JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

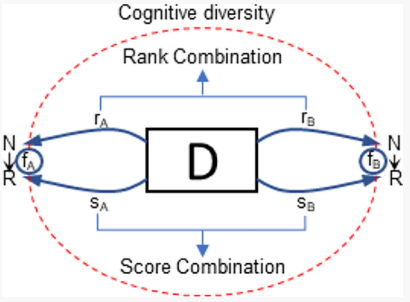
Improving Data and Prediction Quality of High-Throughput Perovskite Synthesis with Model Fusion

Yuanqing Tang, Zhi Li, Mansoor Ani Najeeb Nellikkal, Hamed Eramian, Emory M. Chan, Alexander J. Norquist, D. Frank Hsu, and Joshua Schrier*

Cite This: https://doi.org/10.1021/acs.jcim.0c01307 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Combinatorial fusion analysis (CFA) is an approach for combining multiple scoring systems using the rank-score characteristic function and cognitive diversity measure. One example is to combine diverse machine learning models to achieve better prediction quality. In this work, we apply CFA to the synthesis of metal halide perovskites containing organic ammonium cations via inverse temperature crystallization. Using a data set generated by high-throughput experimentation, four individual models (support vector machines, random forests, weighted logistic classifier, and gradient boosted trees) were developed. We characterize each of these scoring systems and explore 66 possible combinations of the models. When measured by the precision on predicting crystal formation, the majority of the combination models improves the individual model results. The best combination models outperform the best individual models by 3.9 percentage points in precision. In addition to improving prediction quality, we demonstrate how the fusion models can be used to identify mislabeled input data and address issues of data quality. In particular, we identify example cases where all single models and all fusion models do not give the correct prediction. Experimental replication of these syntheses reveals that these compositions are sensitive to modest temperature variations across the different locations of the heating element that can hinder or enhance the crystallization process. In summary, we demonstrate that model fusion using CFA can not only identify a previously unconsidered influence on reaction outcome but also be used as a form of quality control for high-throughput experimentation.



1. INTRODUCTION

High-throughput experimentation (HTE) has been used to accelerate synthesis and characterization[1] for many areas of chemistry[2] and materials science.[3] In addition to merely increasing the rate at which experiments are performed, HTE provides an opportunity to generate larger data sets for use with machine learning and artificial intelligence (ML/AI) methods and to take actions to test model predictions in the laboratory.[4] Furthermore, laboratory automation facilitates the capture of a complete record of experimental successes and failures[5] and enables more systematic sampling of experimental variables that avoids human biases,[6] both of which improve the quality of ML models for chemical reaction prediction.

As one specific example, we consider metal halide perovskites,[7] an emerging class of materials for photovoltaics[8] and optoelectronics.[9] High-throughput approaches have been used to explore perovskite thin-films,[10,11] polycrystalline samples,[12,13] nanocrystals,[14,15] and single crystals (by vapor diffusion[16] and inverse temperature crystallization[17]). Our work has focused on developing high-throughput systems for perovskite single crystal growth (RAPID),[17] utilizing inverse-temperature crystallization (ITC).[18] The ESCALATE[19] software used by our system enables comprehensive data capture and reporting of these experiments. The ESCALATion web dashboard (http://escalation.sd2e.org) automatically trains and evaluates a suite of ML models based on new experimental data, displaying the experimental results and model interpretability insights, as well as tracking and versioning of the data sets and models over time. This provides us with a unique experimental data set that has allowed us to uncover physicochemical features responsible for crystal formation for a diverse set of molecular building units.[20]

An open challenge for scientific HTE applications is the need for quality control. Unlike the traditional quality engineering goal of manufacturing products with known specifications,[21] most scientific experiments do not have a known "right answer". Often the most interesting scientific results involve serendipity,[22] and the most desirable materials are the "extraordinary" ones having extreme properties that exceed previously known examples or compositions that exist outside of the established search domains.[23] The challenge is to distinguish scientifically interesting outliers from experimental anomalies, which can arise from many sources

Received: November 11, 2020

including the following: (i) uncontrolled (but measured) variations in experiment performance that affect entire batches of experiments (e.g., laboratory humidity and temperature variations); (ii) systematic variations that occur across experimental batches (e.g., temperature distribution on a heating element); (iii) uncontrolled and unmeasured variations in experiment performance (e.g., water contents in the reaction solutions and distribution of dispense volumes); (iv) operator errors and variations (e.g., experimental outcome assignments); and (v) the inherently stochastic nature of the process being studied (e.g., crystal growth). The scale of data generation that HTE enables precludes direct human oversight of every individual experiment, so an alternative approach is to use statistical approaches or machine learning models to identify outlier points for re-examination and verification. This presents the classic "chicken and the egg" problem.[24] The model depends upon the training set it is trying to audit, and its ability to detect anomalies may be hindered by the anomalies themselves. Errors that influence large groups of experiments, such as types (i) and (ii) described above, can fool the classifier. Furthermore, legitimate fluctuations like type (v) should be included but might be mistaken for types (iii) and (iv).

Ensemble methods and data fusion have been used in machine learning and AI (ML/AI) strategies and models.[25−30] These include the following: bagging and boosting,[27,29] random forests,[31] conditional mixture models,[32] ensemble models,[29] combining pattern classifiers,[26,33] combining artificial neural nets,[28] data fusion in information retrieval,[34] causal inference and data fusion,[35] and combinatorial fusion analysis.[36] Since each of the individual systems (or models) has strengths and weaknesses across various domains, ensemble methods (or model fusion) have been demonstrated to be successful when individual systems are diverse. However, the notion of "diversity" varies when using various ensemble methods in different domains. These include correlation (or rank correlation) about data distribution in statistics and different diversity measurements in ML/AI. Most of these diversity measurements are related to data items in the data set. Combinatorial fusion analysis (CFA) was proposed to combine multiple scoring systems using the rank-score characteristic (RSC) function and cognitive diversity (CD).[36−39] (RSC and CD are defined in the Methods section.) Instead of defining a performance criterion, the RSC function is used to characterize a scoring system. The CD between two scoring systems A and B is then defined using RSC functions of A and of B to measure the dissimilarity (or variations) between two systems.[37−39] CFA also addresses the issue of rank versus score combination (*vide infra*).[40] This is depicted in the graphical table of contents image, adapted from Figure 5 of ref 38. The CFA framework has been applied to a variety of domains including target tracking and computer vision,[41] ChIP-seq peak detection,[42] information retrieval,[40] brain science,[43] wireless network communication,[44] virtual screening,[45] and reinforcement learning.[46]

Herein, we demonstrate the use of CFA to improve the data quality and prediction quality of HTE materials synthesis experiments. First, we quantify the cognitive diversity of different model types. We demonstrate that combination models constructed by model fusion improve prediction quality metrics. Using the more robust predictions from model fusion, we identify questionable experimental results—focusing on those in which every single model and every fusion model fail to predict the outcome. Using our metal halide perovskite data set, we analyze and identify possible factors associated with these failures. We describe new laboratory experiments for these points that tested both the reproducibility of the original experiments as well as the proposed causative factors. Using these new experimental results, we are able to distinguish the fundamental limits of the current models to predict reaction outcomes and the level of experimental variation to be expected.

2. METHODS

2.1. Model Fusion. Combinatorial fusion analysis (CFA) provides methods and algorithms for data fusion, consensus scoring, preference ranking, and ensemble machine learning.[36−42,45,46] CFA combines multiple scoring systems either at the attribute level (e.g., features, variables, parameters, or cues) or at the system level (e.g., models, modalities, software, or experts). In this paper, we use model fusion, a special case of combinatorial fusion, at the system level where each model is considered as a scoring system.

Let $D = \{d_1, d_2, ..., d_n\}$ be a set of n data items, with each data item, d_i, corresponding to an experiment. The scoring system A on the data set D consists of a score function, s_A, and a rank function, r_A. The score function s_A: $D \rightarrow \mathbb{R}$ assigns a score value (a real number in $\mathbb{R}$; for a binary classifier, this score is the probability of being in the "positive" class) to each data item d_i in D, i.e., $s_A(d_i)$ in $\mathbb{R}$ for each d_i in D. The rank function r_A: $D \rightarrow \mathbb{N}$, where $\mathbb{N} = \{1, 2, ..., n\}$, and n is the cardinality of D, is derived from s_A by sorting the score values into descending order and assigning the rank order of the score value to the data item which has that score value. For a scoring system A, its score function s_A and derived rank function r_A, the rank-score characteristic (RSC) function f_A: $\mathbb{N} \rightarrow \mathbb{R}$ is defined as

$$f_A(i) = s_A(r_A^{(-1)}(i)), \quad \text{for } i \text{ in } \mathbb{N} \tag{1}$$

The RSC function of the scoring system A, f_A, was defined by Hsu et al. in 2002[39] and subsequently used to define the notion of cognitive diversity in a variety of domain applications in ML/AL and data fusion.[38,46] The RSC function characterizes the scoring (or ranking) behavior of the underlying scoring system A. Moreover, since f_A is a function from ranks in $\mathbb{N}$ to scores in $\mathbb{R}$, it does not rely on specific data set items from sampling or experiments, so long as they are sampled from the same pool of potential data items. The RSC function values are normalized to the interval [0, 1]; this allows us to compare different scoring systems, as otherwise they would have their own score interval. In this paper, *cognitive diversity*, CD(A, B), which provides the diversity measurement between two models A and B, is calculated from the difference between RSC functions f_A and f_B[38,40,45]

$$\mathrm{CD(A, B)} = \sqrt{\sum_{i=1}^{n} (f_A(i) - f_B(i))^2} \tag{2}$$

where n = cardinality of D, f_A, f_B: $\mathbb{N} \rightarrow \mathbb{R}$.[45] CD differs from the traditional correlation or rank correlation in statistics (which measures association between two data distributions) or diversity measurement in machine learning and ensemble methods.[26,28,29,33] Correlation and rank correlation measures depend upon having a shared set of data set items. In contrast, CD can be applied even when the test data sets are different (such as different experiments or sampling strategies). The

https://doi.org/10.1021/acs.jcim.0c01307
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

diversity strength, ds(A), of the scoring system A is defined as the arithmetic average of CD between A and other scoring systems.

Methods of combination play an important role in the performance of the combined system. Traditional approaches use either score combination, such as in regression or Bayesian networks, or rank combination, such as in rank aggregation or consensus ranking.[47−50] Hsu and Taksa[40] compared rank and score combination methods for data fusion in information retrieval. They showed that under certain conditions, which include a relatively large cognitive diversity, rank combination can achieve better results than score combination. The CFA framework, which combines multiple scoring systems, allows researchers to take advantage of both worlds either by combining the score functions in the parametric Euclidean score space ($\mathbb{R}^n$) or the rank functions in the permutation rank space ($\mathbb{N}^n$), where D = the set of all data items, cardinality of D = n, with score function s_A: $D \rightarrow \mathbb{R}$ and rank function: r_A: D $\rightarrow$ $\mathbb{N}$. Let A_1, A_2, ..., A_t be a set of t scoring systems and w_1, w_2, ..., w_t be the weights assigned to each of the scoring systems. Each scoring system A_j has score and rank functions, s_{A_j} and r_{A_j}, respectively. The weighted score combination, SC(A_j, j = 1 to t), and weighted rank combination, RC(A_j, j = 1 to t), of the t scoring systems are defined as

$$s_{SC}(d_i) = \left(\sum_{j=1}^{t} w_j * s_{A_j}(d_i)\right) / \sum_{j=1}^{t} w_j \tag{3}$$

and

$$s_{RC}(d_i) = \left(\sum_{j=1}^{t} w_j * r_{A_j}(d_i)\right) / \sum_{j=1}^{t} w_j \tag{4}$$

for d_i in D. The rank functions of the combined scoring systems SC(A_j), RC(A_j), $r_{SC}(d_i)$, and $r_{RC}(d_i)$ can be obtained from $s_{SC}(d_i)$ and $s_{RC}(d_i)$, respectively. Although eqs 3 and 4 look explicitly like linear combinations, they are implicitly in two different combinatorial solution spaces. When evaluating the performance difference between the combined system and each of its individual systems, eq 3 operates on the parametric Euclidean score space. However, eq 4 operates on the set of all permutations, S_n, the symmetric group of order n, with a metric properly defined.[40,46,47] We further note that other methods of combinations are possible. For example, nonlinear combination using majority voting and convex combination using the mixed group rank are used in combining multiple classifier systems.[51,52]

2.2. Performance Metrics. In the binary classification, each data item is classified as either positive or negative; in our experiments, this corresponds to the formation of, or failure to form, a high quality single crystal. Accuracy, precision, and recall are used to evaluate the binary classifiers that comprise each single machine learning model.

To generalize these measures to rank-based systems, it is helpful to recall that each scoring system assigns a score, and a numerical threshold is applied to distinguish positive and negative predictions. If t is the rank of the data item which has this threshold as its score value, then data items ranked from 1 to t are predicted positive, and data items with a rank greater than t are predicted negative. Accuracy, recall, and precision measures can be generalized to evaluate rank-based systems in the following way: If there are k actual positives in the test set, a perfect model should predict all true positives at the top k of the single rank-score characteristic function and predict all true negatives for every data item after these first k items. Therefore, one way to calculate the precision for a ranking system is to take the first k items as predicted positives and then determine the number of true positives contained in that set. The abbreviation "**Pre@k**" denotes the precision of the first k number of data items, where k is the number of actual positives in the test set.[38] Pre@k is commonly used to characterize the retrieval quality of rank models for information retrieval tasks, such as search engines.[40] For both score combination and rank combination models, the model predicts the top k data items as predicted positives and then calculates the Pre@k to evaluate performance.

2.3. Computational Implementation. The work described here used a data set of 9387 inverse temperature crystallization perovskite synthesis experiments divided among 45 organoammonium cation species, reflecting the state of the project on November 27, 2019, assigned the internal label "dataset#44 (DS#44)". A complete transcript of these data is available via the Materials Data Facility[53] and via an interactive browser.[54] This data set includes a set of 75 physicochemical features (e.g., concentrations, temperature, stir rate) and organic property descriptors (e.g., molecular weight, atoms number, functional groups), described in Tables S1 and S2 in the Supporting Information.

These data were used to train four binary classifier models, where 1 ("positive") is the production of a large, high quality single crystalline product, and 0 ("negative") is any other outcome (e.g., polycrystalline sample, precipitation of starting materials, no reaction) and a prediction probability in the range of [0,1]. A 80/20% random train-test split was performed and used for all models to facilitate comparison. Files containing the exact training and test data, as well as the model predictions, are found at https://github.com/tyq0330/Model_Fusiong. Using these data, four classifier models were constructed using an automated model Test Harness (https://github.com/SD2E/test-harness) system implemented in Python 3.6.8 using the scikit-learn 0.22.1[55] implementation of each classifier.

The specific models are (A) support vector radial basis classifier (SVM) (hyperparameters: regularization parameter C = 100000, rbf kernel, gamma = 0.1); (B) random forest classification (RF) (hyperparameters: 361 trees in forest, criterion = "entropy" for the information gain, min_samples_leaf = 13, balanced class weight); (C) weighted logistic classifier (WLC) (cost function: balanced class weight); and (D) gradient boosted tree (GBT) (hyperparameters: learning rate = 1, max_depth = 10, max_features = "auto", n_estimators = 100). The thresholds for RF, WLC, and GBT are set to 0.5, and the threshold for SVM is a variable; in DS#44, it is 0.413.

In this paper, our model fusion combines four single models in pairs (6), triples (4), and quadruple (1). Each of these 11 combined models is then considered using both score and rank combinations. The score combination (SC) combines score values of the score functions from each of the underlying single models (eq 3). Likewise, the rank combination (RC) combines the rank numbers of rank function from each of the underlying single models (eq 4). This results in a total of 22 possible combined models. Finally, the scores and ranks of the different models can be weighted according to three different weighting schemes: average combination (AC), weighted combination by

https://doi.org/10.1021/acs.jcim.0c01307
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

performance (WCP), and weighted combination by diversity strength (WCDS), where

$$w = \frac{\text{weight of model } j}{\text{sum of weights}} = \begin{cases} \frac{1}{n}, & \text{AC} \\ \frac{P_j}{\sum_1^n P_j}, & \text{WCP} \\ \frac{ds_j}{\sum_1^n ds_j}, & \text{WCDS} \end{cases}$$

for AC, WCP, and WCDS, respectively. We note that in the average combination, every model j of the n models is given the same weight $1/n$. In the weighted combination by performance and by diversity strength, we use the performance criterion precision at k, Pre@k, and diversity strength ds(A) as the weight of each individual model A, respectively. Since there are 11 different models in each of the score and rank combinations of three different weight combinations, we have a total of 66 different combined models using the CFA framework. Predictions of the four classifier systems for each of the test set items are provided as input to the CFA analysis which were performed using Python 3.6.8. This code is available at https://github.com/tyq0330/Model_Fusiong.

2.4. Experimental Method. The experimental procedures, material characterizations, and chemical discoveries for the high-throughput inverse temperature crystallization (ITC) synthesis of metal halide perovskite single crystals are described in our previous work.[17] In brief, an automated liquid handling robot pipettes four different types of stock solutions into glass vials on a 96-well microplate (see Figure S1).[17] These stock solutions consist of (a) lead(II) iodide and the selected organoammonium iodide in solvent, (b) just the selected organoammonium iodide in solvent, (c) the neat solvent (most commonly gamma-butyrolactone, GBL), and (d) neat formic acid. The liquid handling robot dispenses the reagent stock solutions and then vortexes and heats the microplates to mix the reagent solutions. After vortexing is complete, the resulting perovskite solutions are heated without vortexing for 2.5 h to allow for crystal growth. For the reactions performed in this study, the heating temperature was typically set to a nominal 105 °C setting, which corresponds to an actual average temperature of 95 °C as measured by IR thermometry. The historical data set, DS#44, was mostly performed at this setting but also contains reactions performed at other temperatures (e.g., 80 °C, 67 °C). After reaction completion, the resultant crystals are scored by visual inspection into four outcome classes: (1) clear solution without any solid; (2) fine powder; (3) small crystallites (average crystal dimension <0.1 mm); and (4) large (>0.1 mm) crystals suitable for structure determination by single crystal X-ray diffraction. Of these, outcome class "4" corresponds to "positive" in our binary classification machine learning task. In addition to visual inspection at the time of experiment, we also capture photographs of the reaction outcomes that are stored with the data. Our past work has indicated that visual inspection with this rubric was more accurate and reproducible across operators than computer vision approaches for this system. To validate model fusion predictions, we performed 92 additional reactions with specific microplate locations across seven chemical systems (ethylammonium iodide/PbI_2, n-butylammoniuon iodide/PbI_2, dimethylammonium iodide/PbI_2, and isobutylammonium iodide/PbI_2, imidazolium iodide/PbI_2 acetamidinium iodide/PbI_2, and guanidinium iodide/PbI_2) for this paper.

3. RESULTS AND DISCUSSION

3.1. Binary Classification Performance of Individual Models. Table 1 shows the mean and standard deviation of prediction metrics for each single model from data set #30 (DS#30) to data set #43 (DS#43). The variations reflect the evolving performance of the models as more training items are added. These variations include the performance as different random test sets are used for evaluation and the changing chemical species being studied over the course of 13 weeks of experimentation. To focus on a single set of these models, Table 2 shows prediction metrics for each individual model for DS#44 where the threshold and its rank t for the four models SVM(A), RF(B), WLC(C), and GBT(D) are 0.413 at t = 380, 0.5 at t = 527, 0.5 at t = 747, and 0.5 at t = 327, respectively. All models, with the exception of GBT, predict more positive outcomes than contained in the test set and have varying capabilities at identifying the true positives present. Although the WLC (C) classifier has the lowest accuracy and precision, its true positive and recall rates are comparable to some of the better models and thus can be useful in an exploratory project, where the cost of low specificity is small (a few extra experiments) but the cost of low sensitivity is high (a missed discovery).

Table 1. Prediction Quality Metrics (Mean and Standard Deviation) of Individual Models for Data Sets DS#30−DS#43

model	accuracy	precision	recall
SVM (A)	0.887 ± 0.006	0.736 ± 0.028	0.753 ± 0.020
RF (B)	0.844 ± 0.008	0.598 ± 0.024	0.877 ± 0.015
WLC (C)	0.689 ± 0.010	0.388 ± 0.033	0.715 ± 0.023
GBT (D)	0.885 ± 0.006	0.745 ± 0.029	0.716 ± 0.025

Table 2. Prediction Quality Metrics of Individual Models for Data Set DS#44[a]

model (threshold)	rank t	true positive	accuracy	precision	recall
SVM (A) (0.413)	380	253	0.877	0.666	0.709
RF (B) (0.5)	527	312	0.862	0.592	0.874
WLC (C) (0.5)	747	271	0.701	0.363	0.759
GBT (D) (0.5)	321	234	0.888	0.729	0.655

[a] t is the rank number of the threshold for each model.

3.2. Model Fusion Performance. To assess the quality of ranking-based fusion models, we first establish a shared baseline with the individual models. An appropriate measurement suitable for characterizing rank-based models is precision at k (Pre@k).[40] As there are 357 actual positives among the 1878 experiments in the test set, the precision at 357 (Pre@357) quantifies the extent to which the highest ranked items correspond to the "best" (i.e., positive outcome) reaction selection. Table 3 shows the number of true positives (#TP) and the precision at rank 357 (Pre@357) for each of the individual models trained and tested on DS#44. The best performance is achieved by the GBT model, followed closely by RF and SVM. The WLC model performs much lower than the other three models, consistent with the lowest performance

https://doi.org/10.1021/acs.jcim.0c01307
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. Number of True Positives, #TP, Found and Precision at Rank 357, Pre@357, Calculated for Each of the Four Individual Models

model	#TP@357	Pre@357
SVM (A)	243	0.681
RF (B)	249	0.697
WLC (C)	157	0.440
GBT (D)	251	0.703

with respect to accuracy and precision in the score-based metrics observed in both Tables 1 and 2. The rank-score characteristic (RSC) function, eq 1, is plotted for each of the individual models in Figure 1. The shape of each RSC function

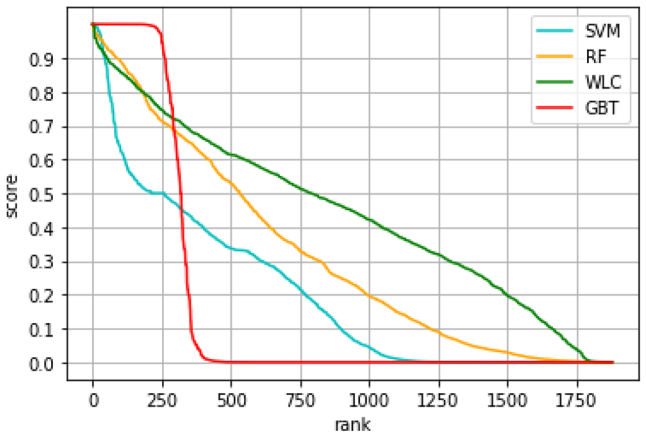


Figure 1. Rank-score characteristic (RSC) function graph (eq 1) for each of the four models: SVM (A), RF (B), WLC (C), and GBT(D) for 1878 test items in DS#44.

characterizes the scoring (or ranking) behavior of that model. A model whose RSC function graph is a hypothetical diagonal line from point (0, 1.0) to point (1878, 0) corresponds to a simple linear relationship between score and rank. In Figure 1, the RSC function graph of the WLC (C) model, which is closest to the diagonal line DL, assigns a score in [0, 1.0] to a rank in [1, 1878] in proportionally decreasing order. An RSC function above this hypothetical diagonal line, such as the first 300 data items predicted by the GBT model (red), corresponds to assigning higher scores than to the corresponding ranks. Steep changes in the RSC function indicate abrupt score assignment changes to subsequently ranked items. A model with the RSC function graph below the hypothetical diagonal lines, such as the SVM model (blue), gives relatively lower scores. The cognitive diversity (CD) between two models (eq 2) describing the area between their RSC functions is shown in Figure 2a. The *diversity strength* of a model, defined as the average of the cognitive diversities to the other three models, is shown in Figure 2b. The model WLC (C) has the largest diversity strength among these four models, followed by GBT (D). The model RF (B) and model SVM (A) have very similar values on diversity strength.

As noted in the Methods section, model fusion considers the 11 combinations of four single models, two methods of combinations (i.e., score combination (SC) and rank combination (RC)), and three weighting schemes (average combination (AC), weighted combination by performance (WCP) using Pre@357, and weighted combination using diversity strength (WCDS)). In general, a good practice when

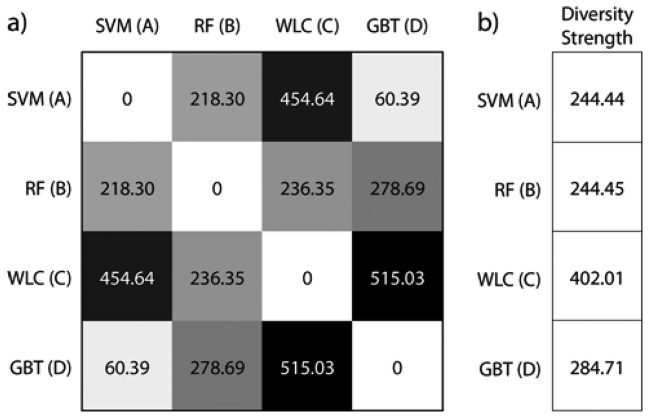


Figure 2. (a) Cognitive diversity (eq 2) between SVM, RF, WLC, and GBT models and (b) diversity strength of each of these individual models.

selecting single models to include in a CFA analysis is to include scoring systems that are relatively "good" (make predictions that are better than chance) and "different" (have a large diversity strength relative to the other models);[45] the four models described in the previous section satisfy these properties. In general, previous virtual screening studies applying CFA[45] and idealized numerical studies[59] have found that combinations of three or four different individual models (with sufficiently large diversity strength) suffice for most of the performance gains, after which there are only diminishing improvements.

Complete results for these combination schemes AC, WCP, and WCDS are included in Table S3 in the Supporting Information. Overall, the majority (39/66) of these new models (13, 15, and 11 cases for these combination methods AC, WCP, and WCDS, respectively) using DS#44 have Pre@357 better than or equal to the best of the individual models. Among the 66 combination models, the best one is RC(ABCD) under the WCP with Pre@357 = 0.742 which is 3.9% points higher than the best individual model GBT(D) with Pre@357 = 0.703 (Table 3 in section 2 and Table S3 in the Supporting Information). This is followed by RC(ABD) under AC with Pre@357 = 0.734, RC(ABD) under WCP with Pre@357 = 0.731, and RC(ABD) under WCDS with Pre@357 = 0.728. Performance of the 22 model fusion results using WCP is depicted in Figure 3. (Figures S2 and S3 show corresponding versions of this plot for the AC and WCDS weighting schemes.) In addition to showing the precision of each of the 22 combined models, the single model results are denoted by three horizontal lines at y = Pre@357 = 0.703(GBT(D)), 0.697(RF(B)), and 0.681(SVM(A)). Not shown is 0.440 (WLC(C)). Most of the high-performing rank combinations are better than the comparable score combination. Despite the greater diversity strength of model C compared to the other models, combinations involving model C tend to have lower prediction performance than other combined models. This is not surprising as model C has much lower performance compared to the other three models. In contrast, combinations involving B but not C perform better in most combined models' cases, as it has a relatively high performance and diversity strength. This explains why the best results under both AC and WCDS are achieved by rank combination of models A, B, and D, specifically RC(ABD). However, the performance of these models is lower than

https://doi.org/10.1021/acs.jcim.0c01307
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

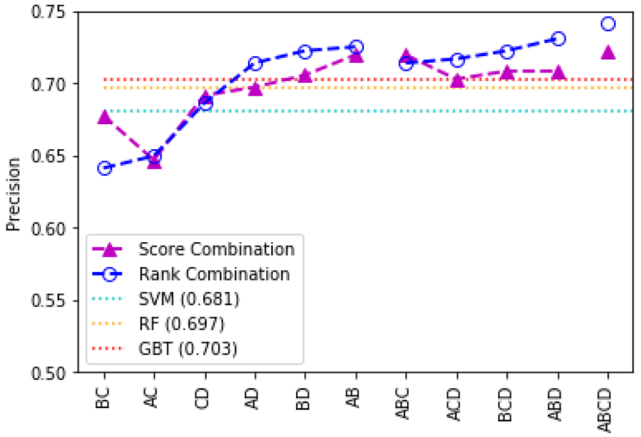


Figure 3. Precision (Pre@357) of each of the 22 combined models using weighted combination by performance (WCP) (points with "○" for rank combination and "▲" for score combination) compared with the single model (horizontal line) precisions (Pre@357) for the models SVM(A), RF(B), and GBT(D) is 0.681, 0.697, and 0.703, respectively. The line for WLC(C)'s Pre@357 = 0.44 is not shown.

RC(ABCD) under WCP discussed above. AC and WCDS are degraded by including (relatively poor performing) model C, because they either equally weight its predictions or overweight its predictions (because C has higher diversity strength), respectively. In contrast, WCP takes into account C's lower performance, while still allowing it to contribute diversity strength to the final prediction.

3.3. Extracting Insight from Model Fusion Results. In binary classification, each individual model has its own score as a threshold to classify positive or negatives: 0.413, 0.5, 0.5, and 0.5 for models A, B, C, and D, respectively. In model fusion using combinatorial fusion analysis (CFA), we define precision of an individual model to be Pre@k, where k is the number of positives in the data set. For example, model WLC(C) has threshold rank t = 747 with 271 true positives found (shown in Table 2) but has 157 true positives discovered with respect to Pre@357 (shown in Table 3). The large discrepancies between predicted 271 and 157 TPs are caused by the smaller threshold with higher rank t used by model WLC which produces not only higher true positives but also much higher false positives.

3.3.1. Model Fusion Finds Positives That Are Missed by Individual Models. Results by model fusion using CFA in Section 2 have correctly found nine true positive (TP) data items which would not have been found by any of the individual models. Table S8 and Table S9 show the rank of these data items in various WCDS fusion models for DS#39 and DS#44, respectively. Tables S4 and S5 with nine TPs and Tables S6 and S7 with five TPs show corresponding versions for model fusion for AC and WCP combinations, respectively.) For example, data item "j" in Table S9 was found to be true positive at rank 335 using Pre@357 by RC(CD) but was predicted (incorrectly) to be negative or to be positive but ranked low by each of the single models and ranked at 667, 476, 387, and 456 by each of the four individual models A (t = 380), B (t = 527), C (t = 747), and D (t = 321), respectively. In total, there are nine such TP data items: four in DS#39 {a,b,c,d} (Table S8) and five in DS#44 {e,f,g,h,j} (Table S9). In contrast, all the TP data items found by individual models A, B, C, and D were also correctly predicted by some model fusions in the CFA framework. This demonstrates that model fusion using CFA provides more predictive power than each of the four individual models. In addition to these WCDS results, similar analyses using AC and WCP are included in Tables S4−S5 and S6−S7, respectively. We also examined agreements between single models and each of the combination models AC, WCP, and WCDS. These include 26 false positives (FPs) (Table S10) and 17 false negatives (FNs) (Table S11). These false positive and false negative results are predicted incorrectly by the individual models *and* by the fusion models. This is a surprising anomaly which may suggest a possible problem with these individual experiments. We propose using this type of discrepancy as a criterion for prioritizing experiments for replication.

3.3.2. Experimental Replication by Reproduction and Relocation. In the original experimental data set, each individual experiment was randomly assigned a location on the 96-well plate to avoid any correlations between the vial location and its composition. As shown in the infrared thermal image of the heating block in Figure 4, there are temperature

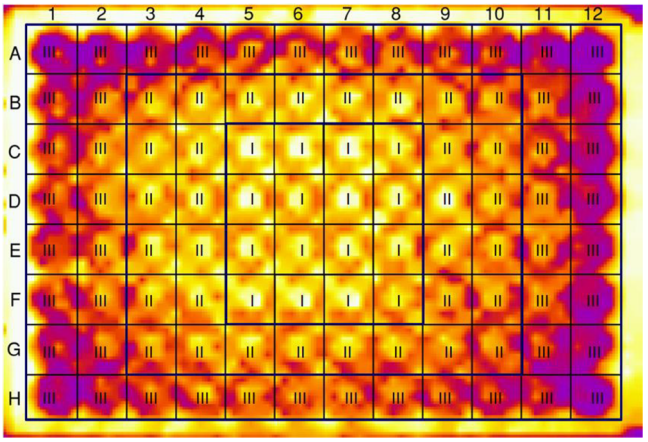


Figure 4. Infrared camera image of the 96-well microplate at a nominal 105 °C setting. The number in each square indicates the types assigned in our analysis; these proceed concentrically from the center. The close-distance infrared images captured at individual vials provide mean temperatures for Types I, II, and III locations of 97.3, 96.7, and 95.9 °C, respectively.

variations between locations in the center areas labeled as Type I (with rows labeled with C, D, E, and F and columns labeled with 5, 6, 7, and 8), in the middle areas labeled with Type II (32 totals), and those in the edge areas on the left/right side and top/bottom labeled with Type III (48 locations total with 24 locations labeled with rows A and H and columns numbered 1−12 and 24 locations with rows from B to G and columns 1, 2, 11, and 12). These variations may alter the equilibrium and/or kinetics of the inverse temperature crystallization process. Although edge effects are well-known in HTE literature,[56,57] they are predominantly treated as a binary distinction between "edge" or "interior". As we will show below, dividing the surface into three regions provides a better explanation.

We considered two different types of replication experiments: *reproduction* experiments are performed at the exact same location as the original experiment, and *relocation* experiments are moved to a different location. In both cases, the composition and the nominal (plate-level) temperature remain fixed. Because a complete electronic record of each experiment is maintained,[17,19] we were able to examine not

https://doi.org/10.1021/acs.jcim.0c01307
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

only expected influences (e.g., composition, temperature, time, etc.) but also unexpected influences (e.g., location of the reaction vials on the heating block). For each of the nine TPs in Tables S8 and S9, we performed reproduction experiments (results in Table S15) and relocation experiments (Table S19). These TP results were correctly predicted by the CFA models, despite being missed by the individual models.

From the 26 FPs and 17 FNs in Tables S10 and S11, we selected 10 and 10 from each of FP and FN groups, respectively, and identified possible commonalities among these anomalous experiments. The goal of selecting a subset of these misclassified experiments was to facilitate reproduction at the same location and experimental replication at different locations, discussed below. In total, there are 29 reproduction results comprising 9 TPs, 10 FPs, and 10 FNs experiments. (See Tables S15, S16, and S17.) There are 63 relocation results, consisting of 18 TPs, 22 FPs, and 23 FNs experiments. (See Tables S19, S20, and S21.) The 29 reproduction and 63 relocation experiments are summarized in Table 4 and Table 5, respectively. These results are tabulated based on whether the outcome of the reproduced/relocated experiment was the same as or different from the original experiment.

Table 4. Status of Different/Identical for 29 Reproduction Experiments[a]

outcome	9 TPs	10 FPs	10 FNs	total
different	4	5	4	13
identical	5	5	6	16
total	9	10	10	29

[a]Conducted at the same location as the original experiment.

Table 5. Status of Different/Identical for 63 Relocation Experiments[a]

outcome	9 TPs	10 FPs	10 FNs	total
different	15	13	13	41
identical	3	9	10	22
total	18	22	23	63

[a]Experiments moved to a different location.

Following the schematic in Figure 4, the 96-well microplate is classified into three types of locations: Type I (16 center locations), Type II (32 middle locations), and Type III (48 edge locations), respectively. Table 6 depicts the number of same and different outcomes, and the type sensitivity ratio (TSR) is defined as the ratio of the number of different to same results (13/16 = 0.81) for the 29 reproduction experiments distributed over the three types of locations: Type I (2/4 = 0.5), Type II (2/5 = 0.4), and Type III (9/7 = 1.29). Table S18 contains a breakdown of experiment outcomes for the 9 TPs, 10 FPs, and 10FNs for these different locations (Types I, II, and III). All three types (TP, FP, and FN) have fewer identical outcomes in Type III locations than in Type I or Type II. All TP reproduction experiments yielded identical results in Type I locations, and reproducibility was higher in Type II locations than in Type III locations. For FP, only identical outcomes were observed in Type I or Type II locations, but there were many different results in Type II locations. This is consistent with increased temperature sensitivity. For FN, there was no clear trend across the different types of locations, suggesting that FN model failures arise from other contributions.

Table 6. Different, Identical, and TSR of the 29 Reproduction Experiments over Locations Type I, Type II, and Type III

	location			
outcome	Type I	Type II	Type III	total
different	2	2	9	13
identical	4	5	7	16
total	6	7	16	29
TSR	0.5	0.4	1.29	0.81

Further insight can be gained by deliberately relocating experiments between different types of locations. Table 7 summarizes the number of different-result, same-result, and type sensitivity ratio (TSR) as the triple (a, b; c) for the 63 relocation experiments, tabulated based on the original position (Type X location) and new position (Type Y location) where {X, Y} = {I, II, III}. The results in Table 6 indicate that Type III locations on the 96-well microplate are more likely to have different results when reproduced; the TSR is greater than the average over all experiments. This is expected, as the edge locations can be as much as 5 °C colder (and on average are 1.4 °C colder) than the interior, which could hinder the inverse-temperature crystallization process. More surprisingly, considering the relocation experiments in Table 7, the TSRs for relocations between middle and edge locations (Type II → Type III or Type III → Type II) are higher than the background of the other experiments. It is known that the onset temperature for inverse temperature crystallization processes is highly dependent upon the composition of the solution. Our results suggest that these specific experiments have compositions where the small temperature variations between the different locations are sufficient to cause or prevent crystal formation. Experiments in Type III locations which were initially incorrectly predicted as FP or FN *would* have been correct predictions if they had been moved to a different location. However, merely augmenting the DS#44 training and testing sets with the location information (provided as a one-hot-encoded vector) did not improve any of the prediction quality of the single models by more than 0.005. This provides additional evidence that these chemical compositions are poorly described by the training data. In this way, we can use the fusion models to provide additional credibility to the predictions and use the discrepancy between predicted and actual outcomes to identify these scientifically interesting anomalies, as distinct from other types of prediction errors.

4. CONCLUSION

Model fusion using combinatorial fusion analysis (CFA), which combines multiple scoring systems (MSSs) using the rank-score characteristic (RSC) function and cognitive diversity (CD), was used to improve the prediction quality of four individual models and enhance the data quality of the HTEs. By combining the four individual models A (SVM), B (random forest), C (weighted logic classifier), and D (gradient boosted tree) in all combinatorial ways (pairs, triples, quadruples) using both score and rank combinations, we have generated 22 fusion models for each of the three combination methods: average combination (AC), weighted combination using performance (WCP), and weighted

https://doi.org/10.1021/acs.jcim.0c01307
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 7. Number of (a) Different, (b) Identical Observed Outcomes, and (c) TSR, Written as Triples (a, b; c) of the 63 Relocation Experiments from Type X to Type Y Location[a]

	Type Y			
Type X	Type I	Type II	Type III	total
I	(0, 1; 0)	(0, 2; 0)	(6, 4; 1.5)	(6, 7; 0.86)
II	(1, 1; 1)	(3, 0; *)[a]	(6, 1; 6)	(10, 2; 5)
III	(3, 3; 1)	(11, 3; 3.66)	(11, 7; 1.57)	(25, 13; 1.92)
total	(4, 5; 0.8)	(14, 5; 2.8)	(23, 12; 1.92)	(41, 22; 1.86)

[a]"*" indicates the number is not applicable.

combination using diversity strength (WCDS). The majority of these 66 fusion models (summarized in Table S3) improves the prediction quality of individual models. Among the 39 fusion models, which improve all single models A, B, C, and D, rank combination of all four models, RC(ABCD), achieves the highest accuracy Pre@357 of 0.742, a 3.9%-point increase over the best single model GBT(D).

An examination of shared attributes of the 26 reactions that are incorrectly predicted as positives (Table S10) or 17 reactions incorrectly predicted as negatives (Table S11) indicated that these reactions are predicted by all of the individual models and the fusion models. Incorrect predictions are more likely to occur on edge sites (Type III in Figure 4) of the reaction plate. Outcomes of 29 reproductions and 63 relocation experiments (comprised of 9 TPs, 10 FPs, and 10 FNs) are shown in Tables 4 and 5. Experiments at the edge locations (Type III) showed many more changes in both reproduction and relocation (Tables 6 and 7). Using a combination of infrared thermometry and experimental replication of 63 experiments to control for location changes, we identified temperature changes of the order of 7 °C at the plate edge as sufficient to change some of the reaction outcomes. The data sets originally used for machine learning model training did not contain this location information and hence could not account for this difference. Merely dividing the locations into "interior" and "edge" is insufficient to describe this trend; rather, division into an interior, middle, and edge region (Types I, II, and III) better explains the results of relocating experiments.

In addition to demonstrating that model fusion can inform and improve data and prediction quality of HTE perovskite synthesis, our work also confirms previous results using CFA framework.[40,45] In agreement with previous theoretical work by Hsu and Taksa,[40] under certain conditions involving cognitive diversity, rank combination can perform better than score combination. In agreement with our numerical findings, when the combination is better than the individual model, the rank combination fusion models (bottom half of Table S3) do perform better than score combination fusion models (top half of Table S3). Results in Yang et al.[45] demonstrate that a combination of scoring systems is better than individual systems only if they are relatively good and different. Three fusion models, AB, BD, and ABD, confirmed this assertion (Table S3). Our work not only builds upon the previous success of this approach to cheminformatics problems on virtual screening and consensus scoring[45,58] but also highlights the ability of the CFA model fusion approach to help improve quality control on high-throughput experimental studies. Based on these results, we plan to incorporate model fusion-based quality control into future versions of the ESCALATE[19] program.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c01307.

Additional details on model input features, fusion model results, and experimental replication results and discussion of shuffling significance tests and alternate analysis of center and edge division (Figures S1−S3 and Tables S1−S21) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

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

Authors

Yuanqing Tang − *Laboratory of Informatics and Data Mining (LIDM), Department of Computer and Information Science, Fordham University, New York 10023, United States*

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

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

Hamed Eramian − *Netrias LLC, Arlington, Virginia 22201, United States*

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

Alexander J. Norquist − *Department of Chemistry, Haverford College, Haverford, Pennsylvania 19041, United States*

D. Frank Hsu − *Laboratory of Informatics and Data Mining (LIDM), Department of Computer and Information Science, Fordham University, New York 10023, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c01307

Notes

The authors declare no competing financial interest.

Code used to generate the single model results and used for the model fusion study and generation of the figures for this article may be accessed at https://github.com/SD2E/test-harness and https://github.com/tyq0330/Model_Fusiong, respectively.

■ ACKNOWLEDGMENTS

We thank Scott Novotny and Nick Leiby (Two Six Technologies) for software engineering of the ESCALATion dashboard and model test system that supported this work.

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). D.F.H. acknowledges support of the LIDM by Mr. Edward Stroz.

REFERENCES

(1) Carson, N. Rise of the Robots. *Chem. - Eur. J.* **2020**, *26*, 3194−3196.
(2) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem., Int. Ed.* **2020**, *59*, 22858.
(3) Stein, H. S.; Gregoire, J. M. Progress and Prospects for Accelerating Materials Science with Automated and Autonomous Workflows. *Chem. Sci.* **2019**, *10*, 9640−9649.
(4) DeCost, B. L.; Hattrick-Simpers, J. R.; Trautt, Z.; Kusne, A. G.; Campo, E.; Green, M. L. Scientific AI in Materials Science: A Path to a Sustainable and Scalable Paradigm. *Mach. Learn. Sci. Technol.* **2020**, *1*, 033001.
(5) 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*, 73−76.
(6) 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.
(7) Saparov, B.; Mitzi, D. B. Organic-Inorganic Perovskites: Structural Versatility for Functional Materials Design. *Chem. Rev.* **2016**, *116*, 4558−4596.
(8) Jena, A. K.; Kulkarni, A.; Miyasaka, T. Halide Perovskite Photovoltaics: Background, Status, and Future Prospects. *Chem. Rev.* **2019**, *119*, 3036−3103.
(9) Quan, L. N.; Rand, B. P.; Friend, R. H.; Mhaisalkar, S. G.; Lee, T.-W.; Sargent, E. H. Perovskites for Next-Generation Optical Sources. *Chem. Rev.* **2019**, *119*, 7444−7477.
(10) 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*, 1437−1451.
(11) 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.; Elliott, M. S.; Haley, T. H.; Dvorak, D. J.; Aspuru-Guzik, A.; Hein, J. E.; Berlinguette, C. P. Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials. *Sci. Adv.* **2020**, *6*, No. eaaz8867.
(12) 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*, 1701543.
(13) Gu, E.; Tang, X.; Langner, S.; Duchstein, P.; Zhao, Y.; Levchuk, I.; Kalancha, V.; Stubhan, T.; Hauch, J.; Egelhaaf, H. J.; Zahn, D.; Osvet, A.; Brabec, C. J. Robot-Based High-Throughput Screening of Antisolvents for Lead Halide Perovskites. *Joule* **2020**, *4*, 1806−1822.
(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*, 9142−9148.
(15) Lignos, I.; Morad, V.; Shynkarenko, Y.; Bernasconi, C.; Maceiczyk, R. M.; Protesescu, L.; Bertolotti, F.; Kumar, S.; Ochsenbein, S. T.; Masciocchi, N.; Guagliardi, A.; Shih, C.-J.; Bodnarchuk, M. I.; deMello, A. J.; Kovalenko, M. V. Exploration of Near-Infrared-Emissive Colloidal Multinary Lead Halide Perovskite Nanocrystals Using an Automated Microfluidic Platform. *ACS Nano* **2018**, *12*, 5504−5517.
(16) 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.
(17) 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.
(18) Liu, Y.; Yang, Z.; Liu, S. F. Recent Progress in Single-Crystalline Perovskite Research Including Crystal Preparation, Property Evaluation, and Applications. *Adv. Sci.* **2018**, *5*, 1700471.
(19) 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.
(20) Pendleton, I. M.; Caucci, M. K.; Tynes, M.; Dharna, A.; Nellikkal, M. A. N.; Li, Z.; Chan, E. M.; Norquist, A. J.; Schrier, J. Can Machines "Learn" Halide Perovskite Crystal Formation without Accurate Physicochemical Features? *J. Phys. Chem. C* **2020**, *124*, 13982−13992.
(21) Ishikawa, K. *Guide to Quality Control*; Industrial engineering & technology; Asian Productivity Organization: Tokyo, 1976.
(22) Yaqub, O. Serendipity: Towards a Taxonomy and a Theory. *Res. Policy* **2018**, *47*, 169−179.
(23) Kauwe, S. K.; Graser, J.; Murdock, R.; Sparks, T. D. Can Machine Learning Find Extraordinary Materials? *Comput. Mater. Sci.* **2020**, *174*, 109498.
(24) Plutarch of Chaeronea, "Symposiacs, Question III". In *Plutarch's Morals*; Goodwin, W. W., Translator; Little, Brown and Company: Boston, 1878; Vol. 3, pp 242−246. https://books.google.com/books?id=zegIAAAAQAAJ&pg=PA242 (accessed 2021-01-09).
(25) Johnson, N. L.; Longmire, V. A. The Science of Social Diversity. *Los Alamos Natl. Lab Theor. Div. T-3 Fluid Dyn.*; LA-UR-99-336; 1999 ..
(26) Kuncheva, L. I. *Combining Pattern Classifiers: Methods and Algorithms*, 2nd ed.; Wiley: Hoboken, NJ, 2014.
(27) Schapire, R. E.; Freund, Y. *Boosting: Foundations and Algorithms*; Adaptive computation and machine learning series; MIT Press: Cambridge, MA, 2012.
(28) Sharkey, A. J. C. *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*; Springer Science & Business Media: 2012; DOI: 10.1007/978-1-4471-0793-4.
(29) Zhou, Z.-H. *Ensemble Methods: Foundations and Algorithms*, 1st ed.; Chapman & Hall/CRC machine learning & pattern recognition series; Chapman and Hall/CRC: Cambridge, UK, 2012; DOI: 10.1201/b12207.
(30) Zhang, Z.; Mansouri Tehrani, A.; Oliynyk, A.; Day, B.; Brgoch, J. Finding the Next Superhard Material through Ensemble Learning. *Adv. Mater.* **2021**, *33*, 2005112.
(31) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(32) Bishop, C. M. *Pattern Recognition and Machine Learning*; Information science and statistics; Springer: New York, 2006.
(33) Kittler, J.; Roli, F. *Multiple Classifier Systems*; Lecture notes in computer science; Springer-Verlag Berlin Heidelberg: 2000; DOI: 10.1007/3-540-45014-9.
(34) Wu, S. *Data Fusion in Information Retrieval*; Springer: Berlin, 2012; DOI: 10.1007/978-3-642-28866-1.
(35) Bareinboim, E.; Pearl, J. Causal Inference and the Data-Fusion Problem. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 7345−7352.
(36) Hsu, D. F.; Chung, Y.-S.; Kristal, B. S. Combinatorial Fusion Analysis: Methods and Practices of Combining Multiple Scoring Systems. In *Advanced Data Mining Technologies in Bioinformatics*; IGI Global: 2006; pp 32−62, DOI: 10.4018/978-1-59140-863-5.ch003.

https://doi.org/10.1021/acs.jcim.0c01307
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(37) Hsu, D. F.; Kristal, B. S.; Schweikert, C. Rank-Score Characteristics (RSC) Function and Cognitive Diversity. In *Brain Informatics*; Yao, Y., Sun, R., Poggio, T., Liu, J., Zhong, N., Huang, J., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2010; pp 42−54, DOI: 10.1007/978-3-642-15314-3_5.
(38) Hsu, D. F.; Kristal, B.; Hao, Y.; Schweikert, C. Cognitive Diversity: A Measurement of Dissimilarity Between Multiple Scoring Systems. *J. Interconnect. Netw.* **2019**, *19*, 1940001.
(39) Hsu, D. F.; Shapiro, J.; Taksa, I. *Methods of Data Fusion in Information Retrieval: Rank vs. Score Combination*; DIMACS TR 2002-58; 2002; p 47.
(40) Hsu, D. F.; Taksa, I. Comparing Rank and Score Combination Methods for Data Fusion in Information Retrieval. *Inf. Retr.* **2005**, *8*, 449−480.
(41) Lyons, D. M.; Hsu, D. F. Combining Multiple Scoring Systems for Target Tracking Using Rank-Score Characteristics. *Inf. Fusion* **2009**, *10*, 124−136.
(42) Schweikert, C.; Brown, S.; Tang, Z.; Smith, P. R.; Hsu, D. F. Combining Multiple ChIP-Seq Peak Detection Systems Using Combinatorial Fusion. *BMC Genomics* **2012**, *13*, S12.
(43) Batallones, A.; Sanchez, K.; Mott, B.; Coffran, C.; Frank Hsu, D. On the Combination of Two Visual Cognition Systems Using Combinatorial Fusion. *Brain Inform.* **2015**, *2*, 21−32.
(44) Kustiawan, I.; Liu, C.-Y.; Hsu, D. F. Vertical Handoff Decision Using Fuzzification and Combinatorial Fusion. *IEEE Commun. Lett.* **2017**, *21*, 2089−2092.
(45) Yang, J.-M.; Chen, Y.-F.; Shen, T.-W.; Kristal, B. S.; Hsu, D. F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134−1146.
(46) Zhong, X.; Hurley, L.; Sirimulla, S.; Schweikert, C.; Hsu, D. F. Combining Multiple Ranking Systems on the Generalized Permutation Rank Space. *5th IEEE Data Comput. Intell. Conf.* **2019**, 123−129.
(47) Diaconis, P. *Group Representations in Probability and Statistics*; Lecture notes-monograph series; Institute of Mathematical Statistics: Hayward, CA, 1988; Vol. *11*.
(48) Fligner, M. A.; Verducci, J. S. *Probability Models and Statistical Analyses for Ranking Data*; Lecture Notes in Statistics; Springer New York: New York, NY, 1993; Vol. *80*, DOI: 10.1007/978-1-4612-2738-0.
(49) Marden, J. I. *Analyzing and Modeling Rank Data*, 1st ed.; Monographs on statistics and applied probability; Chapman & Hall: London; New York, 1995; DOI: 10.1201/b16552.
(50) Gibbons, J. D.; Chakraborti, S. Nonparametric Statistical Inference. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin, Heidelberg, 2011; pp 977−979, DOI: 10.1007/978-3-642-04898-2_420.
(51) Chung, Y.-S.; Hsu, D. F.; Tang, C. Y. On the Relationships Among Various Diversity Measures in Multiple Classifier Systems. In *2008 International Symposium on Parallel Architectures, Algorithms, and Networks (i-span 2008)*; IEEE: Sydney, Australia, 2008; pp 184−190, DOI: 10.1109/I-SPAN.2008.46.
(52) Melnik, O.; Vardi, Y.; Zhang, C.-H. Mixed Group Ranks: Preference and Confidence in Classifier Combination. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 973−981.
(53) Pendelton, 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; DOI: 10.18126/LYK3-QACE (accessed 2021-03-31).
(54) Li, Z.; Najeeb, M. A.; Alves, L.; Sherman, A.; Shekar, V.; Parrilla, P. C.; Pendleton, I. M.; Nega, P. W.; Zeller, M.; Schrier, J.; Norquist, A. J.; Chan, E. M. Interactive Data Visualization and Analysis Interface. https://github.com/darkreactions/rapid (accessed 2021-03-31).
(55) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; 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.
(56) Grosch, J.-H.; Sieben, M.; Lattermann, C.; Kauffmann, K.; Büchs, J.; Spieß, A. C. Enzyme Activity Deviates Due to Spatial and Temporal Temperature Profiles in Commercial Microtiter Plate Readers. *Biotechnol. J.* **2016**, *11*, 519−529.
(57) Burt, S. M.; Carter, T. J. N.; Kricka, L. J. Thermal Characteristics of Microtitre Plates Used in Immunological Assays. *J. Immunol. Methods* **1979**, *31*, 231−236.
(58) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of Data Fusion Methods in Virtual Screening: Similarity and Group Fusion. *J. Chem. Inf. Model.* **2006**, *46*, 2206−2219.
(59) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Model.* **2001**, *41*, 1422−1426.

https://doi.org/10.1021/acs.jcim.0c01307
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX