AI Informed Toxicity Screening of Amine Chemistries used in the Synthesis of Hybrid Organic-Inorganic Perovskites

An Su[1], Haotian Xue[1], Yuanbin She[1], and Krishna Rajan[2]

[1]Zhejiang University of Technology
[2]University at Buffalo

September 24, 2021

Abstract

This paper describes a machine learning guided framework for screening the potential toxicity impact of amine chemistries used in the synthesis of hybrid organic-inorganic perovskites. Using a combination of a probabilistic molecular fingerprint technique that encodes bond connectivity (MinHash) coupled to non-linear data dimensionality reduction methods (UMAP), we develop an "Amine Atlas'. We show how the Amine Atlas can be used to rapidly screen the relative toxicity levels of amine molecules used in the synthesis of 2D and 3D perovskites and help identify safer alternatives. Our work also serves as a framework for rapidly identifying molecular similarity guided, structure-function relationships for safer materials chemistries that also incorporate sustainability/ toxicity concerns.

AI Informed Toxicity Screening of Amine Chemistries used in the Synthesis of Hybrid Organic-Inorganic Perovskites

An Su[1,2], Haotian Xue[3], Yuanbin She[1], and Krishna Rajan[2]*

1. College of Chemical Engineering, Zhejiang University of Technology, Hangzhou, Zhejiang, 310014, China

2. Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260-1660, United States

3. Collaborative Innovation Center of Yangtze River Delta Region Green Pharmaceuticals, Zhejiang University of Technology, Hangzhou, Zhejiang, 310014, China

*Corresponding Author:

Krishna Rajan

orcid.org/0000-0001-9303-2797

Email: *krajan3@buffalo.edu*

Abstract

This paper describes a machine learning guided framework for screening the potential toxicity impact of amine chemistries used in the synthesis of hybrid organic-inorganic perovskites. Using a combination of a probabilistic molecular fingerprint technique that encodes bond connectivity (MinHash) coupled to non-linear data dimensionality reduction methods (UMAP), we develop an "Amine Atlas'. We show how the Amine Atlas can be used to rapidly screen the relative toxicity levels of amine molecules used in the synthesis

of 2D and 3D perovskites and help identify safer alternatives. Our work also serves as a framework for rapidly identifying molecular similarity guided, structure-function relationships for safer materials chemistries that also incorporate sustainability/ toxicity concerns.

Topical Heading

AIChE Journal Special Issue on "Artificial Intelligence in Chemical Engineering".

Keywords

artificial intelligence, machine learning, hybrid organic-inorganic perovskite, amine chemistry, toxicity screening

Introduction

In recent years, organic-inorganic perovskites have received a huge amount of attention due to their promise for photovoltaic application[1-3]. Generally, perovskite refers to the ABX3 three-dimensional structural frameworks with "A" as an organic cation (usually the cation of amines), "B" as a metal, and "X" as a halogen.[1,2]. The chemical diversity of amines used in perovskites for PV applications is large as they appear in both 3D and 2D perovskite structures.[4,5] For instance, in a recent review study by Saparov and Mitzi, more than 60 distinct perovskite amine structures have been mentioned for the design of versatile perovskite materials.[6]

With the rapid development of new classes of perovskite chemistry with promising physical properties, the concerns in identifying amine chemistries that have a minimal environmental/ health footprint become more challenging.[7] In looking at the full life cycle of perovskites used in PV applications, we now need to consider the toxicity of the organic molecules, along with metal elements such as Pb, at all stages of the materials synthesis and degradation.[8-17] For example, previous studies on the toxicity of amines have concluded that many aromatic amines (e.g. aniline, diphenylamine) are potential carcinogens, while the aliphatic amines are less hazardous.[18-21] However, in aquatic environments, there are potential formations of toxic compounds including nitrosamines and nitramines from the reaction between amines and nitrite oxidants.[19,21] As the perovskite materials become more and more popular, it is critical to perform a systematic study on the toxicity of both the existing amines that have been part of the 3-D and lower-dimensional perovskites and the potential perovskite amines that have similar structures to the existing ones. In this study, we present a machine learning aided molecular structure-toxicity analysis to screen the potential toxicity of amines used for the synthesis of hybrid organic-inorganic perovskites.

Our training and test database of perovskite amines is based on open source literature along with a structural similarity search on PubChem (https://pubchem.ncbi.nlm.nih.gov), a well-acknowledged database for chemical structure and functionality[22]. For toxicity data of perovskite amines, instead of fetching data from different literature without a common standard, we performed searches on PubChem Bioactivity Assays database[23] which is based on similar data standards that are more suitable to study at a systematic level. This study aims to establish a structure-toxicity relationship of perovskite amines, help identify safer alternatives for use in perovskite structures.

Methods

As shown in Figure 1, the research consists of five main steps: the curation of perovskite amines database, the two-level classifications of perovskite amines, the chemical informatics and machine learning computations, the search for toxicity data, and the visualization of Amine Atlas.

Curation of perovskite amines database. The amines corresponding to the perovskite ammonium cations mentioned in recently published reviews[6,24] and database[25]are sorted into the "existing perovskite amines" list (e.g., ethylamine corresponds to ethyl ammonium). This list is the basis of the perovskite amines

database. Next, the database is further expanded by including "potential perovskite amines", which have similar structures to existing amines. The PubChem similar structure search is performed on each existing perovskite amine, and the similarity threshold is 95%. By removing ions, non-amines, and existing structures in the database, the amines in the search results are further screened. In addition, these amines must have been tested by at least one activity assay in PubChem BioAssay database[23,26] in order to be included as potential perovskite amines. The above search and filtering steps are completed by our programming tool based on the open-source Python packages PubChemPy (*https://pubchempy.readthedocs.io/en/latest/*) and RDKit (*https://www.rdkit.org/docs/api-docs.html*). In the final database, each amine has its PubChem Compound ID (CID), name, SMILES, and a list of its corresponding PubChem Bioassay ID (AID). The CID and name of their corresponding ammonium cations are also included.

Two-level classification. The amines are first classified according to their aromaticity and the position of their amine group (e.g. on aromatic ring, directly attached to the aromatic ring, or on the alkyl substituent of the aromatic ring). Further subclass classifications are established based on more detailed structures such as functional groups and linearity of alkyl chains). The identification of functional groups and chemical fragments is achieved through our RDKit-based programming tool (provided in supplementary material). It is worth noting that the purpose of classification here is not to establish a new standard of amine classification but to distinguish the amines in our database as much as possible.

Chemical informatics and machine learning computations. The MHFP6 fingerprint[27] is calculated for each perovskite amine molecule using the open-source Python package MHFP (*https://github.com/reymond-group/mhfp*). The dimensionality of the fingerprint is then reduced by Uniform Manifold Approximation and Projection (UMAP)[28] with the open-source Python package UMAP (*https://umap-learn.readthedocs.io/en/latest/index.html*). The parameters of these two tools, including the number of permutations of MHFP and the number of neighborhoods and the minimum distance of UMAP, are optimized for the clustering of different amine classes and subclasses. The data processing steps during the computations are achieved with the open-source Python packages Pandas (*https://pandas.pydata.org/docs/*) and Scikit-learn[29]. The code for the computations is provided in the supplementary material.

Search for toxicity data. The detailed information of all the bioassays with AID recorded in our perovskite amines database is retrieved from PubChem Bioassay Database[23,26] using our PubChemPy-based programming tool. Only the bioassays with more than one perovskite amine showing "active" are kept. In addition, the bioassays showing bioactivity other than toxicity are eliminated. Finally, a table of PubChem Bioassays with their AID, number of perovskite amines tested as active, and assay name is obtained.

Visualization of Amine Atlas. The Amine Atlas is visualized using Plotly, a Python open-source graphing library (*https://plotly.com/python/*). The Amine Atlas can be viewed with or without amine toxicity data. In Amine Atlas, each data point represents an amine—UMAP calculation results are used as the two-dimensional coordinates of the data point, and the classification results or the hit ratio of the compound is displayed in the color of the data point. The detailed information of the corresponding compounds of the data point is displayed in the hover data tab, such as CID, SMILES, type (existing or potential), and classifications. All Amine Atlas shown in the following sections have corresponding interactive versions in the supplementary material.

Results and Discussion

The first part of our curated perovskite amine database consists of 184 amines that correspond to the ammonium cations in literature, named "existing perovskite amines". The structural similarity search on PubChem and further screening process give an additional 264 amines that are considered "potential perovskite amines" —the amines that have similar structures to existing ones. Finally, the curated perovskite amine database contains 448 amine structures. The full table of the database is provided in the supplementary material. The main reason for expanding the database is to make full use of data on amines that have

been tested for bioactivity or toxicity, regardless of whether they have been studied as perovskite amines. As more amines are included in the analysis, it may be easier to find toxicity trends and their relationship to the amine structure.

The introduction of artificial intelligence to the creation of Amine Atlas and toxicity screening of amine chemistries involves the calculation of MinHash fingerprint, up to six bonds (MHFP6) and Uniform Manifold Approximation and Projection (UMAP). MHFP6 is an improved version of the extended connectivity fingerprint (ECFP)[27] that lowers the dimensionality needed to describe the detailed molecular substructures as well as increases the performance of the nearest neighbor search.[30] The MHFP6 fingerprint has been used in recently published chemistry databases[31,32] and data visualization tool[33] with big data settings. MinHash is a locality sensitive hashing (LSH) scheme that applies a family of hashing functions to the substrings in molecular shingling and stores the minimum hash generated from each hashing function in a set. These sets, containing the minimum hash values, have the interesting property that they can be indexed by an LSH algorithm for approximate nearest neighbor search (ANN), removing the curse of dimensionality.[30]MinHash allows for the indexing of chemical structures in extremely sparse Jaccard (Tanimoto) space, a metric more appropriate for fingerprint-based similarity calculations. [30] On the other hand, UMAP is a recently developed non-linear dimensionality reduction algorithm[28] that has been used to analyze various types of scientific data, mainly in the field of biological sciences including genome aggregation[34], single-cell mass flow cytometry[35], and single-cell RNA sequencing (scRNA-seq)[35-37]. UMAP is a manifold learning method that preserves local and global structure of the high-dimensional data points by minimizing data/information loss. It explores the network connectivity using K-nearest neighbor distance (KNN) over a high-dimensional hyperplane and then estimates a low-dimensional coordinate system that replicates the same graph structure, preserving the edge connectivity of the high-dimensional by keeping graphical representation intact in the low-dimensional space. Compared with the more frequently used t -distributed stochastic neighborhood embedding (t-SNE) algorithm which has limited capability to represent the global structure of the data, it is found that UMAP retains the local and global structure of the data by simultaneously capturing the small differences and the continuity between the data subsets.

The higher level of classification gets amines categorized into aliphatic amines (cyclic and noncyclic), heterocyclic aromatic amines, and other aromatic amines including phenylalkyl amines and anilines. Combining this classification information with the results of the UMAP on the MHFP6 fingerprint of perovskite amines, the clustering of these amine classes can be observed on Amine Atlas. The optimized clustering is reached when MHFP permutation number, UMAP number of neighbors, UMAP minimum distance are set to 2048, 50, and 0.25, respectively. Using this combination of parameters, the main classes are well-separated from each other on the Amine-Atlas (Figure 2), and the same parameters are used for all the Amine Atlas below.

For each amine class, the Amine-Atlas can display further classifications as subclasses. The subclasses of heterocyclic aromatic amines are shown in Figure 3. This class of amines is clearly divided into common nitrogen-containing aromatics, including pyrrole, imidazole, pyridine, and thiazole, and sulfur-containing thiophene. No overlap is observed between the clusters, which may be due to the effectiveness of MHFP6 fingerprint in capturing the characteristics of common aromatic compounds.

Similarly, for the class of phenylalkyl amines, the subclasses are well-separated in Amine-Atlas (Figure 4). This figure shows the power of UMAP in capturing both the local and global structure of the data. Here, the UMAP captures subtle differences between subclasses (such as those with the same carbon number) by dividing them into different clusters (e.g. 1-phenylethylamines (C_6H_5-C(C)NH_3) and phenylethylamines (C_6H_5-CCNH_3)). At the same time, the UMAP shows the continuity of close subclasses by placing them in adjacent positions, such as the benzylamines (C_6H_5-CNH_3) and phenylethylamines (C_6H_5-CCNH_3) whose alkyl substituents differ in chain length by 1.

Due to the complex structure of branched alkyl chains, the noncyclic aliphatic amines have some clusters with less organization (Figure 5). However, the trend still exists in the amines with linear alkyl chains, such as the linear diamines (purple) and linear monoamines (orange) subclasses, where the length of the alkyl chain decreases along the UMAP-1 axis. In addition, amines that have functional groups in addition to

amine groups (dark green) are distant from unsubstituted amines (purple and orange).

One important purpose of this study is to screen the relative hazard of amines being used in 2D and 3D perovskite synthesis – those most hazardous and those not so. We retrieve the toxicity data of perovskite amines from PubChem Bioassay Database[23,26], an open-source repository holding a collection of bioactivity and toxicity data of small molecules—these molecules are cross-linked to the data of their chemical structures stored in PubChem Compounds Database[22]. After a search using our programming tools, we summarized a list of PubChem Bioassays that focus on the toxicity of chemicals and in the meantime include perovskite amines as test substances, and the complete list of assays is provided in the supplementary material. Examples of the toxicity effects and corresponding AID are shown in Table 1.

Table 1. Examples of selected PubChem Bioassays and the toxicity effect they study

AID	Toxicity effect
743122	Activator of the aryl hydrocarbon receptor (AhR)
1224892	Agonists of the constitutive androstane receptor (CAR)
1347033	Agonists of Human pregnane X receptor (PXR)
720637	Disruptors of the mitochondrial membrane potential (MMP)
743219	Agonists of the antioxidant response element (ARE)
1159553	Agonists of the retinoic acid receptor (RAR)
743079	Agonists of the estrogen receptor alpha (ER-alpha)
743078	Antagonists of the estrogen receptor alpha (ER-alpha)
1259247	Antagonists of the androgen receptor (AR)
1259396	Antagonists of the estrogen receptor beta (ER-beta)

According to the test results of PubChem Bioassays, we use the hit ratio (the ratio of active substances to the total number of screening targets) to indicate the overall toxicity of our perovskite amines. By plotting hit ratio data on Amine Atlas (Figure 6), rapid screening of structure-toxicity relationships can be established. By comparing Figure 6 with Figure 2 and Figure 3, it is found that most of the toxic perovskite amines are in the anilines cluster, while a few pyrroles and pyridines in the heterocyclic aromatics also have a hit ratio > 0.3. Meanwhile, the aliphatic amines are less toxic, and most of the toxic aliphatic amines are cyclic. It should be noted that compared with other amines, phenylalkyl amines have very little toxicity data, so care should also be taken when using these amines.

In addition to providing the information extracted from the Amine Atlas, we also provide statistical analysis on the toxicity data from PubChem Bioassays without the involvement of AI. We recommend that researchers and manufacturers use the 10 safest existing perovskite amines (Table 2) and potential perovskite amines (Table 3) as they have the lowest hit ratio, and, if possible, avoid using amines with high hit ratio (Table 4). It is worth noting that our recommendation of the safer potential perovskites is entirely based on our research from the perspective of toxicity—people should determine the chemical or physical properties of these amines according to their scientific or industrial needs.

Table 2. Ten safest existing perovskite amines ranked by hit ratio (the fraction of actives out of the total number of screened targets)

CID	Name	SMILES	Class	Subclass	H
674	Dimethylamine	CNC	Aliphatic (noncyclic)	Short amine (heavy atoms < 5)	0/
1146	Trimethylamine	CN(C)C	Aliphatic (noncyclic)	Short amine (heavy atoms < 5)	0/
1390	1-Methylimidazole	Cn1ccnc1	Heterocyclic aromatic	Imidazole	0/
24874	2-Butanamine	CCC(C)N	Aliphatic (noncyclic)	Branched monoamine	1/
10009	3-Aminopyridine	Nc1cccnc1	Heterocyclic aromatic	Pyridine	1/
31018	3-(Aminomethyl)pyridine	NCc1cccnc1	Heterocyclic aromatic	Pyridine	1/

CID	Name	SMILES	Class	Subclass	H
774	Histamine	NCCc1cnc[nH]1	Heterocyclic aromatic	Pyrrole	1/
7515	N-Methylaniline	CNc1ccccc1	Aniline	Aniline	2/
9257	1,2,4-Triazole	c1nc[nH]n1	Heterocyclic aromatic	Pyrrole	2/
4091	Metformin	CN(C)C(=N)N=C(N)N	Aliphatic (noncyclic)	Multi-amine	2/

Table 3. Ten safest potential perovskite amines ranked by hit ratio

CID	Name	SMILES	Class	Subclass	Hit Ratio
13032	N-Isopropylaniline	CC(C)Nc1ccccc1	Aniline	Aniline	0/47
119716	Isononanamine	CC(C)CCCCCCN	Aliphatic (noncyclic)	Branched monoamine	0/44
7096	1-Phenylpiperazine	c1ccc(N2CCNCC2)cc1	Aniline	Aniline	0/32
60993	N,N-Diethylethylenediamine	CCN(CC)CCN	Aliphatic (noncyclic)	Branched diamine	0/32
1727	4-Aminopyridine	Nc1ccncc1	Heterocyclic aromatic	Pyridine	0/48
7973	2-Bromopyridine	Brc1ccccn1	Heterocyclic aromatic	Pyridine	0/27
14310	N-Butylaniline	CCCCNc1ccccc1	Aniline	Aniline	0/44
1049	Pyridine	c1ccncc1	Heterocyclic aromatic	Pyridine	1/48
6115	Aniline	Nc1ccccc1	Aniline	Aniline	1/48
13195	4-Methylimidazole	Cc1cnc[nH]1	Heterocyclic aromatic	Pyrrole	1/47

Table 4. Ten most hazardous existing perovskite amines ranked by hit ratio

CID	Name	SMILES	Class	Subclass	Hit Ratio
7472	N,N-Dimethyl-p-phenylenediamine	CN(C)c1ccc(N)cc1	Aniline	Aniline	35/47
7814	p-Phenylenediamine	Nc1ccc(N)cc1	Aniline	Aniline	29/44
7111	Benzidine	Nc1ccc(-c2ccc(N)cc2)cc1	Aniline	Aniline	28/47
7807	4-Bromoaniline	Nc1ccc(Br)cc1	Aniline	Aniline	24/48
7102	4-Aminobiphenyl	Nc1ccc(-c2ccccc2)cc1	Aniline	Aniline	23/47
16720	1,5-Naphthalenediamine	Nc1cccc2c(N)cccc12	Aniline	Aniline	21/48
4837	Piperazine	C1CNCCN1	Aliphatic (cyclic)	Piperazine	19/44
13583	Dodecylamine	CCCCCCCCCCCCN	Aliphatic (noncyclic)	Linear monoamine	18/48
70248	1,8-Diamino-3,6-dioxaoctane	NCCOCCOCCN	Aliphatic (noncyclic)	Functionalized amine	12/44
7812	4-Chloroaniline	Nc1ccc(Cl)cc1	Aniline	Aniline	13/49

Conclusions

In this study, we have provided a new data-driven / AI framework for environmentally conscious selection of amine chemistries used in the synthesis of hybrid organic-inorganic perovskites. The selection strategy is based on exploring high dimensional data capturing structure-function- toxicity driven by molecular-scale information. To the best of our knowledge, this is the first such study to critically explore AI methods to rank toxicity impact from the perspective of molecular descriptors; and to harness this information to identify safer alternatives that also have been shown to be preserving the functional performance of such perovskites for photovoltaic applications. By coupling new probabilistic-based molecular descriptors with advanced data dimensionality such as UMAP, we have also established a database resource to explore other families of yet unexplored amine chemistries that may be used for hybrid perovskite structures. The need for searching and identifying alternative and safer chemistries for establishing a "benign-by-design' has long been recognized, our work provides an example of how AI coupled to foundational materials chemistry principles can actually facilitate an a priori approach to select chemicals for materials synthesis that meet the structure-function and sustainability metrics.

Acknowledgments

The authors acknowledge the support from NSF Award# 1640867 – DIBBs: EI: Data Laboratory for Materials Engineering and the Collaboratory for a Regenerative Economy (CoRE center) in the Dept of Materials Design and Innovation – University at Buffalo**.**

References

1. Green MA, Ho-Baillie A. Perovskite Solar Cells: The Birth of a New Era in Photovoltaics. *ACS Energy Letters.* 2017;2(4):822-830.

2. Green MA, Ho-Baillie A, Snaith HJ. The emergence of perovskite solar cells. *Nature Photonics.* 2014;8(7):506-514.

3. Park N-G. Organometal Perovskite Light Absorbers Toward a 20% Efficiency Low-Cost Solid-State Mesoscopic Solar Cell. *The Journal of Physical Chemistry Letters.* 2013;4(15):2423-2429.

4. Mitzi DB. Templating and structural engineering in organic–inorganic perovskites. *Journal of the Chemical Society, Dalton Transactions.* 2001(1):1-12.

5. Kieslich G, Sun S, Cheetham AK. Solid-state principles applied to organic–inorganic perovskites: new tricks for an old dog.*Chemical Science.* 2014;5(12):4712-4715.

6. Saparov B, Mitzi DB. Organic–Inorganic Perovskites: Structural Versatility for Functional Materials Design. *Chemical Reviews.*2016;116(7):4558-4596.

7. Zhang J, Gao X, Deng Y, Zha Y, Yuan C. Comparison of life cycle environmental impacts of different perovskite solar cell systems.*Solar Energy Materials and Solar Cells.* 2017;166:9-17.

8. Babayigit A, Ethirajan A, Muller M, Conings B. Toxicity of organometal halide perovskite solar cells. *Nature Materials.*2016;15(3):247-251.

9. Poglitsch A, Weber D. Dynamic disorder in methylammoniumtrihalogenoplumbates (II) observed by millimeter-wave spectroscopy. *The Journal of Chemical Physics.*1987;87(11):6373-6378.

10. Upadhyaya A, Negi CMS, Yadav A, Gupta SK, Verma AS. Synthesis and Characterization of Methylammonium Lead Iodide Perovskite and its Application in Planar Hetero-junction Devices. *Semiconductor Science and Technology.* 2018;33(6):065012.

11. Li Y, Galisteo-López JF, Calvo ME, Míguez H. Facile Synthesis of Hybrid Organic–Inorganic Perovskite Microcubes of Optical Quality Using Polar Antisolvents. *ACS Applied Materials & Interfaces.*2017;9(41):35505-35510.

12. Niu G, Guo X, Wang L. Review of recent progress in chemical stability of perovskite solar cells. *Journal of Materials Chemistry A.* 2015;3(17):8970-8980.

13. Latini A, Gigli G, Ciccioli A. A study on the nature of the thermal decomposition of methylammonium lead iodide perovskite, CH3NH3PbI3: an attempt to rationalise contradictory experimental results.*Sustainable Energy & Fuels.* 2017;1(6):1351-1357.

14. Babayigit A, Duy Thanh D, Ethirajan A, et al. Assessing the toxicity of Pb- and Sn-based perovskite solar cells in model organism Danio rerio. *Scientific Reports.* 2016;6(1):18721.

15. Ke W, Kanatzidis MG. Prospects for low-toxicity lead-free perovskite solar cells. *Nature communications.* 2019;10(1):965.

16. Lyu M, Yun J-H, Chen P, Hao M, Wang L. Addressing Toxicity of Lead: Progress and Applications of Low-Toxic Metal Halide Perovskites and Their Derivatives. *Advanced Energy Materials.* 2017;7(15):1602512.

17. Bae S-Y, Lee SY, Kim J-w, et al. Hazard potential of perovskite solar cell technology for potential implementation of "safe-by-design" approach. *Scientific Reports.* 2019;9(1):4242.

18. El-Eswed B. Effect of basicity and hydrophobicity of amines on their adsorption onto charcoal. *Desalination and Water Treatment.*2016;57(41):19227-19238.

19. Poste AE, Grung M, Wright RF. Amines and amine-related compounds in surface waters: A review of sources, concentrations and aquatic toxicity. *Science of The Total Environment.* 2014;481:274-279.

20. Weisburger EK, Russfield AB, Homburger F, et al. Testing of twenty-one environmental aromatic amines or derivatives for long-term toxicity or carcinogenicity. *J Environ Pathol Toxicol.*1978;2(2):325-356.

21. Lee D, Wexler AS. Atmospheric amines – Part III: Photochemistry and toxicity. *Atmospheric Environment.* 2013;71:95-103.

22. Kim S, Thiessen PA, Bolton EE, et al. PubChem Substance and Compound databases. *Nucleic Acids Research.* 2016;44(D1):D1202-D1213.

23. Wang Y, Suzek T, Zhang J, et al. PubChem BioAssay: 2014 update.*Nucleic Acids Research.* 2013;42(D1):D1075-D1082.

24. Li X, Hoffman JM, Kanatzidis MG. The 2D Halide Perovskite Rulebook: How the Spacer Influences Everything from the Structure to Optoelectronic Device Efficiency. *Chemical Reviews.*2021;121(4):2230-2291.

25. Marchenko EI, Fateev SA, Petrov AA, et al. Database of Two-Dimensional Hybrid Perovskite Materials: Open-Access Collection of Crystal Structures, Band Gaps, and Atomic Partial Charges Predicted by Machine Learning. *Chemistry of Materials.* 2020;32(17):7383-7388.

26. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research.* 2009;37(suppl_2):W623-W633.

27. Rogers D, Hahn M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling.* 2010;50(5):742-754.

28. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426.* 2018.

29. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *The Journal of machine Learning research.*2011;12:2825-2830.

30. Probst D, Reymond J-L. A probabilistic molecular fingerprint for big data settings. *Journal of Cheminformatics.* 2018;10(1):66.

31. Awale M, Sirockin F, Stiefl N, Reymond J-L. Medicinal Chemistry Aware Database GDBMedChem. *Molecular Informatics.*2019;38(8-9):1900031.

32. Bühlmann S, Reymond J-L. ChEMBL-Likeness Score and Database GDBChEMBL. *Frontiers in Chemistry.* 2020;8(46).

33. Probst D, Reymond J-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics.*2020;12(1):12.

34. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans.*Nature.* 2020;581(7809):434-443.

35. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology.*2019;37(1):38-44.

36. Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature.*2019;566(7745):496-502.

37. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods.*2019;16(12):1289-1296.