

Summer 2023

## Machine-Learning Classification to Predict the Dimensionality of Hybrid Organic-Inorganic Halide Perovskites

Yatiwelle Koralalage Samuditha Sandaru Yatiwella

Follow this and additional works at: <https://scholarcommons.sc.edu/etd>

 Part of the [Chemistry Commons](#)

---

### Recommended Citation

Yatiwella, Y.(2023). *Machine-Learning Classification to Predict the Dimensionality of Hybrid Organic-Inorganic Halide Perovskites*. (Master's thesis). Retrieved from <https://scholarcommons.sc.edu/etd/7492>

This Open Access Thesis is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact [digres@mailbox.sc.edu](mailto:digres@mailbox.sc.edu).

MACHINE-LEARNING CLASSIFICATION TO PREDICT THE DIMENSIONALITY OF  
HYBRID ORGANIC-INORGANIC HALIDE PEROVSKITES

by

Yatiwelle Koralalage Samuditha Sandaru Yatiwella

Bachelor of Science  
University of Colombo 2017

---

Submitted in Partial Fulfillment of the Requirements  
for the Degree of Master of Science in  
Chemistry  
College of Arts and Sciences  
University of South Carolina  
2023

Accepted by:

Christopher Sutton, Director of Thesis

Andrew Greytak, Reader

Ann Vail, Dean of the Graduate School

## ACKNOWLEDGMENTS

Words cannot express my gratitude to Professor Christopher Sutton, who gave me valuable instructions on my journey; for his support. I also could not have undertaken this journey without the support of Professor Andrew Greytak and Professor Mark Berg, who generously provided knowledge and expertise. Additionally, this endeavor would not have been possible without the generous support from my group colleagues Sourin, Dipannoy, Jack, Dr. Adihkari, and Dr. Karimitari. I also thank our outside collaborator Professor Steve for his feedback on our project.

I am also grateful to my teachers, the University of South Carolina; Prof. Sophya Garashchuk, Prof. Michael L. Myrick, Prof. Dmitry V. Peryshkov, and the teachers of my undergraduate university; the University of Colombo, SriLanka and all the other teachers who impacted and inspired me.

Most importantly I would like to thank my parents and parents-in-law and my siblings for their incredible support and encouragement and for all they have done to make my life happier. My heartfelt gratitude goes to my dear wife, who supported me in numerous ways to accomplish this target.

Lastly, I would be remiss in not mentioning my university colleagues from the Department of Chemistry and all my friends for their invaluable help and friendship.

## ABSTRACT

Low dimensional hybrid perovskites have demonstrated remarkable performance in photovoltaic applications, primarily due to their exceptional optical and electronic properties. As the search for potential candidates for novel materials continues, understanding the structure of these materials is crucial for investigating their stability. In this study, we implement a framework to find novel material based on a machine-learning model. The machine learning model was trained to predict the dimensionality of the polyhedral network based on the connectivity of the polyhedral network in different directions. The polyhedral connectivity of low-dimensional structures can be classified into three dimensions: 0D, 1D, and 2D. This ML model was trained on 588 unique experimentally reported structures and these structures include halide perovskites as well as related, non-perovskite halometallate structures. These structures contain 65 unique B-X combinations and 315 different organic cations. The ML model achieves a precision of 96%, 87%, and 95% for 0D, 1D, and 2D, respectively for a stratified 20% test set using a graph representation from self-supervised message passing transformer (GROVER) to represent the organic cations and a set of 7 features to represent the inorganic/halide elements. This ML model is then used to fill in the gap of the missing compounds. In total, the ML model is used to predict the dimensionality of 40556 new HOIPs based on the combination of organics, inorganic/halide combinations across different element compositions and B/X ratios. Finally, to identify interesting organic cations that have not yet been examined for HOIPs, we used the GROVER representation and performed a similarity search to select potentially interesting organic cations and predict their corresponding dimensionality.

We screened a set of 6 molecules to find the similarity from two different molecular cases to find novel materials. This organic cation identification is demonstrated on: chiral molecules and large conjugated systems derived from pyrene, thiophenes and cyclopentadiene molecules.

# TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . . .	ii
ABSTRACT . . . . .	iii
LIST OF TABLES . . . . .	vii
LIST OF FIGURES . . . . .	ix
CHAPTER 1 INTRODUCTION . . . . .	1
1.1 Motivations for materials science . . . . .	1
1.2 Motivations for machine learning . . . . .	8
CHAPTER 2 METHODS . . . . .	10
2.1 Dataset generation . . . . .	10
2.2 Feature generation . . . . .	11
2.3 Model Training . . . . .	13
2.4 Evaluation of Machine Learning models . . . . .	14
CHAPTER 3 RESULTS . . . . .	16
3.1 Dimensionality predictions . . . . .	16
3.2 Similarity Search . . . . .	22
3.3 Conclusions and outlook . . . . .	27

BIBLIOGRAPHY . . . . .	28
APPENDIX A SUPPLEMENTARY INFORMATION . . . . .	36

## LIST OF TABLES

Table 1.1	Atomic features used in GROVER model training [41] . . . . .	9
Table 1.2	Bond Features used in GROVER model training.[41] . . . . .	9
Table 2.1	organic molecule features calculated from rdkit package and features calculated using other packages . . . . .	11
Table 2.2	Features generated based on the inorganic ( <i>B</i> ) and halide ( <i>X</i> ) units	12
Table 2.3	Distribution of data in dimensionality. The numeric values in this Table represent the number of samples in different dimensionalities and how they were used in testing and training. . . . .	13
Table 3.1	The precision of the different classification models based on specific feature combinations for the 20% stratified test set according to Table 2.3, in order to ensure the same distribution in the training and test set for every class. The results here are the average of 100 runs. . . . .	17
Table 3.2	Organic inorganic model performance on different inorganic units with multidimensional classes(precision) . . . . .	21
Table 3.3	Predicted dimensionalities of similar cations in reference to selected cations and inorganic units. Rank presents how similar is given cation to the reference. Occurrence shows the number of counts the cation appears in the model training/testing dataset. *Experimental structure contains the predicted dimensionality . .	26
Table A.1	The precisions of the different classification models with different feature combinations for rdkit feature set(rdkit includes the hull volume of the cations) . . . . .	38
Table A.2	The precisions of the different classification models with different feature combinations for the 20% stratified test set for the upsample dataset . . . . .	38



Table A.3	Magpie features used in model training [56]	39
-----------	---	----

## LIST OF FIGURES

Figure 1.1	Different dimensionalities in HOIPS structures (a) 0D structure composed of face sharing octahedrons clusters (b) 1D structure network through edge-sharing and corner-sharing octahedra (c) 2D network of corner-sharing octahedra. (d) 3D network of corner-sharing octahedra . . . . .	1
Figure 3.1	Data distribution of LDA applied to (a) 200 GROVER features (b) 14 RDkit features . . . . .	17
Figure 3.2	Count of BX combination in different dimensionalities . . . . .	18
Figure 3.3	The relation of dimensionality 0D, 1D, and 2D as a function of stoichiometry ratio (SR) is defined as the ratio of the number of X site elements per B site in the given structure. . . . .	19
Figure 3.4	Confusion matrix of the median model of the different feature set in Table 3.1 (a)organic cation (b)inorganic unit features (c) organic cation and inorganic unit features . . . . .	21
Figure 3.5	Ratio between A-site cation and inorganic unit . . . . .	22
Figure 3.6	Probability of each composition to get predicted as a 0D,1D,2D. probability calculated from existing structural data . . . . .	23
Figure 3.7	Variation of the probability of a specific dimensionality in a given BX composition. Three colors red green and blue indicate the dimensionality as 0D, 1D and 2D. The probabilities were scaled and coded with the RGB decimal code to indicate the probability of each dimensionality in a given composition (a)Distribution of 0D probabilities in the predicted dataset. (b) Distribution of 1D probabilities in the predicted dataset. (C) Distribution of 2D probabilities in the predicted dataset. . . . .	24
Figure A.1	Similarity search result with rdkit features without normalizing the features . . . . .	36

Figure A.2	Similarity search result with rdkit features with standardized features . . . . .	36
Figure A.3	Similarity search result with rdkit features with min-max normalized features . . . . .	37
Figure A.4	Similarity search result with grover features . . . . .	37
Figure A.5	Feature importance calculated using SHAP values for two models inorganic only model and inorganic+organic model . . . . .	38

# CHAPTER 1

## INTRODUCTION

### 1.1 MOTIVATIONS FOR MATERIALS SCIENCE

Due to the high structural tunability, hybrid organic-inorganic perovskite materials (HOIPs) have gained significant attention over the past decade as suitable candidates for optoelectronic and photovoltaic applications. In general, these halide perovskites have different compositions such as  $ABX_3$ ,  $A'_2A_{n-1}B_nX_{3n+1}$  or  $A'A_{n-1}B_nX_{3n+1}$ . In compositions where A and A' is a cation, B is an inorganic cation and X is a halide for example A= methylammonium(MA), B-site is an inorganic cation such as  $Pb^{2+}$  and X is a halide( $X=F^-$ ,  $Cl^-$ ,  $Br^-$ ,  $I^-$ ). These different compositions have different phases, these generalized compositions vary with the change of phases such as Ruddlesden-Popper and Dion-Jacobson.

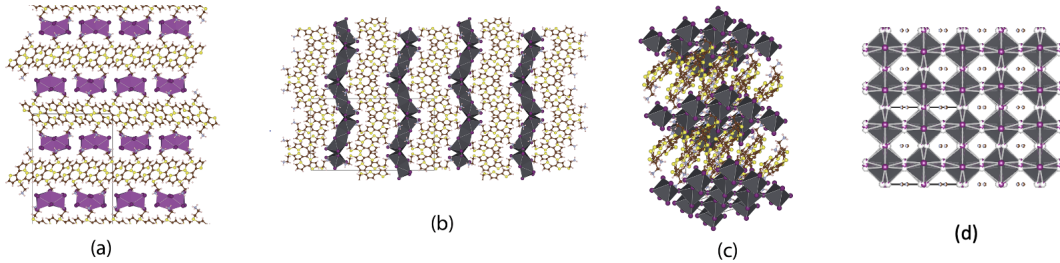


Figure 1.1 Different dimensionalities in HOIPS structures (a) 0D structure composed of face sharing octahedrons clusters (b) 1D structure network through edge-sharing and corner-sharing octahedra (c) 2D network of corner-sharing octahedra. (d) 3D network of corner-sharing octahedra

In general, these structures possess special arrangements in B and X sites. The B-site of the perovskite structure is surrounded by halides in a six-fold coordina-

tion environment to make an octahedron. These octahedral structures have different connectivity types such as corner-sharing, edge-sharing, and face-sharing. In these connectivities, halide atoms are shared between octahedron units and these octahedra share different connectivity to create an octahedral network. These octahedral networks in the structure govern the dimensionality based on the model of sharing. Basically, there are four types of dimensionality, namely 0D,1D,2D, and 3D, Figure 1.1 shows different dimensionalities observed in perovskite structures. The 0D,1D, and 2D materials are referred to as low-dimensional structures. These structures have extraordinary electronic and optical properties which make them suitable for optoelectronic and photovoltaic applications, by changing the structural dimensionality these materials can be tuned to be used in various applications. The unique properties of 2D halide perovskites have enabled a variety of applications, such as LEDs, lasers, and PVs [1],[2],[3]. The high dielectric constant between organic and inorganic moieties in the structure leads to higher electron-hole binding energy (Eb) in 2D perovskites and these excitons can be stabilized by the multiple quantum well structure in ambient temperatures [4]. This 2D quantum well effect drives possibilities to use these materials as optoelectronic applications. The reduction of structural dimensionality results in an increase in the bandgaps of low-dimensional structures. This increase in bandgap in 0D and 1D material makes them less favorable for photovoltaic applications [5], [6]. The 1D and 0D structures show large stokes shift in the range of 100-200(nm) for 1D and 100-350(nm) [6] for 0D and also show broadband emission spectrum, which makes them advantageous in down conversion white LED applications as phosphors [7]. The strong quantum confinement effect in 0D structures leads to higher photoluminescence quantum yield(PLQY), Sun et al. reported the highest efficient blue-violet (392 nm) light emission of 0D [BAPrEDA]PbCl<sub>6</sub> · (H<sub>2</sub>O)<sub>2</sub> with a PLQY of 21.3% [8]. Also, these materials are candidates for optical transistors, switching devices, coherent light sources and lasing [9],[10], [11], [12]

Another significant factor regarding low-dimensional structures is that they possess tunable bandgaps which is an electronic property of this material. In determining the electronic properties in a perovskite structure, the arrangement of the  $\text{BX}_6$  octahedral unit plays a crucial role. Strong interaction between the metal cation and halide anion leads to the formation of electronic bands. The bandgap of these perovskites is mainly determined by the atomic orbitals on B and X. When the dimensionality decreases from 3D to 0D, the bandgap increases, and bandwidth is reduced [13]. This is due to the fact that the lattice constant increases leading to the localization of electrons. The increase in lattice constant is a result of a reduced dimensionality which increases the distance between octahedral units, leading to a weaker intermolecular overlap between octahedral units. This weaker overlap contributes to the reduced bandwidth[14]. Another factor that determines the bandgap is the octahedral connectivity as we talked about earlier in this thesis which leads to different dimensionalities. Kamminga et al. in their study demonstrates the correlation between the connectivity of the octahedral unit and the change in the bandgap. In the study, DFT calculations were used with and without spin-orbit coupling(SOC). The results show that the bandgap increases with decreasing dimensionality. Moreover, as the connectivity between octahedra varies from corner-sharing - to edge-sharing- to face-sharing, the bandgap increases. These trends were held for calculations with and without SOC [15]. In one of their works, Li-Ming Wu et al. generalizes the structure-property relationship of iodoplumbate and iodobismuthate compounds, focusing on how the connectivity of the inorganic units influences the inorganic unit composition and correlates inorganic unit compositions to dimensionality [16]. Lermer et al. as an extension to one of their research on lead halide perovskites, the influence of dimensionality on the band gap was explored. They compared the band gaps of 1D and 2D structures. The 1D structure exhibited a band gap of 2.44 eV, while the 2D structure showed a lower band gap of 1.99 eV. This demonstrates that

the dimensionality of the perovskite structure plays a significant role in determining the band gap values. Therefore it can be stated that dimensionality is an important parameter in designing novel materials[17].

The bandgap depends on several other factors such as the choice of B and X site and BX composition[15][18][19] [20] [21][22]. The X-site contributes to the valence band in the band structure[23] and the B-site contributes to the conduction band in band structure [24]. The observed bandgaps for materials change with the halide substitution; from Cl to Br to I the valence band composition changes from 3p to 4p to 5p this reduces the band gap due to the monotonic decrease in binding energy(low ionization potential). When considering the ionic radii of halide atoms, large halide ions(iodide) result in a weaker bonding between the halide ion and the B-site ion giving rise to a smaller band gap compared to larger bandgaps caused by smaller halide ion [25]. For example, the substitution of Br by I in FAPbX<sub>3</sub> reduces the band gap from 2.23 eV to 1.48 eV [26]. Lermer et al. conducted experimental research on benzimidazolium lead halide perovskites, investigating the impact of halide substitution and dimensionality on the band gap [17]. Their findings reveal a systematic decrease in the band gap as the halide substitution changes from chloride to bromide to iodide. Specifically, the band gap values were found to be 3.08 eV for chloride, 2.60 eV for bromide, and 1.99 eV for iodide.

Oligothiophenes and chiral molecules are two distinct organic molecules, each offering unique characteristics. Oligothiophenes are intriguing due to their structural tunability, allowing for various modifications and adaptations. On the other hand, chiral molecules possess optical activity. These organic molecules are investigated to be used in perovskite structures due to their optoelectronic properties. A study by Denis et al. shows the ability to modify the optoelectronic properties of oligothiophenes by functionalizing and varying the conjugation length of the oligomer, it also shows substitution of a small amount of iodide with bromide reduced the presence

of lower-dimensional hybrids in thin films made out of  $(\text{Bit-C3})_2\text{PbX}_4$  (with  $\text{X} = \text{Cl}, \text{Br}, \text{and I}$ ) material[27]. The low-dimensional chiral hybrid perovskite which incorporates a chiral ligand in the A-site, exhibited both spin-polarized absorption and spin-polarized photoluminescence, even in the absence of an external magnetic field. These properties make these structures more suitable for applications in circularly polarized electronics, photonics, and spintronics [28]. Fu et al shows in their study how to use chiral 1D hybrid perovskite  $[(\text{R/S})\text{-3-aminopiperidine}]\text{PbI}_4$  to distinguish circularly polarized light(CPL) which directly contributes to the optical activity of the material [29]. Using theoretical calculation on  $\text{CHFCINH}_3\text{PbI}_3$  Long et al shows the thermal stability of (R) and (S) enantiomers are similar and also exhibit similar bandgap values. When compared to (RS)- $\text{CHFCINH}_3\text{PbI}_3$  the thermal stability is similar. However, the achiral molecule shows a slightly lower bandgap 1.24 eV, and the chiral molecules process 1.51 eV. This shows that the bandgap of a material can be tuned by substituting a chiral molecule. This gives the ability to tune the bandgap without changing the material structure and properties [30].

The development of computational techniques has increased the use of these techniques to investigate the underlying chemistry in different chemical systems. This opens up new ventures in in-silico material discovery. High-throughput calculations, machine learning, and structure prediction-based calculations provide leads for the synthesis of new materials. The computer-aided material discovery is useful in finding potential materials for a particular task by rationalizing the experimental processes. The use of computer-aided frameworks has eased the burden of expensive material costs and has saved time in the synthesis and optimization of properties. As a result, the successful synthesis of new materials over the past decades has increased and can be seen in the compilation of material databases. Machine learning and deep learning have been used in discovering novel materials and finding relations between different properties of materials. Development in machine learning and the availability of



data sets improve the accuracy of computer-aided material discovery. The work of Lu et al. builds a framework for identifying stable lead-free hybrid organic-inorganic perovskites using machine learning [31]. In this study, a supervised machine learning method was utilized to generate a list of potential hybrid perovskite candidates, and these new structures were screened based on their thermal stability to find potential materials. In the study, they used DFT calculation on 346 HOIPs in training the machine learning algorithm. They successfully identified six lead-free HOIPs with proper bandgaps for solar cells and thermal stability out of 5158 unexplored HOIPs structures. Ai et al. designed a framework to predict the dimensionality in templated oxides to understand the formation principles of templated oxides [32]. They tried to articulate the formation of templated oxides by predicting the dimensionality by training a neural network model on 3725 structures. They predicted the dimensionality with an accuracy of 71% using the reactant identities.

While hybrid perovskites have demonstrated impressive performance, they have few drawbacks for their use in large-scale commercial devices. The most common are toxicity (in Pb-based perovskites) and poor stability under environmental conditions. Exploration of new materials is required in finding solutions to these issues while maintaining the device’s performance. In this regard, HOIPs have emerged as a class of materials that offer a variety of possibilities for modification. In modification and fine-tuning the properties of perovskites. The structural confinement of 3D structures limits the property tunability, while the tunability is increased with an extra degree of freedom gained by low-dimensional structures [33] [34]. Therefore low dimensional HOIPs have huge diversity, which allows the incorporation of a wide range of organic materials into their structures. This versatility enables the exploration of different compositions and structures for designing and tailoring HOIPs to meet specific application requirements [35] [36] [37] [38].

This thesis describes the building of a framework to discover novel HOIP materials

to be used in different applications based on their composition in A-site, B-Site, and X-site. As a first step, a dimensionality-predicting machine learning model to screen novel materials was built to limit the search space before conducting high throughput Density Functional Theory(DFT) calculations. Since the band gap of materials are significantly impacted by the dimensionality, understanding the dimensionality of a novel material provides a preliminary understanding of the potential range of band structure and bandgap values. This saves and prioritizes the computational work on some potential materials over other materials.

## 1.2 MOTIVATIONS FOR MACHINE LEARNING

Machine Learning (ML) methods can generally be categorized into two types: supervised and unsupervised learning. In supervised learning, the ML algorithm is provided with labeled input data for training the model and testing. The algorithm’s task is to learn the underlying function that best represents the relationship between inputs and outputs in a generalized way. These algorithms are commonly used for classification and regression tasks, resulting in trained models capable of making predictions on new inputs. In unsupervised learning, the algorithm learns patterns and structures without any guidance from labeled data to discover hidden patterns and relationships within the data. The advantage of using machine learning is to be able to use a large amount of data to infer relationships without detailed knowledge about the problem.

An essential part in the process of training a precise machine learning model is the formulation of a set of input features or representations, which are used in the regression step. Although several known representations, including SOAP [39], MBTR [40], and others, are available for use, these methods compile a set of numerical values related to specific properties using an experimental structure. Finding a structure for a novel material requires a significant amount of experimental and computational work. In order to simplify this work, representation learning can be used to learn about and represent chemical systems. Representation learning is the use of computational models to extract the properties of different systems and use these properties to generate a representation of the corresponding system. In supervised learning, the training data needs to be labeled. In general, the outcome of the representation learning model is biased towards the input dataset. In order to resolve the biases toward the input dataset, representation learning can be unsupervised or self-supervised which requires unlabeled data [41]. This study was conducted with the aim of designing novel materials, by unbiased feature representation of organic cations. In order to achieve this

objective, GROVER (Graph Representation frOm self-superVised mEessage passing tRansformer), a self-supervised graph neural network was utilized.

GROVER uses Mask Language Modelling for training on unlabeled 10 Million molecules and thus the generated embedding of this model is more generalized [41]. Because, this model is trained on a big corpus of molecules and the representations become very useful for different downstream classification and regression problems. GROVER takes the SMILES string as input and returns 200 numerical values to represent the SMILES string. These values were used in representing the A-site cations. GROVER generates its features from two methods. 1) Node/edge feature extraction, which uses rdkit to extract features from atoms and bonds Table 1.1,1.2. 2) Molecular-level feature extraction. The molecular level feature extraction follows the works by Kevin Yang et al. and Zhenqin Wu et al. and produces 200 molecular-level features using Rdkit for each molecule [42] [43]. These extracted features were concatenated with the output of the self-attentive readout to go through a multi-layer perception (MLP) for the final prediction of 200 GROVER features(embeddings of the GNN).

Table 1.1 Atomic features used in GROVER model training [41]

Features	size	Description
Atom type	100	Types of atoms (e.g., C, N, O) by atomic number
Formal charge	5	Integer electronic charge assigned to atom
Number of bonds	6	Number of bonds the atom is involved in
Chirality	5	Number of bonded hydrogen atoms
Number of H	5	Number of bonded hydrogen atoms
Atomic mass	1	Mass of the atom, divided by 100
Aromaticity	1	Whether this atom is part of an aromatic system
Hybridization	5	sp, sp <sup>2</sup> ,sp <sup>3</sup> ,sp <sup>3d</sup> or sp <sup>3d2</sup>

Table 1.2 Bond Features used in GROVER model training.[41]

Features	size	Description
Bond type	4	Single, double, triple, or aromatic
Stereo	6	None, any, E/Z or cis/trans
In ring	1	Whether the bond is a part of a ring
conjugated	1	Whether the bond is conjugated

## CHAPTER 2

### METHODS

#### 2.1 DATASET GENERATION

As the first step, we compiled a total of 944 hybrid perovskites from three sources: the 2D perovskites database of the laboratory of new materials for solar energy (NMSE) [44] (725 samples, queried on 12/2021); Cambridge Structural Database (203 samples, queried on 03/2022); and research article by Tremblay *et al.* [45] (16 samples) were compiled. Since the 2D perovskites dominate the NMSE database, the CSD database was mostly queried for searching lower dimensional perovskites (1D and 0D) and includes 86 samples reported in Ref [46] and 5 reported in Ref [47]. From the above collection of structures, 35 structures were excluded as the organic cation was unrecognizable due to high disorder in the structure. The remaining 909 samples, we labeled based on the dimensionality of the octahedra as 0D, 1D, 2D, or 3D. From this collection, 23 samples were removed due to the following reasons: (a) had only inorganic cations (12 samples), (b) had 3D connectivity of the octahedra (8 samples), (c) had no octahedra (2 samples) and (d) without halogen atom at the  $X$ -site (1 sample). Further 298 structures out of the remaining 886 samples were removed as the organic and inorganic units and dimensionality were duplicated in the dataset. This contracted the dataset size to 588 unique samples. For validation purposes and to investigate how the ML model generalizes, an additional 37 samples from the HybriD<sup>3</sup> materials database [48] were also extracted but never used during the initial model training/testing.

## 2.2 FEATURE GENERATION

The utilization of ML models in screening new materials can be limited if they depend on having a specific structure as input data to make a prediction. To address this, distinct features have been incorporated from both organic and inorganic constituents. This approach essentially allows an independent assessment of materials using the properties of these units, which can then be combined in an arbitrary manner. The organic cations were represented by the GROVER package not adhering to a specific structure. [41].

To compare with the predictions from GROVER, we also represented the organic molecule using basic molecular descriptors from Rdkit, an open-source cheminformatics package in Python. We chose thirteen molecular features from Rdkit, along with an additional feature representing cation hull volume. The fourteen features derived from this second feature set are listed in Table 2.1. Out of the 588 unique structures in this dataset, the number of different organic cations species incorporated on the *A*-site ( $N_A$ ) varies from  $N_A = 1$ -3. For structures with  $N_A > 1$  on the *A*-site, cation features were averaged.

Table 2.1 organic molecule features calculated from rdkit package and features calculated using other packages

Type	Label	Definition	Unit	Reference
organic molecule features	TPSA	Topological Polar Surface Area	–	[49]
	ASA	Accessible Surface Area	–	[50]
	MolWt	Molecular weight	–	–
	MollogP	Molecular log P	–	[51]
	ecc	Eccentricity	–	[52]
	rgy	Radius of gyration	Å	[52]
	isf	Inertial shape factor	–	[53]
	asp	Asphericity	–	[54]
	SphIndex	Sphericity Index	–	[53]
	aring	Number of aromatic ring	–	–
	ring	Number of ring structures	–	–
	hacceptor	Number of H-bond acceptors for a molecule	–	–
	sp3	The fraction of C atoms that are SP3 hybridized	–	–
	hull_vol	Hull volume of the organic cation	–	–

The descriptors of the organic component were combined with the features derived from the combination of the inorganic metal (*B*-site)/halide (*X*-site), seven features

were used, which are listed in Table 2.2, such as the octahedral factor [55], the average volume of the inorganic unit, and 5 of the highest ranked features out of 47 from Matminer [56]. The octahedral factor measures the ability of the  $B$ -site cation to form stable octahedra with the  $X$ -site anion and it is given by the equation:

$$\mu = \left(\frac{r_b}{r_x}\right) \quad (2.1)$$

where  $r_b$  denotes the radius of the  $B$ -site, and  $r_x$  denotes for the radius of the  $X$ -site. The Shannon radii values were used to calculate  $\mu$ . [57]

The averaged volume for the inorganic layer of each compound in the dataset is given by this expression:

$$V = \left(\frac{\sum C_i \times V_i}{\sum C_i}\right) \quad (2.2)$$

The atomic volume used in this calculation was extracted from Mendeleev (version v0.12.1). In the equation 2.2  $C$  is the number of atoms and  $V_i$  is the atomic volume of atom  $i$ . To augment these two features, which were generated using domain knowledge, features extracted from the Magpie library in Matminer were used. Initially, 132 features were selected from Matminer [56], which were down-selected to 47 based on 0D and 1D precision. After evaluating the features based on mean precision of 0D and 1D for each model, 47 features were initially selected to be used as inorganic features, then the top 5 features were selected using SHapley Additive exPlanations (SHAP) [58] to use as inorganic features (see Table 2.2).

Table 2.2 Features generated based on the inorganic ( $B$ ) and halide ( $X$ ) units

Type	Label	Definition	Range across 588 unique structures	Reference
inorganic unit features	$\mu$	Octahedral factor	0.345-0.657	2.1
	$V$	mean atomic volume for $B$ and $X$ site elements	16.3-25.2	2.2
	avg_dev Column		0.39-3.2	[56]
	mean NpUnfilled		0.7-2.0	[56]
	mean NpValence		3.5-4.8	[56]
	mean NValence		7-21	[56]
	mean AtomicWeight		39- 154	[56]

Table 2.3 Distribution of data in dimensionality. The numeric values in this Table represent the number of samples in different dimensionalities and how they were used in testing and training.

Dimensionality	After filtration	Baseline		Upsampling	
		Test	Train	Test	Train
0D	86	17	69	17	350
1D	64	13	51	13	350
2D	438	88	350	88	350
Total	588	118	470	118	1050

### 2.3 MODEL TRAINING

Compounds in the dataset were then labeled as 2D (438/588 samples), 1D (64/588) or 0D (86/588) (see Table 2.3). Because the dataset is significantly imbalanced where 2D entries have nearly a factor of 3 more frequent than either 1D or 0D samples, the dataset was stratified such that each dimensionality class is represented by at least 20% of the compounds. As a result, out of 86-0D samples in the total dataset of 588 samples, stratifying ensures that a randomly selected 17-0D samples are included in the test set. The training set therefore has 350-2Ds, 51-1Ds, and 69-0Ds (the ML model trained to this set is referred to as the baseline model). Stratification of the dataset, however, does not solve the issue of under-representation of 1D and 0D in the training set. To assess the potential effect of this bias in the training set on model accuracy, an additional analysis where the training data set was up sampled was conducted. This was done to ensure equal representation of each class. In this up-sampling process, instances from the 0D and 1D classes were randomly replicated until the count for each class reached 350, achieving a balanced training set.

In this work, a random forest classifier from the scikit-learn library was used [59]. In training, 100 models were trained with 100 different splits of data with hyper-parameters tuning. The number of estimators and the maximum depth of the tree were used as hyper-parameters and both were tuned from a range of 10 and 100 with an interval of 20. The average of 100 runs was used in evaluating the model’s



performance. In order to carry out further analysis and downstream tasks, the median model was selected based on averaged precision of 0D and 1D for each run.

## 2.4 EVALUATION OF MACHINE LEARNING MODELS

The evaluation of machine learning algorithms plays a crucial role in assessing their performance. However, it can be challenging, especially in scenarios where limited or no access to real-world data exists. In such cases, additional human effort is required to assess the model performance. In classification tasks, the evaluation is typically carried out by splitting the dataset into a training set and a test set. The machine learning algorithm is trained on the training set, while the test set is used to calculate performance indicators that assess the performance of the model. One common challenge faced by machine learning algorithms is the availability of limited training and test data. This can impact the algorithm’s generalization capabilities and lead to overfitting or underfitting. It’s crucial to have enough data for training and testing to have a good machine-learning model. Evaluating the performance of a machine learning model involves considering multiple factors. While there is no perfect indicator applicable to every scenario, several important factors are considered. These factors include:

- Accuracy: Measures the proportion of correctly classified instances, providing an overall assessment of the algorithm’s performance.
- Precision: Evaluates the algorithm’s ability to correctly identify positive instances within a given class, indicating its effectiveness in minimizing false positives.
- Recall: Assesses the algorithm’s ability to identify all positive instances within a given class, indicating its effectiveness in minimizing false negatives.

- F1 score: Combines precision and recall into a single metric, providing a balanced measure of a classifier's performance.

It is important to select appropriate evaluation metrics based on the specific problem domain and objectives of the machine learning algorithm. The choice of evaluation metrics should align with the desired outcomes and provide meaningful insights into the algorithm's performance. The performance of the machine learning model, could be compromised due to the issues such as class imbalance, for example, if the data set is dominated by the class-2D over the classes 0D and 1D. In this case, it is more advantageous to have a model that is able to predict the positive instances for each of those classes with high accuracy rather than using a metric that assesses overall predictions. Accuracy measures the overall correctness of the predictions, which is not a suitable performance metric when class distribution is imbalanced. In recall the score is calculated on the true positives and false negatives values. False negatives describe the ability of the model to identify the positive instances correctly. Recall is not crucial for this type of classification but may be important for certain scenarios such as medical diagnosis, where false negative class is more important. In this classification model, the focus is on achieving accurate predictions for each class that best match the definitions of precision where incorrectly predicted positive instance gives a significant reduction in the model performance. Therefore, precision over recall and accuracy were selected as performance metrics.

## CHAPTER 3

### RESULTS

#### 3.1 DIMENSIONALITY PREDICTIONS

Dimensionality prediction of perovskite structures with varying organic cations was investigated using Redkit generated features and GROVER generated features. The Rdkit features have a precision of 0.60, 0.63, 0.84 for 0D, 1D, and 2D, respectively, compared to GROVER features 0.67, 0.60, 0.86 for 0D, 1D, and 2D (see Table A.1). The performance values are similar and this makes it much harder to choose one feature set over the other. However, the Linear Discriminant Analysis (LDA) plots show significant differences in the two feature sets for the class distribution Figure 3.1. Figure 3.1 demonstrates how the dimensionalities of structures are distributed in a reduced feature space. In the purpose of reducing features, each feature set was reconstructed to two features, so that it could be visualized in 2D space. This gives an intuition about the higher dimensional data distribution. In Figure 3.1 class separation is clearly visible with the GROVER features with lesser overlaps between classes. This shows that GROVER has sufficient discriminatory power in classifying the dimensionality. In the Rdkit feature set, the classes are not well separated and there is a significant amount of overlap between classes. Therefore, GROVER was used to analyze organic cation features.

In the analysis of individual features from the inorganic unit feature set, it was found that the octahedral factor alone could not be relied upon for the prediction of dimensionality of 1D structures. In fact, the model’s performance in predicting

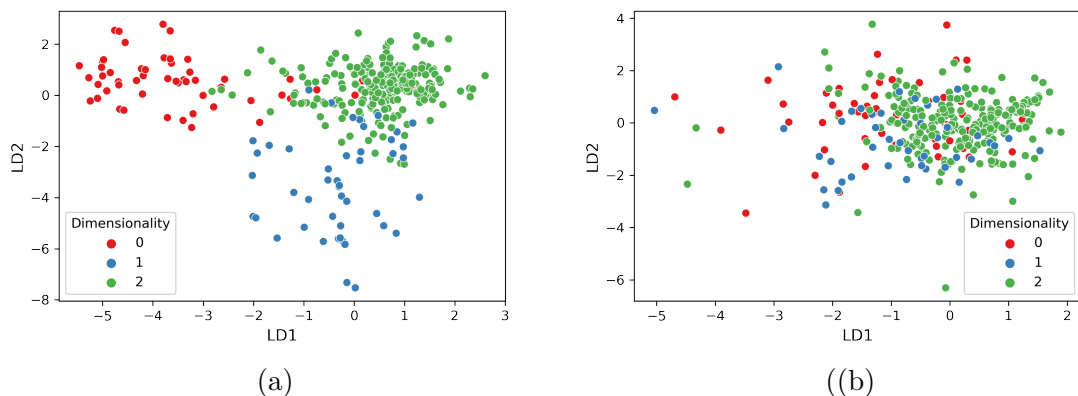


Figure 3.1 Data distribution of LDA applied to (a) 200 GROVER features (b) 14 RDkit features

1D structures consistently yielded a score of 0 across 100 runs Table A.1. This suggests that the octahedral factor does not possess sufficient discriminatory power to differentiate 1D structures from other dimensionalities accurately. Figure 3.2 shows the structural distribution for BX combinations from the dataset. 2D structures are more frequent than other dimensions in all the BX combinations except for SnBr which gave a majority class from 0D structures in one of the cases. 1D class is a major class that does not have a significant number of structures compared to 2D class. This leads the model to always predict 1D structures as another class which yields a score of 0 across all runs when using the octahedral factor.

Table 3.1 The precision of the different classification models based on specific feature combinations for the 20% stratified test set according to Table 2.3, in order to ensure the same distribution in the training and test set for every class. The results here are the average of 100 runs.

Data with different features	No of Features	0D	1D	2D
Organic cations	200	0.67(0.12)	0.60(0.17)	0.86(0.02)
Inorganic features	7	0.96(0.05)	0.94(0.08)	0.94(0.02)
organic cation + Inorganic features	207	0.96(0.04)	0.87(0.09)	0.95(0.02)

When using 7 inorganic unit features, (See Table 2.2) precision  $> 0.94$  was achieved across 0D, 1D, and 2D represented in Table 3.1. The relatively high precision of the ML model trained on inorganic features alone is potentially due to our dataset having a strong dependence of stoichiometry combinations of *B*-site cations and *X*-

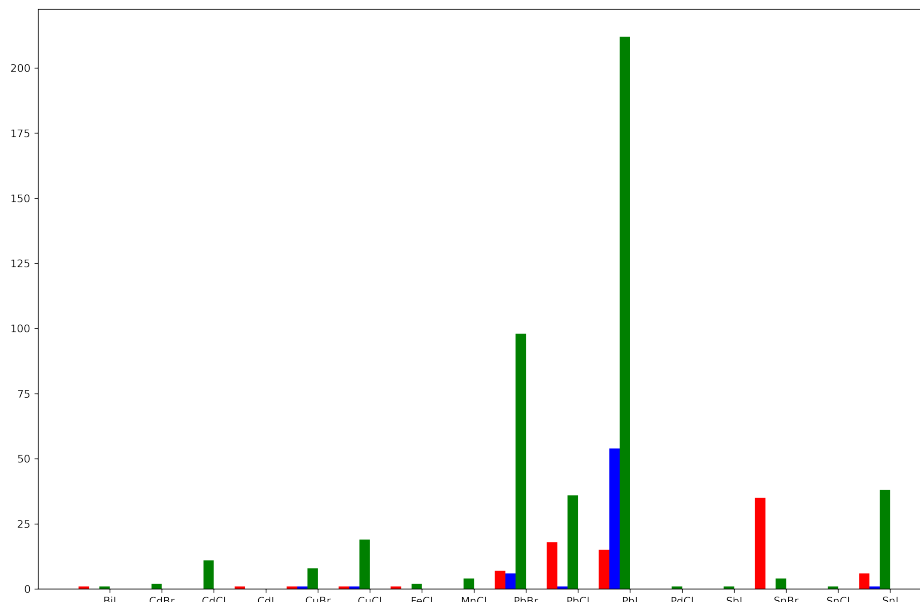


Figure 3.2 Count of BX combination in different dimensionalities

site halides with dimensionality class. For example,  $\text{PbI}_6$  and  $\text{PbBr}_6$  are always founded to be in 0Ds and  $\text{PbI}_3$  is always associated with 1Ds. Indeed, a pronounced separation of the dimensionality is founded with the stoichiometry ratio (SR); defined as the ratio of the number of B-site metals to the X-site halogen atoms represented in inorganic units (Fig 3.3). For  $\text{SR}=6$ , 100% of the perovskites in our data set are 0Ds while no perovskites with  $\text{SR}<4$  have a 0D connectivity. Similarly, 1D are more common for  $\text{SR}=3$  while 2Ds are more common for  $\text{SR}=4$ . However, there are some compositions such as  $\text{PbI}_4$  ( $\text{SR}=4$ ) which included several different dimensionalities. Similarly,  $\text{Pb}_3\text{I}_{10}$  ( $\text{SR}=3.33$ ) is common in both 1D and 2D perovskites (Fig 3.3).

In contrast, when training a machine learning model solely using organic cation features and excluding the inorganic unit, the resulting model demonstrates poor performance. The precision for this model related to 0D, 1D, and 2D structures are 0.67, 0.60, and 0.86 respectively in training with organic unit features (Table 3.1). These

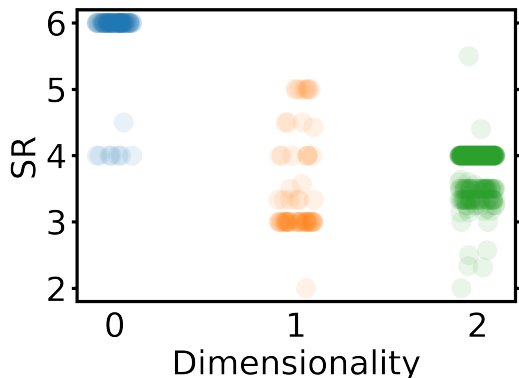


Figure 3.3 The relation of dimensionality 0D, 1D, and 2D as a function of stoichiometry ratio (SR) is defined as the ratio of the number of X site elements per B site in the given structure.

precisions indicate that relying solely on organic cation features is insufficient for the prediction of the dimensionality of HOIPs accurately. When visualizing the organic features using LDA, it was observed that the well-separated dimensionality classes Figure 3.1, indicates that organic features carry useful information for differentiating the dimensionalities. This suggests that use of all the organic cation features introduces some redundancy to the model. Also it can be concluded that all the features do not contribute equally to the prediction task. Figure A.5 shows the importance of the organic and inorganic features for the median model of the inorganic+organic. The inorganic features have significant importance compared to organic features. The top 20 features have the highest effect on the 2D class and the least effect on the 1D class.

When the model was trained by combining organic and inorganic unit features, the model performance increases especially for 0D and 2D classes. Then the resultant precision of 0D, 1D, and 2D, were 0.96, 0.87, and 0.95 respectively. These results show that the inclusion of organic features increases the model performance. As observed from the LDA plots, GROVER features have important information which increases the model performance even though it cannot absolutely predict the dimensionality.

Although upsampling was used to address the class imbalance, performances were

0.53, 0.40, and 0.88 for 0D, 1D and 2D respectively. When organic cation features were used, this was increased to 0.73, 0.61, and 0.93 and when inorganic cation features were incorporated the performance was improved to 0.87,0.77,0.94 for the dimensionalities 0D,1D and 2D respectively.(see Table A.2). The precision was decreased in all the upsampled datasets compared to its baseline dataset.

The confusion matrix in figure 3.4a reveals that the organic features struggle in predicting 0D and 1D structures. Specifically, it misclassifies 8 0D structures as 2D and 7 1D structures as 2D. The model does not have significant issues in differentiating between 0D and 1D but with 2D. This misclassification decreases when the feature set changes to inorganic unit features. Misclassification of 0D as 2D drops to 2 and 1D and 2D to 3. There is a significant improvement in 0D predictions, from 8 correct predictions to 15 correct predictions. combining both feature sets results in an improvement in the number of correct predictions and a reduction in incorrect predictions across each class. For 0D, 1D, and 2D the number of correct predictions increases to 15/17, 10/13, 86/88 samples (3.4c (c)). The combination of features increases the accuracy and reliability of the trained models. When comparing the inorganic model and combined model no significant improvement was observed between the two models. The misclassification between 1D and 2D occurs specifically in materials with a particular inorganic unit composition which usually has a single connectivity type or exhibits multiple connectivity types, such as edge-sharing and corner-sharing. This complexity in connectivity can lead to misclassifications between 1D and 2D structures.

To further investigate dimensionality prediction the organic-inorganic model was employed to predict the dimensionality for the dataset. The individual inorganic units within the dataset were examined, specifically focusing on those that exhibit different dimensionalities 3.2. The 0D inorganic units  $\text{CuCl}_4$ ,  $\text{CuBr}_4$  and  $\text{FeCl}_4$  were observed to have tetrahedral inorganic units rather than octahedra inorganic units.

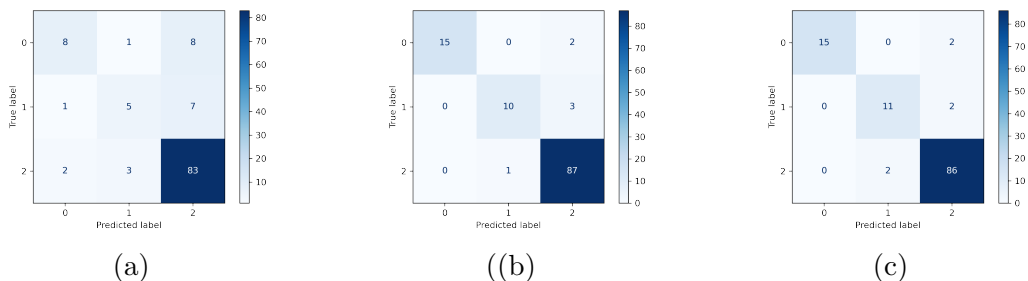


Figure 3.4 Confusion matrix of the median model of the different feature set in Table 3.1 (a)organic cation (b)inorganic unit features (c) organic cation and inorganic unit features

Certain structures from these inorganic units exhibit elongated bonds, which can be attributed to the Jahn-Teller effect. This causes distortions in the coordination geometry. However when a bond is extended beyond the cutoff radius(  $\text{CuCl}=2.75295\text{\AA}$ ,  $\text{FeCl}= 2.86293\text{\AA}$ and  $\text{CuBr}= 2.89295\text{\AA}$ ), it is defined that the octahedra always resulted in a polyhedron with a lower coordination number than an octahedra.

Table 3.2 Organic inorganic model performance on different inorganic units with multidimensional classes(precision)

Inorganic unit	0D	1D	2D
CuCl <sub>4</sub>	0.0	1.0	0.95
CuBr <sub>4</sub>	0.0	1.0	0.89
PbBr <sub>4</sub>	-	0.5	0.99
Pb <sub>2</sub> I <sub>7</sub>	-	0.0	0.96
PbI <sub>4</sub>	-	1.0	1.0
FeCl <sub>4</sub>	1.0	-	1.0
Pb <sub>3</sub> I <sub>10</sub>	-	1.0	0.96

To study the B:X effect on the dimensionality, the probability map was created and is shown in Figure 3.6. In the probability map, the BX combinations in the data set are plotted against the B/X percentage, each cell color and color intensity shows the most probable dimensionality of a given composition. As the octahedral factor increases and the BX percentage decreases, the likelihood of obtaining lower-dimensional structures would be increased. The probability map was calculated from experimental data where probabilities were not available for some compositions. This



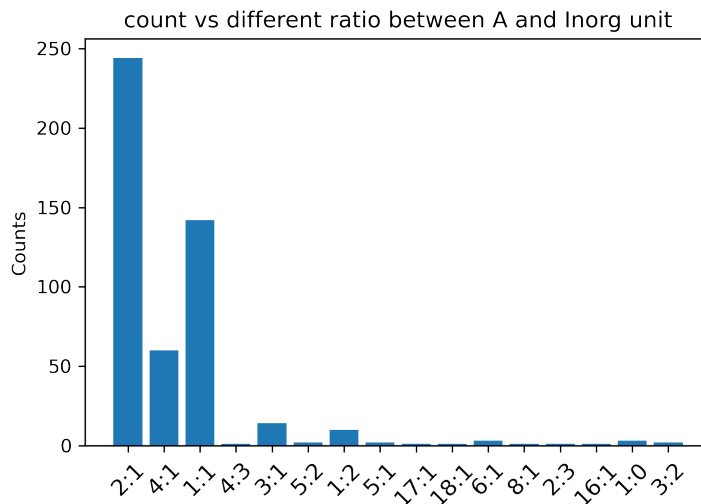


Figure 3.5 Ratio between A-site cation and inorganic unit

makes it challenging to generalize the relationship between dimensionality and composition of the inorganic unit. In order to complete the missing gaps in the probability map it is necessary to gather additional data. Therefore, a new approach was utilized to map the probability of dimensionality for different compositions by using a trained model. The dataset containing 310 charged cations, 16 BX Element combinations, and 42 different B/X compositions were screened and filtered to get 40556 structures. In this process, new materials were filtered based on several factors. These factors included the structure being a single perovskite and considering the most abundant A-site to inorganic unit ratios, which were determined as 1:1, 2:1, 4:1 (see Figure 3.5.

### 3.2 SIMILARITY SEARCH

The aim of the similarity search is to find potential cations that can be used in replacing an existing cation in a perovskite structure. The rationale behind selecting a similar cation is based on the expectation that both cations will exhibit similar properties, this can be identified as A-site engineering. In finding potential cations, similarity between cations was evaluated based on the structural similarity, in order to assess the similarity as a first step GROVER features(cation features ) were used to

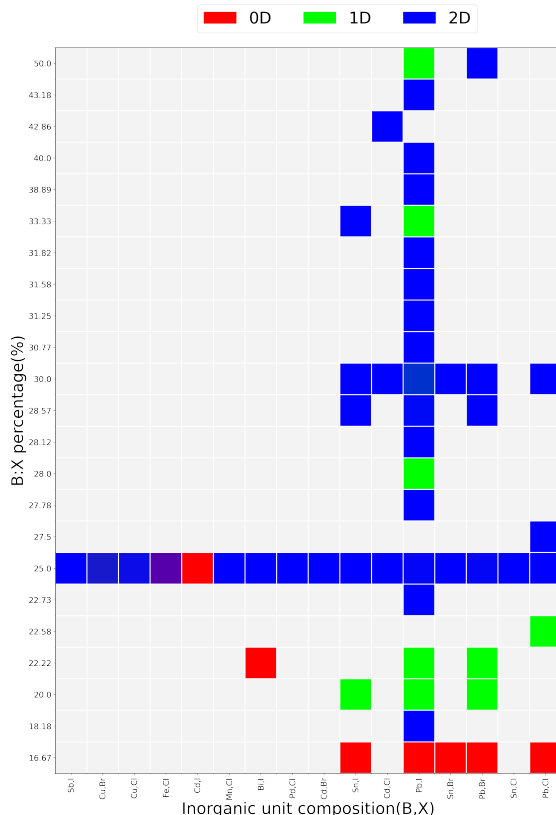


Figure 3.6 Probability of each composition to get predicted as a 0D,1D,2D. probability calculated from existing structural data

represent the structural features of the molecules. GROVER features were generated for the dataset of 17.5 million molecules [60]. The SMILES strings for the dataset were extracted from zinc12, ChEMBL and unique cations in HOIPs dataset. All the smiles were combined together and invalid molecules were filtered by converting the smiles to canonical smiles. Since the GROVER features were generated using a GNN, these features were already normalized. Therefore another normalization was not essential. Figure A.1,A.2,A.3 and A.4 shows the comparison of similarity search results with and without normalization and how it differs from one normalization to another normalization.

The similarity between the two molecules’ feature sets was calculated using different distance matrices such as L1, L2, and cosine. Finding a suitable matrix that calculates the best similarity was based on the performance of the following chemical

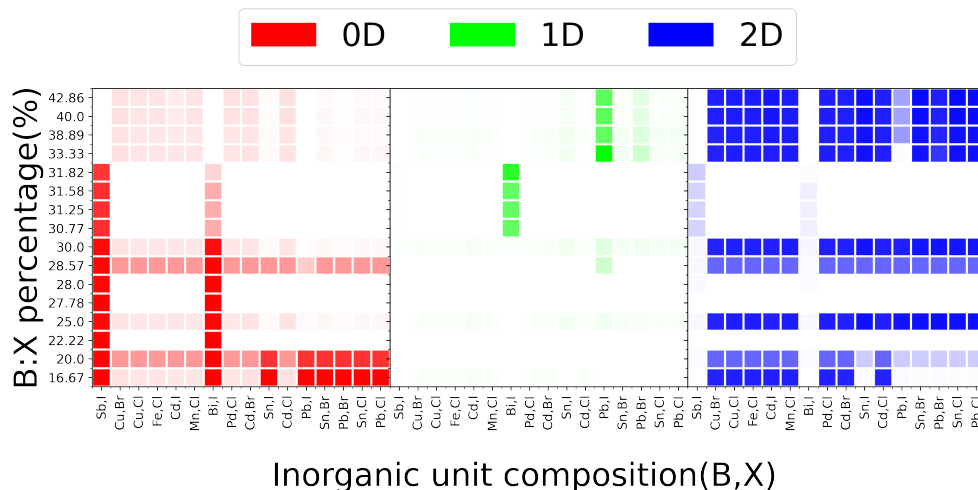


Figure 3.7 Variation of the probability of a specific dimensionality in a given BX composition. Three colors red green and blue indicate the dimensionality as 0D, 1D and 2D. The probabilities were scaled and coded with the RGB decimal code to indicate the probability of each dimensionality in a given composition (a) Distribution of 0D probabilities in the predicted dataset. (b) Distribution of 1D probabilities in the predicted dataset. (c) Distribution of 2D probabilities in the predicted dataset.

systems.

- Aliphatic structures
- Cyclic structures
- Benzene structures
- Sulphur Substitute structures
- Oxygen substituted structures

The L2 and cosine matrices give the most reasonable results and it was harder to draw a line between L2 and cosine results since both results were equally good. For further analysis, cosine distance matrix was used in finding the similarity between molecules. During the study, two chemical systems were taken into consideration for finding new A-site cations to synthesize novel materials.

The L2 and cosine matrices yielded the most reasonable results, and it was challenging to distinguish between the L2 and cosine results as both were equally good.



















Additionally, there was a lack of a quantitative parameter to measure the performance of distance matrices. This made the selection of a distance matrix more challenging.

For further analysis, the cosine distance matrix was utilized to determine the similarity between molecules. During the study, two chemical systems were considered in the search for new A-site cations to synthesize novel materials.

- perovskite structures with chiral cations
- perovskite structures with oligothiophenes

These two types of chemical systems’ organic cations exhibited distinct properties. Chiral cations display optical activity, while oligothiophene cations exhibit fluorescence and demonstrate the ability to tune the electronic properties by functionalizing and extending the length of the oligomer. Six structures were selected from both chemical systems, and their cations were subjected to a similarity search to find the most similar cations. Table 3.3 was constructed after performing a similarity search on the six organic cations and predicting the dimensionality of the structures when these cations were combined with a desired inorganic unit ( $\text{PbI}_6$ ,  $\text{PbCl}_6$ ,  $\text{PbBr}_6$ ,  $\text{PbCl}_4$ ,  $\text{PbBr}_4$ , and  $\text{PbI}_4$ ). In Table 3.3, the term ‘freq’ represents the frequency of occurrence of the organic molecule in the HOIPs dataset, while ‘rank’ indicates the molecule’s similarity ranking among 17 million molecules. Combining these organic and inorganic units in various compositions can yield structures with predicted dimensionality.

Table 3.3 Predicted dimensionalities of similar cations in reference to selected cations and inorganic units. Rank presents how similar is given cation to the reference. Occurrence shows the number of counts the cation appears in the model training/testing dataset. \*Experimental structure contains the predicted dimensionality

Cations from similarity search	freq	Rank	PbCl <sub>6</sub>	PbBr <sub>6</sub>	PbI <sub>6</sub>	PbCl <sub>4</sub>	PbBr <sub>4</sub>	PbI <sub>4</sub>
	2	ref	0D	0D	0D	2D	2D	2D
	0	4	0D	0D	0D	2D	2D	2D
	0	6	0D	0D	0D	2D	2D	2D
	6	ref	0D	0D	0D	2D	2D	2D
	0	4	0D	0D	0D	2D	2D	2D
	0	6	0D	0D	0D	2D	2D	2D
	2	ref	2D	2D	2D	2D	2D	2D
	0	2	0D	0D	2D	2D	2D	2D
	1	15	0D	0D	0D	2D	2D	2D
	2	ref	2D	2D	2D	2D	2D	2D
	1	6	2D	2D	2D	2D	2D	2D
	0	12	0D	0D	0D	2D	2D	2D
	1	ref	0D	0D	0D	2D	2D	2D
	0	2	0D	0D	0D	2D	2D	2D
	0	4	0D	0D	0D	2D	2D	2D
	1	ref	2D	2D	2D	2D	2D	2D
	1	2	2D	2D	2D	2D	2D	2D
	1	3	0D	0D	0D	2D	2D	2D

### 3.3 CONCLUSIONS AND OUTLOOK

In training the machine learning model it is observed that the inorganic unit features are robust features in representing perovskite structures and they have the ability to act as standalone features. Since inorganic features were significant to determining dimensionality the effect of organic features do not have the same impact as inorganics, but organic features increase the model performance in predicting dimensionality and is also a good representation model to represent organic molecules. In the probability distributions, it shows when the BX percentage reduces simultaneously to the increase of octahedral factors the probability of getting a 0D structure increases. In the similarity search, the organic cations with zero frequency are the most probable to make novel materials. It is worth noting that some of these specific organic molecules have not been identified in the CCDC database as perovskites or non-perovskite halometallates, which indicates that novel structures containing them should be synthetically accessible in low-dimensional structures.

Future work will entail expanding the dataset by incorporating additional structures to enhance the performance of the dimensionality-predicting model, DFT calculations will utilize to evaluate candidate structures from similarity search and incorporate more screening steps in generating probability maps and similarity searches. The DFT calculations will determine band gaps and other relevant properties. Incorporating additional screening steps, such as having an advanced charge matching procedure will increase the possibility of the inclusion of double perovskites and also contribute to more precise results for the probability map.

## BIBLIOGRAPHY

- [1] Ling, Y.; Yuan, Z.; Tian, Y.; Wang, X.; Wang, J. C.; Xin, Y.; Hanson, K.; Ma, B.; Gao, H. Bright light-emitting diodes based on organometal halide perovskite nanoplatelets. *Adv. Mater.* **2016**, *28*, 305–311.
- [2] Zhang, Q.; Su, R.; Liu, X.; Xing, J.; Sum, T. C.; Xiong, Q. High-quality whispering-gallery-mode lasing from cesium lead halide perovskite nanoplatelets. *Adv. Funct. Mater.* **2016**, *26*, 6238–6245.
- [3] Zhao, Y.; Zhu, K. Efficient planar perovskite solar cells based on 1.8 eV band gap CH<sub>3</sub>NH<sub>3</sub>PbI<sub>2</sub>Br nanosheets via thermal decomposition. *J. Am. Chem. Soc.* **2014**, *136*, 12241–12244.
- [4] Guo, Z.; Wu, X.; Zhu, T.; Zhu, X.; Huang, L. Electron–phonon scattering in atomically thin 2D perovskites. *ACS Nano* **2016**, *10*, 9992–9998.
- [5] Yin, J.; Maity, P.; De Bastiani, M.; Dursun, I.; Bakr, O. M.; Brédas, J.-L.; Mohammed, O. F. Molecular behavior of zero-dimensional perovskites. *Sci. Adv.* **2017**, *3*, e1701793.
- [6] Sun, S.; Lu, M.; Gao, X.; Shi, Z.; Bai, X.; Yu, W. W.; Zhang, Y. 0D perovskites: unique properties, synthesis, and their applications. *Adv. Sci.* **2021**, *8*, 2102689.
- [7] Zhou, C.; Lin, H.; Tian, Y.; Yuan, Z.; Clark, R.; Chen, B.; van de Burgt, L. J.; Wang, J. C.; Zhou, Y.; Hanson, K., et al. Luminescent zero-dimensional organic metal halide hybrids with near-unity quantum efficiency. *Chem. Sci.* **2018**, *9*, 586–593.

- [8] Sun, C.; Jiang, K.; Han, M.-F.; Liu, M.-J.; Lian, X.-K.; Jiang, Y.-X.; Shi, H.-S.; Yue, C.-Y.; Lei, X.-W. A zero-dimensional hybrid lead perovskite with highly efficient blue-violet light emission. *J. Mater. Chem. C* **2020**, *8*, 11890–11895.
- [9] Gong, S.-H.; Ko, S.-M.; Jang, M.-H.; Cho, Y.-H. Giant rabi splitting of whispering gallery polaritons in GaN/InGaN core-shell wire. *Nano Lett.* **2015**, *15*, 4517–4524.
- [10] Byrnes, T.; Kim, N. Y.; Yamamoto, Y. Exciton-polariton condensates. *Nat. Phys.* **2014**, *10*, 803–813.
- [11] Imamog, A.; Ram, R.; Pau, S.; Yamamoto, Y., et al. Nonequilibrium condensates and lasers without inversion: Exciton-polariton lasers. *Phys. Rev. A* **1996**, *53*, 4250.
- [12] Lanty, G.; Zhang, S.; Lauret, J.; Deleporte, E.; Audebert, P.; Bouchoule, S.; Lafosse, X.; Zuñiga-Pérez, J.; Semond, F.; Lagarde, D., et al. Hybrid cavity polaritons in a ZnO-perovskite microcavity. *Phys. Rev. B* **2011**, *84*, 195449.
- [13] Han, Y.; Yue, S.; Cui, B.-B. Low-dimensional metal halide perovskite crystal materials: structure strategies and luminescence applications. *Adv. Sci.* **2021**, *8*, 2004805.
- [14] Reshmi Varma, P. In *Perovskite Photovoltaics*; Thomas, S., Thankappan, A., Eds.; Academic Press, 2018; pp 197–229.
- [15] Kamminga, M. E.; de Wijs, G. A.; Havenith, R. W.; Blake, G. R.; Palstra, T. T. The role of connectivity on electronic properties of lead iodide perovskite-derived compounds. *Inorg. Chem* **2017**, *56*, 8408–8414.
- [16] Wu, L.-M.; Wu, X.-T.; Chen, L. Structural overview and structure-property relationships of iodoplumbate and iodobismuthate. *Coord. Chem. Rev.* **2009**, *253*, 2787–2804.



- [17] Lermer, C.; Harm, S. P.; Birkhold, S. T.; Jaser, J. A.; Kutz, C. M.; Mayer, P.; Schmidt-Mende, L.; Lotsch, B. V. Benzimidazolium lead halide perovskites: effects of anion substitution and dimensionality on the bandgap. *Z. Anorg. Allg. Chem.* **2016**, *642*, 1369–1376.
- [18] Kamminga, M. E.; Fang, H.-H.; Filip, M. R.; Giustino, F.; Baas, J.; Blake, G. R.; Loi, M. A.; Palstra, T. T. Confinement effects in low-dimensional lead iodide perovskite hybrids. *Chem. Mater.* **2016**, *28*, 4554–4562.
- [19] Knutson, J. L.; Martin, J. D.; Mitzi, D. B. Tuning the band gap in hybrid tin iodide perovskite semiconductors using structural templating. *Inorg. Chem.* **2005**, *44*, 4699–4705.
- [20] Filip, M. R.; Eperon, G. E.; Snaith, H. J.; Giustino, F. Steric engineering of metal-halide perovskites with tunable optical band gaps. *Nat. Commun.* **2014**, *5*, 5757.
- [21] Smith, M. D.; Jaffe, A.; Dohner, E. R.; Lindenberg, A. M.; Karunadasa, H. I. Structural origins of broadband emission from layered Pb–Br hybrid perovskites. *Chem. Sci.* **2017**, *8*, 4497–4504.
- [22] Mao, L.; Wu, Y.; Stoumpos, C. C.; Wasielewski, M. R.; Kanatzidis, M. G. White-light emission and structural distortion in new corrugated two-dimensional lead bromide perovskites. *J. Am. Chem. Soc.* **2017**, *139*, 5210–5215.
- [23] Brivio, F.; Walker, A. B.; Walsh, A. Structural and electronic properties of hybrid perovskites for high-efficiency thin-film photovoltaics from first-principles. *APL Mater.* **2013**, *1*.
- [24] Noel, N. K.; Stranks, S. D.; Abate, A.; Wehrenfennig, C.; Guarnera, S.; Haghighirad, A.-A.; Sadhanala, A.; Eperon, G. E.; Pathak, S. K.; Johnston, M. B., et al.

- Lead-free organic–inorganic tin halide perovskites for photovoltaic applications. *Energy Environ. Sci.* **2014**, *7*, 3061–3068.
- [25] Walsh, A. Principles of chemical bonding and band gap engineering in hybrid organic–inorganic halide perovskites. *J. Phys. Chem. C* **2015**, *119*, 5755–5760.
- [26] Eperon, G. E.; Stranks, S. D.; Menelaou, C.; Johnston, M. B.; Herz, L. M.; Snaith, H. J. Formamidinium lead trihalide: a broadly tunable perovskite for efficient planar heterojunction solar cells. *Energy Environ. Sci.* **2014**, *7*, 982–988.
- [27] Denis, P.-H.; Mertens, M.; Van Gompel, W. T.; Van Hecke, K.; Ruttens, B.; D’Haen, J.; Lutsen, L.; Vanderzande, D. Directing the Self-Assembly of Conjugated Organic Ammonium Cations in Low-Dimensional Perovskites by Halide Substitution. *Chem. Mater.* **2021**, *33*, 5177–5188.
- [28] Dong, Y.; Zhang, Y.; Li, X.; Feng, Y.; Zhang, H.; Xu, J. Chiral perovskites: promising materials toward next-generation optoelectronics. *Small* **2019**, *15*, 1902237.
- [29] Fu, D.; Xin, J.; He, Y.; Wu, S.; Zhang, X.; Zhang, X.-M.; Luo, J. Chirality-Dependent Second-Order Nonlinear Optical Effect in 1D Organic–Inorganic Hybrid Perovskite Bulk Single Crystal. *Angew. Chem. Int. Ed.* **2021**, *60*, 20021–20026.
- [30] Long, G.; Zhou, Y.; Zhang, M.; Sabatini, R.; Rasmita, A.; Huang, L.; Lakhwani, G.; Gao, W. Theoretical prediction of chiral 3D hybrid organic–inorganic perovskites. *Adv. Mater.* **2019**, *31*, 1807628.
- [31] Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **2018**, *9*, 1–8.

- [32] Ai, Q.; Williams, D. M.; Danielson, M.; Spooner, L. G.; Engler, J. A.; Ding, Z.; Zeller, M.; Norquist, A. J.; Schrier, J. Predicting inorganic dimensionality in templated metal oxides. *J. Chem. Phys.* **2021**, *154*, 184708.
- [33] Gao, P.; Bin Mohd Yusoff, A. R.; Nazeeruddin, M. K. Dimensionality engineering of hybrid halide perovskite light absorbers. *Nat. Commun.* **2018**, *9*, 5028.
- [34] Akkerman, Q. A.; Nguyen, T. P.; Boehme, S. C.; Montanarella, F.; Dirin, D. N.; Wechsler, P.; Beiglböck, F.; Rainò, G.; Erni, R.; Katan, C., et al. Controlling the nucleation and growth kinetics of lead halide perovskite quantum dots. *Science* **2022**, *377*, 1406–1412.
- [35] Jahanbakhshi, F.; Mladenović, M.; Dankl, M.; Boziki, A.; Ahlawat, P.; Rothlisberger, U. Organic Spacers in 2D Perovskites: General Trends and Structure-Property Relationships from Computational Studies. *Helv. Chim. Acta* **2021**, *104*, e2000232.
- [36] Vasileiadou, E. S.; Hadar, I.; Kepenekian, M.; Even, J.; Tu, Q.; Malliakas, C. D.; Friedrich, D.; Spanopoulos, I.; Hoffman, J. M.; Dravid, V. P., et al. Shedding Light on the Stability and Structure–Property Relationships of Two-Dimensional Hybrid Lead Bromide Perovskites. *Chem. Mater.* **2021**, *33*, 5085–5107.
- [37] Mitzi, D. B.; Feild, C.; Harrison, W.; Guloy, A. Conducting tin halides with a layered organic-based perovskite structure. *Nature* **1994**, *369*, 467–469.
- [38] Stoumpos, C. C.; Cao, D. H.; Clark, D. J.; Young, J.; Rondinelli, J. M.; Jang, J. I.; Hupp, J. T.; Kanatzidis, M. G. Ruddlesden–Popper hybrid lead iodide perovskite 2D homologous semiconductors. *Chem. Mater.* **2016**, *28*, 2852–2867.
- [39] Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

- [40] Huo, H.; Rupp, M. Unified representation of molecules and crystals for machine learning. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045017.
- [41] Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems* **2020**, *33*, 12559–12571.
- [42] Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513–530.
- [43] Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.
- [44] 2D Perovskites Database - The Laboratory of New Materials for Solar Energetics[<http://pdb.nmse-lab.ru/>]. <http://pdb.nmse-lab.ru/>.
- [45] Tremblay, M.-H.; Boyington, A.; Rigin, S.; Jiang, J.; Bacsá, J.; Al Kurdi, K.; Khrustalev, V. N.; Pachter, R.; Timofeeva, T. V.; Jui, N., et al. Hybrid Organic Lead Iodides: Role of Organic Cation Structure in Obtaining 1D Chains of Face-Sharing Octahedra vs 2D Perovskites. *Chem. Mater.* **2022**, *34*, 935–946.
- [46] Lyu, R.; Moore, C. E.; Liu, T.; Yu, Y.; Wu, Y. Predictive Design Model for Low-Dimensional Organic–Inorganic Halide Perovskites Assisted by Machine Learning. *J. Am. Chem. Soc.* **2021**, *143*, 12766–12776.
- [47] Keerthisinghe, N.; Christian, M. S.; Berseneva, A. A.; Morrison, G.; Klepov, V. V.; Smith, M. D.; Zur Loye, H.-C. Investigation of Metastable Low Dimensional Halometallates. *Molecules* **2022**, *27*, 280.
- [48] Laasner, R.; Du, X.; Tanikanti, A.; Clayton, C.; Govoni, M.; Galli, G.; Ropo, M.; Blum, V. MatD<sup>3</sup>: A Database and Online Presentation Package for Research

Data Supporting Materials Discovery, Design, and Dissemination. *Journal of Open Source Software* **2020**, *5*, 1945.

- [49] Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714–3717.
- [50] Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464–477.
- [51] Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.
- [52] Arteca, G. A. *Rev. Comput. Chem.*; John Wiley I& Sons, Ltd, 1996; pp 191–253.
- [53] Todeschini, R.; Consonni, V. *Handbook of Chemoinformatics*; John Wiley I& Sons, Ltd, 2003; Chapter VIII.2, pp 1004–1033.
- [54] Baumgärtner, A. Shapes of flexible vesicles at constant volume. *J. Chem. Phys.* **1993**, *98*, 7496–7501.
- [55] Li, C.; Soh, K. C. K.; Wu, P. Formability of ABO<sub>3</sub> perovskites. *J. Alloys Compd.* **2004**, *372*, 40–48.
- [56] Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2016**, *2*, 1–7.
- [57] Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1976**, *32*, 751–767.
- [58] Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global

understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 2522–5839.

[59] Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

[60] Irwin, J. J.; Shoichet, B. K. ZINC- a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

# APPENDIX A

## SUPPLEMENTARY INFORMATION

### A.0.1 SIMILARITY SEARCH

Rdkit features No normalization

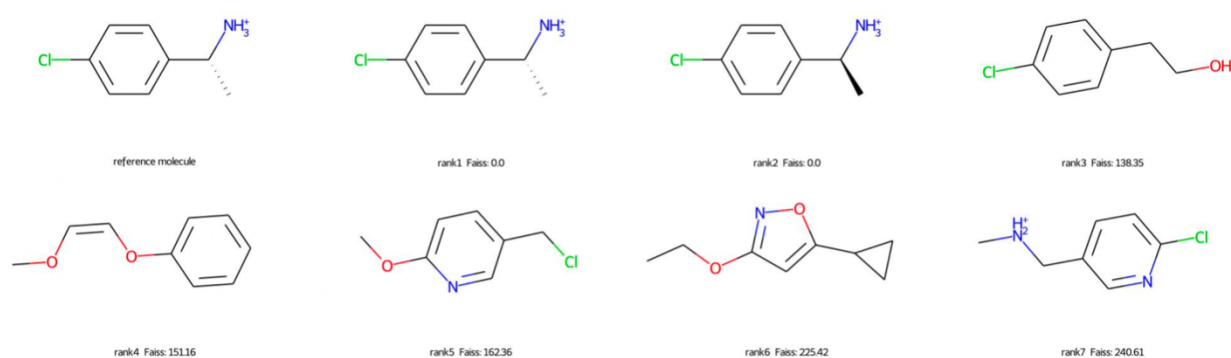


Figure A.1 Similarity search result with rdkit features without normalizing the features

Standardization

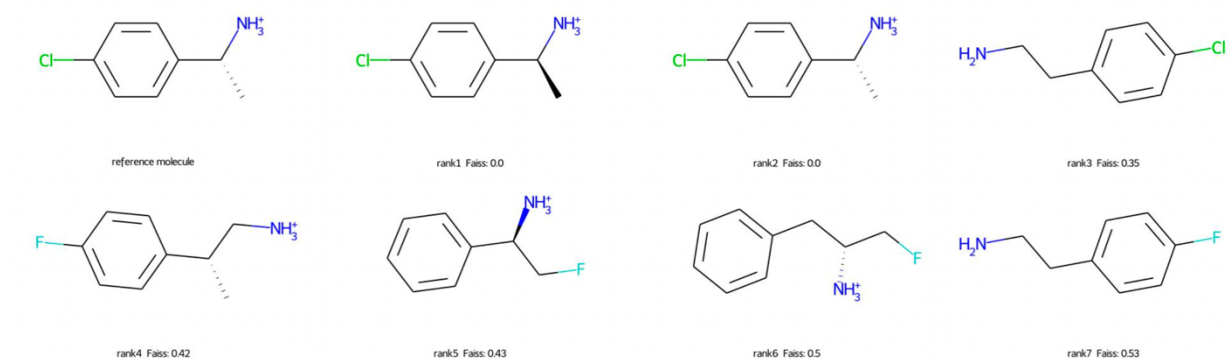


Figure A.2 Similarity search result with rdkit features with standardized features

In considering the datasets of rdkit features we can qualitatively see the min-

max normalization and standardization increases the accuracy of similarity search. Standardization is much more reliable than min-max normalization chemically when we compare the rank 3 and rank 7 molecules. The chemical environment in ranks 3 and 7 in min-max normalization is very different from the reference molecule. When we compare rdkit and Grover in similarity search, qualitatively we can say the Grover dataset performs better.

#### Minmax normalization

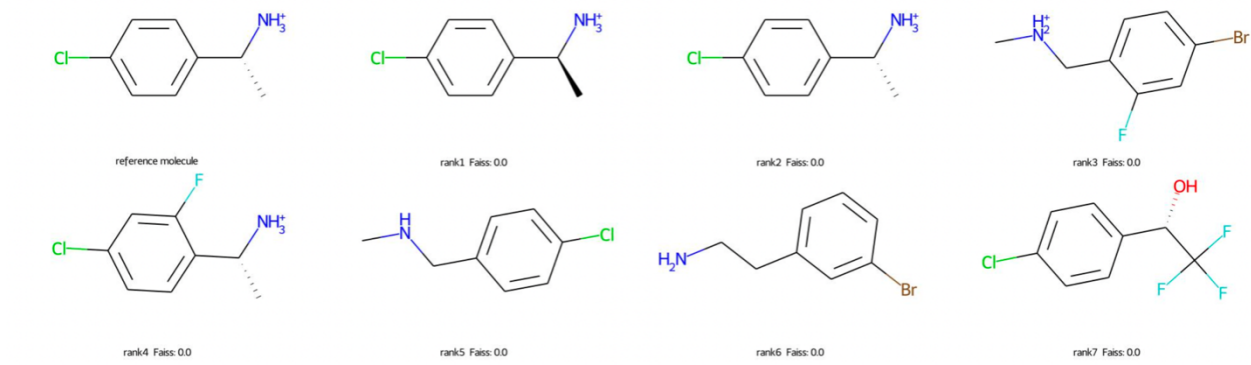


Figure A.3 Similarity search result with rdkit features with min-max normalized features

#### Grover features

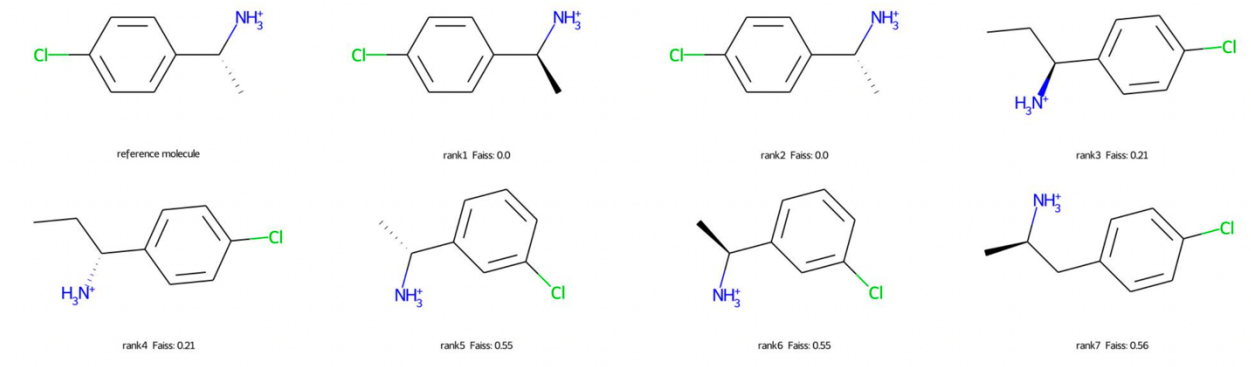


Figure A.4 Similarity search result with grover features

47 features to represent the inorganic unit, in selecting 47 features from 132 a selection criterion was used. Each feature from Matminer is added to the feature list individually and 10 models were trained. The selection was based on the averaged



Table A.1 The precisions of the different classification models with different feature combinations for rdkit feature set(rdkit includes the hull volume of the cations)

Data with different features	No of Features	0D	1D	2D
rdkit only	14	0.63	0.60	0.84
rdkit+inorganic	17	0.85	0.78	0.91
rdkit+inorganic+magpie	64	0.88	0.89	0.93
charge only	1	0.58	0.85	0.91
oct_factor only	1	0.87	0.00	0.77
atomic volume only	1	0.81	0.80	0.92

Table A.2 The precisions of the different classification models with different feature combinations for the 20% stratified test set for the upsample dataset

Data with different features	No of Features	0D	1D	2D
Organic cations	200	0.53	0.40	0.88
Inorganic features	7	0.73	0.61	0.93
organic cation + Inorganic features	207	0.87	0.77	0.94

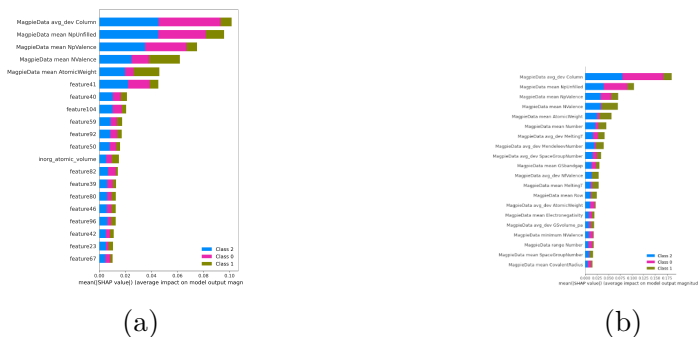


Figure A.5 Feature importance calculated using SHAP values for two models inorganic only model and inorganic+organic model

0D and 1D precision of the test set for the trained model. When the average 0D and 1D precision for 10 runs are above a certain threshold value (the threshold is selected from the baseline model performance before adding the matminer features) the feature is selected to represent the inorganic unit. The selected features are in SI Table A.3.

Table A.3 Magpie features used in model training [56]

Magpie features	MagpieData minimum Row	MagpieData mean NValence
MagpieData range Number	MagpieData mean Row	MagpieData range NsUnfilled
MagpieData mean Number	MagpieData minimum CovalentRadius	MagpieData mean NpUnfilled
MagpieData avg_dev Number	MagpieData mean CovalentRadius	MagpieData avg_dev NdUnfilled
MagpieData minimum MendelevNumber	MagpieData mean Electronegativity	MagpieData mode NfUnfilled
MagpieData mean MendelevNumber	MagpieData mode Electronegativity	MagpieData minimum NUnfilled
MagpieData avg_dev MendelevNumber	MagpieData avg_dev NsValence	MagpieData maximum GSvolume_pa
MagpieData mean AtomicWeight	MagpieData maximum NpValence	MagpieData avg_dev GSvolume_pa
MagpieData avg_dev AtomicWeight	MagpieData range NpValence	MagpieData minimum GSbandgap
MagpieData mode AtomicWeight	MagpieData mean NpValence	MagpieData range GSbandgap
MagpieData minimum MeltingT	MagpieData mode NpValence	MagpieData mean GSbandgap
MagpieData mean MeltingT	MagpieData mean NdValence	MagpieData avg_dev GSbandgap
MagpieData avg_dev MeltingT	MagpieData maximum NfValence	MagpieData avg_dev GSmagmom
MagpieData range Column	MagpieData avg_dev NfValence	MagpieData mean SpaceGroupNumber
MagpieData avg_dev Column	MagpieData minimum NValence	MagpieData avg_dev SpaceGroupNumber
MagpieData mode Column	MagpieData range NValence	MagpieData mode SpaceGroupNumber