Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PBE-assisted machine learning of HSE06+SOC photovoltaic efficiency in halide perovskites

Azhar Ali Haidry

aa.haidry@nuaa.edu.cn

Nanjing University of Aeronautics and Astronautics

Courtney Rutendo Mandebvu

Nanjing University of Aeronautics and Astronautics

Qawareer Fatima

Nanjing University of Aeronautics and Astronautics

Research Article

Keywords: Halide perovskites, Machine learning, Multi-fidelity, Spectroscopic limited maximum efficiency, Hybrid functional, Grouped cross-validation

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11482596/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Hybrid-functional calculations with spin–orbit coupling (HSE06 + SOC) provide more accurate electronic properties of halide perovskites than semilocal PBE calculations but are substantially more expensive. Here, we assess whether low-cost PBE information improves machine-learning (ML) prediction of HSE06 + SOC band gaps and spectroscopic limited maximum efficiency (SLME). Using a public dataset of 243 pure and mixed ABX_3 perovskites, we compare seven ML model families using nested cross-validation with random, chemical-system-grouped, and B-site-grouped splits. PBE inputs substantially improve SLME prediction: the mean absolute error decreases from 0.68 to 0.45 percentage points on chemical-system-grouped splits, with the best model achieving 0.36 percentage points. Δ-learning provides no advantage over direct learning. However, performance deteriorates when the B-site chemistry is absent from training, and conformal intervals become under-calibrated. An audit of the source labels further shows that the HSE06 + SOC SLME is strongly constrained by the HSE06 + SOC band gap and approaches a saturation near 16%, because the reported spectrum was generated by shifting the PBE absorption spectrum. These findings highlight both the value and limitations of PBE-assisted ML for multi-fidelity prediction of perovskite properties.

1 Introduction

Halide perovskites ABX_3 absorb light strongly, have long carrier diffusion lengths and can be processed from solution, and the efficiency of perovskite solar cells has risen quickly since the first reports [1, 2]. Mixing at the A, B and X sites tunes the band gap and the stability, and Sn, Ge and alkaline-earth cations are studied as replacements for Pb to reduce toxicity [3]. The composition space of mixed perovskites is large, so first-principles screening is widely used to choose candidates before synthesis. A common screening metric is the spectroscopic limited maximum efficiency (SLME), which extends the Shockley–Queisser limit by using the computed absorption spectrum and the nature of the band gap [4, 5].

Screening results depend on the exchange–correlation functional. Semi-local PBE [6] underestimates band gaps systematically. The screened hybrid HSE06 [7, 8], combined with spin–orbit coupling (SOC), which is essential for Pb- and Sn-based perovskites [9], is more reliable but much more expensive. Multi-fidelity ML tries to bridge the two levels by learning high-fidelity properties from cheaper data [10, 11], either by Δ-learning, in which the model predicts the correction from the low- to the high-fidelity value [12], or by training a single model on several fidelities. Yang, Manganaris and Mannodi-Kanakkithodi published a high-throughput dataset of 495 pure and mixed ABX_3 halide perovskites computed with PBE and, for subsets, with HSE06 and HSE06 + SOC [13]. They later trained multi-fidelity random forests on it for the decomposition energy, band gap and photovoltaic efficiency and used them for large-scale screening and inverse design [14]. The same group has also worked on halide perovskite alloy design and on multi-fidelity band gaps for photocatalysis [15, 16]; broader reviews of ML for perovskites and other materials are available [17, 18]. In our own earlier work we used composition descriptors to screen cubic and orthorhombic perovskites [19, 20].

Two questions about this kind of data are still open. The first is how much the reported accuracy depends on the validation split. A high-throughput set of mixed perovskites consists of families that differ only in the fraction of one species. A random split therefore places near-identical alloys in both the training and the test set, and the resulting accuracy can overstate what the model does for new chemistry [21, 22]. The second question is what the labels contain. The SLME is computed from an absorption spectrum, and for hybrid functionals this spectrum is often derived from a cheaper calculation rather than computed directly. In that case the label carries less independent information than its name suggests.

We use the public dataset of Ref. [13] as a controlled test bed and ask three questions. How much does a PBE calculation help ML predict the HSE06 + SOC SLME and band gap, compared with composition alone, with the PBE value used as-is and with a linear recalibration of it? Is Δ-learning better than direct learning, and do more flexible learners, including a transformer for tabular data, change the answer? And what limits do the labels set on any such model? To answer them we benchmark seven model families under nested cross-validation with random, chemical-system-grouped and B-site-grouped splits, compare 12 learning algorithms together with a re-implementation of the multi-fidelity random forest of Ref. [14], evaluate the models on the compounds that matter for screening (ranking among high-SLME compounds, held-out SHAP attributions, conformal intervals and learning curves) and audit the labels themselves. We do not propose a new screening tool. The HSE06 + SOC labels are computed values with their own approximations, and the predictions for 207 PBE-only alloys given at the end are hypotheses that need hybrid-functional calculations to confirm. Figure 1 summarises the workflow.

2 Data, targets and label audit

2.1 Data set

All data come from the open dataset of Yang et al. [13]. It contains 495 cubic or pseudo-cubic ABX_3 compounds with A = MA, FA, Cs, Rb or K, B = Ca, Sr, Ba, Ge, Sn or Pb and X = Cl, Br or I. Alloys are mixed at one site at a time and modelled with special quasirandom structures in 2 × 2 × 2 supercells. For each compound the dataset gives the PBE decomposition energy E_d (the energy relative to the AX and BX_2 phases, including a mixing-entropy term at 300 K, per formula unit), the band gap E_g, the SLME for a 5 µm absorber and the deviations of the relaxed PBE cell from cubic symmetry. Subsets were also computed with HSE06 after HSE relaxation (HSE-rel, 299 compounds), with HSE06 + SOC on the HSE-relaxed structures (HSE-rel-SOC, 282) and with HSE06 + SOC on the PBE-relaxed structures (HSE-PBE-SOC, 244). The workbook distributed with the dataset has 488 PBE rows. We corrected 26 formula strings with a misplaced underscore (e.g. CsBaI_2.25_Cl_0.75) and found no duplicate compositions. After matching on composition, 243 compounds have both PBE and HSE-PBE-SOC values and 281 have both PBE and HSE-rel-SOC values. The 207 PBE compounds without any HSE06 + SOC value were kept aside for prospective prediction, see Section 4.8.

2.2 Targets

The primary targets are the HSE06 + SOC SLME and band gap computed on the PBE geometry (HSE-PBE-SOC, n = 243); the values after HSE relaxation (n = 281) are secondary targets. The label audit in Sect. 2.3 explains this choice. The seven model families (Sect. 3.2) were compared on the SLME at both geometries and on the HSE-relaxed gap, and the learning algorithms (Sect. 4.3) on the SLME and gap at the PBE geometry. SLME is expressed as a percentage, and differences in SLME are presented in percentage points; please refer to Table 1 for a summary of the data.

Table 1
Data sets and targets. Chemical systems are sets of species irrespective of fractions; B-site sets are sets of B-site species

Set	n	Chemical systems	B-site sets	Pb-free	Role
PBE (all compounds)	488	376	56	335	inputs
HSE06 + SOC, PBE geometry (HSE-PBE-SOC)	243	222	53	153	primary targets: SLME, E_g
HSE06 + SOC, HSE-relaxed (HSE-rel-SOC)	281	251	55	172	secondary targets: SLME, E_g
PBE only (no HSE06 + SOC value)	207	148	16	163	prospective prediction

2.3 How the SLME labels were made, and what follows

The SLME of Yu and Zunger [4] is the maximum power conversion efficiency of a single-junction cell whose absorber has a thickness L and the computed absorption coefficient α(E). Under AM1.5G illumination the current–voltage relation and the efficiency are

$$J(V) = J_{sc} - J_0[\exp(eV/k_BT) - 1],\ \eta_{SLME} = \max_V[J(V)\,V] / P_{in} \quad (1)$$

$$J_{sc} = e \int a(E)\,\Phi_{AM1.5}(E)\,dE,\ a(E) = 1 - \exp[-2\alpha(E)L] \quad (2)$$

$$J_0 = (e\pi/f_r) \int a(E)\,\Phi_{bb}(E, T)\,dE,\ f_r = \exp(-\Delta/k_BT) \quad (3)$$

where e is the elementary charge, k_B the Boltzmann constant, T the cell temperature, P_{in} the incident power density of the AM1.5G spectrum, $\Phi_{AM1.5}$ its photon flux per unit energy, Φ_{bb} the photon flux per unit energy and solid angle of a black body at T, a(E) the absorptivity of a film with a reflecting back contact, f_r the fraction of radiative recombination and Δ the difference between the lowest direct-allowed gap and the fundamental gap. J_{sc} and J_0 are current densities. In the dataset L = 5 μm [13].

In the source data the hybrid-functional absorption spectra were not computed with hybrid functionals. Instead, the PBE spectrum was shifted rigidly by the difference between the hybrid-functional and PBE gaps [13]:

$\alpha_{HSE}(E) = \alpha_{PBE}(E - \delta)$, $\delta = E_g^{HSE} - E_g^{PBE}$ (4)

Through Eqs. (1)–(3), the HSE06 + SOC SLME therefore depends only on the HSE06 + SOC gap and on the shape of the PBE absorption edge. Three properties of the labels follow (Fig. 2).

The SLME is close to a function of the gap. The HSE06 + SOC SLME is almost a monotonic function of the HSE06 + SOC gap (Spearman $\rho = -0.99$), and a seven-nearest-neighbour regression on the gap alone reproduces it with a chemical-system-grouped cross-validated R^2 of 0.985 (MAE 0.45 points). Predicting the HSE06 + SOC SLME is therefore close to predicting the HSE06 + SOC gap and mapping it through the PBE absorption profile.

SLME saturates nearly 16%. No HSE06 + SOC SLME exceeds 16.1%, whereas the PBE SLME reaches 25% (Fig. 2c). Compounds with HSE06 + SOC gaps between 0.8 and 1.3 eV all lie between 13.2 and 16.1%. Yang et al. attribute this ceiling to the spectral shift of Eq. (4): for the compounds with the highest PBE SLME, whose PBE gaps are near 1 eV, the hybrid-functional gaps differ from the PBE gaps, and after the shift the SLME is always lower and peaks at 15–16% [13]. The labels therefore cannot rank compounds within the top group, and neither can a model trained on them.

HSE relaxation sometimes changes the structure strongly. For 20 of the 243 compounds with both HSE06 + SOC values (8%), relaxation with HSE06 raises the HSE06 + SOC gap by more than 1 eV relative to the same functional at the PBE geometry, while the median change is 0.05 eV (Fig. 2d). Of these 20 compounds, 17 contain MA or FA, 14 contain Sn and 12 are mixed halides (Table S5). For $MASnI_3$ the HSE-relaxed gap is 2.47 eV, whereas the gap at the PBE geometry is 1.14 eV; measured gaps of $MASnI_3$ are about 1.2–1.3 eV [23, 24]. The HSE-relaxed labels thus include large structural distortions for a minority of compounds, which is why we use the PBE-geometry labels as primary targets.

3 Methods

3.1 Descriptors

Each formula was parsed into site fractions x_i of the 5 A, 6 B and 3 X species (X-site fractions normalised to one). From these we computed 35 composition descriptors: the 14 site fractions; the fraction-weighted ionic radii $\bar{r}_A$, $\bar{r}_B$ and $\bar{r}_X$ and their variances; the Goldschmidt tolerance factor t and the octahedral factor μ [25,26]; the weighted Pauling electronegativities of the B and X sites and their difference; the weighted first ionisation energy of the B site, the weighted electron affinity of the X site and the weighted atomic numbers of both sites; the fraction of ns^2 cations (Ge, Sn, Pb) on the B site; the organic fraction on the A site; the configurational mixing entropy of each site; and the number of species. With $\bar{r}_s = \Sigma_{i \in s} x_i r_i$ for site s,

$t = (\bar{r}_A + \bar{r}_X) / [\sqrt{2}\,(\bar{r}_B + \bar{r}_X)]$, $\mu = \bar{r}_B / \bar{r}_X$, $S_s = -\Sigma_{i \in s} x_i \ln x_i$ (5)

where S_s is the ideal mixing entropy of site s in units of k_B per site. The radii are Shannon values [27] with 2.17 and 2.53 Å for MA and FA [28]; for Sn^{2+}, which has no six-coordinate Shannon radius, 1.10 Å was used. The PBE inputs were E_g, SLME and E_d and the five deviations from cubicity of the PBE-relaxed cell. No HSE quantity was used as an input.

3.2 Models

Seven model families were compared: (1) the training-set mean; (2) the PBE value used as-is; (3) a ridge-regression recalibration of the PBE gap, SLME and E_d; (4) ML on the eight PBE inputs only; (5) ML on the composition descriptors only; (6) direct ML on composition plus PBE inputs; and (7) Δ-learning [12]. With x the composition descriptors, p the PBE inputs and y_{PBE} the PBE value of the target, the direct and Δ-learning predictions are

$\hat{y}_{direct} = f(x, p), \hat{y}_{\Delta} = y_{PBE} + g(x, p)$ (6)

where f is trained on the HSE06 + SOC values and g on the differences $y_{HSE} - y_{PBE}$. In families (4)–(7) the learner was chosen in the inner loop among extremely randomised trees [29], random forest [30] and gradient boosting [31], implemented in scikit-learn [32]. Negative predictions were set to zero.

In a second comparison, the direct model was trained with 12 learning algorithms: ridge regression; kernel ridge regression and support-vector regression with RBF kernels; a Gaussian process with an automatic-relevance Matérn kernel (v = 5/2, one length scale per standardised input) [33]; a two-hidden-layer perceptron; random forest; extremely randomised trees; gradient boosting; XGBoost [34]; LightGBM [35]; CatBoost [36]; and an FT-Transformer [37], in which each feature is embedded as a token and two self-attention layers [38] mix the tokens (32-dimensional embeddings, four heads, AdamW, early stopping on an internal 15% split, three-member ensemble). We also re-implemented the multi-fidelity random forest of Ref. [14]: one random forest trained on all PBE rows and on the HSE06 + SOC rows of the training compounds, with composition descriptors and a fidelity indicator as inputs, queried at the high-fidelity level. The original model also used site-averaged elemental properties and a crystal-phase indicator and included experimental data as a third fidelity, so our version reproduces its design rather than the exact model. We tried to include the TabPFN tabular foundation model [39], but its weights require a separate licence agreement and were not used.

3.3 Validation

All results come from nested cross-validation [40]: five outer folds repeated three times with different seeds. Within each outer training set, hyper-parameters were chosen by a randomised search with 12 candidates and three inner folds, scored by the MAE; no tuning, scaling or feature selection used the outer test fold. Three outer split types were used. In random splits each compound is its own group. In chemical-system-grouped splits all compounds with the same set of species (e.g. Cs–Ba–Pb–I, regardless of fractions) fall in the same fold (222 groups for the primary target). In B-site-grouped splits all compounds with the same set of B-site species fall in the same fold (53 groups), which tests transfer

to B-site chemistry absent from training [21]. Inner folds used the same grouping as the outer folds. Models were compared on the same J = 15 outer folds with the corrected resampled t-test of Nadeau and Bengio [41],

$$t = \bar{d} / \sqrt{[(1/J + n_{te}/n_{tr})\, s_d^2]} \quad (7)$$

where $\bar{d}$ and s_d^2 are the mean and variance of the fold-wise MAE differences and n_{te} and n_{tr} the sizes of the test and training sets; t is compared with a t distribution with J − 1 degrees of freedom. The second term in the brackets corrects for the overlap of training sets between folds. The p values are descriptive and were not corrected for multiple comparisons.

3.4 Screening metrics, interpretation and uncertainty

Screening selects compounds at the top of a ranking. Besides the MAE, RMSE and R^2 of the pooled out-of-fold predictions, we therefore report how many of the true top-10 and top-20 compounds are recovered in the predicted top 10 and top 20, and the MAE and Spearman ρ among compounds whose HSE06 + SOC SLME is at least 10%. Feature attributions are TreeSHAP values [42, 43] of an extremely-randomised-trees model, computed on the outer test folds only (held-out SHAP) and averaged over the three repeats. Prediction intervals were obtained by split-conformal prediction [44]. Within each outer training set, 25% of the data (grouped where applicable) were held out as a calibration set of size n_c, and the interval for a test compound x was

$$C(x) = [\hat{y}(x) - \hat{q}, \hat{y}(x) + \hat{q}],\ \hat{q} = \lceil (n_c + 1)(1 - \alpha) \rceil\text{-th smallest of } |y_i - \hat{y}(x_i)| \quad (8)$$

with the absolute residuals taken over the calibration set and α = 0.1 or 0.2 for nominal coverages of 90% and 80%. Coverage is guaranteed only when calibration and test compounds are exchangeable. Y-randomisation [45] repeated the grouped cross-validation 30 times with permuted targets. Learning curves used random subsets of each outer training set.

3.5 Prospective prediction

The final direct model was tuned with grouped inner folds on all 243 labelled compounds and applied to the 207 PBE-only compounds. Each prediction carries a 90% interval whose half-width is the quantile of Eq. (8) computed from the chemical-system-grouped out-of-fold residuals, and an applicability-domain flag: a query lies inside the domain if its mean standardised distance to the five nearest training compounds is below the 95th percentile of the same quantity within the training set [46]. Following Ref. [13], compounds with negative PBE decomposition energy are labelled stable. A negative E_d is a necessary but not a sufficient condition for stability, because competing phases other than AX and BX_2, segregation and kinetics are not considered [13].

4 Results

4.1 PBE versus HSE06 + SOC

Figure 2 compares the PBE and HSE06 + SOC values on the PBE geometry. HSE06 + SOC raises the gap of 85% of the compounds, by 0.67 eV on average. The remaining 15% all contain Pb, for which the SOC-induced lowering of the conduction band outweighs the opening by the hybrid functional [9]. Over all compounds the mean change is + 0.53 eV, and the PBE gap used as an estimate of the HSE06 + SOC gap has an MAE of 0.60 eV. The PBE SLME is correlated with the HSE06 + SOC SLME (Spearman ρ = 0.92 on the PBE geometry and 0.87 after HSE relaxation), but the difference between them is not a small offset. For the 92 compounds with a PBE SLME above 10% the median change is − 5.2 points, and the difference is anti-correlated with the PBE SLME itself (ρ = −0.58 on the PBE geometry and − 0.71 after HSE relaxation).

4.2 Benchmark of model families

Table 2 and Fig. 3 give the results for the primary SLME target. Used as-is, the PBE SLME has an MAE of 2.45 points and R^2 = 0.60, and a linear recalibration reduces the MAE only to 1.6 points. Every ML family does much better. The direct model with composition and PBE inputs gives an MAE of 0.44 ± 0.02 points on random splits and 0.45 ± 0.05 on chemical-system-grouped splits (R^2 = 0.98), so grouping by chemical system costs little. Grouping by B-site set roughly doubles the error, to 0.86 ± 0.20 points (R^2 = 0.94). Transfer to B-site chemistry absent from training is the hardest case, and there the advantage of the PBE inputs disappears (0.86 against 0.86 points for composition alone). Figure 4 shows parity plots for the chemical-system-grouped splits.

Table 2

MAE (percentage points) and R^2 of the pooled out-of-fold predictions of the HSE06 + SOC SLME at the PBE geometry (n = 243), mean ± standard deviation over three repeats of nested five-fold cross-validation

Model	Random	Grouped by chemical system	Grouped by B-site set
Mean	5.90 ± 0.01 (-0.01)	5.90 ± 0.02 (-0.01)	6.16 ± 0.05 (-0.11)
PBE as-is	2.45 ± 0.00 (0.60)	2.45 ± 0.00 (0.60)	2.45 ± 0.00 (0.60)
Linear recal.	1.60 ± 0.01 (0.88)	1.61 ± 0.01 (0.88)	1.78 ± 0.05 (0.85)
ML (PBE only)	0.86 ± 0.03 (0.95)	0.89 ± 0.02 (0.94)	1.25 ± 0.17 (0.89)
ML (comp. only)	0.63 ± 0.04 (0.96)	0.68 ± 0.04 (0.96)	0.86 ± 0.01 (0.94)
ML direct (comp.+PBE)	0.44 ± 0.02 (0.98)	0.45 ± 0.05 (0.98)	0.86 ± 0.20 (0.94)
ML Δ-learning	0.53 ± 0.04 (0.98)	0.54 ± 0.05 (0.98)	0.95 ± 0.20 (0.94)

The secondary targets show the same pattern (Table S2). For the HSE-relaxed SLME the direct model reaches 0.66 points on chemical-system-grouped splits, against 3.65 points for the PBE value, which has

a negative R^2 (− 0.30) on this target. For the HSE-relaxed gap the direct model reaches 0.148 eV, against 0.189 eV for composition alone. For the gap at the PBE geometry, which was evaluated in the algorithm comparison, the direct model reaches 0.096 eV with extremely randomised trees and 0.059 eV with the Gaussian process (Sect. 4.3), against 0.60 eV for the PBE gap used as-is.

4.3 Learning algorithms, transformer and multi-fidelity random forest

Table 3 and Fig. 5 compare the 12 learning algorithms for the direct model on the primary targets. The tree ensembles (extremely randomised trees, random forest, gradient boosting, XGBoost, LightGBM and CatBoost) form a tight group, with SLME errors of 0.45–0.53 points on chemical-system-grouped splits and no significant differences from extremely randomised trees. The Gaussian process with automatic relevance determination is the most accurate algorithm: 0.36 points for the SLME (better than extremely randomised trees in 13 of 15 folds, corrected p = 0.04) and 0.059 eV for the gap, against 0.096 eV for extremely randomised trees (14 of 15 folds, p = 0.005). Its advantage is largest for the gap of B-site chemistries absent from training. There it reaches 0.080 eV, against 0.17–0.23 eV for the tree ensembles, and it is better than extremely randomised trees in all 15 folds (p = 0.002). For the SLME on B-site-grouped splits its MAE is also the lowest (0.63 against 0.73 points), but the difference is not significant and its ranking of the top compounds is weaker (4 of the true top 10; Table S10), so no single algorithm dominates there.

A plausible reason for the advantage of the Gaussian process is the smoothness of its kernel. Fitted to all labelled compounds, its shortest length scales belong to a compact set of inputs: the PBE SLME, the Pb fraction, the PBE gap, the ns^2 fraction, the X-site electron affinity, the b-axis deviation from cubicity and the Ge fraction. This set agrees with the SHAP analysis below. A smooth function of a few physically meaningful inputs can interpolate between B-site chemistries, whereas axis-aligned tree splits produce piecewise-constant predictions. Consistent with this, even linear ridge regression predicts the gap of unseen B-site chemistries (0.151 eV) as well as or better than the tree ensembles (Table S9).

The FT-Transformer reaches 0.61 points on chemical-system-grouped splits, worse than the tree ensembles (p = 0.02), and 0.78 points on B-site-grouped splits, within their range. With fewer than 200 training compounds per fold the attention model has no advantage over simpler learners, consistent with the finding that no deep architecture is uniformly better than gradient-boosted trees on tabular data [37]. Kernel ridge regression, support-vector regression and the multilayer perceptron are less accurate than the tree ensembles, and linear ridge regression is the least accurate for the SLME (1.0 point), so the mapping to the SLME is strongly non-linear.

Our re-implementation of the multi-fidelity random forest of Ref. [14] gives 0.87 points for the SLME and 0.175 eV for the gap on chemical-system-grouped splits, and 1.18 points and 0.265 eV on B-site-grouped splits. For the SLME on chemical-system-grouped splits, this is 1.6–1.9 times the error of the tree ensembles and 2.5 times that of the Gaussian process. In this design the PBE value of a query

compound enters only as a training row with a fidelity flag, so the model has to infer the high-fidelity value from compositions similar to the query. The direct model receives the PBE gap and SLME of the query as inputs. When PBE outputs are available for the compound of interest, using them as features exploits them more effectively.

Table 3
MAE of the direct model (composition + PBE inputs) with different learning algorithms on chemical-system-grouped and B-site-grouped nested cross-validation. p compares each algorithm with extremely randomised trees using the corrected resampled t-test of Eq. (7) on the 15 outer folds

Algorithm	SLME (chem. system)	SLME (B-site set)	E_g (chem. system)	E_g (B-site set)	p (SLME / chem. system)
Gaussian process	0.35	0.63	0.059	0.080	0.04
Gradient boosting	0.45	0.76	0.099	0.169	0.77
Extra trees	0.46	0.73	0.096	0.178	–
LightGBM	0.48	0.74	0.100	0.173	0.74
XGBoost	0.49	0.79	0.103	0.182	0.45
CatBoost	0.51	0.81	0.111	0.170	0.21
Random forest	0.53	0.89	0.126	0.225	0.16
FT-Transformer	0.61	0.78	0.166	0.239	0.02
Kernel ridge (RBF)	0.68	0.92	0.130	0.240	< 0.01
SVR (RBF)	0.71	0.89	0.113	0.184	< 0.01
Multi-fidelity RF [Yang 2024]	0.87	1.18	0.175	0.265	< 0.01
MLP (2 hidden layers)	0.88	1.04	0.298	0.396	< 0.01
Ridge (linear)	1.00	1.15	0.113	0.151	< 0.01

4.4 Δ-learning versus direct learning

Δ-learning was not better than direct learning on any SLME target or split. On the primary target its MAE was 0.54 against 0.45 points on chemical-system-grouped splits, with direct learning better in 12 of 15 outer folds (corrected p = 0.11). On the HSE-relaxed SLME the difference was larger (0.83 against 0.66 points). For the HSE-relaxed gap the two were tied (0.150 against 0.148 eV). The learning curves (Fig. 7b) show the same ordering at every training-set size, so the result is not a small-data effect. The reason is the anti-correlation noted in Sect. 4.1. The difference between the HSE06 + SOC and PBE SLME is large and depends strongly on where the PBE gap lies relative to the SLME optimum, so the function g in Eq. (6) is not easier to learn than f. Δ-learning helps when the low-fidelity value is a good first

approximation and the correction is small and smooth [12]; for the SLME of these compounds neither condition holds.

4.5 Screening performance

The difference between the PBE value and the ML models is largest where screening needs accuracy (Table 4). Among the 83 compounds with an HSE06 + SOC SLME of at least 10% at the PBE geometry, the PBE SLME has an MAE of 4.6 points and a Spearman ρ of only 0.08 with the HSE06 + SOC value, whereas the direct model has an MAE of 0.56 points and ρ = 0.92. Ranking by the PBE SLME places only one of the ten compounds with the highest HSE06 + SOC SLME in its top ten, and the linear recalibration does no better; the direct model recovers nine and Δ-learning six. Over all compounds the PBE ranking looks respectable (ρ = 0.92) only because it separates wide-gap compounds with negligible SLME from the rest. This global correlation says little about the decision a screening study has to make.

Table 4
Screening metrics for the HSE06 + SOC SLME at the PBE geometry on chemical-system-grouped cross-validation: true top-k compounds recovered in the predicted top-k, and MAE and Spearman ρ among the 83 compounds with HSE06 + SOC SLME ≥ 10%

Model	Top-10 hits	Top-20 hits	MAE (SLME ≥ 10%)	ρ (SLME ≥ 10%)	ρ (all)
PBE as-is	1	8	4.56	0.08	0.92
Linear recal.	1	7	2.33	0.24	0.93
ML (PBE only)	2	9	1.31	0.62	0.95
ML (comp. only)	8	16	0.73	0.92	0.99
ML direct (comp.+PBE)	9	18	0.56	0.92	0.99
ML Δ-learning	6	13	0.70	0.86	0.97

4.6 What the model uses: held-out SHAP

Figure 6 shows held-out SHAP values of the direct model for the primary target. The six largest contributions come from the PBE SLME (mean |SHAP| 1.39 points), the ns^2 fraction on the B site (1.21), the PBE gap (1.14), the X–B electronegativity difference (0.78), the B-site electronegativity (0.67) and the B-site ionisation energy (0.42). The PBE inputs account for 38% of the total attribution and the composition descriptors for 62%. The dependence plots are physically sensible. The SHAP value of the PBE gap is large and positive below about 1.5 eV, crosses zero between 2.0 and 2.5 eV and levels off at about − 1 point above 3 eV, where the SLME is negligible at either level of theory (Fig. 6c). The ns^2 fraction raises the predicted SLME monotonically (Fig. 6b). In the ns^2 perovskites the valence-band maximum has B-s/X-p antibonding character and the conduction-band minimum B-p character, which gives strong absorption at the band edge [3, 9]; the alkaline-earth cations lack this configuration and give much wider gaps. A small X–B electronegativity difference, that is, a more covalent B–X bond, raises the

prediction (Fig. 6d), consistent with the narrower gaps of iodides and of ns^2 cations. The contribution of the PBE SLME rises steeply up to about 10% and more slowly above it (Fig. 6e), the same compression of the high-SLME range that the 16% ceiling of the labels imposes. The structural PBE descriptors (deviations from cubicity) contribute little.

4.7 Uncertainty and robustness

Split-conformal 90% intervals of the direct model reached their nominal coverage on random splits (0.92 ± 0.05) and on chemical-system-grouped splits (0.92 ± 0.05), with mean widths of 3.7 and 4.2 points (Fig. 7a). On B-site-grouped splits they under-covered (0.78 ± 0.18). The calibration residuals come from B-site chemistries seen in training and underestimate the errors on unseen ones, so the exchangeability assumed by Eq. (8) does not hold. The 80% intervals behaved in the same way (0.84, 0.82 and 0.64). Learning curves on chemical-system-grouped splits (Fig. 7b) place the direct model below Δ-learning and the composition-only model at every training-set size, with the largest margin for small data (MAE 0.98 against 1.45 and 1.24 points with 39 training compounds). All ML curves are still decreasing at 194 compounds, so more hybrid-functional data would help, whereas the linear recalibration does not improve beyond 1.6 points. In Y-randomisation the grouped-CV R^2 fell from 0.97 for the real targets to −0.23 ± 0.08 (maximum − 0.08) for 30 permuted targets (Fig. 7c). The error is larger for alloys than for compounds with a single species on every site (0.49 against 0.29 points), as expected from the larger configurational space of alloys. A k-nearest-neighbour regressor on the site fractions alone has an MAE of 2.1–2.7 points on the same splits (Table S6), so the models learn more than proximity in composition.

4.8 Prospective predictions for PBE-only alloys

Of the 207 compounds without HSE06 + SOC values, 154 lie inside the applicability domain and 64 have a negative PBE decomposition energy. The query set is chemically close to the training set: 99.5% of the queries have a B-site set that occurs among the 243 training compounds, but only 16% share an exact chemical system with them. The predictions therefore interpolate between known chemistries rather than extrapolate to new ones, and the chemical-system-grouped intervals are the appropriate ones. Table 5 lists the Pb-free compounds that are inside the domain and have negative E_d, ranked by the predicted HSE06 + SOC SLME (Fig. 8). The six leading predictions lie between 14.7 and 15.3%, close to the 16% ceiling of the training labels, and their 90% intervals (± 1.3 points) overlap, so the list is a shortlist for hybrid-functional calculations rather than a ranking. The Gaussian process of Sect. 4.3 gives almost the same ranking (Spearman ρ = 0.99 with the extremely-randomised-trees predictions inside the domain) and predicts these six compounds at 14.7–16.0%. Several entries are mixed iodide–bromide Sn compounds. Four of the 20 compounds whose gap changes by more than 1 eV on HSE relaxation belong to this class, including the largest change ($RbSnI_{1.5}Br_{1.5}$, 2.7 eV; Sect. 2.3), which is a further reason to verify them explicitly.

Table 5
Pb-free PBE-only alloys inside the applicability domain with negative PBE decomposition energy, ordered by the predicted HSE06 + SOC SLME at the PBE geometry (direct model), with 90% intervals (Sect. 3.5)

Composition	PBE E_d (eV f.u.$^{-1}$)	PBE E_g (eV)	$SLME_{PBE}$ (%)	Predicted SLME (%)	90% low	90% high
$CsSnI_{2.625}Br_{0.375}$	−0.14	0.47	8.6	15.3	14.0	16.6
$CsSnI_{2.25}Br_{0.75}$	−0.14	0.46	8.4	15.1	13.8	16.4
$K_{0.375}Rb_{0.375}Cs_{0.25}GeBr_3$	−0.10	0.66	13.2	15.1	13.8	16.4
$CsSnI_{1.875}Br_{1.125}$	−0.16	0.55	10.9	15.0	13.7	16.3
$CsGe_{0.5}Sn_{0.5}Br_3$	−0.28	0.67	13.6	14.7	13.4	16.0
$K_{0.125}Rb_{0.875}SnBr_3$	−0.04	0.57	10.2	14.7	13.4	16.0
$RbSnBr_{1.875}Cl_{1.125}$	−0.05	0.62	11.2	13.3	12.0	14.6
$RbGeBr_{1.5}Cl_{1.5}$	−0.18	0.77	14.7	12.9	11.6	14.2

5 Discussion

5.1 What the PBE calculation adds

The answer depends on the target and the split. Used directly, the PBE SLME is a poor proxy for the HSE06 + SOC SLME, particularly where screening matters: it barely ranks the compounds with high HSE06 + SOC SLME. As inputs to an ML model the PBE outputs help. For the PBE-geometry SLME they lower the error by about a third relative to composition alone on random and chemical-system-grouped splits, and for the HSE-relaxed gap by about a fifth. They do not help the tree models for B-site chemistries absent from training, where the composition model is as good. The SHAP attributions are consistent with this: the PBE SLME and gap carry information about the band edges of a given chemistry, which a tree model can exploit only when similar chemistries were seen in training. The Gaussian process, which uses the PBE gap through a smooth kernel, is the exception for the band gap and the most transferable model we found.

5.2 What the labels allow

Because of Eq. (4), the HSE06 + SOC SLME in this dataset is a near-deterministic function of the HSE06 + SOC gap. A model that predicts this SLME accurately is, in effect, an accurate model of the HSE06 + SOC gap combined with the PBE absorption profile. This does not make the SLME results wrong, but SLME accuracy should not be read as evidence that a model has learned optical physics beyond the gap. Hybrid-functional absorption spectra, which change the oscillator strengths as well as the onset, would make the SLME a more independent target. The 16% ceiling has a more practical consequence. It

compresses the top of the distribution, so labels and predictions both lose resolution where a screening decision is made; users of the dataset should treat SLME ranks within the top group as ties, as our shortlist does. The HSE-relaxation effect is a separate issue. For a minority of compounds the HSE-relaxed geometry departs strongly from the PBE geometry and gives gaps far above experiment, so results on HSE-relaxed targets should be reported separately and with caution.

5.3 Comparison with earlier work on this dataset

Table 6 places our results next to the earlier multi-fidelity study of the same dataset [14]. That study trained multi-fidelity random forests on PBE, HSE06 + SOC and experimental data and reported an RMSE of 1.29% for the HSE-level photovoltaic efficiency from repeated random 80/20 splits. The numbers cannot be compared directly, because the data subsets, input features, fidelities and splits differ. Our re-implementation of their design gives a higher RMSE under chemical-system-grouped splits (1.63%). This difference should not be attributed to split optimism: in our benchmark random and chemical-system-grouped splits gave almost the same error for the direct model (0.44 and 0.45 points), and the remaining differences in features and data are a more likely cause. What can be compared is the effect of the model design on the same folds: models that take the PBE outputs of the query compound as inputs reduce the RMSE to 0.87% (extremely randomised trees) and 0.54% (Gaussian process). For multi-fidelity screening of these materials the practical message is that, once a PBE calculation of a candidate is available, its outputs are most useful as descriptors rather than as a separate fidelity level.

Table 6
Earlier ML results on this dataset compared with the present study. The studies differ in data subsets, features, fidelities and validation, so the numbers are not a performance ranking

Study	Target(s)	Model	Validation	Reported error for hybrid-functional SLME
Yang et al. 2024 [14]	E_d, E_g, PV efficiency (PBE, HSE06 + SOC, experiment)	Multi-fidelity random forest	Repeated random 80/20 splits	RMSE 1.29% (HSE level)
Present study	HSE06 + SOC SLME and E_g (PBE and HSE geometry)	7 model families, 12 algorithms	Nested CV; random and grouped splits	RMSE 0.54% (Gaussian process), 0.87% (extremely randomised trees) and 1.63% (multi-fidelity random forest re-implementation), chemical-system-grouped

5.4 Limitations

All targets are computed values from one dataset with its own approximations: an SLME from a shifted PBE spectrum, special quasirandom structures for alloys and pseudo-cubic cells. The ceiling of the SLME labels and the distortions introduced by HSE relaxation (Sect. 2.3) limit the conclusions. With 243 labelled compounds and six B-site elements, the B-site-grouped results rest on few groups per fold and carry a large uncertainty (standard deviation 0.2 points across repeats). The corrected t-test is conservative and the p values were not adjusted for the number of comparisons, so we describe

differences by their size and by the number of folds in which they hold. The direct model needs a PBE calculation of the query compound, including its absorption spectrum; it saves the hybrid-functional calculation, not the PBE one. The multi-fidelity random forest was re-implemented with our descriptors and without experimental data. Finally, the predictions for PBE-only compounds were not validated with hybrid-functional calculations in this work.

6 Conclusion

On a public dataset of 243 pure and mixed halide perovskites, a PBE calculation helps ML predict the HSE06 + SOC SLME and band gap, but not in the way it is often used. Taken at face value, the PBE SLME is a poor screening metric, with a 2.45-point MAE and almost no ranking power among the compounds with high HSE06 + SOC SLME. Used as input features together with composition, the PBE outputs give a model with a 0.45-point MAE and good ranking on chemical-system-grouped splits, about a third better than composition alone. Δ-learning, often assumed to be the natural way to use low-fidelity data, was not better than direct learning. Of 12 learning algorithms, a Gaussian process with automatic relevance determination was the most accurate (0.36 points for the SLME and 0.059 eV for the gap), particularly for the gap of B-site chemistries absent from training (0.080 eV against 0.17–0.23 eV for tree ensembles). A tabular transformer did not beat the tree ensembles on 243 compounds, and the multi-fidelity random forest design of earlier work, which receives the PBE information only through a shared training set, was less accurate than any tree ensemble that takes the PBE outputs as inputs. For the SLME of B-site chemistries absent from training, the error of the tree models doubled and conformal intervals under-covered. The label audit shows that the hybrid-functional SLME in this dataset is close to a function of the gap because it was derived from a shifted PBE spectrum, that it saturates near 16%, and that HSE relaxation raises the gap by more than 1 eV for 8% of compounds. Datasets with hybrid-functional optics and with the PBE and hybrid calculations done on the same geometry would allow a stronger test.

Declarations

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

Not applicable.

Declaration of generative AI use During the preparation of this work the authors used Claude (Anthropic) to support code development and text preparation. The authors checked all analyses, results and conclusions, reviewed and edited the text, and take full responsibility for the content of the article.

Funding

This work was supported by the National Natural Science Foundation of China (grant 51850410506) and the Central University Basic Scientific Research Business Expenses Special Funds (NG2020002).

Author Contribution

A.A. Haidry: conceptualization, methodology, software, data curation, validation, visualization, project administration, supervision, resources, writing – original draft. C.R. Mandebvu: data curation, formal analysis. Q. Fatima: formal analysis, funding acquisition, supervision, writing – original draft, writing – review and editing. All authors read and approved the final manuscript.

Data availability

The processed data, all result tables and the Python code needed to reproduce the analysis and figures are provided as Supplementary Information and are available from the corresponding author on reasonable request.

References

1. Kojima, A., Teshima, K., Shirai, Y., Miyasaka, T.: Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. J. Am. Chem. Soc. **131**, 6050–6051 (2009). https://doi.org/10.1021/ja809598r
2. Jena, A.K., Kulkarni, A., Miyasaka, T.: Halide perovskite photovoltaics: background, status, and future prospects. Chem. Rev. **119**, 3036–3103 (2019). https://doi.org/10.1021/acs.chemrev.8b00539
3. Giustino, F., Snaith, H.J.: Toward lead-free perovskite solar cells. ACS Energy Lett. **1**, 1233–1240 (2016). https://doi.org/10.1021/acsenergylett.6b00499
4. Yu, L., Zunger, A.: Identification of potential photovoltaic absorbers based on first-principles spectroscopic screening of materials. Phys. Rev. Lett. **108**, 068701 (2012). https://doi.org/10.1103/PhysRevLett.108.068701
5. Shockley, W., Queisser, H.J.: Detailed balance limit of efficiency of p-n junction solar cells. J. Appl. Phys. **32**, 510–519 (1961). https://doi.org/10.1063/1.1736034
6. Perdew, J.P., Burke, K., Ernzerhof, M.: Generalized gradient approximation made simple. Phys. Rev. Lett. **77**, 3865–3868 (1996). https://doi.org/10.1103/PhysRevLett.77.3865
7. Heyd, J., Scuseria, G.E., Ernzerhof, M.: Hybrid functionals based on a screened Coulomb potential. J. Chem. Phys. **118**, 8207–8215 (2003). https://doi.org/10.1063/1.1564060
8. Krukau, A.V., Vydrov, O.A., Izmaylov, A.F., Scuseria, G.E.: Influence of the exchange screening parameter on the performance of screened hybrid functionals. J. Chem. Phys. **125**, 224106 (2006).

https://doi.org/10.1063/1.2404663

9. Even, J., Pedesseau, L., Jancu, J.-M., Katan, C.: Importance of spin–orbit coupling in hybrid organic/inorganic perovskites for photovoltaic applications. J. Phys. Chem. Lett. **4**, 2999–3005 (2013). https://doi.org/10.1021/jz401532q
10. Pilania, G., Gubernatis, J.E., Lookman, T.: Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput. Mater. Sci. **129**, 156–163 (2017). https://doi.org/10.1016/j.commatsci.2016.12.004
11. 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). https://doi.org/10.1038/s43588-020-00002-x
12. Ramakrishnan, R., Dral, P.O., Rupp, M., von Lilienfeld, O.A.: Big data meets quantum chemistry approximations: the Δ-machine learning approach. J. Chem. Theory Comput. **11**, 2087–2096 (2015). https://doi.org/10.1021/acs.jctc.5b00099
13. Yang, J., Manganaris, P., Mannodi-Kanakkithodi, A.: A high-throughput computational dataset of halide perovskite alloys. Digit. Discov. **2**, 856–870 (2023). https://doi.org/10.1039/d3dd00015j
14. Yang, J., Manganaris, P., Mannodi-Kanakkithodi, A.: Discovering novel halide perovskite alloys using multi-fidelity machine learning and genetic algorithm. J. Chem. Phys. **160**, 064114 (2024). https://doi.org/10.1063/5.0182543
15. Mannodi-Kanakkithodi, A., Chan, M.K.Y.: Data-driven design of novel halide perovskite alloys. Energy Environ. Sci. **15**, 1930–1949 (2022). https://doi.org/10.1039/d1ee02971a
16. Biswas, M., Desai, R., Mannodi-Kanakkithodi, A.: Screening of novel halide perovskites for photocatalytic water splitting using multi-fidelity machine learning. Phys. Chem. Chem. Phys. **26**, 23177–23188 (2024). https://doi.org/10.1039/d4cp02330g
17. Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A.: Machine learning for molecular and materials science. Nature. **559**, 547–555 (2018). https://doi.org/10.1038/s41586-018-0337-2
18. 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). https://doi.org/10.1038/s41524-019-0221-0
19. Fatima, Q., Zhang, H., Ahmad, U., Haidry, A.A.: A machine learning framework for screening and band gap prediction of stable cubic halide perovskites for energy applications. J. Power Sources. **679**, 240165 (2026). https://doi.org/10.1016/j.jpowsour.2026.240165
20. Fatima, Q., Haidry, A.A., Ahmad, U., Zhang, H.: The critical role of tilting descriptors in data-driven property prediction of orthorhombic perovskites. Adv. Theory Simul. **9**, e70513 (2026). https://doi.org/10.1002/adts.70513
21. Meredig, B., Antono, E., Church, C., Hutchinson, M., Ling, J., Paradiso, S., et al.: Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. Mol. Syst. Des. Eng. **3**, 819–825 (2018). https://doi.org/10.1039/C8ME00012C

22. Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. Patterns. **4**, 100804 (2023). https://doi.org/10.1016/j.patter.2023.100804
23. Stoumpos, C.C., Malliakas, C.D., Kanatzidis, M.G.: Semiconducting tin and lead iodide perovskites with organic cations: phase transitions, high mobilities, and near-infrared photoluminescent properties. Inorg. Chem. **52**, 9019–9038 (2013). https://doi.org/10.1021/ic401215x
24. Hao, F., Stoumpos, C.C., Cao, D.H., Chang, R.P.H., Kanatzidis, M.G.: Lead-free solid-state organic–inorganic halide perovskite solar cells. Nat. Photonics. **8**, 489–494 (2014). https://doi.org/10.1038/nphoton.2014.82
25. Goldschmidt, V.M.: Die Gesetze der Krystallochemie. Naturwissenschaften. **14**, 477–485 (1926). https://doi.org/10.1007/BF01507527
26. Bartel, C.J., Sutton, C., Goldsmith, B.R., Ouyang, R., Musgrave, C.B., Ghiringhelli, L.M., et al.: New tolerance factor to predict the stability of perovskite oxides and halides. Sci. Adv. 5. eaav0693 (2019). https://doi.org/10.1126/sciadv.aav0693
27. Shannon, R.D.: Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. Acta Crystallogr. A. **32**, 751–767 (1976). https://doi.org/10.1107/S0567739476001551
28. Kieslich, G., Sun, S., Cheetham, A.K.: Solid-state principles applied to organic–inorganic perovskites: new tricks for an old dog. Chem. Sci. **5**, 4712–4715 (2014). https://doi.org/10.1039/C4SC02211D
29. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1
30. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
31. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451
32. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
33. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA (2006)
34. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785
35. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al.: LightGBM: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. **30**, 3146–3154 (2017)
36. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. **31**, 6638–6648 (2018)
37. Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. Adv. Neural Inf. Process. Syst. **34**, 18932–18943 (2021)

38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 5998–6008 (2017)
39. Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., et al.: Accurate predictions on small data with a tabular foundation model. Nature. **637**, 319–326 (2025). https://doi.org/10.1038/s41586-024-08328-6
40. Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC Bioinform. **7**, 91 (2006). https://doi.org/10.1186/1471-2105-7-91
41. Nadeau, C., Bengio, Y.: Inference for the generalization error. Mach. Learn. **52**, 239–281 (2003). https://doi.org/10.1023/A:1024068626366
42. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. **30**, 4765–4774 (2017)
43. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., et al.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. **2**, 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9
44. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. J. Am. Stat. Assoc. **113**, 1094–1111 (2018). https://doi.org/10.1080/01621459.2017.1307116
45. Rücker, C., Rücker, G., Meringer, M.: y-Randomization and its variants in QSPR/QSAR. J. Chem. Inf. Model. **47**, 2345–2357 (2007). https://doi.org/10.1021/ci700157b
46. Tropsha, A.: Best practices for QSAR model development, validation, and exploitation. Mol. Inf. **29**, 476–488 (2010). https://doi.org/10.1002/minf.201000061

Figures

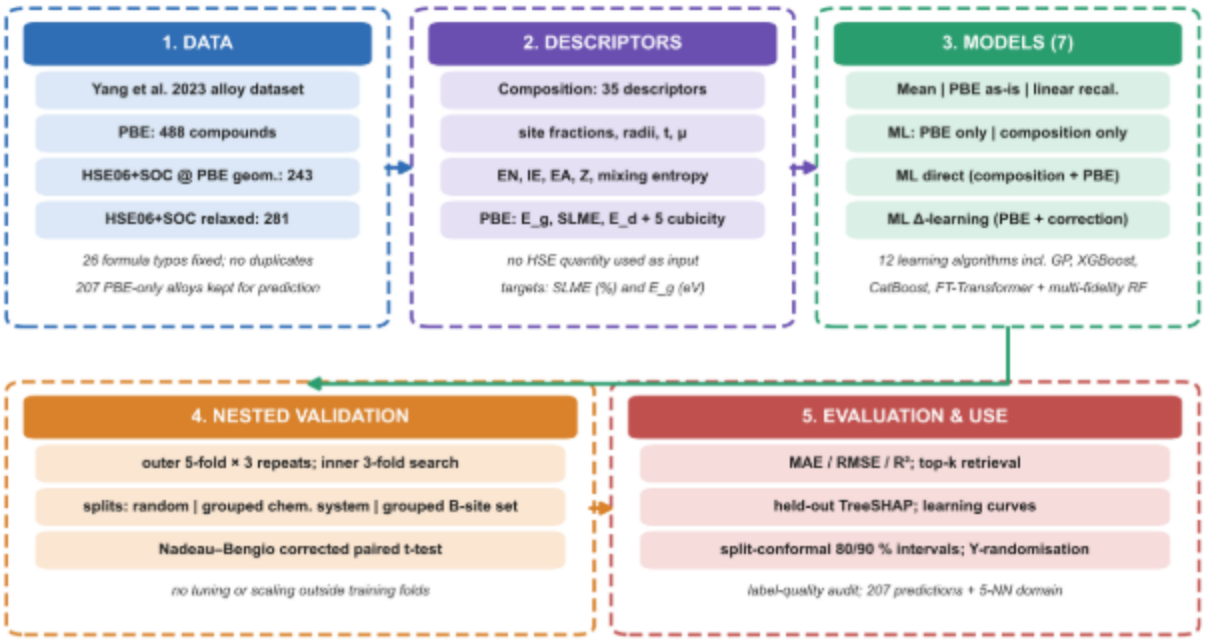


Figure 1

Workflow of the study: data and label audit, descriptors, model families and learning algorithms, nested validation with three split types, and evaluation.

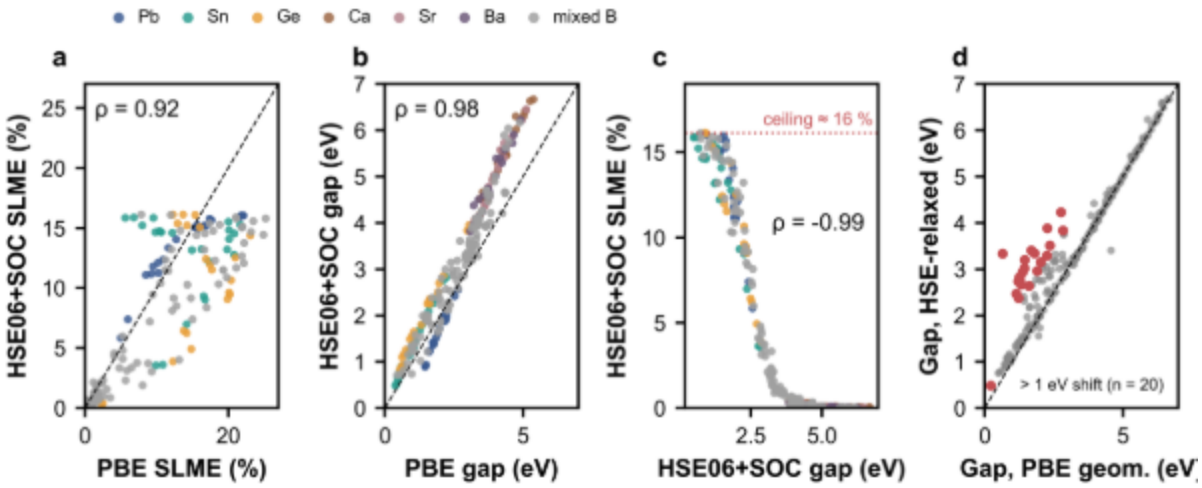


Figure 2

Data and label audit for the 243 compounds with PBE and HSE06+SOC (PBE geometry) values. (a) HSE06+SOC versus PBE SLME and (b) band gap, coloured by B-site species (grey: mixed B site). (c) HSE06+SOC SLME versus HSE06+SOC gap: the SLME is close to a monotonic function of the gap and saturates near 16 % (dotted line). (d) HSE06+SOC gap after HSE relaxation versus the gap at the PBE geometry; red points change by more than 1 eV.

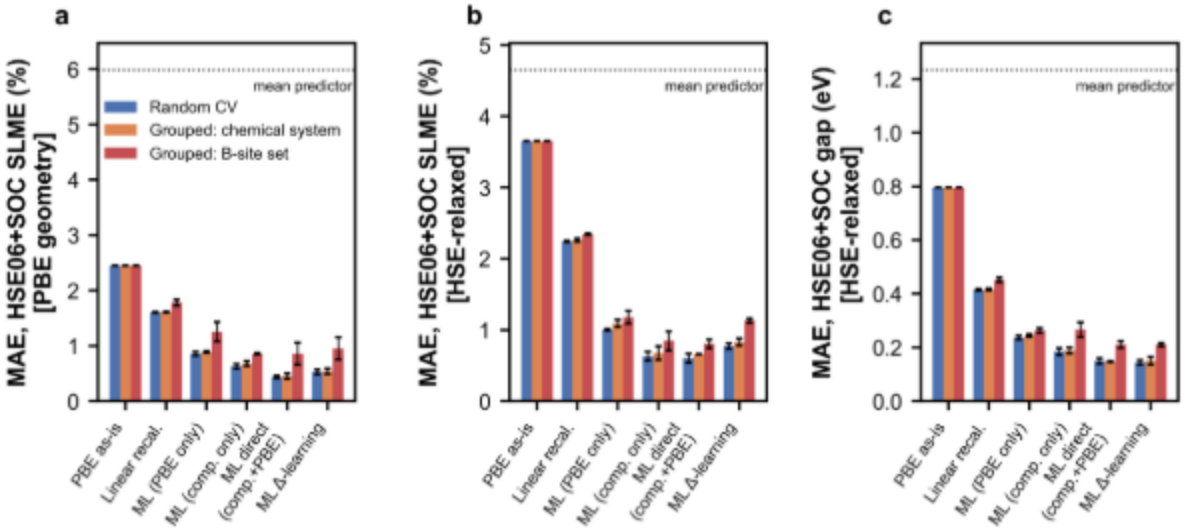


Figure 3

MAE of the model families for (a) the HSE06+SOC SLME at the PBE geometry, (b) the HSE06+SOC SLME after HSE relaxation and (c) the HSE06+SOC gap after HSE relaxation, for random, chemical-system-

grouped and B-site-grouped nested cross-validation. Bars are means and error bars standard deviations over three repeats; the dotted line is the mean predictor.

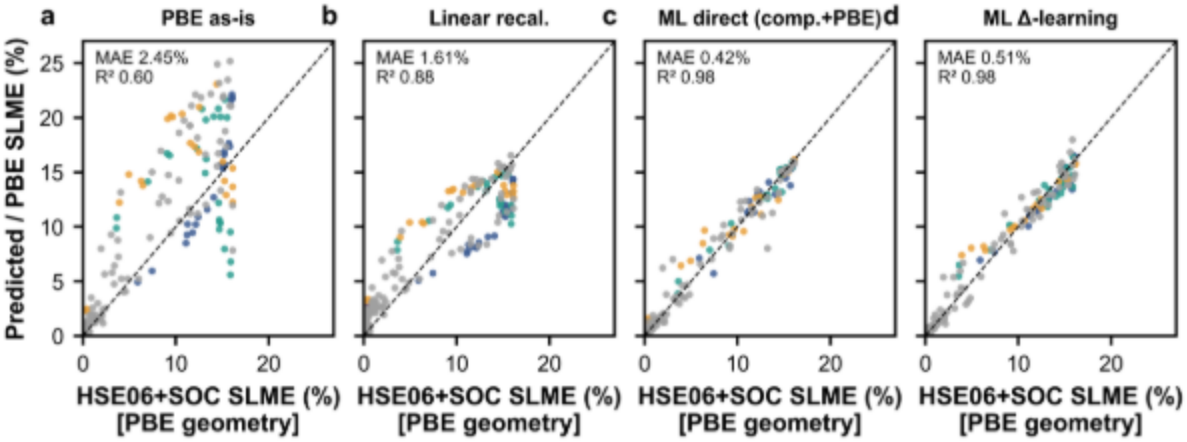


Figure 4

Out-of-fold predictions versus HSE06+SOC SLME at the PBE geometry on chemical-system-grouped splits, averaged over three repeats: (a) PBE SLME used as-is, (b) linear recalibration, (c) direct model with composition and PBE inputs and (d) Δ-learning. Colours as in Fig. 2.

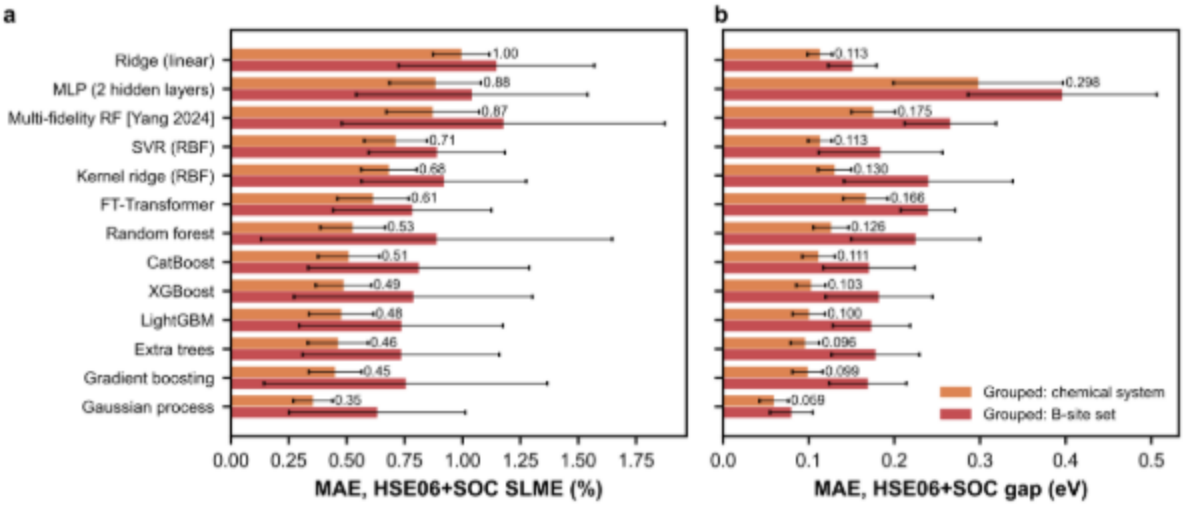


Figure 5

MAE of the direct model (composition + PBE inputs) with different learning algorithms for (a) the HSE06+SOC SLME and (b) the HSE06+SOC gap at the PBE geometry, on chemical-system-grouped (orange) and B-site-grouped (red) nested cross-validation. Error bars are standard deviations over the 15 outer folds; numbers give the chemical-system-grouped MAE.

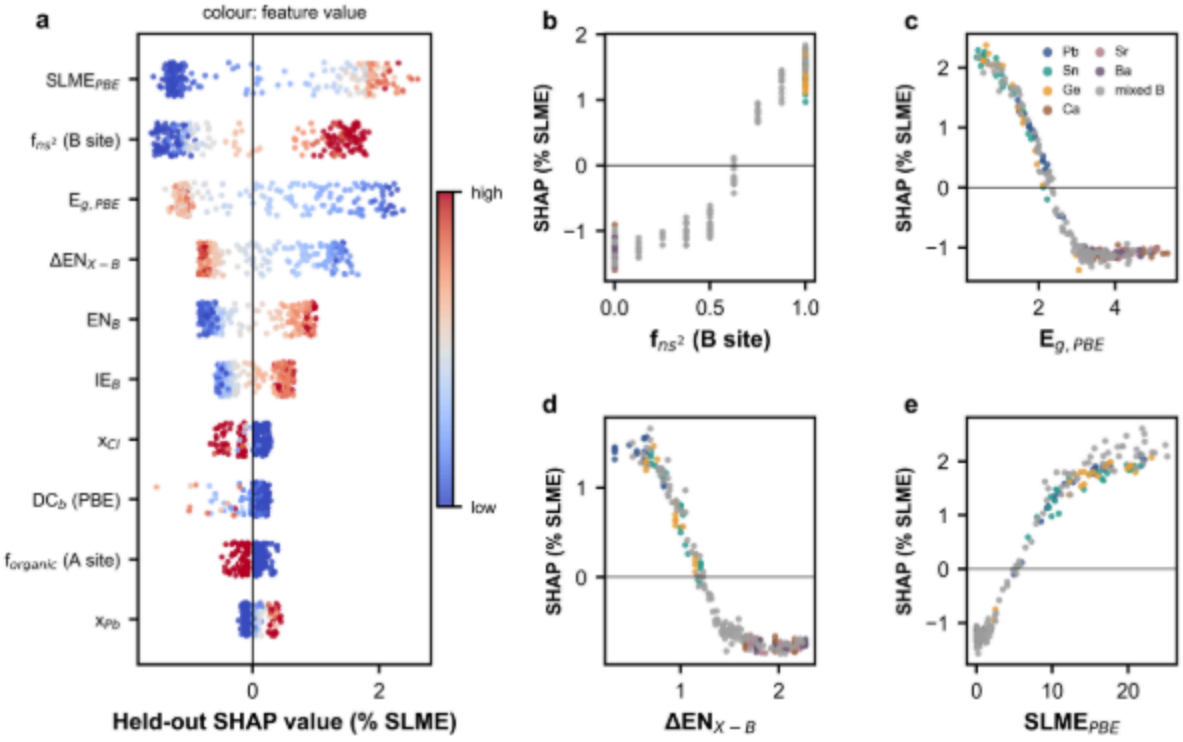


Figure 6

Held-out SHAP analysis of the direct model for the HSE06+SOC SLME at the PBE geometry (chemical-system-grouped folds, averaged over three repeats). (a) SHAP values of the ten most important features; colour gives the feature value. (b–e) Dependence of the SHAP value on the ns^2 fraction of the B site, the PBE gap, the X–B electronegativity difference and the PBE SLME, coloured by B-site species.

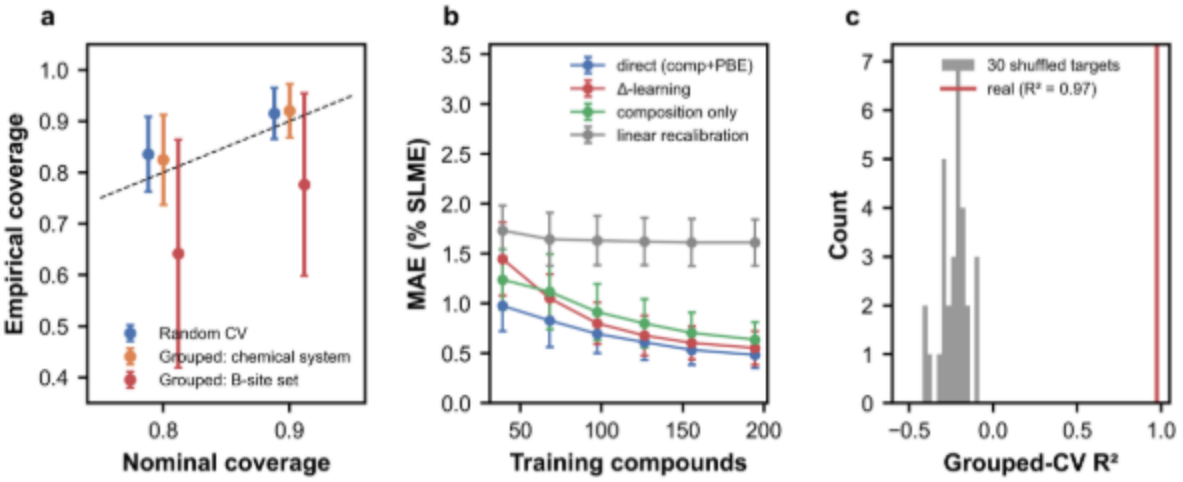


Figure 7

Uncertainty and robustness of the direct model for the HSE06+SOC SLME at the PBE geometry. (a) Empirical coverage of split-conformal 80 % and 90 % intervals for the three split types (mean ± standard

deviation over 15 outer folds). (b) Learning curves on chemical-system-grouped splits. (c) Grouped-CV R^2 for the real target and for 30 permuted targets.

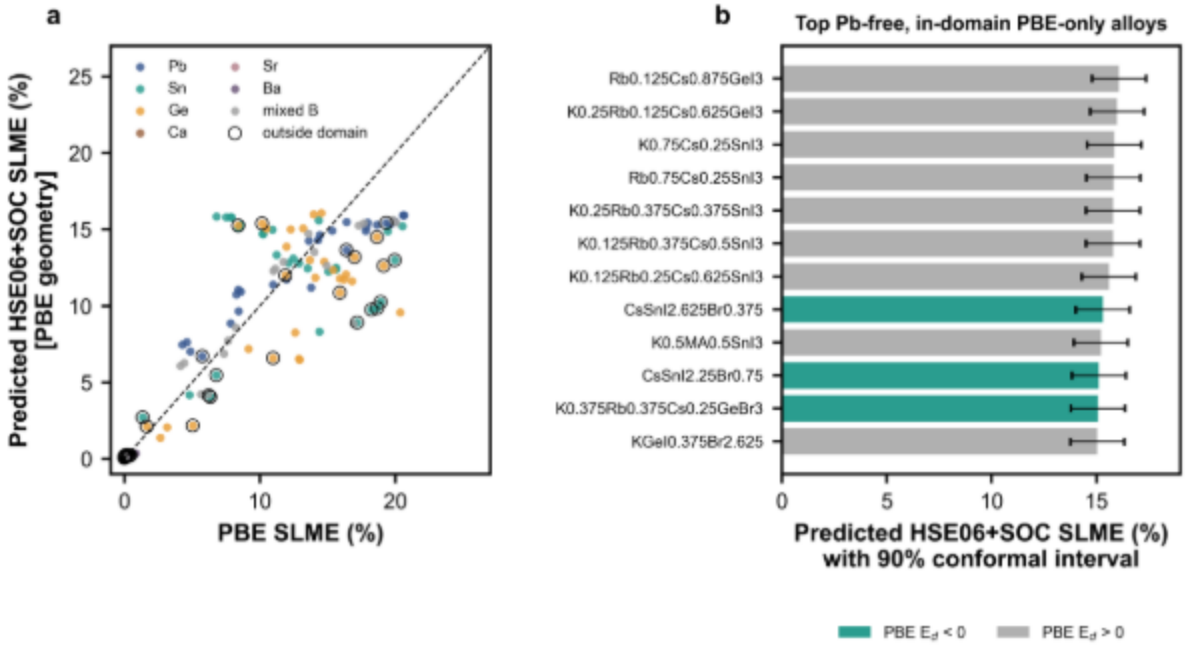


Figure 8

Prospective predictions for the 207 PBE-only compounds. (a) Predicted HSE06+SOC SLME (PBE geometry) versus PBE SLME; open circles mark compounds outside the applicability domain. (b) The twelve Pb-free compounds inside the domain with the highest predicted SLME, with 90% intervals from the grouped cross-validation residuals (Sect. 3.5); teal bars have negative PBE decomposition energy.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesallresults.xlsx
- SupplementaryInformation.docx
- Codeandprocesseddata.zip