Computational Materials Science 172 (2020) 109286

Contents lists available at ScienceDirect

Computational Materials Science

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

A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap

Abhirup Patra[1], Rohit Batra[1], Anand Chandrasekaran, Chiho Kim, Tran Doan Huan, Rampi Ramprasad*

School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, GA 30332, USA

ARTICLE INFO

Keywords:
Density functional theory
Machine learning
Multi-fidelity learning
Polymers
Bandgap

ABSTRACT

The fidelity of data is of paramount importance in the construction of reliable and accurate machine learning (ML) models. Low-fidelity data, although noisy, can usually be obtained for a large number of material samples. High-fidelity data, on the other hand, is time-consuming and oftentimes, only available for a limited number of target samples. While the former can provide useful information to help generalize the ML models over large materials space, the latter is useful to build more accurate surrogate models. Information fusion schemes that utilize the data available at multiple levels of fidelity can outperform traditional single fidelity based ML methods, such as Gaussian process regression. In this work, a variant of the multi-fidelity information fusion scheme, namely multi-fidelity co-kriging, is used to build powerful prediction models of polymer bandgaps. To benchmark this strategy, we utilize a bandgap dataset of 382 polymers, obtained at two levels of fidelity: using the Perdew-Burke-Ernzerhof (PBE) exchange-correlational functional (“low-fidelity”) and the Heyd-Scuseria-Ernzerhof (HSE06) functional(“high-fidelity”) of density functional theory. The multi-fidelity model, trained on both PBE and HSE06 data, outperforms a single-fidelity Gaussian process regression model trained on just HSE06 band-gaps in a number of scenarios and is also able to generalize better to a more diverse chemical space.

1. Introduction

Machine-learning (ML) [1] has emerged as an important tool [2–4] to solve many materials science and engineering problems, including classifying materials phases [5], predicting material properties 6–8, uncovering possible hidden structure-property or property-property correlations [9], or suggesting potential materials synthesis routes [10,11]. A critical requirement common across all such ML methods is the availability of reliable and accurate data to “train” these predictive models. Referred to as the training set, this data is obtained from either physics-based computations or collected from past experiments.

Since the accuracy of a ML model is bounded by the quality of the underlying training data, it is highly desired that such data is prepared at the highest possible level of fidelity (or accuracy). However, in most scenarios, especially in materials science, the available data is quite diverse in terms of fidelity. For example, for a given material property, data produced using different computational techniques and/or experiments often differ in the level of quality. Further, higher the quality of a data source, higher is the cost associated with the measurement, which in turn, severely restricts the availability of large volumes of high-fidelity data. Computation of the bandgap (E_g) of insulator is a classic example. On the one hand, E_g computed using the relatively inexpensive Perdew-Burke-Ernzerhof (PBE) [12] exchange-correlation functional of density functional theory (DFT) are typically significantly underestimated relative to experiments. On the other hand, E_g computed using the Heyd-Scuseria-Ernzerhof (HSE06) [13] functional provides a more accurate estimate, although at a much higher computational cost. Consequently, for a large number of materials, PBE E_g values have been computed, while high-fidelity HSE06 E_g for only a relatively small fraction of materials is known. Thus, within this context, restricting oneself to building ML models based only on limited high-fidelity data can have serious limitations in terms of generalizability of the model. On the other hand, ML models developed using low-fidelity data (although plentiful) will be limited in accuracy. In fact, a multi-fidelity information scheme that utilizes information available at different levels of fidelity could be a more optimal way to build predictive surrogate models.

Several information fusion schemes have been introduced to

* Corresponding author.
E-mail address: rampi.ramprasad@mse.gatech.edu (R. Ramprasad).

[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.commatsci.2019.109286
Received 4 June 2019; Received in revised form 12 September 2019; Accepted 14 September 2019

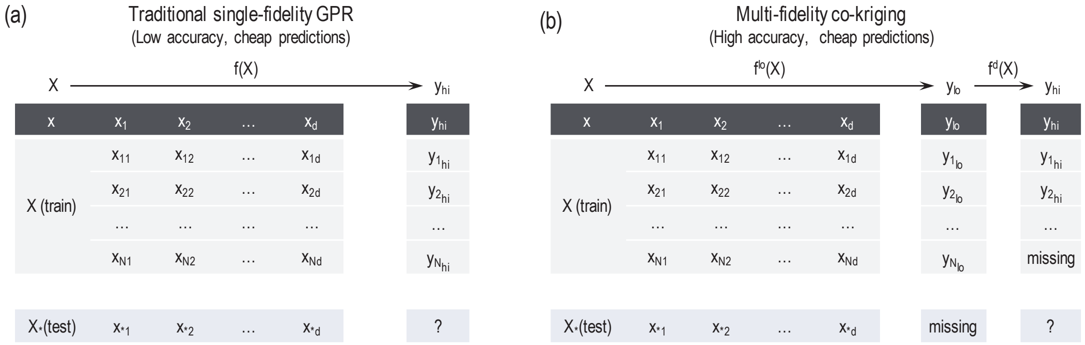
(a)
Traditional single-fidelity GPR
(Low accuracy, cheap predictions)
X
f(X)
y_{hi}
x
x_1 x_2 … x_d
y_{hi}
X (train)
x_{11} x_{12} … x_{1d}
x_{21} x_{22} … x_{2d}
… … … …
x_{N1} x_{N2} … x_{Nd}
$y_{1_{hi}}$
$y_{2_{hi}}$
…
$y_{N_{hi}}$
X_*(test)
x_{*1} x_{*2} … x_{*d}
?
(b)
Multi-fidelity co-kriging
(High accuracy, cheap predictions)
X
$f^{lo}(X)$
y_{lo}
$f^d(X)$
y_{hi}
x
x_1 x_2 … x_d
y_{lo}
y_{hi}
X (train)
x_{11} x_{12} … x_{1d}
x_{21} x_{22} … x_{2d}
… … … …
x_{N1} x_{N2} … x_{Nd}
$y_{1_{lo}}$
$y_{2_{lo}}$
…
$y_{N_{lo}}$
$y_{1_{hi}}$
$y_{2_{hi}}$
…
missing
X_*(test)
x_{*1} x_{*2} … x_{*d}
missing
?

Fig. 1. Comparison of the datasets and mapping functions learned using the SF and the MF approach. The SF maps the fingerprint/feature space to the targeted HSE bandgap (y_{hi}) via $f(\mathbf{x}) \to y_{\text{hi}}$. On the other hand, the MF model utilizes the same fingerprint space to predict low-fidelity bandgap (y_{lo}) via $f^{\text{lo}}(\mathbf{x}) \to y_{\text{lo}}$ and the difference between these two fidelities using the $f^d(\mathbf{x}) \to \Delta = y_{\text{hi}} - y_{\text{lo}}$ mapping relation.

combine knowledge available at multiple levels of fidelity for a common target property of interest [14]. These include, Δ-learning (which uses low-fidelity data as a fingerprint/feature) and the multi-fidelity co-kriging (MF) approach. Recently, Batra and co-workers [14] showed the superiority of the MF approach over other methods, especially when the size of high-fidelity dataset is particularly limited. Fig. 1 illustrates the difference between the traditional single-fidelity (SF) based learning and the MF approach. In the SF scheme, only the target property obtained at the highest level of fidelity (*i. e.*, y_{hi}) is used to learn a mapping from the material fingerprint ($\mathbf{x}$). On the other hand, in the MF approach, property values obtained from two (or more) levels of fidelity, i.e., y_{lo} and y_{hi}, are used simultaneously to construct the ML model. It is important to note that the MF scheme also learns from instances where just the low-fidelity value is known, and the corresponding high-fidelity value is unknown – one such case is marked as “missing” in the column y_{hi} in Fig. 1(b). Owing to its advantages over traditional SF scheme, the MF method has been recently employed in some chemical or materials science problems. For example, Pilania et al. [15] used the MF framework to learn bandgap in perovskite materials obtained at two levels of fidelity using different DFT functionals.

In this work, we explore the utilization of the MF strategy in polymer informatics. We previously demonstrated [16,17] the power of the SF scheme in learning from computed [18] and measured data for several properties of polymers, e.g., bandgap E_g, dielectric constant ϵ, glass transition temperature T_g, etc. culminating in a polymer informatics platform named *Polymer Genome* (PG). Several polymer designs have emerged from this effort [19–21]. Since the chemical or configurational space spanned by polymers is enormous, a MF learning approach is quite appropriate wherein large regions of the space are explored at a low-fidelity level, while a few interesting cases are explored at a higher-fidelity level (the above example of bandgap computation is particularly valid for the case of polymers). Advanced approached such as these can significantly expand and extend recent polymer discovery efforts [22–24]. Our results clearly show that utilizing less accurate PBE bandgap in conjunction with only a few HSE06 bandgap values to train the MF model resulted in much better learning performance when compared to the Gaussian Process Regression (GPR) models trained using just the high-fidelity HSE06 bandgap data. Thus, the MF approach can clearly help to accelerate the polymer (or materials) discovery process by effective use of limited resources aimed at large number of exploratory and cheap low-fidelity points, along with only a few expensive, but accurate, high-fidelity values.

This paper is organized as follows. In the “Methodology” section we describe the dataset utilized in this work, and provide theoretical details on its generation. The SF and the MF methods are discussed next in the same section. In “Results & Discussions” we benchmark the performance of the SF and MF approaches in a number of different scenarios. Finally, in the “Conclusion” section we summarize the insights gained in this work and provide a perspective of how such MF approaches could be of relevance in other aspects of the rational-design of polymeric materials.

2. Methodology

2.1. Dataset

An important property of a polymer that describes its usefulness as an optical, electronic or energy storage material is its electronic bandgap. Thus, we considered a DFT computed bandgap dataset of 382 polymers, calculated using both PBE and HSE06 functionals [18]. Details of how these polymer crystal structures were constructed and their properties were computed can be found in earlier works [18,23,25]. Fig. 2 shows the correlation between the low-fidelity PBE bandgaps (E_g^{PBE}) and the high-fidelity HSE06 bandgaps (E_g^{HSE}). As expected, the PBE bandgaps are underestimated compared to those calculated using the HSE06 functional but this underestimation is not strictly linear. Further, the dataset can be seen to span a fairly large range of bandgap values (0.75–10 eV), containing polymers with very high bandgap, such as Polyoxymethylene ($[\text{-CH}_2\text{-O-}]_n$) and Poly-2,3,3-Trifluoroacryloyl fluoride ($[\text{-C}_3\text{F}_4\text{O-}]_n$) with bandgap of 9 and 10 eV, respectively, and polymers like Polythiophene ($[\text{-C}_4\text{H}_2\text{S-}]_n$) with a bandgap of just 0.75 eV. In terms of chemical diversity, the 382 polymers are composed of six atomic species: C, H, O, N, S, and F, and the following building-blocks: CH_2, CO, CS, NH, C_6H_4, C_4H_2S, CF_2, CHF and O.

In order to better understand the merits of the MF approach, we first divide our entire dataset of 382 polymers, denoted as D382, into two categories, D351 and D31 consisting of 351 and 31 polymers, respectively. This particular choice is based on the number of building blocks present in the polymers. Polymers in D351 comprise less than or equal to six building blocks, while the polymers in D31 have more than six building blocks in their repeat unit. Further, SF and MF models constructed using the D351 dataset were evaluated on the D31 dataset, thereby providing a unique opportunity to test our models on not only completely unseen data, but also on cases which are expensive to compute using DFT [26] computations. As illustrated in Fig. 2b, within each of the D351 and D31 data sets, both PBE and HSE06 E_g values were available, resulting in four subsets, namely, $\text{D351}^{\text{HiFi}}$, $\text{D351}^{\text{LoFi}}$, D31^{HiFi} and D31^{LoFi}. These subsets of data were further utilized to demonstrate the superiority of the MF approach under different scenarios.

2.2. Fingerprinting

In order to establish a ML-based mapping between the polymer and its bandgap, a numerical representation or fingerprint of the polymer is required, represented as a *d*-dimensional vector $\mathbf{x}$ shown in Fig. 1. In

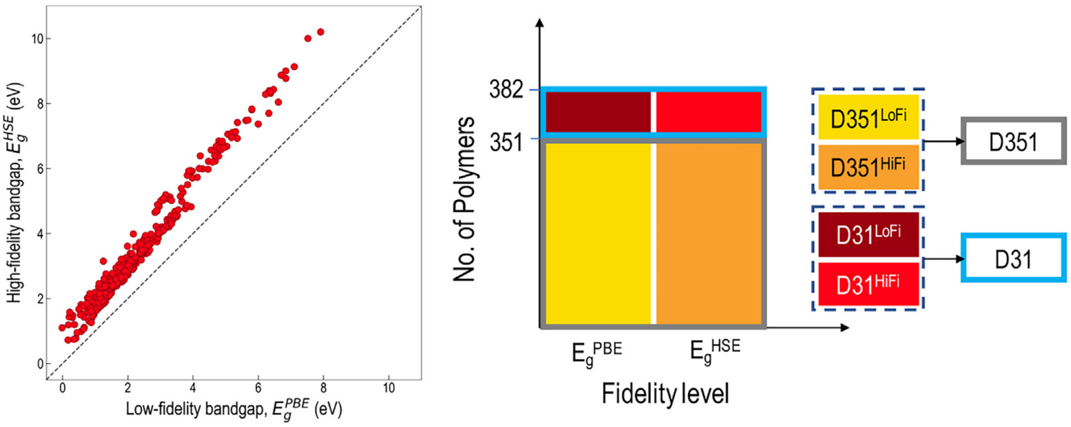

Fig. 2. (a) The correlation between the two levels of fidelities (PBE and HSE06) present in the bandgap dataset of 382 polymers. (b) Division of the dataset into different subsets based on the fidelity of the data and the length of the polymer chains. Here, D351$^{\text{HiFi}}$ and D351$^{\text{LoFi}}$ denote two division of D351 dataset based on HSE06 and PBE bandgap values respectively. Likewise, D31 dataset is also divided into D31$^{\text{HiFi}}$ and D31$^{\text{LoFi}}$ datasets with HSE06 and PBE bandgap values of 31 polymers.

accordance with our previous work [17], a hierarchical three-level fingerprinting scheme—composed of atomic [22], quantitative structure-property relationship (QSPR) and morphological components—was used to represent the chemical and morphological information of the polymers. In total, a 187-dimensional fingerprint was used to fingerprint every polymer in this dataset. However, it is helpful to reduce the dimensionality of fingerprints to improve the performance of our model. To do so, we used the recursive feature elimination (RFE) algorithm [27] and removed those fingerprint components which possessed a minimal contribution to the overall prediction accuracy of the targeted property (HSE band gap in our case). This reduced the dimensionality of our fingerprint to 116 and this "filtered" version of the fingerprint was used to train both the SF and the MF models. Details of this procedure has been described in detail elsewhere [17].

2.3. Machine learning models

2.3.1. Single-fidelity model

To serve as a benchmark method against the MF model, we used the commonly employed SF based GPR algorithm. This is also the ML technique using which the current predictive models of Polymer Genome have been built. GPR, also known as *kriging*, is a widely used regression technique in applied sciences, including materials science [28,29]. It is a kernel based ML algorithm that uses the Bayesian statistical framework. The GPR model is built by fitting a Gaussian process to the training data, which is then used to obtain a distribution at any new point. In this work, the radial basis function (RBF) was used as the covariance function. The hyper parameters of the RBF kernel were determined by maximizing the log-likelihood function using the training data and 5-fold cross validation. It is important to note that, GPR or kriging is a SF model, which was trained using only the high-fidelity HSE06 bandgap data as the target property. Thus, we refer to models constructed using this scheme as SF GPR models.

2.3.2. Multi-fidelity co-kriging model

First proposed by Kennedy and O'Hagan [30], MF is a natural extension of the kriging method with data available at multiple levels of fidelity [31]. As shown in Fig. 1 (b), the high-fidelity prediction in the MF approach not only depends on the low-fidelity function ($f^{\text{lo}}(\mathbf{x})$) but also on the difference between the low-fidelity and the high-fidelity functions ($f^{\text{d}}(\mathbf{x})$). In the two-level MF approach, this is achieved by expressing the overall ML model as $\mathbf{Z}_{\text{hi}}(\mathbf{x}) = \rho\mathbf{Z}_{\text{lo}}(\mathbf{x}) + \mathbf{Z}_{\text{d}}(\mathbf{x})$, where, $\mathbf{Z}_{\text{lo}}$ is the low-fidelity estimation and $\mathbf{Z}_{\text{d}}$ is the Gaussian process related to the difference between the two fidelities of data. In the context of this work, $\mathbf{Z}_{\text{lo}}(.)$ and $\mathbf{Z}_{\text{hi}}(.)$ are the Gaussian processes representing the low-fidelity (PBE bandgap) and the high-fidelity (HSE06 bandgap) datasets for the polymers under consideration. However, the MF approach assumes that for the polymers for which HSE06 bandgaps (y_{hi}) are available, their respective PBE bandgaps (y_{lo}) are also known. Nonetheless, it is important to note that when the prediction for a new case is to be made, only the feature vector $\mathbf{x}$ is required.

Table 1 shows the various GPR and MF models that were constructed using different subsets of the polymer bandgap dataset. The philosophy behind creating a series of such ML models is to evaluate the performance of the SF and the MF approaches under different scenarios of systematically increasing low as well as high-fidelity data. For both the SF GPR and MF approaches, 5-fold cross-validation was adopted to ensure that the models do not over-fit the training data. The performance of the two approaches was compared in terms of the root mean square error (RMSE) and the correlation coefficient evaluated on different test sets, as explained in the next section.

3. Results and discussions

Before bench-marking the models described in the earlier section, we first sought to comprehensively study the performance of the MF method with respect to variation in the number of low-fidelity (N_{lo}) and high-fidelity (N_{hi}) data points (assuming $N_{\text{hi}} \subseteq N_{\text{lo}}$). These models were trained using subsets of the D351 dataset and the test-points were the remaining D351$^{\text{HiFi}}$ points (not in the training set). Averaged test RMSE over 50 runs for each of these models are shown in Fig. 3(a). Notably, the positive effect of larger N_{lo} can be seen on the test RMSE – the test error decreases systematically as we increase the number of low-fidelity data points to train our model. This clearly suggests that although low-fidelity points are not accurate, they still contain enough information to improve the ML model performance. This is also reflected in the first panel of Fig. 3(b), where we show learning-curves of the test and training errors as a function of number of high-fidelity points in the

Table 1
The ML models constructed using different subsets of the E_g dataset in order to evaluate the accuracy of the MF approach in comparison to GPR under different scenarios of available training data.

Model	Training data
GPR$_{351}$	D351$^{\text{HiFi}}$
GPR$_{382}$	D351$^{\text{HiFi}}$+ D31$^{\text{HiFi}}$
MF$_{351}$	D351$^{\text{LoFi}}$ + D351$^{\text{HiFi}}$
MF$_{382a}$	D351$^{\text{LoFi}}$ + D351$^{\text{HiFi}}$ + D31$^{\text{LoFi}}$
MF$_{382b}$	D351$^{\text{LoFi}}$ + D351$^{\text{HiFi}}$+ D31$^{\text{LoFi}}$ + D31$^{\text{HiFi}}$

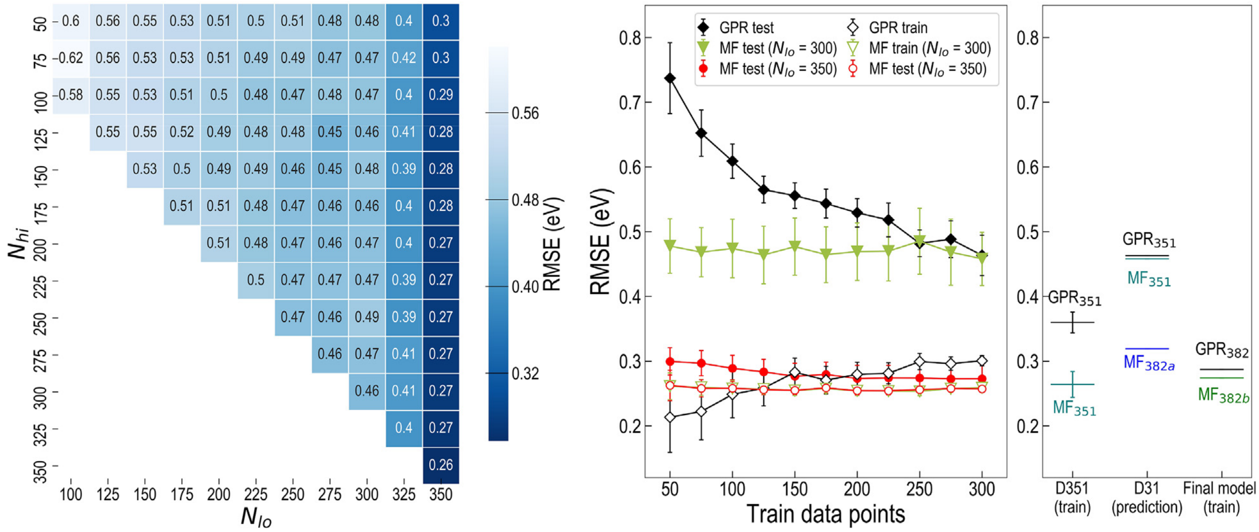

Fig. 3. Figure (a) shows averaged test RMSE of MF model for different combination of low-(N_{lo}) and high-fidelity (N_{hi}) data points. These RMSE's were calculated by taking the mean of the test RMSE from 50 different runs. First panel of figure (b) depicts a comparison of learning-curves for the GPR and MF model. Train errors of both GPR and MF model are shown in the next panel when both are trained with D351 dataset. Prediction accuracy of D31 dataset is shown in the third panel, whereas the last panel shows the train error of the MF model trained with the full dataset of 382 polymers.

training set.

To further highlight the flexibility and the associated effect on the performance of the MF approach, two sets of learning curves are shown: one with 300 and other with 350 known low-fidelity PBE bandgap values. In each case the test set consisted of the remaining HSE06 points. The following important observations can be made: first, both test and train errors of the MF model are smaller compared to that of the GPR model for a particular training set size (i.e. N_{hi} or HSE bandgap data size), especially when the training set size is fairly small. We again note that GPR represents a benchmark case just trained using the high-fidelity data. Second, test error in the MF model decreases when N_{lo} is higher or larger number of low-fidelity data is known during training (see the $N_{lo} = 300$ and $N_{lo} = 350$ curves). Thus, the MF learning-curve for the model trained with 350 low-fidelity data reaches the test RMSE of 0.27 eV compared to the value of 0.47 eV test RMSE of the SF GPR model when tested for the case of 200 high-fidelity bandgap values. Both the above observations are in-line with the previous remark that the knowledge of the low-fidelity data, although somewhat inaccurate, can still improve the accuracy of the ML models. Finally, we note that the number of low-fidelity and high-fidelity data required to improve the performance of the MF approach in comparison to the SF model is problem-specific. It depends on the accuracy of the LF data, noise in the HF data, and the ratio of the cost required to obtain the data points at different levels of fidelity.

The training error of the two models on the entire D351 dataset, namely GPR_{351} and MF_{351}, are depicted comparatively on the leftmost side of the 2^{nd} panel of Fig. 3(b). In the middle of the 2nd panel, we show the predictive capabilities of these models on the hold-out set of 31 polymers (D31). We also show how this test error is drastically reduced when low-fidelity data is introduced for the D31 dataset (see MF_{382a}). Finally, 5-fold cross-validation is used on the entire dataset of 382 polymers to construct the models GPR_{382} and MF_{382b}. We notice that the MF models outperform GPR models in all scenarios and possess tremendous utility when making predictions in new chemical spaces. We note that the MF_{382b} built using the entire dataset can be considered as the best model for future use. Fig. 4 shows a parity plot, comparing the predicted vs the DFT computed HSE06 bandgap values of this final model. Favorable performance across the entire 0.75–10 eV range can be seen.

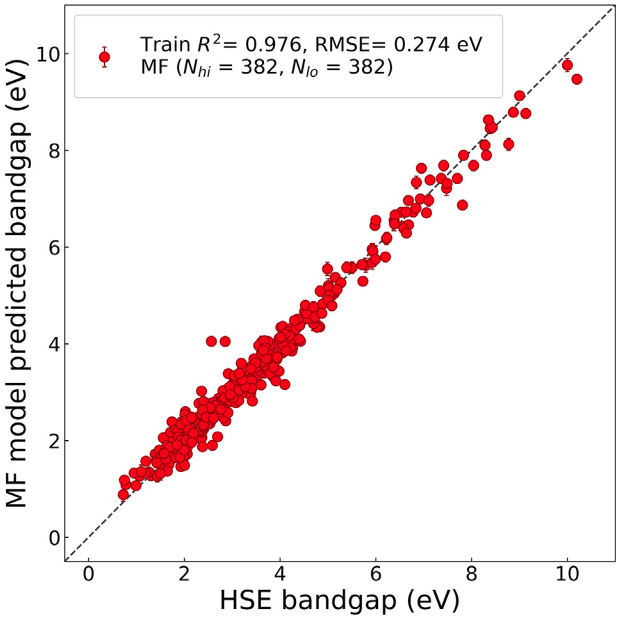

Fig. 4. As reported earlier (in the third panel of Fig. 3 (b), the inclusion of more PBE bandgaps in the training data remarkably improved the prediction accuracy of MF model. In this figure, we show a parity plot of that model (MF_{382b}).

4. Conclusion

In conclusion, we demonstrated that fusing information available at multiple levels of fidelity can indeed be useful for building powerful and accurate predictive models. Using a bandgap dataset of 382 polymers, computed at two different levels of fidelity by employing different DFT functionals, i.e. low-fidelity PBE and high-fidelity HSE06, we built a series of multi-fidelity co-kriging models, and evaluated their performance against traditional Gaussian process regression based models which use information only at a single level of fidelity. The multi-fidelity models were found to consistently outperform the traditional single fidelity models, especially in scenarios when large volumes of low-fidelity data was available. Further, for cases where the low-fidelity information was available, we found that the predictions from the multi-fidelity models were much more accurate than the single fidelity models, which rely on just the high-fidelity data. This clearly demonstrates that although low-fidelity data is not accurate, it contains enough information to be fused with limited high-fidelity data, and allows better generalization of ML models. Thus, for problems involving exploration over large chemical space with particularly high cost associated with high-fidelity measurements/computations, this multi-fidelity approach is expected to provide a cost-effective pathway to generate accurate surrogate models. In this regard, we intend to implement MF algorithms as part of the continuously expanding prediction toolkit of *Polymer Genome* (https://www.polymergenome.org).

CRediT authorship contribution statement

Abhirup Patra: Investigation, Writing - original draft, Data curation, Formal analysis. **Rohit Batra:** Writing - original draft, Formal analysis. **Anand Chandrasekaran:** Resources, Validation. **Chiho Kim:** Resources, Validation. **Tran Doan Huan:** Resources, Validation. **Rampi Ramprasad:** Supervision, Project administration.

Data availability

The raw data required to reproduce these findings are available to download from [*Khazana* (https://khazana.gatech.edu)]. The processed data required to reproduce these findings are available to download from [*Khazana* (https://khazana.gatech.edu)].

Acknowledgement

The authors acknowledge support of this work by the Toyota Research Institute through the Accelerated Materials Design and Discovery program.

References

[1] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[2] T. Mueller, A.G. Kusne, R. Ramprasad, Rev. Comput. Chem. (2016) 186–273.
[3] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, npj Comput. Mater. (2017) 3.
[4] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Nature 559 (2018) 547–555.
[5] J. Carrasquilla, R.G. Melko, Nat. Phys. 13 (2017) 431.
[6] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O.A. von Lilienfeld, New J. Phys. 15 (2013) 095003.
[7] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, J. Phys. Chem. Lett. 6 (2015) 2326–2331.
[8] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, R. Ramprasad, npj Computat. Mater. 5 (2019) 22.
[9] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, Nat. Commun. 8 (2017) 15679.
[10] E. Kim, K. Huang, S. Jegelka, E. Olivetti, npj Comput. Mater. 3 (2017) 53.
[11] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, Chem. Mater. 29 (2017) 9436–9444.
[12] J.P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. 77 (1996) 3865–3868.
[13] A.V. Krukau, O.A. Vydrov, A.F. Izmaylov, G.E. Scuseria, J. Chem. Phys. 125 (2006) 224106.
[14] R. Batra, G. Pilania, B.P. Uberuaga, R. Ramprasad, ACS Appl. Mater. Interfaces (2019) in press.
[15] G. Pilania, J.E. Gubernatis, T. Lookman, Comput. Mater. Sci. 129 (2017) 156–163.
[16] A. Mannodi-Kanakkithodi, A. Chandrasekaran, C. Kim, T.D. Huan, G. Pilania, V. Botu, R. Ramprasad, Mater. Today 21 (2018) 785–796.
[17] C. Kim, A. Chandrasekaran, T.D. Huan, D. Das, R. Ramprasad, J. Phys. Chem. C 122 (2018) 17575–17585.
[18] T.D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, R. Ramprasad, Sci. Data 3 (2016) 160012.
[19] R.G. Lorenzini, W.M. Kline, C.C. Wang, R. Ramprasad, G.A. Sotzing, Polymer 54 (14) (2013) 3529–3533.
[20] R. Ma, V. Sharma, A.F. Baldwin, M. Tefferi, I. Offenbach, M. Cakmak, R. Weiss, Y. Cao, R. Ramprasad, G.A. Sotzing, J. Mater. Chem. A 3 (28) (2015) 14845–14852.
[21] R. Ma, A.F. Baldwin, C. Wang, I. Offenbach, M. Cakmak, R. Ramprasad, G.A. Sotzing, ACS Appl. Mater, Interfaces 6 (13) (2014) 10445–10451.
[22] T.D. Huan, A. Mannodi-Kanakkithodi, R. Ramprasad, Phys. Rev. B 92 (2015) 014106.
[23] A. Mannodi-Kanakkithodi, G. Pilania, T.D. Huan, T. Lookman, R. Ramprasad, Sci. Rep. 6 (2016) 20952.
[24] T.D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, R. Ramprasad, NPJ Comput. Mater. 3 (2017) 37.
[25] C.S. Liu, G. Pilania, C. Wang, R. Ramprasad, J. Phys. Chem. A 116 (2012) 9347–9352.
[26] W. Kohn, L.J. Sham, Phys. Rev. 140 (1965) A1133.
[27] F. Pedregosa, et al., J. Mach. Learn. Res. 12 (2011) 2825–2830.
[28] A.P. Bartók, G. Csányi, Int. J. Quantum Chem. 115 (2015) 1051–1057.
[29] A. Seko, T. Maekawa, K. Tsuda, I. Tanaka, Phys. Rev. B 89 (2014) 054303.
[30] M.C. Kennedy, A. O'Hagan, Biometrika 87 (2000) 1–13.
[31] A.I. Forrester, A. Sóbester, A.J. Keane, Proc. R. Soc. A: Math. Phys. Eng. Sci. 463 (2007) 3251–3269.