frontiers

Supplementary Material

An Evolutionary Variational Autoencoder for Perovskite Discovery

Ericsson Tetteh Chenebuah[1, 2*], Michel Nganbe[1], Alain Beaudelaire Tchagang[1,2]

[1]Department of Mechanical Engineering, University of Ottawa, 161 Louis-Pasteur, Ottawa, ON, K1N 6N5, Canada

[2]Digital Technologies Research Center, National Research Council of Canada, 1200 Montréal Road, Ottawa, ON, K1A 0R6, Canada

*** Correspondence:** echen013@uottawa.ca

1 Image-based perovskite representation and pixel attribution

As illustrated in the main article (Fig. 2), the image-based descriptor design consists of two sections that play important roles in relating the crystallographic properties and thermochemistry behavior of the described perovskite material in the training set. Considering the chemical formula $KCrF_3$ (Materials Project ID: 566131 and ICSD ID: 27690) for a crystal geometry (lattice and basis) with five atoms in the unit cell for example, the stacking arrangement for developing the input image is guided by the representation for the ABX_3 stoichiometry. In the original form, the crystallographic properties section aggregates into a (152×7) matrix, while the thermochemistry section aggregates into a (130×7). The first three columns in the thermochemistry section corresponds to the discrete features with respect to each distinctive ionic occupancy for site-A, site-B, and site-X. Besides, the current study utilizes similar thermochemistry properties as suggested in the Crystal Graph Convolutional Neural Network (CGCNN) model (Xie and Grossman, 2018) for target property prediction. However, new features that are not considered in the conventional CGCNN are infused into the present descriptor design. The newly introduced features are: average ionic radius, polarizability, specific heat, and thermal conductivity. The proven effect of these additional features for improved target prediction modeling has been demonstrated in our previous research (Chenebuah et al., 2022). Table S1 outlines all discretized properties (including label atomic numbers) and shows the range in boundary values, in addition to the number of bins for discretization (i.e. one-hot encoding). Adjoining both crystallographic and discrete thermochemistry properties along their row axes results into a (282×7) matix. In order to fit the universal description that equally adopts other complex forms of perovskite stoichiometries (i.e. $A_2BB'X_6$ and $AA'BB'X_6$), the $KCrF_3$ descriptor is augmented (zero-padded) and normalized to finally aggregate into (282×8) matrix. For illustration, Fig. S1 reveals fully developed (282×8) images for describing $KCrF_3$, Sr_2FeOsO_6 and $KFePbOF_6$ perovskites in the training set. The denser contrast in pixel values for $KFePbOF_6$ indicate the contributing effect of more chemical elements used to develop the input image. For ML training however, all perovskite images are uniformly reshaped into a $(94 \times 8 \times 3)$ *RGB* descriptor form, which serves as inputs in the generative and target property simulation experiments.

Table S1. Thermochemistry properties used in developing the image-based input descriptor. Each property is one-hot encoded based on the discrete number of bins and numerical range in real values.

Ref	Thermochemistry property	Unit	Range	Number of bins
-	Atomic number	-	1,2,…,103	103
-	Group number	-	1,2,…,18	18
-	Row number	-	1,2,…,9	9
D. Lide, 2004.	Pauling electronegativity	Pauling	0.7 - 3.98	10
Cordero et al., 2008.	Covalent radius	Angstrom	0.28 - 2.6	10
-	Valence	-	1,2,…,9	9
D. Lide, 2004.	First ionization energy (log scale)	eV	3.89 - 24.59	10
D. Lide, 2004.	Electron affinity	eV	-2.33 - 5.94	10
-	Block	-	*s,p,d,f*	4
Ong et al., 2013.	Molar volume (log scale)	cm^3	4.39 - 70.94	10
Ong et al., 2013.	Average ionic radius	Angstrom	0 - 1.94	10
D. Lide, 2004.	Static average electric dipole Polarizability	$10^{-24}\,cm^3$	0.21 - 59.42	10
D. Lide, 2004.	Specific heat (log scale)	KJ/kg.K	0.06 - 14.3	10
Ho et al., 1972.	Thermal conductivity (log scale)	W/m.K	0.0036 - 430	10

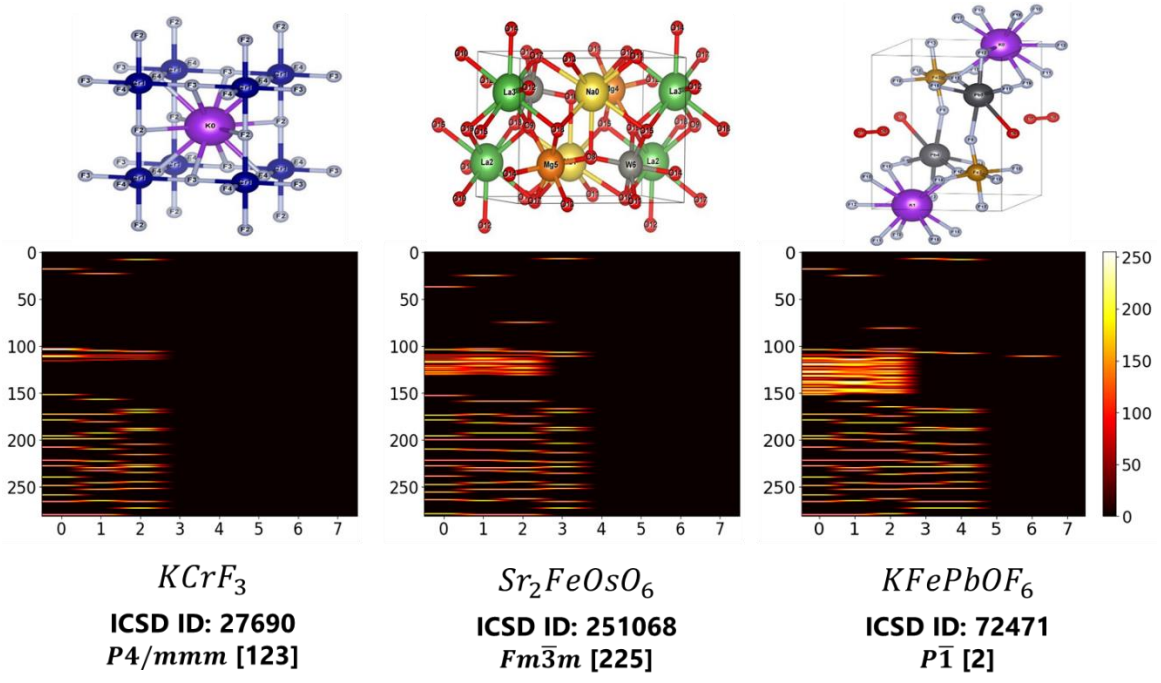

Figure S1. Fully developed (282×8) input image from the normalization of feature embedding for $KCrF_3$, Sr_2FeOsO_6 and $KFePbOF_6$ perovskites in training set. Hot intensities on input image indicate higher pixel values.

To analyze which pixels are critical in the forward modeling process, pixel attribution is investigated on the image-based descriptor. As such, figures S2 (A) and (B) display average saliency maps from the predictions of the formation energy ($\boldsymbol{E_f}$) and energy above convex hull ($\boldsymbol{E_{hull}}$), respectively. The concept of gradient-based predictive approach is applied for generating the maps with respect to the outputted gradients from each contributing input feature (Molnar, 2022). Upon inspection, both saliency maps are seen to be similar with bright intensities suggesting the higher importance of that input pixel/feature in the prediction exercise. This is to be expected moreover, as the trained regressive models regard both target's ML assimilation processes to be strongly interrelated from thermodynamic calculations (Emery and Wolverton, 2017; Ishikawa and Miyake, 2020; Chenebuah et al., 2021). In general, the most influential pixel region on the saliency maps with the best effect on the prediction process can be observed to be within rows 109 to 132 (i.e. image height). This highly influential region accommodates unit cell features in the lattice parameters (i.e. 3D inter-axial angles and lengths) and fractional atomic coordinates. Moreover, both unit cell features are conjointly responsible for describing the geometrical/crystallographic uniqueness of each perovskite sample in the training dataset, and as such, greatly influence the prediction process. On further examination, it also appears that almost all thermochemistry input features (i.e. rows 152 to 282) play crucial roles in the prediction process. This therefore supports the rationale on including the four additional features (i.e. average ionic radius, polarizability, specific heat, and thermal conductivity) of which were originally omitted in the traditional CGCNN model for accurate target property prediction. To shed more light into the monotonic relationships among all thermochemistry properties, figure S3 displays Spearman's correlation (ρ) as calculated using the average values of the distinctive input features used to build the perovskite crystal.

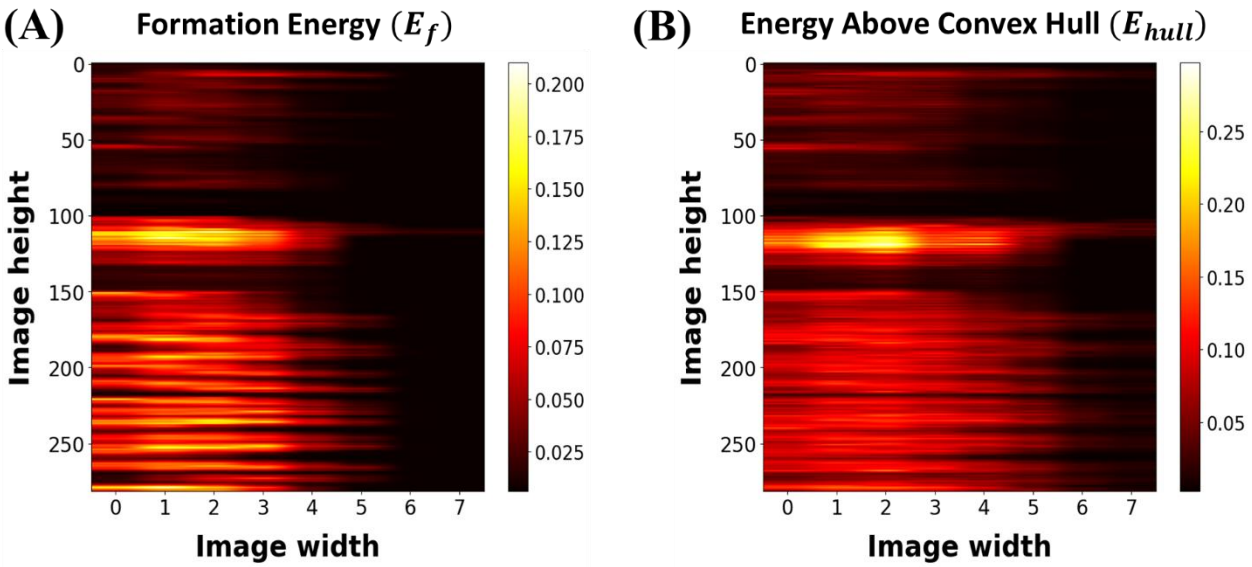

Figure S2. Average pixel attribution (saliency maps) on image-based features that are highly influential towards the predictions of: (A) formation energy and (B) energy above convex hull. The assembly in image dimension follows the same design pattern previously illustrated using the Figure 2 in the main article.

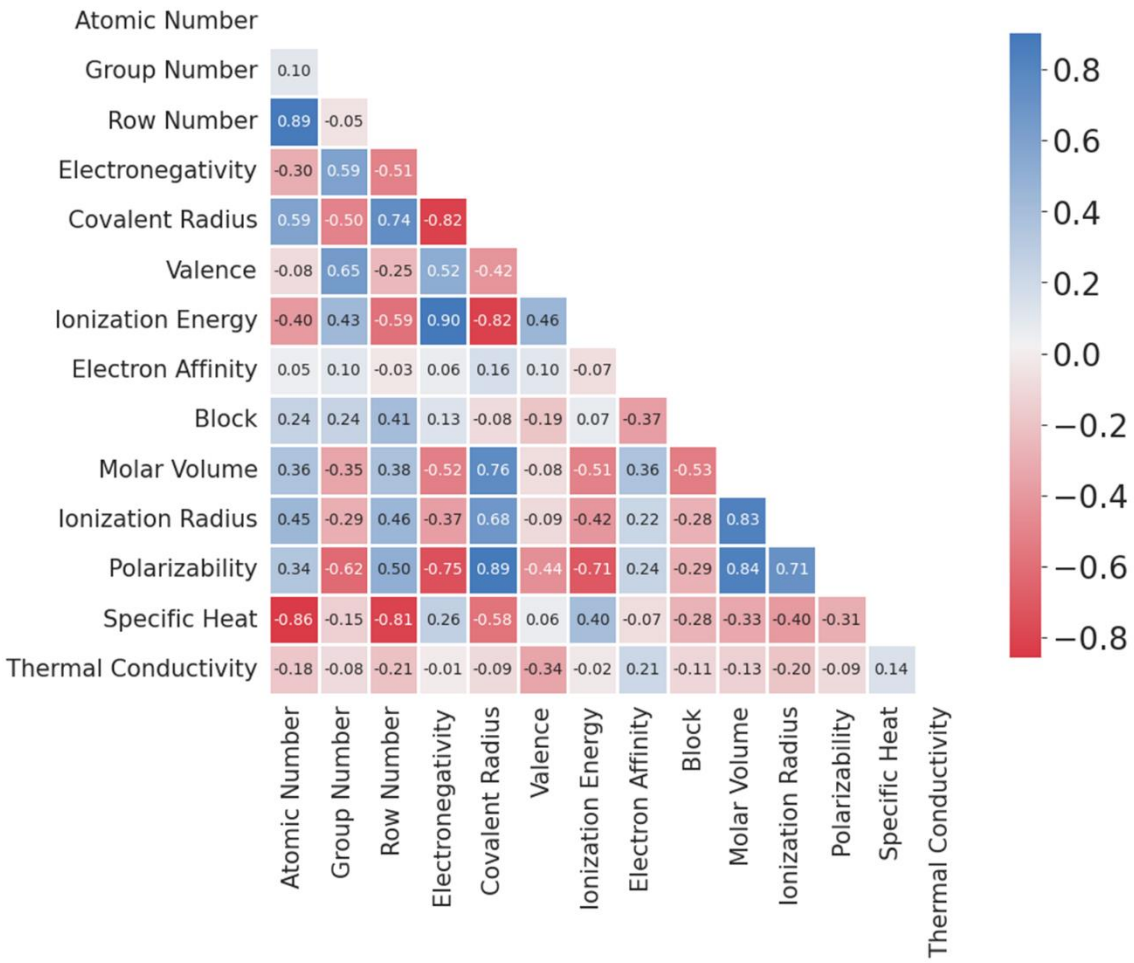

Figure S3. Spearman's ranked correlation coefficient among all discretized input features, as represented using phase-field descriptor approach.

2 Model architecture and learning curves

The general Evolutionary Variational Autoencoder for Perovskite Discovery (EVAPD) model consists of several inter-related deep learning frameworks that are used to solve the forward and inverse design challenge. For addressing the inverse design, a semi-supervisory variational autoencoder (SS-VAE) is developed, which comprises three main sections, namely: encoder, target-learning regressor, and decoder. Through the incorporation of a feed-forward multi-layer perceptron (MLP), the target-learning regressor influences the overall training losses by mapping encoded sampling vectors ($\{\boldsymbol{z_i}\} \subseteq \boldsymbol{Z}$) to their corresponding formation energy targets. As such, Table S2 details the architecture for designing the SS-VAE model and shows hyperparameter specifications for guiding the training process. In addition, Fig. S4 displays the error-losses curves for simulating VAE and MLP learning processes, as evaluated on training and validation sets. The curves are seen to descend uniformly with increasing number of epochs. For updating the feedback function of the genetic algorithm, independent deep learning frameworks are further architectured, which aids the prediction of the energy above convex hull ($\boldsymbol{E_{hull}}$) and classification of inorganic crystal structure

database (*ICSD*) label targets. Likewise, Table S3 outlines the modeling architecture and hyperparameters for directly mapping deterministic target variables using two-dimensional convolutional neural networks (Conv2D). Besides, the same Conv2D forward design architecture is also applied for determining energy above convex hull and formation energy targets of all newly discovered perovskites, as outlined in the main article (Table 2). All architectural ML designs were scripted using the Python programing language on a Keras functional API (Gulli and Pal, 2017) within a TensorFlow backend (Abadi et al., 2016). The codes are made available at github.com/chenebuah/EVAPD.

Table S2. Modeling architecture and hyperparameters for designing the generative SS-VAE model of the EVAPD design.

Network	Layer API class	Filter/Units	Kernel	Stride	Padding	Activation
Encoder (VAE)	Input shape	$(94 \times 8 \times 3)$				
	Conv2D	32	3×3	1	same	LeakyReLU (α=0.2)
	Conv2D	64	3×3	2	same	LeakyReLU (α=0.2)
	Flatten	12032				
	Dense	1024				Sigmoid
	$\mathbf{z_{mean}}$ Dense	256				Linear
	$\mathbf{z_{var}}$ Dense	256				Linear
	z (encoded sampling vector)	$\mathbf{z_{mean}} + \mathbf{exp}(\mathbf{0.5} \times \mathbf{z_{var}}) \cdot \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$				
Regressor (MLP)	Input					ReLU
	Dense	256				ReLU
	Dense	128				ReLU
	Dense	64				ReLU
	Dense	32				ReLU
	Output Dense	1				Linear
Decoder	Dense	1024				Sigmoid
	Dense	12032				LeakyReLU (α=0.2)
	Reshape	$(47 \times 4 \times 64)$				
	Conv2DTranspose	32	3×3	2	same	LeakyReLU (α=0.2)
	Reconstructed Conv2DTranspose	3	3×3	1	same	Sigmoid
	Output shape	$(94 \times 8 \times 3)$				
Hyperparameters						
Learning rate		1.00e-04				
Decay		1e-4/200				
Batch size		32				
Epoch		1500				
Optimizer		Adam				

Table S3. Conv2D model architecture and hyperparameters for predicting all relevant target variables.

Layer API class	Filter/Units	Kernel	Stride	Padding	Activation
Input shape	$(94 \times 8 \times 3)$				
Conv2D	8	3×3	1	same	ReLU
Conv2D	16	3×3	2	same	ReLU
Conv2D	32	3×3	2	same	ReLU
MaxPooling2D		2×2	None	same	Linear
Dropout	rate $=0.2$				
Flatten	384				
Dense	256				ReLU
Dense	64				ReLU
Dense	16				ReLU
Dense	4				ReLU
Output Dense	1		Linear (if predicting E_{hull} or E_g values) Sigmoid (if classifying *ICSD* labels)		
Hyperparameters					
Learning rate	1.00e-03				
Decay	1e-3/200				
Batch size	32				
Epoch	150				
Optimizer	Adam				
Loss	Mean squared error (if predicting E_{hull} or E_g values) Binary cross entropy (if classifying *ICSD* labels)				

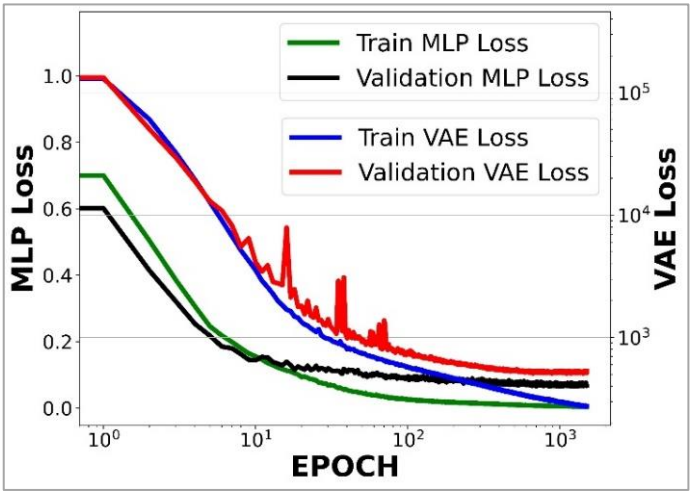

Figure S4. Learning curves on modeling behavior of distinctive MLP and VAE during training and validation.

3 Genetic Algorithm hyperparameters

For obtaining high-quality solutions, the study applies genetic algorithm (GA) to search for the most promising perovskite candidates. The GA algorithm as implemented in the study is based on the PyGAD (Gad, 2021) module, which is an open-source Python library for performing evolutionary learning and optimizing ML algorithms. Table S4 outlines the crucial hyperparameter settings as executed in the present research.

Table S4. Hyperparamter settings for implementing search operations using the PyGAD genetic algorithm module.

Individual vector length (i.e. gene space)	**256**
Number of generations	**100**
Number of parents mating per batch population	**50**
Cross-over type	**Single-point**
Mutation type	**Adaptive**
Ideal mutation percent for low-quality solutions	**5%**
Ideal mutation percent for high-quality solutions	**2.5%**
Parent selection type	**Steady-state selection**
Fitness function	**Multi-objective user defined**

4 More details on forward design modeling results

Figure S5 displays additional results from the target prediction exercises of *ICSD* labeling and energy above convex hull ($\boldsymbol{E_{hull}}$), in addition to the band gap ($\boldsymbol{E_g}$). It shall be noted however, that $\boldsymbol{E_g}$ does not affect the simulation process or the EVAPD model in general. The $\boldsymbol{E_g}$ results are presented for purely assessing the further predictive capability of the present image-based descriptor design. In future study, concrete plans are being considered towards integrating the $\boldsymbol{E_g}$ target into the EVAPD model to produce a more robust framework that equally assimilates concise electrical behavior. On taking a closer look into the average overall modeling performance using five-fold cross-validation for $\boldsymbol{E_{hull}}$ and $\boldsymbol{E_g}$, the regressive fitting (i.e. R^2) scores are determined at 54.40 ± 8.91 % (Fig. S5 (A)) and 80.05 ± 0.88 % (Fig. S5 (B)), respectively. The current study acknowledges that the prediction accuracy for all regressive targets (i.e. $\boldsymbol{E_{hull}}$, $\boldsymbol{E_f}$ and $\boldsymbol{E_g}$) may be further improved upon the consideration of more sophisticated descriptors and/or ML architectural designs (Chenebuah et al., 2021; Ren et al., 2022; Chenebuah et al., 2023). Such sophistication in descriptor concept are expected to be at higher computational training cost due to the denser pixels from the enhanced descriptor image. The current study seeks to keep the modeling frameworks for both descriptor and deep learning architectures as computationally simple as possible while ensuring that efficiency - as prioritized on novel perovskite generation - is not compromised. For this reason, such enhanced descriptor designs were not implemented in the present research. Moreover, Figs. S5 (C) and (D) display classification results from the modeling of *ICSD* label targets. It should be noted that prior to conducting the *ICSD* classification exercise, equal numbers of samples from both classes (i.e. *ICSD* vs *non-ICSD* labels) were used for training in order to avoid the biased adjudication of the model towards a majority class. The achieved results confirm the average Receiver Operating Characteristic (ROC) curve at 84.35 % with respect to Area Under Curve (AUC).

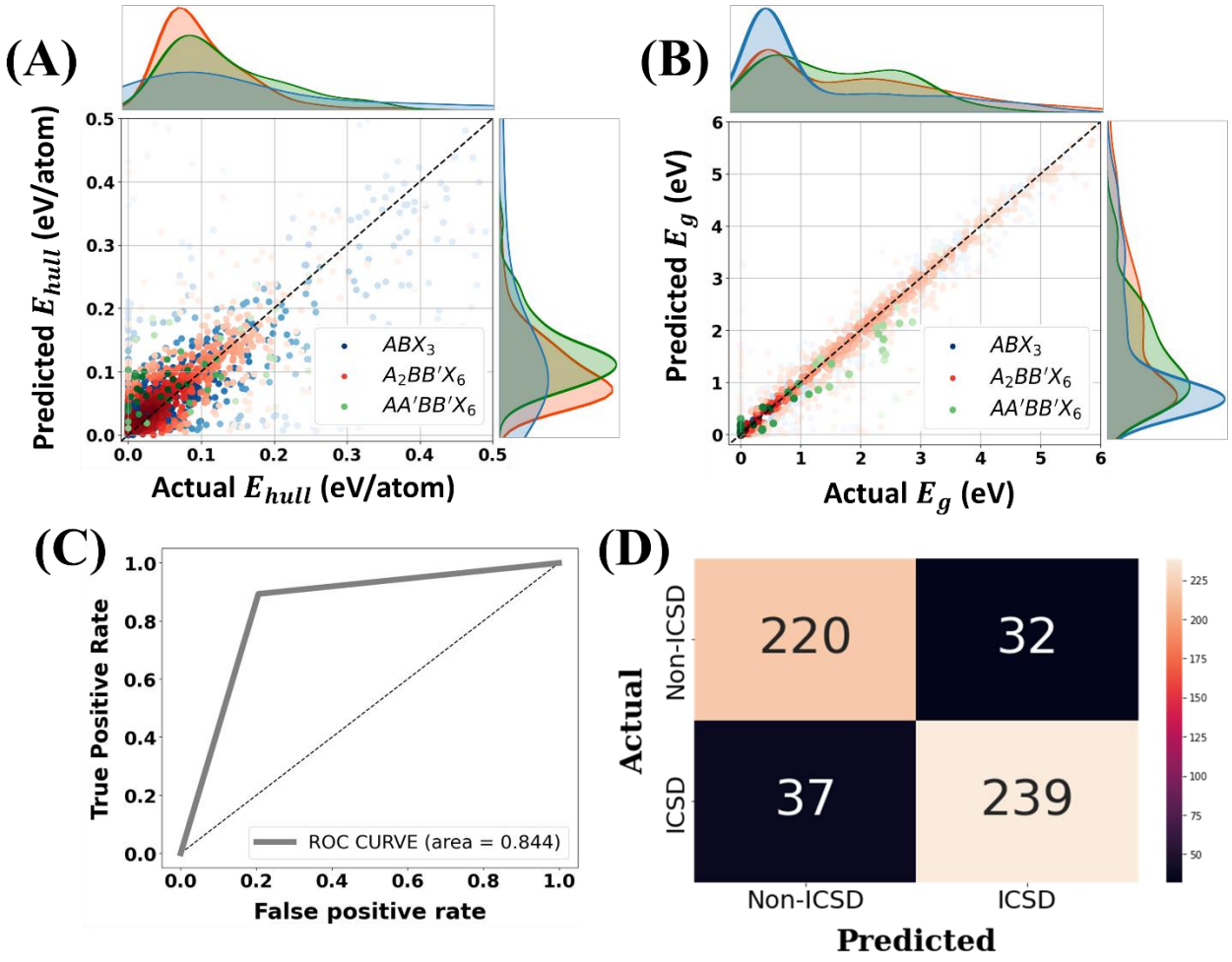

Figure S5. Error evaluation on interested target properties. (A) Energy above convex hull ($\boldsymbol{E_{hull}}$) with regression fitting at 54.40 % R^2; (B) Energy band gap ($\boldsymbol{E_g}$) with regression fitting at 80.05 % R^2; (C) Receiver Operating Characteristic (ROC) curve showing classification measurements for *ICSD* versus *non-ICSD* perovskites at different threshold settings. Area Under Curve (AUC) is reported at 84.35 %; (D) Confusion matrix on *ICSD* classification analysis.

5 Target-property analysis of newly discovered perovskites

The current study presents 114 $A_2BB'X_6$ and 23 $AA'BB'X_6$ new materials discovered by the EVAPD model that successfully underwent Density Functional Theory (DFT) validation. The newly discovered materials are inspected in order to correlate specific trends with respect to stability and functionalization. As such, Fig. S6 (A) plots DFT-determined $\boldsymbol{E_g}$ against Conv2D model predicted $\boldsymbol{E_g}$. The Conv2D model is demonstrated to accurately predict $\boldsymbol{E_g}$ by about 40% of all perovskites within ±0.3 eV and 50% within ±0.5 eV standard deviation. Perovskites predicted with very high accuracy (i.e. ±0.001) were mostly observed to be metallic materials with $\boldsymbol{E_g = 0\ eV}$, such as $SrFeIrRuO_6$ (CIF ID: #125), Ca_2YOsO_6 (CIF ID: #1), Sr_2LuReO_6 (CIF ID: #67) and In_2YOsO_6

(CIF ID: #2). On investigating highly stable compounds, Fig. S6 (B) reveals the distribution in data clustering with respect to model predicted $\boldsymbol{E_{hull}}$ versus $\boldsymbol{E_f}$ (i.e. formation energy). From the investigation, about 73% of all newly discovered materials are within prescribed stability and synthesizability, which were previously defined at $\boldsymbol{E_{hull}} \leq 0.08$ eV/atom (Singh et al., 2019) and $\boldsymbol{E_f} \leq -1.5$ eV/atom (Ren et al., 2022) in the main article. Examples of some highly stable perovskites with $\boldsymbol{E_{hull}} \leq 0.01$ eV/atom include: Sr_2MgIrO_6 (CIF ID: #71), K_2MgVO_6 (CIF ID: #7), K_2LiAlF_6 (CIF ID: #4), and Sr_2MgRuF_6 (CIF ID: #76). On investigating DFT-deterministic functional properties, Figs. S6 (C) and (D) reveal band gap and magnetic target property distributions, respectively, with respect to stability. It can be observed that a high proportion of newly discovered materials exhibit metallic and diamagnetic properties. The new materials are archived in NOMAD repository (http://www.doi.org/10.17172/NOMAD/2023.05.31-1) for interested users.

6 DFT-computed relative energies of newly discovered perovskites with potential for photovoltaic and optoelectronic applications

In addition to model predicted formation energy ($\boldsymbol{E_f}$), the stability of three highly regarded materials were further assessed in the main article using DFT-determined relative energy ($\boldsymbol{E_{rel}}$) (see **section 4.1**). Thermodynamically, $\boldsymbol{E_{rel}}$ accounts for the difference in total DFT-computed energies between the relaxed perovskite bulk crystal and the sum of the isolated constitutive elements, at the same level of theory as the bulk crystalline material calculation (Kim et al., 2017; Emery and Wolverton, 2017). In order to establish some form of comparison between model predicted $\boldsymbol{E_f}$ and DFT-determined $\boldsymbol{E_{rel}}$, the computation is extended to fourteen additional materials with bandgaps that are equally suitable for photovoltaic and/or optoelectronic applications. Considering the perovskite stoichiometries of interest, the equations for computing $\boldsymbol{E_{rel}}$ are expressed as follows:

$$\boldsymbol{E_{rel}(A_2BB'X_6) = E_{total}(A_2BB'X_6) - 2E(A) - E(B) - E(B') - 3E(X_2)} \quad \textbf{(1)}$$

$$\boldsymbol{E_{rel}(AA'BB'X_6) = E_{total}(AA'BB'X_6) - E(A) - E(A') - E(B) - E(B') - 3E(X_2)} \quad \textbf{(2)}$$

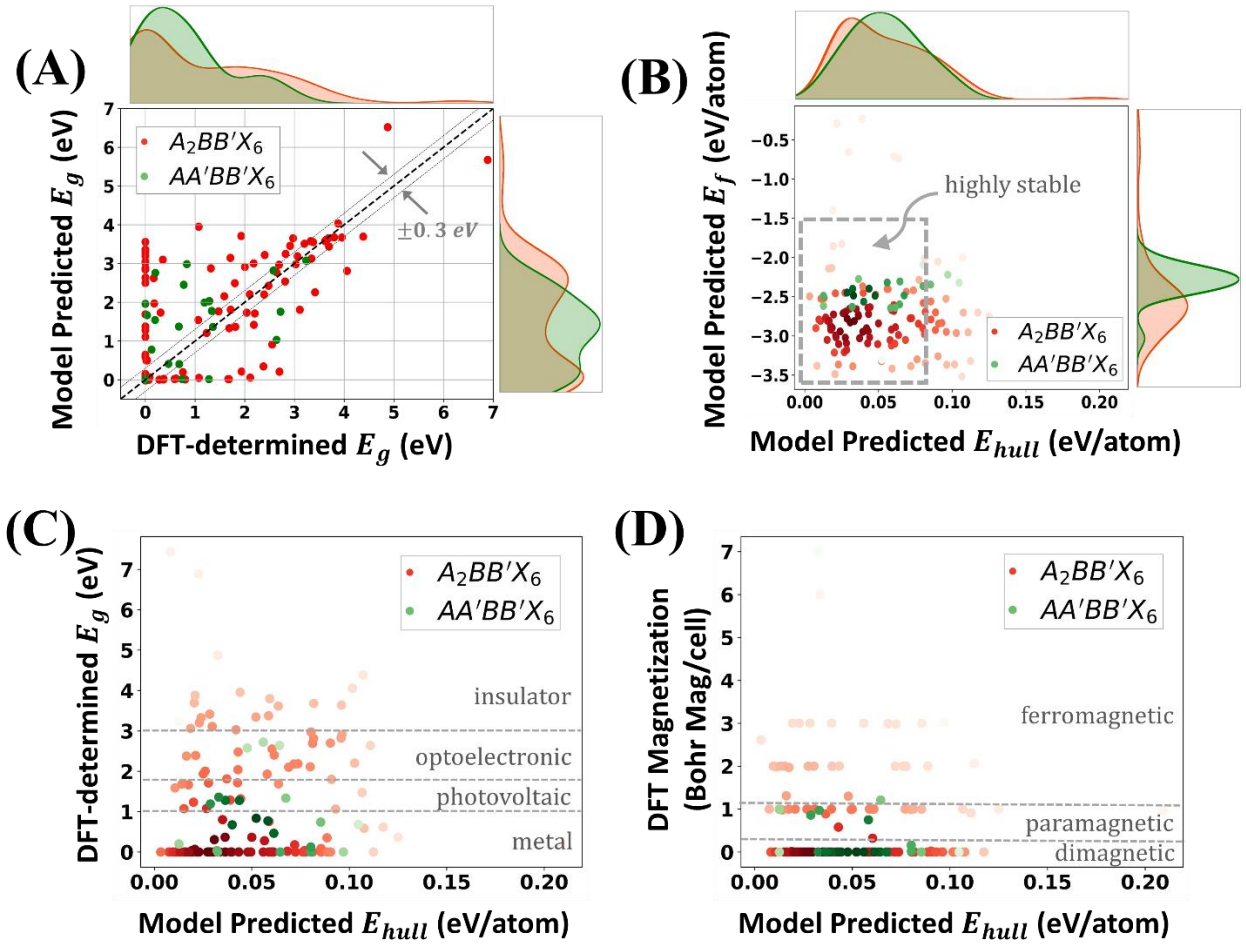

Figure S6. Correlation between DFT-determined and model predicted target properties on newly discovered perovskites. (A) DFT-determined versus model predicted energy bandgap ($\boldsymbol{E_g}$). Dashed lines surrounding the fitted y=x line are deviated at ±0.3 eV. (B) Model predicted $\boldsymbol{E_{hull}}$ versus model predicted $\boldsymbol{E_f}$ revealing some highly stable perovskite points; (C) Model predicted $\boldsymbol{E_{hull}}$ versus DFT-determined $\boldsymbol{E_g}$ showing distribution in electrical behavior with respect to stability; (D) Model predicted $\boldsymbol{E_{hull}}$ versus DFT-determined magnetization (in units of Bohr Mag/cell) showing distribution in approximated magnetic behavior with respect to stability.

$\boldsymbol{E_{total}}(.)$ is the total DFT computed energy of the perovskite chemical compound; $\boldsymbol{E}(.)$ is the DFT computed energy of the isolated constituent ions. For all computations, the lattice coordinates and inter-axial angles are consistent with the final relaxed perovskite structure. Isolated ions are positioned at the center of the bulk crystal lattice, and for diatomic anions (i.e. $\boldsymbol{X_2}$ in oxides, hydrides and halides), two bonded atoms are positioned at the center. Figure S7 reveals the results from the computations of DFT-determined $\boldsymbol{E_{rel}}$ and model predicted $\boldsymbol{E_f}$, as expressed using similar units (i.e. eV/atom). For most materials, the predicted $\boldsymbol{E_f}$ is substantially more than $\boldsymbol{E_{rel}}$. This is consistent with

nominal practice, given that $\boldsymbol{E_{rel}}$ overestimates stability by not accounting for Hubbard-U corrections or fittings with experimental formation energies, notwithstanding the possibility of inherent errors that may be associated with the DFT software (Cohen et al., 2012). Moreover, the experimented dataset from the Materials Project (Jain et al., 2013), which is used for training the predictive $\boldsymbol{E_f}$ model, obtains their DFT properties via Vienna Ab initio Simulation Package (VASP), whereas computations for $\boldsymbol{E_{rel}}$ is done using Quantum Espresso (QE). These discrepancies are possibly some of the reasons for the apparent contrast in reported stability values between $\boldsymbol{E_f}$ and $\boldsymbol{E_{rel}}$. The purpose of presenting both quantities is to show the similarity in estimation pattern, and also to verify the consistency in the used predictive forward design model. From Fig. S7, it can therefore be observed that a similar trend in estimation pattern exists between $\boldsymbol{E_f}$ and $\boldsymbol{E_{rel}}$ for most materials. The average marginal difference between $\boldsymbol{E_f}$ and $\boldsymbol{E_{rel}}$ was estimated at 0.972 eV/atom.

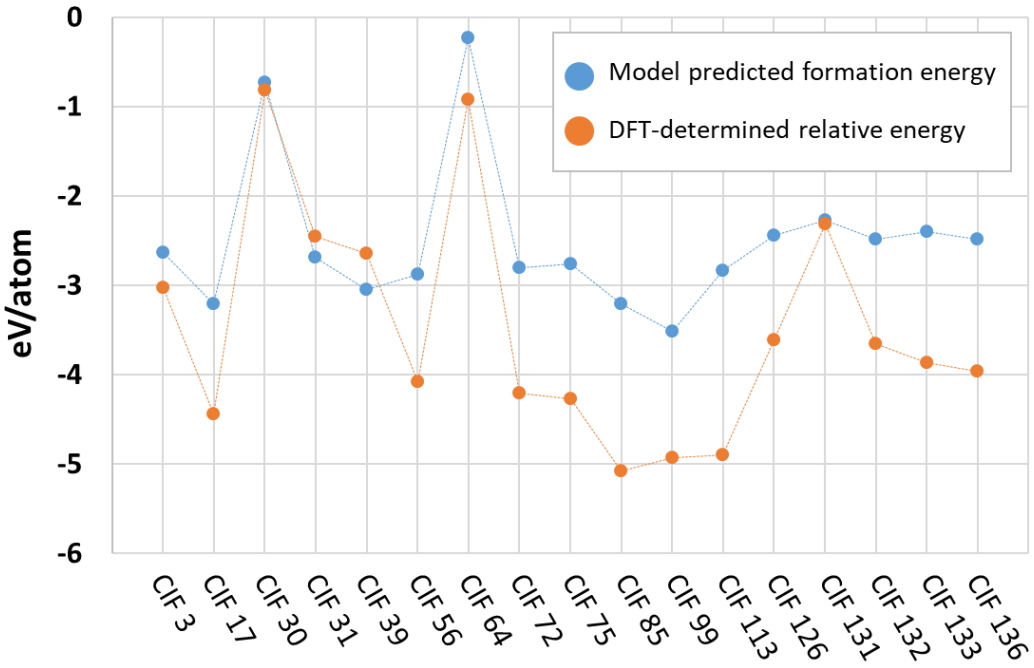

Figure S7. Comparison in stability assessments between model predicted formation energy ($\boldsymbol{E_f}$) and DFT-determined relative energy ($\boldsymbol{E_{rel}}$) for some highly regarded perovskites. Note that generally, with Quantum Espresso (Giannozzi et al., 2009), total energies for computing $\boldsymbol{E_{rel}}$ are evaluated in Rydberg, while the presented results are converted to similar units of electron volt (eV).

7 Standard perovskite forms used as reference in similarity analytical assessment

The generated candidates emerging from the Spherical Linear Interpolation (SLERP) and Genetic Algorithm (GA) stages are screened to ensure their geometrical atomic coordination is in close similarity with some standard perovskite forms. As such, Table S5 outlines the twelve standard

perovskite forms from the Materials Project database (Jain et al., 2013) that are used as references in the current study. The frequency in calculated dissimilarity values ($\boldsymbol{\mathcal{F}}$) between the outlined standards and each generated candidate is illustrated in the Fig. 10 of the main article.

Table S5. Standard perovskite forms used in the similarity analytical assessment of feasible chemical compounds. All standards are primitive or singular formula unit perovskites, and as such, contain ten atoms in their unit cells.

Materials Project ID	Perovskite formula	Crystal system	Spacegroup	$\boldsymbol{E_f}$ (eV/atom)	$\boldsymbol{E_{hull}}$ (eV/atom)	*ICSD* label
$A_2BB'X_6$ standards						
mp-1079615	Ba_2UCdO_6	cubic	$Fm\bar{3}m$	-3.121	0	True
mp-1091394	Ba_2NaIO_6	cubic	$Fm\bar{3}m$	-2.220	0	True
mp-6183	Ba_2PrRuO_6	monoclinic	$C2/m$	-2.752	0	True
mp-22531	Ba_2GdNbO_6	tetragonal	$I4/m$	-3.433	0	True
mp-1078551	Ba_2NdMoO_6	triclinic	$P\bar{1}$	-3.032	0	True
mp-13356	Ba_2SrTeO_6	trigonal	$R\bar{3}$	-2.740	0	True
$AA'BB'X_6$ standards						
mp-1218109	$SrPrFeCoO_6$	triclinic	$P\bar{1}$	-2.461	0	False
mp-1227701	$BaSrCaWO_6$	monoclinic	Cm	-3.028	0.004	False
mp-39249	$BaLaMgRuO_6$	tetragonal	$I\bar{4}$	-2.769	0.006	False
mp-1223441	$KBaLiZnF_6$	cubic	$F\bar{4}3m$	-3.301	0.007	False
mp-1227325	$BaSrMgTeO_6$	cubic	$F\bar{4}3m$	-2.695	0.007	False
mp-1227677	$BaSrBiSbO_6$	trigonal	$R3$	-2.462	0.012	False

References

Abadi, M., Barham, P., Chen, J., et al. (2016). TensorFlow: a system for large-scale machine learning. arXiv:1605.08695v2 [cs.DC]. doi:10.48550/arXiv.1605.08695

Chenebuah, E. T., Nganbe, M., and Tchagang, A. B. (2021). Comparative analysis of machine learning approaches on the prediction of the electronic properties of perovskites: A case study of ABX_3 and $A_2BB'X_6$. *Mater. Today Commun.* 27, 102462. doi:10.1016/j.mtcomm.2021.102462

Chenebuah, E. T., Nganbe, M., and Tchagang, A. B. (2023). A Fourier-transformed feature engineering design for predicting ternary perovskite properties by coupling a two-dimensional convolutional neural network with a support vector machine (Conv2D-SVM), *Mater. Res. Express*. 10, 026301. doi:10.1088/2053-1591/acb683

Cohen, A. J., Mori-Sánchez, P., and Yang, W. (2012). Challenges for Density Functional Theory. *Chem. Rev*. 112 (1), 289-320. doi:10.1021/cr200107z

Cordero, B., Gómez, V., Platero-Prats, A. E., Revés, M., Echeverría, J., Cremades, E., Barragán, F., and Alvarez, S. (2008). Covalent radii revisited. *Dalton Trans*., 21, 2832–2838. doi:10.1039/B801115J

Emery, A., and Wolverton, C. (2017). High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO_3 perovskites. *Sci Data*. 4, 170153. doi:10.1038/sdata.2017.153

Gad, A. F. (2021). PyGAD: An Intuitive Genetic Algorithm Python Library. arXiv:2106.06158v1 [cs.NE]. doi:10.48550/arXiv.2106.06158

Giannozzi, P., Baroni, S., Bonini, N., Calandra, M., Car, R., Cavazzoni, C., et al. (2009). QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials, J. Phys.:Condens. Matter. 21 (39), 395502. doi:10.1088/0953-8984/21/39/395502

Gulli, A. and Pal, S. (2017). Deep learning with keras. *Packt Publishing Ltd*.

Ho, C. Y., Powell, R. W., and Liley, P. E. (1972). Thermal conductivity of the elements. *J. Phys. Chem. Ref. Data*. 1, 279-421. doi:10.1063/1.3253100

Ishikawa, T., and Miyake, T. (2020). Evolutionary construction of a formation-energy convex hull: Practical scheme and application to a carbon-hydrogen binary system, Phys. Rev. B. 101, 214106. doi:10.1103/PhysRevB.101.214106

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., et al., (2013). The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*. 1 (1), 011002. doi:10.1063/1.4812323

Kim, C., Huan, T. D., Krishnan, S., and Ramprasad, R. (2017). A hybrid organic-inorganic perovskite dataset. *Sci Data.* 4, 170057, (2017). doi:10.1038/sdata.2017.57

Lide, D. (2004), CRC Handbook of Chemistry and Physics, 85th Edition. *Taylor & Francis*.

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., and Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* 68, 314–319. doi:10.1016/j.commatsci.2012.10.028

Ren, Z., Tian, S. I. P., Noh, J., Oviedo, F., Xing, G., Li, J., et al. (2022). An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*. 5 (1), 314-335. doi:10.1016/j.matt.2021.11.032

Singh, A. K., Montoya, J. H., Gregoire, J. M., and Persson, K. A. (2019). Robust and synthesizable photocatalysts for CO_2 reduction: a data-driven materials discovery. *Nat. Commun.* 10, 443. doi:10.1038/s41467-019-08356-1

Xie, T., and Grossman, J. C. (2018). Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. Phys. Rev. Lett. 120 (14), 145301. doi:10.1103/physrevlett.120.145301.