Supplementary Information (SI) for Journal of Materials Chemistry A.

Supplementary Information

High-Throughput Screening of Piezoelectric Hybrid Organic–Inorganic Perovskites via Density-Functional Theory and Machine Learning

Jianxin He,[a] Tianle Yue,[a] Yu Lim Kim,[a] Ying Li[a]*

[a]*Department of Mechanical Engineering, University of Wisconsin-Madison, Madison, WI 53706, United States*

**Corresponding author: yli2562@wisc.edu*

S1. Benchmark

Table S1. Calculated piezoelectric stress constants (e_{ij}), elastic constants (C_{ij}), and piezoelectric strain constants (d_{ij}) for zinc oxide (ZnO). Other theoretical and experimental results are also shown for comparison.

	$e_{ij}(C/m^2)$			$c_{ij}(C/m^2)$						$d_{ij}(pC/N)$		
	e_{31}	e_{33}	e_{15}	c_{11}	c_{33}	c_{44}	c_{66}	c_{12}	c_{13}	d_{31}	d_{33}	d_{15}
Ours	-0.564	1.12	-0.412	204.77	202.42	34.33	37.60	129.57	105.42	-5.11	10.85	-12.01
Cal	-0.63	1.22	-0.46	202	219	36	39	125	110	-5.7	11.3	-12.8
Cal	-0.55	1.19	-0.46	246	246	56	-	127	105	-3.7	8.0	-8.2
Exp	-0.62	0.96	-0.37	209	216	44	-	120	104	-5.1	12.3	-8.3
ExP	-0.61	1.15	-	-	-	-	-	-	-	-	-	-

Table S2. Comparison between computed and experimental piezoelectric coefficients d_{33} for representative lead-halide hybrid organic–inorganic perovskites (HOIPs) from dataset of 1,346 ABX_3 HOIPs. For each composition, the computed $|d_{33}|$ corresponds to the lowest-energy structure in the dataset. Experimental values (in pm/V) were measured by piezoresponse force microscopy (PFM) and are reproduced from the indicated references. Note that 1 pC/N is numerically equivalent to 1 pm/V.

Composition	Experimental d_{33} (pm/V)	This work $\|d_{33}\|$ (pC/N)	Ref.
$MAPbCl_3$	3.4	5.9	[1]
$MAPbBr_3$	4.9	5.9	[1]
$MAPbI_3$	5.0	5.6	[2]

Table S3. Γ-point phonon mode summary for representative $MAPbX_3$ (X = I, Br, Cl) structures from the dataset. DFPT Calculations performed with VASP on the fully relaxed structures.

Composition	Real modes	Lowest real mode (cm^{-1})	Highest real mode (cm^{-1})	Imaginary modes (cm^{-1})
$MAPbI_3$	33	20.2	3243.6	1.20, 2.64, 6.56
$MAPbBr_3$	33	28.0	3286.9	2.03, 2.61, 3.45
$MAPbCl_3$	33	33.9	3306.5	4.83, 6.50, 7.21

The three imaginary modes per composition correspond to the acoustic branches at the Γ-point, with magnitudes within the numerical noise expected from finite-precision structural relaxation. No significant imaginary mode indicative of dynamical instability is observed.

S2. Distribution of DFT dataset for ML pretrain in this work

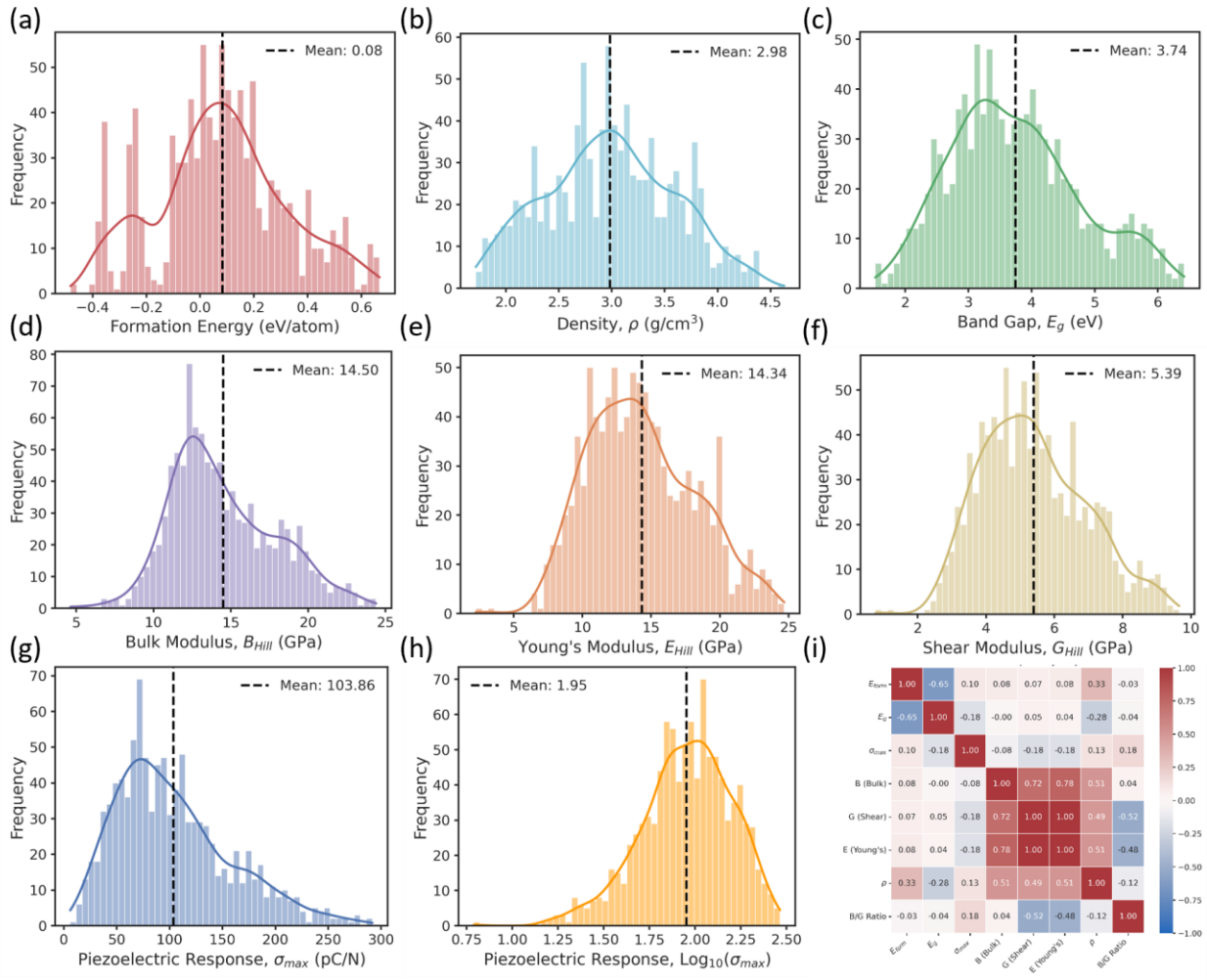

Fig. S1. Distributions of formation energy, density, band gap, elastic moduli, and piezoelectric response (σ_{max}), along with their pairwise correlations, for the 1,346 HOIP structures studied.

S3. Statistical distributions of full piezoelectric strain tensor components and descriptor

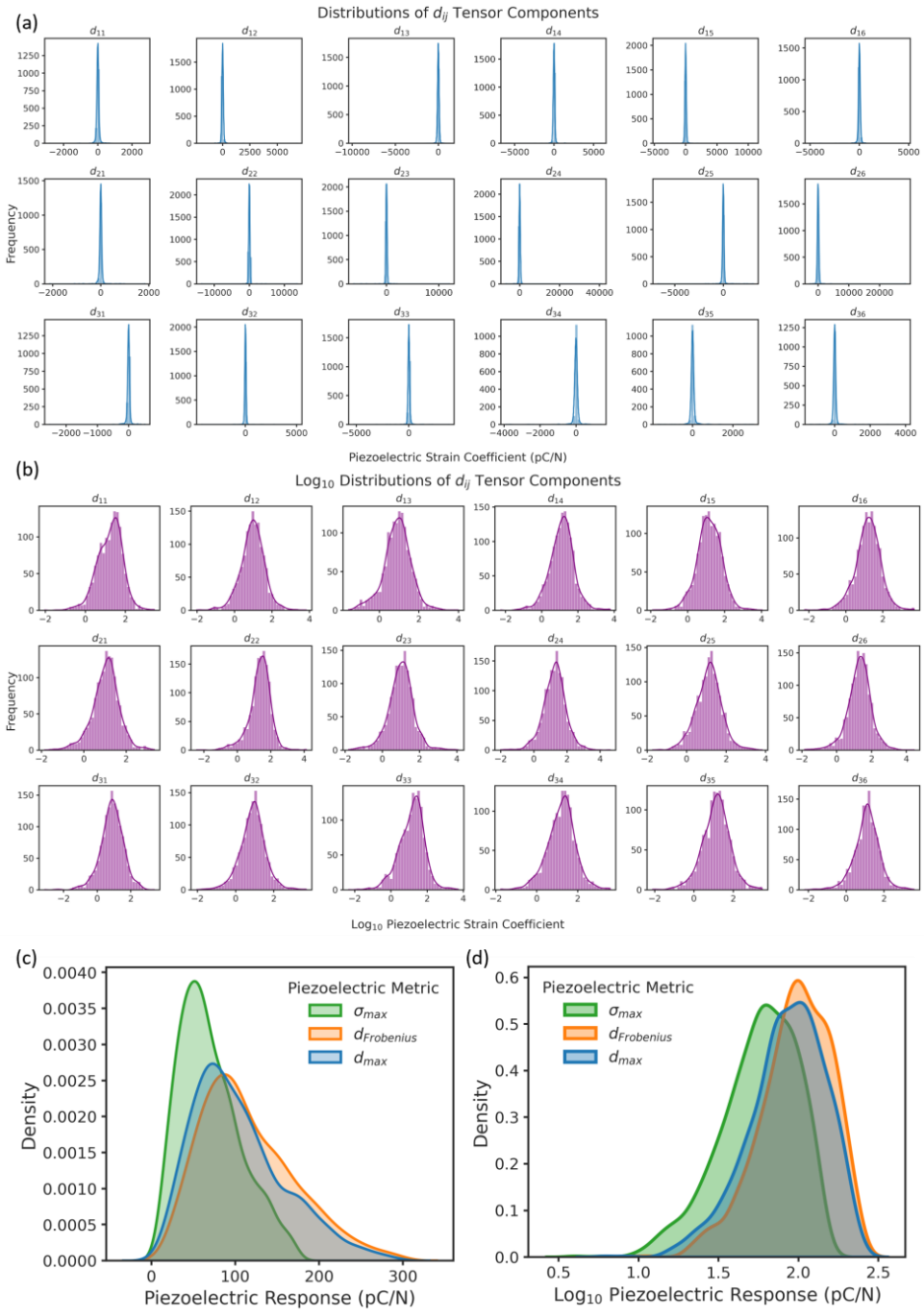

Fig. S2 Statistical distributions of piezoelectric strain tensor components (d_{ij}) and comparison of scalar descriptors (σ_{max}, d_{Frob}, d_{max}) for the HOIP dataset.

S4. ML for HOIPs properties prediction

Table S4 summarizes all the properties involved in the machine learning approaches used in this study, along with their corresponding sources.

Table S4. Summary of the properties included in the HOIP dataset and their sources.

	Property	Sources
#1	Atomization energy, Bandgap, Density, Dielectric constant (electronic), Dielectric constant (ionic), Dielectric constant (total), Refractive index, Relative energy1, Relative energy2, Volume of the unit cell.	*Scientific data*[3]
#2	Bulk modulus, Young's Modulus, Shear modulus, Piezoelectric descriptor	This study

In this study, we employed four atomistic feature engineering approaches to represent the crystal structures: the Atom-Centered Symmetry Functions (ACSF), the Ewald sum matrix, the sine matrix, and the Many-Body Tensor Representation (MBTR).

The sine matrix is similar in form to the Coulomb matrix but is specifically designed for periodic solids. The Coulomb matrix is calculated using the equation:

$$M_{ij}^{\text{Coulomb}} = \begin{cases} 0.5Z_i^{2.4} & \text{for } i = j \\ \dfrac{Z_i Z_j}{R_{ij}} & \text{for } i \neq j \end{cases}$$

where Z_i denotes the nuclear charge and R_{ij} denotes the distance between nuclei i and j.The diagonal elements represent the interaction of an atom with itself and can be regarded as a polynomial fit of the atomic energies to the nuclear charge Z_i, while the off-diagonal elements correspond to the Coulombic repulsion between nuclei i and j. In the sine matrix, the actual interatomic distances are replaced by sine functions, which effectively simulate the repeating interactions of the periodic crystal lattice. The matrix elements are defined by:

$$M_{ij}^{\text{sine}} = \begin{cases} 0.5Z_i^{2.4} & \text{for } i = j \\ \dfrac{Z_i Z_j}{\left|\mathbf{B} \cdot \sum_{k=\{x,y,z\}} \hat{\boldsymbol{e}}_k \sin^2\left(\pi \hat{\boldsymbol{e}}_k \mathbf{B}^{-1} \cdot \left(\boldsymbol{R}_i - \boldsymbol{R}_j\right)\right)\right|} & \text{for } i \neq j \end{cases}$$

where $\boldsymbol{B}$ is a matrix formed by the lattice vectors and $\hat{\boldsymbol{e}}_k$ are the cartesian unit vectors. Despite lacking an explicit physical interpretation, this functional form reproduces essential features of the Coulomb interaction, such as the periodic behavior of the crystal lattice and the singularity in energy at atomic overlap.

The Ewald sum matrix can be viewed as a natural extension of the Coulomb matrix to periodic systems, capturing the electrostatic interactions between atoms in periodic crystalline structures.

In periodic systems, every atom is periodically replicated along the three lattice vectors (a, b, and c), leading to an electrostatic interaction between atoms that can be expressed as

$$\phi_{ij} = \sum_{\boldsymbol{n}} \frac{Z_i Z_j}{\left|\boldsymbol{R}_i - \boldsymbol{R}_j\right| + \boldsymbol{n}}$$

where $\sum_{\boldsymbol{n}}$ is the sum over all lattice vectors $\boldsymbol{n} = h\boldsymbol{a} + k\boldsymbol{b} + l\boldsymbol{c}$.

The ACSF framework characterizes the local atomic environment through a fingerprint composed of several two- and three-body symmetry functions, each designed to capture specific structural motifs. For each function type, multiple parameterizations are employed to encode various regions of the chemical environment. Two- body symmetry functions include:

$$G_i^{1,Z_1} = \sum_j^{|Z_1|} f_c(R_{ij}),$$

$$G_i^{2,Z_1} = \sum_j^{|Z_1|} e^{-\eta\left(R_{ij}-R_s\right)^2} f_c(R_{ij})$$

$$G_i^{3,Z_1} = \sum_j^{|Z_1|} \cos(\kappa R_{ij})\, f_c(R_{ij})$$

where the summation for j runs over all atoms with atomic number Z_1, η, R^s and κ are user-defined parameters, $R_{ij} = \left|\boldsymbol{R}_i - \boldsymbol{R}_j\right|$ and $f_c(r) = \frac{1}{2}\left[\cos\left(\pi \frac{r}{r_{cut}}\right) + 1\right]$, where r_{cut} is a cutoff radius. Furthermore, three-body symmetry functions are utilized to capture characteristic motifs involving three atoms, one of which acts as the central atom. These functions incorporate both the angular dependence between atomic triplets within the cutoff and their interatomic distances.

The Many-Body Tensor Representation (MBTR) encodes both finite and periodic atomic structures by decomposing them into distributions of structural motifs of different orders, such as single atoms, pairs, and triplets, and grouping these according to the chemical species involved. For each order k , a geometry function g_k transforms the atomic configuration into a scalar value that represents a particular geometric property: atomic number for $k = 1$, (inverse) interatomic distance for $k = 2$, and angles or cosines of angles for $k = 3$. These scalar values are broadened through kernel density estimation using a Gaussian kernel to produce continuous distributions. The final MBTR descriptor is obtained by taking weighted sums of these distributions for all relevant combinations of atomic species.[4] Fig. S3 illustrates the performance of the four feature engineering methods on Dataset #1.

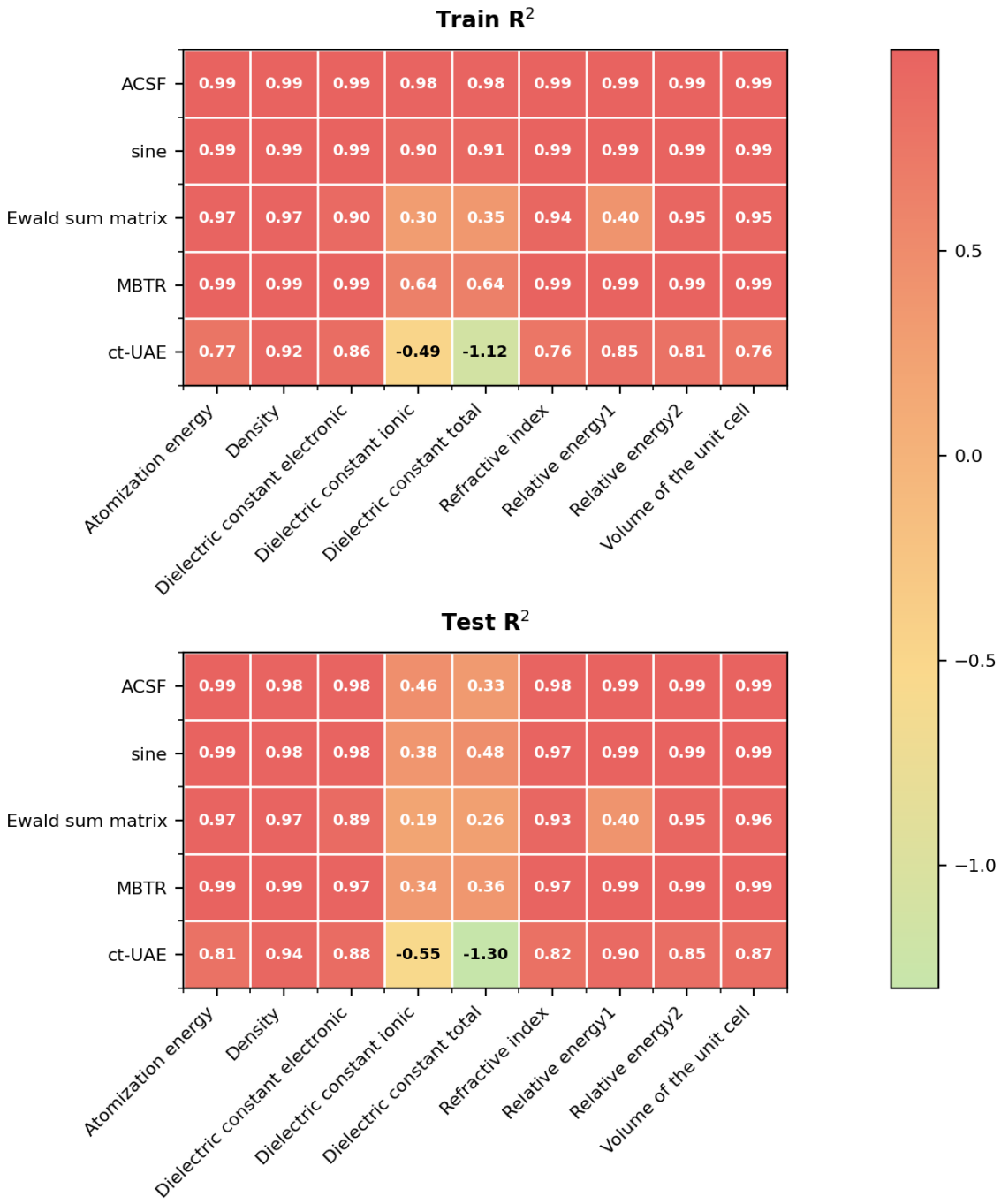

Fig. S3. R^2 for the training and test sets of all target properties using four different structural representations with FNN models, as well as results from the fine-tuned Transformer-based ct-UAE model.

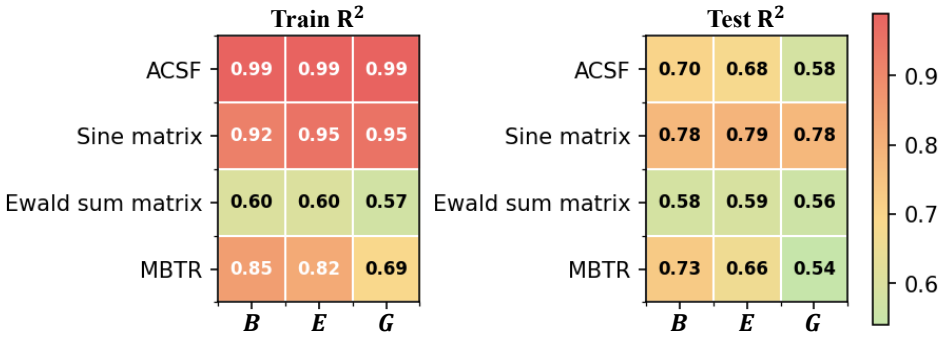

Fig. S4. R^2 for the training and test sets of the three mechanical properties B, E, and G.

Detailed comparisons reveal that when ACSF was used as the input representation, the FNN models achieved relatively high R^2 values on the training set for all three mechanical properties. When the sine matrix representation was employed, the training R^2 values were slightly lower than those obtained with ACSF but remained within an acceptable range, indicating strong fitting capability. In contrast, the Ewald sum matrix and MBTR representations produced comparatively lower training R^2 values. However, a markedly different trend was observed on the test set. The ACSF-based models exhibited a substantial drop in R^2 values for all three moduli, indicating severe overfitting. Although the MBTR and Ewald sum matrix representations did not display pronounced overfitting behavior, their overall predictive performance remained unsatisfactory. By comparison, the sine matrix representation maintained relatively high R^2 values for B, E, and G on the test set, demonstrating superior generalization capability.

Reference

1. D. B. Kim, K. S. Jo, K. S. Park and Y. S. Cho, Adv. Sci., 2023, 10, 2204462.
2. M. Coll, A. Gomez, E. Mas-Marza, O. Almora, G. Garcia-Belmonte, M. Campoy-Quiles and J. Bisquert, J. Phys. Chem. Lett., 2015, 6, 1408-1413.
3. C. Kim, T. D. Huan, S. Krishnan and R. Ramprasad, Sci. Data, 2017, 4, 1-11.
4. L. Himanen, M. O. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, Comput. Phys. Commun., 2020, 247, 106949.