Supplementary Information: Topological Feature Engineering for Machine Learning Based Halide Perovskite Materials Design

D Vijay Anand,[1] Qiang Xu,[2] JunJie Wee,[1] Kelin Xia,[1,*] and Tze Chien Sum[2,*]

[1]Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371

[2] Division of Physics & Applied Physics, School of Physical and Mathematical Sciences Nanyang Technological University, Singapore 637371

[*]To whom correspondence should be addressed; E-mail: Xiakelin@ntu.edu.sg; Tzechien@ntu.edu.sg.

Supplementary Methods

OIHP-based PFs

Topological representations

There exist strong correlations between molecular structures and their functions. Efficient structural representation holds the key to enhancing the performance of QSAR/QSPR and machine learning (ML) models. The traditional molecular topological representations are graphs and simplicial complexes. Graphs (or networks) characterize pair-wise interactions within and/or between structures. Simplicial complexes are generalization of graphs into higher-dimensional counterparts by the incorporation of "many-body interactions" as simplexes.

Graph The most widely used topological models are graphs. For graph-based molecular models, atoms and bonds are usually represented as vertices and edges respectively. Mathematically, a graphical representation can be denoted as $G(V, E)$, where $V = \{v_i; i = 1, 2, ..., N_v\}$

are vertex set with $N_v = |V|$ the total number, and $E = \{e_i = (v_{i_1}, v_{i_2}); 1 \leq i_1 < i_2 \leq N_v\}$ denotes the edge set. Molecular structural, dynamic and functional properties can be modeled and analyzed through various graph measurements, such as node degree, shortest path, clique, cluster coefficient, closeness, (various types of) centrality, betweenness, Cheeger constant, modularity, graph Laplacian, graph spectral, Erdős number, percolation, etc. Establishing proper structure-function relationships is key to all these models.

Simplical complex A simplicial complex can be viewed as a generalization of a graph into its higher-dimensional counterpart. Simplexes are the building blocks for a simplicial complex. Each simplex is a finite set of vertices, and can be viewed geometrically as, a point (0-simplex), an edge (1-simplex), a triangle (2-simplex), a tetrahedron (3-simplex), and their k-dimensional counterpart (k-simplex).

A k-th chain group C_k is an Abelian group generated by oriented k-simplexes $\sigma^k = [v_0, v_1, v_2, \cdots, v_k]$, which are simplexes together with an orientation, i.e., ordering of their vertex set. The boundary operator ∂_k ($C_k \to C_{k-1}$) for an oriented k-simplex σ^k can be denoted as,

$$\partial_k \sigma^k = \sum_{i=0}^{k} (-1)^i [v_0, v_1, v_2, \cdots, \hat{v_i}, \cdots, v_k]. \tag{1}$$

Here $[v_0, v_1, v_2, \cdots, \hat{v_i}, \cdots, v_k]$ is an oriented $(k-1)$-simplex, which is generated by the original set of vertices except v_i. The boundary operator maps a simplex to its faces and it guarantees that $\partial_{k-1}\partial_k = 0$. There are various kinds of simplical complexes, including Vietoris-Rips complex, Čech Complex, Alpha complex, Clique complex, Neighborhood complex, Dowker complex, etc. Here we consider Vietoris-Rips complex for OIHP structure representation. An example of simplicial complexes for $CsSnBr_3$ can be found in Supplementary Figure **1A**.

Filtration-based multiscale representation Multiscale representations are key to the characterization of molecular structures and interactions. In general, molecules have hierarchial struc-

tures from different scales, ranging from atomic, residual, secondary structural, chain scale, to molecular monomer, and molecular complex. Molecular interactions are also composed of various scales from strong covalent bonds to related week interactions, such as hydrogen bonds, Van der Waals forces and electro-static interactions. The filtration process is a key component of persistent models, including persistent homology/cohomology, persistent spectral, and persistent Ricci curvature. A series of nested molecular topological representations, including graphs, simplicial complexes and hypergraphs, are generated from different scales during the filtration process. Supplementary Figure **1A** illustrates the filtration process for perovskite $CsSnBr_3$. The filtration parameter (f) is chosen as the sphere diameter for each atom.

Persistent functions

Persistent Homology As the key component of TDA, persistent homology has three unique properties, including simplicial complex representation, topological invariant, and geometric measurement-based topology. Firstly, persistent homology is based on the simplicial complex representation, thus it describes higher dimensional topological information. Secondly, persistent homology studies the structures with topological invariant of Betti numbers, which characterizes intrinsic and global information of structures. Finally, using a filtration process, persistent homology incorporates geometrical measurements into topological invariants, and thus provides a balance between geometric complexity and topological simplification.

From the boundary operation, one can define the k-th cycle group Z_k and the k-th boundary group B_k as follows,

$$Z_k = \text{Ker } \partial_k = \{d \in C_k \mid \partial_k d = 0\}, \tag{2}$$

$$B_k = \text{Im } \partial_{k+1} = \{d \in C_k \mid \exists d' \in C_{k+1} : d = \partial_{k+1} d'\}. \tag{3}$$

With $\partial_{k-1}\partial_k = 0$, cycle and boundary groups satisfy $B_k \subseteq Z_k$, and their quotient group $H_k = Z_k/B_k$ is known as the k-th homology group. The rank of k-th homology group H_k is called

k-th Betti number and satisfies,

$$\beta_k = \text{rank } H_k = \text{rank } Z_k - \text{rank } B_k. \tag{4}$$

Geometrically, we can regard β_0 as the number of isolated components, β_1 the number of one-dimensional loops, circles, or tunnels and β_2 the number of two-dimensional voids or holes.

The nested sequence of complexes from the filtration process can be expressed as,

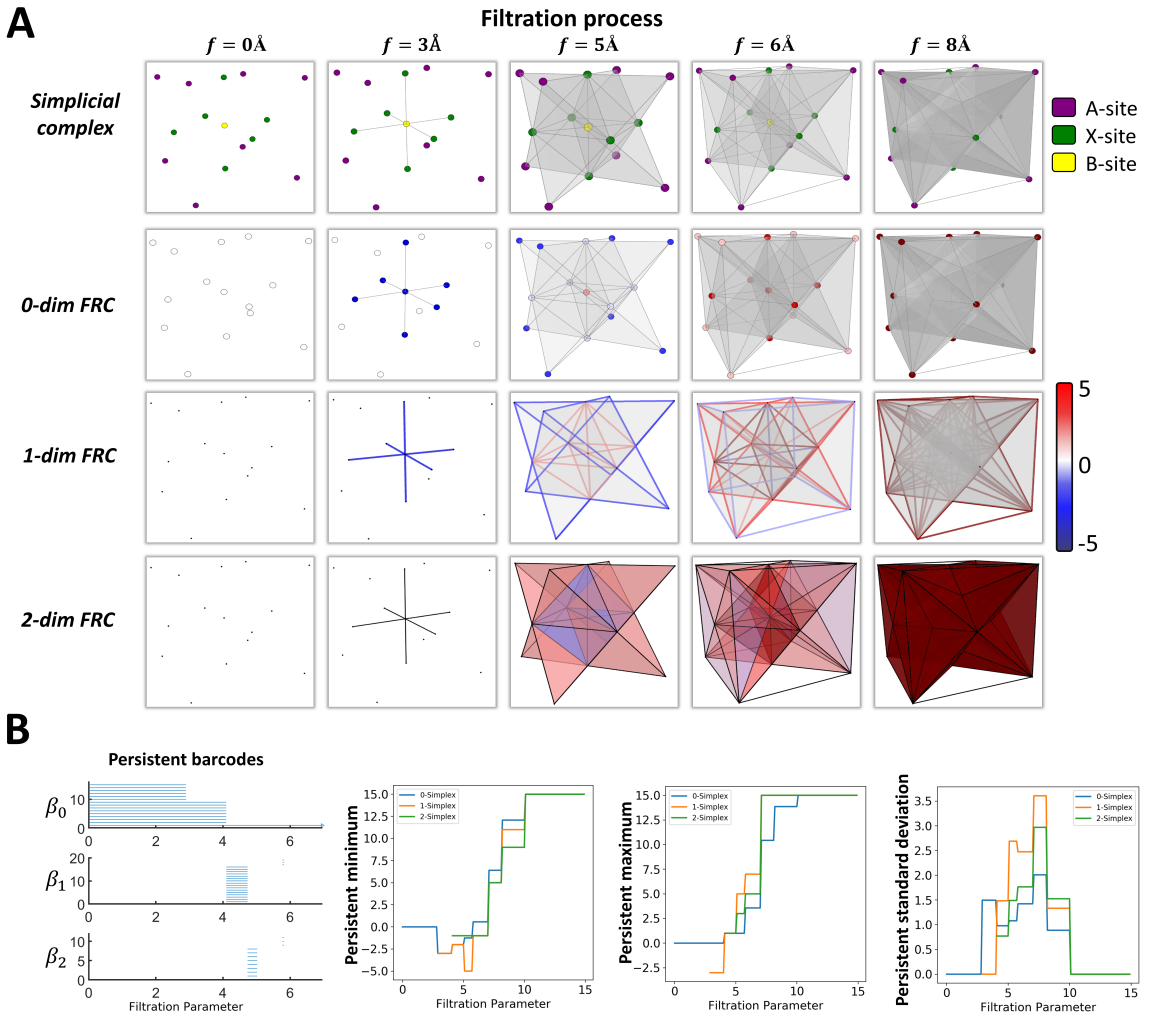
$$\varnothing = K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m = K. \tag{5}$$

And the p-persistent k-th homology group at filtration time i can be represented as

$$H_k^{i,p} = Z_k^i / (B_k^{i+p} \cap Z_k^i). \tag{6}$$

Essentially, the persistence of homology group can be viewed as a geometric measurement. Supplementary Figure **1B** illustrates persistent barcodes for $CsSnBr_3$ structure.

Persistent Ricci curvature As one of the fundamental concepts in differential geometry and theoretical physics, Ricci curvature describes local geometric quantities. Two discrete Ricci curvature forms, i.e., Ollivier Ricci curvature and Forman Ricci curvature, have been developed recently. Among them, ORC is defined as the Wasserstein distance between two associated probability measurements on metric spaces. It captures clustering and coherence properties of local structures in networks. In contrast, FRC is defined as a combinatorial properties of upper-adjacent, lower-adjacent and parallel simplexes on CW complexes. This combinatorial curvature can be directly derived from the combinatorial Bochner-Weitzenböck decomposition. It characterizes geodesics dispersal property and algebraic topological information within networks. Even though the two discrete forms can have totally different values, sometimes even signs, for network substructures, they are found to be highly correlated in various complex networks. In general, ORC can be better suited for analyzing network probabilistic properties, while Forman-Ricci curvature is better for network combinatorial property analysis.



Supplementary Figure 1: Persistent function (PF)-based material representations. **A** Illustration of the filtration process and its corresponding Forman Ricci curvatures (FRC) for $CsSnBr_3$ cubic perovskite. A sequence of Viertoris-Rips complexes at 5 different filtration size (f =0, 3, 5, 6, 8Å) are illustrated together with their 0-dimensional (0-dim) FRC, 1-dimensional (1-dim) FRC, and 2-dimensional (2-dim) FRC. **B** Illustration of PH-based persistent barcodes and PFC-based persistent attributes, including persistent minimum, persistent maximum, and persistent standard deviation, for the filtration process in **A**.

Here we focus on the Forman Ricci curvature. Let σ^k and $\bar{\sigma}^k$ be k-dimensional simplexes of a simplicial complex K, $\alpha^{k+1} > \sigma^k$ denotes a $(k+1)$-simplex α^{k+1} that has σ^k as a face and $\gamma^{k-1} < \sigma^k$ denotes $(k-1)$-simplex γ^{k-1} as a face of σ^k. Two simplexes σ^k and $\bar{\sigma}^k$ are upper adjacent, if they share a $(k+1)$-simplex α^{k+1}, that is, there exists a α^{k+1} in K such that $\alpha^{k+1} > \sigma$ and $\alpha^{k+1} > \bar{\sigma}^k$. Two simplexes σ^k and $\bar{\sigma}^k$ are lower adjacent, if σ^k and $\bar{\sigma}^k$ share a $(k-1)$-simplex γ^{k-1}, that is, there exists a γ^{k-1} in K such that $\gamma^{k-1} < \sigma^k$ and $\gamma^{k-1} < \bar{\sigma}^k$. More importantly, if σ^k and $\bar{\sigma}^k$ are either upper adjacent or lower adjacent but not both, they are parallel.

Let $k > 0$ and K be a weighted simplicial complex with w the positive weight defined on simplexes of K. The Forman Ricci Curvature for σ^k is defined as [13],

$$\mathcal{F}_k(\sigma^k) = w_{\sigma^k} \left\{ \left[\sum_{\alpha^{k+1} > \sigma^k} \frac{w_{\alpha^{k+1}}}{w_{\sigma^k}} + \sum_{\gamma^{k-1} < \sigma^k} \frac{w_{\gamma^{k-1}}}{w_{\sigma^k}} \right] - \sum_{\bar{\sigma}^k \neq \sigma^k} \left| \sum_{\substack{\alpha^{k+1} > \sigma^k \\ \alpha^{k+1} > \bar{\sigma}^k}} \frac{\sqrt{w_{\sigma^k} w_{\bar{\sigma}^k}}}{w_{\alpha^{k+1}}} - \sum_{\substack{\gamma^{k-1} < \sigma^k \\ \gamma^{k-1} < \bar{\sigma}^k}} \frac{w_{\gamma^{k-1}}}{\sqrt{w_{\sigma^k} w_{\bar{\sigma}^k}}} \right| \right\}. \tag{7}$$

In particular, if the simplicial complex is unweighted, FRC in Eq. (7) can be simplified as,

$$\mathcal{F}_k^{\sharp}(\sigma^k) = \sharp\{\alpha^{k+1} > \sigma^k\} + \sharp\{\gamma^{k-1} < \sigma^k\} - \sharp\{\text{parallel neighbours of } \sigma^k\}. \tag{8}$$

The term $\sharp\{\alpha^{k+1} > \sigma^k\} = \sharp\{\text{Upper adjacent simplex of } \sigma^k\}$ is the number of upper adjacent simplexes, also known as the upper degree of σ^k, denoted as $d(\sigma^k)$. The term $\sharp\{\gamma^{k-1} < \sigma^k\} = \sharp\{\text{Lower adjacent simplex of } \sigma^k\}$ is the number of lower adjacent simplexes, also known as the lower degree of σ^k.

For graphs or networks, there exist only vertex (0-simplex) and edge (1-simplex) FRCs [12, 15, 4]. For edge FRCs, Eq.(7) is reduced to as follows,

$$\mathcal{F}_1(\sigma^1) = w_{\sigma^1} \left\{ \left[\sum_{\gamma^0 < \sigma^1} \frac{w_{\gamma^0}}{w_{\sigma^1}} \right] - \sum_{\bar{\sigma}^1 \neq \sigma^1} \sum_{\substack{\gamma^0 < \sigma^1 \\ \gamma^0 < \bar{\sigma}^1}} \frac{w_{\gamma^0}}{\sqrt{w_{\sigma^1} w_{\bar{\sigma}^1}}} \right\}. \tag{9}$$

The vertex FRC is usually defined as the average edge FRCs as follows[12, 15, 14],

$$\mathcal{F}_0(\sigma^0) = \frac{1}{D(\sigma^0)} \sum_{\alpha^1 > \sigma^0} \mathcal{F}_1(\alpha^1). \tag{10}$$

Here $D(\sigma^0)$ is the (upper) degree of vertex σ^0, which is the number of edges that are connected to σ^0. Note that for an isolated vertex with no neighbors, we define its FRC as 0. Supplementary Figure **1A** illustrates FRCs at different dimensions for the OIHP structure.

Persistent Forman Ricci curvature is defined on a filtration process. More specifically, based on the simplicial complex sequence $\{K^i\}$, a series of sets of Ricci curvatures can be obtained. The persistence and variation of Ricci curvatures on this nested sequence of simplicial complexes are defined as PFC.

PF-GBT for OIHP data analysis

OIHP data preprocessing

The dataset consists of 1,346 OIHP structures and bandgap values calculated from Density Functional Theory (DFT) models [3]. The unit cell along with periodic boundary conditions or lattice vectors is provided. In our PF models, super cells are generated from unit cell for a better comparison. Three types of super cell structures are considered with unit size n =3, 4, and 5. Note that super cell is of size $n \times n \times n$.

OIHP molecular dynamics simulation In our MD simulations, we consider the all-electron-like projector augmented wave (PAW) method[5] and the Perdew-Burke-Ernserhof (PBE) exchange correlation potential [6] as implemented in the VASP code [7]. The semicore of Pb atoms (the orbital of 5d electrons) is treated as valence electrons, that is, 14 valence electrons for Pb ($5d^{10}6s^26p^2$) atom. All initial crystal structures for the 9 types of OIHPs are fully optimized. The structures are illustrated in Supplementary Figure **4**. The optimized lattice constants are list in Supplementary Table **10**. In the structure optimizing and self consistent calculation,

the cut-off energy for the plane wave expansion of the wave functions is 500 eV, and 300 eV for MD calculations. The Hellman-Feynman forces are less than of 0.01 eV/Å. The method of Tkatchenko and Scheffler is used to calculate the VDW interactions. For the self consistent calculations, the spin-orbital coupling is included. In MD calculations, we employed $2 \times 2 \times 1$ supercells for orthorhombic and tetragonal phases. For the cubic phase, we rotate the unit cell by 45^o in x-y plane, then build a $2 \times 2 \times 1$ supercell for keeping the same total number of atoms with orthorhombic and tetragonal phases. The Gamma $k-$points for Brillouin zone integration [7] was used in all calculations. Based on the optimized structures at 0 K, the supercells were heated up to various temperatures list in Supplementary Table**10**, which depends on the crystal phases. A $20ps$ MD simulation was carried out in the canonical ensemble with a $2fs$ time step. The MD structures are sampled every $20fs$. Thus, we obtain 1000 structures for each OIHP structure.

OIHP molecular features

Traditional molecular features used for OIHPs A total of 32 traditional pervoskite descriptors [1] have been used for machine learning models in OIHP data analysis. These TPDs are as follows,

- Packing Factor (P_f)
- Tolerance Factor (T_f)
- Octahedral Factor (O_f)
- Sum of atomic orbital
- Sum of atomic group number
- Sum of atomic mass

- Sum of atomic number
- Atomic radii for B- and X-site atoms
- Crystal radii for B- and X-site atoms
- Sum of the s and p orbital radii for B- and X-site atoms
- Martynov-Batsanov electronegativity scales (χ) for B- and X-site atoms
- Covalent radii for B- and X-site atoms
- Atomic number for A-, B- and X-site atoms
- Relative atomic mass for A-, B-, and X-site atoms
- Group number for the A-, B- and X-site atoms
- Atomic orbital for the A-, B- and X-site atoms
- Ion radii for the A-, B- and X-site atoms

Other than the above molecular features, other molecular features have also been considered, such as,

- Total number of ionic charge for B-site atoms
- Electron affinity for B-site atoms
- Ionic polariability for A- and B-site atoms
- Highest occupied molecular orbital (HOMO) for A-site
- Lowest unoccupied molecular orbital (LUMO) for A-site

All these features are based on structural and physical properties.

PH-based molecular features To characterize the detailed interactions within OIHP structures at molecular level, we consider the element-specific molecular representations [8, 2]. In our models, OIHP super cell structures are decomposed into a series of atom-sets, and PH analysis is carried out on these individual sets. Each OIHP super cell contains organic compound at the A-site composed of C, N, H, and O, cation atom at B-site including Ge, Pb, and Sn, and halide anion at X-site including Br, Cl, F, and I. A total number of 30 atom-sets can be generated from the different atom combinations, including one-element sets, [A_C, A_H, A_N, A_O, B, X], two-element combinations, [BX, A_CB, A_HB, A_NB, A_OB, A_CX, A_HX, A_NX, A_OX], three-element combinations [A_CBX, A_HBX, A_NBX, A_OBX], four-element combinations [A_{CH}BX, A_{CN}BX, A_{CO}BX, A_{HN}BX, A_{HO}BX, A_{NO}BX], five-element combinations [A_{CNH}BX, A_{COH}BXC, A_{ONH}BX, A_{CNO}BX] and finally all-element combination [A_{CNOH}BX].

The topological information for the 30 atom-sets from OIHP super cell structures is calculated by using GUDHI [9]. The filtration value of $f = 15$Å is set for individual elements [B, X, A_C, A_H, A_N, A_O] and is reduced to $f = 10$Å for remaining element combinations. Since the persistent barcode or persistent diagram results can be directly used as input for machine learning models, Betti curves are generated by counting the bars at each filtration, that is the Betti number at each filtration value. Computationally, the filtration region is divided into 100 bins and Betti numbers are obtained from each bin to form the molecular descriptors. The size of PH-based molecular fingerprint is 6000=30 (atom-combinations)$\times$ 2 (β_0 and β_1) $\times$ 100 (bins). To reduce the size of the feature matrix the columns which have completely zeros (due to β_1) or ones (due to β_0) are removed after assembling the feature matrix for the entire dataset. The final feature size of PH-based fingerprint for each OIHP structure reduces to 4319. Computationally, it has been found a grid size of 0.1Å to 0.25Å is good enough for persistent models to characterize the geometric and topological information embedded in the molecular structures and interactions [2, 11]. In our PH-based models, we have reduced the number of bins from

100 to 50, that is to increase the grid size from 0.15Å (0.1Å) to 0.30Å (0.2 Å). The results for PCC change from 0.9689 to 0.9633 and R^2 changes from 0.9372 to 0.9245. This is consistently with our previous findings and demonstrates that our models are relatively robust.

PFC-based molecular features In our PFC models, we also consider the 30 atom combinations and similar filtration values. At each filtration value, a 2-dimensional simplicial complex is constructed, the corresponding 0-, 1-, and 2-dimensional Forman Ricci curvatures can be calculated systematically. Since 0-, 1-, and 2-dimensional FRCs are based on vertices, edges and triangles respectively, the total number of FRCs are dramatically different for different dimensions and filtration values. To ensure a fixed-sized molecular fingerprint, 10 FRC attributes are considered. More specifically, for a set of FRCs $\{c_1, c_2, \cdots, c_N\}$, the 10 FRC descriptors are as follows [10, 11],

- $\min\{c_1, c_2, \cdots, c_N\}$
- $\max\{c_1, c_2, \cdots, c_N\}$
- $\bar{c} = \frac{1}{N}\sum_{i=1}^{N} c_j$
- $\sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(c_j - \bar{c})^2}$
- $\sum_{j=1}^{N} c_j \mathbb{1}_{\{c_j>0\}}$ ($\mathbb{1}_{c_j>0}$ is the indicator function.)
- $\sum_{j=1}^{N} |c_j - \bar{c}|$
- $\sum_{j=1}^{N} c_j^2$
- $\sum_{j=1}^{N} c_j^2 \mathbb{1}_{\{c_j>0\}}$
- $\sum_{j=1}^{N} |c_j - \bar{c}|^3$

- $\log\left(N\sum_{j=1,c_j>0}^{N^+}\frac{1}{c_j}+1\right)$ (N^+ is the total number of positive elements.)

The persistence and variation of the these PFC attributes during the filtration process is called persistent attributes. PFC-based molecular fingerprint is obtained from the discretization of these 10 persistent attributes. The PFC features are generated for sixteen filtration values ($f = 0$Å to 15Å) for individual elements [B, X, A_C, A_H, A_N, A_O] and eleven filtration values ($f = 0$Å to 10Å) is used for rest of the element combinations. Therefore we have 2880 = 16(filtration values) $\times$ 3(dimension) $\times$ 10(persistent attributes) for individual 6 elements and 7920 = 11(filtration values) $\times$ 3(dimension) $\times$ 10(persistent attributes) for the remaining element combinations. This leads to a combined long feature vector of size 10800=2880+7920 PFC-based descriptor for each OIHP structure. Supplementary Figure **1B** illustrates PFC-based persistent attributes, including persistent minimum, persistent maximum, and persistent standard deviation for perovskite $CsSnBr_3$.

PF-GBTs for OIHP bandgap energy prediction

Since our PF-based molecular fingerprint is of relatively large sizes, gradient boosting tree model is considered to circumvent the overfitting problem. The detailed parameter setting for GBT model is listed in Supplementary Table **1**.

In our Leave One Group Out cross validation, the OIHP dataset is divided into different groups based on the type of site atoms. For A-site organic cation, a total of 16-types of cations are considered namely., (Ammonium, Methylammonium, Dimethylammonium, Trimethylammonium, Tetramethylammonium, Ethylammonium, Propylammonium, Isopropylammonium, Butylammonium, Hydroxylammonium, Formamidinium, Acetamidinium, Hydrazinium, Guanidinium, Azetidinium, Imidazolium). For B-site, there are 3 types of group-IV cations, including Germanium, Lead, and Tin. For X-site, a total of 4 types of halide anions are used, including Bromide, Chloride, Fluoride, and Iodide. Based on A-, B- and X-sites, we have three categories

of LOGO cross validation. For A-site, we choose one type of organic cation as test set, and use all the other OIHP structures from the rest 15 types of cations in the training set. Similar procedure is considered for B- and X-site. In our machine learning model, each cross validation is repeated for 10 times and only the median value is reported as the final results.

Supplementary Table 1: Setting parameters for our GBT model.

No. of estimators	Max depth	Minimum sample split	Learning rate
20000	8	2	0.001
Loss function	Max features	Subsample size	Repetition
least square	square root	0.7	10 times

Supplementary Results

Supplementary Table 2: Results for the 10-fold cross-validation for OIHP bandgaps from 1346 OIHP database. Five different error measurements, including Coefficient of determination (R^2), Pearson correlation coefficient (PCC), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE) are considered. The average results for traditional perovskite descriptor based gradient boosting regression (TPD-GBR) are adopted from reference 14. Persistent homology and persistent Ricci curvature based GBT models are calculated on super-cell with size 3*3*3, which is denoted as subscript a. We have also compared with SchNet[17] and MEGnet[16]. Note that SchNetPack, the updated version of SchNet, is used in performing the 10-fold CV. Both potential energies and dipole moments are used as output features in SchNetPack. Results of the two best models are in bold.

Error measurements	**TPD-GBT**	**PRC_a-GBT**	**SchNet**	**PH_a-GBT**	**MEGnet**
R^2	0.8270	0.9068	0.9274	**0.9372**	**0.9673**
PCC	——	0.9531	0.9656	**0.9689**	**0.9840**
MAE	0.3770	0.2620	0.2112	**0.2032**	**0.1396**
MSE	0.2010	0.1083	0.0792	**0.0729**	**0.0370**
RMSE	0.4483	0.3288	0.2800	**0.2697**	**0.1924**

Supplementary Table 3: Leave One Group Out (LOGO) cross validation test for A-site. Note that PH_a-GBT, PH_b-GBT, and PH_c-GBT are same model with features from super-cell of sizes 3*3*3, 4*4*4, and 5*5*5, respectively. Results of best models are in bold.

A-site (R^2)	TPD-GBT	FRC_a-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT
1	0.7829	0.9109	0.9358	0.9365	**0.9390**
2	0.7514	0.8147	0.8666	0.8799	**0.8870**
3	0.7314	0.8409	0.8729	0.8703	**0.8705**
4	0.9349	0.9332	**0.9543**	0.9535	0.9534
5	0.7972	0.9372	0.9683	0.9685	**0.9688**
6	0.7816	0.8030	**0.8531**	0.8491	0.8468
7	0.4463	0.7405	0.7166	0.7528	**0.7803**
8	0.6233	0.8890	0.9331	0.9331	**0.9353**
9	0.7842	0.8329	0.8520	0.8602	**0.8614**
10	0.8362	0.8268	0.8656	**0.8669**	0.8658
11	0.5872	0.8644	0.9147	0.9139	**0.9172**
12	0.8830	0.9362	0.9371	0.9383	**0.9386**
13	0.6048	0.8502	0.8726	0.8764	**0.8779**
14	0.8570	0.9061	0.9394	0.9464	**0.9471**
15	0.8825	0.8430	**0.9105**	0.9096	0.9073
16	0.8774	0.9024	0.9088	**0.9102**	0.9096
Average	0.7601	0.8645	0.8938	0.8979	**0.9004**
A-site (PCC)	TPD-GBT	FRC_a-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT
1	0.8976	0.9586	0.9678	0.9684	**0.9696**
2	0.8879	0.9227	0.9427	0.9520	**0.9544**
3	0.8653	0.9495	0.9579	0.9586	**0.9591**
4	0.9676	0.9761	0.9820	0.9822	**0.9826**
5	0.8943	0.9696	0.9845	0.9852	**0.9856**
6	0.8853	0.9045	**0.9306**	0.9289	0.9277
7	0.8198	0.8738	0.8731	0.8882	**0.8978**
8	0.9214	0.9674	0.9767	0.9758	**0.9767**
9	0.8912	0.9305	0.9333	0.9384	**0.9389**
10	0.9252	0.9479	0.9525	0.9545	**0.9557**
11	0.8631	0.9419	0.9599	0.9593	**0.9608**
12	0.9562	0.9775	0.9827	0.9835	**0.9835**
13	0.8525	0.9321	0.9509	0.9530	**0.9536**
14	0.9281	0.9533	0.9696	0.9732	**0.9735**
15	0.9442	0.9595	0.9908	0.9906	**0.9908**
16	0.9402	**0.9587**	0.9556	0.9566	0.9564
Average	0.9025	0.9452	0.9569	0.9593	**0.9604**

A-site (MAE)	TPD-GBT	FRC_a-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT
1	0.4078	0.2436	0.2051	0.2074	**0.2047**
2	0.3473	0.3529	0.2915	0.2811	**0.2731**
3	0.4237	0.3092	**0.2732**	0.2768	0.2780
4	0.2354	0.2133	**0.1790**	0.1792	0.1833
5	0.3554	0.2227	0.1454	**0.1441**	0.1446
6	0.3845	0.3538	**0.2932**	0.2953	0.2981
7	0.6347	0.4253	0.4101	0.3783	**0.3598**
8	0.5839	0.3298	0.2293	0.2289	**0.2287**
9	0.4122	0.3637	0.3371	0.3293	**0.3290**
10	0.3428	0.3863	0.3265	0.3229	**0.3225**
11	0.4805	0.2521	0.2014	0.2008	**0.1990**
12	0.3252	0.2316	0.2402	0.2365	**0.2360**
13	0.5240	0.3565	0.3046	**0.3006**	0.3018
14	0.3278	0.2508	0.2035	0.1911	**0.1909**
15	0.3153	0.2916	**0.2597**	0.2632	0.2691
16	0.3363	0.2751	0.2753	**0.2745**	0.2772
Average	0.4023	0.3036	0.2609	0.2569	**0.2560**

A-site (MSE)	TPD-GBT	FRC_a-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT
1	0.2232	0.0916	0.0660	0.0653	**0.0634**
2	0.2305	0.1718	0.1237	0.1113	**0.1069**
3	0.2572	0.1524	**0.1217**	0.1242	0.1245
4	0.0738	0.0758	**0.0519**	0.0527	0.0537
5	0.2277	0.0705	0.0356	0.0353	**0.0353**
6	0.2539	0.2290	**0.1708**	0.1754	0.1789
7	0.5573	0.2612	0.2852	0.2488	**0.2280**
8	0.5507	0.1622	0.0978	0.0978	**0.0966**
9	0.2620	0.2029	0.1708	0.1698	**0.1688**
10	0.2131	0.2253	0.1748	**0.1731**	0.1756
11	0.3220	0.1058	0.0665	0.0672	**0.0651**
12	0.1583	0.0863	0.0850	0.0834	**0.0834**
13	0.4852	0.1839	0.1564	0.1517	**0.1512**
14	0.1516	0.0995	0.0643	0.0568	**0.0560**
15	0.1374	0.1835	**0.1046**	0.1057	0.1106
16	0.1609	0.1281	0.1197	**0.1178**	0.1195
Average	0.2666	0.1518	0.1184	0.1148	**0.1136**

A-site (RMSE)	TPD-GBT	FRC_a-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT
1	0.4725	0.3027	0.2569	0.2555	**0.2518**
2	0.4801	0.4145	0.3517	0.3336	**0.3269**
3	0.5072	0.3904	**0.3489**	0.3525	0.3529
4	0.2716	0.2752	**0.2278**	0.2296	0.2317
5	0.4771	0.2655	0.1888	0.1880	**0.1879**
6	0.5039	0.4786	**0.4133**	0.4188	0.4230
7	0.7465	0.5111	0.5341	0.4988	**0.4775**
8	0.7421	0.4028	0.3127	0.3127	**0.3109**
9	0.5119	0.4504	0.4240	0.4121	**0.4108**
10	0.4616	0.4746	0.4181	**0.4160**	0.4191
11	0.5675	0.3253	0.2579	0.2592	**0.2551**
12	0.3978	0.2938	0.2916	0.2888	**0.2888**
13	0.6965	0.4288	0.3955	0.3895	**0.3888**
14	0.3894	0.3155	0.2536	0.2384	**0.2366**
15	0.3706	0.4284	**0.3234**	0.3250	0.3325
16	0.4012	0.3580	0.3460	**0.3432**	0.3457
Average	0.4998	0.3822	0.3340	0.3289	**0.3275**

Supplementary Table 4: Leave One Group Out (LOGO) cross validation test for B-site.

B-site (R^2)	TPD-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT	FRC_a-GBT
1	0.4297	0.8584	0.8577	**0.8594**	0.8503
2	0.5945	0.5902	0.5725	0.5657	**0.6760**
3	0.5672	0.7229	0.7237	**0.7242**	0.7141
Average	0.5305	0.7238	0.7180	0.7164	**0.7468**
B-site (PCC)	TPD-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT	FRC_a-GBT
1	0.8966	**0.9659**	0.9650	0.9647	0.9617
2	0.9007	**0.9701**	0.9699	0.9697	0.9625
3	0.8757	0.9673	**0.9679**	0.9678	0.9673
Average	0.891	**0.9678**	0.9676	0.9674	0.9638
B-site (MAE)	TPD-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT	FRC_a-GBT
1	0.7327	0.3214	0.3200	**0.3181**	0.3341
2	0.5592	0.6272	0.6408	0.6464	**0.5307**
3	0.5143	0.4508	0.4505	**0.4500**	0.4622
Average	0.6021	0.4665	0.4704	0.4715	**0.4424**
B-site (MSE)	TPD-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT	FRC_a-GBT
1	0.7262	0.1804	0.1810	**0.1787**	0.1903
2	0.4531	0.4572	0.4769	0.4844	**0.3614**
3	0.4224	0.2700	0.2693	**0.2688**	0.2786
Average	0.5339	0.3025	0.3091	0.3106	**0.2768**
B-site (RMSE)	TPD-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT	FRC_a-GBT
1	0.8522	0.4247	0.4254	**0.4228**	0.4362
2	0.6731	0.6762	0.6906	0.6960	**0.6012**
3	0.6499	0.5196	0.5189	**0.5185**	0.5278
Average	0.7251	0.5402	0.5450	0.5458	**0.5218**

Supplementary Table 5: Leave One Group Out (LOGO) cross validation test for X-site.

X-site (R^2)	TPD-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT	FRC_a-GBT
1	-1.037	0.5755	0.5920	**0.5936**	0.4666
2	-0.1962	0.0352	0.0384	**0.0342**	0.1632
3	-4.8948	-4.137	-4.0442	-4.0352	**-3.2516**
4	-4.2578	-0.8163	-0.9055	-0.9388	**-0.5746**
Average	-2.5965	-1.0857	-1.0798	-1.0866	**-0.7991**
X-site (PCC)	TPD-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT	FRC_a-GBT
1	0.2296	0.7703	0.7781	**0.7785**	0.7212
2	0.4225	0.7271	0.7208	0.7173	**0.7567**
3	**0.5322**	-0.098	-0.0967	-0.0877	0.1859
4	0.1912	0.7211	0.7128	0.7059	**0.8127**
Average	0.3439	0.5301	0.5288	0.5285	**0.6191**
X-site (MAE)	TPD-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT	FRC_a-GBT
1	0.5928	0.2716	0.2663	**0.2656**	0.3169
2	0.4807	0.4448	0.4444	0.4451	**0.4081**
3	1.2463	1.1220	1.1107	1.1097	**1.0134**
4	0.9478	0.5367	0.5503	0.5556	**0.5137**
Average	0.8169	0.5938	0.5929	0.5940	**0.5630**
X-site (MSE)	TPD-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT	FRC_a-GBT
1	0.5548	0.1157	0.1111	**0.1107**	0.1453
2	0.3302	0.2663	0.2654	0.2666	**0.2310**
3	1.7857	1.5561	1.528	1.5252	**1.2879**
4	1.1078	0.3827	0.4015	0.4085	**0.3318**
Average	0.9446	0.5802	0.5765	0.5778	**0.4990**
X-site (RMSE)	TPD-GBT	PH_a-GBT	PH_b-GBT	PH_c-GBT	FRC_a-GBT
1	0.7449	0.3393	0.3334	**0.3327**	0.3812
2	0.5746	0.5160	0.5152	0.5163	**0.4806**
3	1.3363	1.2474	1.2361	1.2350	**1.1348**
4	1.0525	0.6186	0.6336	0.6391	**0.5760**
Average	0.9271	0.6803	0.6796	0.6808	**0.6431**

Additional 10-fold cross-validation for OIHP properties The 10-fold cross-validation of the models for the four other types of OIHP properties, namely the refractive index n and three dielectric constants (electronic (ϵ_{elec}), ionic (ϵ_{ion}) and total ($\epsilon_{elec} + \epsilon_{ion}$)). Five different error measurements, including Coefficient of determination (R^2), Pearson correlation coefficient (PCC), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE) are considered. The average results for traditional perovskite descriptor based gradient boosting tree (TPD-GBT) are obtained by re-applying the models to train and predict the 4 properties. Persistent homology and persistent Ricci curvature based GBT models (PRC_a-GBT and PH_a-GBT) are calculated using a super-cell of size 3*3*3, which is denoted as subscript a. Overall, our models still showed a more robust predictive power as compared to traditional molecular descriptors when predicting these 4 properties.

The dielectric constants are calculated using density functional perturbation theory (DFPT)[3]. For ϵ_{elec}, corresponding to the high frequency dielectric constants (ϵ_∞), the calculations are accurate. However, ϵ_{ion} has more deviation based on the DFPT method. Thus, this also causes the total dielectric constant ($\epsilon_{elec} + \epsilon_{ion}$), i.e., static dielectric constant (ϵ_0), to be less accurate. It is possible to improve the accuracy of the calculations using a correction method based on many-body perturbation theory. However, it is too computationally costly to build a large data set of the dielectric constants. This is beyond the scope of the present work.

Supplementary Table 6: Results for the 10-fold cross-validation of the three models on the prediction of OIHP refractive index n.

n	TPD-GBT	PRC_a-GBT	PH_a-GBT
R^2	0.9283	0.9528	**0.9698**
PCC	0.9643	0.9771	**0.9853**
MAE	0.0314	0.0293	**0.0224**
MSE	0.0025	0.0016	**0.0011**
RMSE	0.0499	0.0403	**0.0322**

Supplementary Table 7: Results for the 10-fold cross-validation of the three models on the prediction of OIHP dielectric constant (electronic contribution ϵ_{elec}).

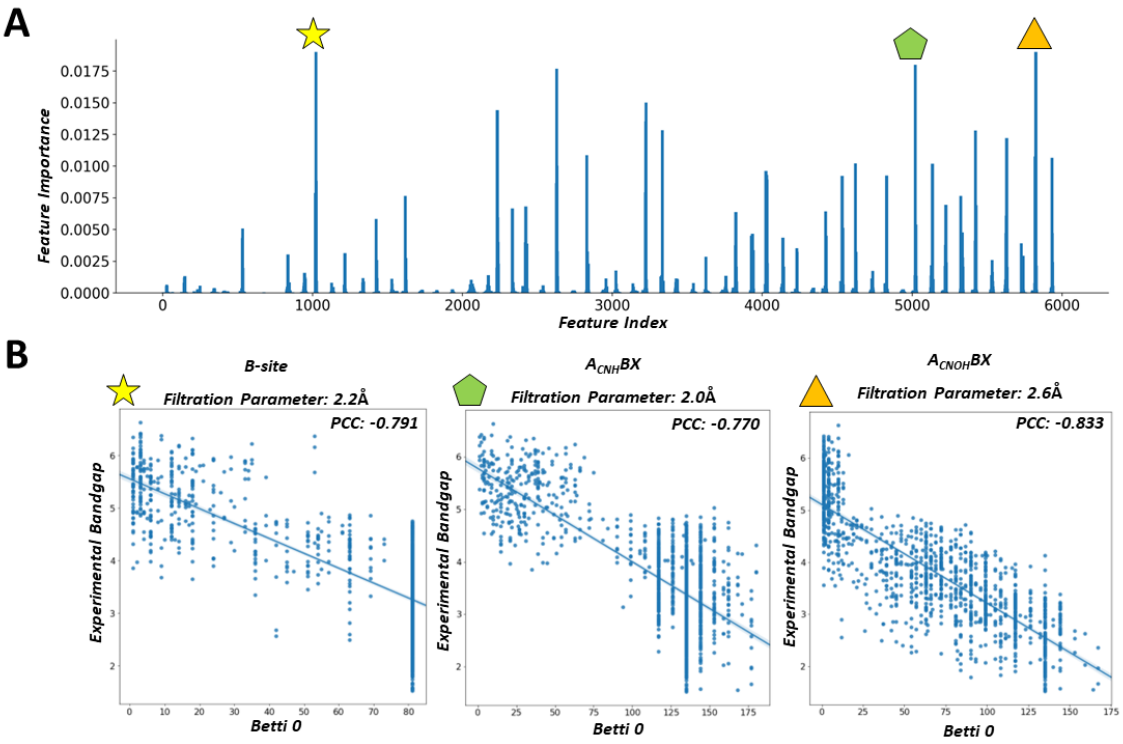
ϵ_{elec}	TPD-GBT	PRC_a-GBT	PH_a-GBT
R^2	0.9152	0.9503	**0.9656**
PCC	0.9583	0.9761	**0.9836**
MAE	0.1282	0.1158	**0.0908**
MSE	0.0464	0.0271	**0.0187**
RMSE	0.2120	0.1640	**0.1364**

Supplementary Table 8: Results for the 10-fold cross-validation of the three models on the prediction of OIHP dielectric constant (ionic contribution ϵ_{ion}).

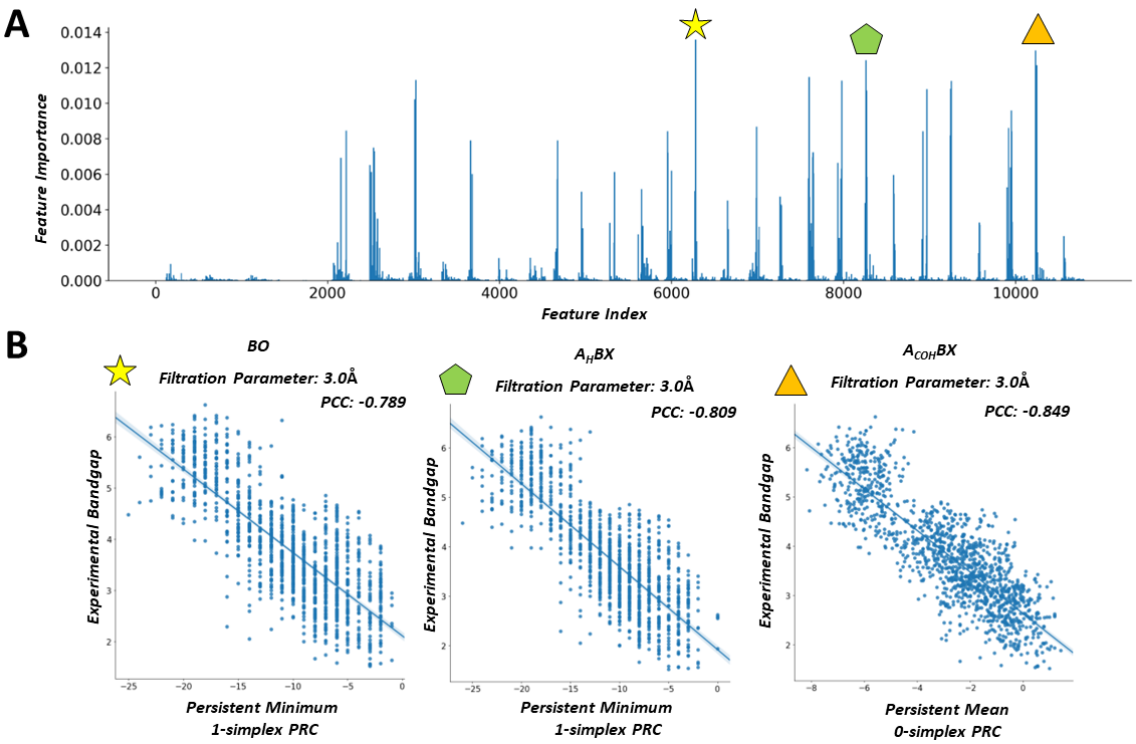
ϵ_{ion}	TPD-GBT	PRC_a-GBT	PH_a-GBT
R^2	0.0202	0.3243	**0.4179**
PCC	0.4960	0.5840	**0.6581**
MAE	4.1703	4.1446	**3.4170**
MSE	75.1413	56.3928	**48.3550**
RMSE	8.5999	7.2608	**6.7773**

Supplementary Table 9: Results for the 10-fold cross-validation of the three models on the prediction of OIHP total dielectric constant ($\epsilon_{elec} + \epsilon_{ion}$).

Total Dielectric Constant ($\epsilon_{elec} + \epsilon_{ion}$)	TPD-GBT	PRC_a-GBT	PH_a-GBT
R^2	0.0720	0.3360	**0.4087**
PCC	0.5269	0.6004	**0.6581**
MAE	4.2278	4.2355	**3.5482**
MSE	79.6327	57.4920	**50.6806**
RMSE	8.7952	7.4900	**7.0145**



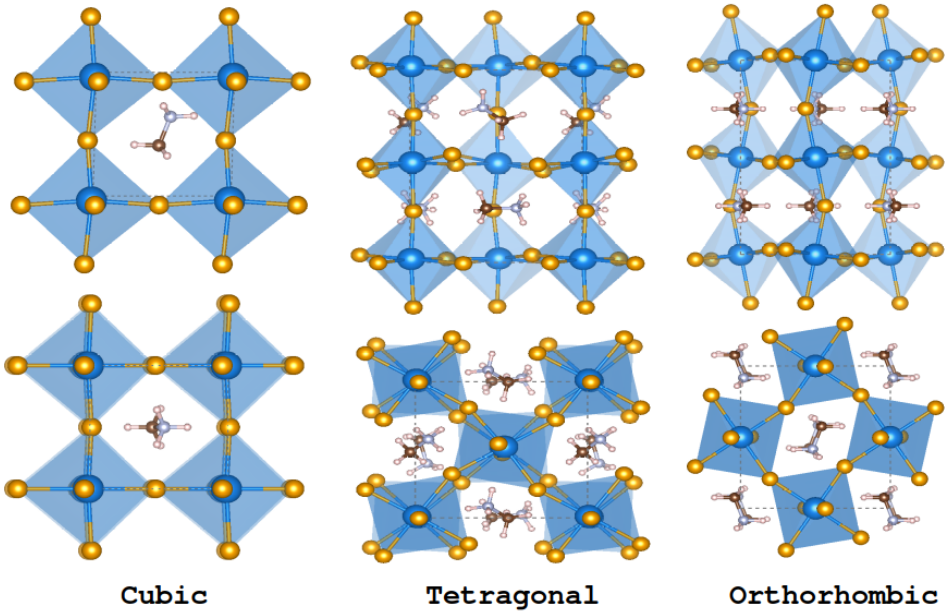
Supplementary Figure 2: The feature importance of PH-based molecular features and the correlation of the top three most important descriptors with the experimental bandgap. Note that the the feature index for the X-axis is arranged according to the atom combinations (from one-element to all-element) as the above discussions. All 6000 features (including all zero-value ones) in the PH-based molecular fingerprint are considered. (**A**) Feature importance of PH-based molecular features. (**B**) Illustration of the correlation between the top three most important PH-based features and bandgap values. Note that strong (inverse) correlation can be found between our descriptors and bandgap values.



Supplementary Figure 3: The feature importance of PRC-based molecular features and the correlation of the top three most important descriptors with the experimental bandgap. Note that the the feature index for the X-axis is arranged according to the atom combinations (from one-element to all-element) as the above discussions. All 10800 features in the PH-based molecular fingerprint are considered. (**A**) Feature importance of PRC-based molecular features. (**B**) Illustration of the correlation between the top three most important PRC-based features and bandgap values. Note that strong (inverse) correlation can be found between our descriptors and bandgap values.

Supplementary Data

Crystal structures of OIHP The chemical formula of hybrid halide perovskite is commonly expressed as ABX_3 (X=Cl, Br, and I). A site is occupied by cation or organic molecule with +1 valence, such as Cs^+ and $CH_3NH_3^+$. The metal cation with +2 valences (Pb^{2+}, Sn^{2+}, etc) is at the B site. X site is usually filled by the halide anions, i.e., Cl^-, Br^-, and I^-. The small cation B^{2+} is surrounded by six-nearest neighbor X^- to form an octahedron BX_6. The larger cation A^+ is coordinated with maximum twelve-nearest neighbor X^{2+}. For the crystal structures of $CH_3NH_3PbX_3$ ($MAPbX_3$, X=Cl, Br, and I), they have flexible feature because of their perovskite framework. In addition, the various origination of the double bridging halide ion (B-X-B) make them easily phase transitions upon application of temperature, pressure, etc. From the high to low temperature, their phases transfer from Cubic phase ($Pm\bar{3}m$) to Tetragonal phase ($I4mcm$) to Orthorhombic phase ($Pnma$), and the Orthorhombic phase is their stable phase. The critical temperature of phase transition and optimized lattice constants are listed in Supplementary Table **10**. In Fig. **4**, we show the three phases of $CH_3NH_3PbX_3$ ($MAPbX_3$, X=Cl, Br, and I). Obviously, there are distortion of the octahedron commonly known as Jan-Teller Effect. In addition, because an organic molecule occupies the A site, the vdW interactions and HB effects play an important role in the stability and deformation of hybrid halide perovskites.



Supplementary Figure 4: The crystal structures of $CH_3NH_3PbX_3$ ($MAPbX_3$, X=Cl, Br, and I). The side-view figures are shown in the top row, and the top-view figures are shown in the bottom row. The blue ball represents the cation at the B site, the yellow ball represents the halide anion at the X site. The light grey ball is Nitrogen, the brown ball is Carbon, and the light red ball is Hydrogen.

Supplementary Table 10: The parameters of the crystal structures of $CH_3NH_3PbX_3$ ($MAPbX_3$, X=Cl, Br, and I).

	phase	Symmetry	T(K)[†]	Optimized Lattice Constants (Å)
	α	$Pm\bar{3}m$	> 178.8	$a = 5.92, b = 5.74, c = 5.76$
$MAPbCl_3$	β	$I4/mcm$	$172.9 \sim 178.8$	$a = 8.13, b = 8.13, c = 11.84$
	δ	$Pnma$	< 172.9	$a = 8.54, b = 11.57, c = 7.66$
	α	$Pm\bar{3}m$	> 236.9	$a = 6.16, b = 6.01, c = 6.03$
$MAPbBr_3$	β	$I4/mcm$	$155.1 \sim 236.9$	$a = 8.53, b = 8.51, c = 12.05$
	δ	$Pnma$	< 144.5	$a = 8.84, b = 12.10, c = 8.07$
	α	$Pm\bar{3}m$	> 327.4	$a = 6.53, b = 6.41, c = 6.44$
$MAPbI_3$	β	$I4/mcm$	$162.2 \sim 327.4$	$a = 8.96, b = 8.91, c = 13.18$
	δ	$Pnma$	< 162.2	$a = 9.27, b = 12.89, c = 8.64$

[†]Poglitsch A. and Weber D., J. Chem. Phys. 87, 6373 (1987).

Additional 48 Crystal structures of ABX_3 The chemical formula for the 48 crystal structures used have the formula ABX_3 (X=O, and N) with A and B site having metal cations. These 48 structures, with diverse range of A sites and B sites, are also a subset of 209 insulators[‡].

Supplementary Table 11: The parameters of the 48 crystal structures of ABX_3 (ABX_3, X=O, N).

Name	F_b (MV/m)	E_g (eV)	ω_{max} (THz)	a (Å)	b (Å)	c (Å)	α ()	β ()	γ ()	Space Group
$AlBO_3$	1986.75	4.96	39.3	3.91	3.92	3.92	82.17	82.16	82.16	R3m
$AsAlO_3$	820.28	5.3	23.48	3.78	3.78	3.78	87.9	87.9	87.9	R3m
$AsGaO_3$	693.43	4.82	23.38	3.92	3.92	3.92	87.24	87.24	87.24	R3m
$AsInO_3$	607.55	4.5	23.12	4.25	4.25	4.25	86	86.04	86.04	R3m
AsYO3	821.79	5.51	22.62	4.37	4.37	4.37	87.73	87.74	87.74	R3m
$BaHfO_3$	437.22	5.1	16.44	4.12	4.12	4.12	90	90	90	Pm-3m
$BaSiO_3$	403.97	3.01	26.3	4.87	3.52	3.52	90	90	90	P4mm
$BaSnO_3$	231.62	2.67	19.06	4.1	4.1	4.1	90	90	90	Pm-3m
$BaTeO_3$	144.03	2.19	14.48	4.31	4.31	4.31	91.38	91.42	91.42	R3m
$BaTiO_3$	212.89	3.16	14.95	3.96	3.96	3.96	89.97	89.97	89.97	R3m
$BaZrO_3$	380.29	4.65	16.31	4.17	4.17	4.17	90	90	90	Pm-3m
$BeSiO_3$	735.13	4.02	29.05	3.54	3.55	3.55	87.33	87.38	87.38	Cm
$BiAlO_3$	319.58	3.08	21.63	4.25	3.62	3.62	90	90	90	P4mm
$BiBO_3$	1284.21	4.27	37.05	3.87	3.87	3.87	81.68	81.7	81.7	R3m
$BiWN_3$	49.18	0.18	27.05	3.82	3.82	4.63	90	90	90	P4mm
$BiYO_3$	419.58	4.3	18.94	4.39	4.39	4.39	88.91	88.95	88.95	R3m
$CaGeO_3$	303.32	2.84	22.5	3.72	3.72	3.72	90	90	90	Pm-3m
$CaHfO_3$	602.6	5.32	19.46	4.16	3.99	3.99	90	90	90	P4mm
$CaSiO_3$	1062.7	6.22	23.07	3.54	3.54	3.54	90	90	90	Pm-3m
$CaTiO_3$	358.14	3.48	20.88	3.95	3.79	3.79	90	90	90	P4mm
$CdSiO_3$	240.11	2.32	22.7	3.55	3.55	3.55	90	90	90	Pm-3m
$CsIrN_3$	27.58	0.01	28.32	3.36	4.97	4.97	89.01	89.74	89.74	Cm
$CsMnN_3$	35.36	0.05	26.42	4.15	4.16	4.16	89.95	89.97	89.97	R3m
$CsTaO_3$	263.11	3.7	15.38	4.09	4.09	4.09	89.94	89.94	89.94	R3m
$GePbO_3$	39.09	0.12	19.05	4.27	4.27	4.27	87.69	87.73	87.73	R3m
$GeSnO_3$	221.55	2.32	21.13	4.05	4.05	4.05	89.02	89.02	89.02	R3m
$GeTiO_3$	284.73	2.93	20.76	3.88	3.88	3.88	89.22	89.22	89.22	R3m
$HfBeO_3$	652.62	4.65	23.36	4.22	3.27	3.27	90	90	90	P4mm
$KSbO_3$	89.33	0.79	21.34	3.94	3.94	3.94	90	90	90	Pm-3m
$LaAlO_3$	607.88	4.94	21.08	3.74	3.74	3.74	90	90	90	Pm-3m
$LaBO_3$	2784.29	5.89	38.39	3.89	3.89	3.89	81.08	81.08	81.08	R3m
$MgSiO_3$	658.97	4.13	26.46	3.29	3.29	4.34	90	90	90	P4mm
$PbSiO_3$	391.3	3.02	25.68	3.45	4.67	3.45	90	90	90	P4mm

$PbSnO_3$	276.05	2.91	20.33	4.06	4.06	4.06	89.67	89.67	89.67	R3m
$RbNbO_3$	266.59	3.46	16.61	4.02	4.02	4.02	89.93	89.93	89.93	R3m
$SbAlO_3$	443.78	3.99	21.23	3.79	3.79	3.79	89.08	89.08	89.08	R3m
$SbBO_3$	1293.17	4.38	36.26	3.92	3.92	3.92	80.68	80.69	80.69	R3m
$SbGaO_3$	529.09	4.5	21.2	3.93	3.93	3.93	88.88	88.88	88.88	R3m
$SbInO_3$	512.13	4.39	21.25	4.25	4.25	4.25	88.17	88.22	88.22	R3m
$SbYO_3$	540.88	4.69	20.62	4.38	4.38	4.38	89.21	89.21	89.21	R3m
$SiGeO_3$	28.65	0.01	21.52	3.82	3.84	3.84	87.63	87.67	87.67	Cm
$SnTiO_3$	155.9	1.48	23.31	3.79	3.79	4.27	90	90	90	P4mm
$SrGeO_3$	185.86	2.12	19.61	3.78	3.78	3.78	90	90	90	Pm-3m
$SrSiO_3$	574.72	5.14	19.55	3.62	3.62	3.62	90	90	90	Pm-3m
$SrSnO_3$	357.32	3.39	21.4	4.03	4.03	4.03	90	90	90	Pm-3m
$TeBeO_3$	908.12	4.76	27.67	3.94	3.94	3.94	83.3	83.32	83.32	R3m
$YGaO_3$	574.62	4.72	21.29	4.66	3.59	3.59	90	90	90	P4mm
$ZrBeO_3$	537.77	4.16	23.17	4.2	3.3	3.3	90	90	90	P4mm

Supplementary References

[1] Wu, T. & Wang, J. Global discovery of stable and non-toxic hybrid organic-inorganic perovskites for photovoltaic systems by combining machine learning method with first principle calculations. *Nano Energy* **66**, 104070 (2019).

[2] Cang, Z. X., Mu, L. & Wei, G. W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **14**, e1005929 (2018).

[3] Kim, C., Huan, T. D., Krishnan, S. & Ramprasad, R. A hybrid organic-inorganic perovskite dataset. *Sci. data* **4**, 170057 (2017).

‡Kim, Chiho, Ghanshyam Pilania, and Rampi Ramprasad. "Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX3 perovskites." The Journal of Physical Chemistry C 120, no. 27 (2016): 14575-14580.

[4] Saucan, E. & Weber, M. Forman's ricci curvature-from networks to hypernetworks. In *International Conference on Complex Networks and their Applications*, 706-717 (Springer Nature Switzerland AG, 2019).

[5] Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953-17979 (1994).

[6] Perdew, J. P., Burke, K. & Ernzerhof M. Perdew, burke, and ernzerhof reply. *Phys. Rev. Lett.* **80**, 891 (1998).

[7] Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169-11186 (1996).

[8] Cang, Z. X. & Wei, G. W. TopologyNet: Topology based deep convolutional and multitask neural networks for biomolecular property predictions. *PLoS Comput. Biol.* **13**, e1005690 (2017).

[9] The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board (2015).

[10] Wee, J. J. & Xia, K. L. Ollivier Persistent Ricci Curvature-Based Machine Learning for the ProteinLigand Binding Affinity Prediction. *J. Chem. Inf. Model.* **61**, 1617-1626 (2021).

[11] Wee, J. J. & Xia, K. L. Forman persistent Ricci curvature (FPRC)-based machine learning models for proteinligand binding affinity prediction. *Brief. Bioinform.* **22**, bbab136 (2021).

[12] Samal, A. et al. Comparative analysis of two discretizations of ricci curvature for complex networks. *Sci. Rep.* **8**, 8650 (2018).

[13] Forman, R. Bochner's method for cell complexes and combinatorial Ricci curvature. *Discrete Comput. Geom.* **29**, 323-374 (2003).

[14] Sreejith, R. P., Mohanraj, K., Jost, J., Saucan, E. & Samal, A. Forman curvature for complex networks. *J. Stat. Mech.: Theory Exp.* **2016**, 063206 (2016).

[15] Saucan, E., Sreejith, R. P., Vivek-Ananth, R. P., Jost J. & Samal, A. Discrete ricci curvatures for directed networks. *Chaos Solit. Fractals* **118**, 347-360 (2019).

[16] Chi, C., Weike Y., Yunxing Z., Zheng C. & Shyue Ping O. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 9, 3564-3572 (2019).

[17] Schütt, K., Kindermans, P. J., Sauceda F., Huziel E., Chmiela S., Tkatchenko A. & Müller, K.R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *J. Chem. Phys.* **30**, (2017).