CRYSTAL GROWTH & DESIGN

pubs.acs.org/crystal

Article

Principled Exploration of Bipyridine and Terpyridine Additives to Promote Methylammonium Lead Iodide Perovskite Crystallization

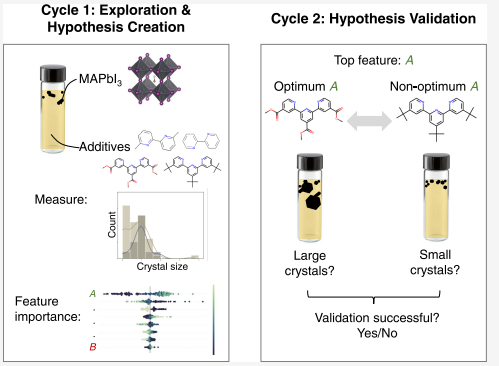
Noor Titan Putri Hartono,* Mansoor Ani Najeeb, Zhi Li, Philip W. Nega, Clare A. Fleming, Xiaohe Sun, Emory M. Chan, Antonio Abate, Alexander J. Norquist, Joshua Schrier, and Tonio Buonassisi*

Cite This: https://doi.org/10.1021/acs.cgd.2c00522

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Additives in the precursor solution can promote lead-halide perovskite (LHP) crystallization. We present a systematic exploration of nine (9) bipyridine- and terpyridine-based additives selected from 29 candidates using high-throughput single-crystal growth. To combat selection bias and generate hypotheses for future experimental cycles of learning, we featurize candidate additives using Mordred descriptors and compare similarity metrics. A previously unreported additive, 6,6′-dimethyl-2,2′-dipyridyl, is shown to work particularly well (the highest top 10th percentile is ∼3.8 mm, in comparison to ∼1.9 mm without additive) in improving the crystallization of prototypical methylammonium lead iodide ($MAPbI_3$). Our strategy of machine-learning-guided high-throughput experimentation is generally applicable to other crystal growth problems.



1. INTRODUCTION

Lead-halide perovskites (LHP) are an emerging class of material with applications in photovoltaics,[1] photodetectors,[2] and transistors.[3] Research on LHPs has focused on improving the optical and electrical properties of polycrystalline thin films due to their convenient solution processability (e.g., spin-coating). However, such films exhibit a high density of structural defects and grain boundaries that are difficult to control.[4] In contrast, LHP single crystals exhibit a lower density of structural defects due to a more ordered lattice, reduced defects, and lack of internal grain boundaries.[5] Identifying the conditions for growing high-quality single crystals can be challenging. One strategy for improving crystal growth is the use of additive molecules in precursor solutions.[6] A large variety of additives have been used to control LHP growth such as CH_3COONa,[7] choline bromide,[8] and 3-(decyldimethylammonio)-propane-sulfonate inner salt,[9] which have been demonstrated to improve the growth of $Cs_2AgBiBr_6$, $CsPbBr_3$, and $MAPbI_3$ single crystals, respectively.

In this study, we investigate bipyridine- and terpyridine-based additive molecules for the growth of methylammonium lead iodide ($MAPbI_3$) LHP derivatives. We focused on these two classes of additives because bidentate ligand 2,2′-bipyridine and tridentate ligand 2,2′:6′,2″-terpyridine, which serve as Lewis bases, have been shown to improve the crystallization and morphology of $FA_{0.88}Cs_{0.12}PbI_3$ thin-film perovskites[10] but have not been studied with other forms of LHPs (bulk samples or nanomaterials). However, considering the vast number of possible bipyridine- and terpyridine-based additive molecules with various functional groups, a systematic exploration of additive molecules coupled with a method to better understand the results is still needed. Previous studies on single-crystal LHPs using high-throughput experiments solely focused on exploring different A-site cations and reaction conditions;[11,12] only recently have additives such as water[13,14] been explored in the same manner.

The diverse set of LHP crystallization additives suggests that we are far from having a strong theoretical framework for selecting additives. In this early stage of research, human selection bias can limit the chemical diversity of explored compounds and, in turn, limit the conclusions we draw from that data.[15] We introduce a framework to combat selection bias and promote structured hypothesis generation, which ensures diversity in each experimental cycle of learning. In this framework, we evaluate similarity metrics (mainly Tanimoto similarity, but also t-distributed stochastic neighbor embedding, t-SNE) and ensure a principled optimization approach by down-selecting features using a recursive feature elimination (RFE) process. We apply this framework to the second experimental cycle of learning to systematically screen additive molecules for LHP single crystals, specifically methylammonium lead iodide ($MAPbI_3$). We discuss the merits and limitations of this approach.

Received: May 6, 2022
Revised: July 18, 2022

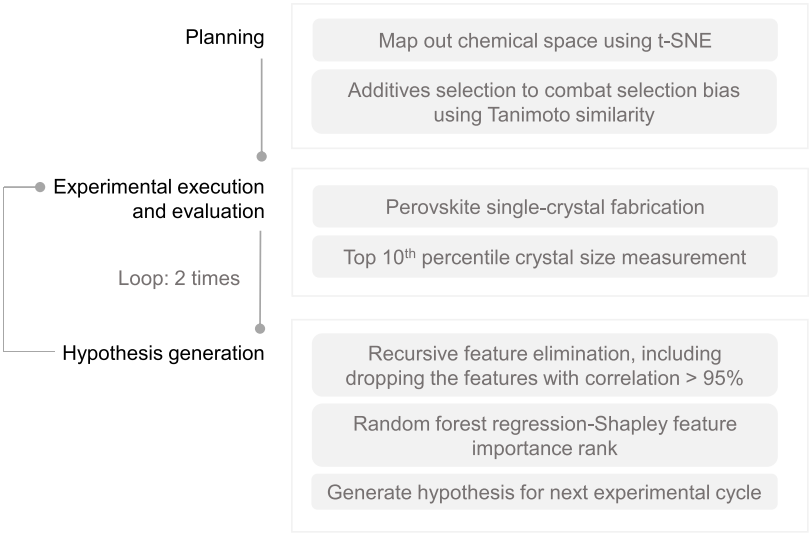


Figure 1. Workflow of the hypothesis generation and simulation. It consists of 3 groups: planning, experimental execution and evaluation, and hypothesis generation.

Table 1. List of Additive Molecule Candidates[a]

#	additive molecules	note
0	2,2′-bipyridyl	cycle 1
1	4,4′-dimethyl-2,2′-dipyridyl	insoluble
2	5,5′-dimethyl-2,2′-dipyridyl	cycle 1
3	6,6′-dimethyl-2,2′-dipyridyl	cycle 1, best
4	6-bromo-2,2′-bipyridine	
5	6,6′-dibromo-2,2′-dipyridyl	
6	4,4′-dimethoxy-2,2′-bipyridine	cycle 1
7	2,2′-bipyridine-4,4′-dicarboxaldehyde	
8	2,6-bis(2-pyridyl)-4(1*H*)-pyridone	cycle 1
9	2-(2-pyridinyl)quinoline	
10	2,2′-biquinoline	insoluble
11	4′-chloro-2,2′:6′,2″-terpyridine	
12	6,6″-dibromo-2,2′:6′.2″-terpyridine	insoluble
13	2,2′-bipyridine-3,3′-diol	cycle 1
14	4,4′-di-*tert*-butyl-2,2′-dipyridyl	cycle 1
15	2,2′-bipyridine-4-carboxylic acid	
16	2,2′-bipyridine-4,4′-dicarboxylic acid	
17	4,4-dinonyl-2,2′-dipyridyl	
18	5,5′-bis(trifluoromethyl)-2,2′-bipyridine	
19	4′,4‴-(1,4-phenylene)bis(2,2′:6′,2″-terpyridine)	
20	4,4′,4″-tri-*tert*-butyl-2,2′:6′,2″-terpyridine	cycle 2
21	2,2′-bipyridine-5,5′-dicarboxylic acid	
22	(2,2′)bipyridinyl-4,4′-dicarboxylic acid dimethyl ester	
23	4′-(4-methylphenyl)-2,2′:6′,2″-terpyridine	
24	4′-(4-chlorophenyl)-2,2′:6′,2″-terpyridine	
25	trimethyl 2,2′:6′,2″-terpyridine-4,4′,4″-tricarboxylate	cycle 2
26	caerulomycin A	
27	2,2′-bipyridine-3,3′-dicarboxylic acid	insoluble
28	2-pyridin-2-yl-quinoline-4-carboxylic acid	

[a]The list shows the additive candidates, including the ones that are insoluble and added for $MAPbI_3$ crystal growth in cycles 1 and 2. Candidates with a blank "Note" column were not tested in this study.

2. MATERIALS AND EXPERIMENTAL METHODS

2.1. Workflow of the Study. The general workflow of the study consists of three steps, as shown in Figure 1: planning, experimental execution and evaluation, and hypothesis generation. Experimental execution and evaluation are automated to improve reproducibility. After the first hypothesis generation step, the hypothesis is refined by repeating the planning and experimental steps. These steps are repeated in two cycles.

During the planning step, additive toxicity and availability drove the selection of additive molecules in the first cycle. To improve diversity, in the second cycle, Tanimoto similarity[16,17] is calculated for all 29 candidate additive molecules, sorting each of the candidates from less to more similar with the remainder of the data set. To visualize these

https://doi.org/10.1021/acs.cgd.2c00522
Cryst. Growth Des. XXXX, XXX, XXX−XXX

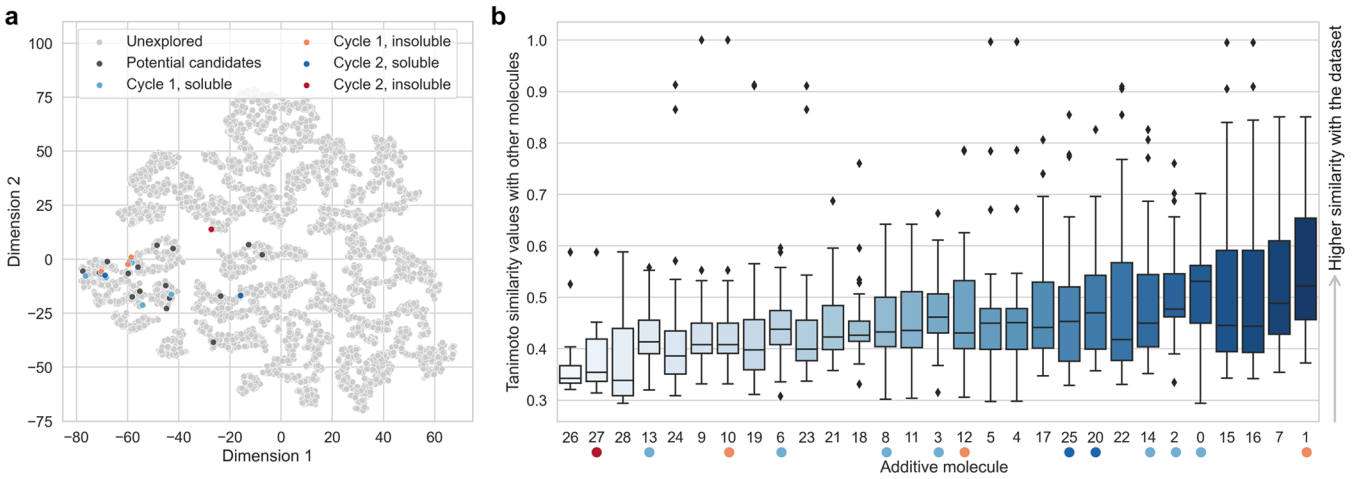


Figure 2. Diversity of additive molecule candidates. (a) t-SNE plot showing where the 29 additive molecule candidates are within the chemical space with n = 5000 randomly drawn chemicals from the eMolecules database. (b) Distribution of Tanimoto similarity between one additive candidate with other candidates, sorted from low to high mean values; the boxes indicate the lower and upper quartiles, with the median indicated by a line inside the box. The lines outside the box indicate the lowest and highest value, and the dots indicate outliers.

results, we performed t-distributed stochastic neighbor embedding (t-SNE). t-SNE informs the researchers in the subsequent steps to prioritize and select as diverse additive molecules as possible to be synthesized.

The next step is experimental execution and evaluation. Once an additive has been selected, a batch of 24 $MAPbI_3$ single crystals is grown. Experiments are created by randomly sampling the volume ratios of stock solutions (the stock solutions are pure γ-butyrolactone (GBL), $MAPbI_2$ mixture in GBL, MA in GBL, and additive in GBL), followed by image capture of the results and image processing to quantify the size distribution for each additive and processing condition. The images are processed computationally to extract the crystal size distribution using ImageJ,[18] and the top 10^{th} percentile crystal size is used as the figure of merit for whether the conditions result in large single crystals.

After repeating two cycles of this process and before the hypothesis generation step is performed, we preprocess the data by transforming the simplified molecular-input line-entry system (SMILES)[19] of additive molecules into Mordred descriptors.[20] We then input the data set into a pipeline (see Figure 1) containing recursive feature elimination (RFE) including dropping the features with Pearson's correlation higher than 95%, random forest regression, Shapley feature importance rank, result evaluation by human, and finally, hypothesis generation for the next experimental cycle.

2.2. Additive Candidate Selection and Materials. The 29 bipyridine- and terpyridine-based additive molecule candidates are screened based on the availability from the manufacturer (Sigma-Aldrich) and further selected based on safety/toxicity level. The complete list of the additive molecule candidates is shown in Table 1.

Out of 29 bipyridine-based and terpyridine-based additive molecule candidates, the solubilities of 4 molecules do not reach target thresholds (0.005 mol/L), leaving 25 candidates (see Table 1). The first cycle was focused on gathering initial data on a diverse set of species and tested seven additive molecules. The second cycle was focused on testing the hypotheses learned from the first cycle, and so 2 additives were tested, one with a predicted positive effect and one with a predicted negative effect on crystallization. The general single-crystal synthesis process follows the established inverse temperature crystallization protocol.[12] For each additive, 24 single-crystal growth experiments were performed, sampling randomized volumes of MAI, PbI_2, additive precursor solutions, and solvents. In previous studies, random sampling has been shown to be a resource-efficient way to initialize the learning process, relative to, e.g., full factorial grid search, which would have required a larger number of samples to obtain the same quality of information.[21] Without knowing the optimum volumes for the salts for each additive, constraining to a single salt volume (i.e., only sampling different additive concentrations) might result in missing an effect in one additive or another. The single crystals were synthesized and images of the single crystals were taken in two locations, Haverford College and Lawrence Berkeley National Laboratory, to deliberately introduce noise due to a slight difference in laboratory conditions and to achieve a more robust result.

All reagents were purchased from commercial sources and used without further purification. Lead(II) iodide (99%), γ-butyrolactone (GBL, ≥99%), dichloromethane (DCM, ≥99.8%), and the nine additives used in this study (6,6′-dimethyl-2,2′-dipyridyl (98%), 5,5′-dimethyl-2,2′-dipyridyl (98%), 2,2′-bipyridyl (98%), 4,4′-dimethoxy-2,2′-bipyridine (97%), 2,2′-bipyridine-3,3′-diol (98%), 2,6-bis(2-pyridyl)-4(1*H*)-pyridone (98%), 4,4′-di-*tert*-butyl-2,2′-dipyridyl (98%), 4,4′,4″-tri-*tert*-Butyl-2,2′:6′,2″-terpyridine (95%), and trimethyl 2,2′:6′,2″-terpyridine-4,4′,4″-tricarboxylate (98%)) were purchased from Sigma-Aldrich. Methylammonium iodide (MA, 99.99%) organic ammonium salt was purchased from GreatCell Solar. While impurities in reagents may affect the outcome of experiments in general, in our study, the effect of additive precursor purity on these specific experiments appears to be small. All additives in this study had similar purities (ranging between 95 and 98%), and the purity level is not among the top eight features surviving RFE (see Figure S1 and surrounding discussion).

2.3. Solubility Measurement. The solubility of PbI_2 depends on the concentration of coexisting methylammonium iodide (MA) in the solution. To determine the maximum solubility of the $MAPbI_2$ precursor solution, we weighed fixed masses of PbI_2 and methylammonium iodide and gradually added γ-butyrolactone (GBL) solvent. The suspension was placed in a heated oil bath at 75 °C and 450 rpm. Neat GBL was added gradually until the solid was completely dissolved. Maximum solubility limits of pristine amine and additive were also determined using the same method. The solutions were then brought to room temperature to ensure their stability. This step also ensured that the solution remained homogeneous during the robotic run and that the stock did not precipitate, which could disrupt the pipetting by the liquid handler. The solubility was then calculated as the total moles of solute divided by the total volume of the solutions. A concentration threshold for the additives of 0.005 mol/L was established to ensure adequate reactant concentrations. Reactions in which this threshold could not be achieved were not performed. The concentrations of $MAPbI_2$ mixture solution were 0.95 mol/L (PbI_2) and 0.66 mol/L (MA). The concentration of the MA-only solution was 2.14 mol/L. The maximum concentrations of the additive-only solutions are given in Supporting Information, Table S1.

2.4. Equipment: Liquid Handling Robot for Single-Crystal Synthesis. An enclosed Hamilton Microlab NIMBUS4 was used at

https://doi.org/10.1021/acs.cgd.2c00522
Cryst. Growth Des. XXXX, XXX, XXX−XXX

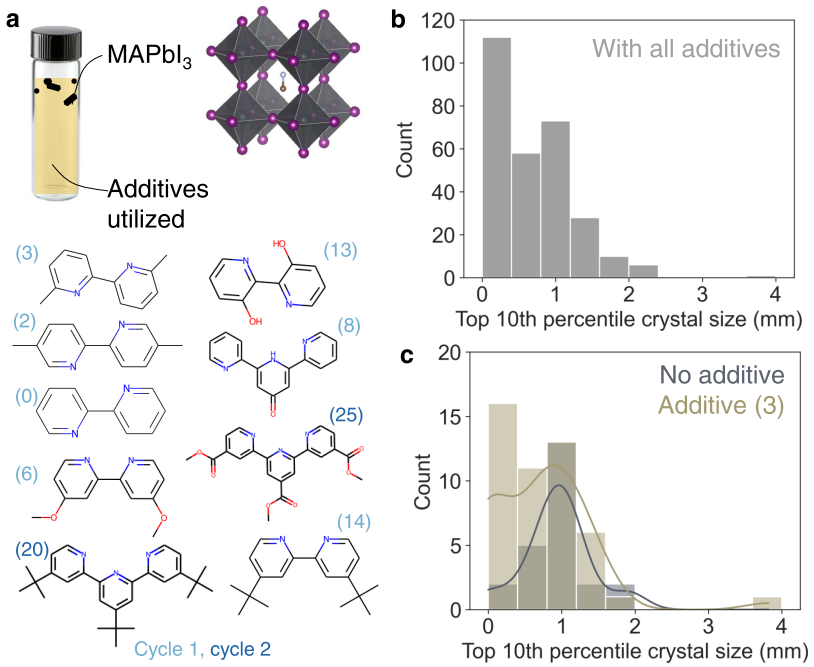


Figure 3. Experimental results of the perovskite single-crystal growth with various additives. (a) After screening the additives based on toxicity and solubility, $MAPbI_3$ single crystals were grown with seven additives in the first cycle: (3) 6,6′-dimethyl-2,2′-dipyridyl, (2) 5,5′-dimethyl-2,2′-dipyridyl, (0) 2,2′-bipyridyl, (6) 4,4′-dimethoxy-2,2′-bipyridine, (13) 2,2′-bipyridine-3,3′-diol, (8) 2,6-bis(2-pyridyl)-4(1*H*)-pyridone, and (14) 4,4′-di-*tert*-butyl-2,2′-dipyridyl; and two additives in the second cycle: (20) 4,4′,4″-tri-*tert*-butyl-2,2′:6′,2″-terpyridine and (25) trimethyl 2,2′:6′,2″-terpyridine-4,4′,4″-tricarboxylate. The histogram of top 10th percentile crystal size for each processing condition for all additives (b) and the comparison between single crystals with additive (3) and no additive (c).

Haverford College and a Hamilton Microlab NIMBUS4 liquid handling robot was used at Lawrence Berkeley National Laboratory in this study for the high-throughput antisolvent vapor-assisted crystallization (HT-ASVC) synthesis of $MAPbI_3$-additive crystals.[14] ASVC is a slow process that relies upon variations of the perovskite solubility in different solvents. In other words, supersaturation can be attained by exposing a solution of the product to another solvent in which the product is sparingly soluble (otherwise called antisolvent). The slow diffusion of AS vapors into the perovskite precursor solution drastically reduces the solubility and results in precipitation.[22] The effectiveness of this method, including the quality and size of the crystals, depends on the solvent and antisolvent used, the presence of additives, the volume and concentration ratio, the reaction time, and the diffusion rate.

The robot used in this study features four independent pipetting channels for transferring liquid. The pipettors aspirate reagent stock solutions were stored in polypropylene containers organized in racks placed in programmatically defined positions on the robot deck. Stock solutions were used on the same day as they were prepared (within 8 h) to avoid any possible solution degradation. New pipette tips were used for each stock solution. Solutions for HT-ASVC reactions were prepared on a Hamilton Heater and Shaker (HHS) module, which can be heated up to 105 °C (actual temperature solution temperature reaches 95 °C) and can vortex microplates up to 2000 rpm. Robotic protocols were programmed in the Hamilton Method Editor software; reaction time, shaking speed, etc. were imported from spreadsheets generated by ESCALATE.[23] Detailed descriptions of the software and step-by-step synthetic protocols are in the "Robotic workflow" section of the Supporting Information.

3. RESULTS AND DISCUSSION

3.1. Planning: Additive Molecules Selection and Exploration. To estimate where the 29 additive molecules are within the chemical space, a t-SNE analysis was performed. t-SNE (t-distributed stochastic neighbor embedding) is a statistical method for visualizing high-dimensional data in a two-/three-dimensional map, and it can show where additive molecule candidates are within the chemical space in relation to other commercially available organic molecules,[24] although several recent studies caution against overinterpreting distance directly from a t-SNE plot.[25] The t-SNE plot for the additive molecule candidates is shown in Figure 2a, which shows the 29 additive molecule candidates in a chemical space of n = 5000 randomly drawn organic molecules from the eMolecules database.[26] We first generated Mordred descriptors from the SMILES of the eMolecules database and the list of additive candidates. Then, the t-SNE analysis was carried out using scikit-learn version 1.0.[27] As expected, the bipyridine and terpyridine additive molecule candidates are clustered in one region, yet span a large spread within the cluster. While the results derived from this study might only be applicable within this region/type of additive molecules, this type of map may help identify the potential region to explore for future additive studies.

To perform the experimental part efficiently in the second experiment cycle, we prioritized the types of additive molecules based on the important feature(s) found in the first cycle and the similarity between each additive molecule with the rest of the candidates. The Tanimoto similarity of all additive molecule pair combinations in the list was calculated,[16,17] and the distribution of these values for each additive is shown in Figure 2b. This value helps us to see how each additive molecule is similar/dissimilar to the rest of the molecules in the list. Additives (26) caerulomycin A and (1) 4,4′-dimethyl-2,2′-dipyridyl have the lowest and the highest mean of Tanimoto similarity, respectively, in comparison to other candidates. Not all of these are suitable for laboratory testing; the additive most similar to the others, (1) 4,4′-dimethyl-2,2′-dipyridyl, and one of the additives least similar to the others,

https://doi.org/10.1021/acs.cgd.2c00522
Cryst. Growth Des. XXXX, XXX, XXX−XXX

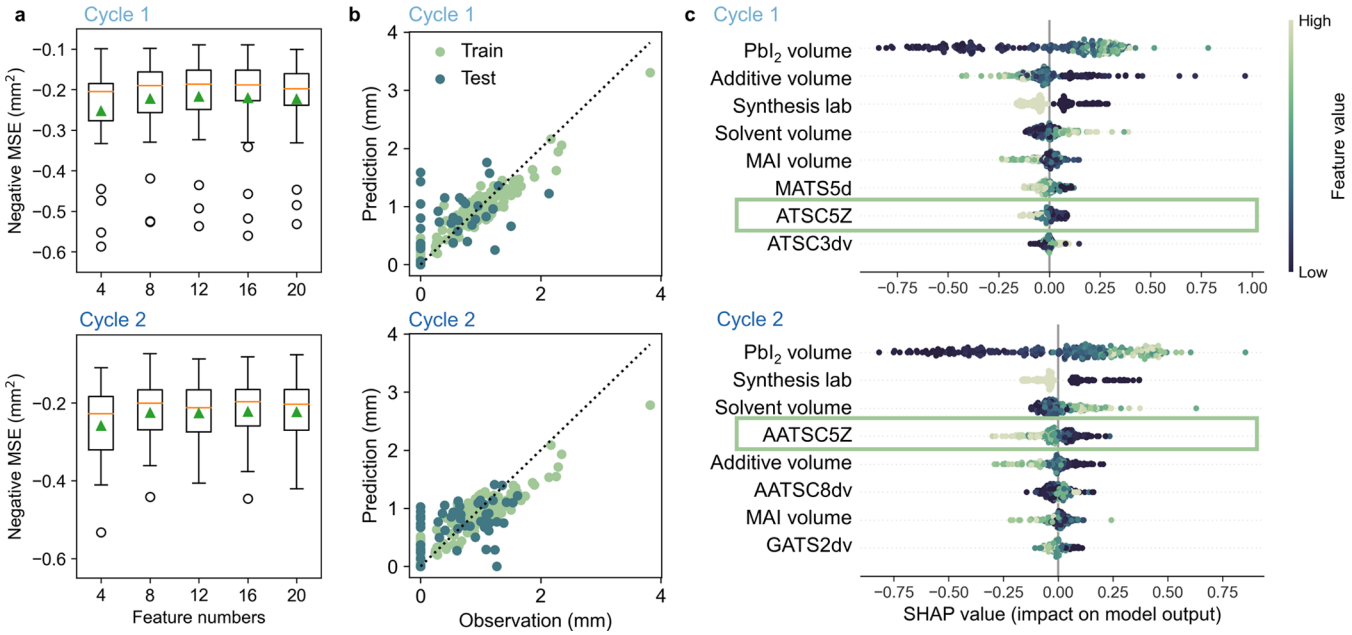


Figure 4. Hypothesis generation step driven by the sequence of algorithms. The recursive feature elimination negative mean squared error (MSE) distribution results for 4, 8, 12, 16, and 20 features in cycles 1 and 2 (a). Then, the selected features are trained in random forest regression (b), which can be further analyzed using Shapley values feature importance rank (c) in both cycles 1 and 2.

(27) 2,2′-bipyridine-3,3′-dicarboxylic acid, are both insufficiently soluble. Despite the insolubility of some molecules, we synthesized $MAPbI_3$ single crystals with a diverse set of additive molecules within two experimental cycles, demonstrating a way to combat selection bias using t-SNE coupled with Tanimoto similarity.

3.2. Experiment: Perovskite Single-Crystal Fabrication and Crystal Size Distribution Measurement. Seven additives were tested in the first cycle of single-crystal synthesis (light blue dots in Figure 2): (3) 6,6′-dimethyl-2,2′-dipyridyl, (2) 5,5′-dimethyl-2,2′-dipyridyl, (0) 2,2′-bipyridyl, (6) 4,4′-dimethoxy-2,2′-bipyridine, (13) 2,2′-bipyridine-3,3′-diol, (8) 2,6-bis(2-pyridyl)-4(1*H*)-pyridone, and (14) 4,4′-di-*tert*-butyl-2,2′-dipyridyl; and two additives in the second cycle: (20) 4,4′,4″-tri-*tert*-butyl-2,2′:6′,2″-terpyridine and (25) trimethyl 2,2′:6′,2″-terpyridine-4,4′,4″-tricarboxylate, as shown in Figure 3a. After the single crystals were synthesized, their images were taken. (Example images are shown in Supporting Information, Figure S2.) The distribution of single crystals was then measured manually using ImageJ,[18] from the front vial images, assuming that there was a negligible crystal size deviation due to the convex surface of the vial. Then, we extracted the top 10th percentile crystal size for each single-crystal growth condition for each additive, as shown in Figure 3b. The comparison between the distribution of top 10th percentile crystal size with no additive and with additive (3) 6,6′-dimethyl-2,2′-dipyridyl (best of all nine explored herein) is shown in Figure 3c. The distribution of crystal sizes with no additive tends to center around 1.2 mm with a relatively tight distribution. Despite only very few crystals with a size of ∼4 mm appeared with additive (3), it was sufficient to distinguish the "good crystal size" region from the "bad crystal size" region based on the t-SNE analysis shown in Supporting Information, Figure S3. In the presence of additives, the crystal size distribution expanded, increasing the probability of finding a large single crystal.

3.3. Machine Learning as a Tool for Hypothesis Generation and Refinement. Usually, machine learning regression is used as a predictive tool; in that case, extensive training with large data sets is conducted to minimize loss or error. Here, we did not have large training sets, so we could not use regressors as a predictive tool. However, a regressor trained on sparse data could still generate testable hypotheses for root causes to explore as well as future experiments to perform, the latter is the premise of active, or sequential, learning. In this study, 9 among the 29 candidate additives were studied in detail, and the hypotheses regarding the relationship between additive molecular structure and effect on crystal size were generated.

To represent the additive molecules in a regressor, we utilized Mordred descriptors, a set of more than 1800 two- and three-dimensional descriptors that encoded molecular properties and molecular topology.[20] This was in contrast to Morgan fingerprints, which represent the presence of certain atoms or functional groups.[28,29] The Mordred descriptors were generated from the SMILES representation of the additive molecules using the Mordred package.[20] Then, the Pearson's correlation of each feature was calculated and combined with the processing conditions of the fabrication. The features with high correlation (>0.95) were excluded to reduce the computational time and focus on the features, which were not highly correlated.

Figure 4a shows the negative mean squared error (MSE) for cycles 1 and 2 during the recursive feature elimination (RFE) step for various numbers of features. After eight features, the negative MSE does not improve, suggesting that eight features might be the optimum number of features.

After that, both the top 10th percentile crystal size and the input consisting of the Mordred descriptors of the additives combined with the processing conditions were trained using random forest regression. The fitting result for both cycles is shown in Figure 4b. Note that there were a few false positives in which small/medium crystal size is predicted to be >0 mm,

https://doi.org/10.1021/acs.cgd.2c00522
Cryst. Growth Des. XXXX, XXX, XXX−XXX

contradicting the experimental observations. However, this false positive result was reduced from cycle 1 to cycle 2, indicating a better fit when we have a larger data set.

To guide the selection of our additives, we iteratively used feature importance rank to select the next cycle of additive molecules. Shapley value analysis[30] is shown in Figure 4c, highlighting that all of the precursor volumes (PbI_2, MAI, additive, and solvent) survive the recursive feature elimination process and are revealed as the top features affecting the crystal size. We also observed that the synthesis laboratory was a top feature. Table S2 in the Supporting Information shows which additive is used in single-crystal synthesis in each laboratory.

After repeating the RFE-random forest regression-Shapley value analysis sequence, as shown in Supporting Information, Figure S4, we observe that one Mordred feature continues to rank as an important feature, ATSC5Z. This is the centered Moreau-Broto autocorrelation of lag 5 (the distance/lag d of five atomic steps (or bonds) between atomic pairs indexed as i and j) weighted by atomic number (Z), defined in eq 1, where d is the lag of autocorrelation, w_i and w_j are properties of the molecule at positions i and $j = (i + d)$, respectively, δ_{ij} is the Kronecker delta, and $\delta_{ij} = 1$ when $j = i + d$.[31]

$$\mathrm{ATSC}_{d,w} = \sum_{i=1}^{A}\sum_{j=1}^{A}\delta_{ij}(w_i - \overline{w})(w_j - \overline{w}) \quad (1)$$

ATSC5Z describes how identical pairs of elements are separated by five bonds within the molecule; or, more generally, how a given atomic number (element) is distributed along the topological structure. In the second cycle, AATSC5Z appeared consistently among the top features. AATSC5Z is the averaged ATSC5Z, factoring in the normalization factor for various sizes of additive molecules.

The specific choice of autocorrelation factor "5" has no clear physical significance, and its relationship to the additive behavior is purely empirical. We examined this factor from the dual perspectives of methodological robustness and chemical significance. From a methods point of view, the "5Z" parameter consistently appears. However, if we force the regressor to adopt "NZ" features (where N = 3−7), predictive accuracy (RMSE) does not suffer significantly (see Figure S5 in the Supporting Information). Additionally, the survival of 5Z through the feature downselection process depends on the order in which features are listed (an idiosyncrasy of scikit-learn). Lastly, we tested by excluding other autocorrelation effects, e.g., when feature ATSCNZ is included in the final feature set, the feature importance ranking rates it low (see Figure S6 in the Supporting Information). Therefore, we conclude that autocorrelation on the basis of elemental proximity and similarity, perhaps in the range of three to seven atoms, correlates most strongly with additive performance.

The first cycle of learning was designed to propose a hypothesis for the underlying additive feature governing crystal size, resulting in the identification of ATSC5Z. The second cycle of learning was designed to validate the hypothesis using additives with low and high ATSC5Z values (factual and counterfactual), see Figure S7 in the Supporting Information. The results confirmed our hypothesis: we observed larger crystals with the additive of low ATSC5Z (additive 25). To increase crystal size further, further experiments would be needed with this specific goal, potentially focusing on exploring further additives with ATSC5Z values around the optimal.

Alternatively, we also considered a different framework without the recursive feature elimination step. The Shapley value analysis with our small data set (seven additive molecules in the first cycle and extra two molecules in the second cycle) shows that between the first and second cycle, no Mordred feature has a higher rank than processing condition features, and the Mordred features' influence on crystal size is significantly lower, as shown in Supporting Information, Figure S8. The inconsistency of the top Mordred features rank also makes it harder to pinpoint the future exploration direction, showing the importance of having the complete sequence of RFE-RF-SHAP.

Although the chemical significance of these features for crystal growth is unclear, similar cheminformatics descriptors have been previously used successfully in drug discovery.[32−34] Future work, perhaps including simulation and/or theory, may clarify the possible relationship between autocorrelation features, additive performance, and underlying chemical mechanism.

4. CONCLUSIONS

We identified a family of 29 commercially available additive molecules to help the crystallization of $MAPbI_3$ single crystals. We used high-throughput experimentation guided by the data analysis to prioritize laboratory testing of nine additives in two experimental cycles. This approach yielded one hitherto unreported additive, (3) 6,6′-dimethyl-2,2′-dipyridyl, which resulted in large crystal sizes in our study (the highest top 10th percentile single crystal size of $MAPbI_3$ under our specific processing conditions). Applying Tanimoto similarity to the additive molecule candidates had helped in combatting selection bias and prioritizing which molecules to synthesize among a large candidate list. We built a data workflow for hypothesis generation by combining recursive feature elimination, classification, and Shapley feature importance rank algorithms. Using this workflow, we obtained and tested a hypothesis related to the topological property of the additive to the resulting crystal size distribution.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.cgd.2c00522.

Further discussions on AATSC5Z, hypothesis testing simulation description, robotic workflow summary, effects of additive impurities on the crystal size outcome, additives solubility table, the crystal image examples, the root mean square error (RMSE) for AATSCNZ (N = 3, 4, 5, 6, 7) and their Shapley feature importance rank, the summary of ATSC5Z values, and the fabricated additives' crystal size distribution (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Noor Titan Putri Hartono − *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Berlin 12489, Germany;* orcid.org/0000-0002-0748-0620; Email: noortitan@alum.mit.edu

Tonio Buonassisi − *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;*

https://doi.org/10.1021/acs.cgd.2c00522
Cryst. Growth Des. XXXX, XXX, XXX−XXX

orcid.org/0000-0001-8345-4937; Email: buonassisi@mit.edu

Authors

Mansoor Ani Najeeb − *Department of Chemistry, Haverford College, Haverford, Pennsylvania 19041, United States;* orcid.org/0000-0002-8258-0613

Zhi Li − *Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States*

Philip W. Nega − *Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States*

Clare A. Fleming − *Department of Chemistry, Haverford College, Haverford, Pennsylvania 19041, United States*

Xiaohe Sun − *Department of Chemistry, Haverford College, Haverford, Pennsylvania 19041, United States*

Emory M. Chan − *Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States;* orcid.org/0000-0002-5655-0146

Antonio Abate − *Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Berlin 12489, Germany*

Alexander J. Norquist − *Department of Chemistry, Haverford College, Haverford, Pennsylvania 19041, United States;* orcid.org/0000-0003-1040-5880

Joshua Schrier − *Department of Chemistry, Fordham University, The Bronx, New York 10458, United States;* orcid.org/0000-0002-2071-1657

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.cgd.2c00522

Author Contributions

M.A.N., Z.L., P.N., C.A.F., and X.S. synthesized the perovskite single crystals. N.T.P.H. and J.S. performed the machine learning analysis. E.C., A.A., A.N., J.S., and T.B. supervised the study. N.T.P.H., M.A.N., and T.B. contributed to the writing of the paper. All the authors discussed the results and implications of the work and edited the paper.

Notes

The authors declare the following competing financial interest(s): Although our laboratory has IP filed covering photovoltaic technologies and materials informatics broadly, we do not envision a direct COI with this study, for which code & data are open sourced. One of the authors (T.B.) own equity in a startup company, Xinterra, which applies machine learning and high-throughput experimentation to develop new materials, and two authors (A.J.N. and J.S.) are scientific advisors for a startup company, Atinary, commercializing machine-learning enhanced experiment planning for materials. The code and experimental data are available in the GitHub repository: https://github.com/PV-Lab/PvkAdditives. The Mordred descriptors are implemented using Mordred package version 1.2.0.[20] The recursive feature elimination and the random forest regression are implemented using scikit-learn package,[27] and the feature importance rank is implemented using the SHAP package.[30]

■ ACKNOWLEDGMENTS

This study is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under contract no. HR001118C0036. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA. Work at the Molecular Foundry was supported by the Office of Science, Office of Basic Energy Sciences of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. J.S. acknowledges the Henry Dreyfus Teacher-Scholar Award (TH-14-010).

■ REFERENCES

(1) Zhang, W.; Eperon, G. E.; Snaith, H. J. Metal Halide Perovskites for Energy Applications. *Nat. Energy* **2016**, *1*, 1−8.

(2) Wang, K.; Wu, C.; Yang, D.; Jiang, Y.; Priya, S. Quasi-Two-Dimensional Halide Perovskite Single Crystal Photodetector. *ACS Nano* **2018**, *12*, 4919−4929.

(3) Chin, X. Y.; Cortecchia, D.; Yin, J.; Bruno, A.; Soci, C. Lead Iodide Perovskite Light-Emitting Field-Effect Transistor. *Nat. Commun.* **2015**, *6*, No. 7383.

(4) Ball, J. M.; Petrozza, A. Defects in Perovskite-Halides and Their Effects in Solar Cells. *Nat. Energy* **2016**, *1*, 1−13.

(5) Lei, Y.; Chen, Y.; Xu, S. Single-Crystal Halide Perovskites: Opportunities and Challenges. In *Matter*; Cell Press, 2021; pp 2266−2308. DOI: 10.1016/j.matt.2021.05.002.

(6) Lewis, A.; Marcelo, S.; Herman, K.; van Gerda, R. *Industrial Crystallization: Fundamentals and Applications*; Cambridge University Press, 2015.

(7) Zhang, W.; Zhu, H.; Pan, S.; Li, H.; Zhang, J.; Gong, Z.; Zhang, Y.; Pan, J. Growth and Properties of Centimeter-Sized Lead Free All Inorganic Perovskite $Cs_2AgBiBr_6$ Crystal by Additive CH_3COONa. *J. Cryst. Growth* **2020**, *532*, No. 125440.

(8) Feng, Y.; Pan, L.; Wei, H.; Liu, Y.; Ni, Z.; Zhao, J.; Rudd, P. N.; Cao, L. R.; Huang, J. Low Defects Density CsPbBr3single Crystals Grown by an Additive Assisted Method for Gamma-Ray Detection. *J. Mater. Chem. C* **2020**, *8*, 11360−11368.

(9) Liu, Y.; Zheng, X.; Fang, Y.; Zhou, Y.; Ni, Z.; Xiao, X.; Chen, S.; Huang, J. Ligand Assisted Growth of Perovskite Single Crystals with Low Defect Density. *Nat. Commun.* **2021**, *12*, No. 1686.

(10) Chen, J.; Kim, S. G.; Ren, X.; Jung, H. S.; Park, N. G. Effect of Bidentate and Tridentate Additives on the Photovoltaic Performance and Stability of Perovskite Solar Cells. *J. Mater. Chem. A* **2019**, *7*, 4977−4987.

(11) Kirman, J.; Johnston, A.; Kuntz, D. A.; Askerka, M.; Gao, Y.; Todorović, P.; Ma, D.; Privé, G. G.; Sargent, E. H. Machine-Learning-Accelerated Perovskite Crystallization. *Matter* **2020**, *2*, 938−947.

(12) Li, Z.; Najeeb, M. A.; Alves, L.; Sherman, A. Z.; Shekar, V.; Cruz Parrilla, P.; Pendleton, I. M.; Wang, W.; Nega, P. W.; Zeller, M.; Schrier, J.; Norquist, A. J.; Chan, E. M. Robot-Accelerated Perovskite Investigation and Discovery. *Chem. Mater.* **2020**, *32*, 5650−5663.

(13) Nega, P. W.; Li, Z.; Ghosh, V.; Thapa, J.; Sun, S.; Hartono, N. T. P.; Nellikkal, M. A. N.; Norquist, A. J.; Buonassisi, T.; Chan, E. M.; Schrier, J. Using Automated Serendipity to Discover How Trace Water Promotes and Inhibits Lead Halide Perovskite Crystal Formation. *Appl. Phys. Lett.* **2021**, *119*, No. 041903.

(14) Li, Z.; Nega, P. W.; Nellikkal, M. A. N.; Dun, C.; Zeller, M.; Urban, J. J.; Saidi, W. A.; Schrier, J.; Norquist, A. J.; Chan, E. M. Dimensional Control over Metal Halide Perovskite Crystallization Guided by Active Learning. *Chem. Mater.* **2022**, *34*, 756−767.

(15) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J. Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis. *Nature* **2019**, *573*, 251−255.

(16) Tanimoto, T. *An Elementary Mathematical Theory of Classification and Prediction*; International Business Machines Corp., 1958; http://dalkescientific.com/tanimoto.pdf.

(17) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.

(18) Schneider, C. A.; Rasband, W. S.; Eliceiri, K. W. NIH Image to ImageJ: 25 Years of Image Analysis. *Nat. Methods* **2012**, *9*, 671−675.

(19) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(20) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminf.* **2018**, *10*, 1−14.
(21) Bergstra, J.; Ca, J. B.; Ca, Y. B. Random Search for Hyper-Parameter Optimization Yoshua Bengio. *J. Mach. Learn. Res.* **2012**, *13*, 281−305.
(22) Zhang, H.; Liu, X.; Dong, J.; Yu, H.; Zhou, C.; Zhang, B.; Xu, Y.; Jie, W. Centimeter-Sized Inorganic Lead Halide Perovskite $CsPbBr_3$ Crystals Grown by an Improved Solution Method. *Cryst. Growth Des.* **2017**, *17*, 6426−6431.
(23) Pendleton, I. M.; Cattabriga, G.; Li, Z.; Najeeb, M. A.; Friedler, S. A.; Norquist, A. J.; Chan, E. M.; Schrier, J. Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): A Software Pipeline for Automated Chemical Experimentation and Data Management. *MRS Commun.* **2019**, *9*, 846−859.
(24) Van Der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2625.
(25) Chari, T.; Banerjee, J.; Pachter, L. The Specious Art of Single-Cell Genomics. *bioRxiv* **2021**, 1−25.
(26) eMolecules Inc. eMolecules, 2022. https://www.emolecules.com/ (accessed February 18, 2022).
(27) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(28) Sanchez-Lengeling, B.; Wei, J. N.; Lee, B. K.; Gerkin, R. C.; Aspuru-Guzik, A.; Wiltschko, A. B. *Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules*. 2019, arXiv:1910.10685. arXiv.org e-Print archive. http://arxiv.org/abs/1910.10685 (accessed June 30, 2022).
(29) Leach, A. R.; Gillet, V. J. *An Introduction to Cheminformatics*; Springer, 2007.
(30) Lundberg, S. M.; Lee, S. I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process Syst.* **2017**, 4766−4775.
(31) Moreau, G.; Broto, P. Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv. J. Chim.* **1980**, *4*, 359−360.
(32) Srisongkram, T.; Weerapreeyakul, N. Route of Intracellular Uptake and Cytotoxicity of Sesamol, Sesamin, and Sesamolin in Human Melanoma SK-MEL-2 Cells. *Biomed. Pharmacother.* **2022**, *146*, No. 112528.
(33) Minnich, A. J.; Mcloughlin, K.; Tse, M.; Deng, J.; Weber, A.; Murad, N.; Madej, B. D.; Ramsundar, B.; Rush, T.; Calad-Thomson, S.; Brase, J.; Allen, J. E. AMPL: A Data-Driven Modeling Pipeline for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 1955−1968.
(34) Ponzoni, I.; Sebastián-Pérez, V.; Requena-Triguero, C.; Roca, C.; Martínez, M. J.; Cravero, F.; Díaz, M. F.; Páez, J. A.; Arrayás, R. G.; Adrio, J.; Campillo, N. E. Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery. *Sci. Rep.* **2017**, *7*, No. 2403.

https://doi.org/10.1021/acs.cgd.2c00522
Cryst. Growth Des. XXXX, XXX, XXX−XXX