

ARTICLE OPEN



A simple denoising approach to exploit multi-fidelity data for machine learning materials properties

Xiaotong Liu^{1,2}, Pierre-Paul De Breuck³, Linghui Wang² and Gian-Marco Rignanesi^{3,4}✉

Machine-learning models have recently encountered enormous success for predicting the properties of materials. These are often trained based on data that present various levels of accuracy, with typically much less high- than low-fidelity data. In order to extract as much information as possible from all available data, we here introduce an approach which aims to improve the quality of the data through denoising. We investigate the possibilities that it offers in the case of the prediction of the band gap using both limited experimental data and density-functional theory relying on different exchange-correlation functionals. After analyzing the raw data thoroughly, we explore different ways to combine the data into training sequences and analyze the effect of the chosen denoiser. We also study the effect of applying the denoising procedure several times until convergence. Finally, we compare our approach with various existing methods to exploit multi-fidelity data and show that it provides an interesting improvement.

npj Computational Materials (2022)8:233; <https://doi.org/10.1038/s41524-022-00925-1>

INTRODUCTION

With the considerable increase in available data, materials science is undergoing a revolution as attested by the multiplication of data-driven studies in recent years^{1–5}. Among all these investigations, the prediction of properties from the atomic structure (or sometimes just the chemical composition) is an extremely important topic⁶. Indeed, having an accurate predictor can be very useful for accelerating high-throughput material screening^{7,8}, generating new molecules³, or classifying reactions⁹. To develop such predictors, different kinds of structural descriptors have been used relying on the Coulomb matrix¹⁰, graphs¹¹, voxels¹², and so on. Furthermore, various advanced machine learning techniques have been employed in the predictor model, such as *attention*¹³, *graph convolution*¹¹, *embedding*¹⁴, or *dimensionality reduction*¹⁵. However, the characteristics of the data itself are seldom discussed.

In particular, many databases consist of simulated data, hence intrinsically poorer in quality compared to the experimental results. Indeed, it is basically always necessary to resort to approximations to limit the computational power required to solve the quantum mechanical equations describing a molecule or a solid. For instance, density-functional theory (DFT) relies on an approximate functional to model the exchange-correlation (XC) energy (see e.g., ref. ¹⁶). This necessarily comes at the cost of a reduction of the accuracy of the predicted properties. For example, it is well known that properties calculated within DFT may present systematic errors^{17–19}. Nonetheless, many computational works focus on the relative property values for different structures, so that the systematic errors can be canceled and the trend between the structures is not affected²⁰. In contrast, if the focus is on the absolute property value for a given structure, the difference between the experimental and calculated values cannot be neglected. For instance, for the band gap of solids, the DFT calculations relying on local and semi-local XC functionals

typically lead to a systematic underestimation of 30–100% with respect to the experimental results²¹.

The available data for one property may present various levels of accuracy, depending on the approach adopted to collect them, be it computational or experimental. Obtaining computational results faster usually requires to resort to more important approximations and hence leads, as a general rule, to a lower accuracy (cost vs. accuracy trade-off). Obtaining experimental results generally necessitates even more time. Therefore, available databases typically contain orders of magnitude less high-accuracy results than low-accuracy ones.

To take full advantage of all these different accuracy data, various multi-fidelity approaches have recently been developed for studying all kinds of properties, such as molecular optical peaks²², formation energies^{23,24}, and band gaps^{25,26} which will also be studied in this work. In the same line of thought, various methods have also been considered, including information fusion algorithms²³, Bayesian optimization^{24,27}, directed message passing neural networks²², and transfer learning^{28,29}.

In the multi-fidelity approaches, different values of the property (varying in accuracy) may be provided for the same compound. The systematic errors that they may present (from the data-producer standpoint) are included in the *noise* as defined from a machine-learning (ML) perspective (i.e., from the data-consumer standpoint), see Supplementary Information about bias-variance-noise decomposition. Here, we consider the experimental data as the true value (even though it may vary depending on the technique). Note that the ML noise (data-consumer standpoint) covers not only systematic errors but also random noise (data-producer standpoint).

In the present work, we take a different route to exploit all available information: we focus on decreasing the noise on the data (i.e., reducing the input errors of the different models), taking inspiration from O2U-Net, a recent effort for improving the model performance in image classification³⁰. The concept of noisy data dates back to the 80s and reasonable strategies were developed

¹Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, No. 35 Beisihuan Middle Road, Beijing 100101 Beijing, P. R. China. ²School of Computer, Beijing Information Science and Technology University, No. 35 Beisihuan Middle Road, Beijing 100101 Beijing, P. R. China. ³Institute of Condensed Matter and Nanosciences, UCLouvain, Chemin des Étoiles 8, Louvain-la-Neuve 1348, Belgium. ⁴School of Materials Science and Engineering, Northwestern Polytechnical University, No. 127 Youyi West Road, Xi'an 710072 Shaanxi, P. R. China. ✉email: gian-marco.rignanesi@uclouvain.be

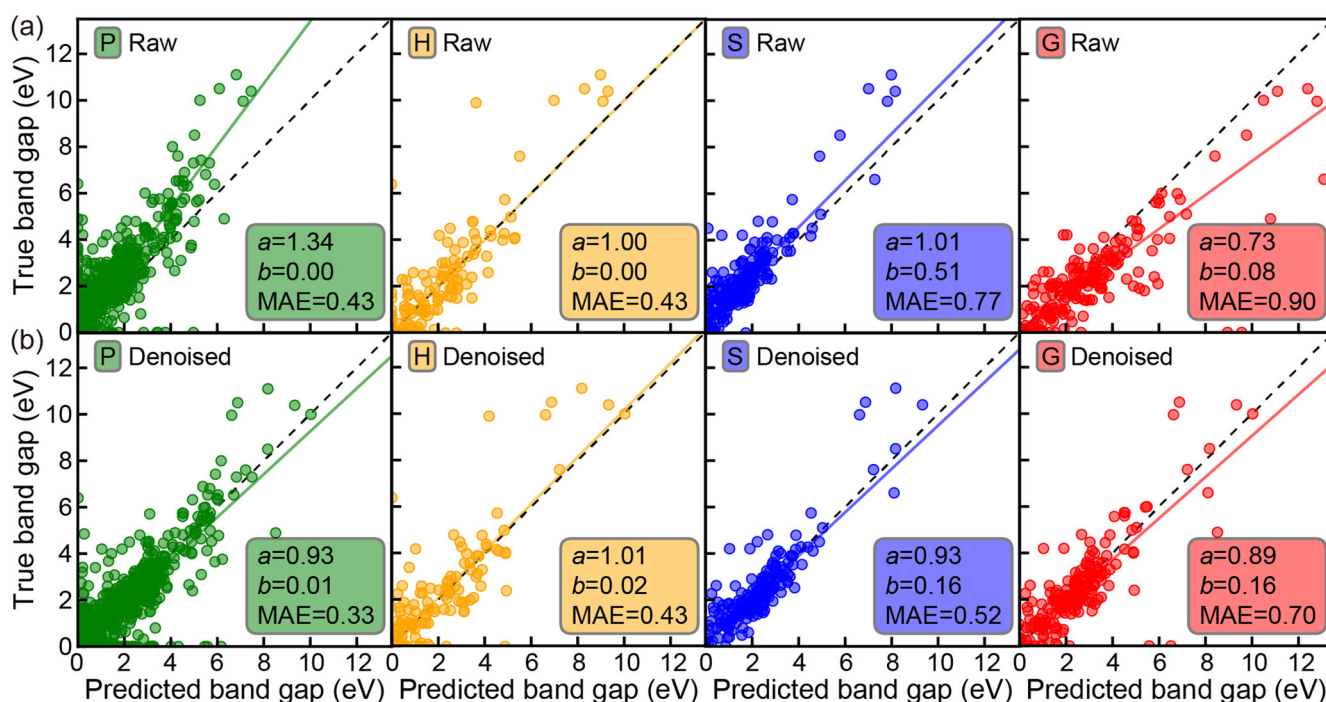


Fig. 1 True (i.e., experimental) vs. predicted band-gap parity plots. **a** The raw DFT data for different functionals (P, H, S, and G) and **b** the corresponding results obtained by applying the denoising procedure. A first-order polynomial $T = aP + b$ is used to fit the true T and predicted P data points. The dashed line corresponds to a perfect match between P and T (i.e., $a = 1, b = 0$ with a null MAE).

to handle labeling errors, provided that these affect a limited amount of the samples^{31,32}. A straightforward approach is to first remove as much noise as possible, and then train the model with the *cleaned* dataset³³. Some other works rely on curriculum learning³⁴ to gradually train the model using the complete dataset ordered in a meaningful sequence^{35,36}.

Theoretically, an additive normal distribution noise ($N \sim (\mu, \sigma^2)$) can be denoised effectively by soft-thresholding with Stein's Unbiased Risk Estimate (SURE)^{37,38} (A good example can be found at: <https://github.com/ilkercbayram/SURE>). However, the quantum mechanics errors are probably more like a combination of additive noise and multiplicative noise (see below). It is thus interesting to design a way to decrease the noise. This work investigates how such a denoising approach can be applied in materials science to improve the performance of a typical machine learning model.

RESULTS

Raw data

In this paper, we focus on various datasets (four with DFT predictions and one with experimental measurements) available for the band gap (see the sub-section "Data" in Methods). As a starting analysis, it is interesting to compare how DFT with different functionals P, H, S, and G performs with respect to the experiments (E), considered as the true values. This can obviously only be done for the compounds that belong to the intersections $P \cap E$, $H \cap E$, $S \cap E$, and $G \cap E$. As can be seen in Fig. 1 and Supplementary Figure 11, the systematic errors depend on the XC energy functional: some underestimate the band gap with a mean error (ME) of -0.34 eV for P, -0.09 eV for H, and -0.68 eV for S; while others overestimate it with a ME of 0.65 eV for G. Note that these statistics are clearly affected by the compounds that have been considered (see discussion below). The corresponding mean absolute errors (MAE) are reported in blue in Fig. 2, as well as in Supplementary Table 1. For a deeper understanding, we also indicate the MAEs corresponding to three categories of

compounds: metals, as well as small-gap ($E_g < 2$) and wide-gap ($E_g \geq 2$) semiconductors.

In Supplementary Figure 1, we provide complementary information about the different datasets based on this decomposition. It is clear that the dataset E contains an important fraction of metals (51%). The accuracy of the DFT predictions (as measured by the global MAE) is thus very sensitive to the accuracy for metals. It turns out that P and H are doing a very good job for metals with 91% and 90% accuracy, respectively (see confusion matrix in Supplementary Figure 1), leading to a MAE of 0.03 and 0.07 eV, respectively. In contrast, S and G are doing a rather poor job for metals. For the small-gap ($E_g < 2$) semiconductors, all the functionals have a very similar accuracy and MAE. For the wide-gap ($E_g \geq 2$) semiconductors, P clearly provides the worst predictions while the other three functionals have roughly the same accuracy.

Another important remark is that the distribution of compounds between the three different categories varies for the different intersections ($P \cap E$, $H \cap E$, $S \cap E$, and $G \cap E$). In $P \cap E$ and $H \cap E$, it is not too different from the actual distribution in the dataset E. That is clearly not the case for $S \cap E$ and $G \cap E$ in which metals are strongly underrepresented. Furthermore, in $G \cap E$, the wide-gap ($E_g \geq 2$) semiconductors are largely overrepresented. In fact, this dataset was created to analyze how the corresponding functional performs for correcting the systematic underestimation of the band gap.

This remark is also important in the framework of the machine learning training process. Indeed, a basic underlying assumption of such approaches is that the training dataset has a similar distribution to the test dataset. This is a reasonable assumption for the dataset H and to a lesser extent for the dataset P, but not at all for the datasets S and G. Given that the whole point here (and of multi-fidelity approaches) is to take advantage of all available data to overcome the lack of experimental data, we have to accept to deal with datasets with all kinds of distributions. But it is clear that the underlying distribution will impact the ML models.

In the Supplementary Information, the interested reader will find further analyses of the raw data including the elemental distribution, the overlaps between the different datasets, PCA

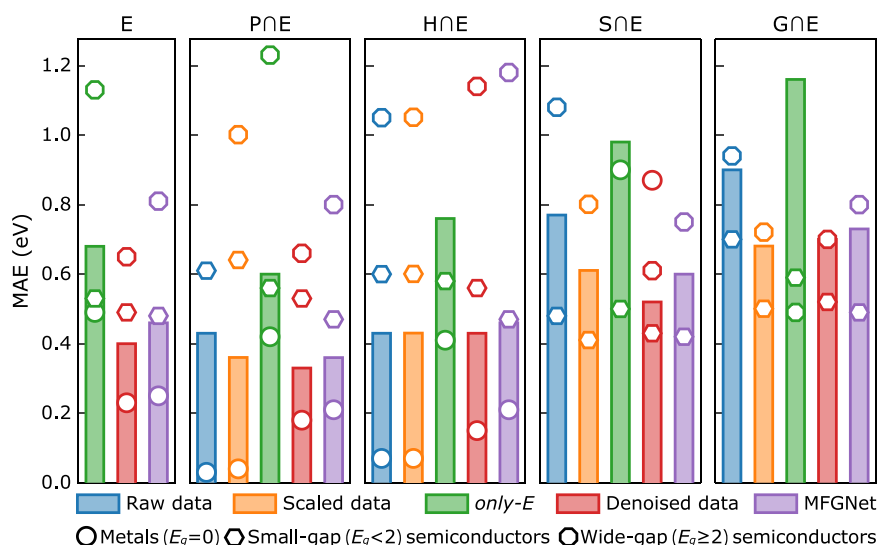


Fig. 2 MAE for the different approaches. The subplot on the left shows the MAE for all the available experimental data, the other subplots report the MAE for the different intersections $P \cap E$, $H \cap E$, $S \cap E$, and $G \cap E$. In each subplot, the MAEs are also indicated separately for the metal (circles), the small-gap ($E_g < 2$) semiconductors (hexagons), and the wide-gap ($E_g \geq 2$) ones (octagons). The corresponding MAE values are provided in Supplementary Tables 1–5.

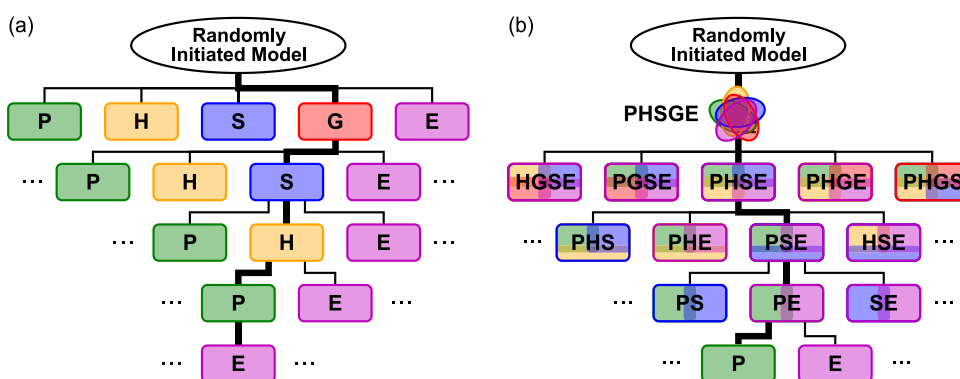


Fig. 3 Schematic representation of the training sequence. **a** One-by-one approach and **b** union approach. Both involve five training steps. The thicker line in each diagram shows one possible training sequence. For five datasets, each tree contains 120 ($=5!$) different branches and 325 (resp. 206) nodes in the *one-by-one* (resp. *union*) approach. Each node in the tree represents a training step. As the datasets P, H, S, G, and E appear (resp. disappear) with the same probability in the trees, there are $325/5=61$ (resp. $206-205/5=165$) nodes containing P (or any other letter) in the *one-by-one* (resp. *union*) tree.

visualizations and the distribution of the band gap predictions and errors for the different XC functionals. We did not identify any other specific bias in the raw data than the one towards metal compounds.

Scaled data

It has been long known that DFT predictions may present systematic deviations (from a data-producer standpoint). Methods for analyzing these errors have therefore already been considered previously³⁹. The simplest model assumes that a perfect linear correlation exists between the experimental true value T and the DFT prediction P , possibly with a random noise δ centered around a zero mean: $T = aP + b + \delta$. The parameters a and b can be determined by a linear regression (by minimizing the MAE, following a 2-fold testing procedure). Figure 1(a) shows the results of such an analysis for the different intersections ($P \cap E$, $H \cap E$, $S \cap E$, and $G \cap E$).

These relations can be used to scale the band gaps predictions from DFT²¹. The resulting MAEs are reported in orange in Fig. 2, as well as in Supplementary Table 2. The scaled DFT data are improved compared to those of the raw ones, except for the

metals in $P \cap E$ and $H \cap E$. In fact, this simple scaling approach mainly corrects the wide-gap compounds. Though it may reduce the systematic error for a given fidelity, it is not very widely used.

Training approaches

As already indicated, the experimental values are assumed to be the *true values*. When training models, the main objective is to minimize the MAE between the outputs and the true values. Before presenting our denoising procedure, we introduce the following training approaches:

1. *only-E*: the training data consists only of the experimental data;
2. *all-together*: the training data consists in the union of all datasets (P, H, S, G, and E training fold) regardless of their different fidelity;
3. *one-by-one*: the training data is sequentially changed in a selected sequence based on fidelity (e.g., $G \rightarrow S \rightarrow H \rightarrow P \rightarrow E$ as illustrated in Fig. 3a). Weights are transferred from one training stage to the other with no change in the hyperparameters.

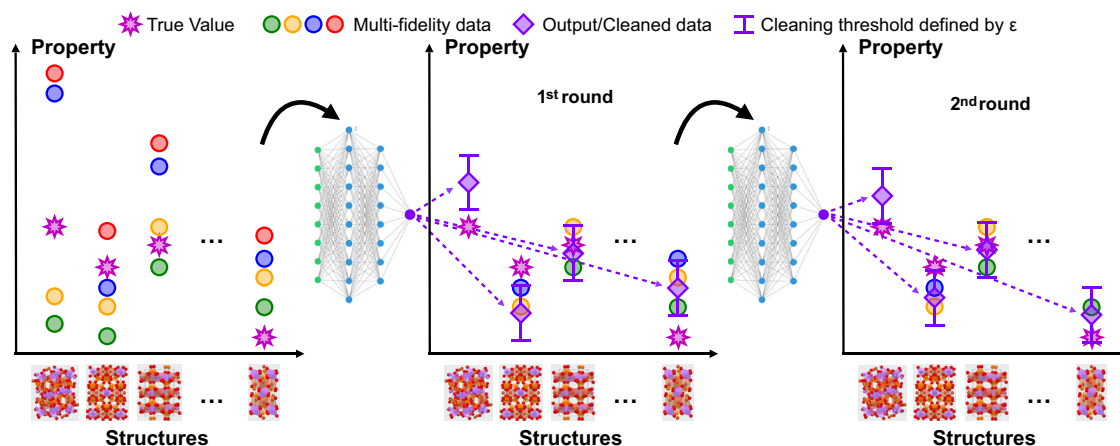


Fig. 4 Schematic representation of a typical data cleaning process. At each step, the multi-fidelity data (different property values for a given structure) are represented using different colors (the same as those adopted throughout the paper) and symbols (a star for the true experimental value and circles for the data with noise). Using the data, a predictive model (here symbolized by a neural network) is trained and predictions are made. The low-fidelity data which are too far (i.e., outside the interval defined by the cleaning threshold ϵ) from the output of the predictive model are replaced by the latter values. After this cleaning step, the procedure can be repeated until convergence.

4. *onion*: the training data consists of, first, the combination of all five datasets (P, H, S, G, and E) and, then, those obtained by removing one dataset at a time in a selected sequence (e.g., PHSGE $\rightarrow$ PHSE $\rightarrow$ PSE $\rightarrow$ GE $\rightarrow$ E as illustrated in Fig. 3b). Again, weights are transferred from one training stage to the other with no change in the hyperparameters.

The *only-E* approach is a single-fidelity approach that could have been used if only experimental data were available. The *one-by-one* approach is a form of curriculum learning.

There are 120 (=5!) different possible sequences for both the *one-by-one* and *onion* training approaches. In this work, we consider all those alternatives systematically. These can be represented as a tree, a part of which is shown in Fig. 3, highlighting one potential choice. In what follows, we adopt the Environment for Tree Exploration (ETE) Toolkit⁴⁰ to display the complete tree of the different results. By investigating all those options, which is very time consuming, we aim to analyze the sensitivity of the methods to the selected sequence. Ideally, one would like to avoid to take them all into account for actual ML problems. It is thus important to devise a method that is as little sensitive as possible to the selected sequence.

Note that the *all-together* approach is the first step of the *onion* tree, while the *only-E* approach is the first step of a part of the *one-by-one* tree.

Denoising procedure

Let T_D be the target value (which includes noise) for a given sample in dataset D . P is the corresponding prediction by a reasonable model. A denoising procedure typically consists in replacing T_D by $\hat{T} = f(T_D, P)$ where f is any function of T_D and P and is usually referred to as the *denoiser*. Note that this can be an iterative procedure. Given that the type of noise in our DFT datasets is unclear, it is not obvious to select an existing denoiser. In this work, we adopt a rather straightforward one:

$$\hat{T} = \begin{cases} T_D & \text{if } |T_D - P| \leq \epsilon \\ P & \text{otherwise} \end{cases} \quad (1)$$

where ϵ is a hyper-parameter to be determined (e.g., by grid search, random search, or Bayesian optimization). The whole denoising process is schematically represented in Fig. 4.

We are still left with the choice of the reasonable model to be used for making the prediction P . Typically, the best model from the *one-by-one* and *onion* approaches is used. The effect of

choosing poorer models will also be discussed. It is important to note that the denoiser model can be updated in an iterative process which improves the model performance until convergence is achieved.

Training on the raw data

We first test the different approaches on the raw data (i.e., without applying the denoising procedure). The most representative results are summarized in Table 1, while the complete results of the *one-by-one* and *onion* approaches are shown in Figs. 5 and 6.

The *only-E* approach is the reference scenario. It leads to a MAE of 0.680 eV as reported in green in Fig. 2, as well as in Supplementary Table 3. It is higher than most of the results obtained with any other approach. This can be traced back to the small dataset size.

The *all-together* approach leads to a MAE of 0.501 eV. That is a significant improvement by 26%, which can be attributed to a better prediction of metals thanks to the much larger size of the dataset. This can be understood by analyzing the results obtained by training only on the dataset P (*only-P*). This approach leads to a MAE of 0.595 eV, which is already an improvement by 13% compared to the *only-E* approach despite the fact that PBE is known to underestimate the band gap. In fact, 72% of the experimental data points correspond to a band gap lower than 2 eV and 51% are actually metals. If we focus on the intersection $P \cap E$ (containing 1765 compounds), we see that 59% of the compounds are metallic and P is actually correct in 91% of the cases. The underestimation of the band gap only leads to 9% of false metallic compounds. Now, moving to the rest of the datasets $P \setminus E$, we see that, out of the 50583 compounds, 18775 (37%) are metals. This number is basically one order of magnitude larger than the 1384 metallic compounds present in the whole dataset E. So, the ML model can better learn to predict metals. Adding the fact that another 15218 compounds have a band gap smaller than 2 eV for which the PBE error is not going to be very big, we can easily understand the nice improvement in MAE. For the datasets S and G, the number of new metallic systems added compared to E (11, and 0, respectively) is much smaller. So, not surprisingly, *only-S* and *only-G* suffer much more from the noise due to the XC functionals than the *only-P* one leading a MAE of 1.446 and 1.406 eV, respectively. For the *all-together* approach, the improvement results from both the effect of the number of metallic samples and an averaging of the noise of the different XC functional. The *only-H*

Table 1. Most representative results for the different training approaches and sequences without any denoising.

Approach	Sequence	MAE			
		Global	$E_g = 0$	$E_g < 2$	$E_g \geq 2$
<i>only-E</i>	E	0.680	0.490	0.534	1.131
<i>all-together</i>	PHSGE	0.501	0.150	0.567	1.091
<i>one-by-one</i>	$S \rightarrow G \rightarrow P \rightarrow E \rightarrow H$ (best)	0.484	0.228	0.571	0.884
	$H \rightarrow P \rightarrow S \rightarrow E \rightarrow G$ (worst)	0.993	1.039	0.733	1.097
	$H \rightarrow P \rightarrow G \rightarrow S \rightarrow E$ (worst*)	0.599	0.434	0.510	0.964
<i>onion</i>	PHSGE $\rightarrow$ PHSE $\rightarrow$ HSE $\rightarrow$ HE $\rightarrow$ E (best)	0.438	0.239	0.515	0.743
	PHSGE $\rightarrow$ PHSG $\rightarrow$ PSG $\rightarrow$ SG $\rightarrow$ G (worst)	0.916	0.937	0.634	1.083
	PHSGE $\rightarrow$ PHSG $\rightarrow$ PSG $\rightarrow$ PG $\rightarrow$ G (2 nd -worst)	0.889	0.883	0.665	1.064
	PHSGE $\rightarrow$ PSGE $\rightarrow$ SGE $\rightarrow$ GE $\rightarrow$ E (worst*)	0.495	0.338	0.485	0.790

For the *one-by-one* and *onion* approaches, we report the results for the best training sequence, the worst one, and the worst one among those ending with E (indicated with a star). The 2-fold MAEs on the predicted band gap are expressed in eV.

results are somewhere in between with a MAE of 0.796 eV. Indeed, the number of new metals in $H \setminus E$ (2599) is only the double than in E (compared to more than 10 times in P). So, the effect of the better prediction of metals is more limited compared to P.

For the *one-by-one* and *onion* approaches, the results vary depending on the training sequence. The best and worst results are reported in Table 1. In order to analyze the effect of the training sequence, we have produced two plots for both approaches in Figs. 5 and 6. In the first part of those figures, the sequences are classified according to the first dataset used or removed (P, H, S, G, or E); while, in the second part, they are ordered depending on the last dataset used.

In both figures, each class presents much more variation around its mean in the first plot than in the second one. In other words, the final dataset used seems to matter much more than the first one used (resp. removed) in the *one-by-one* (resp. *onion*) approach. It is, however, also clear that using the dataset G first leads to better results and not surprisingly finishing the training with it produces the worst results by far. For the *one-by-one* approach, the best results on average are obtained for the sequences finishing with H. They are slightly better than those finishing with E. For the *onion* approach, it is actually the reverse: the best results on average being achieved for the sequences finishing with E. As a general rule, in order to limit the number of models to be tested, one can clearly focus on the latter sequences (i.e., those finishing with the available true values) and, for further restriction, one can concentrate on those which end with PE or HE given that P and H have the lowest MAE (i.e., the highest fidelity) in Supplementary Table 1.

Training on the denoised data

We now turn to the analysis of the results that can be obtained when denoising the data. Given that we have already considered all the possible training sequences, the natural choice to clean the data is to use the best model obtained with the raw data. Once again, we first analyze the effect of the training sequence. The results obtained for both *one-by-one* and *onion* approaches are reported in Figs. 7 and 8. The striking difference with respect to the results obtained on the raw data is that the training sequence has a much smaller impact on the results. This is a really important point in order to avoid the burden of having to compute all the different training sequences. The second important observation is that, once again, the *onion* approach produces better results than *one-by-one*. So, from now on, we

focus on the *onion* approach to analyze the effects of the cleaning procedure.

Given that in a normal investigation the best possible model will not be known *a priori* (it only can *a posteriori* once all sequences have been considered), we investigate the importance of the choice of the denoiser. Here, we have plenty of models at hand differing by the training sequence in the raw data. Besides the one already considered, we select four other denoiser models for comparison:

- PHSGE $\rightarrow$ PHSG $\rightarrow$ PSG $\rightarrow$ SG $\rightarrow$ G which leads to the worst performance among all training paths: MAE = 0.916 eV (Supplementary Figure 12),
- PHSGE $\rightarrow$ PHSG $\rightarrow$ PSG $\rightarrow$ PG $\rightarrow$ G which leads to the second-worst performance among all training paths: MAE = 0.889 eV (Supplementary Figure 13),
- PHSGE $\rightarrow$ PHGE $\rightarrow$ PGE $\rightarrow$ GE $\rightarrow$ E which has a rather poor performance among all training path ending with E: MAE = 0.483 eV (Supplementary Figure 14),
- PHSGE $\rightarrow$ PHSE $\rightarrow$ PHE $\rightarrow$ HE $\rightarrow$ E which has a rather good performance among all training path ending with E: MAE = 0.443 eV (Supplementary Figure 15),

The complete results obtained after the denoising procedure based on these four different models are shown in Supplementary Figures 12–15.

In all four cases, the denoising procedure improves the global average of the MAE for the whole tree, as well as the average MAE of all the sequences ending with E compared to the results of the denoiser model itself. However, when the worst or the second-worst model is used as the denoiser, the results are worse than with the raw data.

Basically, we observe that the better the denoiser model the better the cleaning effects, which translates not only into a lower MAE but also into a lower variance with respect to the training sequence. Therefore, the choice of the denoiser is quite critical.

It would be cheating to use the final results as an indicator to choose the denoiser model. However, we note that, as soon as a model whose training sequence ends with E is chosen as the denoiser (even the rather poor performance one), the results are clearly improved with respect to those obtained based on the raw data. Therefore, based on the observations of the previous section, we recommend as heuristic to use a denoiser for which the training sequence is in increasing fidelity of the data (i.e., decreasing MAE and/or absolute ME with respect to the true values), ending with the true values. This would have led to the sequence PHGSE $\rightarrow$ PHSE $\rightarrow$ PHE $\rightarrow$ HE $\rightarrow$ E, which is not the best

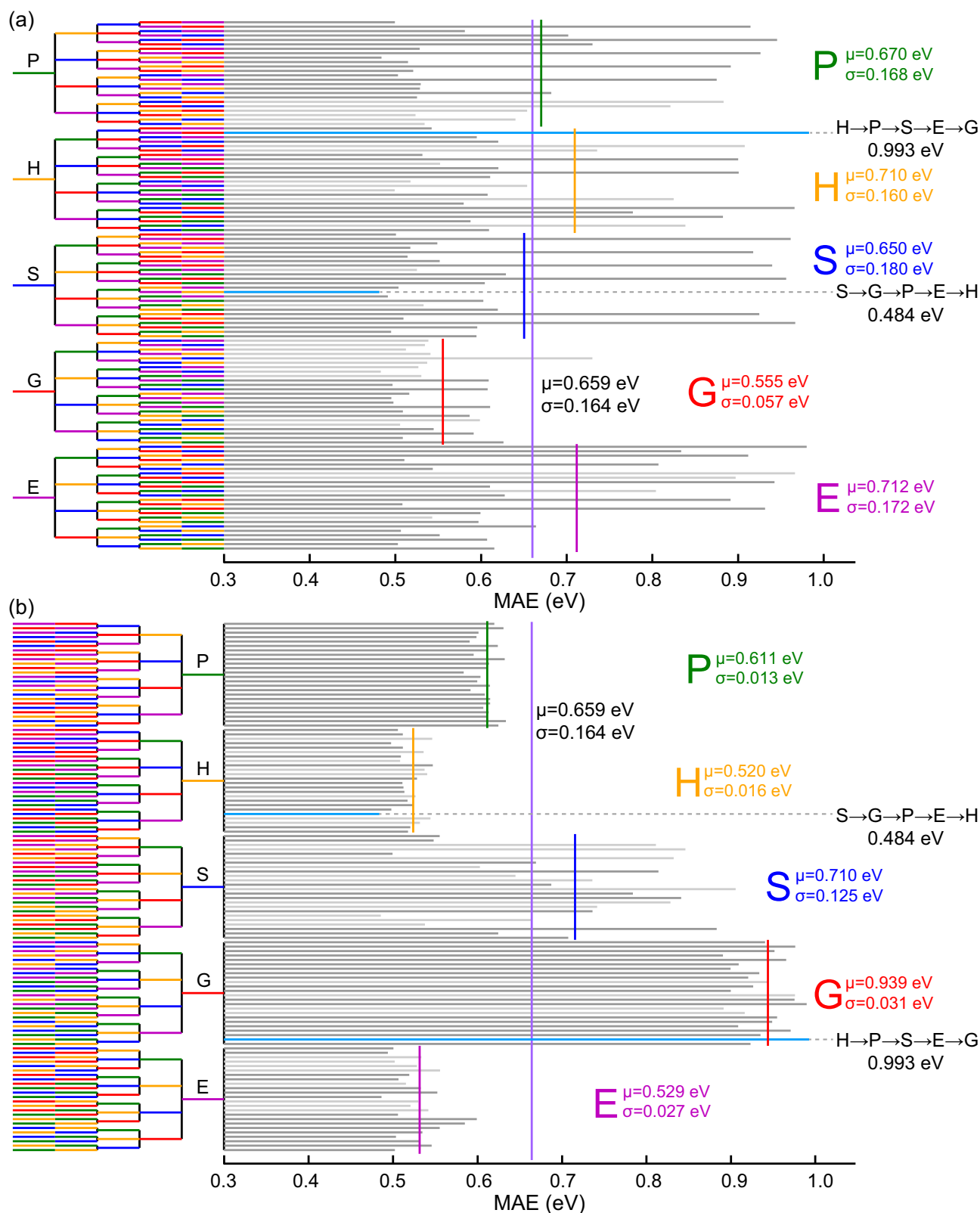


Fig. 5 MAE results on the raw data using one-by-one training approach. All possible dataset sequences are gathered **a** according to the first dataset used and **b** following the last dataset used. The global average of the MAE is shown by a vertical solid purple line ($\mu = 0.659$ eV), while the group averages are indicated by their corresponding color (P in green, H in orange, S in blue, G in red, and E in magenta). The corresponding standard deviations (σ) are also indicated accordingly. The best and worst training sequences are highlighted in light blue. The training sequences that produce NaN for one of the folds (so the MAE is only that of the other fold) are indicated by a lighter gray bar, while those that lead to NaN for both folds are left blank.

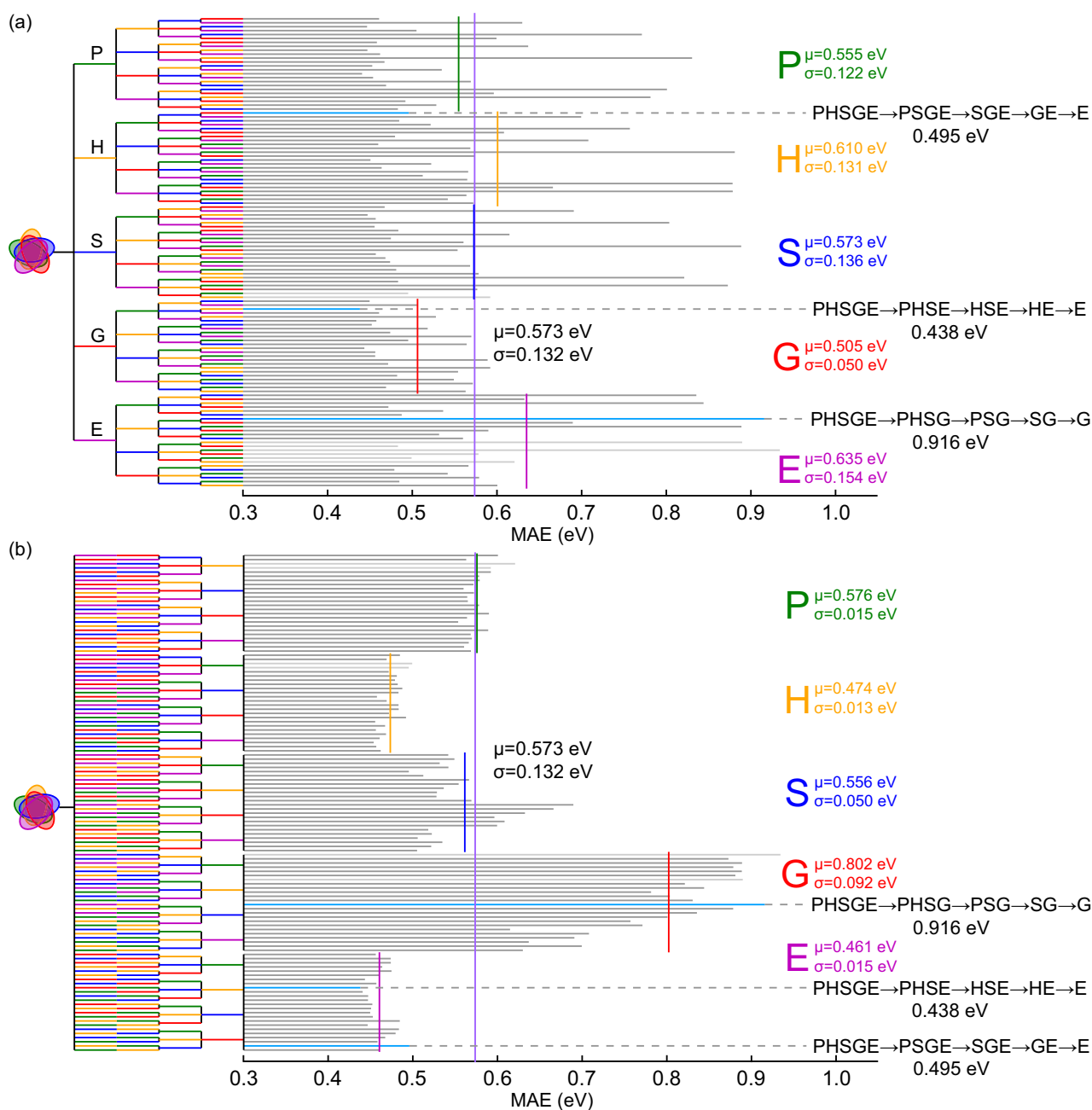


Fig. 6 MAE results on the raw data using the *onion* training approach. All possible dataset orders are gathered **a** according to the first dataset used and **b** following the last dataset used. The global average of the MAE is shown by a vertical solid purple line ($\mu = 0.573$ eV), while the group averages are indicated by their corresponding color (P in green, H in orange, S in blue, G in red, and E in magenta). The corresponding standard deviations (σ) are also indicated accordingly. The best and worst training sequences, as well as the worst one ending by E, are highlighted in light blue. The training sequences that produce NaN for one of the folds (so the MAE is only that of the other fold) are indicated by a lighter gray bar, while those that lead to NaN for both folds are left blank.

one but really close to it. In other words, the proposed method is not sensitive to small changes in the training sequence, as long as it respects the previously stated heuristic based on MAE. Note that the evaluation of the fidelity might be tuned depending on the importance to the different kind of compounds (metals, small-gap, and wide-gap semiconductors). For instance, P has the lowest MAE for metals while S has the lowest MAE for semiconductors.

The final results obtained after one step of the denoising procedure are represented in Fig. 9 for all the samples in dataset E (together with the distribution of errors). For the samples in the intersections $P \cap E$, $H \cap E$, $S \cap E$, and $G \cap E$, the analogous parity

plots can be found in Fig. 1b. Compared to the raw data in Fig. 1a, a clear reduction of noise is observed. This also translates in the corresponding MAE which are reported in red in Fig. 2 and in Supplementary Table 4. As a last assessment of our denoising procedure, we compare the final results obtained after one step of the denoising procedure with those computed using MFGNet²⁵, which is a multi-fidelity model with good performance on the band gap problem. The corresponding results are reported in purple in Fig. 2, as well as in Supplementary Table 5. The global MAE of our approach (0.40 eV) is improved by more than 10% compared to that of MFGNet (0.46 eV). Nonetheless, for the small-

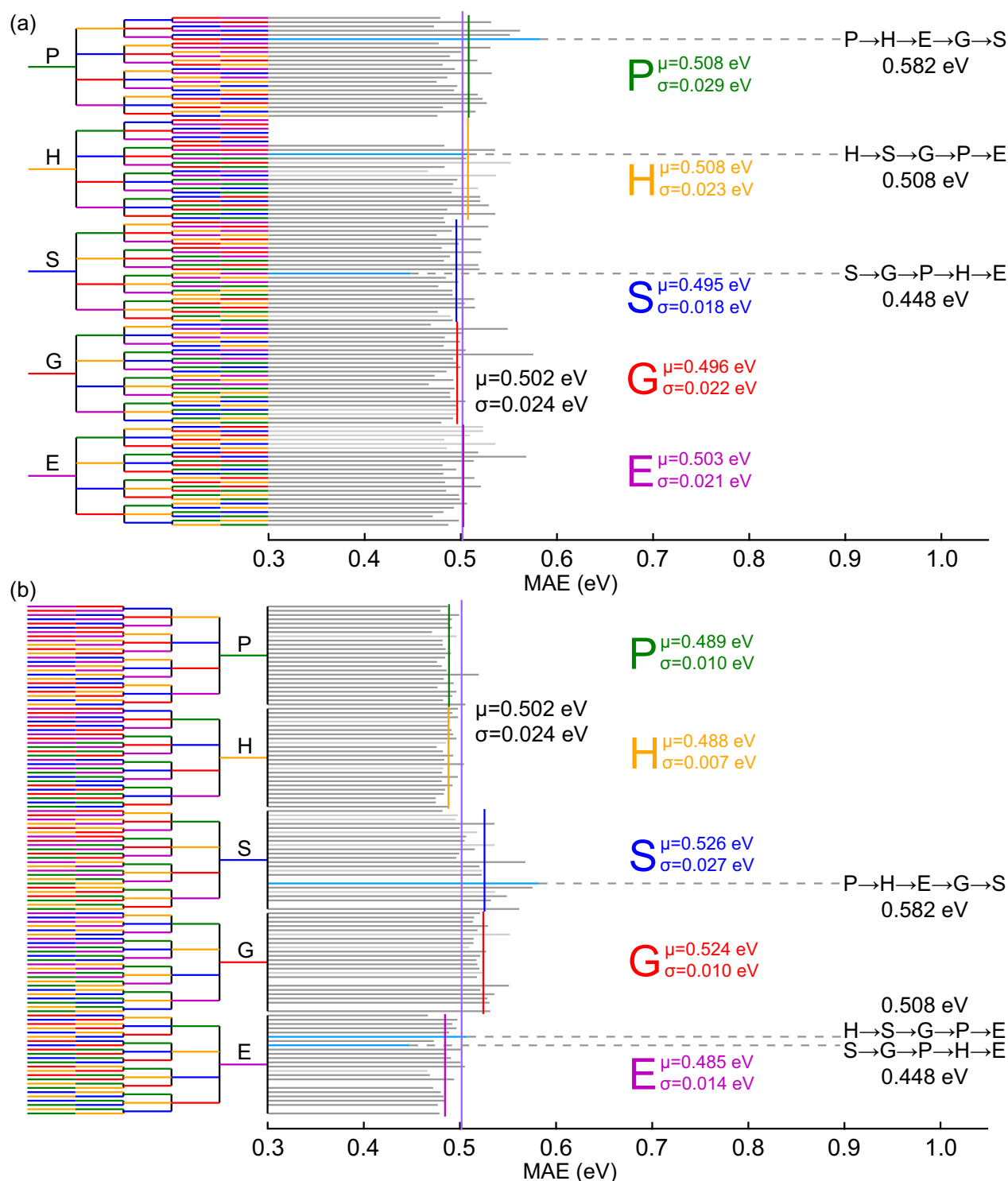


Fig. 7 MAE results on the cleaned data using the *one-by-one* training approach. All possible dataset sequences are gathered **a** according to the first dataset used and **b** following the last dataset used. The data was cleaned using the best model of Fig. 5. The global average of the MAE is shown by a vertical solid purple line ($\mu = 0.502$ eV), while the group averages are indicated by their corresponding color (P in green, H in orange, S in blue, G in red, and E in magenta). The corresponding standard deviations (σ) are also indicated accordingly. The best and worst training sequences, as well as the worst one ending by E, are highlighted in light blue. The training sequences that produce NaN for one of the folds (so the MAE is only that of the other fold) are indicated by a lighter gray bar, while those that lead to NaN for both folds are left blank.

gap ($E_g < 2$) semiconductors (hexagons in Fig. 2), MFGNet shows a better accuracy.

As already indicated, the cleaning procedure can be iterated toward convergence. In Fig. 10, we show the evolution of the

results as a function of the iteration for some representative training sequences. PHSGE → PHSE → HSE → HE → E leads to the lowest MAE (0.394 eV at the 6th iteration). Compared with *one-by-one* and *all-together* approaches, the *onion* training not only shows

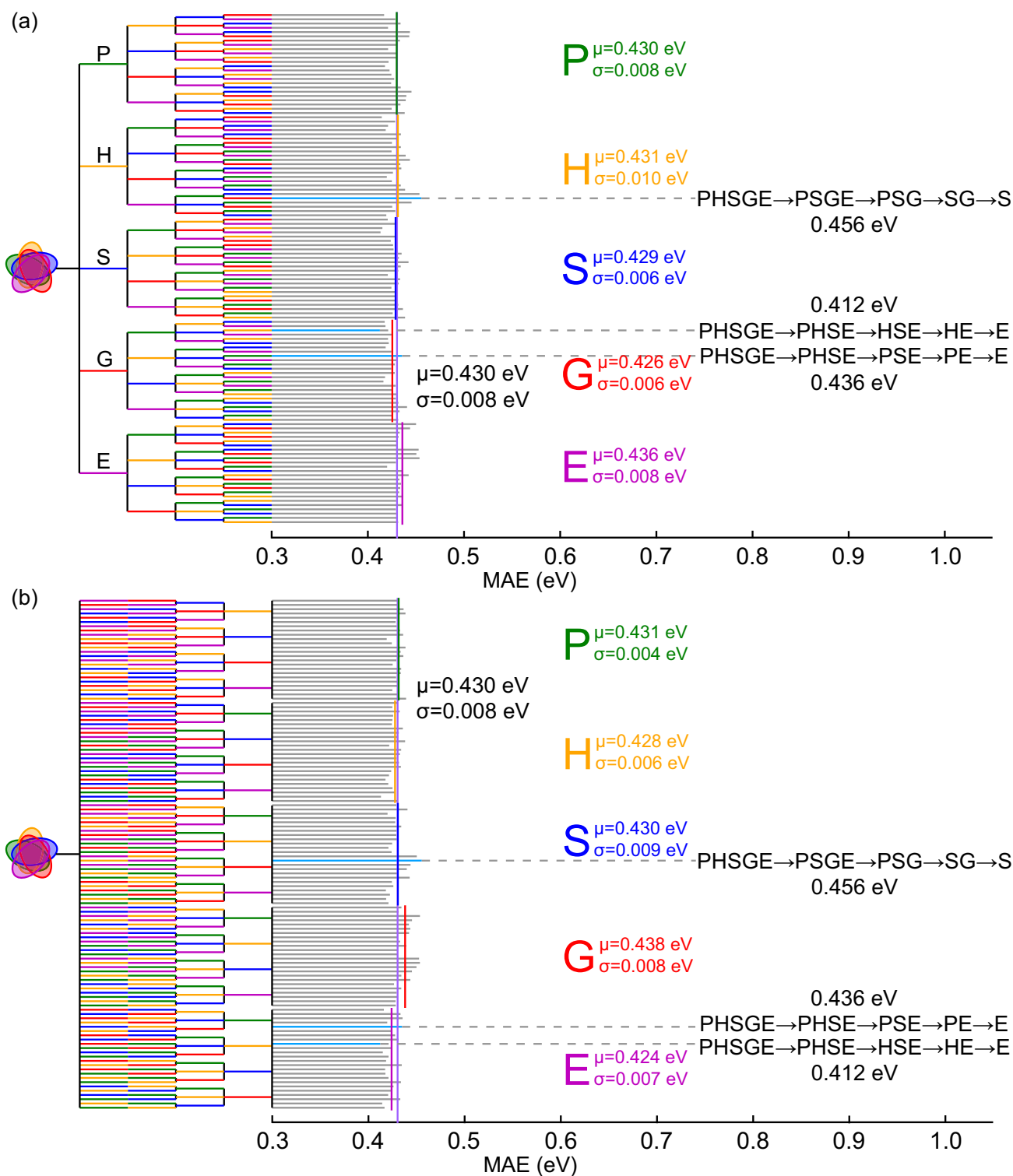


Fig. 8 MAE results on the cleaned data using the *onion* training approach. All possible dataset sequences are gathered **a** according to the first dataset used and **b** following the last dataset used. The data was cleaned using the best model of Fig. 6. The global average of the MAE is shown by a vertical solid purple line ($\mu = 0.430 \text{ eV}$), while the group averages are indicated by their corresponding color (P in green, H in orange, S in blue, G in red, and E in magenta). The corresponding standard deviations (σ) are also indicated accordingly. The best and worst training sequences, as well as the worst one ending by E, are highlighted in light blue. The training sequences that produce NaN for one of the folds (so the MAE is only that of the other fold) are indicated by a lighter gray bar, while those that lead to NaN for both folds are left blank.

the best performance at the starting point, but it also has the greatest potential for improvement. The *one-by-one* training results can actually hardly be improved by the cleaning procedure due to the lack of a real synergetic effect by the different datasets.

DISCUSSION

To assess the generality of our denoising procedure, we further apply it using MODNet¹⁵ as the machine-learning model. The latter is among the best models of the MatBench test suite⁶.

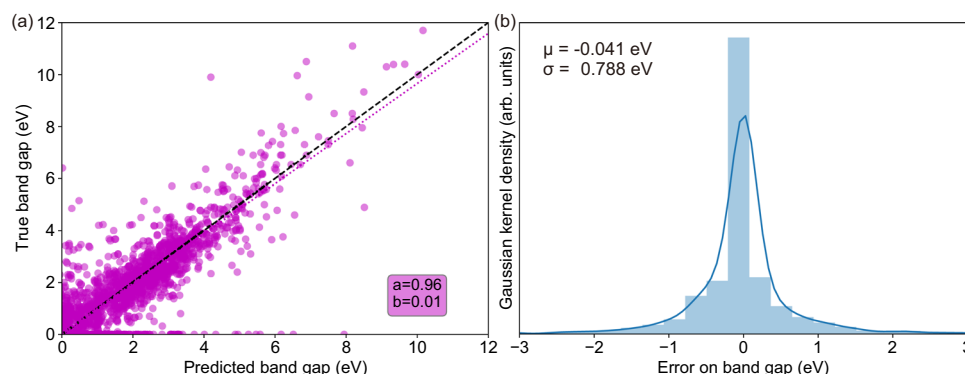


Fig. 9 Final results obtained by applying the denoising procedure to the whole dataset E. **a** True (i.e., experimental) vs. predicted band-gap parity plots. The corresponding plots produced by splitting the data according to the intersections with datasets P, H, S, and G are given in Fig. 1 and compared with the raw DFT data. **b** Corresponding error distribution plot.

Table 2. Most representative results for the different training approaches with and without denoising.

Approach	Denoiser	MAE (eV)		
		MODNet	MEGNet	MFGNet
<i>only-E</i>	—	0.446	0.701/0.844	
<i>all-together</i>	—	0.477	0.467/0.727	0.423/0.686
<i>all-together</i> (3-fi)	—	0.533	0.487/0.768	0.416/0.662
<i>onion</i>	—	0.403	0.397/0.734	0.395/0.725
<i>onion</i> (3-fi)	—	0.429	0.388/0.754	0.400/0.638
<i>onion</i>	<i>onion</i>	0.396	0.373/0.630	

The “3-fi” results are obtained using only the three datasets P, H, and E. The training sequence for the onion approach is PHSGE → PHSE → HSE → HE → E when all five datasets are considered and PHE → HE → E otherwise. For MEGNet and MFGNet, we report both the results for ordered/disordered compounds. The 5-fold MAEs on the predicted band gap are expressed in eV.

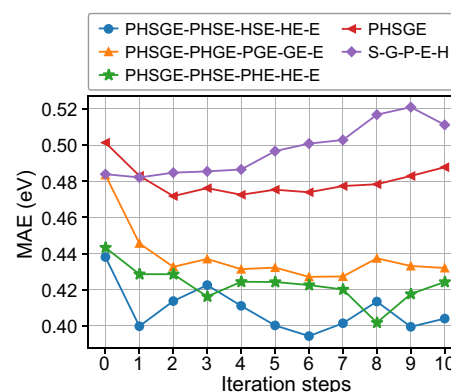


Fig. 10 Application of an iterative procedure for some representative training sequences. The first point of each line is the result obtained with the raw data (i.e., without data cleaning). Every subsequent point is obtained using the model corresponding to the previous point as the denoiser for cleaning the data.

Finally, a zero-shot learning (ZSL) is performed on the same set of crystals with site disorder as in ref. ²⁵. This is particularly interesting to analyze how the different methods can gain general knowledge about crystal structures. Unfortunately, this ZSL test cannot be performed with MODNet since the featurization process relies on Matminer⁴¹ which cannot handle crystals with partial occupancy.

From here on, we do not consider anymore all the possible training sequences and focus on the one that was found to be the best with MEGNet. Note that we have no clue whether it is also the best with MODNet, but it meets the heuristic defined above for obtaining a reasonable denoiser. As we only consider one route, we can afford a 5-fold train/test splitting of the dataset.

In ref. ²⁵, the 4-fidelity model (not using the dataset G) was found to perform better than the one with using all five datasets. Therefore, here, based on our analysis of the raw data, we also considered a 3-fidelity model relying only on the datasets P, H, and E. In this case, the training sequence was chosen to be PHE → HE → E.

Based on the interesting results obtained with the *onion* approach, we also investigated a variation of the original MFGNet approach. While the latter uses all the datasets simultaneously for training the model (just like in the *all-together* approach), we propose to adopt the *onion* approach as well.

The results obtained by applying our denoising procedure with both MODNet and MEGNet, as well as those calculated using MFGNet are summarized in Table 2. The results obtained with

only-E are indicated as a reference. Just like in ref. ²⁵, the 3-fidelity results of MFGNet are better than the 5-fidelity ones, both for the ordered/disordered crystals. While the *all-together* results with MEGNet are better than with the *only-E* approach (just like with the 2-fold train/test splitting), they are worse with MODNet for which the increase in the number of available compounds is less important than the noise in the data. Note that they are even worse in the case of the 3-fidelity model.

As already observed with MEGNet, the *onion* approach improves the predictions of MODNet compared to the *only-E* and *all-together* approaches. These improvements are less impressive for MODNet than for MEGNet given that the reference results (*only-E*) were already reasonably good. Very interestingly, combining the *onion* approach with MFGNet also leads to a clear improvement of the results (reduction of the MAE by 4–6%).

Finally, the denoising procedure also leads to a further improvement of the results with a reduction of the MAE by 6% for MEGNet and by 2% for MODNet. Compared to the *only-E* results, the MAE is reduced by 11% for MODNet and 46% for MEGNet (25% for the disordered compounds). It should be mentioned that the use of MEGNet leads to the best final results. This is consistent with the fact that it is better suited for large datasets while MODNet targets smaller datasets.

For the sake of completeness, we finally compare these results with those that can be obtained using transfer learning. The latter is known to improve the performance of machine learning models for small datasets. The transfer process can typically be performed

between different properties²⁶ or different fidelities²⁸. Here, we tested both options using AtomSets from the MAML project²⁹, which adopts MEGNet as a feature extractor followed by a simple MLP (Multi-Layer Perception) as regressor. We obtained an MAE of 0.668 eV when transferring from the formation energy and of 0.597 eV when transferring from all the DFT data used in this work. The results are better than the *only-E* ones, but they are worse than any of those obtained with *onion* approach in Table 2.

In this paper, we have introduced a method to take full advantage of the availability of multi-fidelity data and tested it thoroughly for the prediction of the band gap based on the structure. The method is based on an appropriate combination of all the data into a multistep training sequence and on a simple denoising procedure. For combining the data, we have compared four different training approaches (*only-E*, *all-together*, *one-by-one*, and *onion*). It turned out that the best one consists in training the model successively on different datasets resulting from, first, the combination of all available datasets and, then, of those obtained by removing one dataset at a time by increasing fidelity (from the poorest to the highest fidelity, hence, finishing with the true data). For the denoising procedure, we have tested a simple technique by which target values are replaced by the output of the selected denoiser when the former are too far (i.e., outside the interval defined by a cleaning threshold) from the latter. Other denoising procedures resulting in better results might be existing, but is left for future work. We have found that the denoising procedure improves the final results provided that a reasonable denoiser is chosen. Furthermore, based on our observations, we proposed a simple heuristic for the denoiser. Finally, we have investigated the effect of applying the denoising procedure several times until convergence.

The method proposed here provides a sensible way to improve the results that can be achieved when multi-fidelity data are available which is basically often the case in materials science given that accuracy in the data always comes at a cost. It thus has considerable potential of applications.

METHODS

Data

We use the same band gap datasets (four with DFT predictions and one with experimental measurements) as in ref. ²⁵. The four DFT datasets consist of calculations performed with the Perdew-Burke-Ernzerhof (PBE)⁴², Heyd-Scuseria-Ernzerhof (HSE)^{43,44}, strongly constrained and appropriately normed (SCAN)⁴⁵, and Gritsenko-Leeuwen-Lenthe-Baerends (GLLB)^{46,47} exchange-correlation functionals for 52348⁴⁸, 6030⁴⁴, 472⁴⁹, and 2290⁵⁰ crystalline compounds from the Materials Project, respectively. For the sake of simplicity, these datasets will be referred to as P, H, S, and G, respectively (i.e., using the first letter of the corresponding functional). The experimental dataset (referred to as E) comprises the band gaps of 2703 ordered crystals⁵¹, out of which 2401 could be assigned a most likely structure from the Materials Project⁵². The dataset of ref. ⁵¹ also contains 278 crystals with site disorder, which are used as a final ZSL test set.

Machine learning model

In the “Results” section, we use MEGNet v1.2.3 with its default hyperparameters and MFGNet²⁵ (v1.2.9), which is a multi-fidelity version of MEGNet. Given the large amount of calculations (all the training sequences), a 2-fold training-testing procedure is adopted. Training (including an inner 8:2 validation split for early stopping, *patience* = 10) is performed exclusively on one fold, while testing is done on the other hold-out fold. On a NVIDIA Tesla P100 graphics card, one *onion* tree training costs about 6 days while the *one-by-one* tree training costs about 35 days. The few

samples resulting in a Not-a-Number (NaN) with MEGNet are ignored.

For the denoising procedure, ϵ is treated as a hyper-parameter to be optimized. Here, $\epsilon = 0.3$ was found to be the most appropriate value.

It is worth noting that using the scaled DFT data for the denoising procedure (see Supplementary Table 6) or scaling the denoised data (see Supplementary Table 7) does not lead to any improvement compared to the results of obtained with the denoising approach using the raw DFT data. This was to be expected since the neural network that is used in the underlying model already contains multiple linear transformations, which are optimized during the training process.

In the “Discussion” section, we use MODNet¹⁵ (v0.1.12), MEGNet (v1.2.9), MFGNet²⁵ (v1.2.9), and MAML²⁹ (v2022.6.11) to validate our observations about the effect of denoising and to compare with transfer learning. Default hyperparameters are used for MEGNet, MFGNet, and MAML, while MODNet uses a Genetic Algorithm to optimize the hyperparameters⁵³. Given that there are fewer calculations, a 5-fold training-testing procedure is adopted here. For the denoising procedure, $\epsilon = 0.4$ was found to be the most appropriate value.

For models that include different training stages (*one-by-one* and *onion*), weights are transferred from one training stage to the other with no change in hyperparameters.

DATA AVAILABILITY

We use the same band gap datasets (four with DFT predictions and one with experimental measurements) as in ref. ²⁵. They are available at <https://doi.org/10.6084/m9.figshare.13040330>.

CODE AVAILABILITY

The source code is available on <https://github.com/liuxiaotong15/denoise> for the tree training with MEGNet, on https://github.com/ppdebreuck/onion_modnet for MODNet validation.

Received: 21 April 2022; Accepted: 1 November 2022;

Published online: 11 November 2022

REFERENCES

- Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
- Lusher, S. J., McGuire, R., van Schaik, R. C., Nicholson, C. D. & de Vlieg, J. Data-driven medicinal chemistry in the era of big data. *Drug Discov. Today* **19**, 859–868 (2014).
- Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
- Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* **5**, 83 (2019).
- Choudhary, K. et al. Recent advances and applications of deep learning methods in materials science. *Npj Comput. Mater.* **8**, 59 (2022).
- Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *Npj Comput. Mater.* **6**, 138 (2020).
- Cao, G. et al. Artificial intelligence for high-throughput discovery of topological insulators: the example of alloyed tetradymites. *Phys. Rev. Mater.* **4**, 034204 (2020).
- Pyzer-Knapp, E. O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J. & Aspuru-Guzik, A. What is high-throughput virtual screening? A perspective from organic materials discovery. *Annu. Rev. Mater. Res.* **45**, 195–216 (2015).
- Ghiandoni, G. M. et al. Development and application of a data-driven reaction classification model: comparison of an electronic lab notebook and medicinal chemistry literature. *J. Chem. Inf. Model* **59**, 4167–4187 (2019).
- Rupp, M., Tkatchenko, A., Müller, K.-R. & Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).

11. Tsubaki, M. & Mizoguchi, T. Fast and accurate molecular property prediction: learning atomic interactions and potentials with neural networks. *J. Phys. Chem. Lett.* **9**, 5733–5741 (2018).
12. Kuzminykh, D. et al. 3d molecular representations based on the wave transform for convolutional neural networks. *Mol. Pharmaceutics* **15**, 4378–4385 (2018).
13. Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *Npj Comput. Mater.* **7**, 77 (2021).
14. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
15. De Breuck, P.-P., Hautier, G. & Rignanese, G.-M. Materials property prediction for limited datasets enabled by feature selection and joint learning with modnet. *Npj Comput. Mater.* **7**, 83 (2021).
16. Maurer, R. J. et al. Advances in density-functional calculations for materials modeling. *Annu. Rev. Mater. Res.* **49**, 1–30 (2019).
17. Perdew, J. P. & Levy, M. Physical content of the exact Kohn-Sham orbital energies: band gaps and derivative discontinuities. *Phys. Rev. Lett.* **51**, 1884 (1983).
18. Hautier, G., Ong, S. P., Jain, A., Moore, C. J. & Ceder, G. Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability. *Phys. Rev. B* **85**, 155208 (2012).
19. Bartel, C. J., Weimer, A. W., Lany, S., Musgrave, C. B. & Holder, A. M. The role of decomposition reactions in assessing first-principles predictions of solid stability. *Npj Comput. Mater.* **5**, 4 (2019).
20. Bartel, C. J. et al. A critical examination of compound stability predictions from machine-learned formation energies. *Npj Comput. Mater.* **6**, 97 (2020).
21. Morales-García, Á., Valero, R. & Illas, F. An empirical, yet practical way to predict the band gap in solids by using density functional band structure calculations. *J. Phys. Chem. C* **121**, 18862–18866 (2017).
22. Greenman, K. P., Green, W. H. & Gomez-Bombarelli, R. Multi-fidelity prediction of molecular optical peaks with deep learning. *Chem. Sci.* **13**, 1152–1162 (2022).
23. Batra, R., Pilania, G., Uberuaga, B. P. & Ramprasad, R. Multifidelity information fusion with machine learning: a case study of dopant formation energies in hafnia. *ACS Appl. Mater. Interfaces* **11**, 24906–24918 (2019).
24. Egorova, O., Hafizi, R., Woods, D. C. & Day, G. M. Multifidelity statistical machine learning for molecular crystal structure prediction. *J. Phys. Chem. A* **124**, 8065–8078 (2020).
25. Chen, C., Zuo, Y., Ye, W., Li, X. & Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat. Comput. Sci.* **1**, 46–53 (2021).
26. Gupta, V. et al. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nat. Commun.* **12**, 1 (2021).
27. Tran, A., Tranchida, J., Wildey, T. & Thompson, A. P. Multi-fidelity machine-learning with uncertainty quantification and bayesian optimization for materials design: Application to ternary random alloys. *J. Chem. Phys.* **153**, 074705 (2020).
28. Hutchinson, M. L. et al. Overcoming data scarcity with transfer learning. Preprint at <http://arxiv.org/abs/1711.05099> (2017).
29. Chen, C. & Ong, S. P. Atomsets as a hierarchical transfer learning framework for small and large materials datasets. *Npj Comput. Mater.* **7**, 1 (2021).
30. Huang, J., Qu, L., Jia, R. & Zhao, B. O2u-net: a simple noisy label detection approach for deep neural networks. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 3325–3333 (2019).
31. Oja, E. On the convergence of an associative learning algorithm in the presence of noise. *Int. J. Syst. Sci.* **11**, 629–640 (1980).
32. Angluin, D. & Laird, P. Learning from noisy examples. *Mach. Learn.* **2**, 343–370 (1988).
33. Han, B. et al. Co-teaching: robust training of deep neural networks with extremely noisy labels. in *Proceedings of the 32nd International Conference on Neural Information Processing Systems* 8536–8546 (2018).
34. Bengio, Y., Louradour, J., Collobert, R. & Weston, J. Curriculum learning. in *Proceedings of the 26th Annual International Conference on Machine Learning* 41–48 (2009).
35. Guo, S. et al. Curriculumnet: Weakly supervised learning from large-scale web images. *Proceedings of the European Conference on Computer Vision (ECCV)* 135–150 (2018).
36. Jiang, L., Zhou, Z., Leung, T., Li, L.-J. & Fei-Fei, L. Mentornet: learning data-driven curriculum for very deep neural networks on corrupted labels. in *International Conference on Machine Learning*, 2304–2313 (PMLR, 2018).
37. Donoho, D. L. De-noising by soft-thresholding. *IEEE Trans. Inf. Theory* **41**, 613–627 (1995).
38. Donoho, D. L. & Johnstone, J. M. Ideal spatial adaptation by wavelet shrinkage. *biometrika* **81**, 425–455 (1994).
39. Lejaeghere, K., Van Speybroeck, V., Van Oost, G. & Cottenier, S. Error estimates for solid-state density-functional theory predictions: an overview by means of the ground-state elemental crystals. *Crit. Rev. Solid State Mater. Sci.* **39**, 1–24 (2014).
40. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
41. Ward, L. et al. Matminer: an open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60 (2018).
42. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
43. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **118**, 8207–8215 (2003).
44. Jie, J. S. et al. A new MaterialGo database and its comparison with other high-throughput electronic structure databases for their predicted energy band gaps. *Sci. China Tech. Sci.* **62**, 1423–1430 (2019).
45. Sun, J., Ruzsinszky, A. & Perdew, J. P. Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.* **115**, 036402 (2015).
46. Gritsenko, O., van Leeuwen, R., van Lenthe, E. & Baerends, E. J. Self-consistent approximation to the Kohn-Sham exchange potential. *Phys. Rev. A* **51**, 1944 (1995).
47. Kuisma, M., Ojanen, J., Enkovaara, J. & Rantala, T. Kohn-Sham potential with discontinuity for band gap materials. *Phys. Rev. B* **82**, 115106 (2010).
48. Jain, A. et al. Commentary: The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
49. Borlido, P. et al. Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids. *J. Chem. Theory Comput.* **15**, 5069–5079 (2019).
50. Castelli, I. E. et al. New light-harvesting materials using accurate and efficient bandgap calculations. *Adv. Energy Mater.* **5**, 1400915 (2015).
51. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
52. Kingsbury, R. et al. Performance comparison of r2SCAN and SCAN metaGGA density functionals for solid materials via an automated, high-throughput computational workflow. *Phys. Rev. Mater.* **6**, 013801 (2022).
53. De Breuck, P.-P., Heymans, G. & Rignanese, G.-M. Accurate experimental band gap predictions with multifidelity correction learning. *J. Mater. Inf.* **2**, 10 (2022).

ACKNOWLEDGEMENTS

X.T.L. is grateful for the funding support from National Natural Science Foundation of China (No. 22002008, 22203008). P.-P.D.B. and G.-M.R. are grateful to the F.R.S.-FNRS for financial support. We also thank to the authors of MEGNet and MFGNet. They inspired us and help us fix some issues on GitHub. X.T.L. is also thankful to Prof. Ning Li, Dr. Tao Yang, Prof. Xiaodong Wen, Dr. Yurong He, and Mr. Enhu Diao for providing feedback and help on this work.

AUTHOR CONTRIBUTIONS

X.T.L. and G.-M.R. conceived the idea and designed the work. X.T.L. implemented the models and performed the analysis. P.-P.D.B. provided the data and validated the denoising performance with MODNet. L.H.W. helped with the data analysis and figure plotting. G.-M.R. supervised the project. All authors wrote the manuscript and contributed to the discussion and revision.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-022-00925-1>.

Correspondence and requests for materials should be addressed to Gian-Marco Rignanese.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022