Supporting Information

Uncertainty-Aware Dimensionality Prediction of Low-Dimensional Hybrid Metal Halides by Integrating Bayesian Modeling and Experiments

Migon Choi[a], Dongkyu Derek Cho[b], Richard Sheridan[a], David B. Mitzi[a,c], and L.Catherine Brinson[a*]

[a]Department of Mechanical Engineering and Materials Science, Duke University, Durham, North Carolina, 27708, USA

[b]Department of Statistical Science, Duke University, Durham, North Carolina, 27708, USA

[c]Department of Chemistry, Duke University, Durham, North Carolina 27708, USA

*email: cate.brinson@duke.edu

1. Single Crystal Growth

Crystals in this study were synthesized using the slow evaporation method. 3,3-dimethylbutylamine (95%, Acros Organics), 2,3-dimethylbutan-2-amine (95%, Advanced ChemBlocks), N-methylbutan-2-amine (Biosynth), lead iodide (99.9999%, TCI Chemicals), methanol (VWR Chemicals), hydroiodic acid (99.95%, Sigma Aldrich) were used as received. For all crystallizations, 0.1 mmol of PbI_2 was used as the lead precursor, while the amount of spacer cation varied depending on the targeted molar ratio: 0.1 mmol for 1:1, 0.2 mmol for 2:1, and 0.4 mmol for 4:1. For 3,3-dimethylbutylamine, identical crystals were obtained across all ratios. In the 1:1 ratio, 0.1 mmol of PbI_2 and 0.1 mmol of 3,3-dimethylbutylamine were dissolved in 1 mL of HI and 1 mL of methanol. For N-methylbutan-2-amine, the 2:1 and 4:1 ratio yielded the same

crystals, whereas the 1:1 ratio resulted in a different crystal form. The 1:1 ratio solution was prepared by dissolving 0.1 mmol of PbI_2 and 0.1 mmol of N-methylbutan-2-amine in 0.15 - 0.2 mL of HI and 2mL of methanol. For the 2:1 and 4:1 ratios, 0.1 mmol of PbI_2 and 0.2 mmol of N-methylbutan-2-amine were dissolved in 0.2 mL and 0.3 mL of HI and 0.3 - 0.4 mL of methanol. In the case of 2,3-dimethylbutan-2-amine, the 1:1 and 2:1 ratios yielded identical crystals, whereas the 4:1 ratio resulted in a distinct crystal form. The 1:1 ratio solution was prepared by dissolving 0.1 mmol of PbI_2 and 0.1 mmol of 2,3-dimethylbutan-2-amine in 0.15 mL of HI and 0.15 mL of methanol, while for the 4:1 ratio, 0.1 mmol of PbI_2 and 0.4 mmol of 2,3-dimethylbutan-2-amine were dissolved in 0.3 mL of HI and 0.4 mL of methanol. To inhibit solution oxidation, 10-20% of H_3PO_3 was added. All precursor solutions were prepared and stored in a nitrogen-filled glovebox until crystallization was complete.

2. Characterization

2.1. Powder X-ray Diffraction (XRD)

Powder XRD analysis was performed for the powder samples using a PANalytical Empyrean powder X-ray diffractometer operating at 45 kV and 40 mA, with Cu Kα radiation. The powder samples were prepared by finely grinding the crystals on glass slides (**Figs. S1 – S3**).

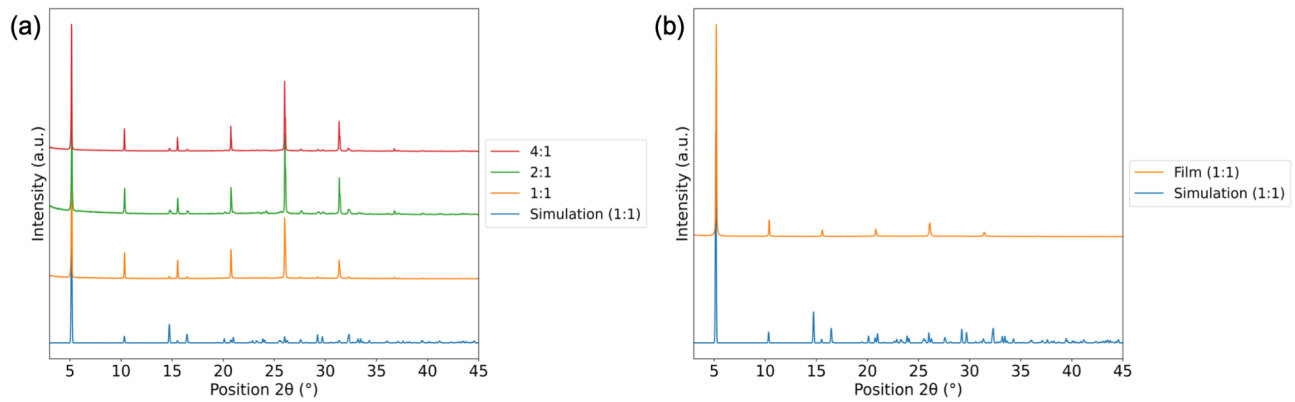

Figure S1. PXRD patterns of 3,3-dimethylbutylamine-derived samples with varying precursor ratios: (a) crystals formed by mixing 3,3-dimethylbutylamine and PbI_2; (b) thin film deposited using a 1:1 molar ratio of amine to PbI_2. The simulated PXRD pattern was generated in CrystalDiffract.

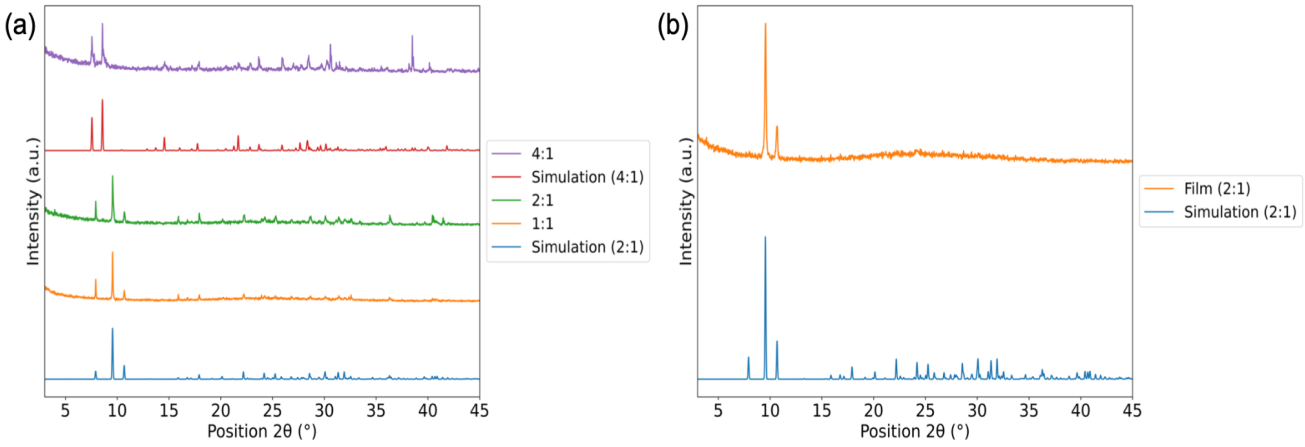

Figure S2. PXRD patterns of 2,3-dimethylbutan-2-amine-derived samples with varying precursor ratios: (a) crystals formed by mixing 2,3-dimethylbutan-2-amine and PbI_2; (b) thin film deposited using a 2:1 molar ratio of amine to PbI_2. The simulated PXRD pattern was generated in CrystalDiffract.

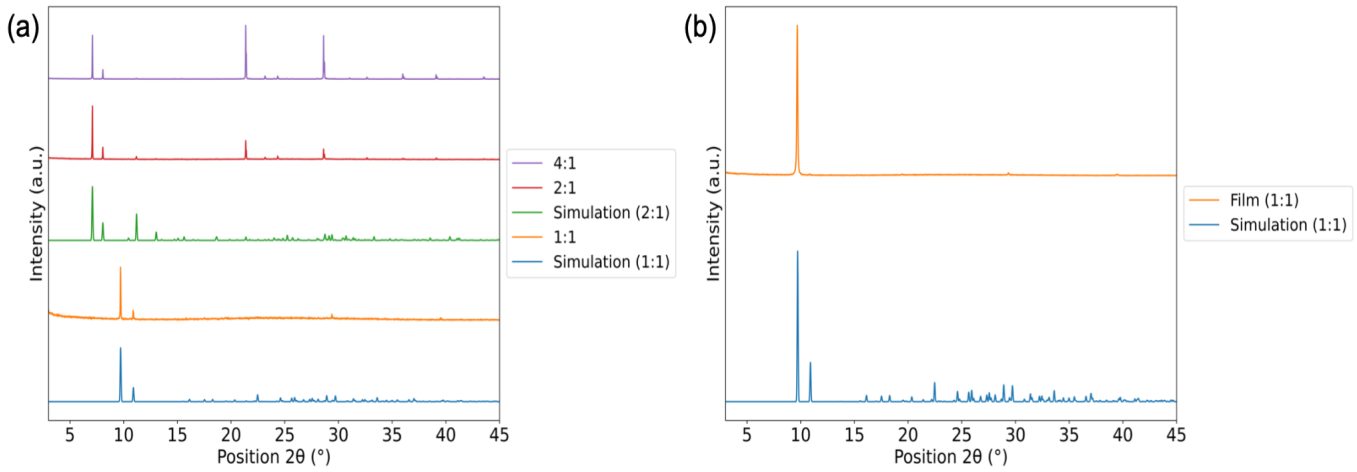

Figure S3. PXRD patterns of N-methylbutan-2-amine-derived samples with varying precursor ratios: (a) crystals formed by mixing N-methylbutan-2-amine and PbI_2; (b) thin film deposited using a 1:1 molar ratio of amine to PbI_2. The simulated PXRD pattern was generated in CrystalDiffract.

2.2. Optical Properties

2.2.1. Ultraviolet-visible Spectroscopy

UV-Vis-NIR absorbance measurements were conducted using a Shimadzu UV-3600i UV-Vis-NIR Spectrophotometer equipped with an integrating sphere (ISR-603) to account for diffuse reflectance (**Figs. S4 – S6**). The baseline was corrected to eliminate instrumental artifacts. Absolute intensities were normalized. Single crystals were dissolved in DMF, targeting a concentration of 0.2 M (±0.07 M). The solutions were spin-coated onto 1 cm × 1 cm glass substrates at 3000 rpm for 30 s, followed by annealing at 100°C for 10 min.

2.2.2. Photoluminescence Spectroscopy

PL measurements were performed using a Horiba Jobin Yvon LabRam ARAMIS Raman/PL Spectrometer, equipped with four excitation lasers (325 nm, 442 nm, 633 nm, and 785 nm) (**Figs. S4 – S6**). Data collection was conducted using the LabSpec software package. PL measurements were conducted on both thin films and single crystals depending on the sample. For 3,3-dimethylbutan-2-amine perovskite, a thin-film sample was prepared following the same procedure as UV-Vis spectroscopy. For the other two hybrid metal halide (HMH) samples, single crystals were used for PL analysis without further processing. For 3,3-dimethylbutan-2-amine perovskite, PL spectra were recorded using a 442 nm excitation laser with a 50% power filter. For all other HMH samples, PL spectra were acquired using a 325 nm excitation laser, with either a 50% or 100% power filter, depending on the sample.

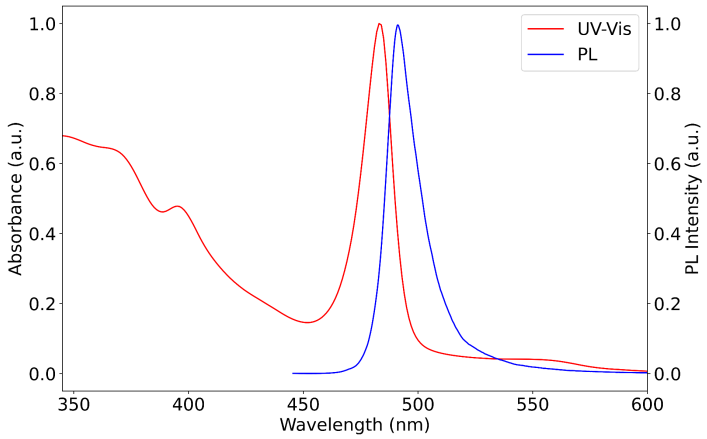

Figure S4. Normalized (a) UV–Vis absorption and (b) photoluminescence (PL) spectra of (3,3-DMBA·H)$_2$PbI$_4$.

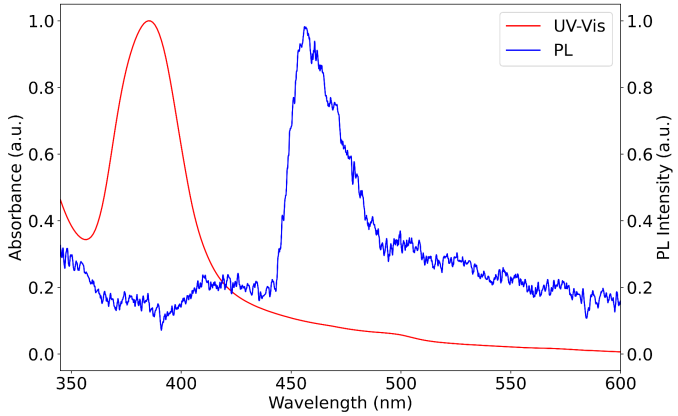

Figure S5. Normalized (a) UV–Vis absorption and (b) PL spectra of (2,3-DMB2A·H)PbI_3.

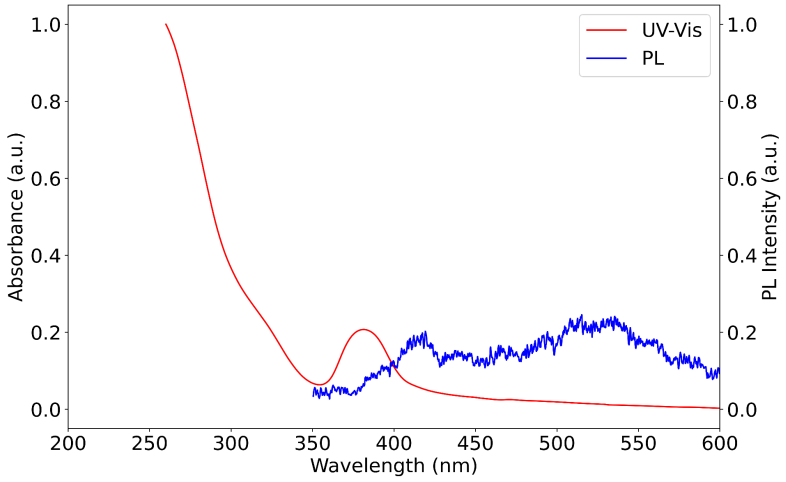

Figure S6. Normalized (a) UV–Vis absorption and (b) PL spectra of (NMB2A·H)PbI_3.

2.3. Single Crystal X-ray Diffraction (SCXRD)

SCXRD characterization was performed on a Rigaku XtaLAB Synergy-S diffractometer using Mo Kα radiation (λ = 0.71073 Å) and operating at 50 kV and 30 mA at room temperature (298 K). Peak hunting, data reduction, and numerical absorption correction for all collected data were performed using CrysAlisPro. The crystal structures were solved and refined using SHELXS direct

methods and SHELXL least-squares method within the Olex2 software package. Crystallographic and structural refinement data, along with summaries of lattice parameters and structural distortion metrics, are provided for all synthesized structures in **Table S1** and **Table S2**.

To probe the effect of precursor stoichiometry, additional samples with 2:1 and 4:1 spacer-to-lead ratios were characterized. PXRD and SCXRD analysis suggest that these variations can lead to changes in structural connectivity (**Fig. S1 - S3**), 1D corner sharing (**Fig. S7**) and 1D trimer (**Fig. S8**) structures, indicating the potential sensitivity of perovskite architecture to compositional tuning. SCXRD analysis of N-methylbutan-2-ammonium 2:1 ratio with lead iodide crystals indicated the presence of solvent molecules in the crystal structure (**Fig. S8**). To confirm this, thermogravimetric analysis (TGA) was performed, which verified solvent incorporation (**Fig. S9**). To further investigate solvent incorporation, Fourier-transform infrared (FTIR) spectroscopy was performed on the synthesized crystals (**Fig. S10**). T he FTIR spectrum of N-methylbutan-2-ammonium 2:1 ratio with lead iodide crystals (Fig**. S10(a)**) exhibits a broad absorption band centered around 3200-3400 cm^{-1} (the orange highlighted-region in Fig. S10(a)), arising from overlapping N-H and O-H stretching vibrations. This feature, together with the small mass loss below ~150 °C observed in TGA (**Fig. S9(a)**), suggests the presence of a small amount of hydrogen-bonded water or residual solvent within the crystal structure. In contrast, the spectrum of $(3,3\text{-DMBA}\cdot H)_2PbI_4$ (**Fig. S10(b)**) shows only weak features in this region, and its TGA profile lacks a low-temperature weight-loss step (**Fig. S9(b)**), indicating negligible solvent inclusion.

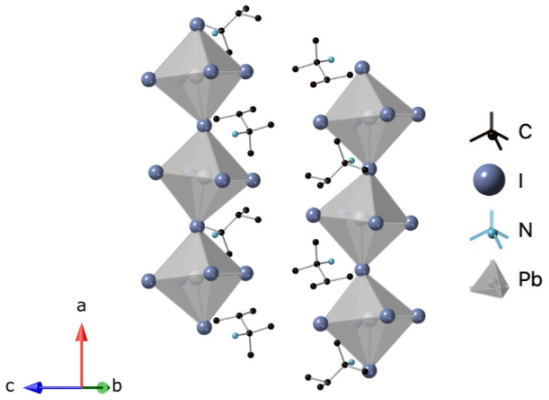

Figure S7. Schematic single-crystal structures of 2,3-dimethylbutan-2-ammonium 4:1 ratio with lead iodide ($(2,3\text{-DMB2A}\cdot H)_3PbI_5$).

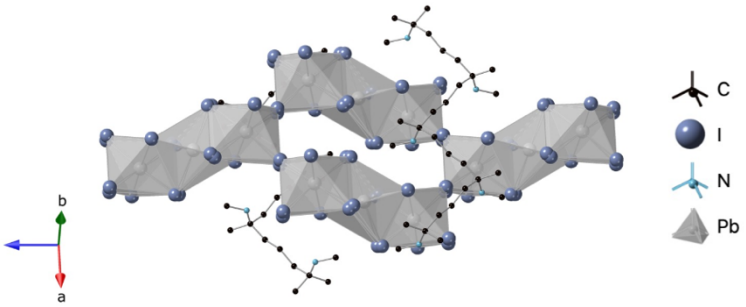

Figure S8. Schematic single-crystal structures of N-methylbutan-2-ammonium 2:1 ratio with lead iodide.

Table S1. Crystallographic and structural refinement data

	3,3-DMBA	NMB2A	NMB2A	2,3-DMB2A	2,3-DMB2A
Amine : Pb Ratio	1:1 Ratio	1:1 Ratio	2:1 Ratio (includes defect)	4:1 Ratio	2:1 Ratio
Empirical formula	$C_{12}H_{32}I_4N_2Pb$	$C_5H_{14}I_3NPb$	$C_{3.69}H_{4.31}I_{3.03}N_{0.62}Pb_{0.92}$	$C_{18}H_{48}I_5N_3Pb$	$C_{48}H_{128}I_{24}N_8Pb_8$
Formula weight	919.18	676.06	641.68	1148.28	5520.7
Temperature/K	296.8(4)	296.92(17)	293	297.1	297.2

Crystal system	monoclinic	orthorhombic	orthorhombic	orthorhombic	orthorhombic
Space group	*C*2/*c*	$P2_12_12_1$	*Cmce*	*Pnma*	*Pbca*
a/Å	34.2229(19)	8.0027(3)	8.9213(3)	12.5943(3)	16.5610(5)
b/Å	8.4548(3)	10.9820(5)	24.9062(11)	23.3662(6)	7.9958(3)
c/Å	8.8239(3)	16.2078(6)	23.0237(8)	11.4813(3)	22.3017(7)
α/°	90	90	90	90	90
β/°	90.303(4)	90	90	90	90
γ/°	90	90	90	90	90
Volume/Å^3	2553.14(19)	1424.43(10)	5115.8(3)	3378.73(15)	2953.16(17)
Z	4	4	13	4	1
ρcalc/ g/cm^3	2.391	3.152	2.673	2.257	3.104
μ/mm^{-1}	11.44	18.302	15.801	9.571	17.659
F(000)	1648	1168	3474	2096	2400
Crystal size/mm^3	0.03 × 0.03 × 0.005	0.2 × 0.1 × 0.1	0.2 × 0.05 × 0.05	0.05 × 0.03 × 0.015	0.2 × 0.1 × 0.1
Radiation	Mo Kα (λ = 0.71073)	Mo Kα (λ = 0.71073)	MoKα (λ = 0.71073)	Mo Kα (λ = 0.71073)	Mo Kα (λ = 0.71073)
2Θ range for data collection/ °	4.762 to 52.74	4.48 to 54.964	3.718 to 52.744	3.486 to 61.612	5.946 to 48.97
Index ranges	-42 ≤ h ≤ 40, -10 ≤ k ≤ 9, -10 ≤ l ≤ 10	-9 ≤ h ≤ 9, -14 ≤ k ≤ 14, -20 ≤ l ≤ 19	-10 ≤ h ≤ 11, -31 ≤ k ≤ 31, -28 ≤ l ≤ 28	-16 ≤ h ≤ 16, -26 ≤ k ≤ 33, -14 ≤ l ≤ 14	-19 ≤ h ≤ 19, -9 ≤ k ≤ 9, -26 ≤ l ≤ 24
Reflections collected	10227	12190	25807	36564	19749
Independent reflections	2599 [R_{int} = 0.0355, R_{sigma} = 0.0296]	3162 [R_{int} = 0.0289, R_{sigma} = 0.0276]	2760 [R_{int} = 0.0335, R_{sigma} = 0.0178]	4569 [R_{int} = 0.0331, R_{sigma} = 0.0215]	2428 [R_{int} = 0.0427, R_{sigma} = 0.0240]
Data/restraints/parameters	2599/0/56	3162/157/63	2760/77/98	4569/64/144	2428/0/105
Goodness-of-fit on F^2	1.037	1.031	1.067	1.065	1.169
Final R indexes [I>=2σ (I)]	R_1 = 0.0363, wR_2 = 0.0923	R_1 = 0.0317, wR_2 = 0.0798	R_1 = 0.0399, wR_2 = 0.0986	R_1 = 0.0296, wR_2 = 0.0630	R_1 = 0.0243, wR_2 = 0.0565

Final R indexes [all data]	R_1 = 0.0463, wR_2 = 0.0971	R_1 = 0.0496, wR_2 = 0.0856	R_1 = 0.0561, wR_2 = 0.1060	R_1 = 0.0447, wR_2 = 0.0669	R_1 = 0.0270, wR_2 = 0.0575
Largest diff. $^{-3}$ peak/hole / e Å $^{-3}$	1.96/-1.21	0.86/-0.70	1.93/-0.58	1.77/-1.91	0.63/-1.88
Flack parameter		-0.009(6)			

Table S2. Summary of lattice parameters and structural distortion parameters

	3,3-DMBA	NMB2A	NMB2A	2,3-DMB2A	2,3-DMB2A
Amine : Pb Ratio	1:1 Ratio	1:1 Ratio	2:1 Ratio	4:1 Ratio	2:1 Ratio
Lattice Parameters	34.2229(19) Å 8.4548(3) Å 8.8239(3) Å	8.0027(3) Å 10.9820(5) Å 16.2078(6) Å	8.9213(3) Å 24.9062(11) Å 23.0237(8) Å	12.5943(3) Å 23.3662(6) Å 11.4813(3) Å	16.5610(5) Å 7.9958(3) Å 22.3017(7) Å
Octahedral bond distortion Δd	1.06×10^{-5}	1.74×10^{-4}	Pb1: 5.52×10^{-4} Pb2: 4.01×10^{-5}	6.63×10^{-4}	6.88×10^{-4}
Octahedral angle variance σ^2	8.04×10^{0}	2.47×10^{1}	Pb1: 1.79×10^{1} Pb2: 1.98×10^{1}	1.57×10^{1}	5.47×10^{1}

2.4. Thermogravimetric Analysis (TGA)

For N-methylbutan-2-amine 2:1 ratio, TGA measurements were performed on a TA Q50 instrument at a ramp rate of 2 °C/min from 30 °C to 350 °C. For 3,3-dimethylbutan-2-amine crystals, TGA measurements were performed on a TA Instruments TGA5500, at a ramp rate of 20 °C/min from room temperature to 350 °C.

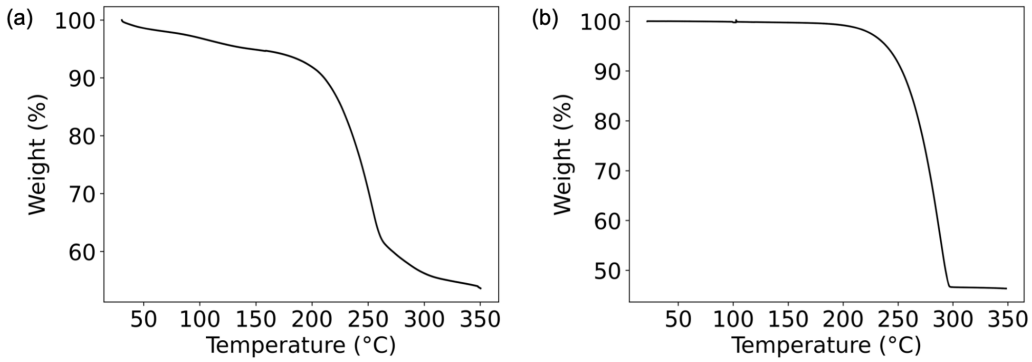

Figure S9. Thermogravimetric analysis (TGA) of (a) N-Methylbutan-2-amine crystal with a 2:1 molar ratio and (b) $(3,3\text{-DMBA}\cdot H)_2PbI_4$.

2.5. Fourier-transform infrared (FTIR) spectroscopy

FTIR spectroscopy is conducted using a Thermo Scientific Nicolet iS50 FTIR spectrometer equipped with a RaptIR+ microscope (Thermo Fisher Scientific, USA). Spectra were collected in the mid-IR range (4000–400 cm^{-1}).

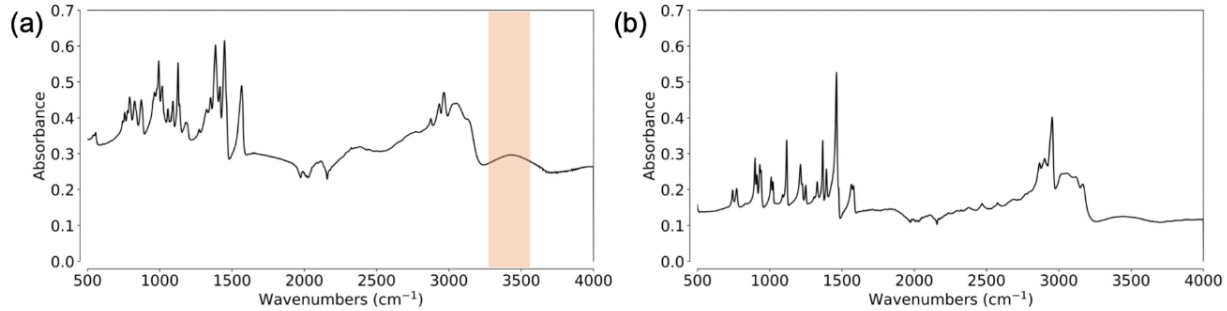

Figure S10. FTIR spectra of HMH crystals: (a) N-methylbutan-2-ammonium 2:1 ratio with lead iodide and (b) $(3,3\text{-DMBA}\cdot H)_2PbI_4$.

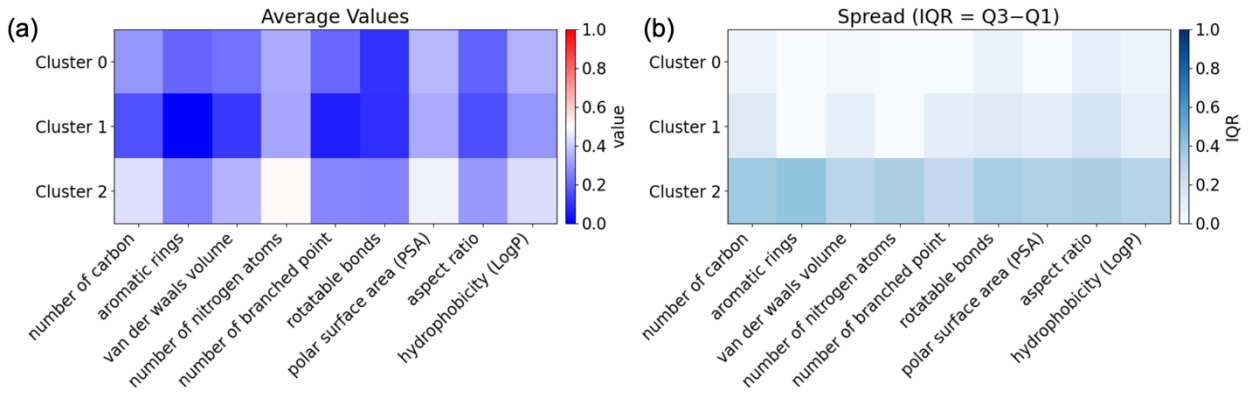

Figure S11. Descriptor statistics for each cluster identified in the PaCMAP embedding. (a) Average normalized descriptor values for each cluster, and (b) spread of each descriptor, represented as the interquartile range (IQR = Q3 – Q1).

3. Data Handling

We retrieved entries from the Hybrid^3 database [1] and applied the following filtering criteria to ensure consistency and relevance: (1) the inorganic component is lead iodide, (2) the structure is either 1D or 2D, (3) the sample is a single crystal (excluding films, powders, or bulk), (4) the organic spacer's IUPAC name could be converted to a SMILES string via OPSIN or Leskoff [2, 3]. During the filtering process, approximately 20 entries were excluded in step (3) and (4), mainly due to insufficient structural annotation and issues converting names to SMILES. To address the limited number of 1D structures (11 in Hybrid^3), we manually curated additional 1D samples from the literature using the same selection process. We constructed a dataset comprising 113 perovskite samples, including 76 with 2D and 37 with 1D structures. Data were split into train and test set (Train:84, Test: 29) for 3.2.1. Model Performance and Uncertainty Quantification and 3.2.2. Feature Importance) and into train, test, and pool set (Train: 69, Test: 29, Pool: 15) for 3.2.3. Active

Learning. For Active learning, out of 15 data in pool set, 6 highest information gain samples were added to the train set.

4. PaCMAP

To visualize and compare molecular descriptors, we employed PaCMAP (Pairwise Controlled Manifold Approximation)[4], a nonlinear dimensionality reduction method designed to preserve both local neighborhood structure and global data geometry. The dataset was first embedded into a 2D latent space using PaCMAP for visualization. To interpret the structure of the embedding, we applied post hoc clustering techniques tailored to each analysis stage. For the full dataset, the PaCMAP map was used purely for visualization of the descriptor-based relationships among samples. The PaCMAP visualization separated the samples into three groups, which informed our understanding of the underlying chemical space. To assign convenient labels to these groups, we applied K-means clustering. Setting K=3 reproduced the same grouping observed from PaCMAP, providing a consistent labeling scheme for reference (**Fig. 3**). In contrast, for the test set, DBSCAN, a non-parametric density-based clustering, was used to capture naturally occurring density-based patterns, as the smaller sample size and more heterogeneous distribution made density-based clustering more appropriate (**Fig. 7**). Categorical descriptor, "nitrogen atom position" is omitted for PaCMAP analysis. To further interpret the embedding, we analyzed molecular descriptor statistics for each cluster (**Fig. S11**). The average and interquartile ranges of normalized descriptors highlight distinct chemical tendencies across clusters, supporting the clustering results and clarifying the positioning of the synthesized compounds.

5. Descriptors

Molecular descriptors, which were not included in the Hybrid3 database, were calculated using the RDKit cheminformatics software[5]. SMILES strings of the compounds were converted into molecular objects, and the following descriptors were computed: number of carbon atoms, nitrogen atom positions, number of nitrogen atoms, number of rotatable bonds, number of aromatic rings, hydrophobicity (LogP), polar surface area (PSA), number of branched points, aspect ratio, and van der Waals volume (see **Table S3** and codes in the provided GitHub repository, see Data & Code Availability in the main text).

Table S3. Molecular descriptors and their definitions. Crystal names and dimensionality data were collected from Hybrid[3] database, published literature, and synthesized structures, with molecular descriptors computed using RDKit.

Descriptor	**Definition**	**Implementation (RDKit function / method)**
number of carbon	Counts the number of carbon (C) atoms in the molecule.	Manual count: atom.GetAtomicNum() == 6.
nitrogen atom position	Returns the indices (positions) of nitrogen (N) atoms in the molecule.	Manual: [atom.GetIdx() for atom in mol.GetAtoms() if atom.GetSymbol() == 'N'].
number of nitrogen atoms	Counts the number of nitrogen (N) atoms in the molecule.	Manual count: atom.GetSymbol() == 'N'.
rotatable bonds	Number of rotatable bonds in the molecule.	rdkit.Chem.Descriptors.NumRotatableBonds(mol).
aromatic rings	Number of aromatic rings in the molecule.	rdkit.Chem.rdMolDescriptors.CalcNumAromaticRings(mol).
hydrophobicity (LogP)	Molecular hydrophobicity, calculated as the octanol–water partition coefficient (LogP).	rdkit.Chem.Descriptors.MolLogP(mol).
polar surface area (PSA)	Polar surface area (TPSA) of the molecule, with additional correction for positively charged sulfur (S^+) atoms.	rdkit.Chem.Descriptors.TPSA(mol) with additional correction for positively charged sulfur (S^+) atoms.

number of branched point	Approximated by the number of aliphatic heterocycles in the molecule.	rdkit.Chem.rdMolDescriptors.CalcNumAliphaticHeterocycles(mol)
aspect ratio	Ratio of molecular length to width, derived from 3D coordinates of atoms.	rdkit.Chem.AllChem.EmbedMolecule(mol, AllChem.ETKDG()) + manual length/width calculation,
van der Waals volume	Estimated molecular volume, calculated by summing the van der Waals radii of atoms.	Manual calculation with predefined radii dictionary.

6. Random Forest Classifier (RF)

A random forest[6] is an ensemble learning method that combines many decision trees to improve predictive performance and reduce overfitting. Each tree is trained on a bootstrap-resampled subset of the data, and at each split only a random subset of features is considered, which decorrelates the trees. Predictions from all trees are then aggregated by majority vote for classification to produce the final output. This classifier produces a binary true or false categorization. In a Random Forest, multiple trees independently make predictions (1D or 2D), and the final decision is determined by majority voting. While this ensemble strategy improves performance, it yields only a single estimate without quantifying predictive confidence (**Fig. S14**). In contrast, the BART framework provides a full posterior distribution over outcomes, enabling direct assessment of uncertainty through predictive probabilities and credible intervals. This probabilistic formulation offers a more informative and interpretable representation of model predictions (**Fig. 4(a)**). Further details on BART are presented in the following section. All computations were conducted in Python 3.9.20 using NumPy 1.24.4, Matplotlib 3.9.4, and scikit-learn 1.5.1.

7. Bayesian Additive Regression Trees (BART)

We employ Bayesian Additive Regression Trees (BART) to model the probabilistic relationship between the binary outcome $y \in \{0,1\}$ and a set of predictor variables x. In this study, the predictor

variables x correspond to the molecular and structural descriptors, where the binary outcome y denotes the crystal dimensionality (1D or 2D).

BART models the latent response as an additive ensemble of J regression trees,

$$f(x) = \sum_{j=1}^{J} g(x; T_j, M_j),$$

where each tree T_j defines a partition of the predictor space into non-overlapping terminal regions (also known as terminal nodes or leaves), as $R_{j1}, R_{j2}, \ldots, R_{jB_j}$ and $M_j = \{\mu_{j1}, \mu_{j2}, \ldots, \mu_{jB_j}\}$ denotes the corresponding mean parameters associated with each region. The contribution of the j-th tree is then

$$g(x; T_j, M_j) = \mu_{jb} \; if \;\; x \in R_{jb}.$$

Using a logistic link function $\phi(u) = 1/(1 + exp\,(-u))$, the model estimates the mean of the Bernoulli distribution as

$$Y \sim Ber[\,\phi\{\textstyle\sum_j \quad g(x; T_j, M_j)\}]$$

For the prior distributions, we follow the guidelines proposed by Chipman et al.[7] and Quiroga et al[8]. For the completeness of the paper, we provide details of the prior choices. Here, the prior on each tree structure T_j controls tree depth and variable selection. Specifically, the probability that a node at depth $\delta$is non-terminal is $0.95(1 + \delta)^{-2}$, while both the splitting variable and splitting rule are drawn from uniform distributions. Terminal node parameters $\mu_{jb}$are assigned independent normal priors,

$$\mu_{jb} \sim N(0, \sigma_\mu^2)$$

Where we choose $\sigma_{\mu}^{2} = \frac{9}{J}$ which controls the spread of the distribution. These hierarchical priors encourage shallow trees and ensure the additive ensemble remains well-regularized.

For posterior inference, we employed the PGBART sampler proposed by Quiroga et al.[8] with 10,000 samples. Convergence diagnostics were assessed using the effective sample size (ESS) and the Gelman–Rubin statistic ($\hat{R}$). As shown in **Figure S12**, ESS values were sufficiently large (>3000), and $\hat{R}$ values were close to 1 for all parameters, indicating satisfactory convergence and efficient mixing across Markov chains. All computations were conducted in Python 3.9.20 using NumPy 1.24.4, Matplotlib 3.9.4, scikit-learn 1.5.1, PyMC 5.12.0, and pymc-bart 0.5.11.

Figure S13 shows the marginal posterior distributions across. In the rank plot, each color represents a different MCMC chain. The accompanying rank plots display approximately uniform ranks for all chains, confirming that posterior samples are well-mixed and evenly distributed across the parameter space. Together, these diagnostics provide strong evidence of reliable posterior inference.

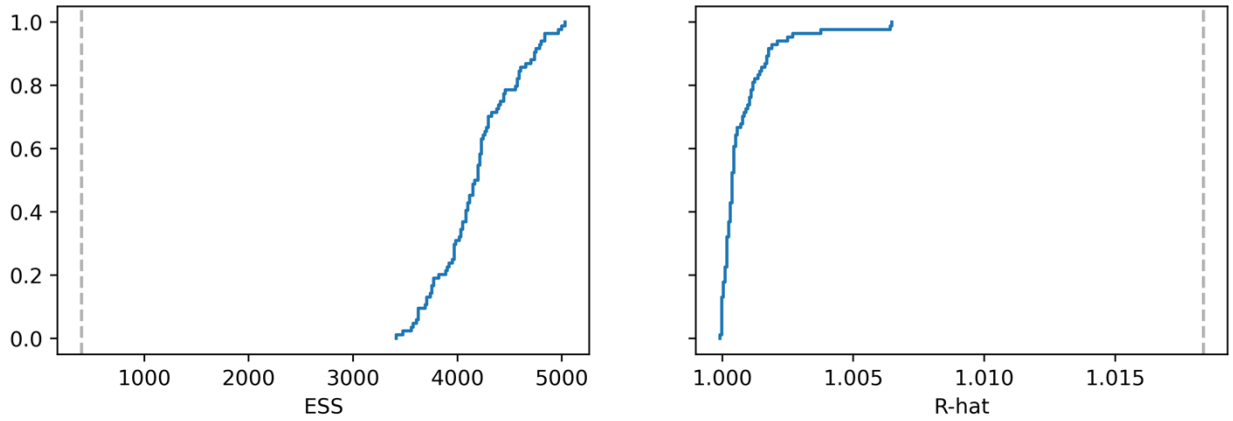

Figure S12. Convergence diagnostics of the Bayesian model. Left: effective sample size (ESS) distribution across parameters. Right: R-hat statistic.

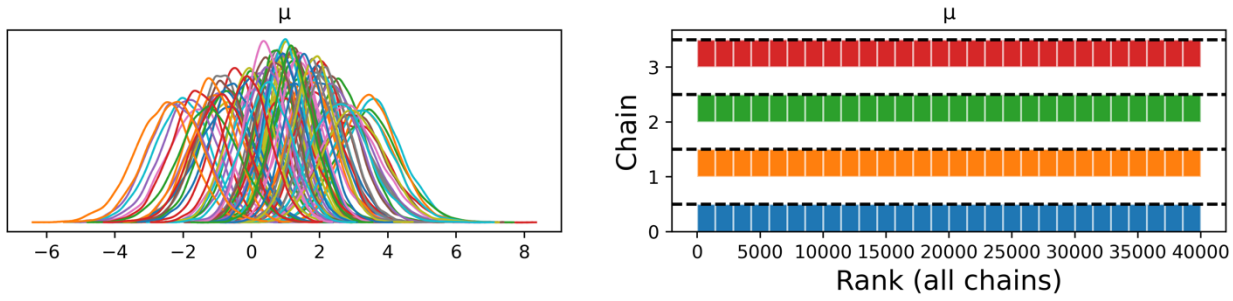

Figure S13. Posterior distribution diagnostics for parameter μ. Left: marginal posterior densities across Markov Chain Monte Carlo (MCMC) chains. Right: rank plots demonstrating mixing across chains.

Given a new observation x^*, the predictive probability under BART is obtained by using posterior samples drawn from the PGBART sampler. Each posterior draw $k = 1, \ldots, K$ provides a realization of the ensemble of regression trees $\{T_j^{(k)}, M_j^{(k)}\}_{j=1}^{J}$. For a given draw, the additive prediction is

$$f^{(k)}(x^*) = \sum_{j=1}^{J} g(x^*; T_j^{(k)}, M_j^{(k)}),$$

which is then transformed through the logistic link function $\phi(z)$ to yield the predictive probability

$$\pi^{(k)}(x^*) = \phi\{f^{(k)}(x^*)\}.$$

The posterior predictive distribution of the probability is represented by the collection $\{\pi^{(k)}(x^*)\}_{k=1}^{K}$, and summary statistics such as the posterior mean or credible interval can be computed from these draws. **Figure S14** presents simple schematic illustraction of the Random Forest (RF) and Bayesian Additive Regression Trees (BART).

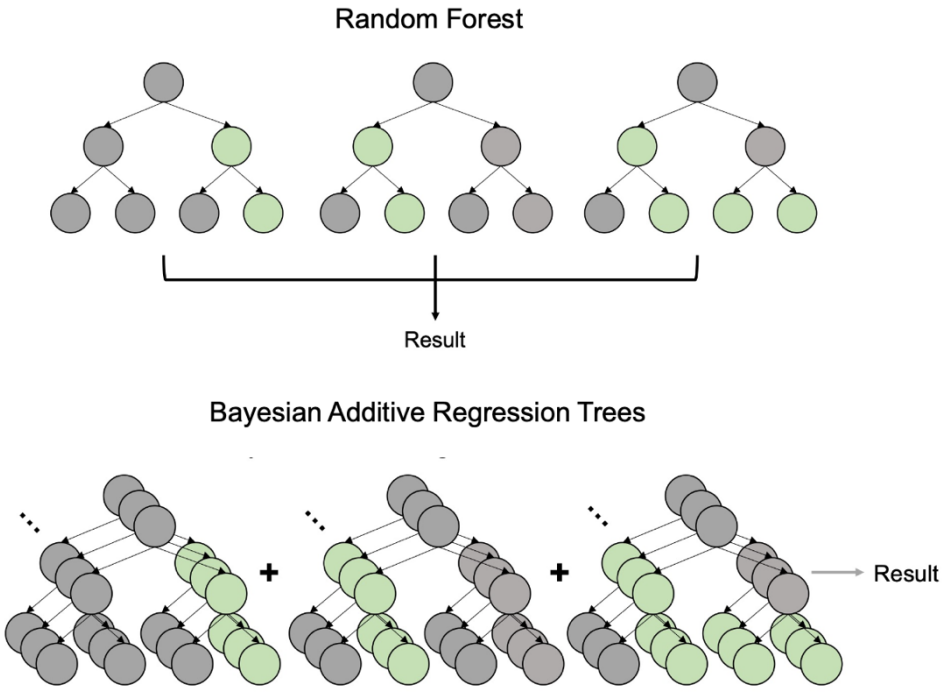

Figure S14. Schematic illustration of the Random Forest (RF) and Bayesian Additive Regression Trees (BART) models. In RF classifier, the final prediction is obtained by majority voting among the outcomes of multiple trees. In contrast, BART models the latent function as an additive ensemble of decision trees and infers a posterior distribution over their combined outputs, yielding a continuous probabilistic estimate for binary classification.

8. Model Interpretation and Active Learning Evaluation

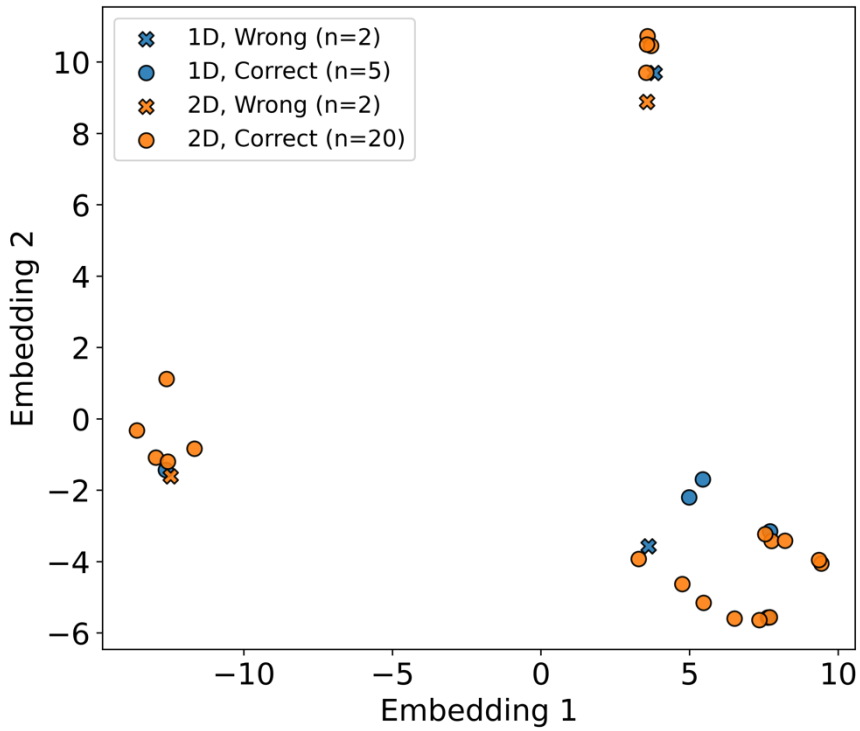

Figure S15. PaCMAP embedding of the test samples (compare to **Fig. 3** containing all samples). Colors indicate ground-truth structure (blue = 1D, orange = 2D), and markers denote prediction outcomes (o = correct, × = wrong). For example an orange x indicates the ground truth structure is 2D but prediction from BART was incorrect (1D).

To qualitatively assess model behavior on the test set, we projected all test samples into a 2D PaCMAP embedding (**Fig. S15**), coloring points by their ground-truth dimensionality and marking prediction correctness.

Feature importance quantifies the relative contribution of each input feature to the model's predictive performance[9]. To identify the molecular features that most strongly influence the dimensionality outcome, feature importance was examined using the trained models: built-in

feature importance from BART (**Fig. S16**), permutation importance from BART and RF (**Fig. 5 and Fig. S17**), and SHAP summary from RF (**Fig. S18**).

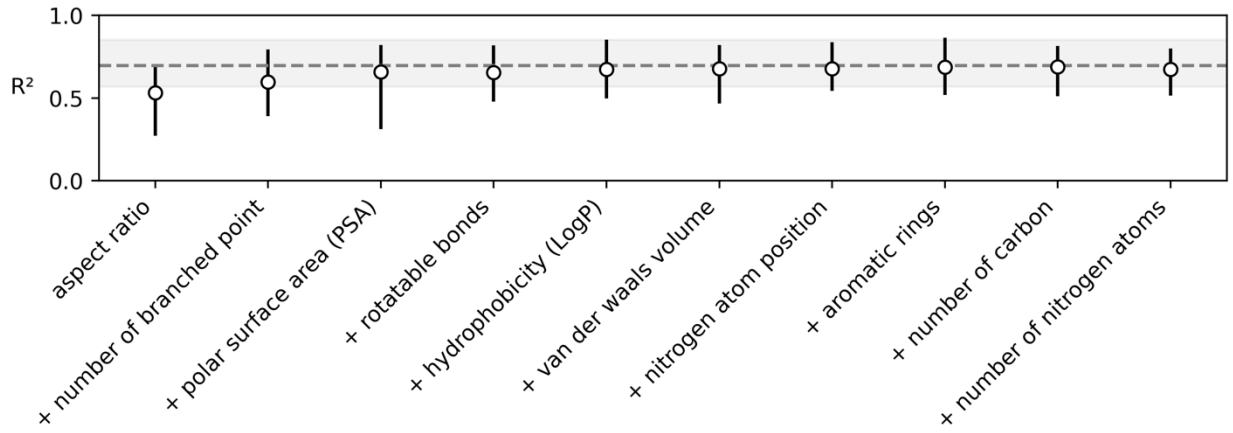

Figure S16. Incremental R^2 contributions of molecular descriptors. Each point shows the posterior mean and 94% highest density interval (HDI) of R^2 as descriptors are sequentially added to the model.

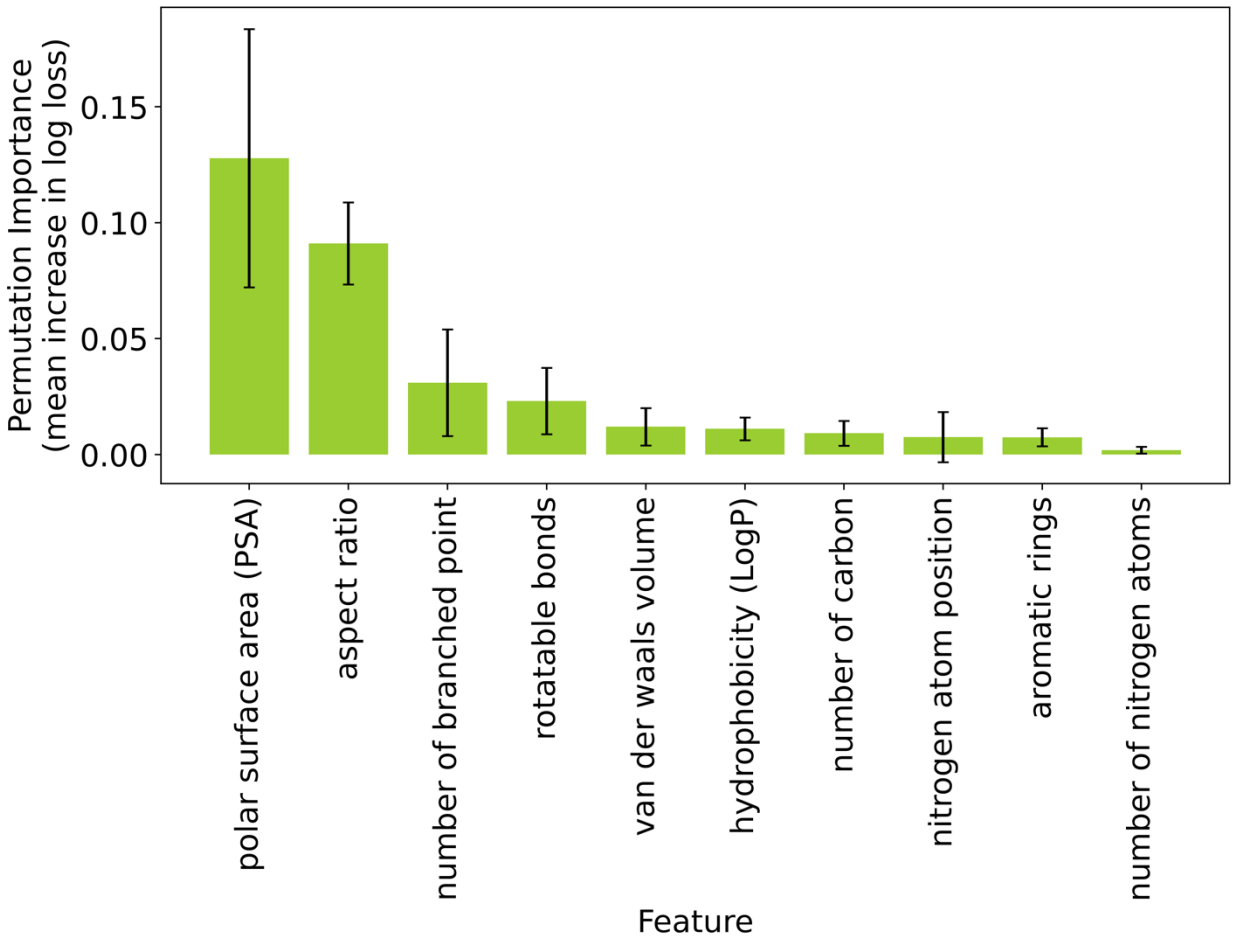

Figure S17. Permutation feature importance of the random forest classifier. Mean increase in log loss after feature permutation is shown for each molecular descriptor, with error bars representing ±1 standard deviation across cross-validation folds.

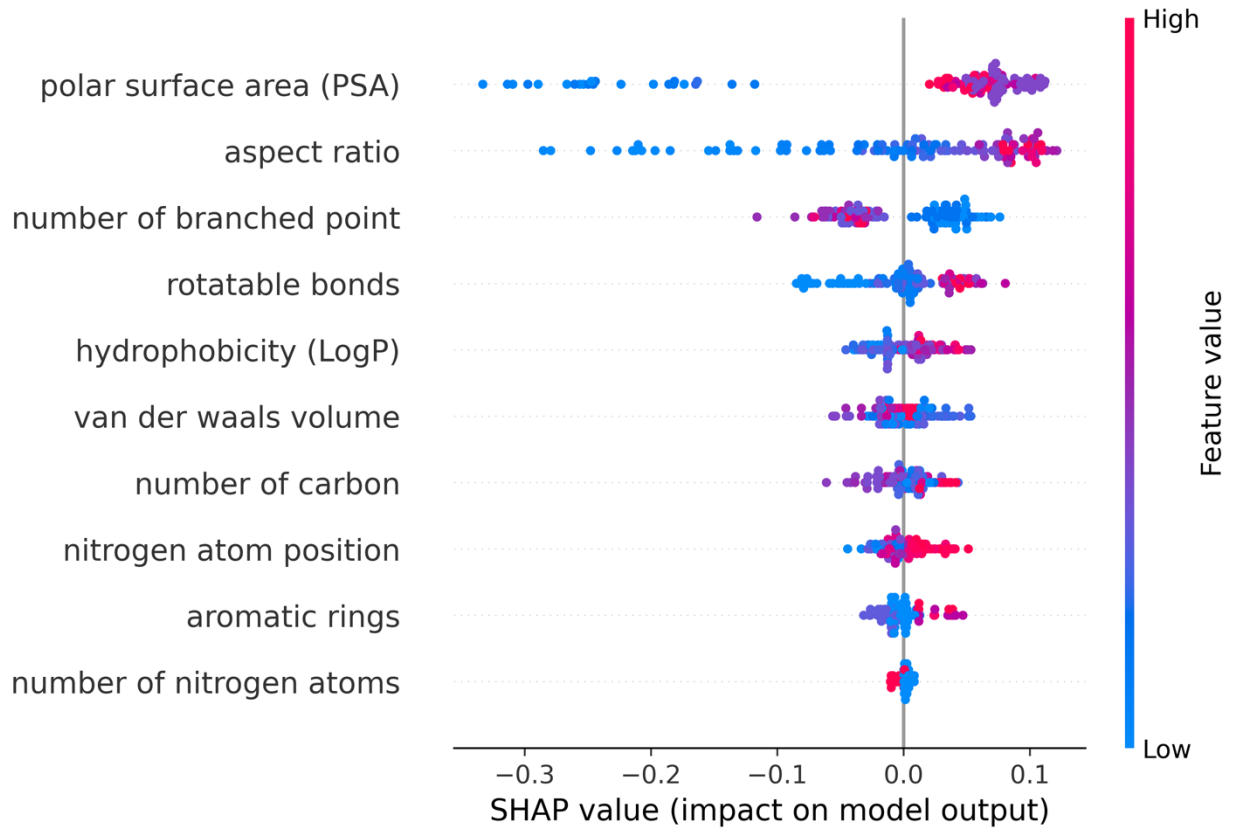

Figure S18. SHAP summary plot for the random forest classifier. Each dot represents the SHAP value of a single sample for a given molecular descriptor. Colors indicate the relative feature value (blue = low, red = high), showing both the magnitude and direction of feature contributions to model output.

To assess how active learning improved or worsened the model, we examined the changes in mutual information (ΔMI) and log loss (ΔLL). Mutual information measures how much knowledge about the target variable is gained from the model predictions, while log loss evaluates how well the predicted probabilities align with the true outcomes. Together, these metrics capture both the informativeness of newly acquired data and the overall calibration of the Bayesian model

during each active learning iteration[10, 11] (**Fig. S19**). After retraining, among the 29 test samples, 14 samples showed improved performance while 3 samples worsened.

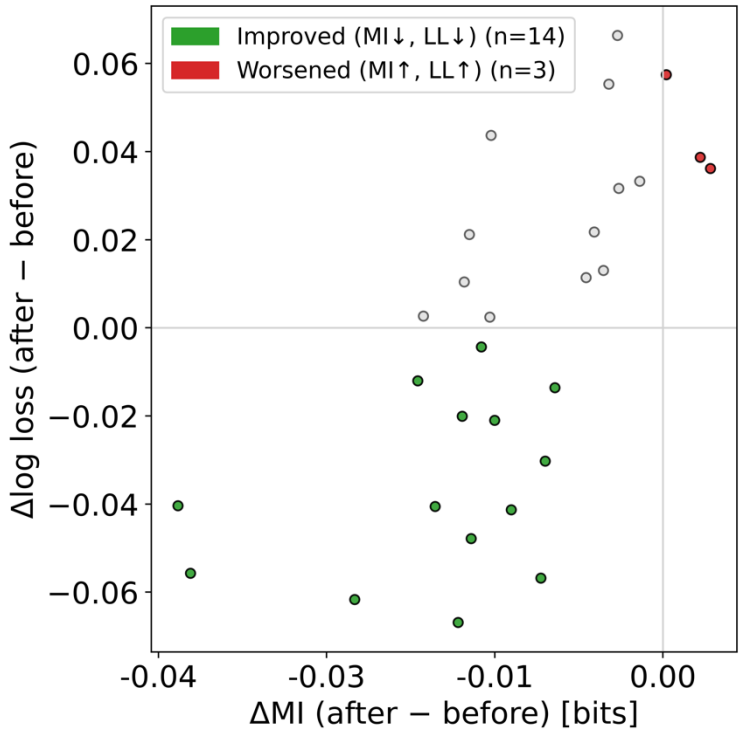

Figure S19. Active learning results plotted in quadrants of changes in mutual information (ΔMI) and changes in log loss (ΔLL). Green points indicate improved samples (ΔMI < 0, ΔLL < 0), and red points indicate worsened samples (ΔMI > 0, ΔLL > 0).

References

1. HybriD³ materials database, Duke University, https://materials.hybrid3.duke.edu/
2. D. M. Lowe, P. T. Corbett, P. Murray-Rust and R. C. Glen, Journal of Chemical Information and Modeling, 2011, 51, 739-753
3. SMILES to IUPAC, Leskoff, https://www.leskoff.com/s01814-0

4. Y. Wang, H. Huang, C. Rudin and Y. Shaposhnik, Journal of Machine Learning Research, 2021, 22, 1-73
5. RDKit, http://www.rdkit.org/
6. L. Breiman, Machine Learning, 2001, 45, 5-32
7. H. A. Chipman, E. I. George and R. E. McCulloch, 2010, arXiv:0806.3286
8. M. Quiroga, P. G. Garay, J. M. Alonso, J. M. Loyola and O. A. Martin, 2022, arXiv:2206.03619
9. F. K. Ewald, L. Bothmann, M. N. Wright, B. Bischl, G. Casalicchio and G. König, 2024, arXiv:2404.12862
10. C. E. Shannon, A Mathematical Theory of Communication, 1948, 27, 3, 379-423
11. I. J. Good, Journal of the Royal Statistical Society: Series B (Methodological), 2018, 14, 107-114