FRONT MATTER

Title

Serendipity based recommender system for perovskites material discovery: balancing exploration and exploitation across multiple models

[short] Serendipity recommender for perovskites discovery

Authors

Venkateswaran Shekar,[1] Vincent Yu,[1] Benjamin J. Garcia[2], David Benjamin Gordon[2], Gemma E. Moran[3], David M. Blei[3], Loïc M. Roch[4], Alberto García-Durán[4], Mansoor Ani Najeeb[5], Margaret Zeile[5], Philip W. Nega[6], Zhi Li[6], Mina A. Kim[6], Emory M. Chan[6], Alexander J. Norquist[5], Sorelle Friedler,[1]* Joshua Schrier[7]

Affiliations

[1] Department of Computer Science, Haverford College, Haverford, PA, USA
[2] Synthetic Biology Center, Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[3] Data Science Institute, Columbia University, New York, NY, USA
[4] Atinary Technologies, Lausanne, Vaud, Switzerland
[5] Department of Chemistry, Haverford College, Haverford, PA, USA
[6] Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[7] Department of Chemistry, Fordham University, The Bronx, NY, USA

*Corresponding author: sorelle@cs.haverford.edu

Abstract

Machine learning is a useful tool for accelerating materials discovery, however it is a challenge to develop accurate methods that successfully transfer between domains while also broadening the scope of reaction conditions considered. In this paper, we consider how active- and transfer-learning methods can be used as building blocks for predicting reaction outcomes of metal halide perovskite synthesis. We then introduce a serendipity-based recommendation system that guides these methods to balance novelty and accuracy. The model-agnostic recommendation system is tested across active- and transfer-learning algorithms, using laboratory experiments for training and testing and a time-separated hold out that includes four different chemical systems. The serendipity recommendation system achieves high accuracy while increasing the scope of the synthesis conditions explored.

Teaser

- Serendipity recommendation system guides laboratory tests of perovskite synthesis experiments to success and exploration.

MAIN TEXT

Introduction

Machine-learning provides many new tools for advancing experimental science.(*1*) One specific application involves *autonomous experimentation systems (AES),* in which algorithms specify an iterative sequence of new experiments. These experiments are

conducted in an automated fashion, and the results captured with minimal human intervention. Recent reviews discuss progress on AES in materials science,(*2*, *3*), organic chemistry,(*4*) inorganic chemistry,(*5*) nanoscience,(*6*) biomaterials,(*7*) and formulations.(*8*)

Beyond the significant software and hardware engineering challenges, better algorithms are needed for directing autonomous experimentation systems.(*1*) Such algorithms would ideally utilize historical information and physical theories to make more accurate predictions,(*9*) and use recommendation systems that use these models to determine experimental choices. To date, recommendation systems used in chemical domains have largely focused on achieving high recommendation accuracy,(*10*, *11*) with a few focusing instead on encouraging exploration.(*12*, *13*) However, advances in recommendation systems now allow for the balancing of objectives to capture the desire to explore the chemical space while simultaneously achieving high accuracy.(*14*, *15*)

Towards that end, this paper compares the performance of an exploitative recommendation system that optimizes solely based on accuracy, and a *serendipity*-based recommendation system, which balances between accuracy and diversity measure based on distances in the chemical space. Eleven candidate active-learning models were considered in a common laboratory test setting, tested across four different chemical systems that made up a hold-out test set. The AES used was an automated system for high-throughput inverse temperature crystallization (ITC) growth of halide perovskites.(*16*) In addition to general interest in metal halide perovskites for high-performance photovoltaics and optoelectronic applications,(*17*) the relatively mild, solution-based syntheses for these materials make them amenable to high-throughput automated experimentation (reviewed in Ref. (*18*)). The ability to incorporate a different organic cations results in a vast, chemically diverse space to explore for new compounds.(*19*) The underlying problem of finding reaction conditions resulting in crystal formation is representative of a broad class of chemistry/materials optimization problems—controlling compositional variables in a highly non-ideal mixture to achieve a desired thermodynamic and kinetic goal, in the absence of a precise models with experimental noise in the reaction outcomes. All models were provided with the same initial data (including uniformly sampled historical data), and then had the ability to request a fixed budget of laboratory experiments as part of training of the active learning models. All models were evaluated using the exploitative recommendation system. Based on these results, four high-performing models and two baseline models were then given another fixed budget of laboratory experiments to recommend using the serendipity recommendation system. Independent of the tests, we acquired an experimental random baseline dataset for assessing the improvement of the algorithms.

While exploitative recommendation systems often succeed in narrowly defined laboratory optimization tasks, they can get trapped in local minima and do not request a diverse set of experiments. This can result in missing potentially new phases—for example, we recently reported a morpholinium lead iodide system where small concentration changes result in phases with distinct structural and optical properties.(*20*) The serendipity-based recommendation system can be applied to any model to increase the recommendation diversity while keeping the probability of success high. Laboratory comparisons indicate that serendipity-directed recommendation improves the diversity of recommendations, which in turn improves the robustness of the recommendation against initialization conditions, without substantially degrading recommendation success.

Results

The underlying modeling problem is of the following form: Given a chemical system comprised of a metal halide, one specific organoammonium halide salt (for brevity, we refer to this as the "amine"), a solvent, and an additive (e.g., formic acid)—find the set of concentrations for each of these species, that result in the formation of a large, high-quality single crystalline product via an inverse temperature crystallization reaction.(*16*, *21*) Concentrations of the three solutes (metal halide, amine, and formic acid) define a 3-dimensional space, and only compositions within the convex hull of the initial stock solutions are feasible.(*22*) Precision limits on the robotic liquid handler result in a discrete state set grid of approximately 20,000 feasible compositions within the composition space for each chemical system. The goal of the machine learning models is to predict which of these compositions result in crystal formation, and the subsequent goal of the recommendation system is to propose composition that result in crystal formation while also exploring the chemical space.

Eleven active learning models were assigned the initial prediction task: Bayesian Additive Regression Trees with Transfer Learning (BART),(*23*, *24*) PLATIPUS (PLT),(*21*, *25*) Bayesian Optimization using Gaussian Processes (BGP),(*26*) Falcon (FAL), Falcon GPBO (G),(*26*, *27*) Falcon DNGO (D),(*28*) Gryffin (GR),(*29*) Gaussian Process with Transfer Acquisition Functions (TA)(*30*), and Falcon with Historical Data (FH), as well as active learning k-Nearest Neighbors (KNN)(*31*, *32*) and Decision tree (DT)(*33*) models which serve as a baseline for model performance. (See Methods for a complete description of each model.) Data and code are available in the Supporting Information.

Fig. 1 shows an overview of the experimental campaign designed to evaluate the model recommendations. The general structure of the campaign, including the historical dataset, experimental details, and performance of the PLATIPUS model were previously described in Ref. (*21*). The campaign is split into model training and refinement, active learning, and recommendation phases. In the initial model training and refinement phase (Fig. 1A), each model has access to prior historical data comprised of 1722 total reactions on 19 amines (see Table S1). In addition to the raw experimental description of concentrations and outcomes, the input features are augmented with physicochemical properties and cheminformatics descriptors relevant to the amine for a total of 54 input features per reaction (Table S2). Models could use this historical information for training and hyperparameter tuning as desired, as well as augment the input features with custom transfer learned features (in the case of BART). Subsequent phases test the ability of each model to operate new, time-separated hold-out chemical systems, specifically four amines (4-chlorophenethylamine, 4-chlorophenylamine, 4-hydroxyphenethylamine, dimethylamine) absent from the historical dataset. To establish a statistical baseline, we first collected experimental data for 96 experiments drawn uniformly from the feasible stateset for each new amine. Success rates for these randomly selected reactions are shown in Table 1. To initialize the active learning phase (Fig. 1B), each model was provided with the same set of 10 initialization samples, drawn randomly from the statistical baseline data. To study the robustness of model performance with respect to initialization samples, each model is provided with two draws of initialization data. This results in a total of $11 \times 4 \times 2 = 88$ total amine-specific models evaluated in the campaign. Each model is then allowed 10 sequential active learning experiment requests. At each iteration, the model identifies one new stateset point, with maximum uncertainty for classification

models (KNN, DT, PLT and BART) and a combination of crystal score and model uncertainty for regression models (FAL, G, D, GR, TA, FH) to aid in the active learning training. Details related to query selection are provided in the methods section. The requested experiments are performed in the laboratory, and then each model is updated with the new result that it requested. The number of initialization and sequential experiments is intentionally small, as this allows us to assess the feasibility of these algorithms for future non-automated experiments. After completing the iterative active learning, each model enters a recommendation phase where it is allowed to request 9 reactions, to be run in parallel.

Recommender systems in their simplest form, generate a ranked list that tries to predict the most relevant items based on a user or application's constraints.(*34*) One common way to create a recommendation system focused on exploitation is to select the experiments with the highest predicted probability of success based on the underlying model; we term this recommendation system *exploitative* (Fig 1C). The *serendipity*-based recommendation system (Fig 1D) we develop is based on the idea of balancing exploitation with exploration. From an information retrieval perspective, the fraction of recommended reactions that successfully result in crystal formation in the laboratory (denoted *success fraction*) corresponds to item relevance or accuracy. Recommendation serendipity quantifies the joint combination of relevance and surprise;(*14*) the latter is quantified by the distance from other examples in the data-item feature space. (See Methods.)

Exploitation Recommender. The accuracy (success fraction) of each model's exploitation recommendations can be measured either globally (over all 8 tasks), on a per-amine basis, or on each draw of initialization data for each amine. (See Table 1.) Evaluating the accuracy in each of these divisions gives a useful perspective on the model's applicability to experimental tasks. Global success fraction gives a sense of how well models perform generally on $9 \times 4 \times 2 = 72$ total recommendations, and thus the best sense of overall recommendation quality over a wide range of problems. Globally, Falcon (FAL, 0.83) gave the best recommendations, followed closely by Bayesian Additive Regression Trees (BART, 0.79) and Bayesian Optimized Gaussian Processes (BGP, 0.75). In general, all models performed better in the global task than random sampling (0.36), except for the decision tree (DT, 0.22) and KNN baselines (KNN, 0.08).

Per-amine success fraction for each model allows us to see how well models fare in tasks of varying difficulty, as the underlying base rate of success for a random reaction differs for these systems on 18 recommendations. No single model was a clear winner, but BART performed best for two amines (dimethylamine and 4-chlorophenylamine), whereas Falcon with historical data (FH) did best on 4-chlorophenethylamine, and BGP did best for 4-hydroxyphenethylamine. In general, BART is always among the top-3 models for all four amines tested; the other models are not as consistent, and FAL and BGP are only in the top-3 for two of the four amines. Several of the models—KNN, Falcon DNGO (D), Gryffin (GR) and Gaussian Process with Transfer Acquisition function (TA)—failed to recommend successful reactions for at least one of the tested amines. As the per-amine results consist of a limited sample of 18 requests, we also quantified the uncertainty by computing the conjugate prior estimate of the success rate given these observations, i.e., the distribution of success rates consistent with the finite observation;(*21*) results are shown in Fig. S1.

Per-draw success fraction tells us about the sensitivity to the initialization data provided to initiate the active learning cycle, evaluated on 9 recommendations each. An ideal model

should be robust, i.e., capable of recommending some successful reactions independent of the chemical system or the initialization data. On a per-draw basis, 6 out of the 11 models fail to recommend at least one success for a given draw. Curiously, this includes the Falcon (FAL) model, which despite the highest global success was unable to successfully identify crystal growth conditions for a single draw of 4-Chlorophenylamine, yet made perfect recommendations in the other draw of this amine. A closer examination of the data showed that Falcon recommended the same experiment as the successful experiment it requested during the sequential active learning stage, as well as other experiments that were immediately adjacent in the stateset. However, none of these recommendations resulted in crystal formation during the recommendation stage during this trial. A similar result was observed with Falcon GPBO (G) having a high average success fraction of 0.736 but was unable to predict successful reaction for a single draw of 4-Chlorophenethylamine, despite being successful in 6 of the 9 recommendations for the other draw of the amine.

Recommendation diversity is important in scientific discovery tasks, but an exploitation-focused recommender does not incorporate this as a criterion. As a representative example, the successful experiments predicted by different models using the exploitation recommender for 4-Hydroxyphenethylamine are shown in Fig. 2, where the axes are the concentration of different chemical components used in the experiment. Many models, including BART, BGP, FAL, FALGP and GRYF, are clustered together in chemical space. This is expected, as the exploitation recommender optimizes for experiments with the highest probability of success, which tend to have similar experimental conditions to those already sampled during the initialization and active learning phases. Further, each model mentioned above has two clusters of points corresponding to the initialization experiments used to train the model. Thus, these models not only explore a small fraction of the chemical space but also cluster in various parts of the space owing to the sensitivity to initialization samples. We selected a subset of the models for further study and experimental testing, specifically the best performing models (FAL, BART, BGP and PLT) and the two baseline models (KNN and DT). To assess the fraction of the chemical space explored, we consider the *volume fraction*—the volume of the compositional space that these experiments cover normalized by the total volume of the achievable convex hull. Fig. 3 shows each model's volume fraction explored versus how successful it was in predicting crystallization averaged over all amines. We see that these models demonstrate an exploration versus exploitation tradeoff; the high-performing models (BART, FAL, and BGP) explore a smaller fraction of space but achieve high success (consistent with our observations regarding Fig. 2), the low-performing models (DT and KNN) better explore the space but perform poorly, and PLT finds an intermediate point in this trade-off. We use this as a baseline to observe the change in success fraction and volume explored when the serendipity recommender is applied to the same models with the goal of better balancing success and volume fraction.

Serendipity recommender. Incorporating the serendipity constraint in the recommendation increases the volume of compositional space explored, while not decreasing the *success fraction* (Table 2). Fig. 4A shows the volume fraction versus success fraction for the experiments selected by the exploitation recommender and serendipity recommender. In general, the average volume of chemical space explored increases when the serendipity recommender is applied to all models except for DT where the average volume fraction decreases. Despite increasing the range of chemical space, the success fraction increases for the DT and KNN baseline models and only slightly changes for BGP, PLT, BART and

FAL models. As expected, the average serendipity metric in recommended experiments increases when this constraint is added (Fig. 4B).

The changes in each of these measures are depicted in Fig. 5 and Fig. 6. In Fig. 5, volume fraction, success fraction, and serendipity measures are indicated by the circle, triangle, and square points, respectively, for the exploitation recommender on the x-axis and the serendipity recommender on the y-axis. Points above the unit slope indicate higher values for the serendipity recommender compared to the exploitation recommender. Fig. 6 is a bar plot showing a side-by-side comparison of the volume fraction, success fraction and serendipity measures for the two recommenders for different models. The serendipity measures (squares, Fig. 6A) improve when the serendipity recommender is used. The fraction of successful reactions predicted by models, indicated by triangular points (Fig. 6B), changes slightly for most models and increases significantly for models such as KNN and DT. The volume fraction explored by models, indicated by circles (Fig. 6C), increases for all models except for DT. Thus, the serendipity recommender forces models to explore a large space of experimental parameters (indicated by the volume fraction and serendipity measures) while maintaining a high success rate.

On a per-amine basis, the probability estimates for each amine remain high for all models, except for 4-chlorophenylamine (Fig. S2). BGP and FAL are consistently top performers with the serendipity recommender. On a per-draw basis—and in contrast to the exploitative recommender results—none of the tested models fail to yield at least one success in every draw when the serendipity recommender is used. This indicates that the serendipity recommender constraint improves robustness against initialization conditions.

Discussion

Although a variety of algorithms have been developed and tested for autonomous chemistry and materials experimentation,(*2*–*8*) there has not yet been a direct comparison of different algorithms in an experimental laboratory setting.(*35*) To date, large-scale comparisons of different algorithmic methods have been performed computationally on previously collected datasets, rather than on live laboratory experiments. Specifically, Rohr et al. benchmarked sequential learning algorithms on a dataset of oxygen evolution reaction catalysts obtained from a historical set of combinatorial experiments,(*36*) and Liang et al. benchmarked Bayesian optimization-based and random forest models on five different materials datasets.(*37*) Comparisons on pre-existing datasets avoids the time and costs of laboratory experimentation and allows for a broad survey of possible algorithms and hyperparameters. However, this restricts algorithm requests to a limited set of pre-existing data. The distribution of these historical data may incorporate anthropogenic or algorithmic biases that degrade model performance.(*38*) The best comparison would approach would be to acquire unbiased, uniformly sampled data and then evaluate the algorithms on new problems based on the experiments that they select. As such, the results described above address this lack of direct comparison.

Unlike the majority of the recommendation systems literature, which considers domains in which there are multiple users who may have different notions of relevance and surprise,(*39*) the recommendation systems we introduce here consider a single-user case. In multi-user cases the determination of relevancy is often subjective based on user opinion and thus hard to assess;(*39*) this work benefits from determining relevancy directly based on reaction success which allows direct assessment of the quality of the resulting

recommendations. With the exploitative recommender, all the tested active learning algorithms are better than random sampling, when judging by reaction success. In general, Falcon (FAL), BART, and BGP are the best across the entire study, and we generally recommend them for future work. Despite comparable performance, these three models take very different approaches to the problem. FAL and BGP use only the small amount of data (10 initialization experiments plus 10 sequentially selected active learning experiments) specific to the current amine being investigated to make their recommendations. BGP is even more restrictive, as it only makes use of the organic, inorganic, and formic acid concentrations as input variables to make the predictions, ignoring the other stoichiometric and physicochemical properties that were provided as possible model inputs. This suggests that in chemical optimization tasks like this one, large historical data sets are not necessarily required to create task specific models with high performance. This is especially relevant for the many scientific research applications where experimental databases do not yet exist.(*1*) In contrast, BART uses all of the input features and historical data, as well as incorporating its own transfer-learned latent representation of the amine molecule structure. This has clear benefits, as BART performs consistently well across all four amines tested, indicating the value of making use of historical information when available. However, here too, the total amount of historical data is relatively small (only 1722 data items for 19 different amines), indicating the applicability of these methods for the broad class of scientific problems where only modest datasets are available.

Exploitative recommendations that try to maximize success alone are not as robust to the possible uncontrolled experiment variations —such as laboratory humidity,(*40*) impurities,(*41*) stock solution preparation and aging,(*42*) liquid handler calibrations(*43*)— that can affect crystal growth. For example, despite Falcon's overall strength, we observed one instance where it failed to find a solution due to non-representative active learning data. As noted above, during one of the trials for the 4-chlorophenylamine iodide system, the model encountered "successes" during the active learning phase which produced a crystal, and then assigned high success probabilities to this and other similar experiments. However, subsequent recommendations based on these data failed to yield crystals. Crystal growth experiments can be strongly affected by subtle differences in reaction conditions, owing to experimental precision limitations, which can lead to imperfect replicability as the process is inherently stochastic.(*44*) This illustrates the need to incorporate diversity into the recommendations. Furthermore, reaction recommendations that purely try to maximize success can often be trapped in local minima possibly due to bias in the initial training data, and do not necessarily sample diverse compositions which might contain new, unexpected outcomes.

Incorporating a definition of reaction serendipity into the recommendations provides a model-agnostic way to improve robustness against noise with respect to initialization data, allowing all models to find some successes regardless of how they are initialized. While this slightly reduces the overall success rate, it encourages more diverse sampling of reaction composition, which in turn reduces the risk of failure and facilitates discovery of unexpected phenomenon. For example, different concentrations in the 4-chlorophenethylamine iodide system result in two different crystalline products, a *red* perovskitoid and a *yellow* non-perovskite (formed from decomposed N,N-dimethylformamide), both of which are considered a "success". Recommending a diverse set of reactions increases the likelihood of identifying both phases, rather than being trapped in a local minimum of one or the other. While we have focused on the problem of

optimizing reaction composition, the serendipity approach could be applied to compositional optimization problems by incorporating a suitable composition-space distance metric(*45*) during in the calculation.

Our serendipity metric closely follows its use in the recommendation systems field.(*14*, *15*) Within this field, "serendipity" is meant to introduce a notion of surprise—here quantified via distance from recommended reactions to historical reactions—that goes beyond novelty, where novelty is meant to indicate that recommendations are previously unknown, relevant, and different.(*14*, *46*) In the sociology of science, "serendipity" is, similarly, used to express the notion of researchers making unexpected and beneficial discoveries, but the precise definition, significance, and broader implications have varied.(*47*) A variety of metrics and tools have been proposed for assessing and aiding scientific serendipity, often with different goals.(*48*) Yaqub's taxonomy distinguishes aspects of scientific serendipity in terms of whether there exists a target line of enquiry and whether the solution is for a pre-existing problem or a "solution waiting for a problem."(*49*) Within this taxonomy, the approach taken here corresponds to Mertonian serendipity, wherein the search is conducted with a defined problem in mind (specifically: finding the parameters that will result in crystal formation for a given chemical system) and where the obtained solution is to that same problem, but in a new way. Operationally, this corresponds to proposing reactions with the broadest variety of possible successes. Incorporating serendipity into the reaction selection phase in this way reduces the number of experiments needed to find surprising results, and thus offers an efficiency improvement over the "accelerated serendipity" in which brute force experimentation is used to identify novel reactions.(*50*) It is also distinct from Walpolian serendipity in which a targeted line of inquiry results in data that solves *another* problem; a recent example of this type of serendipity is the use of data initially collected to identify perovskite crystallization conditions which was subsequently analyzed for latent information about the role of ambient humidity in the crystallization process.(*40*) The serendipity discussed here is also in distinction to the scientific literature retrieval methods reviewed in Ref. (*48*).

Famously, there is "no free lunch" in machine learning—an algorithm that performs well on a certain class of problems may not perform well on other problems.(*51*) Thus, while there is no guarantee about the applicability to chemistry problems in general, we have assessed some aspects of the generality of the serendipity recommendation system and active learning models by benchmarking across four distinct chemical systems and eleven models. The self-imposed limits on the acquisition of task-specific data (10 initialization data, and 10 sequential experiments) makes our results relevant to both automated and non-automated experimentation. To facilitate broader use of these models, we have provided source code which can be used for other problems. In addition, all data used and generated in this study is provided, allowing other users to test their own methods against this benchmark task.

Materials and Methods

Experimental Design. The experimental procedure for the high-throughput inverse temperature crystallization (ITC) synthesis of metal halide perovskitoid single crystals was described in our previous work.(*16*, *21*) A Hamilton Microlab NIMBUS liquid handler

pipettes four different stock solutions into glass vials on a 96-well microplate. These stock solutions consist of (a) lead (II) iodide and the selected organoammonium iodide in solvent, (b) organoammonium iodide in solvent, (c) neat solvent, and (d) neat formic acid. N,N-Dimethylformamide (DMF) was used as the solvent for the experiments described here. For brevity, the text refers to the organoammonium salt by the name of the corresponding freebase amine. Reagent stock solutions are dispensed into the pre-heated (70 °C) vials, the vials are vortexed for 35 min to ensure proper mixing, then heated to 95 °C for 150 min without vortexing to allow for crystal growth. Reaction outcomes are scored by visual inspection into the four classes: (1) no solid observed in the solution; (2) fine powder observed; (3) small crystals observed; (4) large crystals observed (>0.1 mm). Fig. S3 shows representative outcome images for these classes for all the amines tested during the experimental campaign. Visual inspect was more reliable than computational image processing, due to reflections on the glass vials. In this study, the outcomes are reduced to binary values for the classification models, with large (class 4) crystals considered as successful (denoted as a classification outcome of 1) and all other classes are considered as failed experiments (classification outcome of 0). Regression models use the full range of crystal scores to make predictions. A raw data file containing a description of the stock solution concentrations used for each experiment, as well as details of the pipetting instructions, final compositions, and outcomes of each reaction is in the supplementary materials.

For each of the four amines tested in the experimental campaign, we initially acquired 96 experiments sampling the concentrations uniformly in the achievable 3-dimensional composition space(*22*) to serve as a statistical benchmark for the models. To initialize the models for the active learning cycle (Fig. 1B), two draws of 10 experiments were selected using uniform random sampling from this pool. Models requested 10 additional experiments sequentially from the stateset of possible achievable compositions for the amine. Because only one experiment is requested by each model at a time, the requested experiments were dispensed by manual pipetting, but otherwise follow the same experimental process described above. At the conclusion of the experiment, the results were returned to the models. Each ITC experiment requires approximately 4 hours to complete, allowing for 2 active learning rounds per amine per day. At the conclusion of the active learning cycle each fully trained model made 9 recommendations using either the exploitation recommender (Fig. 1C) or the serendipity recommender (Fig. 1D). These experiments were conducted using the liquid handler robot, batching together the recommendations made by all tested models.

Dataset Description. The historical datasets were acquired using the automated ITC experimental procedure described above, and previously described in Ref. (*21*). Each data item describes an inverse-temperature crystallization (ITC) metal halide perovskite synthesis by including concentrations of lead iodide, formic acid, and an organoammonium cation (referred to as the amine), other reaction conditions (such as temperature), and outcomes. From our collected set of historical data, we extracted only experiments performed at the nominal 105 °C (correspond to an actual temperature of 95°C, as measured by infrared thermometry), where the concentrations were sampled uniformly over the achievable convex hull of compositions, and for which at least one successful outcome was observed. Of the 20 amines satisfying these criteria in the historical data, one amine (Dimethylammonium iodide) was held out to be used as part of the laboratory test experiments. In addition, we acquired a uniformly sampled baseline for three additional amines for which we had no previous data, to demonstrate the resulting models on a true

time-separated hold-out set. Table S1 summarizes the amines included in the training and testing phases of the study, and the number of experiments from the historical dataset, and the observed success fraction. The ESCALATE software(*52*) was used to append stoichiometric and physicochemical descriptors from the raw record of reaction conditions and amine structure. In total, each experiment is described by 50 input features: 28 molecular descriptors (number of atoms, rotatable bond counts, etc.), 7 reaction conditions (concentration of acid, organic and inorganic compounds in solvent), and 15 stoichiometric descriptors. The full list of included features can be found in Table S2. The 44 numerical features in the dataset were standardized to zero mean and unit variance on the training data, and these training mean, and variance were used to rescale the values for subsequent active learning experiments. The complete dataset is available at https://github.com/darkreactions/serendipity_recommender

Models Tested

KNN: K-Nearest Neighbors(*31*, *32*) serves as a baseline model, with the prediction generated by the single (k=1) Euclidean nearest example in the training set, using sklearn.neighbors.NearestNeighbors in scikit-learn 0.23.2. Previous work has indicated that 1-NN is an appropriate baseline for model memorization.(*53*). Exploitation recommender results described here were previously reported in Ref. (*21*).

DT: Decision Tree(*33*) serves as a baseline model with a maximum depth of 12, 5 minimum samples per split and 1 minimum sample per leaf and class weight ratio of 0.105 for failure and 0.895 for success. These calculations were performed using the sklearn.tree.DecisionTreeClassifier in scikit-learn 0.23.2. Exploitation recommender results presented here were previously reported in Ref. (*21*).

BART: Bayesian Additive Regression Trees with transfer learning (BART+TL) approach consisted of two steps. First, additional features were obtained for each amine using transfer learning. Second, the probability of crystallization was modeled using Bayesian Additive Regression Trees.(*23*) The transfer-learned features were obtained using the Chemprop neural network model(*24*) which predicts molecular properties by using convolutions centered on molecular bonds. Specifically, the Chemprop model was trained on a dataset of 118,360 molecules curated from the eMolecules.com database. The eMolecules dataset was created by selecting all amines and amides based on the following SMARTS patterns
s=’[#7]’ s!=’[OX2H]’ s!=’[OX2H]’ s!=’[OH]’ s!=’[OX2,OX1-][OX2,OX1-]’
s != ’[#2, #3, #4, #5, #11, #12, #13, #14, #15, #18, #19, #20, #21, #22, #23, #24, #25, #26, #27, #28, #29, #30,#31,#32,#33,#36,#37,#38,#39,#40,#41,#42,#43, #44, #45, #46, #47, #48, #49, #50, #51, #52, #54]’ and with molecular weight less than 370, water accessible surface area between 15.05-149.19 and van der Waals volume between 42.58-229.29. The Chemprop model took as input the SMILES string of each molecule and was trained to predict 44 molecular descriptors (accessible surface area, rotatable bond count, minimal projection area etc). For the Chemprop model, the dimension of the hidden layer was set to 50, otherwise default Chemprop settings were used. After training, the transfer-learned features for each of the 19 historical and 4 evaluation amines were obtained by passing their SMILE strings through the trained Chemprop model and extracting the final layer of the neural network as a set of 50 additional features for each amine considered here. Consequently, the probability of crystallization was modeled using a total of 101 input features: 51 features related to the reaction conditions, stoichiometry, and chemist-curated properties and 50 features from the Chemprop neural network. A BART model was applied to these features to predict the probability of crystallization The BART model was fitted using the dbarts package in R.

PLATIPUS(*25*) (Probabilistic LATent model for Incorporating Priors and Uncertainty in few-Shot learning) extends the model agnostic meta-learning (MAML)(*54*) model which enables a model to quickly adapt to new amines using only a few data points and iterations. PLATIPUS constructs a Bayesian network for data and parameters and estimates uncertainty in key parameters. PLATIPUS computes approximate posterior distributions of amine-specific parameters and meta-parameters by maximizing variational lower bound of log likelihood function. Each amine is considered a meta-task. The application of PLATIPUS to perovskite experiments using the exploitation recommender was previously reported in Ref. (*21*), and the corresponding results described here are taken from that paper.

BGP: A Gaussian process-based Bayesian Optimization (BO) algorithm utilizing spatial constraints was implemented using the GPyOpt package(*55*) with the following parameters: Matern32 kernel, local penalization, automatic relevance determination, jitter, and the expected improvement acquisition function. Constraints for the BO were generated by computing a convex hull from the stateset for each amine using the "ConvexHull" function from scipy. For optimization, the independent variables were: molar organic concentration, molar inorganic concentration, and molar acid concentration. The crystal score was considered as the dependent variable, and the values were treated as negatives for minimization within the BO. For each round of active learning, a location within the stateset was chosen based on the lowest Euclidean distance to the location suggested by the "suggest_next_locations" function. Resulting crystal scores for the tested location were then incorporated into the training data to compute a suggestion for the next round. In the model validation round, CMA-ES(*55*) was utilized to identify points in the stateset likely to exhibit high-quality crystals. CMA-ES was run 20 times, seeding each search with the 20 points measured in the training and active learning rounds (1 point per CMA-ES run). As different seeds would often converge to the same minima, CMA-ES produced too few distinct final guesses. To fill in the remaining guesses, "suggest_next_locations" was executed 100 times. The resulting locations were ranked based on expected crystal score, and the remaining guesses were filled in with the top-scoring points. As before, the locations were mapped to their closest locations in the stateset.

Atinary™ Falcon: Falcon is a general-purpose optimization algorithm developed by Atinary Technologies which can solve optimization problems that include continuous, discrete and/or categorical variables with or without physicochemical descriptors, as well as batch-constrained optimization. Among other enhancements, both Atinary™ Falcon GPBO and Atinary™ Falcon DNGO empower existing Bayesian optimizers based on Gaussian Processes(*26*, *27*) and Neural Networks(*28*), respectively, with the capacity for users to easily deal with arbitrary combinations of continuous, discrete and categorical parameters (with or without descriptors) as well as with batch-constrained optimization problems.

Atinary™ Falcon GPBO: Gaussian Processes are probably the most standard choice of surrogate model for problems that are conceivably optimized with a relatively small number of queries.(*26*, *27*) However, its computational cost scales cubically in the number of queries-which hinders its usage for more challenging problems.

Atinary™ Falcon DNGO: Neural Network based Optimization technique(*28*) that maintains desirable properties of the Gaussian Processes (e.g. management of uncertainty) while improving its scalability from cubic to linear in the number of queries.

Gryffin: Gryffin(*29*) is an extension of Phoenics(*56*) to also handle categorical and discrete parameters. As opposed to GPBO and DNGO, its surrogate model does not provide information about the uncertainty of the predictions. However, this limitation is overcome with a novel acquisition function that still enables balancing between exploration and exploitation.

GPTAF: Gaussian Processes (GPs) with Transfer Acquisition Functions (TAF)(*30*) approach injects prior knowledge directly into the optimization strategy via an acquisition function that leverages independently learned GPs for each of the amines.

Atinary™ Falcon leveraging historical data (FH): This is the same method, but training set to include historical data collected for the 19 amines.

Serendipity based recommender system. The recommender system is modeled as a constrained optimization problem. Given a set of candidate recommendations *C*, the recommender provides a set of recommended reactions *R*, where $|R| \leq |C|$. The elements in *R* are determined by

$$R = \text{argmax} \left(\alpha \sum_{x_i \in R} \mathbf{rel}(x_i) + (1-\alpha)\, \mathbf{obj}\,(R) \right)$$

Where $\mathbf{rel}(x_i)$ is the recommendation relevance of candidate x_i, i.e., the predicted probability of the experiment being successful, and $\mathbf{obj}(R)$ is an additional objective function that captures the diversity and serendipity of the recommended reaction. The weight hyperparameter α is determined through a validation process, described below.

Serendipity of a recommended experiment is the notion that it is unknown to the user, relevant and different from the list of already conducted experiments.(*14*) Serendipity can be measured using a distance or similarity metric between two reactions, that is serendipity between reactions is inversely proportional to their distances. The objective function attempts to maximize the Euclidean distance between a recommended reaction and a historical reaction as well as other previously recommended reactions. These distances are known as *inter-list* serendipity and *intra-list* serendipity respectively. Therefore, the objective function is defined as

$$\mathbf{obj}(R) = \lambda\, \mathbf{S}(H, R) + (1-\lambda)\, \mathbf{S}(R, R)$$

where *R* denotes current list of recommendations and *H* is the list of historical reactions. $\mathbf{S}(H, R)$ is the serendipity of recommendations *R* with respect to historical reactions *H*, defined as

$$\mathbf{S}(H, R) = \frac{1}{|R|} \sum_{x_i \in R} \min_{x_j \in H} \tanh\left(\frac{\|x_i - x_j\|}{n} \right)$$

where $\|x_i\|$ denotes the Euclidean (L2) norm of x_i, n is the number of features describing the experiment (fixed to 50 as hyper-parameters α and λ are determined using the 50 features used in the baseline models) and $\tanh x = (e^x - e^{-x})/(e^x + e^{-x})$ is the hyperbolic tangent. Applying the hyperbolic tangent to the Euclidean distance between items normalizes the non-negative distances to be within the range of $[0, 1]$, preventing either the either the interlist or intralist serendipity values from dominating the final objective function. The serendipity of a recommended reaction with respect to other recommendations in the list is defined as

$$\boldsymbol{S}(R, R) = \frac{1}{|R|} \sum_{x_i \in R} \min_{x_j \in R \setminus \{x_i\}} \tanh\left(\frac{\|x_i - x_j\|}{n} \right)$$

where the notation $R \setminus \{x_i\}$ denotes the elements in R excluding x_i.

We determine the hyperparameter values of the serendipity recommender (α and λ) as follows. First, the PLATIPUS model was trained on 16 of the 19 amines in the historical

dataset. Next, for each value of α and λ within range [0, 1] with steps of 0.05, the recommender is applied to PLATIPUS predictions made on the remaining 3 amines in the historical data. For each combination of parameters, the algorithm recommends 20 candidate reactions. The experiments were performed in the lab and the following dual objective score is calculated for a particular set of recommendations R by

$$\mathrm{D}(R) = \frac{1}{2} \times \frac{\max(SF) - SF(R)}{\max(SF) - \min(SF)} + \frac{1}{2} \times \frac{\max(S) - S(R)}{\max(S) - \min(S)},$$

where SF(R) is the fraction of successful experiments and S(R) is the serendipity value of the set of recommendations R. The dual objective score is the mean of the normalized values of SF(R) and S(R). Fig. S4 shows the contour plot of the above metric of both success percentage and serendipity values, normalized separately for different values of α and λ. The highest dual objective score values were observed for $0.35 \leq \alpha \leq 0.6$ and $0 \leq \lambda \leq 0.1$; the final hyperparameter values selected for use in the study were $\alpha = 0.5$ and $\lambda = 0.1$.

Statistical Analysis. 95% confidence limits in Figures 3-5 are established using bootstrap calculations, as implemented in stats.bootstrap module of the SciPy version 1.7.3 Python library.

Volume fraction calculation. Convex hull of experiments is calculated using scipy.spatial.ConvexHull module with the three dimensions being organic amine, inorganic lead diiodide and formic acid concentrations. Volume fraction is the ratio of volume occupied by the convex hull of recommended points to the volume occupied by the convex hull of the stateset of all possible reactions for a specific amine.

References

1. J. Yano, K. J. Gaffney, J. Gregoire, L. Hung, A. Ourmazd, J. Schrier, J. A. Sethian, F. M. Toma, The case for data science in experimental chemistry: examples and recommendations. *Nat Rev Chem.* **6**, 357–370 (2022).

2. E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, C. P. Gomes, J. M. Gregoire, A. Mehta, J. Montoya, E. Olivetti, C. Park, E. Rotenberg, S. K. Saikin, S. Smullin, V. Stanev, B. Maruyama, Autonomous experimentation systems for materials development: A community perspective. *Matter*. **4**, 2702–2726 (2021).

3. J. H. Montoya, M. Aykol, A. Anapolsky, C. B. Gopal, P. K. Herring, J. S. Hummelshøj, L. Hung, H.-K. Kwon, D. Schweigert, S. Sun, S. K. Suram, S. B. Torrisi, A. Trewartha, B. D. Storey, Toward autonomous materials research: Recent progress and future challenges. *Appl. Phys. Rev.* **9**, 011405 (2022).

4. N. S. Eyke, B. A. Koscher, K. F. Jensen, Toward Machine Learning-Enhanced High-Throughput Experimentation. *Trends in Chemistry*. **3**, 120–132 (2021).

5. N. J. Szymanski, Y. Zeng, H. Huo, C. J. Bartel, H. Kim, G. Ceder, Toward autonomous design and synthesis of novel inorganic materials. *Mater. Horiz.* **8**, 2169–2198 (2021).

6. J. A. Bennett, M. Abolhasani, Autonomous chemical science and engineering enabled by self-driving laboratories. *Curr. Opin. Chem. Eng.* **36**, 100831 (2022).

7. A. L. Ferguson, K. A. Brown, Data-Driven Design and Autonomous Experimentation in Soft and Biological Materials Engineering. *Annu. Rev. Chem. Biomol. Eng.* (2022), doi:10.1146/annurev-chembioeng-092120-020803.

8. L. Cao, D. Russo, A. A. Lapkin, Automated robotic platforms in design and development of formulations. *AIChE Journal*. **67**, e17248 (2021).

9. R. K. Vasudevan, M. Ziatdinov, L. Vlcek, S. V. Kalinin, Off-the-shelf deep learning is not enough, and requires parsimony, Bayesianity, and causality. *npj Comput Mater*. **7**, 16 (2021).

10. P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, Machine-learning-assisted materials discovery using failed experiments. *Nature*. **533**, 73–76 (2016).

11. H. Hayashi, K. Hayashi, K. Kouzai, A. Seko, I. Tanaka, Recommender System of Successful Processing Conditions for New Compounds Based on a Parallel Experimental Data Set. *Chem. Mater.* **31**, 9984–9992 (2019).

12. J. Grizou, L. J. Points, A. Sharma, L. Cronin, A curious formulation robot enables the discovery of a novel protocell behavior. *Sci. Adv.* **6**, eaay4237 (2020).

13. K. Terayama, M. Sumita, R. Tamura, D. T. Payne, M. K. Chahal, S. Ishihara, K. Tsuda, Pushing property limits in materials discovery *via* boundless objective-free exploration. *Chem. Sci.* **11**, 5959–5968 (2020).

14. M. Kaminskas, D. Bridge, Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* **7**, 1–42 (2017).

15. R. J. Ziarani, R. Ravanmehr, Serendipity in Recommender Systems: A Systematic Literature Review. *J. Comput. Sci. Technol.* **36**, 375–396 (2021).

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

17. A. K. Jena, A. Kulkarni, T. Miyasaka, Halide Perovskite Photovoltaics: Background, Status, and Future Prospects. *Chem. Rev.* **119**, 3036–3103 (2019).

18. M. Ahmadi, M. Ziatdinov, Y. Zhou, E. A. Lass, S. V. Kalinin, Machine learning for high-throughput experimental exploration of metal halide perovskites. *Joule*. **5**, 2797–2822 (2021).

19. M. D. Smith, E. J. Crace, A. Jaffe, H. I. Karunadasa, The Diversity of Layered Halide Perovskites. *Annu. Rev. Mater. Res.* **48**, 111–136 (2018).

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

21. V. Shekar, G. Nicholas, M. A. Najeeb, M. Zeile, V. Yu, X. Wang, D. Slack, Z. Li, P. W. Nega, E. M. Chan, A. J. Norquist, J. Schrier, S. A. Friedler, Active meta-learning for predicting and selecting perovskite crystallization experiments. *J. Chem. Phys.* **156**, 064108 (2022).

22. J. Schrier, Solution Mixing Calculations as a Geometry, Linear Algebra, and Convex Analysis Problem. *J. Chem. Educ.* **98**, 1659–1666 (2021).

23. H. A. Chipman, E. I. George, R. E. McCulloch, BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4**, 266–298 (2010).

24. K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).

25. C. Finn, K. Xu, S. Levine, Probabilistic model-agnostic meta-learning. *Advances in Neural Information Processing Systems (NeurIPS 2018)*. **31** (2018) (available at https://proceedings.neurips.cc/paper/2018/file/8e2c381d4dd04f1c55093f22c59c3a08-Paper.pdf).

26. J. Snoek, H. Larochelle, R. P. Adams, Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*. **25**, 2951–2959 (2012).

27. B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, Bayesian reaction optimization as a tool for chemical synthesis. *Nature*. **590**, 89–96 (2021).

28. J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, Md. M. A. Patwary, P. Prabhat, R. P. Adams, Scalable Bayesian optimization using deep neural networks. *International conference on machine learning, PMLR*. **37**, 2171–2180 (2015).

29. F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch, A. Aspuru-Guzik, Gryffin: An algorithm for Bayesian optimization of categorical variables informed by expert knowledge. *Appl. Phys. Rev.* **8**, 031406 (2021).

30. M. Volpp, L. P. Fröhlich, K. Fischer, A. Doerr, S. Falkner, F. Hutter, C. Daniel, Meta-Learning Acquisition Functions for Transfer Learning in Bayesian Optimization. *International Conference on Learning Representations* (2020) (available at https://openreview.net/forum?id=ryeYpJSKwr).

31. E. Fix, J. Hodges, "Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties" (Technical Report ADA800276, University of California, Berkeley, 1951), (available at https://apps.dtic.mil/sti/citations/ADA800276).

32. T. Cover, P. Hart, Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory*. **13**, 21–27 (1967).

33. L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *Classification And Regression Trees* (Routledge, ed. 1, 2017; https://www.taylorfrancis.com/books/9781351460491).

34. F. Ricci, L. Rokach, B. Shapira, P. B. Kantor, Eds., *Recommender Systems Handbook* (Springer US, Boston, MA, 2011; http://link.springer.com/10.1007/978-0-387-85820-3).

35. H. S. Stein, Advancing data-driven chemistry by beating benchmarks. *Trends in Chemistry* (2022), doi:10.1016/j.trechm.2022.05.003.

36. B. Rohr, H. S. Stein, D. Guevarra, Y. Wang, J. A. Haber, M. Aykol, S. K. Suram, J. M. Gregoire, Benchmarking the acceleration of materials discovery by sequential learning. *Chemical Science*. **11**, 2696–2706 (2020).

37. Q. Liang, A. E. Gongora, Z. Ren, A. Tiihonen, Z. Liu, S. Sun, J. R. Deneault, D. Bash, F. Mekki-Berrada, S. A. Khan, K. Hippalgaonkar, B. Maruyama, K. A. Brown, J. Fisher III, T. Buonassisi, Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. *npj Comput Mater*. **7**, 188 (2021).

38. X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist, J. Schrier, Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature*. **573**, 251–255 (2019).

39. M. D. Ekstrand, J. T. Riedl, J. A. Konstan, Collaborative Filtering Recommender Systems. *Foundations and Trends in Human–Computer Interaction*. **4**, 81–173 (2011).

40. P. W. Nega, Z. Li, V. Ghosh, J. Thapa, S. Sun, N. T. P. Hartono, M. A. N. Nellikkal, A. J. Norquist, T. Buonassisi, E. M. Chan, J. Schrier, Using automated serendipity to discover how trace water promotes and inhibits lead halide perovskite crystal formation. *Appl. Phys. Lett.* **119**, 041903 (2021).

41. I. Levchuk, Y. Hou, M. Gruber, M. Brandl, P. Herre, X. Tang, F. Hoegl, M. Batentschuk, A. Osvet, R. Hock, W. Peukert, R. R. Tykwinski, C. J. Brabec, Deciphering the Role of

Impurities in Methylammonium Iodide and Their Impact on the Performance of Perovskite Solar Cells. *Advanced Materials Interfaces*. **3**, 1600593 (2016).

42. M. Jung, S.-G. Ji, G. Kim, S. I. Seok, Perovskite precursor solution chemistry: from fundamentals to photovoltaic applications. *Chem. Soc. Rev.* **48**, 2011–2038 (2019).

43. J. T. Bradshaw, K. J. Albert, in *Practical Approaches to Method Validation and Essential Instrument Qualification* (John Wiley & Sons Inc., 2010), pp. 347–376.

44. E. C. Lee, J. M. Parrilla-Gutierrez, A. Henson, E. K. Brechin, L. Cronin, A Crystallization Robot for Generating True Random Numbers Based on Stochastic Chemical Processes. *Matter*. **2**, 649–657 (2020).

45. S. G. Baird, T. Q. Diep, T. D. Sparks, *Digital Discovery*, in press, doi:10.1039/D1DD00028D.

46. L. Zhang, The Definition of Novelty in Recommendation System. *J. Eng. Sci. Technol. Rev.* **6**, 141–145 (2013).

47. R. K. Merton, E. G. Barber, *The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology of Science* (Princeton Univ. Press, 2006).

48. Y. Shuo, H. D. Bedru, C. Xinbei, Y. Yuyuan, W. Liangtian, X. Feng, Understanding Serendipity in Science: A Survey. *Data Analysis and Knowledge Discovery*. **5**, 16–35 (2021).

49. O. Yaqub, Serendipity: Towards a taxonomy and a theory. *Research Policy*. **47**, 169–179 (2018).

50. A. McNally, C. K. Prier, D. W. C. MacMillan, Discovery of an α-Amino C–H Arylation Reaction Using the Strategy of Accelerated Serendipity. *Science*. **334**, 1114–1117 (2011).

51. D. H. Wolpert, The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*. **8**, 1341–1390 (1996).

52. I. M. Pendleton, G. Cattabriga, Z. Li, M. A. Najeeb, S. A. Friedler, A. J. Norquist, E. M. Chan, J. Schrier, Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): a software pipeline for automated chemical experimentation and data management. *MRS Communications*. **9**, 846–859 (2019).

53. I. Wallach, A. Heifets, Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* **58**, 916–932 (2018).

54. C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks. *Proceedings of the 34th International Conference on Machine Learning*. **70**, 1126–1135 (2017).

55. N. Hansen, Yoshihikoueno, ARF1, K. Nozawa, L. Rolshoven, M. Chan, Youhei Akimoto, Brieglhostis, D. Brockhoff, *CMA-ES/pycma: r3.2.2* (2022; https://zenodo.org/record/2559634).

56. F. Häse, L. M. Roch, C. Kreisbeck, A. Aspuru-Guzik, Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **4**, 1134–1145 (2018).

Acknowledgments

Funding:

Defense Advanced Research Projects Agency (DARPA) contract HR001118C0036 (VS, VY, MAN, MZ, PWN, ZL, MK, EMC, AJN, SF, JS)

Defense Advanced Research Projects Agency (DARPA) contract FA8750-17-C-0229 (BJG, DBG)

Defense Advanced Research Projects Agency (DARPA) contract FA8750-18-C-0130 (GM, DMB)

Office of Basic Energy Sciences, of the U.S. Department of Energy Contract No. DE-AC02-05CH11231. (PWN, ZL, MK, EMC)

Henry Dreyfus Teacher-Scholar Award TH-14-010 (JS)

Author contributions:

Conceptualization: JS, AJN, SF
Methodology: VS, VY, GEM, BJG, LMR, AGD
Investigation: MAN, MZ, PWN, ZL, MK
Visualization: VS
Supervision: JS, SF, EMC, AJN, DBG, DMB
Writing—original draft: VS, JS, SF
Writing—review & editing: AJN, EMC, DBG, VY, GEM, BJG

Competing interests: LR is the co-founder, AGD is an employee, and AJN and JS are on the scientific advisory board of Atinary Technologies Inc., which developed and commercializes the Falcon suite of machine learning algorithms. LR is co-author of the Gryffin algorithm. Falcon and Gryffin used in this paper are included in the Atinary SDLabs platform. Atinary ran the GPTAF model and collected the associated results. All other authors declare they have no competing interests.

Data and materials availability: Complete data and code for this study are available at https://github.com/darkreactions/serendipity_recommender (we will archive this on Zenodo upon acceptance) except for the Atinary models, which are available in the cloud-based SDLabs platform at **home.atinary.com**

Figures and Tables

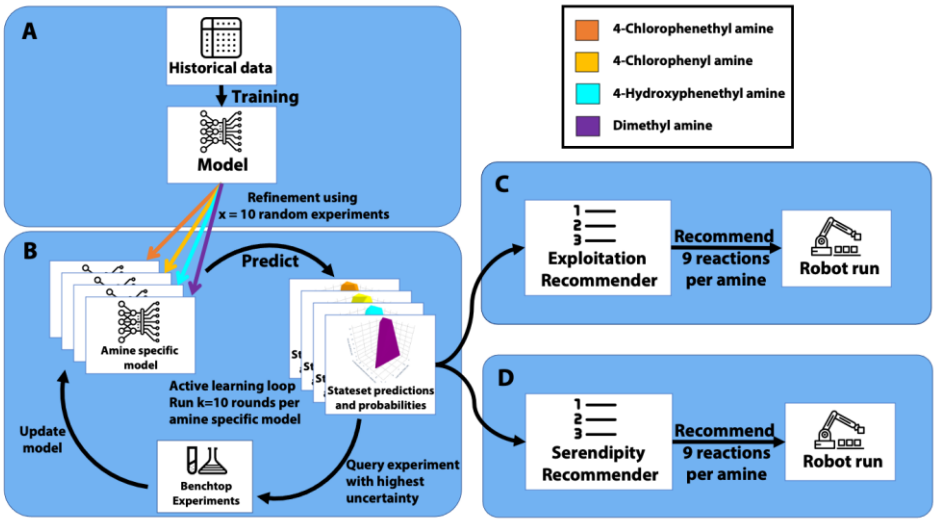

Fig. 1. Overview of recommendation systems with active learning (A) Model are provided with historical data for training, and refinement and initialized with 10 random experiments for each new chemical system to explore. **(B)** Ten active learning rounds enable amine-specific models to be further refined by requesting experiments from a stateset of possible experiments and performing them in the lab. Success probabilities per model are calculated for the stateset and serve as inputs to recommendation systems. **(C)** The exploitative recommender maximizes success probabilities predicted by a model. **(D)** The serendipity recommender maximizes the serendipity measure which attempts to increase diversity in experimental conditions while keeping the success probabilities high.

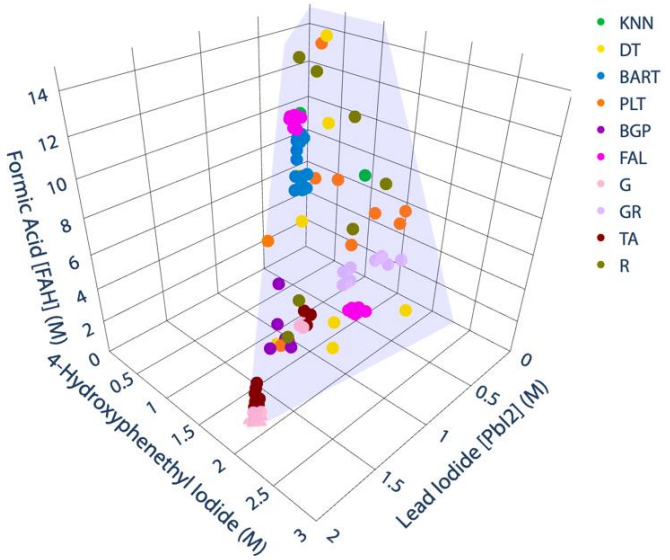

Fig. 2. Convex hull of experiments for 4-Hydroxyphenethyl amine. The blue polygon represents the stateset of all achievable concentrations due to volume and solubility limits. Points represent successful experiments predicted by different models. Models such as BART, BGP, FAL, FALGP and GRYF, are clustered together in chemical space. The exploitation recommender optimizes for experiments with the highest probability of success which tend to have similar experimental conditions to those already sampled. Each model mentioned above has two clusters of points corresponding to the initialization experiments used to train the model. These models explore a small fraction of the chemical space and cluster in different parts of the space due to the sensitivity to initialization samples

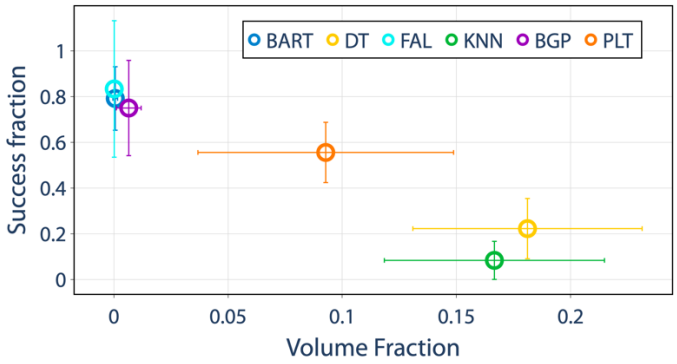

Fig 3. Convex hull volume fraction versus success fraction for the exploitation recommendations. The figure shows fraction of convex hull volume explored by each model versus the crystallization success rate averaged over all amines, with error bars showing the 95% confidence interval. The best performing models, FAL, BART, and BGP models do not have a large convex hull volume formed by the recommended reactions, implying selected experiments are close together in chemical space

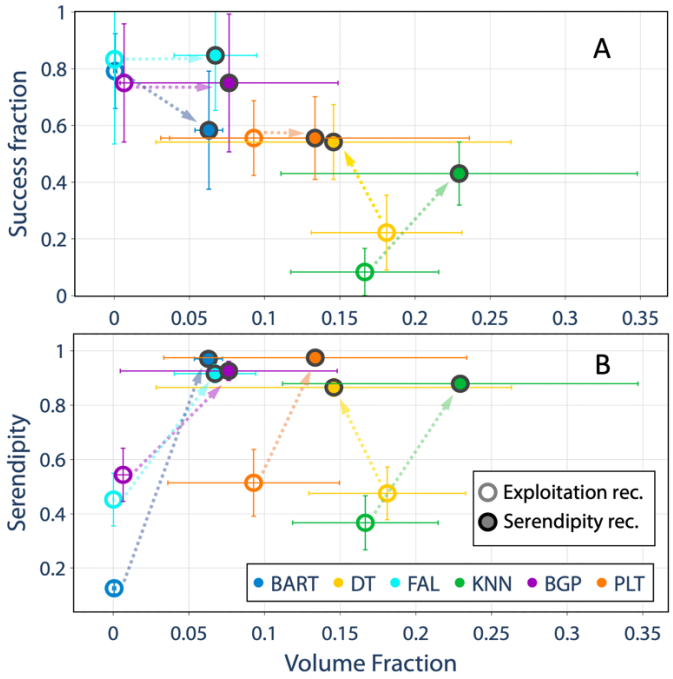

Fig 4. Change in model performance due to serendipity recommender. Error bars show 95% confidence interval over 72 recommended experiments. **(A)** Fraction of chemical space explored (volume fraction) per model averaged over all amines vs. success fraction for exploitation recommender experiment and serendipity plates. All models except for DT move towards the right indicating the recommendations made by the serendipity recommender explore more of the chemical space with very little change in success fraction for the best performing models **(B)** Average volume fraction per model over all amines vs. average serendipity per model over all amines for original experiment and serendipity plates. Serendipity values increase across all models indicating a greater variety of experiments selected by the recommender

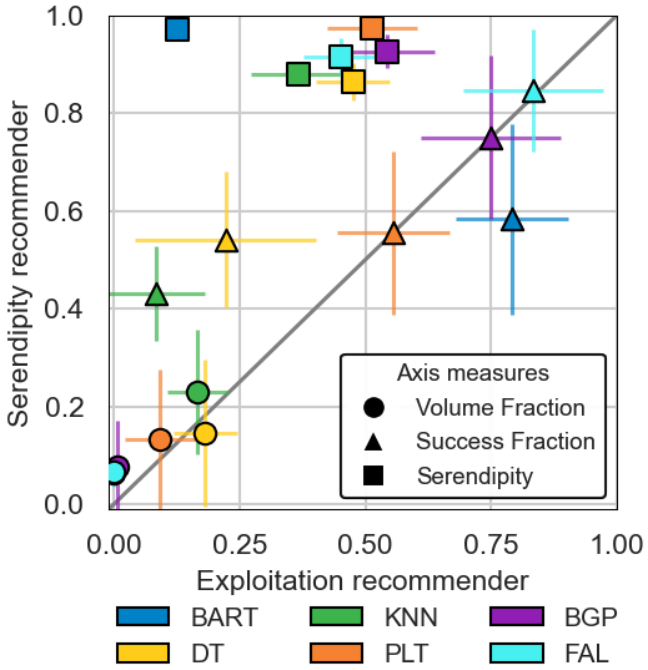

Fig. 5: Overview of recommender measures. Volume fraction, success fraction and serendipity measure comparisons indicated by circle, triangle, and square points, respectively between the serendipity recommendations and exploitation recommender with error bars showing 95% confidence interval. Points above unit slope indicate higher values for the serendipity recommender as compared to the exploitation recommender. Serendipity measures and volume fractions both increase for all models except for DT by applying the serendipity recommender, indicating an increase chemical space explored while maintaining success fractions on average.

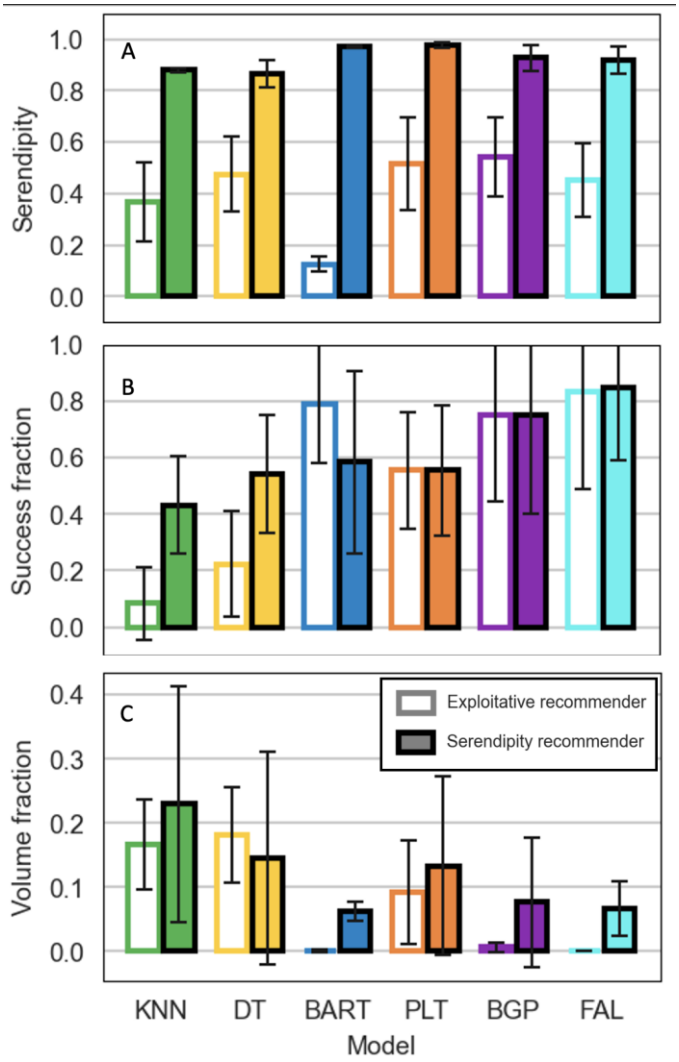

Fig 6: Side by side comparison of measures per model. Each bar show an average over two trials of 4 amines, and error bars show one standard deviation. **(A)** Serendipity measure of recommendations made by models. **(B)** Success fraction of recommendations made by models. **(C)** Fraction of convex hull volume occupied by recommendations of each model.

Table 1. Models used and exploitation recommender results. Table shows the model codes used throughout this paper in the first column and the corresponding model name in the second column. Columns that follow show results for the exploitation recommender, starting with "Successful initialization draws" which is the number of initialization experiment draws that yield at least one success. "Successful for all amines and draws" indicates whether the model predicts at least one success for all amines and draws. Followed by the fraction of experiments correctly predicted and the fraction of the convex hull covered by the recommended experiments

Model Code	Model Name	Successful initialization draws	Successful for all amines and draws	Fraction Success	Convex hull volume fraction
KNN	K Nearest Neighbors	3		0.08	0.166
DT	Decision Tree	8	✓	0.22	0.181
BART	Bayesian Additive Regression Trees with Transfer Learning (BART+TL)	8	✓	0.79	62.7×10^{-5}
PLT	PLATIPUS	8	✓	0.56	0.0927
BGP	Bayesian Optimization using Gaussian Processes	8	✓	0.75	0.00653
FAL	Atinary Falcon	7		0.83	16.9×10^{-5}
G	Falcon GPBO	7		0.74	23.4×10^{-5}
D	Falcon DNGO	5		0.28	8.78×10^{-5}
GR	Gryffin	4		0.39	9.87×10^{-5}
TA	Gaussian Processes with Transfer Acquisition Functions	6		0.67	6.9×10^{-5}
FH	Falcon with Historical data	8	✓	0.47	0.0598
R	Uniform random selection	8	✓	0.36	0.135

Table 2. Models used and serendipity recommender results. Table shows model code followed by the corresponding success fraction, fraction of the convex hull occupied by recommended experiments and the serendipity values for both the serendipity recommender and the exploitation recommender

Model Code	Fraction Success	Convex hull volume fraction	Serendipity value	
			Serendipity rec.	Exploitation rec.
KNN	0.43	0.23	0.88	0.37
DT	0.54	0.14	0.87	0.47
BART	0.58	0.063	0.97	0.12
PLT	0.56	0.133	0.98	0.51
BGP	0.75	0.077	0.93	0.54
FAL	0.84	0.067	0.92	0.45