A simple denoising approach to exploit multi-fidelity data for machine learning materials properties (Supporting Information)

Xiaotong Liu,[†] Pierre-Paul De Breuck,[‡] Linghui Wang,[†] and Gian-Marco Rignanese[*,‡]

[†]*Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, P. R. China*

[‡]*UCLouvain, Chemin des Étoiles 8, Louvain-la-Neuve 1348, Belgium*

E-mail: gian-marco.rignanese@uclouvain.be

A Supplementary datasets analysis

In this section, we provide various supplementary analyses of the datasets. These aim both at checking for possible biases in the latter and at delivering more detail about them.

A.1 Analysis by Element

The distribution of the chemical elements in each dataset is shown in Fig. S1. Given that the number of structures in the dataset P is considerably larger than in the other datasets, we use it as a reference (normalized to one with the actual number being indicated on top) and we adopt logarithm to indicate the other numbers. The PBE dataset contains 87 different chemical elements which cover all the elements present in the other datasets (75 for H, 66 for S, 63 for G, and 80 for E). All the chemical elements in the dataset E are included in at least one of the DFT datasets ensures that the training procedure and the prediction models of the current work are reasonable.

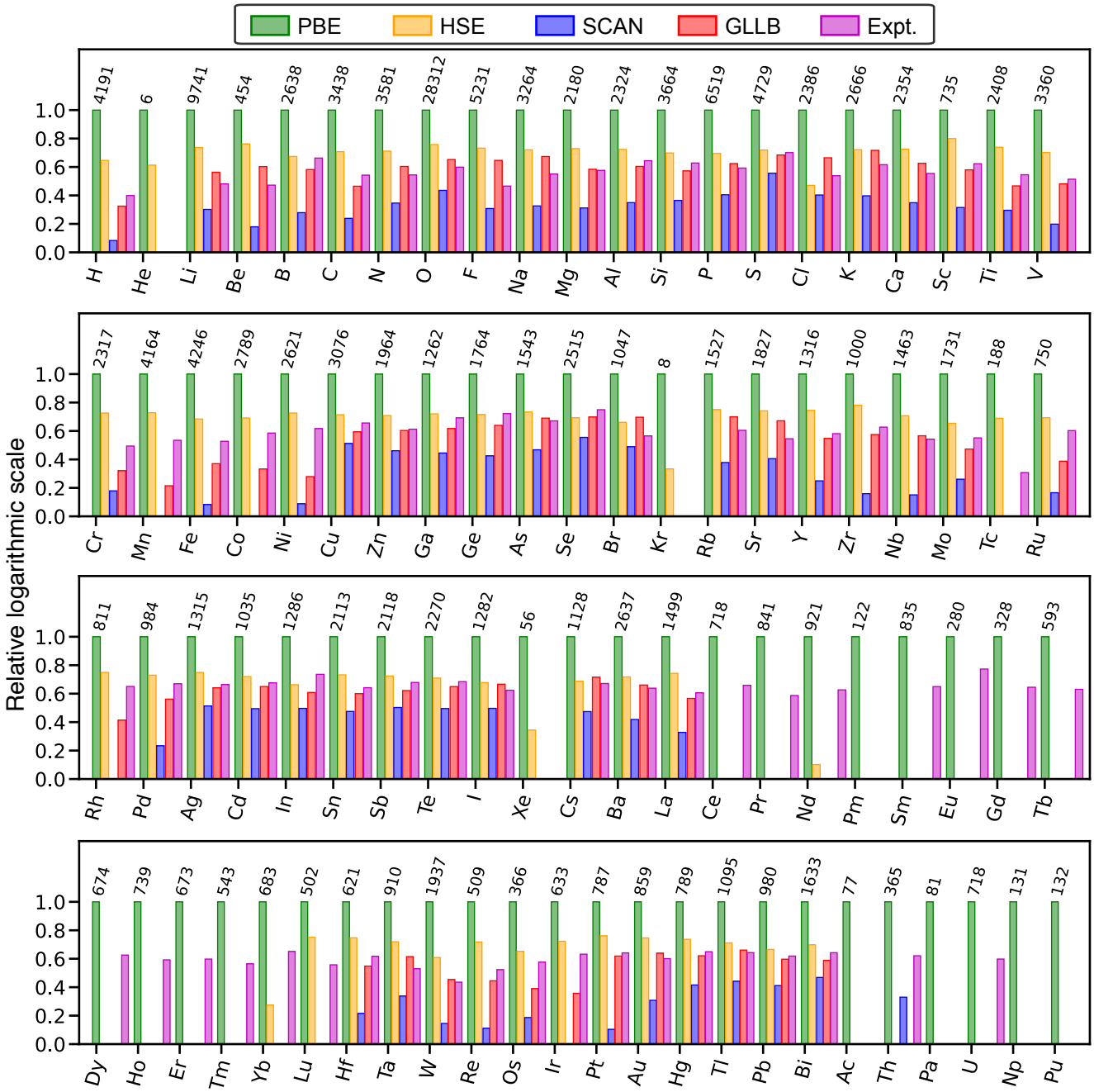

Figure S1: Distribution of the chemical elements in the different datasets. The number over the green bar is the count of structures with that element in the dataset P. For one given element, the height of the bars gives the relative logarithmic ratio of that element in the different datasets.

A.2 Venn diagram and Upset plot analysis

To compare the structures available in the different datasets, we first use their Material Project identificators (MP-ids). For the experimental dataset, 2401 of the 2703 compounds could be assigned a most likely structure from the Materials Project.[1] The size of each dataset and of their intersections are shown in Figs. S2, S3, S4 and S5.

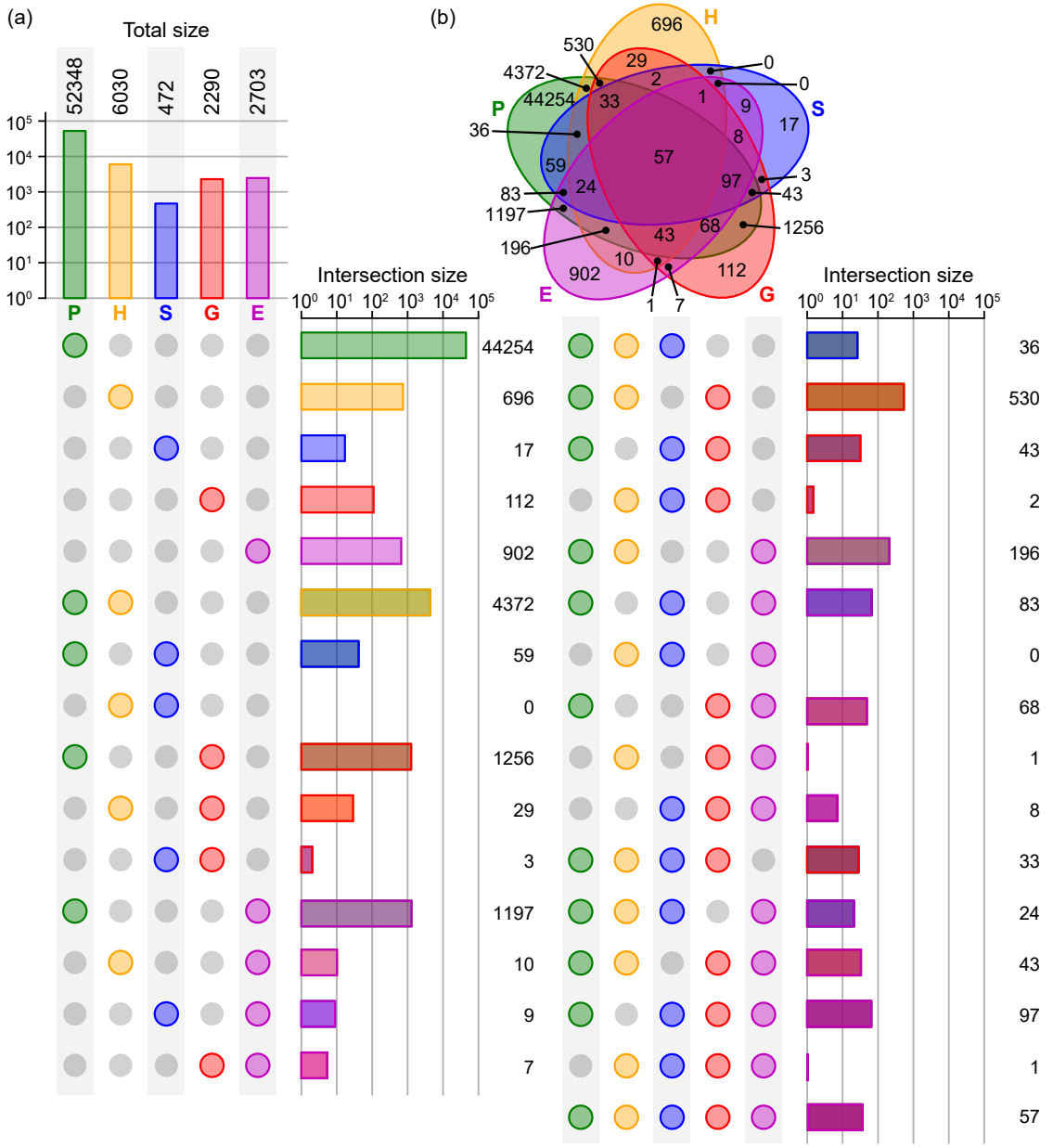

Figure S2: (a) Upset plot and (b) Venn diagram illustrating the intersections between the five datasets (P, H, S, G, and the part of E with likely-mpid). All the bar plots are in logarithmic scale. The intersections are illustrated more clearly in Figs. S3, S4, and S5 below.

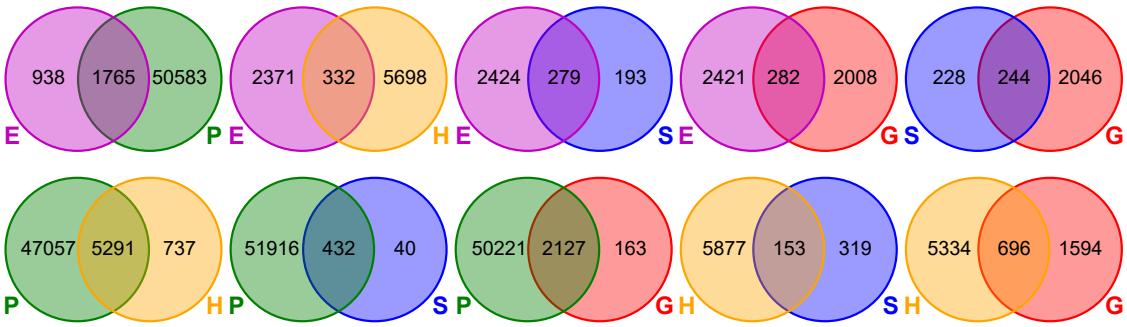

Figure S3: The Venn diagrams for any two of the five datasets.

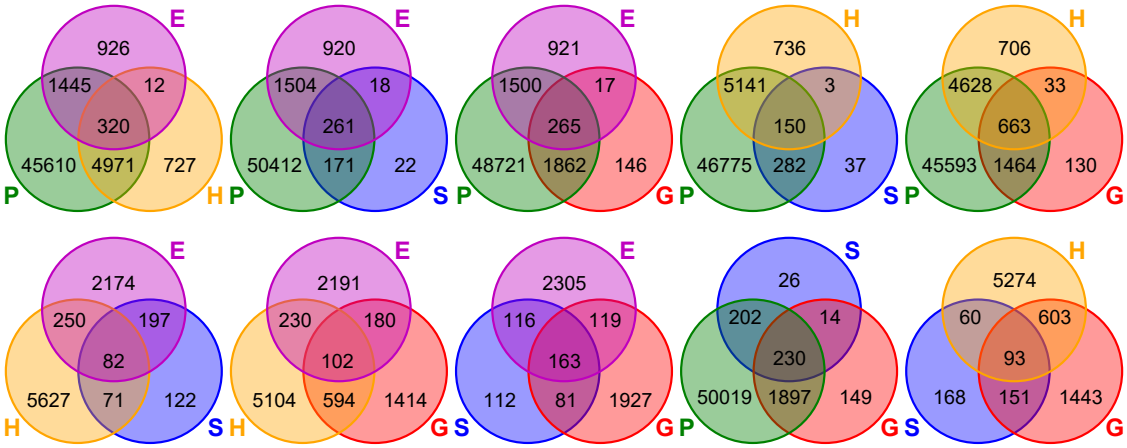

Figure S4: The Venn diagrams for any three of the five different datasets.

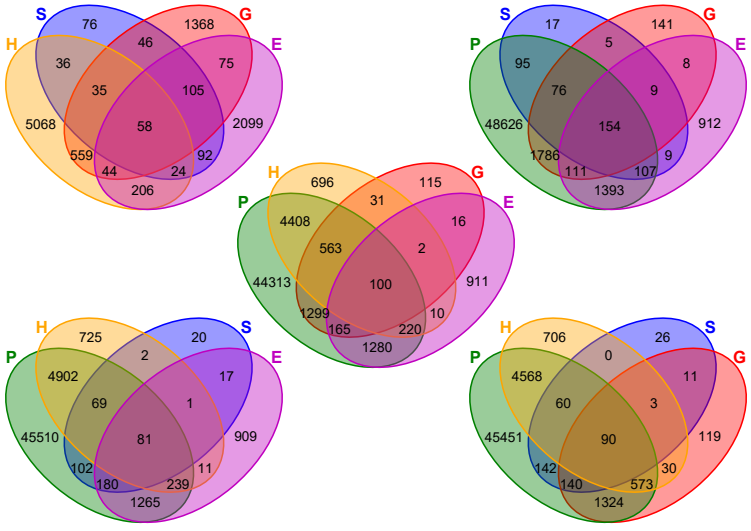

Figure S5: The Venn diagrams for any four of the five different datasets.

A.3 2D distribution of the structures

For the structures that have not been assigned an MP-id, it is, however, not possible to figure out how similar it is compared to other structures in the dataset. To overcome this limitation, we first extract a 96D vector for each structure from a median layer of the MEGNet model. Subsequently , we perform a dimensionality reduction through Principal Component Analysis (PCA) and we plot any two of the top three PCA directions. The resulting distributions of the data points are shown in Figs. S6, S7 and S8. Given that, in the PCA approach, the 0^{th} dimension has the largest variance of all, the mean value in that direction is far from 0.5 after normalization. The plots provide valuable information about the coverage of the chemical space by all the datasets. The same trend emerges that P is the most diverse and it covers almost all structures in the other datasets.

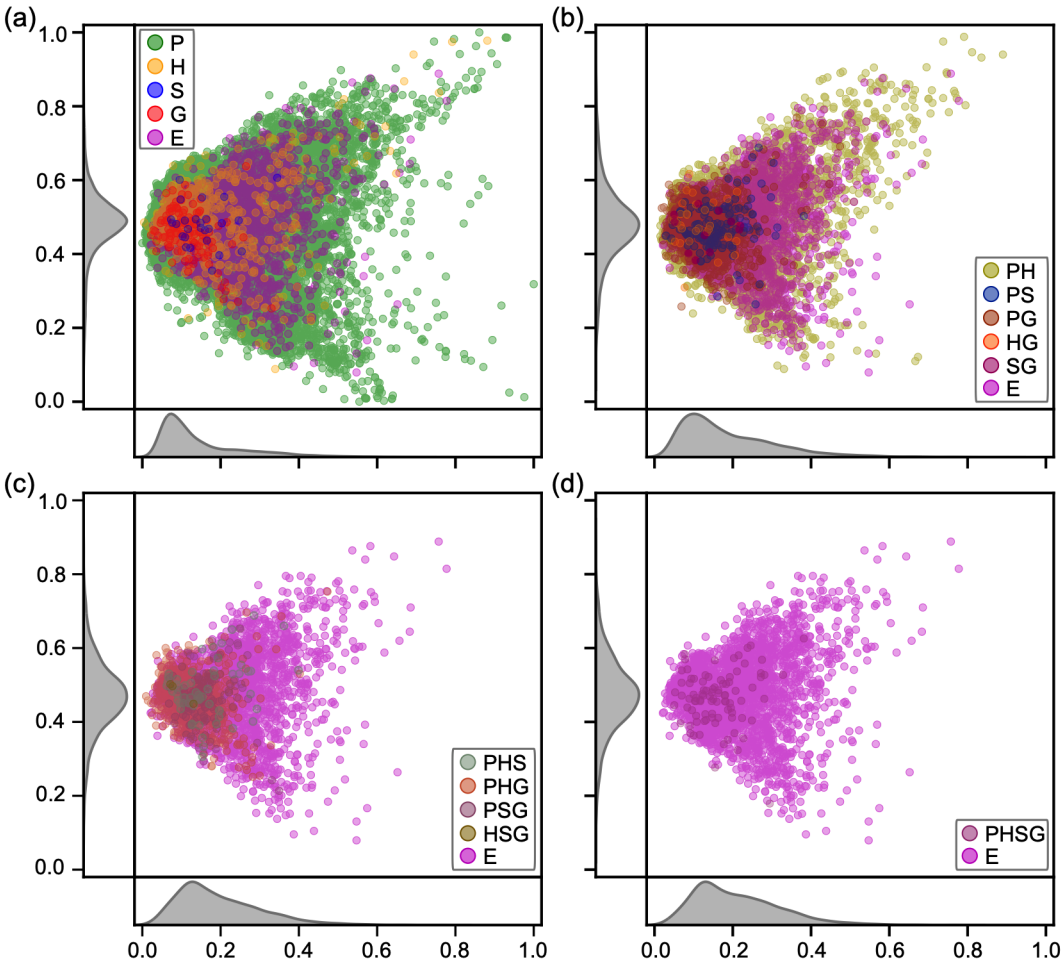

Figure S6: Distribution of the datapoints along the 0^{th} and 1^{st} PCA directions for the different datasets. E corresponds to the complete experimental dataset in each panel. In panel (a), P, H, G, and S refer only to the structures that are not included in the other datasets. The panels. (b), (c), and (d) show the structures that belong to 2, 3, and 4 datasets, respectively. The distributions of the data in each direction are reported in the subplots on the side and at the bottom.

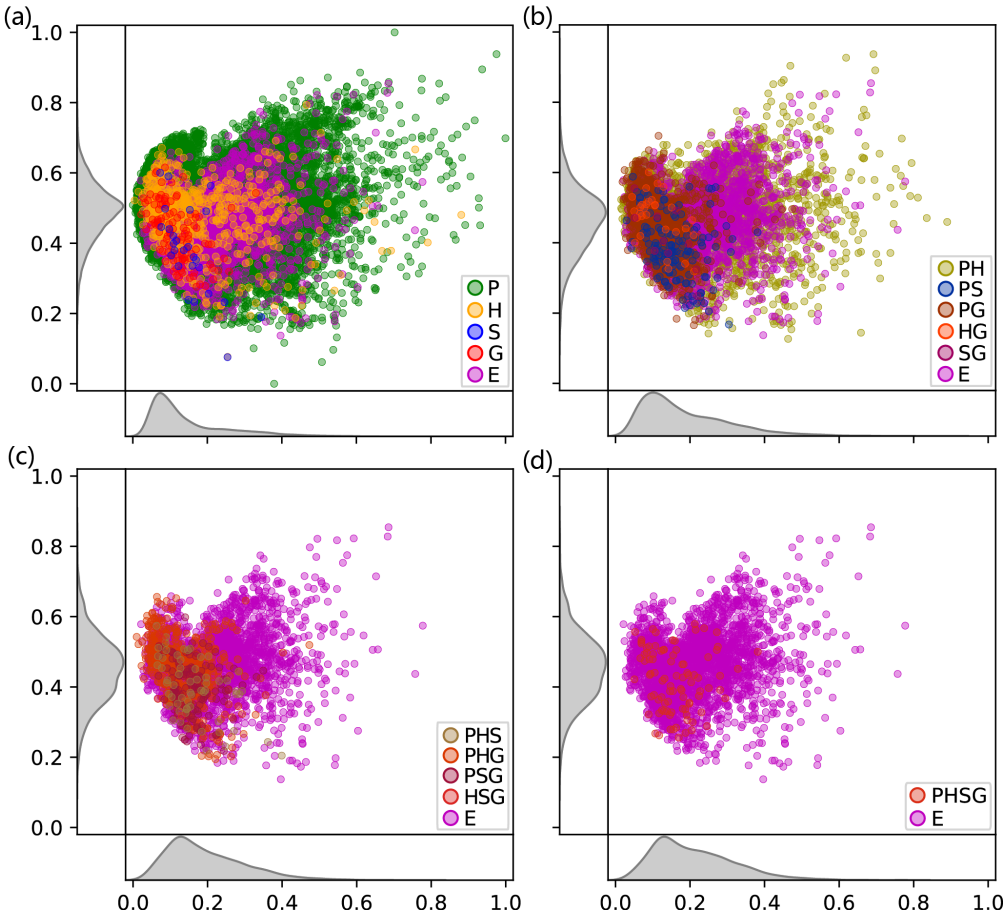

Figure S7: Distribution of the datapoints along the 0^{th} and 2^{nd} PCA directions for the different datasets. E corresponds to the complete experimental dataset in each panel. In panel (a), P, H, G, and S refer only to the structures that are not included in the other datasets. The panels. (b), (c), and (d) show the structures that belong to 2, 3, and 4 datasets, respectively. The distributions of the data in each direction are reported in the subplots on the side and at the bottom.

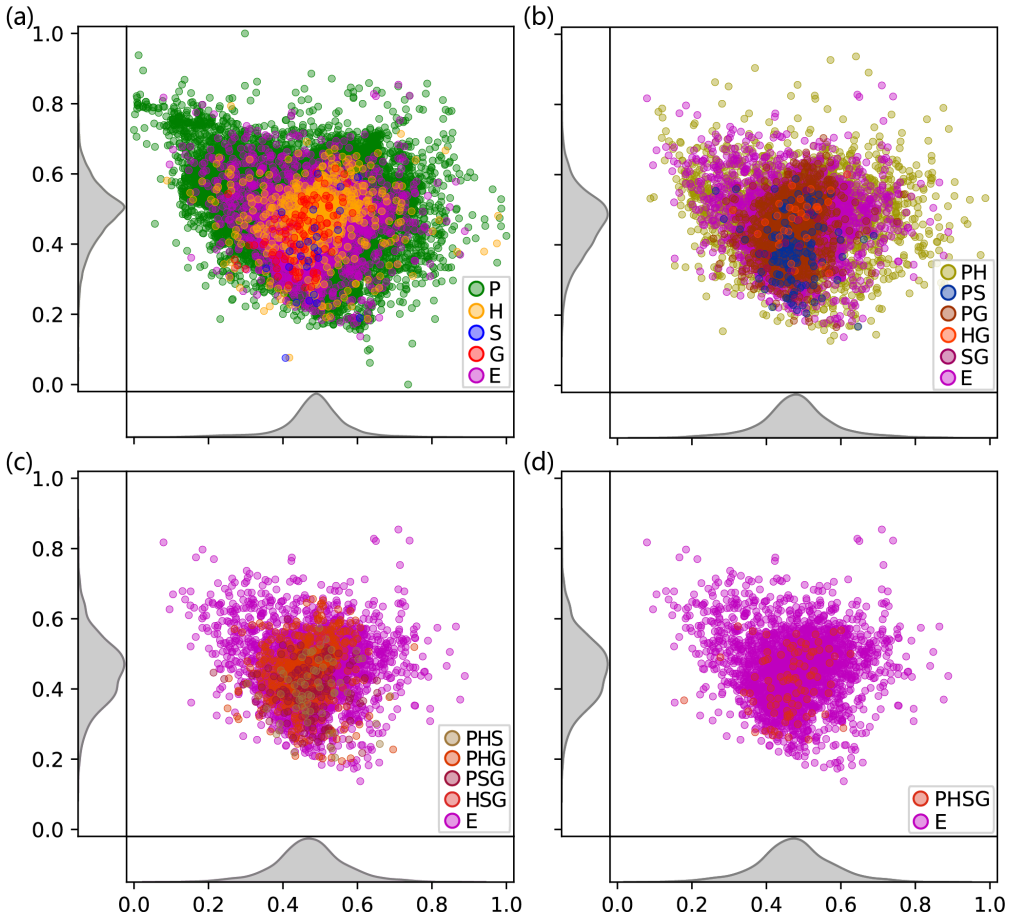

Figure S8: Distribution of the datapoints along the 1st and 2nd PCA directions for the different datasets. E corresponds to the complete experimental dataset in each panel. In panel (a), P, H, G, and S refer only to the structures that are not included in the other datasets. The panels. (b), (c), and (d) show the structures that belong to 2, 3, and 4 datasets, respectively. The distributions of the data in each direction are reported in the subplots on the side and at the bottom.

A.4 KL divergence analysis

Without having to perform dimensionality reduction, the Kullback-Leibler (KL) divergence can also be used to obtain a measure the similarity of two discrete distributions in the previously mentioned 96D space. Typically, for two distributions P and Q in the same probability space χ, the KL divergence $D_{\mathrm{KL}}(P||Q)$ is defined by:

$$D_{\mathrm{KL}}(P||Q) = \sum_{x \in \chi} P(x) \ln \frac{P(x)}{Q(x)}. \tag{S1}$$

Given that $D_{\mathrm{KL}}(P||Q)$ is not equal to $D_{\mathrm{KL}}(Q||P)$, we adopt:

$$D(P,Q) = \frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2} \tag{S2}$$

as the measure the similarity. The smaller $D(P,Q)$, the more two distributions are similar with $D(P,P)$=0. We list some typical $D(P,Q)$ in Table S1.

Table S1: Bi-directional KL divergence between the distribution of the different datasets, as represented by the MEGNet median layer in 96D space. X refers to one of the DFT datasets (P, S, H and G), while $\overline{X}$ indicates the sum of all the datasets but X. $D(E,X)$ is a measure of the similarity between the structures in the experimental and DFT datasets. H with the smallest KL divergence is the most similar to E. $D(\overline{X},X)$ is a measure of the similarity between structures in a DFT dataset and its complementary set. S with the highest KL divergence is the least similar to all other, and hence brings in the most new information.

	X			
	P	S	H	G
$D(E,X)$	0.067	0.105	0.031	0.129
$D(\overline{X},X)$	0.017	0.116	0.018	0.061

A.5 Distribution of the band gap predictions and errors for the different DFT functionals

Finally, it is interesting to compare the different DFT predictions for the same structures (even when the experimental data is not available). An obvious approach to do so is to analyze the distribution of the data points in the different intersections (based on the MP-ids). In Fig. S9, we report the average and standard deviation values for the different intersectionis of the datasets. As a general trend, we observe that P $<$ S $<$ H $<$ G. This trend holds for both the average and standard deviation. It shows the order of absolute value of the band gaps in the different datasets. A similar representation for the distribution of the errors can be found in Fig. S10.

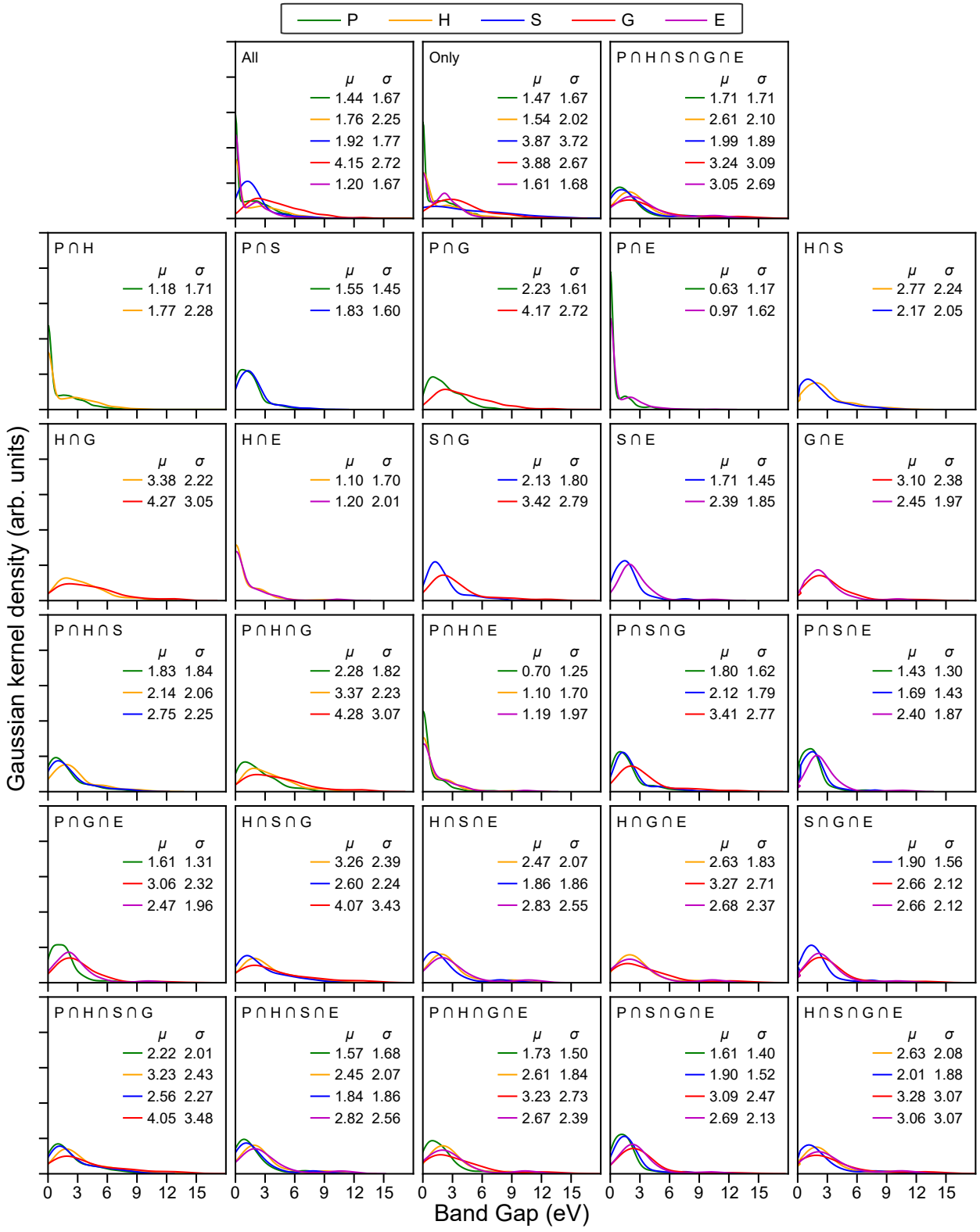

Figure S9: An analysis of band gap value distribution with all full and sub datasets in details.

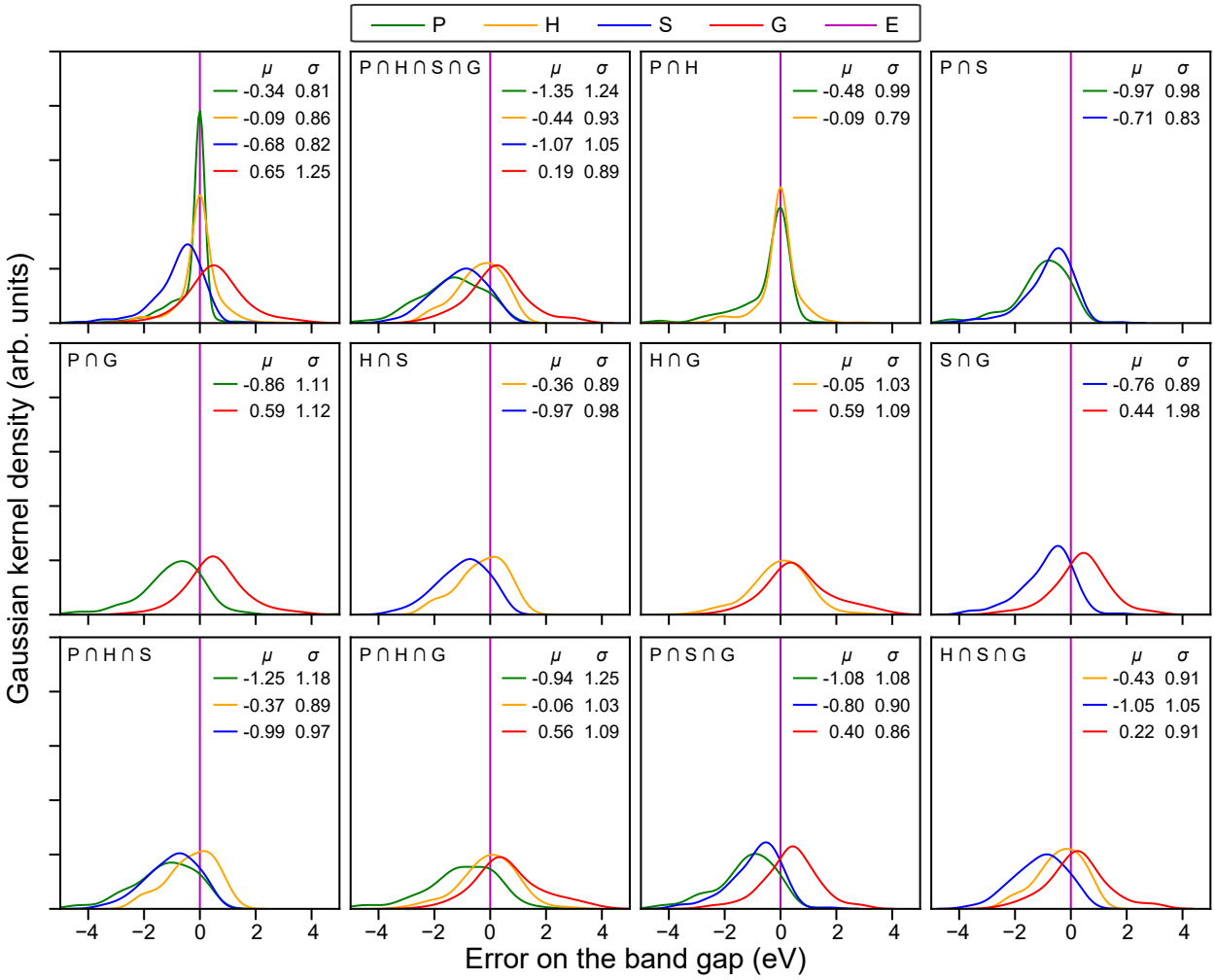

Figure S10: An analysis of band gap prediction error (X - E) distribution with all full and sub datasets in details.

B Onion tree training results from different denoised data

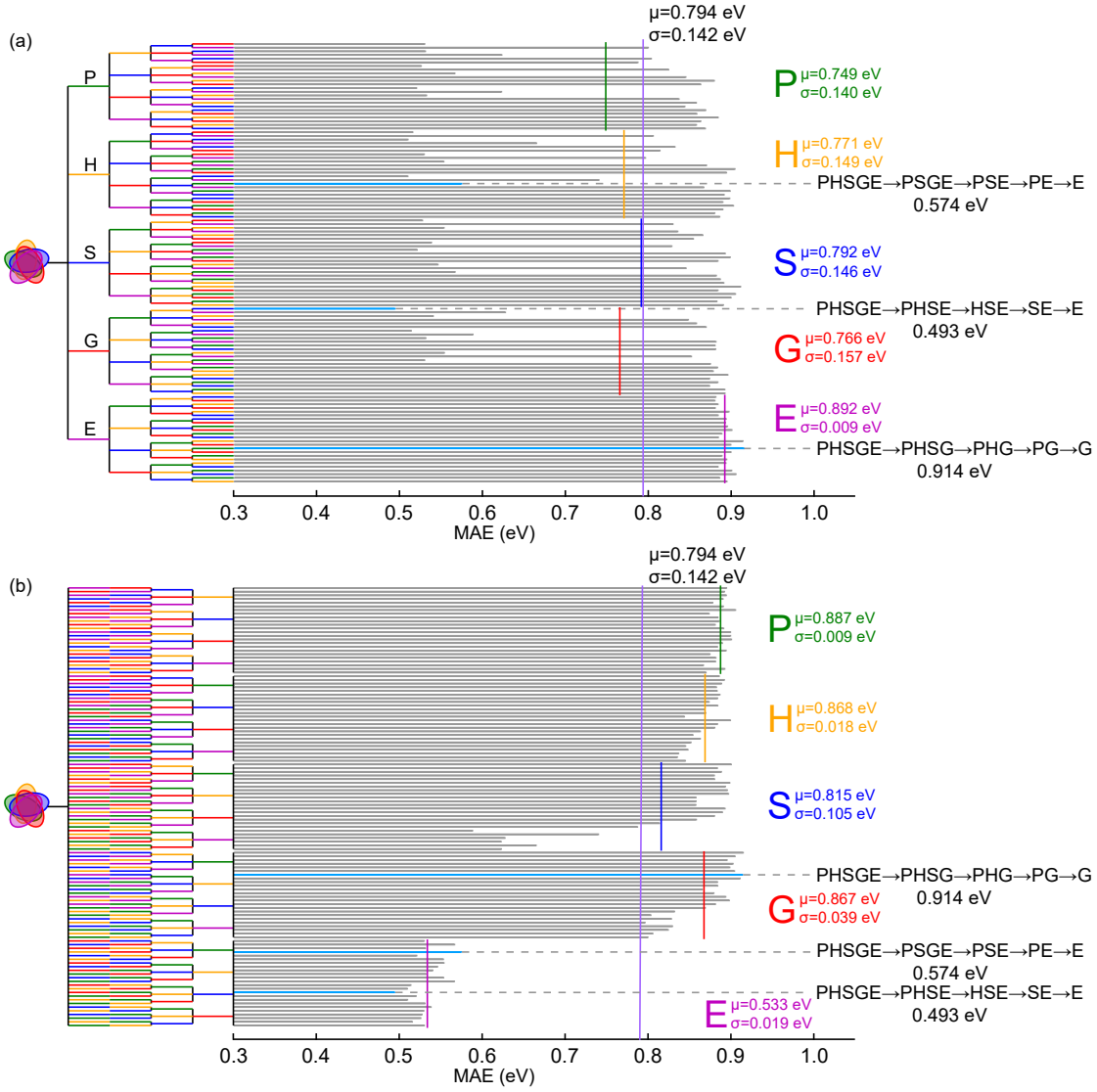

Figure S11: MAE results obtained on the data cleaned with the worst model (PHSGE→PHSG→PSG→SG→G with MAE=0.916 eV) of Fig. 6 using the *onion* training approach for all possible dataset orders: (a) gathered according to the first dataset used and (b) grouped following the last dataset used. The global average of the MAE is shown by a vertical solid purple line (μ=0.794 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.

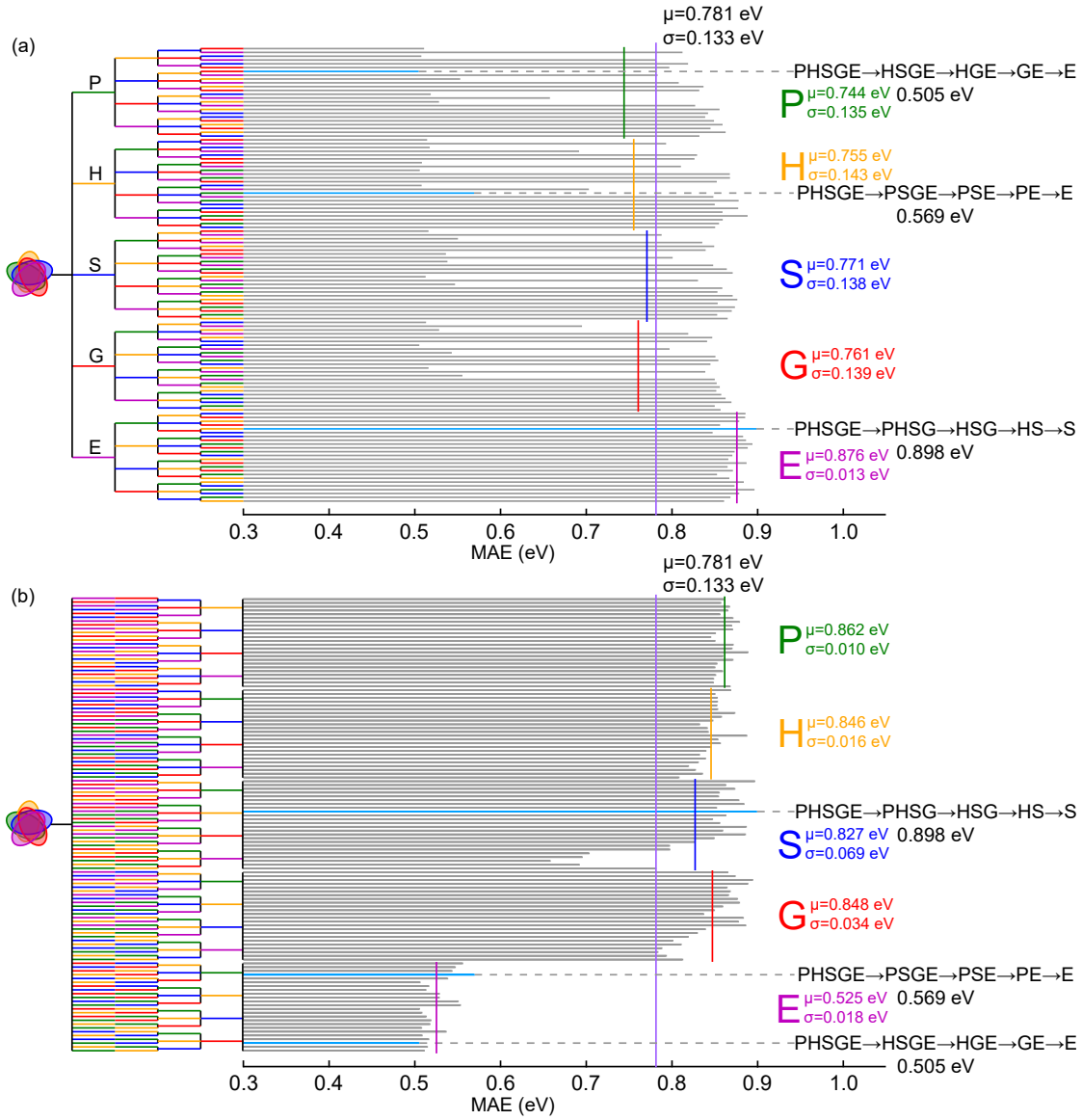

Figure S12: MAE results obtained on the data cleaned with the 2^{nd} worst model (PHSGE→PHSG→PSG→PG→G with MAE=0.889 eV) of Fig. 6 using the *onion* training approach for all possible dataset orders: (a) gathered according to the first dataset used and (b) grouped following the last dataset used. The global average of the MAE is shown by a vertical solid purple line (μ=0.781 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.

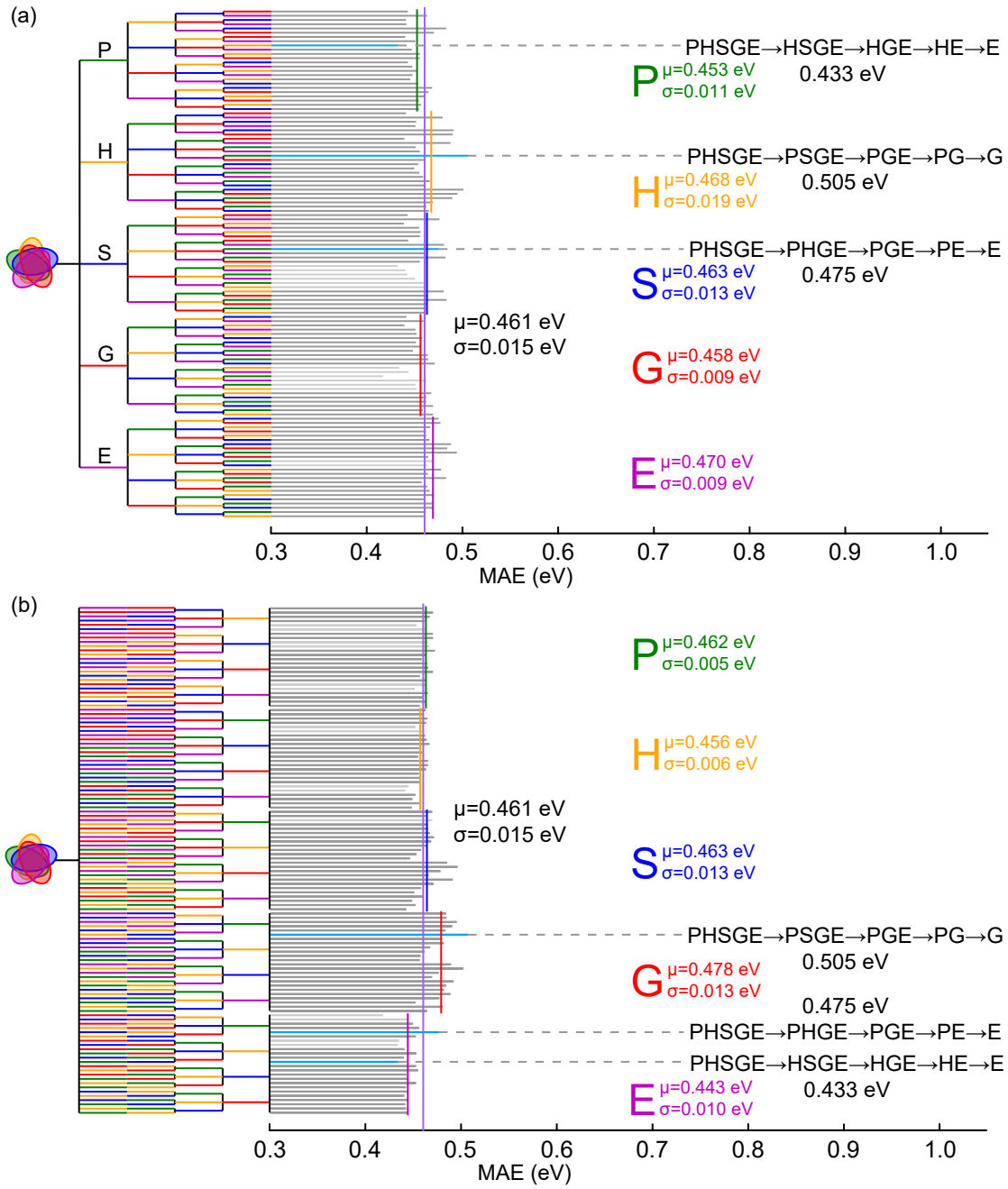

Figure S13: MAE results obtained on the data cleaned with a rather poor model among those ending with E (PHSGE→PHGE→PGE→GE→E with MAE=0.483 eV) of Fig. 6 using the *onion* training approach for all possible dataset orders: (a) gathered according to the first dataset used and (b) grouped following the last dataset used. The global average of the MAE is shown by a vertical solid purple line (μ=0.461 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.

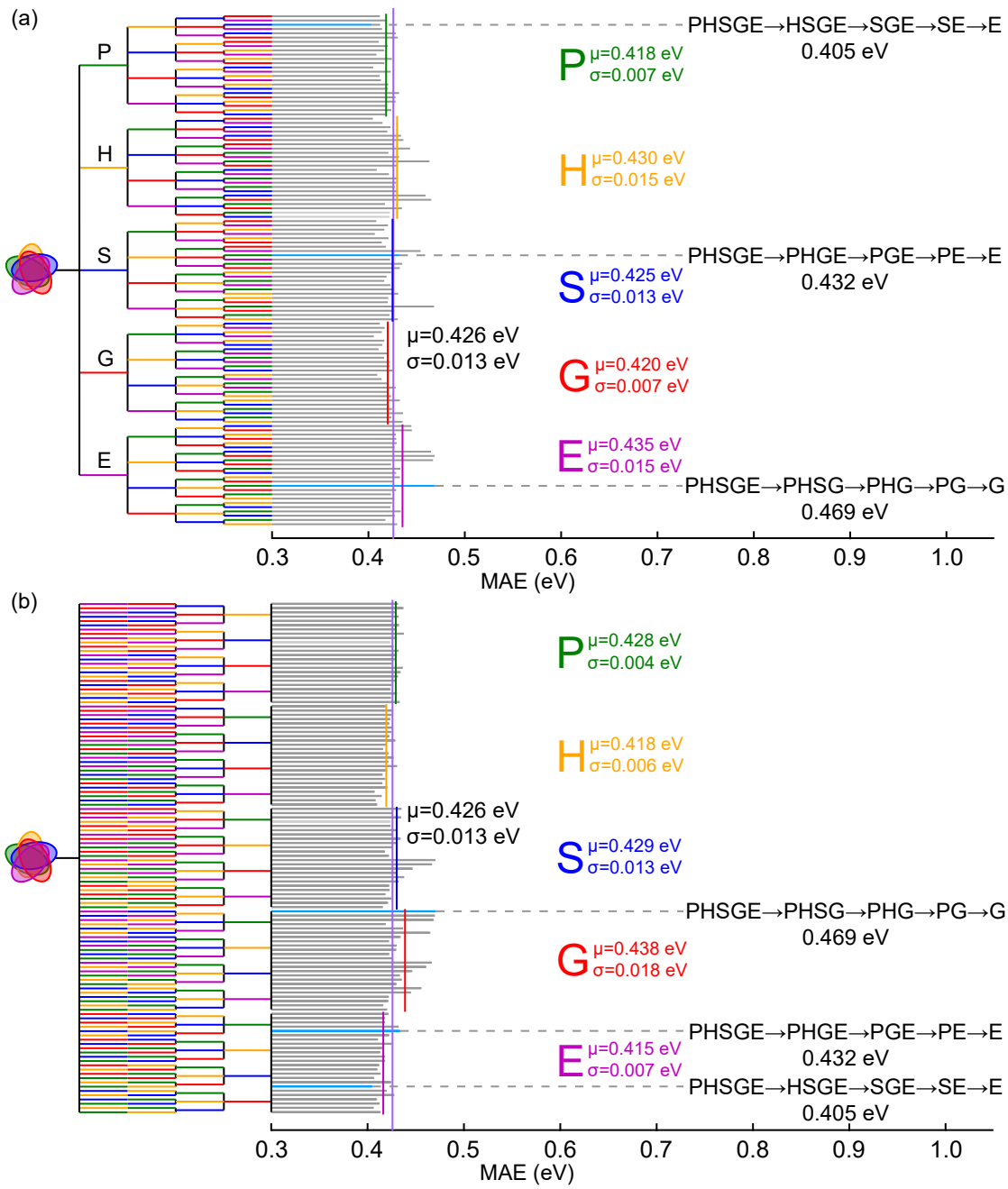

Figure S14: MAE results obtained on the data cleaned with a rather good model among those ending with E (PHSGE→PHSE→PHE→HE→E with MAE=0.443 eV) of Fig. 6 using the *onion* training approach for all possible dataset orders: (a) gathered according to the first dataset used and (b) grouped following the last dataset used. The global average of the MAE is shown by a vertical solid purple line (μ=0.426 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.

References

(1) Kingsbury, R.; Gupta, A. S.; Bartel, C. J.; Munro, J. M.; Dwaraknath, S.; Horton, M.; Persson, K. A. Performance comparison of r2SCAN and SCAN metaGGA density functionals for solid materials via an automated, high-throughput computational workflow. *Phys. Rev. Mater.* **2022**, *6*, 013801.