Wolverton, Performance of the strongly constrained and appropriately normed density functional for solid-state materials, *Phys. Rev. Mater.* **2**, 063801 (2018).
- [41] V. Stevanović, S. Lany, X. Zhang, and A. Zunger, Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies, *Phys. Rev. B* **85**, 115104 (2012).
- [42] S. Grindy, B. Meredig, S. Kirklin, J. E. Saal, and C. Wolverton, Approaching chemical accuracy with density functional calculations: Diatomic energy corrections, *Phys. Rev. B* **87**, 075150 (2013).
- [43] J. Honig and T. Reed, Electrical properties of Ti<sub>2</sub>O<sub>3</sub> single crystals, *Phys. Rev.* **174**, 1020 (1968).
- [44] S.-I. Ohkoshi, Y. Tsunobuchi, T. Matsuda, K. Hashimoto, A. Namai, F. Hakoe, and H. Tokoro, Synthesis of a metal oxide with a room-temperature photoreversible phase transition, *Nat. Chem.* **2**, 539 (2010).
- [45] J. Qiu, W. Zheng, R. Yuan, C. Yue, D. Li, F. Liu, and J. Zhu, A novel 3D nanofibrous aerogel-based MoS<sub>2</sub>@Co<sub>3</sub>S<sub>4</sub> heterojunction photocatalyst for water remediation and hydrogen evolution under simulated solar irradiation, *Appl. Catal., B* **264**, 118514 (2020).
- [46] A. Ginsburg, D. A. Keller, H.-N. Barad, K. Rietwyk, Y. Bouhadana, A. Anderson, and A. Zaban, One-step synthesis of crystalline Mn<sub>2</sub>O<sub>3</sub> thin film by ultrasonic spray pyrolysis, *Thin Solid Films* **615**, 261 (2016).
- [47] A. Bocquet, T. Mizokawa, K. Morikawa, A. Fujimori, S. Barman, K. Maiti, D. Sarma, Y. Tokura, and M. Onoda, Electronic structure of early 3d-transition-metal oxides by analysis of the 2p core-level photoemission spectra, *Phys. Rev. B* **53**, 1161 (1996).
- [48] S. Mattsson and B. Paulus, Density functional theory calculations of structural, electronic, and magnetic properties of the 3d metal trifluorides MF<sub>3</sub> (M= Ti-Ni) in the solid state, *J. Comput. Chem.* **40**, 1190 (2019).
- [49] O. O. Balayeva, A. A. Azizov, M. B. Muradov, A. M. Maharramov, G. M. Eyvazova, R. M. Alosmanov, Z. Q. Mamiyev, and Z. A. Aghamaliyev,  $\beta$ -NiS and Ni<sub>3</sub>S<sub>4</sub> nanostructures: Fabrication and characterization, *Mater. Res. Bull.* **75**, 155 (2016).
- [50] I. T. Sines, R. Misra, P. Schiffer, and R. E. Schaak, Colloidal synthesis of non-equilibrium Wurtzite-Type MnSe, *Angew. Chem., Int. Ed.* **49**, 4638 (2010).
- [51] L. G. J. de Haart, A. De Vries, and G. Blasse, Photoelectrochemical properties of MgTiO<sub>3</sub> and other titanates with the ilmenite structure, *Mater. Res. Bull.* **19**, 817 (1984).
- [52] C. Lokhande, A. Ennaoui, P. Patil, M. Giersig, M. Muller, K. Diesner, and H. Tributsch, Process and characterisation of chemical bath deposited manganese sulphide (MnS) thin films, *Thin Solid Films* **330**, 70 (1998).
- [53] M. Ikeda, K. Itoh, and H. Sato, Electrical and optical properties of CdS-MnS single crystals, *J. Phys. Soc. Jpn.* **25**, 455 (1968).
- [54] L. Gnanasekaran, R. Hemamalini, R. Saravanan, K. Ravichandran, F. Gracia, S. Agarwal, and V. K. Gupta, Synthesis and characterization of metal oxides (CeO<sub>2</sub>, CuO, NiO, Mn<sub>3</sub>O<sub>4</sub>, SnO<sub>2</sub> and ZnO) nanoparticles as photo catalysts for degradation of textile dyes, *J. Photochem. Photobiol., B* **173**, 43 (2017).
- [55] H. Kalt, *General properties, in Optical Properties. Part 2, Landolt-Börnstein - Group III Condensed Matter Vol. 34C2, Springer Materials* (Springer, Berlin, 2004), .

- [56] J. Kang, A. Hirata, L. Kang, X. Zhang, Y. Hou, L. Chen, C. Li, T. Fujita, K. Akagi, and M. Chen, Enhanced supercapacitor performance of  $\text{MnO}_2$  by atomic doping, *Angew. Chem., Int. Ed.* **52**, 1664 (2013).
- [57] M. J. Young, A. M. Holder, S. M. George, and C. B. Musgrave, Charge storage in cation incorporated  $\alpha - \text{MnO}_2$ , *Chem. Mater.* **27**, 1172 (2015).
- [58] M. Lorenz, G. Pettit, and R. Taylor, Band gap of gallium phosphide from 0 to 900 K and light emission from diodes at high temperatures, *Phys. Rev.* **171**, 876 (1968).
- [59] I. Catalano, A. Cingolani, and A. Minafra, Multiphoton transitions at the direct and indirect band gaps of gallium phosphide, *Solid State Commun.* **16**, 417 (1975).
- [60] R. Asahi, T. Morikawa, T. Ohwaki, K. Aoki, and Y. Taga, Visible-light photocatalysis in nitrogen-doped titanium oxides, *Science* **293**, 269 (2001).
- [61] O. Polat, Z. Durmus, F. Coskun, M. Coskun, and A. Turut, Engineering the band gap of  $\text{LaCrO}_3$  doping with transition metals (Co, Pd, and Ir), *J. Mater. Sci.* **53**, 3544 (2018).
- [62] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, Accelerated search for materials with targeted properties by adaptive design, *Nat. Commun.* **7**, 11241 (2016).
- [63] A. M. Gopakumar, P. V. Balachandran, D. Xue, J. E. Gubernatis, and T. Lookman, Multi-objective optimization for materials discovery via adaptive design, *Sci. Rep.* **8**, 3738 (2018).
- [64] A. Gopakumar, K. Pal, and C. Wolverton, Identification of high-dielectric constant compounds from statistical design, *npj Comput. Mater.* **8**, 146 (2022).
- [65] OQMD+, <https://hse.oqmd.org>.
- [66] [https://static.oqmd.org/subdatasets/hse\\_mohan/](https://static.oqmd.org/subdatasets/hse_mohan/).