Computational Materials Science 129 (2017) 156–163

Contents lists available at ScienceDirect

Computational Materials Science

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

Editor's Choice

Multi-fidelity machine learning models for accurate bandgap predictions of solids

G. Pilania [a,*], J.E. Gubernatis [b], T. Lookman [b]

[a] Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[b] Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87544, USA

ARTICLE INFO

Article history:
Received 8 November 2016
Received in revised form 30 November 2016
Accepted 1 December 2016

Keywords:
Double perovskites
Elpasolites
Materials informatics
Information fusion

ABSTRACT

We present a multi-fidelity co-kriging statistical learning framework that combines variable-fidelity quantum mechanical calculations of bandgaps to generate a machine-learned model that enables low-cost accurate predictions of the bandgaps at the highest fidelity level. In addition, the adopted Gaussian process regression formulation allows us to predict the underlying uncertainties as a measure of our confidence in the predictions. Using a set of 600 elpasolite compounds as an example dataset and using semi-local and hybrid exchange correlation functionals within density functional theory as two levels of fidelities, we demonstrate the excellent learning performance of the method against actual high fidelity quantum mechanical calculations of the bandgaps. The presented statistical learning method is not restricted to bandgaps or electronic structure methods and extends the utility of high throughput property predictions in a significant way.

1. Introduction

Owing to its central role in modern device physics, energy harvesting, energy storage, catalysis and other technologically relevant applications [1], the bandgap often serves as a crucial screening parameter in rational design of functional materials [2–4]. As available experimental data for bandgaps is generally limited [5,6], it is not surprising that a number of recent high throughput chemical space explorations, in search of compounds with improved functionalities, have targeted the calculation of the bandgap [7–16].

Given that accurate calculations of bandgaps are time consuming and resource intensive, we demonstrate in this paper the use of a statistical learning method that generates machine-learned models to obviate the costs of such calculations. The multi-fidelity method presented combines many inexpensive lower accuracy computations of bandgaps with fewer expensive higher accuracy computations to predict bandgaps whose accuracies are comparable to those produced by the higher accuracy calculations alone. The greater the difference in the costs of the calculations, the greater is the cost advantage of the method we demonstrate. Because a natural and well documented accuracy hierarchy exists for bandgap calculation methods, bandgap computations are natural for the method presented. In principle, the method can be applied to the predictions of computed or experimental properties where the data is grouped into different levels of accuracy.

Our approach takes available bandgap prediction methods from different levels of theories with different fidelities to estimate bandgaps at the fidelity level of the more accurate and computationally expensive theory. On one end of the fidelity spectrum, we may have surrogate models, perhaps obtained via high throughput computations, but less trustworthy or known to be inaccurate. On the other end, we may have high-fidelity models that enable quite accurate estimates. In the present context of quantum mechanical computations of bandgaps, the variable fidelity can be thought of as bandgaps computed from different levels of exchange correlation functionals within DFT [17–20], as creatively captured in *Jacob's ladder of density functional approximations* put forward by Perdew (Fig. 1) [21] as well as expressed by the spectrum of beyond-DFT approaches such as the GW method [22,23], Møller-Plesset perturbation theory (MP2) [24], and configuration interaction (CI) [25].

We use the statistical learning approach of multi-fidelity co-kriging on a data set of 640 double perovskite halide compounds for which the bandgap energies can be calculated with two levels of fidelity (the lower fidelity Perdew-Burke-Ernzerhoff (PBE) [19] and higher fidelity Heyd-Scuseria-Ernzerhof (HSE06) [17] exchange-correlation functional approximations). We make predictions of the HSE06 bangap energies by considering different

* Corresponding author.
E-mail address: gpilania@lanl.gov (G. Pilania).

http://dx.doi.org/10.1016/j.commatsci.2016.12.004
0927-0256/

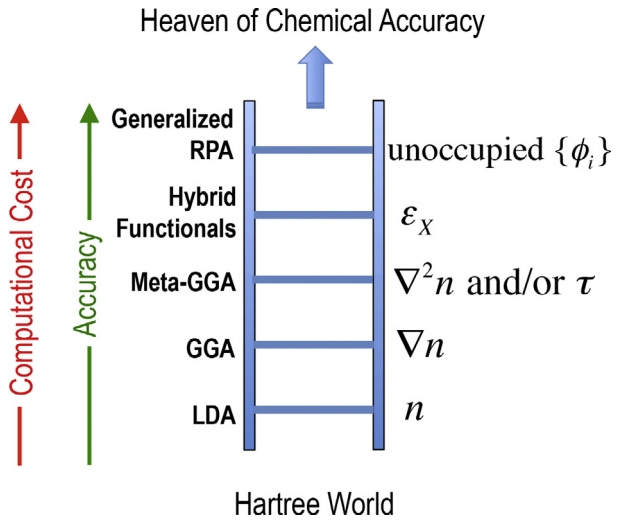

Fig. 1. Jacob's ladder of density functional approximations to the exchange-correlation energy (as put forward by Perdew, adapted from Ref. [21]) presents a prototypical example of multi-fidelity computations of materials properties, where a natural hierarchy exists in both the computational cost and accuracy. The present study demonstrates the effectiveness of multi-fidelity ML approach for a bandgap dataset of elpasolite compounds computed at the GGA and the hybrid functional levels. Symbols n, ϵ_x and ϕ_i in the figure represent the ground state charge density, the exact exchange and Kohn–Sham orbitals, respectively.

numbers of the PBE bandgaps in the combined data in which the number of high fidelity bandgaps is a subset of the low fidelity ones. One of our central conclusions is not surprisingly that the accuracy of the prediction increases as the number of high fidelity bandgaps increases in the training set. A second conclusion is, more importantly, that the accuracy of the predictions increases as the number of low fidelity data increases in the training set. We deduce these conclusions from heat maps of the mean square error in the HSE06 predictions, made by our trained learning model on unseen data, as a function of the number of low fidelity data points and the relative proportion of high fidelity data used.

Our approach is markedly different from other multi-fidelity approaches in the literature which are based on using low fidelity data (*e.g.*, PBE bandgaps) as *features* in the machine learning (ML) model [26] and therefore strictly require low fidelity data for all materials for which predictions are to be made using the trained model. This can be particularly challenging and extremely computationally demanding when faced with a combinatorial problem that targets exploring a vast chemical and configurational space. Here, we present a framework for a multi-fidelity Gaussian process (GP) based ML regression model that seamlessly combines bandgap inputs from two or more levels of fidelities to make accurate predictions of the bandgaps for the highest fidelity. Furthermore, adopting a nested setting for variable-fidelity training data, the model requires high-fidelity training data only on a subset of compounds for which low-fidelity training data is readily available. More importantly, the trained model can make efficient yet accurate predictions for the highest-fidelity bandgaps even in the absence of the low-fidelity bandgap data for the prediction set compounds. In addition to the bandgap predictions obtained with lower cost, the adopted GP-regression framework also allows us to predict the underlying uncertainties as a measure of our confidence in the predictions.

The outline of our paper is as follows. In Section 2 we review multi-fidelity approaches (as they have been studied in the engineering context) and DFT methods. We also introduce and discuss the co-kriging approach which we utilize in this study. This introduction is followed by a description of the data set of double perovskites and the features used to the train the Gaussian model used in co-kriging. We also provide the computational details underpinning our DFT calculations. Section 3 presents and discusses the results of our calculations as we show how our predictions of the bandgaps are successively refined depending on the data sizes of the low and high fidelity data. Finally, Section 4 concludes and discusses the significance and implications of our work.

2. Background and methodology

2.1. Motivation and past work

Fueled by recent advances in methodologies and computational power, density functional theory (DFT) has become a standard workhorse for *ab initio* electronic structure calculations, providing the best trade-off between predictive accuracy and computational efficiency [27]. While the standard implementations of DFT are widely employed to compute structural, electronic, electrical, magnetic and other properties of plethora of materials, they suffer from a well known deficiency (also known as the "bandgap problem") [28] in which DFT within local or semi-local exchange-correlation functionals fails to correctly predict the energy gaps between occupied and unoccupied states. In fact, the experimental bandgap ϵ_{exp} is often severely underestimated by the Kohn-Sham gap ϵ_{KS}. This underestimation is attributed to the inherent lack of derivative discontinuity [29] and delocalization error [30,31] within the local or semi-local exchange-correlation functionals such as the local density approximation (LDA) or the generalized gradient approximation (GGA) [32].

For the exact Kohn-Sham formalism, the physical gap equals ϵ_{KS} plus the derivative discontinuity of the exchange-correlation (xc) energy with respect to the number of electrons [29,33,34]. In a non-exact KS approach with an approximate xc energy functional, the above relation may not be exact and, in addition, the approximate functional may not reproduce the correct xc energy derivatives [30,31]. Correction techniques such as the DFT + U can improve the KS gap only to a limited extent [35]. Recently developed more advanced exchange-correlation functionals, such as the modified Becke-Johnson (mBJ) functional by Tran and Blaha [36] and strongly constrained and appropriately normed (SCAN) [37] meta-GGA, certainly improve over the classic LDA and GGA functionals, but do not completely alleviate the problem. Several other alternative approaches are frequently employed to address the bandgap problem, including the delta self-consistent-field (Δ-SCF) method [38], hybrid functional methods [17] and the quasi-particle GW calculations based on the many-body perturbation theory [39]. The Δ-SCF formalism requires the evaluation of the total energy at three different numbers of electrons (N, $N + \delta$, and $N - \delta$), N being the number of electrons in the neutral ground state. δ – representing the amount of charge added or removed to simulate the excitation process – depends on N and is determined empirically to best yield the experimental bandgap. While the Δ-SCF method is not entirely parameter-free, the latter two approaches – although capable of providing bandgap estimates in good agreement with the corresponding experimental measurements – suffer from high computational costs.

ML methods have recently had phenomenal success in condensed matter physics and materials science for high throughput screening and accelerated property predictions [40–42]. Even with a limited set of prior data to train on, ML-based approaches have been successfully employed, for instance, to accurately estimate a wide range of material properties for molecular [43,44] and periodic systems [7,45], to develop adaptive force fields [46,47], to devise schemes for crystal structure classifications [48–50], to predict dielectric breakdown strength of insulators [51,52] and to enable alternative self-consistent solutions for quantum mechanics [53].