Can machines 'learn' halide perovskite crystal formation without accurate physicochemical features?

*Ian M. Pendleton[1], Mary K. Caucci[2], Michael Tynes[2,3], Aaron Dharna[3], Mansoor Ani Najeeb Nellikkal[1], Zhi Li[4], Emory M. Chan[4], Alexander J. Norquist[1], Joshua Schrier [2]**

[1] Department of Chemistry, Haverford College, 370 Lancaster Avenue, Haverford, Pennsylvania 19041, USA

[2] Department of Chemistry, Fordham University, 441 E. Fordham Road, The Bronx, New York, 10458, USA

[3] Department of Computer and Information Science, Fordham University, 441 E. Fordham Road, The Bronx, New York, 10458, USA

[4]Molecular Foundry, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California, 94720, USA

* jschrier@fordham.edu

Abstract

Discovery of new perovskite materials is motivated by a broad range of materials applications and accelerated by recent advances in machine learning (ML). We herein report dataset augmentation, benchmarking, and interrogation for an ongoing experimental campaign consisting of 9,483 halide perovskite single crystal experiments. To address limitations in previous work, we developed an improved description of the reactant concentrations in the experiments (validated against experimental observations) and performed experiments quantifying the excess volume of mixing of gamma-butyrolactone/formic acid mixtures used in the perovskite syntheses. Combining this improved description of reactant concentration with other physicochemical features of the reactants, we constructed 1108 ML models to elucidate the roles of the algorithm (k-nearest neighbors, linear support-vector machine, and gradient boosted tree), feature set (12 in total), preprocessing regime (e.g., standardization), and training data holdout scheme on ML predictive ability. ML comparisons illustrated that the chemical accuracy of less sophisticated physical models in a dataset do not hinder interpolative model performance. Analysis of feature contributions showed how ML models 'learn' competitive representations for concentration using raw experimental descriptions. Interrogation of the most performant models indicated that the numerical values of physicochemical features were not important, rather these features were being used to identify and interpolate within a particular reactant set. ML models were shown to be capable of making rudimentary extrapolations to untrained chemical systems when compared against basic benchmarks, and models which included the newly developed chemical features were shown to be more reliable than models trained without. These results illustrate how a stepwise comparative approach to machine

learning can provide insight into *what* and *how much* models are ‘learning’ for a given prediction task.

Introduction

Continued research into metal halide perovskites has improved photovoltaic power conversion efficiency from 3.8% in 2009 to 25.2% in 2019,[1–5] and has led to increased chemical diversity,[6,7] novel optoelectronic devices,[8–11] sensors,[12] and batteries.[13] Accelerated discovery of new perovskite compositions has been achieved through high-throughput synthesis of nanomaterials,[14] thin films,[15–18], and single crystals.[19] Single crystal perovskites are especially important as they have longer charge diffusion lengths,[5,20] higher carrier mobilities,[21,22] higher optoelectronic performance than thin film analogues,[23,24] and the atomic structures obtained via single-crystal X-ray diffraction are a starting point for first principles simulations. Predicting perovskite crystallization is difficult because the precursor solutions are concentrated, non-ideal electrolytes,[25–27] subject to reagent compositional variation,[28,29] and sensitive to potentially uncontrolled experimental conditions such as temperature and humidity.[30] As a result, most efforts to grow novel perovskite single crystals proceed through trial and error.

There is precedent for predicting crystal formation using machine learning (ML). For example, a radial basis function (RBF) support vector machine (SVM) trained on a dataset of 35,858 small organic molecules predicted crystallizability of previously unreported compounds with 79% classification accuracy using a dataset consisting of 35,858 small organic molecules,[31] a random forest model trained on 1,948 crystal structures predicted the crystallizability of full-Heusler compounds with a true positive rate of 0.94,[32] and an SVM trained on historical reaction data including successful and failed reactions (3,955 entries) predicted the formation of novel vanadium selenites with 79% classification accuracy.[33] Furthermore, machine learning has been used in an iterative fashion to actively explore and optimize polyoxometalate crystal growth.[34,35] These examples suggest the feasibility of halide perovskite crystal formation prediction.

In our recent work describing a dataset of high-throughput inverse temperature crystallization (ITC) perovskite syntheses (RAPID),[19] we showed that ML models trained on a small number of experiments for a given set of reactants can successfully predict reaction outcomes for new experiments performed using those reactants. However, this type of reaction prediction does not exactly correspond to the problem of discovering new materials. At the most general level, an experiment is described by the reaction *conditions* (e.g., temperature, time, composition) for a set of *reactants* (e.g., lead iodide, methylammonium iodide, solvent). Machine learning (ML) models can be trained to 'learn' the underlying relationships between inputs (*conditions*, *reactants)* and experimental outcomes, and then evaluated on a test set of experimental data. When the same *reactant sets* exist in both the training and testing data, model performance measures only the quality of a prediction at the level of *conditions*. This type of performance is measured when using training and testing sets comprised of only a single reactant set, or performs a randomly selected standard test-train-split (STTS) across experiments. In contrast, the problem of developing a new material corresponds to predicting outcomes for a novel *reactant set* that are not in the training data. This type of performance is measured by a leave-one-out (LOO) test-train split at the level of reactants. A successful "extrapolation" to new *reactant sets* relies upon having a sufficiently rich numerical description such that new experiments can be related to the past experiments, i.e., so that one can interpolate between the descriptions of the different reactants. The models in Ref. 19 successfully predicted reaction outcomes for experiments at novel conditions only when they had training data for the reactants; that is they successfully interpolated only at the level of *conditions*, but failed to extrapolate to new *reactants*. This is surprising because it included molecular features that have previously been used to successfully predict hydrothermal synthesis of amine-templated metal oxides at the level of *reactants*.[19,33,36] This failure suggested a deficiency in the existing experiment description.

One possible experimental description deficiency involves reaction concentrations, which could prevent finding similarities between experiments using different reactants. For example, both nonideality of mixing and inadequacies in how the reactant concentrations are computed following a series of automated dispense steps can be problematic. In principle, corrections for these issues could be "learned" from a sufficient dataset of raw observations, but it is unclear whether they can be learned from a typical experimental dataset. As a ground truth, we used experiment and chemical insight to develop physical based models for both effects. We then performed an extensive set of machine learning calculations to determine: (1) whether ML can 'learn' a way out of a bad physical model of concentration using raw experimental data; (2) what features are utilized to 'learn' these properties; and (3) the extent to which these improved descriptions improve our ability to predict the formation of novel perovskite crystals.

Methods

SolV Model Volumes

Our typical experiment consists of the following steps: (1) preparation of precursor solutions following a volumetric and gravimetric recipe, (2) volumetric dispensing of the precursors into reaction vials by a liquid handling robot, (3) characterization and classification of the experimental outcome based on the size and quality of the crystals formed during the experiment. Determining reactant concentration following step 2 requires calculation of chemical concentrations in the precursor solutions from step 1. The original calculation of the chemical concentration in an experiment, denoted herein as the '**Sol**vent **V**olume only' model or '**SolV**', made two assumptions which simplified early data processing. First, SolV assumes that the volume of precursor mixing is additive. This assumption is valid when the solvent molecules have comparable sizes and shapes, but more often there is some positive or negative excess volume of mixing (i.e., $V_m^E \neq 0$). This assumption can be critiqued by experimental measurements of the excess volume of mixing. Second, SolV neglects volume displacement by solutes in precursor solution. This assumption is valid in the limit of dilute solutions, but fails for concentrated solutions. The "concentration" computed in this way is initially proportional to molality, but when these solutions are subsequently dispensed volumetrically (by an automated liquid handler) to create the reaction composition, the unknown solution density introduces an idiosyncratic deviation in the computed concentrations of the final reaction unique to each stock solution, precluding a direct comparison across different experiments.

SolUD Model Volumes

The SolV model can be critiqued by developing and comparing an alternate model that incorporates this volume change. To this end, we propose the '**Sol**utions **U**sing **D**ensity' or

‘**SolUD’** model which makes the following assumptions: (1) the volume of the solution is the sum of the volumes of the solvents and solutes, (2) a dissociable ionic solute’s volume is the sum of the molecular or ionic volumes of its component ions, (3) deviations from the previous two assumptions are approximately linear, and can be corrected empirically.

The first challenge to development of the SolUD model is determining the most practical means of calculating the displacement volume of a solute. One approach is to use experimental bulk crystal densities. This is especially applicable to lead diiodide which is present in the form of 80-100 nm lead iodide colloids during the ITC process.[25,27] For some of the organoammonium salts considered here, densities are also available from crystallographic databases. Experimental bulk crystal densities provide an upper bound of the total volume displacement, because bulk crystals contain void spaces which lower the density. Alternatively, the volume displacement can be computed from the molecular van der Waals volume of a given organoammonium species combined with experimental ionic radius of the halide counterion. An advantage of this approach is that it is applicable to solutes that do not have known crystal structures. In either case, experimentally determined or calculated, the density can be used to determine the final volume change of a solution upon addition of the solute species and thereby improve the concentration calculation for precursor solutions.

The SolUD model approximates total solution volume as defined by the sum of solvent and solute volumes. The process of calculating SolUD solution volumes can be divided into three steps: (1) calculation of an approximate density using molecular and ionic volumes, (2) application of an empirical correction to the density to compute an approximated bulk crystal density, (3) summation of the species volumes. For the first step, the density derived from volumes of the molecular species (ρ_a) is estimated as

$$\rho_a = \frac{m_o + n_h m_h}{V_o + n_h V_h}, \#(1)$$

where m_o and m_h are the organoammonium cation and halide masses (respectively), V_o is the organoammonium cation van der Waals volume (computed using ChemAxon[37] cxcalc 19.27.0 from the cation's SMILES string), V_h is the ionic volume of halide ion (determined from the tabulated ionic radii),[38] and n_h is the number halide anions present in the salt.

Next, an empirical correction to the calculated ρ_a values was generated through comparison with the subset of organoammonium salts with experimental bulk densities available in the Cambridge Structural Database (CSD).[39] A total of four linear corrections were obtained to transform calculated ρ_a values to the bulk crystal densities (ρ_b) using equation (2): $\rho_{b_{combined}}$, ρ_{b_I}, $\rho_{b_{Br}}$, and $\rho_{b_{Cl}}$.

$$\rho_b = (m * \rho_a) + b \#(2)$$

where m and b are empirical parameters. Using the calculated bulk crystal density values, we generated a dataset describing computed volumes of solutions derived from SolV, $\text{SolUD}_{\text{Mol}}$ (volume derived from molecular volumes), and SolUD (volume derived from ρ_b calculated densities). We compared these values to experimentally observed solution volumes for precursor solutions prepared from lead diiodide, organoammonium iodide and γ-butyrolactone (GBL). The SolUD volumes were calculated by taking the sum of the volumes from each solid species added to the total solvent volume,

$$V_{total} = V_{solvent} + \sum_i \frac{m_i}{\rho_i}, \#(3)$$

where m_i and ρ_i are the added mass and the density for each solute *i*. The values for each organoammonium salt density ρ_a and ρ_b are reported in the file 'OrganoammoniumDensityDataset.xlsx' included in the supporting information. The dataset also includes associated refcodes for each structure obtained from CSD, van der Waal's volumes obtained from ChemAxon, and details of all described calculations. The dataset comparing experimentally observed volumes to SolV and SolUD volumes is included in the supporting information as, 'SolutionVolumeDataset.xlsx'.

V_m^E Experimental Method

Determination of excess molar volume (V_m^E) of mixing was performed following the density-measurement procedure described by Lunelli and Scagnolari.[40] The density, $\rho(i)$, of a FAH/GBL stock solution, *i,* is obtained by sequential mass measurement of a filled Hamilton 1710 TLC analytical syringe using a Mettler Toledo B204S balance. The molar volume of each stock solution, $V_m(i)$, is determined from,

$$V_m(i) = \frac{v_{FAH}(i)\rho_{FAH} + v_{GBL}(i)\rho_{GBL}}{\rho(i)\left[v_{FAH}(i)\rho_{FAH}M_{FAH}^{-1} + v_{GBL}(i)\rho_{GBL}M_{GBL}^{-1}\right]} \#(4)$$

where ρ_x, $v_X(i)$, and M_X are the density, volume added for stock solution *i*, and molecular weight of the neat chemical *X*. The excess molar volume (V_m^E) for each mole fraction,

$$V_m^E(i) = V_m(i) - \left[x_{FAH}(i)\,V_{m,\ FAH} + x_{GBL}(i)V_{m,\ GBL}\right] \#(5)$$

where $x_{FAH}(i)$ and $x_{GBL}(i)$ are the mole fractions of FAH and GBL in a given stock solution and $V_{m,\ GBL}$ and $V_{m,\ FAH}$ are the molar volumes of the neat solutions of GBL and FAH. The dataset of experimental measurements ('ExcessMolarVolumeData.xlsx'), excess volume calculations, and Mathematica 12.0 and Python 3.7 code files used for curve fitting are included with the supporting information.

Experimental Dataset

Experimental data are taken from ongoing work in high-throughput robotically-driven inverse temperature crystallization (ITC) workflow, comprising 9,483 organic-inorganic metal halide perovskites crystallization experiments at the time of this study. A detailed description of the ITC workflow and experimental process can be found in the supporting information as well as in past publications.[19] Experimental data capture and reporting, including precursor preparation, materials monitoring data, and final data augmentation with concentrations and physicochemical descriptors is performed using the ESCALATE 'capture' platform (v2.57).[41] The SolUD and SolV derived concentration values were incorporated into the ESCALATE 'report' code (at v0.7),[42] and were applied to the entirety of the dataset, including retroactive experiments.

For this article, the larger perovskite dataset was filtered to include only: (1) ITC experiments (designated "Workflow 1.1"), (2) reactions that use GBL as a solvent, (3) reactant sets containing at least one instance of 'success' (large single crystal). These restrictions reduce the dataset to 5,049 unique experiments spanning 19 unique reactant sets.[43] Each experiment has 423 features. A complete description of dataset is included as '0045.perovskitedata.csv' in supporting information.

Machine Learning Models for Perovskite Crystallization

The development pipeline for ML models includes a data preprocessing regime followed by the fitting and evaluation of multiple ML models, varying both feature sets data hold out schemes (Figure 1).

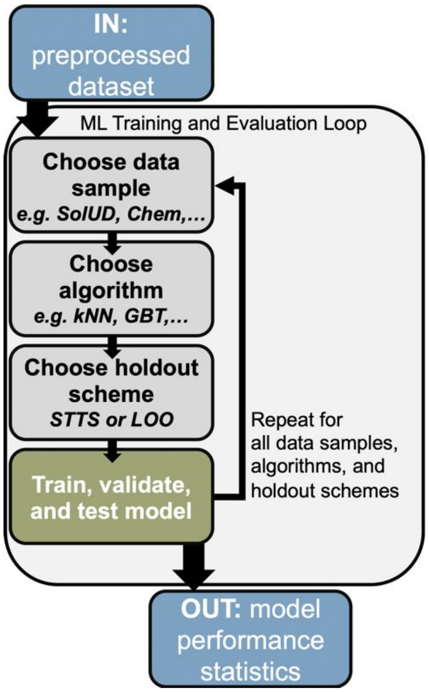
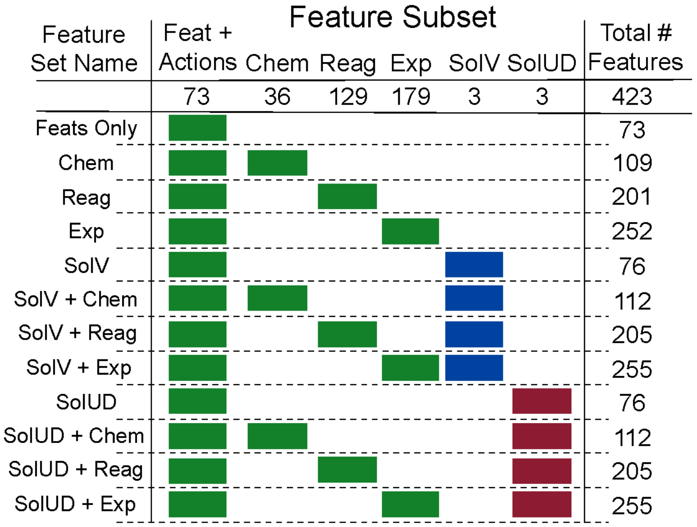

IN: preprocessed dataset
ML Training and Evaluation Loop
Choose data sample
e.g. SolUD, Chem,...
Choose algorithm
e.g. kNN, GBT,...
Choose holdout scheme
STTS or LOO
Train, validate, and test model
Repeat for all data samples, algorithms, and holdout schemes
OUT: model performance statistics

Figure 1. Overview of computational pipeline for machine learning

A total of 1108 models were trained, evenly divided between **s**tandard **t**rain-**t**est **s**plits (STTS) and a **l**eave-**o**ne-amine-**o**ut (LOO) holdout scheme. During a STTS split, datasets were divided into six folds: five folds were used for cross-validation and hyperparameter optimization and one fold was used for testing. Reported results are based on test-fold performance. In the LOO holdout scheme, each ammonium halide salt was treated as the test set while data on the remaining 18 ammonium halide salts were used for hyperparameter tuning and cross-validated model training. Both STTS and LOO holdout sets were constructed through random sampling of the larger dataset without replacement.

To study the role of different types of experimental and computational features for prediction, we grouped the features into four categories or "feature subsets": *"Chemicals"* (*Chem*), *"Reagents"* (*Reag*), "*Experiment*" (*Exp.*), and "*Features*" (*Feat*). *Chem* includes chemical quantities, such as the masses and volumes of each chemical, used in the preparation of "*Reagents*" (i.e., precursor solutions) along with the dispense volume of each reagent into experiments. *Reag* includes information such as the concentration of the final reagent solutions and the volume of each reagent dispensed into experiments. *Exp* contains all data captured by the workflow required to prepare the final experiment solution. Finally, *Feat (Feat + Actions)* includes only the physicochemical descriptors (e.g., surface area, polarity, volume, number of rings) of the organic reagents and the actions associated with experiments. To emphasize the comparison of concentration descriptors central to this article, use of "SolV" and "SolUD" feature subsets are explicitly denoted. From these feature subsets, we constructed 12 feature sets by systematically varying the included feature subsets, as outlined in Figure 2.



Feature Set Name	Feature Subset						Total # Features
	Feat + Actions	Chem	Reag	Exp	SolV	SolUD	
	73	36	129	179	3	3	423
Feats Only	🟩						73
Chem	🟩	🟩					109
Reag	🟩		🟩				201
Exp	🟩			🟩			252
SolV	🟩				🟦		76
SolV + Chem	🟩	🟩			🟦		112
SolV + Reag	🟩		🟩		🟦		205
SolV + Exp	🟩			🟩	🟦		255
SolUD	🟩					🟥	76
SolUD + Chem	🟩	🟩				🟥	112
SolUD + Reag	🟩		🟩			🟥	205
SolUD + Exp	🟩			🟩		🟥	255

Figure 2. Overview of the 12 unique feature sets generated from individual feature subsets of the main perovskite dataset

Machine learning (ML) models were built with Scikit-Learn for Python.[44] Common ML algorithms were chosen, including: gradient boosted trees (GBT), k-nearest neighbors (kNN), and linear support vector machines (L-SVC). A minimum performance baseline was constructed using GBT models trained on two "nonsense" datasets the "Y" and "deep" shuffled datasets, which were generated by shuffling inputs relative to outputs or shuffling all data within each feature set, respectively. Models generated from Y shuffled datasets with high correlation indicate overfittings or spurious correlations between model inputs and the outputs; any seemingly performant model generated on the original dataset would be suspect.[45,46] An additional performance baseline was provided by kNN where k=1, which can be thought of as 'memorization' of the data.[47] Because measures such as precision and accuracy are often deceptive for datasets with large class imbalances, we instead report the Matthews correlation coefficient (MCC), which does not suffer from this problem.[48] Models that only predict the majority class will have an MCC of zero; an MCC of '1' corresponds to perfect prediction of both successes and failures. Preprocessing variations include one hot-encoding (*OHE*) of the organoammonium identity with simultaneous removal of physicochemical descriptors, normalization (*norm*), and standardization (*std*). Normalization and standardization processes were divided into three increasingly general tiers: (0) preprocessing on Feats* only, (1) preprocessing on SolUD and SolV, and (2) the entire dataset. These models were computed using Texas Advanced Computing Center (TACC) resources and infrastructure. Additional details regarding model design, optimized hyperparameters, and related code is documented in the supporting information.

Results and Discussion

There are two competing philosophies for machine learning in science. Historically, most scientists have tried to use domain expertise to improve the underlying descriptions on which the machine learning work is based. This feature development technique is especially important when working with relatively small experimental datasets, as it reduces the complexity of the model. It also has the benefit of facilitating more human-interpretable models,[33,49] but requires a significant investment in human expertise and may introduce anthropogenic biases.[36] Alternatively, a complicated model can be trained to make high quality predictions directly from the raw inputs, given a sufficiently large dataset. However, this approach may introduce a series of potential problems. The models may be uninterpretable, they may extrapolate in nonsensical ways outside the training data, and they may not 'learn' what you think they are learning, but instead focus on scientifically irrelevant details in the dataset.[50]

Here, we have the opportunity to compare these two approaches directly. We first describe the development of improved features for describing reaction composition using the traditional toolbox of physical chemistry. Any models introduced are kept intentionally simple, e.g., correcting systematic trends using linear models. We also quantify other factors that would compete with these descriptions of composition (e.g., nonideality of mixing). This allows us to perform a direct comparison between the physicochemical model approach and the purely data driven approach, which we analyze in the context of ML model generation and evaluations. Finally, an exploration of model performance in the context of leave-one-out (LOO) is performed. This workflow (data analysis, model generation, model performance evaluation) enables critical analysis of the input data as well as elucidating the effect of input data on machine learning model utility.

Determination of bulk density

Densities derived from the organoammonium iodides van der Waals volumes and tabulated iodide ion volumes were calculated and compared to available experimental values from CSD (reported at 295K). Similar linear regression metrics including R-squared, root mean squared error (RMSE), and mean absolute error (MAE), were observed for organoammonium iodide, bromide, and chloride salts (Figure 3, Table S1). A high correlation (R^2 = 0.96) between experimental and computed density values is observed in Figure 3, but the values computed using van der Waals and ionic radii fail to account for void spaces in the bulk crystal and therefore systematically overestimate density of ammonium halide salts. An empirical correction was developed for each halide independently. The correction to the organoammonium iodides ρ_{a_I} provides the computed bulk density ρ_{b_I} via equation (6).

$$\rho_{b_I} = \left(1.02 * \rho_{a_I}\right) - 0.76\#(6)$$

Additional information describing the derivation of bulk density for organoammonium salts can be found in the supporting information.

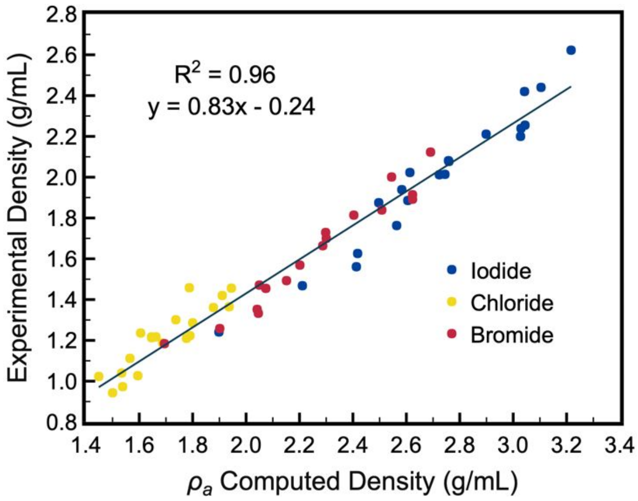


Figure 3. Experimental crystal densities versus computed densities of organoammonium halide salts

Comparison of SolV and SolUD Models

As our dataset contains only organoammonium iodides, we used ρ_{b_I} derived densities and the reported density of PbI_2 for subsequent volume calculations of precursor solutions. Experimentally measured volumes were available for 172 of the 219 precursor solutions. Comparing the observed volume to the calculated volumes illustrates the improvement provided by the SolUD volume model. A linear regression and associated metrics including R-squared, RMSE, and MAE are reported for SolV and SolUD (Table 1, Figure 4). The SolV model underestimates the observed volumes by an average of 30% across all reagents, whereas the SolUD model has one to one correspondence with the observed volumes and has a MAE within the precision of most observations (± 1 mL).

Table 1. Linear regression and model fit metrics comparing observed reagent volumes to those derived from SolV and SolUD volume models.

	SolV	SolUD
Slope (*m*)	1.3	1.0
Intercept *(b)*	1.6	0.2
R^2	0.94	0.96
RMSE[†] (mL)	3.1	2.0
MAE[‡] (mL)	2.2	0.9

[†] RMSE is root mean squared error. [‡] MAE is mean absolute error.

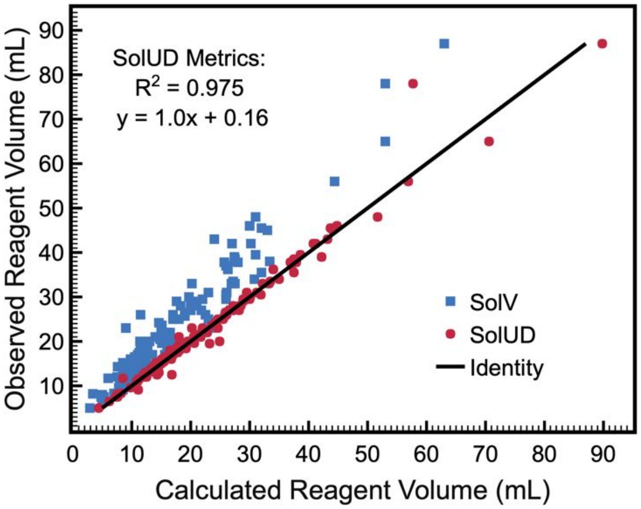


Figure 4. Observed solution volumes versus the SolV and SolUD calculated volumes

SolUD Volume Estimates for Dataset Auditing

Given the overall high performance of the SolUD volume estimates, we suspected that the outliers were the result of data entry problems. We used written laboratory notebooks to confirm the presence of data entry errors and then rectified where possible. Of the 219 solutions prepared, only 172 of them have experimental volume observations. Of these, four (4) had observed volumes that deviated by more than 10% from the SolUD predicted volume, and of these, one (1) was confirmed as a data entry error. An additional four (4) solutions were identified where the actual concentrations exceeded the maximum expected concentrations based on measured solubility limits. Of these, two (2) solution preparation logs were confirmed to have data entry errors. Since the implementation of these validation steps we have audited a total of eight stock solutions (effecting 768 experiments, 8.1% of the dataset) and rectified data entry errors associated with three stock solutions (effecting 288 experiments, 3% of the dataset).

Non-Ideal Mixing of GBL and FAH

Another possible source of error is the non-ideal volume of mixing. These perovskite ITC experiments consist predominately of two solvents: (1) GBL used in the reagent solutions and (2) formic acid (FAH) added in the final step of the experiment. We are unaware of prior experimental measurements of the excess molar volume of GBL:FAH mixtures. From the solution density measurements (eq 4), we calculated the excess molar volume as a function of mole ratios of FAH and GBL (eq 5), fitting the results to a cubic polynomial. These data provide a calculated maximum V_m^E = 1.33 mL / mol for a mole fraction of formic acid of 0.36 (Figure 5).

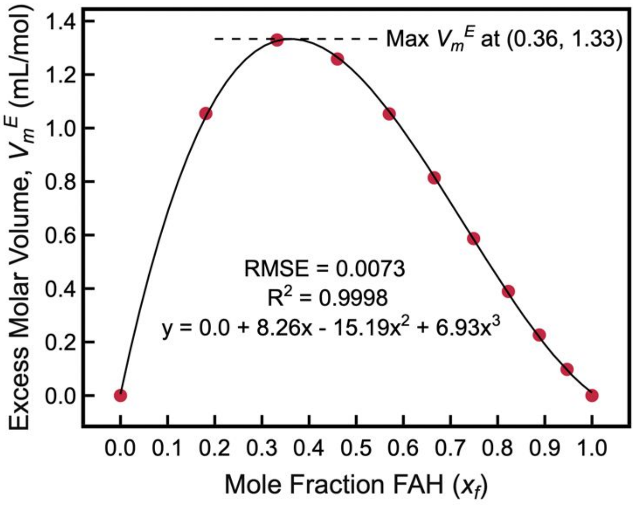


Figure 5. Third order polynomial fit of excess molar volume (V_m^E) versus mole fraction of FAH (x_f). The Maximum $\mathrm{V_m^E}$ of 1.33 mL/mmol occurs at a FAH mole fraction of 0.36.

The V_m^E data can be used to approximate the maximum error incurred for $\Delta \mathrm{V_{mix}}$ in our workflow process. Even in the worst case of 0.36 FAH: 0.64 GBL, neglecting the excess molar volume of mixing only underestimates the true volume by 2.02%, and hence overestimates the solute concentration by 2.06%. In comparison, neglecting the volume displaced by solutes (using SolV instead of SolUD), overestimates the solute concentration by an average of 24.1 ± 9.0% across all experiments, with a maximum and minimum concentration overestimation of 41.7% and 4.5% for the solute, respectively. These data suggest that the volume error due to non-ideal mixing would only marginally improve the SolUD model; therefore, we opted to omit $\Delta \mathrm{V_{mix}}$ contributions when computing SolUD concentrations. Calculations comparing the percent change in concentration values as well as a detailed derivation of V_m^E can be found in the supporting information.

Machine Learning Models for Perovskite Crystallization

The dataset described above —along with the improved concentration models— provides a unique opportunity to benchmark the performance of ML models for predicting perovskite crystallization with and without concentration features. In particular, we aimed to demonstrate: (1) that machine learning performance using an inferior SolV solution model was similar to the improved SolUD solution model (2) raw experimental characteristics could be used to match *or even exceed* model performance of physically accurate concentration features given enough experimental data and physicochemical features.

We first compared MCC for six-fold **s**tandard-**t**est-**t**rain-**s**plit (STTS) models using *SolV, SolUD,* and *SolUD + Chem* feature sets (Figure 6). The benchmark baselines of Y-shuffled or deep-shuffled data yielded an MCC of 0, as expected. Models with access to meaningful features outperform the shuffled baselines. Improving the concentration model alone (*SolV* versus *SolUD*) does not improve model predictions, but the additional inclusion of chemical features (*SolUD+Chem*) does. The GBT model only slightly outperforms 'the memorization' strategy of kNN where k=1. Using the complete training data (Figure 6a) gives better performance than the stratified training sets (Figure 6b) where 96 experiments are sampled from each organoammonium iodide. As this trend holds true for all STTS model comparisons, we will focus the remaining analysis on models trained using only the full dataset without stratification.

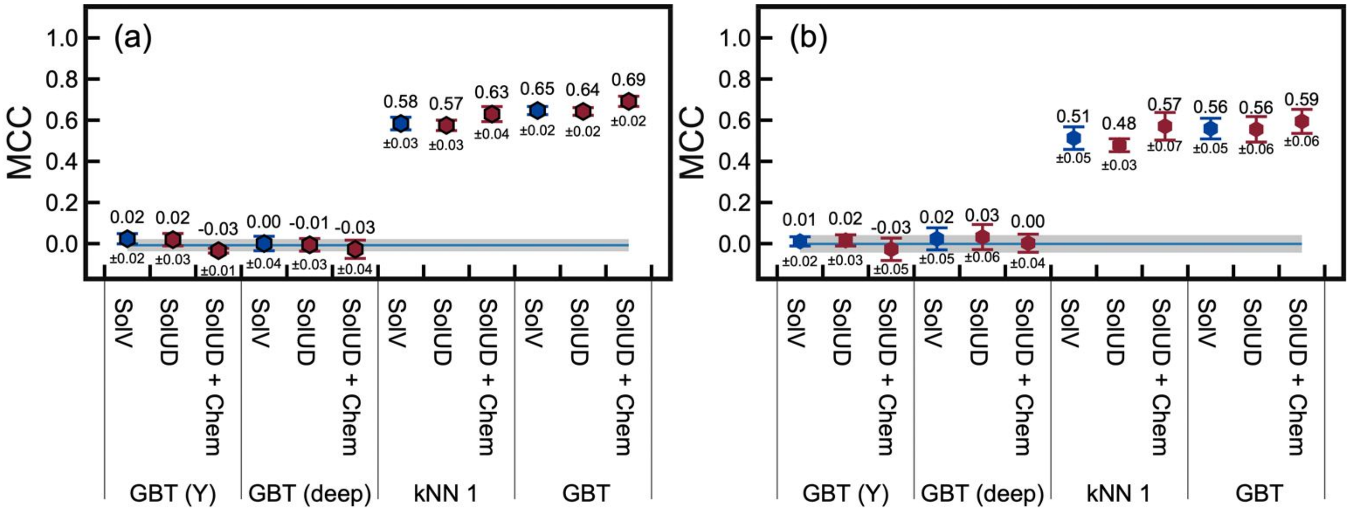


Figure 6. Baseline GBT performance for six-fold **s**tandard-**t**est-**t**rain-**s**plit (STTS) trained on (a) all members of the training set (b) stratified sample of 96 random experiments from each reactant set. Error bars show the standard deviation of performance for held-out data.

Comparing the concentration features (*SolV*, *SolUD*), physicochemical descriptors (*Feats Only*), and chemical descriptions (*Chem*), indicates that GBT is the highest performing algorithm. (Figure 7). Models trained on the *SolV* and *SolUD* features perform similarly despite the latter being a more faithful description of solution concentration. For all three algorithms, the *Chem* features, which consists of quantities of chemicals used in reagents and the volume of the reagents, are the most performant. GBT model performance is not largely impacted by the choice of feature set (Figure S6) nor by standardization or normalization of the input (Figure S8).

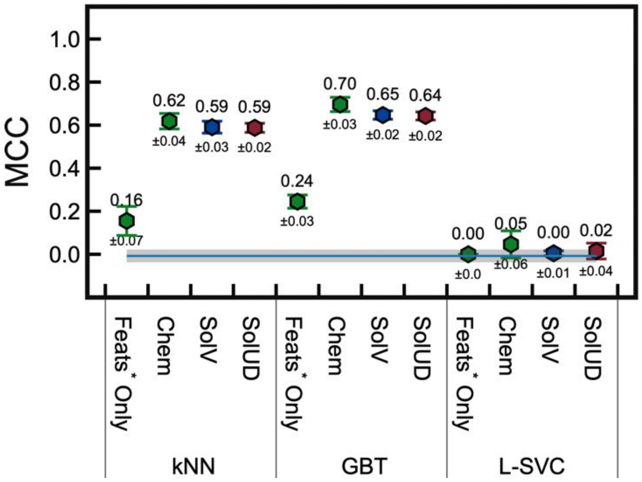


Figure 7. MCC comparison of kNN, GBT, and L-SVC models using STTS on various feature sets. Error bars and labels describe the standard deviation of performance for held-out data.

The best model overall was a GBT model fit on the *SolV* + *Chem* feature set using standardized physicochemical features (MCC = 0.71 ± 0.01, 109 total features). The best kNN models perform nearly as well as the best GBT models (kNN, *Chem*, standardizing all inputs, MCC = 0.66 ± 0.05, 109 total features). However, these additional *Chem* features provide only a small improvement to model performance. For comparison, the best model without them is GBT *Feat* + *SolUD* with MCC = 0.64 ± 0.02, using 76 total features. This small difference indicates that the additional *Chem* information about quantities and volumes provides limited information beyond the concentration features (*SolV* and *SolUD*) .

To better understand which data were most impactful to the model, we compared normalized feature contributions from GBT models targeting the *Chem, SolV, SolUD*, and *SolUD* + *Chem* feature sets (Figure 8).

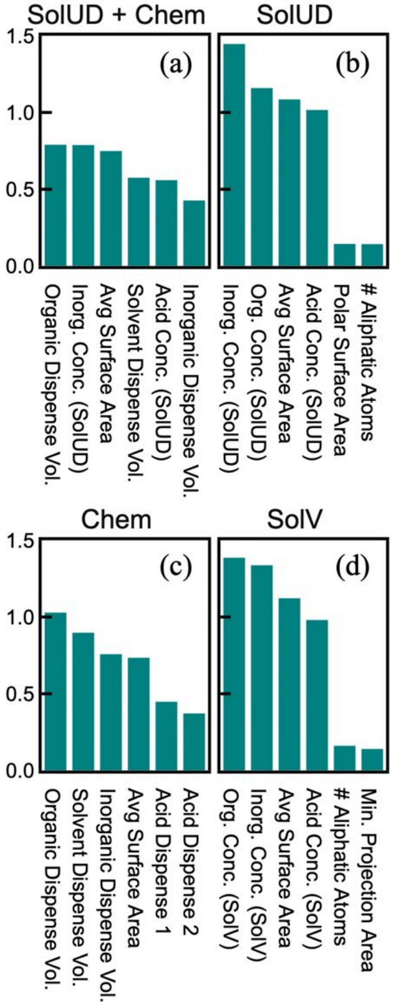
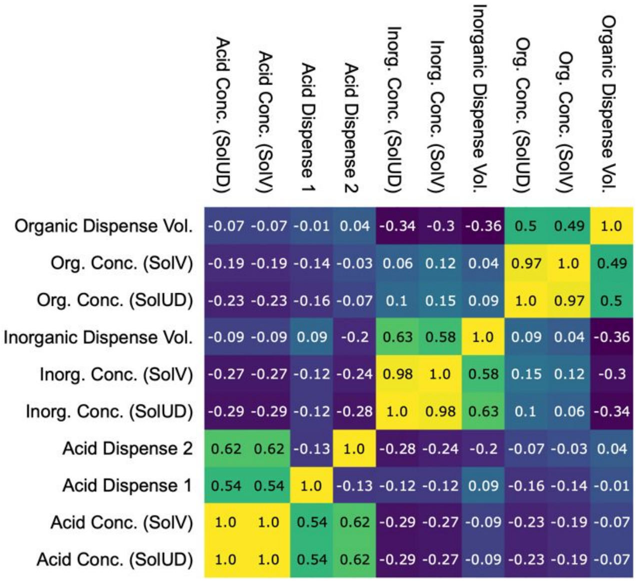


Figure 8. Normalized feature contributions for GBT models trained using a) *SolUD* + *Chem* b) *SolUD* c) *Chem* and d) *SolV* feature sets

When either the *SolV* and *SolUD* concentration features are included, they take significant precedence over the other features (Figure 8a, b, d). The *Chem* feature set lacks an explicit concentration feature, but does contain precursor dispense volumes (Fig 8c) from which the variation in the actual concentration of a species can be accurately inferred. As an example, we observe positive values of Pearson correlation coefficient (PCC) between SolUD-derived inorganic concentration and the inorganic dispense volume (PCC = 0.63) as well as between the SolUD-derived organic concentration and the organic dispense volume (PCC = 0.50) (Figure 9). When no concentration model is present, ML models learn the most from features that are related to computing concentration; out of the 103 features in the *Chem* feature set, all ten features necessary to fully calculate experimental concentration were present in the top twenty features. The comparable performance of the *Chem* feature set with the *SolUD*, along with the feature contribution analysis strongly suggests that the model is 'learning' an equivalent representation of concentration from the raw experimental data.

Models fit on the *SolUD* + *Chem* feature set use both concentration and dispense volume information (Fig 8a), but removing dispense volumes does not reduce model performance. For example, the GBT algorithm on the *SolUD* + *Chem* dataset after removing the dispense volumes features has no impact on model performance (MCC = 0.67 ± 0.04). From this we conclude that dispense volumes do not provide useful additional information, but are likely merely selected because of high covariance with the concentration (Figure 9).[51]



	Acid Conc. (SolUD)	Acid Conc. (SolV)	Acid Dispense 1	Acid Dispense 2	Inorg. Conc. (SolUD)	Inorg. Conc. (SolV)	Inorganic Dispense Vol.	Org. Conc. (SolUD)	Org. Conc. (SolV)	Organic Dispense Vol.
Organic Dispense Vol.	-0.07	-0.07	-0.01	0.04	-0.34	-0.3	-0.36	0.5	0.49	1.0
Org. Conc. (SolV)	-0.19	-0.19	-0.14	-0.03	0.06	0.12	0.04	0.97	1.0	0.49
Org. Conc. (SolUD)	-0.23	-0.23	-0.16	-0.07	0.1	0.15	0.09	1.0	0.97	0.5
Inorganic Dispense Vol.	-0.09	-0.09	0.09	-0.2	0.63	0.58	1.0	0.09	0.04	-0.36
Inorg. Conc. (SolV)	-0.27	-0.27	-0.12	-0.24	0.98	1.0	0.58	0.15	0.12	-0.3
Inorg. Conc. (SolUD)	-0.29	-0.29	-0.12	-0.28	1.0	0.98	0.63	0.1	0.06	-0.34
Acid Dispense 2	0.62	0.62	-0.13	1.0	-0.28	-0.24	-0.2	-0.07	-0.03	0.04
Acid Dispense 1	0.54	0.54	1.0	-0.13	-0.12	-0.12	0.09	-0.16	-0.14	-0.01
Acid Conc. (SolV)	1.0	1.0	0.54	0.62	-0.29	-0.27	-0.09	-0.23	-0.19	-0.07
Acid Conc. (SolUD)	1.0	1.0	0.54	0.62	-0.29	-0.27	-0.09	-0.23	-0.19	-0.07

Figure 9. Covariance analysis comparing dispense volumes, SolV concentrations and SolUD concentrations. Numerical values are equal to the Pearson correlation of the intersecting features.

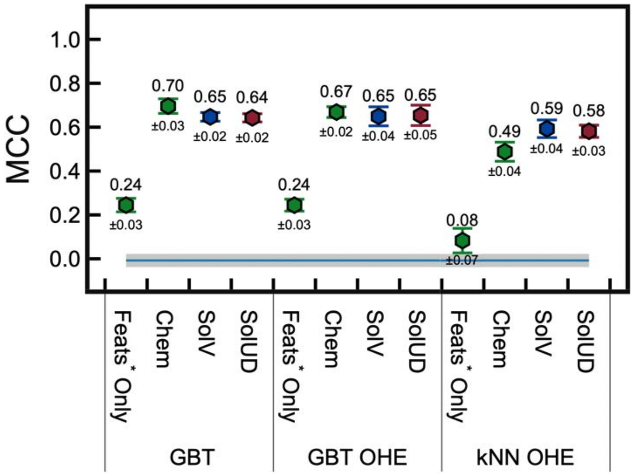


Figure 10. Comparison of model performance for kNN and GBT models fit to one-hot-encoded (OHE) dataset, compared to performance of GBT model fit on data with no preprocessing

To understand if the numerical values of the *Feat* feature set were at all important to the models, we compared kNN and GBT models fit to a dataset without the physicochemical descriptors (*i.e.*, we removed the *Feat* subset) and instead used one-hot encoded (OHE) (Figure 10). One-hot encoding the amine identity reduces the number of input features from 76 (Feat+SolUD) to 25 (OHE+SolUD), without reducing the GBT model performance. This indicates that the underlying models require minimal information from the numerical value of the physicochemical features and instead are using the values as an identity; effectively, the model likely uses the feature to determine the reactant set and then interpolates within the concentration space for that reactant set. These data are further supported by the weighted feature analysis of the OHE GBT method which demonstrates similar feature weighting and performance (Figure S10).

Taken together, these analyses illustrate that ML is capable of 'learning' a way around bad physical models (i.e., *SolV*) and even assembling raw experimental characteristics (i.e., *Chem)* into highly competitive models. ML performance is shown to be independent of the accuracy of the chemical representation (i.e., *SolV* vs *SolUD* vs *Chem*) and dependent upon precise representations of experimental variation. The features governing the similar performance in GBT models were identified through step-wise analysis which showed that the model was 'choosing' data associated directly with the concentration calculation over other extraneous information in the dataset.

Toward Generalizable ML Models (LOO)

The model performance and underlying behavior of physicochemical features in STTS models outline the weakness of current physicochemical descriptors for generalizing to untested amines. While physicochemical features help guide models to concentration relationships for a set of chemicals (similar to OHE), these values seem to be numerically meaningless; attempting to build suitable models for LOO prediction demonstrates the inherent difficulty in predicting crystallization in untrained chemical systems (Figure 11).

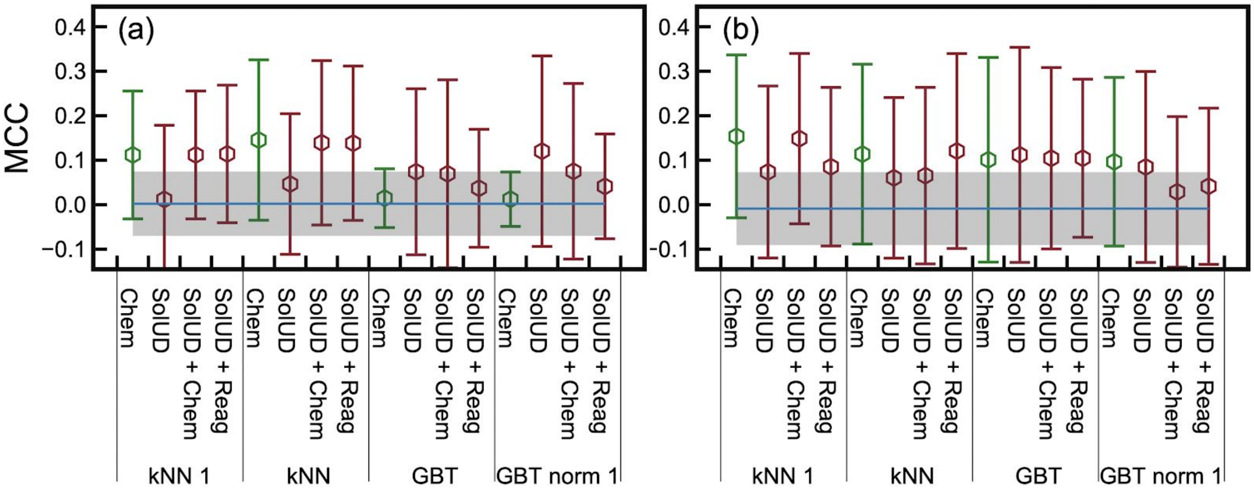


Figure 11. Best performing LOO models across all algorithms, preprocessing and feature sets (a) all available data for each process space (b) stratified random draw from each process space

The highest average MCC was reported for a kNN = 1 model, although the standard deviation of model performance was large (kNN=1, *Chem*, MCC = 0.15 ± 0.18). The success of kNN=1 indicates that memorization of similar reactions is a better strategy than any other attempted algorithm and preprocessing scheme. The small performance improvement when moving from only the description of masses and volumes (*Chem*) and the calculation using the density-corrected concentration (*SolUD)* indicates that the models account for some but not all of the effect. Using all of the training examples (Figure 11a) tends to perform worse than stratifying the training set examples (Figure 11b). A closer examination of each of the resulting models for predicting each of left-out amines revealed that 33% of models built using all training examples predicted *every* test reaction to fail, whereas only 19% of models built using stratified examples predicted 'all fail'. There is no discernable trend in the relationship between amines which correlated with a model making 'all fail' or 'all success' predictions, but kNN models are less susceptible to this pathology. Models which included the SolUD features (independent of algorithm) were also less likely to suffer from the 'all fail' condition and also had the lowest average MCC improvement by excluding 'all fail' (see extended discussion in SI). Furthermore, the best models identified in the STTS work (GBT, *SolUD* + *Chem*) do not suffer from this problem.

Constant predictions yield an MCC score of zero, whereas other common metrics such as precision and recall can be deceptively higher. On the other hand, precision (true positives per predicted positives) can be a valuable metric for experimental choice quality. To better understand this, we focus on the best model algorithms and feature sets identified above. Figure 12 compares the MCC and precision for the stratified LOO problem for (1) the collection of all models (including constant predictions), (2) only models that make non-constant predictions, (3) the hypothetical cases of predicting "all failure", and (4) "all success" for every reaction (with 3-4 indicated as horizontal lines). An "all fail" prediction gives undefined MCC and precision scores, indicated by the solid red line at zero in both plots. (Figure S11 shows the corresponding unstratified training regime results.)

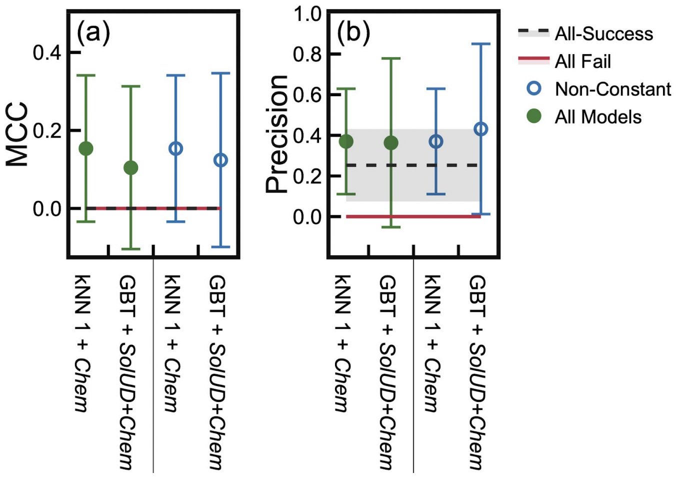


Figure 12. Comparison (a) MCC and (b) precision scores for of algorithm + feature set LOO performance. Error bars indicate standard deviation across all amines. The grey band in (b) indicates the variation in precision for the "all success" model for different test amines.

Excluding constant prediction cases (open blue circle) increases both performance metrics. Excluding 'all fail' examples increased MCC by an average of 0.015 across all model types and feature; for the models using the SolUD features and average MCC improvement of 0.010 was observed. The smaller improvement reflects that previous observation that models using SolUD are less susceptible to the 'all-fail' pathology. Whether or not constant predictions are excluded, the precision and MCC for models are better than the hypothetical "all success" predictions (dashed black line) which models the case of picking experiments at random. A positive correlation was observed between percentage of total successful experiments and precision of all models; a higher proportion of successful experiments is an easier target for models (Table S7).

The STTS and LOO tasks provide a useful benchmark for ML performance using the described perovskite dataset. Although the performance on the LOO task is worse than on the STTS task, our results demonstrate that the improved SolUD concentration model, in combination with the additional experimental information in the perovskite dataset, can predict perovskite crystallization in untrained chemical system better than chance.

Conclusion

Improved concentration representations are applicable both retroactively and proactively to an ongoing high-throughput perovskite crystallization campaign. Volumes calculated using the '**Sol**ution **U**sing **D**ensity' (SolUD) method closely agree with experimental observations (R^2 = 0.975 across 172 precursor solutions). The SolUD model also provided a useful auditing tool which identified possible anomalies covering 8% (768 experiments) of the dataset, of which 37.5% (288 experiments) could be corrected from other laboratory records. We applied machine learning (ML) to the updated perovskite dataset and demonstrated three important outcomes: (1) upfront costs associated with intensive feature engineering can be mitigated by careful experimental design (i.e., constraints on the scale and variability of experimental campaigns), (2) ML can accommodate less sophisticated physical models affording that precise representations of experimental variation are available, and most interestingly (3) ML can 'learn' a proxy for concentration using raw experimental descriptions.

The development of an accurate concentration representation and subsequent ML comparison have provided crucial insight toward development of a generalizable model for prediction of perovskite crystallization. ML models are capable of describing compositional variations for a known set of reactants, and the best gradient boosted tree (GBT) model demonstrated a Matthews Correlation Coefficient (MCC) of 0.71 ± 0.01 using 109 total features. However, most of this performance can be captured by one-hot-encoding chemical identities in combination with SolUD concentration features (MCC = 0.65 ± 0.05). The similar performance of the models implies that molecular property descriptors are used primarily as a means of identifying a particular reactant set, rather than learning generalizable trends. This is further supported by the 'leave one reactant set out' testing, which was less successful (MCC = 0.15 ± 0.18) at predicting reactions involving novel reactant sets. However, even this limited performance is above the baseline MCC = 0 score, indicating the value of chemical features to make extrapolative predictions about reaction outcome. The wide standard deviations for LOO models were attributed to variation in amine model performance; the LOO models which were identified as consistently performant across amines—those which specifically avoiding the pitfall of making 'all fail' predictions—were shown to predominately include the newly developed SolUD features. Although models can use experimental data (masses and volumes) to make interpolative predictions, physically meaningful features (density-corrected concentrations) improves robustness when performing extrapolations.

In summary, a stepwise comparative approach to machine learning can provide insight into *what* and *how much* models are 'learning' for a given prediction task.[50] We aim to use these findings to improve the perovskite dataset with the goal of significantly expanding the ability to predict crystal formation in untested reactant sets.

Abbreviations

GBL: gamma-butyrolactone or γ-butyrolactone LOO: leave one reactant-set out, MAE: mean absolute error, ML: machine learning, MSE: mean squared error, OHE: one-hot encoding, Reag: reagents, RMSE: root mean squared error, STTS: standard train-test splits, SolUD: solution using density, SolV: solvent only volume, MCC: Matthew's correlation coefficient

Supporting Information

The datasets and computational codes in Mathematica 12.0 and Python 3.7 performing the analyses described in this paper are available at [ACS info], and on Material Data Facility[52,53] at https://www.doi.org/10.18126/lyk3-qace.

Acknowledgements

We thank Victor Ghosh and Phillip Nega for help with data collection. We thank Gary Cattabriga for help reviewing this article. 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. Experimental data collection from 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. JS acknowledges the Henry Dreyfus Teacher-Scholar Award (TH-14-010).

References

(1) Best Research-Cell Efficiency Chart. *https://www.nrel.gov/pv/cell-efficiency.html*, 2019.

(2) Herz, L. M. Charge-Carrier Mobilities in Metal Halide Perovskites: Fundamental Mechanisms and Limits. *ACS Energy Lett.* **2017**, *2* (7), 1539–1548. https://doi.org/10.1021/acsenergylett.7b00276.

(3) Tai, Q.; Tang, K.-C.; Yan, F. Recent Progress of Inorganic Perovskite Solar Cells. *Energy Environ. Sci.* **2019**, *12* (8), 2375–2405. https://doi.org/10.1039/C9EE01479A.

(4) Huang, J.; Yuan, Y.; Shao, Y.; Yan, Y. Understanding the Physical Properties of Hybrid Perovskites for Photovoltaic Applications. *Nat. Rev. Mater.* **2017**, *2* (7), 17042. https://doi.org/10.1038/natrevmats.2017.42.

(5) Jena, A. K.; Kulkarni, A.; Miyasaka, T. Halide Perovskite Photovoltaics: Background, Status, and Future Prospects. *Chem. Rev.* **2019**, *119* (5), 3036–3103. https://doi.org/10.1021/acs.chemrev.8b00539.

(6) Saparov, B.; Mitzi, D. B. Organic–Inorganic Perovskites: Structural Versatility for Functional Materials Design. *Chem. Rev.* **2016**, *116* (7), 4558–4596. https://doi.org/10.1021/acs.chemrev.5b00715.

(7) Smith, M. D.; Crace, E. J.; Jaffe, A.; Karunadasa, H. I. The Diversity of Layered Halide Perovskites. *Annu. Rev. Mater. Res.* **2018**, *48* (1), 111–136. https://doi.org/10.1146/annurev-matsci-070317-124406.

(8) Stranks, S. D.; Snaith, H. J. Metal-Halide Perovskites for Photovoltaic and Light-Emitting Devices. *Nat. Nanotechnol.* **2015**, *10* (5), 391–402. https://doi.org/10.1038/nnano.2015.90.

(9) Lozano, G. The Role of Metal Halide Perovskites in Next-Generation Lighting Devices. *J. Phys. Chem. Lett.* **2018**, *9* (14), 3987–3997. https://doi.org/10.1021/acs.jpclett.8b01417.

(10) Smith, M. D.; Karunadasa, H. I. White-Light Emission from Layered Halide Perovskites. *Acc. Chem. Res.* **2018**, *51* (3), 619–627. https://doi.org/10.1021/acs.accounts.7b00433.

(11) Yi, Z.; Ladi, N. H.; Shai, X.; Li, H.; Shen, Y.; Wang, M. Will Organic–Inorganic Hybrid Halide Lead Perovskites Be Eliminated from Optoelectronic Applications? *Nanoscale Adv.* **2019**, *1* (4), 1276–1289. https://doi.org/10.1039/C8NA00416A.

(12) Yao, F.; Gui, P.; Zhang, Q.; Lin, Q. Molecular Engineering of Perovskite Photodetectors: Recent Advances in Materials and Devices. *Mol. Syst. Des. Eng.* **2018**, *3* (5), 702–716. https://doi.org/10.1039/C8ME00022K.

(13) Ahmad, S.; George, C.; Beesley, D. J.; Baumberg, J. J.; De Volder, M. Photo-Rechargeable Organo-Halide Perovskite Batteries. *Nano Lett.* **2018**, *18* (3), 1856–1862. https://doi.org/10.1021/acs.nanolett.7b05153.

(14) Li, J.; Lu, Y.; Xu, Y.; Liu, C.; Tu, Y.; Ye, S.; Liu, H.; Xie, Y.; Qian, H.; Zhu, X. AIR-Chem: Authentic Intelligent Robotics for Chemistry. *J. Phys. Chem. A* **2018**, *122* (46), 9142–9148. https://doi.org/10.1021/acs.jpca.8b10680.

(15) Sun, S.; Hartono, N. T. P.; Ren, Z. D.; Oviedo, F.; Buscemi, A. M.; Layurova, M.; Chen, D. X.; Ogunfunmi, T.; Thapa, J.; Ramasamy, S.; Settens, C.; DeCost, B. L.; Kusne, A. G.; Liu, Z.; Tian, S. I. P.; Peters, I. M.; Correa-Baena, J.-P.; Buonassisi, T. Accelerated Development of Perovskite-Inspired Materials via High-Throughput Synthesis and

Machine-Learning Diagnosis. *Joule* **2019**, *3* (6), 1437–1451. https://doi.org/10.1016/j.joule.2019.05.014.

(16) Chen, S.; Zhang, L.; Yan, L.; Xiang, X.; Zhao, X.; Yang, S.; Xu, B. Accelerating the Screening of Perovskite Compositions for Photovoltaic Applications through High-Throughput Inkjet Printing. *Adv. Funct. Mater.* **2019**, 1905487. https://doi.org/10.1002/adfm.201905487.

(17) MacLeod, B. P.; Parlane, F. G. L.; Morrissey, T. D.; Häse, F.; Roch, L. M.; Dettelbach, K. E.; Moreira, R.; Yunker, L. P. E.; Rooney, M. B.; Deeth, J. R.; Lai, V.; Ng, G. J.; Situ, H.; Zhang, R. H.; Aspuru-Guzik, A.; Hein, J. E.; Berlinguette, C. P. Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials. *ArXiv190605398 Cond-Mat Physicsphysics* **2019**.

(18) Chen, S.; Hou, Y.; Chen, H.; Tang, X.; Langner, S.; Li, N.; Stubhan, T.; Levchuk, I.; Gu, E.; Osvet, A.; Brabec, C. J. Exploring the Stability of Novel Wide Bandgap Perovskites by a Robot Based High Throughput Approach. *Adv. Energy Mater.* **2018**, *8* (6), 1701543. https://doi.org/10.1002/aenm.201701543.

(19) Li, Z.; Najeeb, M. A.; Alves, L.; Sherman, A.; Parrilla, P. C.; Pendleton, I. M.; Zeller, M.; Schrier, J.; Norquist, A. J.; Chan, E. *Robot-Accelerated Perovskite Investigation and Discovery (RAPID): 1. Inverse Temperature Crystallization*; preprint; 2019. https://doi.org/10.26434/chemrxiv.10013090.v1.

(20) Dong, Q.; Fang, Y.; Shao, Y.; Mulligan, P.; Qiu, J.; Cao, L.; Huang, J. Electron-Hole Diffusion Lengths > 175 Mm in Solution-Grown $CH_3NH_3PbI_3$ Single Crystals. *Science* **2015**, *347* (6225), 967–970. https://doi.org/10.1126/science.aaa5760.

(21) Giorgi, G.; Fujisawa, J.-I.; Segawa, H.; Yamashita, K. Small Photocarrier Effective Masses Featuring Ambipolar Transport in Methylammonium Lead Iodide Perovskite: A Density Functional Analysis. *J. Phys. Chem. Lett.* **2013**, *4* (24), 4213–4216. https://doi.org/10.1021/jz4023865.

(22) Brenner, T. M.; Egger, D. A.; Kronik, L.; Hodes, G.; Cahen, D. Hybrid Organic—Inorganic Perovskites: Low-Cost Semiconductors with Intriguing

Charge-Transport Properties. *Nat. Rev. Mater.* **2016**, *1* (1), 15007. https://doi.org/10.1038/natrevmats.2015.7.

(23) Wang, X.; Li, W.; Liao, J.; Kuang, D. Recent Advances in Halide Perovskite Single-Crystal Thin Films: Fabrication Methods and Optoelectronic Applications. *Sol. RRL* **2019**, *3* (4), 1800294. https://doi.org/10.1002/solr.201800294.

(24) Chen, Y.; Orgiu, E. Charge Transport in Halide Perovskite Single Crystals: Experimental and Theoretical Perspectives. *ChemNanoMat* **2019**, *5* (3), 290–299. https://doi.org/10.1002/cnma.201800679.

(25) Fateev, S. A.; Petrov, A. A.; Khrustalev, V. N.; Dorovatovskii, P. V.; Zubavichus, Y. V.; Goodilin, E. A.; Tarasov, A. B. Solution Processing of Methylammonium Lead Iodide Perovskite from γ-Butyrolactone: Crystallization Mediated by Solvation Equilibrium. *Chem. Mater.* **2018**, *30* (15), 5237–5244. https://doi.org/10.1021/acs.chemmater.8b01906.

(26) Jung, M.; Ji, S.-G.; Kim, G.; Seok, S. I. Perovskite Precursor Solution Chemistry: From Fundamentals to Photovoltaic Applications. *Chem. Soc. Rev.* **2019**, *48* (7), 2011–2038. https://doi.org/10.1039/C8CS00656C.

(27) Yan, K.; Long, M.; Zhang, T.; Wei, Z.; Chen, H.; Yang, S.; Xu, J. Hybrid Halide Perovskite Solar Cell Precursors: Colloidal Chemistry and Coordination Engineering behind Device Processing for High Efficiency. *J. Am. Chem. Soc.* **2015**, *137* (13), 4460–4468. https://doi.org/10.1021/jacs.5b00321.

(28) Cao, J.; Jing, X.; Yan, J.; Hu, C.; Chen, R.; Yin, J.; Li, J.; Zheng, N. Identifying the Molecular Structures of Intermediates for Optimizing the Fabrication of High-Quality Perovskite Films. *J. Am. Chem. Soc.* **2016**, *138* (31), 9919–9926. https://doi.org/10.1021/jacs.6b04924.

(29) Moore, D. T.; Sai, H.; Tan, K. W.; Smilgies, D.-M.; Zhang, W.; Snaith, H. J.; Wiesner, U.; Estroff, L. A. Crystallization Kinetics of Organic–Inorganic Trihalide Perovskites and the Role of the Lead Anion in Crystal Growth. *J. Am. Chem. Soc.* **2015**, *137* (6), 2350–2358. https://doi.org/10.1021/ja512117e.

(30) Li, B.; Isikgor, F. H.; Coskun, H.; Ouyang, J. The Effect of Methylammonium Iodide on the Supersaturation and Interfacial Energy of the Crystallization of Methylammonium

Lead Triiodide Single Crystals. *Angew. Chem. Int. Ed.* **2017**, *56* (50), 16073–16076. https://doi.org/10.1002/anie.201710234.

(31) Wicker, J. G. P.; Cooper, R. I. Will It Crystallise? Predicting Crystallinity of Molecular Materials. *CrystEngComm* **2015**, *17* (9), 1927–1934. https://doi.org/10.1039/C4CE01912A.

(32) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chem. Mater.* **2016**, *28* (20), 7324–7331. https://doi.org/10.1021/acs.chemmater.6b02724.

(33) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533* (7601), 73–76. https://doi.org/10.1038/nature17439.

(34) Duros, V.; Grizou, J.; Xuan, W.; Hosni, Z.; Long, D.-L.; Miras, H. N.; Cronin, L. Human versus Robots in the Discovery and Crystallization of Gigantic Polyoxometalates. *Angew. Chem. Int. Ed.* **2017**, *56* (36), 10815–10820. https://doi.org/10.1002/anie.201705721.

(35) Duros, V.; Grizou, J.; Sharma, A.; Mehr, S. H. M.; Bubliauskas, A.; Frei, P.; Miras, H. N.; Cronin, L. Intuition-Enabled Machine Learning Beats the Competition When Joint Human-Robot Teams Perform Inorganic Chemical Experiments. *J. Chem. Inf. Model.* **2019**, *59* (6), 2664–2671. https://doi.org/10.1021/acs.jcim.9b00304.

(36) 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* (7773), 251–255. https://doi.org/10.1038/s41586-019-1540-5.

(37) *ChemAxon's Calculator Plugin Cxcalc Was Used for Calculation of Molecular Features*; CxCalc 19.9.0, ChemAxon (https://www.chemaxon.com).

(38) Ionic Radii for Ions with Various Coordination Numbers. *CRC Handbook of Chemistry and Physics*; Rumble, J. R., Ed.; CRC Press/Taylor & Francis, Boca Raton, FL; Vol. 100th Edition (Internet Version 2019).

(39) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2016**, *72* (2), 171–179. https://doi.org/10.1107/S2052520616003954.

(40) Lunelli, B.; Scagnolari, F. Rapid Microdetermination of Partial Molar and Excess Molar Volumes. *J. Chem. Educ.* **2002**, *79* (5), 626. https://doi.org/10.1021/ed079p626.

(41) 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* (3), 846–859. https://doi.org/10.1557/mrc.2019.72.

(42) *https://github.com/darkreactions/ESCALATE_report*; Dark Reactions Project, 2019.

(43) Pendleton, I. M.; Caucci, M. K.; Tynes, M.; Dharna, A.; Najeeb, M. A.; Chan, E. M.; Norquist, A. J.; Schrier, J. *Untangling How Machines "Learn" Perovskite Crystallization Chemistry Through Stepwise Data Sample Comparisons*; Materials Data Facility, 2020. https://doi.org/10.18126/LYK3-QACE.

(44) 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, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(45) Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G. Response to Comment on "Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning." *Science* **2018**, *362* (6416), eaat8763. https://doi.org/10.1126/science.aat8763.

(46) Tropsha, A.; Gramatica, P.; Gombar, V. The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22* (1), 69–77. https://doi.org/10.1002/qsar.200390007.

(47) Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* **2018**, *58* (5), 916–932. https://doi.org/10.1021/acs.jcim.7b00403.

(48) Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics* **2020**, *21* (1), 6. https://doi.org/10.1186/s12864-019-6413-7.

(49) Nisbet, M. L.; Pendleton, I. M.; Nolis, G. M.; Griffith, K. J.; Schrier, J.; Cabana, J.; Norquist, A. J.; Poeppelmeier, K. R. Machine-Learning-Assisted Synthesis of Polar Racemates. *J. Am. Chem. Soc.* **2020**, *142* (16), 7555–7566. https://doi.org/10.1021/jacs.0c01239.

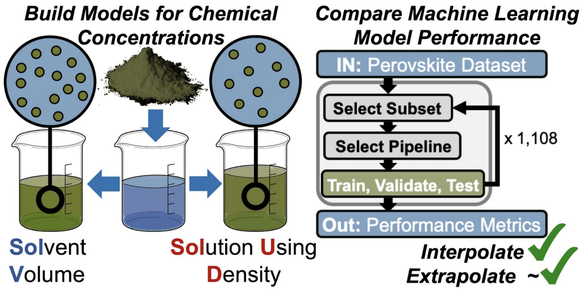
(50) Walker, E.; Kammeraad, J.; Goetz, J.; Robo, M. T.; Tewari, A.; Zimmerman, P. M. Learning To Predict Reaction Conditions: Relationships between Solvent, Molecular Structure, and Catalyst. *J. Chem. Inf. Model.* **2019**, *59* (9), 3645–3654. https://doi.org/10.1021/acs.jcim.9b00313.

(51) A covariance analysis of the physicochemical features illustrates that many are highly covariant, though the average surface area is among the least covariant with the others in the dataset

(52) Blaiszik, B.; Ward, L.; Schwarting, M.; Gaff, J.; Chard, R.; Pike, D.; Chard, K.; Foster, I. A Data Ecosystem to Support Machine Learning in Materials Science. *MRS Commun.* **2019**, *9* (4), 1125–1133. https://doi.org/10.1557/mrc.2019.118.

(53) Blaiszik, B.; Chard, K.; Pruyne, J.; Ananthakrishnan, R.; Tuecke, S.; Foster, I. The Materials Data Facility: Data Services to Advance Materials Science Research. *JOM* **2016**, *68* (8), 2045–2052. https://doi.org/10.1007/s11837-016-2001-3.

TOC FIGURE


Build Models for Chemical Concentrations
Solvent Volume
Solution Using Density
Compare Machine Learning Model Performance
IN: Perovskite Dataset
Select Subset
Select Pipeline
Train, Validate, Test
x 1,108
Out: Performance Metrics
Interpolate ✓
Extrapolate ~✓