The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Assessing the frontier: Active learning, model accuracy, and multi-objective candidate discovery and optimization

Cite as: J. Chem. Phys. 153, 024112 (2020); doi: 10.1063/5.0006124
Submitted: 2 March 2020 • Accepted: 22 June 2020 •
Published Online: 9 July 2020

Zachary del Rosario,[1,a)] Matthias Rupp,[2,b)] Yoolhee Kim,[2,c)] Erin Antono,[2,d)] and Julia Ling[2,e)]

AFFILIATIONS

[1] Olin College of Engineering, 1000 Olin Way, Needham, Massachusetts 02492, USA
[2] Citrine Informatics, 2629 Broadway, Redwood City, California 94063, USA

Note: This paper is part of the JCP Special Topic on Machine Learning Meets Chemical Physics.
[a)] **Author to whom correspondence should be addressed:** zdelrosario@olin.edu
[b)] **Electronic mail:** rupp@mrupp.info
[c)] **Electronic mail:** ykim@citrine.io
[d)] **Electronic mail:** erin@citrine.io
[e)] **Electronic mail:** jling@citrine.io

ABSTRACT

Discovering novel chemicals and materials can be greatly accelerated by iterative machine learning-informed proposal of candidates—active learning. However, standard *global error* metrics for model quality are not predictive of discovery performance and can be misleading. We introduce the notion of *Pareto shell error* to help judge the suitability of a model for proposing candidates. Furthermore, through synthetic cases, an experimental thermoelectric dataset and a computational organic molecule dataset, we probe the relation between acquisition function fidelity and active learning performance. Results suggest novel diagnostic tools, as well as new insights for the acquisition function design.

I. INTRODUCTION

Accelerated design, optimization, and tuning of chemicals and materials via machine learning is receiving increasing interest in science and industry. A major driver of this interest is the potential to reduce the substantial cost and effort involved in manual development, synthesis, and characterization of large numbers of candidates. The primary aim is to reduce the number of both failed candidates and development cycles.

A data-driven approach to achieve this acceleration is *active learning* (AL),[1] an iterative procedure in which a machine-learning model suggests candidates, a selection of which are synthesized, characterized, and fed back into the model to complete a learning iteration. The objective of this procedure varies; in chem- and materials-informatics, the objective is often to identify promising candidates by optimizing properties of interest. Recent work has leveraged AL for candidates when there is a *single objective*.[2–17] However, new issues arise when optimizing multiple objectives simultaneously, as is frequently the case beyond proof-of-principle settings.

Furthermore, the aim to identify promising candidates is distinct from the aim to improve model accuracy, a frequent objective in AL.[18] In fact, model accuracy can be at odds with acquiring optimal candidates: Figure 1 demonstrates that the usual global notion of model accuracy is not necessarily associated with optimal chemicals and materials discovery. In this work, we introduce a notion of model accuracy more closely associated with rapidly discovering new candidates.

We offer two primary contributions: First, we introduce novel concepts to judge the performance of multi-objective AL, with

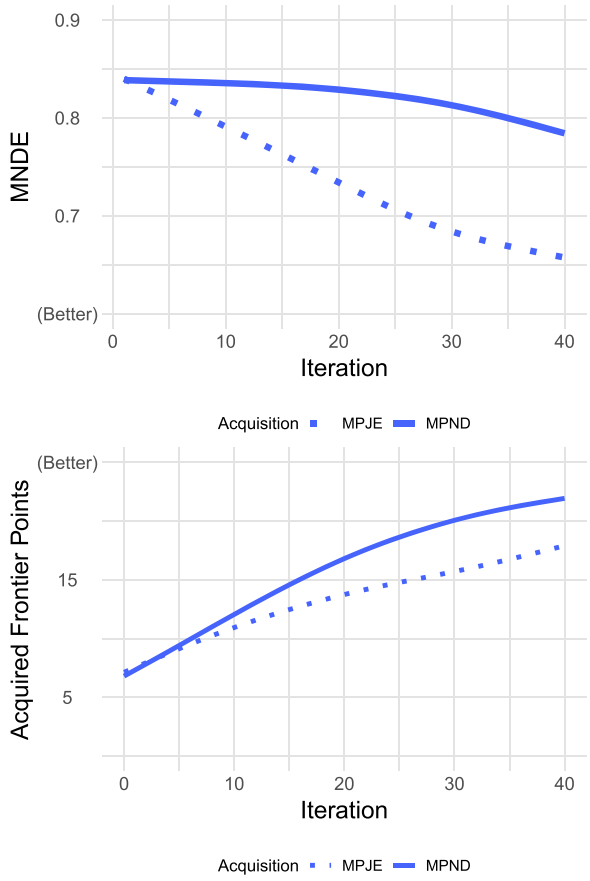

FIG. 1. *Best global error does not guarantee optimal candidate discovery.* Shown are global prediction errors (MNDE, top) and number of optimal candidates found (bottom). The MPJE decision criterion (dashed line) has lower errors but leads to fewer optimal candidates than the MPND criterion (solid line). Simulated active learning on a thermoelectric dataset. Estimated mean performance over repetitions (Sec. IV). MNDE = Mean Non-Dimensional Error (7), MPJE = Maximum Probability of Joint Exceedance (11), and MPND = Maximum Probability Non-Dominated (13).

an aim toward the specific concerns of candidate discovery. In addition to the usual notion of multi-objective optimality[19]—non-dominance—we use criteria informed by the concept of *strata*—recursive non-dominance—from the database literature.[20] We also introduce *scoped* error metrics, which emphasize regions of interest in performance (property) space, particularly bands about the Pareto frontier we call *Pareto shells*. We demonstrate that the usual global error provides less usable signal for candidate discovery performance, while Pareto shell error—under certain conditions—correctly signals when AL will likely identify performant candidates. Second, we compare multi-objective acquisition functions (also known as *improvement criteria*) in terms of their quantitative and qualitative performance. Specifically, we compare a collection of exploitation/exploration-navigating acquisition functions: Probability of Joint Exceedance (PJE), Hyperplane Probability of Improvement (HPI), and Probability Non-Dominated (PND). These acquisition functions differ in terms of the fidelity with which they represent the Pareto frontier, allowing us to study how this affects AL performance.

In this work, we review AL for single-objective candidate discovery and relevant concepts from multi-objective optimization and synthesize scientifically relevant concepts for judging AL performance. We introduce a family of acquisition functions (decision-making strategies) and compare their performance across several synthetic datasets and a real thermoelectric dataset. Our results illustrate how new concepts can help articulate the difference between AL strategies and how model accuracy does—and does not—relate to AL performance in discovering novel candidates.

II. ACTIVE LEARNING FOR SINGLE-OBJECTIVE CANDIDATE DISCOVERY

AL is a specialized problem setting in machine learning related to the optimal experimental design.[1,11,21] In the chemical and materials science context, consider a description $\boldsymbol{x}$ of a candidate with corresponding observed property of interest y, thought to be linked by an unknown function f: $\boldsymbol{x} \mapsto y$, which is expensive to evaluate, for example, synthesizing and characterizing a candidate. To systematically identify novel candidates $\boldsymbol{x}$ with desirable changes in y, a statistical (machine-learning) model $\hat{f}$ is built that predicts the unknown function f. Trained on an initial set of characterized $\tilde{\mathscr{X}}_0 = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_0}\}$ with measured properties $\tilde{\mathscr{Y}}_0 = \{y_1, \ldots, y_{n_0}\}$, the initial model $\hat{f}_0$ is used to select new candidates $\boldsymbol{x}_{n_0+1}, \ldots, \boldsymbol{x}_{n_0+n_1}$. These are characterized and added to the dataset, $\tilde{\mathscr{X}}_1 = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_0+n_1}\}$, which is then used to train a new improved model $\hat{f}_1$. This results in a sequence of datasets $\tilde{\mathscr{X}}_k, \tilde{\mathscr{Y}}_k$ for $k = 0, \ldots, K$. While this cycle can, in principle, be fully automated, new candidates are usually selected by domain experts based on AL suggestions ("human-in-the-loop").

Various objectives can drive the design of AL strategies; for instance, a common objective in general machine learning is to improve the model $\hat{f}$. In informatics, the objective is often to identify an improved material $\boldsymbol{x}$ using as few physical experiments as possible.[6] We call the design space of experimentally accessible, that is, synthesizable and measurable, candidates the *global scope* and call the error computed on this scope *global error*.

For a *single objective*, measuring the relative performance of candidates is straightforward (Fig. 2) as they can be ranked unambiguously based on their scalar performance, for example, the strength-to-weight ratio for structural alloys. One way to measure the performance of AL is to retrospectively simulate AL on a known dataset and count the number of AL iterations necessary to reach a known top candidate.[6] A key decision in designing an AL method is the choice of *acquisition function*. We will discuss these in greater detail in Sec. III; briefly, an acquisition function is a decision rule used to rank potential candidates in an AL context. Such criteria navigate the "exploration-exploitation" trade-off:[22] The algorithm should seek improved candidates but should also try "risky" candidates to improve its model and enable later discoveries.[18] Many improvement criteria for the single-objective case have been proposed and tested in the literature.[2,6,14–16,23] The multi-objective setting, however, requires a more nuanced understanding of candidate ranking.

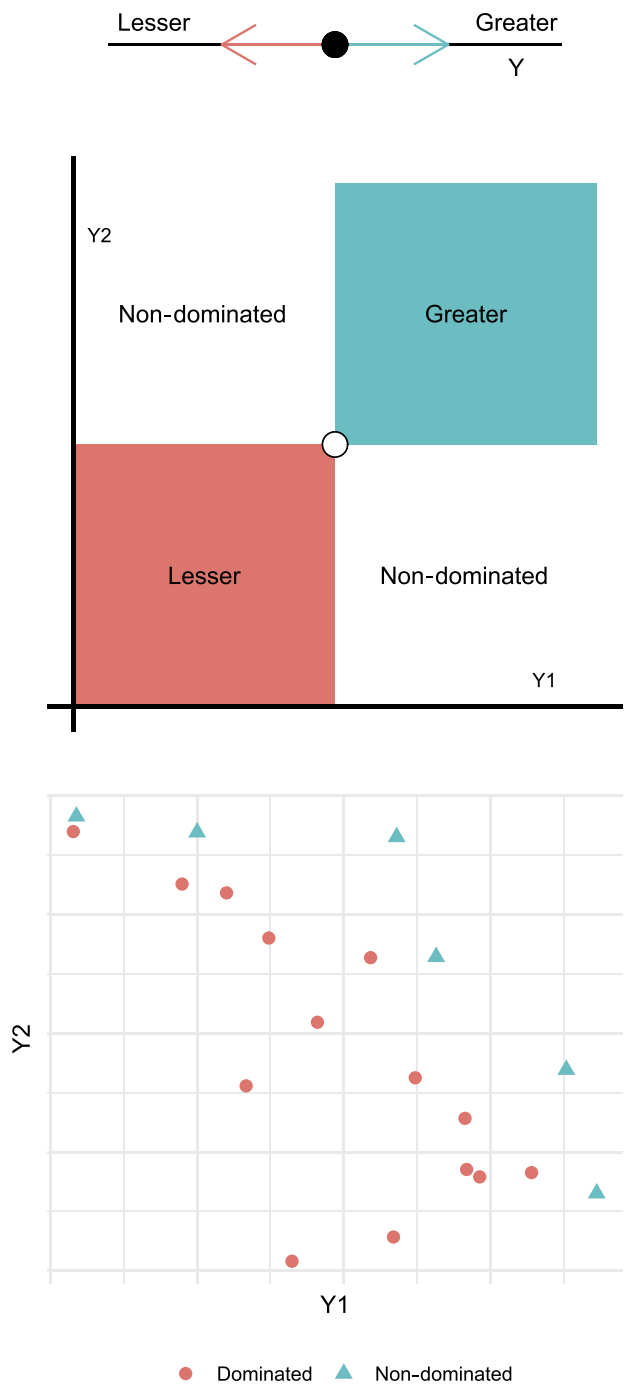

FIG. 2. *Multi-objective optimization requires ranking concepts beyond "greater" and "lesser."* Illustration of candidate ranking settings (top) and an example multi-objective frontier (bottom). In the single-objective setting, a relative ranking between candidates is always possible via the total ordering induced by the single objective. However, in the multi-objective setting, two candidates can be neither greater nor lesser than the other if they are alternately dominant along different objectives—two distinct candidates can be *non-dominated*. Recognizing this issue leads to the notion of a *non-dominated set* of mutually incomparable candidates—the Pareto frontier (see Fig. 12.1 of Ref. 19).

III. METHODS

A. Multi-objective optimization background

In multi-objective optimization, the categories of "greater" and "lesser" are insufficient (Fig. 2). Since two candidates can compete along multiple axes, it is possible for them to be mutually *non-dominated*. Multiple objectives occur naturally in chemical and materials science problems and are often unavoidable, e.g., the strength-toughness trade-off, which arises from fundamental, competing effects.[24] *In lieu* of more preference information, one must navigate the resulting multi-objective trade-off space. Having access to more candidates in this trade-off space enables more complete scientific understanding and more informed engineering decisions. In this section, we review concepts from multi-objective optimization and introduce acquisition functions for the multi-objective setting.

1. Dominance and strata

In the single-objective case, candidates $y, y' \in \mathbb{R}$ can be unambiguously ranked: A candidate y is either lesser $y < y'$, greater $y > y'$, or equal $y = y'$ to another candidate y'. However, the introduction of multiple objectives $\boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^D$ introduces new complexities. Figure 2 illustrates the universe of possible comparisons; in addition to lesser, greater, and equal, in the multi-objective setting, there exist *non-dominated* points[19] (Chap. 12). *Dominance* (in the Pareto sense) is a pairwise relationship; the candidate $\boldsymbol{y}$ is said to *dominate* $\boldsymbol{y}'$ if

$$\begin{aligned} &y_d \geq y'_d \text{ for all } d \in \{1, \ldots, D\} \\ &\text{and } y_d > y'_d \text{ for some } d. \end{aligned} \tag{1}$$

We use the notation $\boldsymbol{y}' \prec \boldsymbol{y}$ to denote that $\boldsymbol{y}$ dominates $\boldsymbol{y}'$. Note that the definitions above pre-suppose that optimization is posed in terms of maximization of all objectives. This is not a limitation—if minimization is desired for a given axis y_d, then one must simply reverse the relevant inequalities in definition 1. Furthermore, proximity to a desired value can be encoded as minimization of the absolute distance from the target value, e.g., bandgap as close as possible to 1.2 eV.

In the multi-objective setting, the possibility of non-dominance implies that a single "best" multi-objective output value $\boldsymbol{y}$ may not exist. Instead, there may be a set of "best values." Given a set of candidates $\mathscr{A} \subseteq \mathbb{R}^D$, the set of non-dominated points is called the *Pareto frontier*, defined by

$$\mathscr{P}(\mathscr{A}) = \{\boldsymbol{y} \in \mathscr{A} \mid \forall \boldsymbol{y}' \in \mathscr{A}, \boldsymbol{y} \not\prec \boldsymbol{y}'\}. \tag{2}$$

The Pareto frontier $\mathscr{P}(\mathscr{A})$ represents the set of trade-offs one must navigate in choosing an optimization candidate. Further selection must be made through external considerations: possibly a prioritization of the objectives or through harder-to-quantify concerns such as the corrosion resistance of a material[25] (Chap. 5).

While the Pareto frontier is an important set of candidates, points outside the frontier are not without utility, particularly if aforementioned external concerns exist. Candidates near the frontier can be useful as training data for a machine learning model, while measurement or model uncertainties may lead to the false classification of a point as dominated. To describe points near the Pareto frontier, we use the notion of *strata*.

The *strata* are defined via a recursive relationship.[20] Let $\mathscr{A}$ be a set of candidates as above and define the s-th stratum $\mathscr{S}_s$ via

$$\begin{aligned} \mathscr{S}_1 &= \mathscr{P}(\mathscr{A}), \\ \mathscr{S}_s &= \mathscr{P}(\mathscr{A} - \mathscr{S}_{s-1}) \text{ for } s = 2, \ldots. \end{aligned} \tag{3}$$

Figure 3 illustrates a few strata on a thermoelectric dataset (introduced in Subsection IV A). Note that by definition, we have $\mathscr{S}_i \cap \mathscr{S}_j = \varnothing$ if $i \neq j$. This allows us to define a *stratum number* for each point $\boldsymbol{y} \in \mathscr{A}$ via

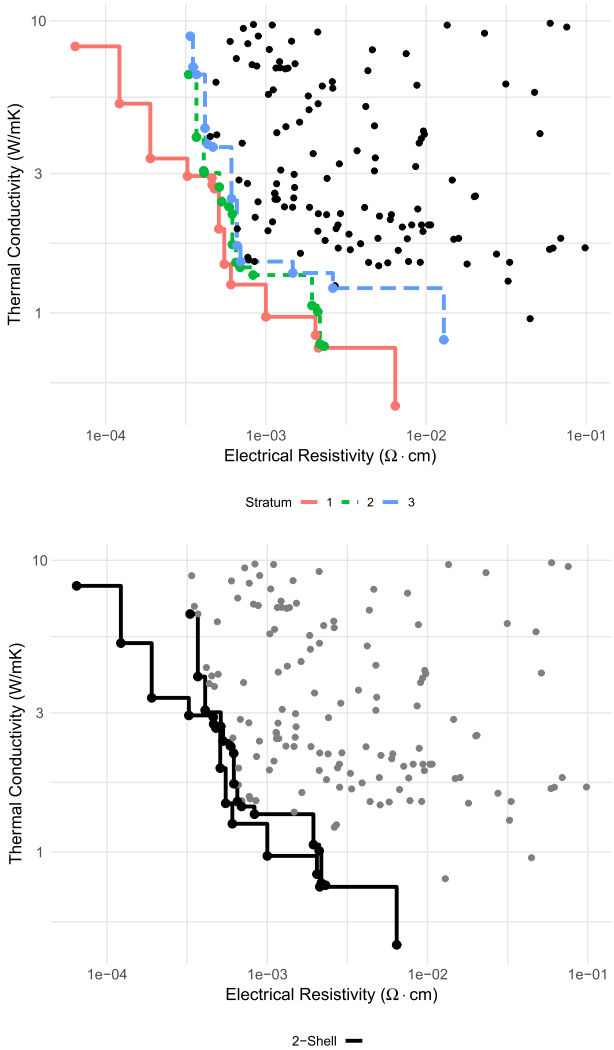

FIG. 3. *A recursive generalization of non-dominance (strata) enables a definition of scope tailored for optimal material candidates.* Example of strata (left) and a Pareto 2-shell (right) for the thermoelectric dataset, considering only ρ, κ for preference ranking (both minimized). The first stratum corresponds to the Pareto frontier, while higher strata are the Pareto frontier of the previous frontier's remainder [Definition (5)]. Considering all strata up to a number s yields the Pareto s-shell. On the right, we visualize an example 2-shell, which consists of all the points in the first two strata. Below, we use the concept of Pareto shells to define an error scope relevant to AL for candidate discovery.

$$s(\boldsymbol{y}) = s \text{ if } \boldsymbol{y} \in \mathscr{S}_s. \tag{4}$$

The candidates along the Pareto frontier then have $s(\boldsymbol{y}) = 1$, while points with a larger stratum number lie further from it. We will use the stratum number to rank the performance of AL with greater resolution than counting frontier points alone.

We also define the *Pareto s-shell* via

$$\mathscr{P}_s = \bigcup_{j=1}^{s} \mathscr{S}_j. \tag{5}$$

This definition allows one to select a "band" of points along and near the Pareto frontier. Below, we will use the Pareto shell $\mathscr{P}_s$ as a targeted scope for computing model accuracy. We use the nomenclature s-shell to denote $\mathscr{P}_s$: Figure 3 illustrates a Pareto 2-shell for the thermoelectric dataset.

B. Proposed performance measures

1. Candidate improvement performance

Prior work assessing AL in informatics has focused primarily on the number of Pareto frontier points acquired during AL.[6,26] While relevant, this "all or nothing" measurement of success provides no accounting for candidates near the Pareto frontier, which may still be of scientific interest. This is particularly troublesome when studying datasets that have very few Pareto frontier points, such as the "sparse" frontier we will consider below. To provide a measure with more granularity, we consider the *mean stratum number* at each iteration of AL $\bar{\mathscr{Y}}_k$ using the ground truth strata from the full dataset.

2. Model accuracy performance

In addition to the candidate improvement performance of AL, we also consider trajectories of model performance. We use the notion of *non-dimensional error* (NDE) across each of the output quantities; given a set of true $\{\boldsymbol{y}_i\}_{i=1}^n = \mathscr{Y}$ and estimated $\{\hat{\boldsymbol{y}}_i\}_{i=1}^n = \hat{\mathscr{Y}}$ response values, we compute the NDE for the output d via

$$\text{NDE}_d = \sqrt{\frac{\sum_{i=1}^n (y_{i,d} - \hat{y}_{i,d})^2}{\sum_{i=1}^n (y_{i,d} - \bar{y}_d)^2}}, \tag{6}$$

where $\bar{y}_d = \frac{1}{n}\sum_{i=1}^n y_{i,d}$ is the sample mean. If we have D output quantities, then there are D NDE values to compute. Note that the NDE is closely related to the *coefficient of determination* R^2 via the expression $\text{NDE} = \sqrt{1 - R^2}$.[27] Since the NDE is dimensionless, we may safely average NDE values across the D outputs to compute a *mean non-dimensional error* (MNDE), given by

$$\text{MNDE} = \frac{1}{D}\sum_{d=1}^{D} \text{NDE}_d. \tag{7}$$

Given the retrospective nature of our test cases, the global error is naturally evaluated by (6) across the entire dataset. However, this *global scope* includes regions of output space that are not emphasized in our frontier-seeking context. To compute error metrics with *targeted scope*, we use the notion of a Pareto shell to compute error on the relevant portions of the output domain (Fig. 3). This notion of scope is closely related to the concept of a *domain of applicability*

(input space), with the contrast that the scope is defined with respect to the range (output space).[28]

C. Acquisition functions for benchmarking

To assess the performance measures above, we must perform active learning with some form of decision criteria that summarizes available multi-objective property predictions and automatically ranks candidates. To this end, we define an *acquisition function* $f_a(\boldsymbol{x}; \mathscr{M})$ as any function that is used to select an optimal candidate via

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \tilde{\mathscr{X}}^C}{\arg\max} f_a(\boldsymbol{x}; \mathscr{M}), \tag{8}$$

where $\tilde{\mathscr{X}}^C$ is the complement of the training set and $\mathscr{M}$ is a trained machine learning model, which returns both a prediction $\hat{\boldsymbol{y}}$ and a predictive distribution $\boldsymbol{Y} \sim \hat{\mathscr{D}}$, both of which are available to the acquisition function f_a.

Many generalizations of single-objective acquisition functions are available from the literature, including the probability of improvement and expected improvement criteria,[29] the max–min criterion,[30] and the expected hyper-volume improvement (EHVI) criterion.[26] Comparing these and other criteria is outside the scope of this work; instead, we are focused on (i) how the fidelity with which the Pareto frontier is represented relates to AL performance and (ii) how optimal candidate selection is (or is not) related to model accuracy. Below, we introduce the acquisition functions to be studied in this work after a remark on dimensional homogeneity.

1. Importance of dimensional homogeneity

Before introducing specific acquisition functions, we first make a remark on the importance of *dimensional homogeneity*. In order to measure the performance of multi-objective AL, we may wish to measure the "distance" of candidates to the Pareto frontier. However, we will see here that a naive notion of distance is problematic as it does not respect dimensional homogeneity—the ranking results are not independent of the analyst's choice of the unit system.

To illustrate, let $\boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^D$, where the y_d potentially have different physical units. The ordinary notion of distance is given by

$$\|\boldsymbol{y} - \boldsymbol{y}'\|_p = \left(\sum_{d=1}^{D} |y_d - y'_d|^p \right)^{1/p}. \tag{9}$$

Note that for all $p \in (0, +\infty)$, this expression involves the addition of terms of potentially different physical units. This expression violates dimensional homogeneity, which introduces an artificial dependence on the chosen unit system. Thus, the ranks computed by distances are not necessarily stable to a change of the unit system—we provide an example of this pathology in the supplementary material. This illustrates that arbitrary choices can drastically affect decisions based on this sort of distance computation. To overcome this issue, we only consider acquisition functions that respect dimensional homogeneity.

2. Uncertainty sampling

On the scale from exploitation to exploration, *uncertainty sampling* leans heavily toward the latter; one chooses candidates based on where the model is most uncertain. In the scalar case, this is easily accomplished by choosing the candidate with the greatest predictive variance $\hat{\sigma}_i^2$.[18] This approach does not immediately generalize to the multi-objective case as the component variances $\hat{\sigma}_{i,d}^2 = \hat{\boldsymbol{\Sigma}}_{i,dd}$ do not necessarily have the same units. To respect dimensional homogeneity, we generalize uncertainty sampling by considering a sum of dimensionless quantities.

Sum of Coefficients of Variation (SCV): The SCV is defined via

$$f_{\mathrm{SCV}}(\boldsymbol{x}_i) = \sum_{d=1}^{D} \mathrm{COV}_{i,d}, \tag{10}$$

where the $\mathrm{COV}_{i,d} = \sigma_{i,d}/|\mu_{i,d}|$ are coefficients of variation and $\hat{\mu}_{i,d}, \hat{\sigma}_{i,d}$ are the $d = 1, \ldots, D$ components of the mean and standard deviation of the predictive distribution $\hat{\mathscr{D}}_i$. Note that this definition is problematic if any of the μ_i are exactly zero. However, the coefficients of variation $\mathrm{COV}_{i,d}$ are dimensionless quantities with a normalizing scale $\hat{\mu}_{i,d}$ set not by the user, but rather by the available data.

3. Frontier modeling strategies

Here, we introduce a family of acquisition functions that seek improvement over an existing Pareto frontier, in the order of increasing fidelity with which they model the Pareto frontier. Each of these strategies is a form of probability statement; by construction, these quantities respect dimensional homogeneity.

Probability of Joint Exceedance (PJE): The PJE is defined via

$$f_{\mathrm{PJE}}(\boldsymbol{x}_i) = \mathbb{P}_{\hat{\mathscr{D}}_i}[Y_{i,d} > \max_{\boldsymbol{y} \in \tilde{\mathscr{Y}}} y_d], \tag{11}$$

where $\boldsymbol{Y}_i \sim \hat{\mathscr{D}}_i$ is the random variable, which follows the predictive distribution $\hat{\mathscr{D}}_i$ for candidate i. In words, definition (11) is the probability that candidate $\boldsymbol{Y}_i$ will exceed the performance observed in our existing training data $\tilde{\mathscr{Y}}$ along every axis of comparison. This is a very "aggressive" acquisition function that ignores much of the structure of the Pareto frontier. Figure 4 illustrates the PJE, alongside the other frontier modeling criteria (Fig. 5).

Hyperplane Probability of Improvement (HPI): The HPI is defined in terms of a hyperplane fit of the Pareto frontier in the available training data. Modeling the Pareto frontier as a hyperplane requires fitting a normal direction $\hat{\boldsymbol{w}} \in \mathbb{R}^D$ and an offset $\hat{b}$. Once these are fit, the HPI is defined in terms of the appropriate Z-score via

$$f_{\mathrm{HPI}}(\boldsymbol{x}_i) = \frac{\hat{\boldsymbol{w}}^\top \hat{\mu}_i - \hat{b}}{\sqrt{\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{\Sigma}}_i \hat{\boldsymbol{w}}}}. \tag{12}$$

In words, the HPI is the probability that a candidate $\boldsymbol{Y}_i$ will present improvement over the hyperplane fit of the existing Pareto frontier.

There are many ways to fit a hyperplane to a set of data. One option is to arbitrarily select one output Y_i and perform linear regression using the remaining outputs as regression variables. We recommend against this approach as it suffers from the *regression fallacy*[31] (Chap. 9.1). In practice, we perform a principal component analysis of the available Pareto frontier data and use the least-variance direction to define the hyperplane direction $\hat{\boldsymbol{w}}$.Together

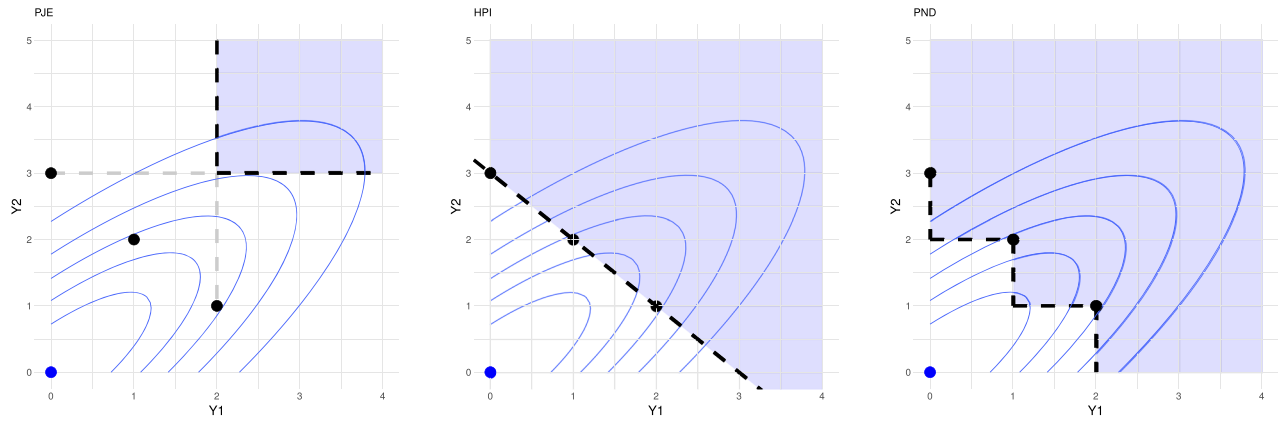

FIG. 4. *Schematics illustrating acquisition functions:* probability of joint exceedence (PJE, left), hyperplane probability of improvement (HPI, center), and probability non-dominated (PND, right) for the same candidate (blue, at origin) and frontier points (black). Both outputs $Y1$, $Y2$ are to be maximized. The blue curves depict equi-likelihood contours for a single candidate's predictive density $\hat{\mathscr{D}}$. The shaded region depicts the region that is integrated for the respective acquisition function. Note that the PJE largely ignores the frontier geometry, the HPI crudely models the Pareto frontier, and the PND accurately considers the Pareto frontier.

with the mean of the Pareto frontier data $\overline{\boldsymbol{Y}}$, we then define the offset via $\hat{b} = \hat{\boldsymbol{w}}^\top \overline{\boldsymbol{Y}}$.

Figure 4 illustrates the HPI: Note that this definition assumes a hyperplane structure, which will not be appropriate for all Pareto frontiers. Furthermore, the HPI "fills in" the staircase structure posed by the true Pareto frontier—the final frontier modeling acquisition function captures this structure.

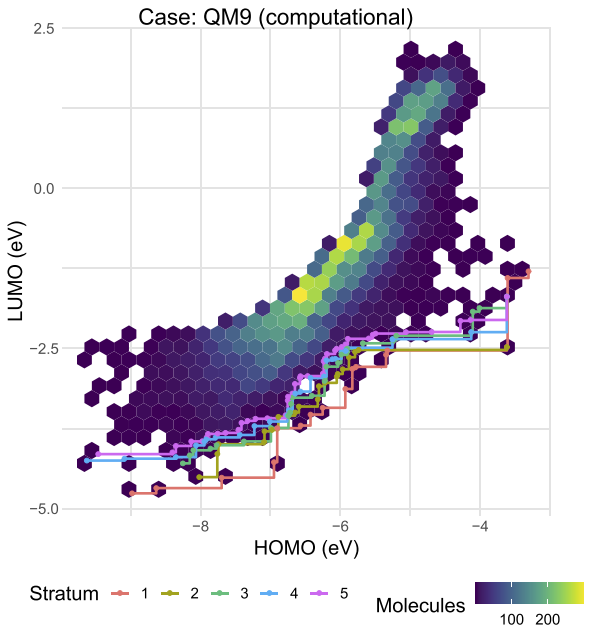

FIG. 5. *Organic molecule objective space.* The HOMO and LUMO property space of the first 25 strata contains 2108 molecules. The strata are derived from maximizing HOMO and minimizing LUMO. These samples vary in density, with far fewer molecules near the Pareto frontier.

Probability Non-Dominated (PND): The PND is defined via

$$f_{\mathrm{PND}}(\boldsymbol{x}_i) = \mathbb{P}_{\mathscr{D}_i}[Y_i \prec \boldsymbol{y},\ \forall \boldsymbol{y} \in \tilde{\mathscr{Y}}], \tag{13}$$

with dominance $\prec$ defined in (1). In words, the PND computes the probability that a given candidate $\boldsymbol{Y}_i$ will be non-dominated with respect to the available data. This criterion is studied in an approximate form in Ref. 29. The PND fully considers the known Pareto frontier, introducing no modeling assumptions. Figure 4 illustrates this acquisition function against the aforementioned definitions.

Note that while PND captures the true geometry of the Pareto frontier, since the frontier is allowed to have quite general structure, no simple analytic expression as for the PJE or HPI exists. Instead, one may approximate the PND via ordinary Monte Carlo, drawing samples from the predictive distribution $\mathscr{D}_i$ and counting the proportion that are non-dominated. This leads to a greater computational expense to evaluate the PND, as compared with the PJE or HPI. The experiments below will demonstrate that this added expense can be valuable if higher AL performance is desired.

IV. RESULTS

A. Test cases

To compare the acquisition functions introduced above, we simulate AL on a collection of synthetic and physical databases. The synthetic cases are constructed to present different Pareto frontier geometries, including a linear frontier, as well as examples of convex and concave frontiers. We also construct a "sparse" frontier containing relatively few low-stratum points. Figure 6 presents these test-cases' two-dimensional output spaces. The functional forms for these models are given in the Appendix.

In the motivating example above (Fig. 1), we considered a published dataset of experimental thermoelectric materials.[32] The performance (property) space consists of thermal conductivity κ, electrical resistivity ρ, and the Seebeck coefficient S. The inputs

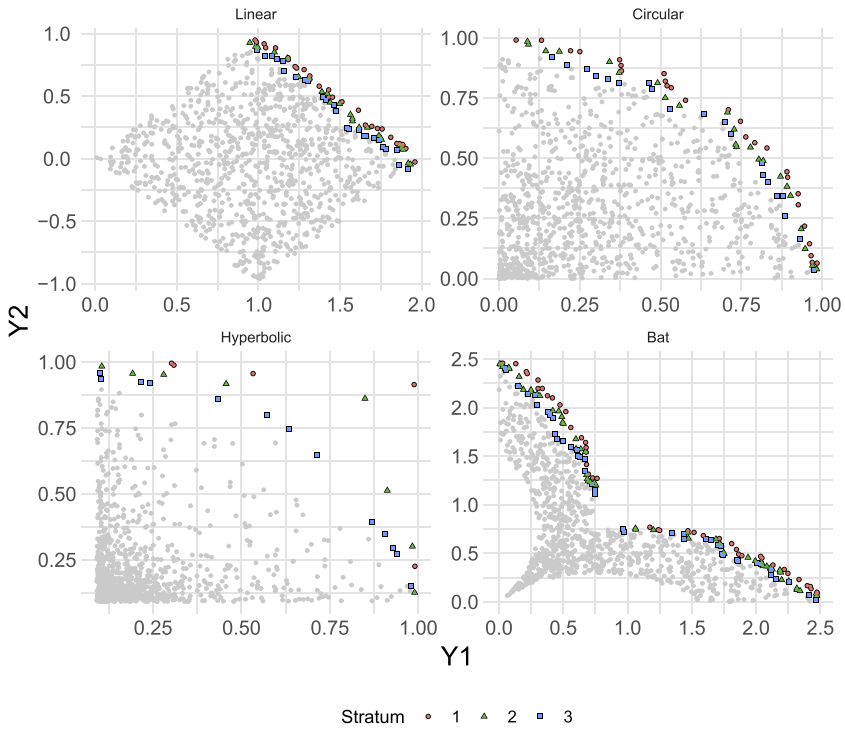

FIG. 6. *Synthetic test cases* in terms of their output (property) spaces. Each output space is two-dimensional, with both outputs to be maximized. Note that each test case has a different Pareto frontier geometry, including convex (`Circular`), concave (`Bat`), and sparse (`Hyperbolic`) examples.

are computed using the Magpie featurization library: We use the `matminer` package to compute features (descriptors) including stoichiometry, valence orbital, and ion properties as inputs.[33,34] Full code to load this dataset and featurization is available in the supplementary material.

We also consider a computational dataset of organic molecules: the QM9 dataset, which contains 134 k small organic molecules with ground-state geometry and several properties computed at the density-functional level of theory.[35] We featurize these data using the Chemistry Development Kit[36] and model the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) properties. This case study is inspired by, but not directly related to, the design of organic photovoltaics.[37] For optimization, we consider maximization of HOMO and minimization of LUMO, corresponding to a minimization of the HOMO–LUMO gap, while seeking a low LUMO value. To promote scalability for replications, we filter the dataset to contain only the first 25 strata, resulting in 2108 molecules. Figure 5 illustrates this frontier.

B. Retrospective active learning experiments

The primary evidence of this work is based on a battery of AL simulations, which support the comparison of different acquisition functions against a common setup. For all experiments, we use the random forest model implemented in the `lolo` package.[38]

Chance phenomena such as initial data selection can affect AL results, implying that single runs of AL are insufficient for measuring performance. We consider an ensemble of runs against a randomized selection of initial data in order to provide a robust estimate of relative performance.

To perform a single run of AL, we carry out the following steps:

1. Choose an initial random subset $\tilde{\mathscr{X}}_0 \subseteq \mathscr{X}$ of size $|\tilde{\mathscr{X}}_0| = C$ for the training data and reveal the paired labels $\tilde{\mathscr{Y}}_0$. Set the iteration counter to $k = 0$.
2. Fit a random forest model to the available paired data $\tilde{\mathscr{X}}_k, \tilde{\mathscr{Y}}_k$. This returns predictions $\hat{\boldsymbol{y}}_i$ and predictive densities $\hat{\mathscr{D}}_i$.
3. Rank the candidates $\boldsymbol{x}_i \in (\mathscr{X} - \tilde{\mathscr{X}}_k)$ remaining in the dataset according to the chosen acquisition function $f_a(\boldsymbol{x}_i)$. Select

TABLE I. *Summary of acquisition functions used in active learning experiments* in this work, ordered by the fidelity with which they represent the Pareto frontier. While the acquisition logic and scalar values (initial candidate pool size C, total iterations K, and replications R) varied across experiments, the model topology and training methods were held constant for all experiments. Thus, differences in results are attributed to randomization (addressed via aggregate comparisons), choice of acquisition function, and choice of test case.

Short	Acquisition function	Fidelity
Rand.	Random selection	NA
MSCV	Maximum sum of coefficients of variation	NA
MPJE	Maximum probability of joint exceedance	Low
MHPI	Maximum hyperplane probability of improvement	Medium
MPND	Maximum probability non-dominated	High

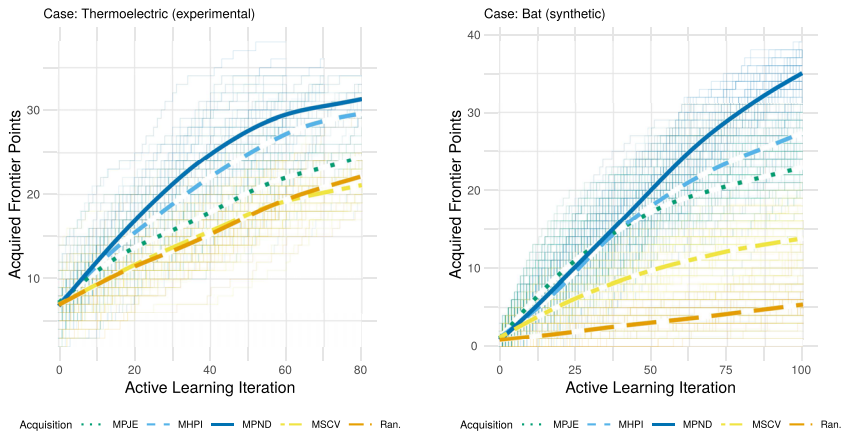

FIG. 7. *Greater acquisition function fidelity leads to more non-dominated candidates.* Total frontier points acquired for the `Thermoelectric` (left) and `Bat` (right) test-cases. In terms of long-run performance, the criteria MPJE, MHPI, and MPND tend to rank in the same order as the fidelity with which they represent the Pareto frontier. This suggests that higher-fidelity acquisition functions enable more non-dominated acquisitions.

the top-performing candidate $\boldsymbol{x}^*$ and add it to the database $\tilde{\mathcal{X}}_{k+1} = \tilde{\mathcal{X}}_k + \{\boldsymbol{x}^*\}$. Reveal the label for $\boldsymbol{x}^*$.

4. Repeat from Step 2 for $k = 1, \ldots, K-1$ total iterations.

To perform an ensemble of AL runs, we select different random subsets $\tilde{\mathcal{X}}_{0,1}, \ldots, \tilde{\mathcal{X}}_{0,R}$ for R repetitions and aggregate the results across these runs. We vary the initial candidate pool size C, the total number of iterations K, the considered test case, and the chosen acquisition function. We performed 150 AL runs for each choice of acquisition function and test case: Table I summarizes experimental details, while the supplementary material provides numerical details on C, K, and R.

C. Candidate improvement

Here, we report results on candidate discovery performance. We consider both the usual metric of the number of non-dominated points (NNDP) acquired (Fig. 7) and the proposed measure of the mean stratum number (Fig. 8). Generally, the acquisition functions that acquire the most NNDP are (1) MPND, (2) MHPI, and (3).

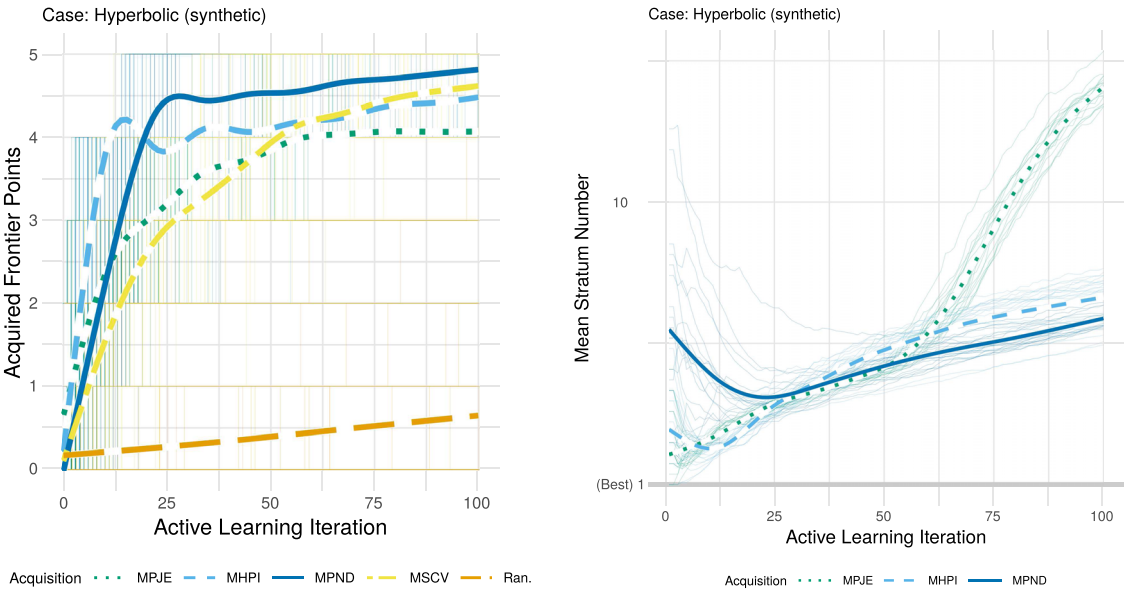

FIG. 8. *Mean stratum number highlights trends not captured by the number of non-dominated points (NNDP, also acquired frontier points).* The three improvement-based criteria result in similar NNDP on the `hyperbolic` case (higher is better). However, studying the mean stratum number reveals strong differences: The MPJE and MHPI approaches tend to more quickly approach the frontier (lower mean stratum is better), while the MPND approach steadily acquires lower-stratum candidates. The MPJE approach saturates in performance around 50 iterations, while we do not observe this in the NNDP results. In cases where an analyst cares not just for Pareto candidates but also for near-optimality, the mean stratum number is a higher resolution metric to judge (multi-objective) the active learning performance.

MPJE—this matches the descending order for the Pareto frontier representation fidelity. The ranking results for the same criteria are mixed for shorter-term performance. These results suggest that capturing the true geometry of the Pareto frontier is most important in "later stage" AL, where the space of candidates is limited. As seen in Fig. 1, MPND also out-performs other criteria on the thermoelectric dataset in terms of long-run performance. These results suggest that incorporating high-fidelity modeling of the Pareto frontier into the

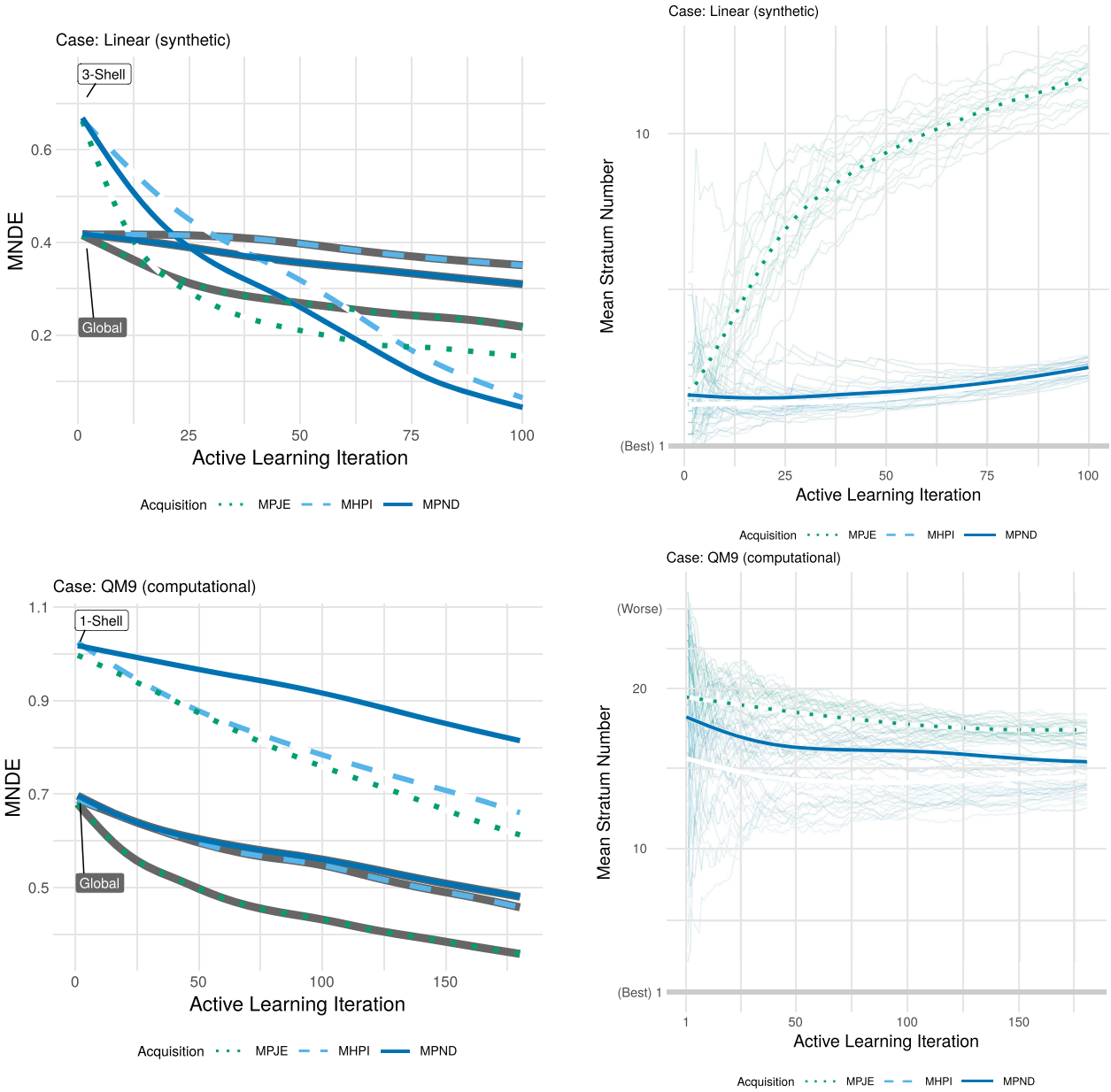

FIG. 9. *Shell error is more associated with candidate improvement than global error.* Different error scopes computed on the same set of active learning replications suggest different rankings of model accuracy, associated with the chosen acquisition function. As seen in Fig. 1 and here, for the `linear` case, global error can be misleading when considering a model for active learning performance. For the `linear` frontier (top row), the MPJE approach initially reduces the 3-shell error but saturates. This trend is reflected in the mean stratum number as the MPJE continues to acquire points distal to the frontier. Conversely, the MHPI and MPND approaches continue to decrease their 3-shell error and maintain a low mean stratum number. Note that this association is not predictive; on the `linear` case, the MHPI and MPND strategies do not overtake MPJE in 3-shell error until well after the latter has saturated. The association between acquisition performance and shell error is not perfect: The QM9 data (bottom row) show that 1-shell error erroneously favors MPJE but correctly favors MHPI. While an improvement over global error, is it clear that more factors are salient in candidate improvement than the frontier-based scope alone.

acquisition function is particularly important in well-studied problems with few possible candidates. Finally, note that the hyperbolic test-case has only five non-dominated candidates total—this fundamentally limits the resolution of AL results for this case. Figure 8 analyzes the same AL results in terms of the NNDP and mean stratum number, demonstrating how the two metrics indicate different trends.

The difference in the "early stage" performance is better revealed by the mean stratum results (Fig. 8). In the early stages of AL, MHPI tends to give the best performance, delivering more candidates at or near the Pareto frontier, while MPND consistently achieves long-run performance. Note that MPJE performance tends to saturate earlier than the other criteria, rather than ranking second when the performance is measured with the NNDP metric—this suggests that MPJE begins selecting candidates far from the Pareto frontier in later-stage AL. The reasons for these differences are explained in Subsection IV E, which analyzes qualitative performance.

D. Accuracy and acquisitions

Here, we report results on how model accuracy relates to acquisitions during AL. Figure 1 illustrates that global scope error can be misleading for judging the relative performance of active learning techniques; Figure. 9 compares shell and global error with acquisition performance. The results on the analytically simple `linear` case indicate that shell scope error is more closely associated with the acquisition of higher-performing candidates.

However, in Fig. 9, we also see complicating factors: Results on the QM9 dataset show that shell error erroneously highlights MPJE as performant but correctly identifies MHPI as performant. Note that the shell scope is based solely on recursive non-dominance, a feature of output property space. Other factors, such as structure–property relationships, are only accounted for indirectly through a shell error calculation. These results indicate that while shell error is more informative for AL performance, the shell scope is evidently (and intuitively) not the only salient factor.

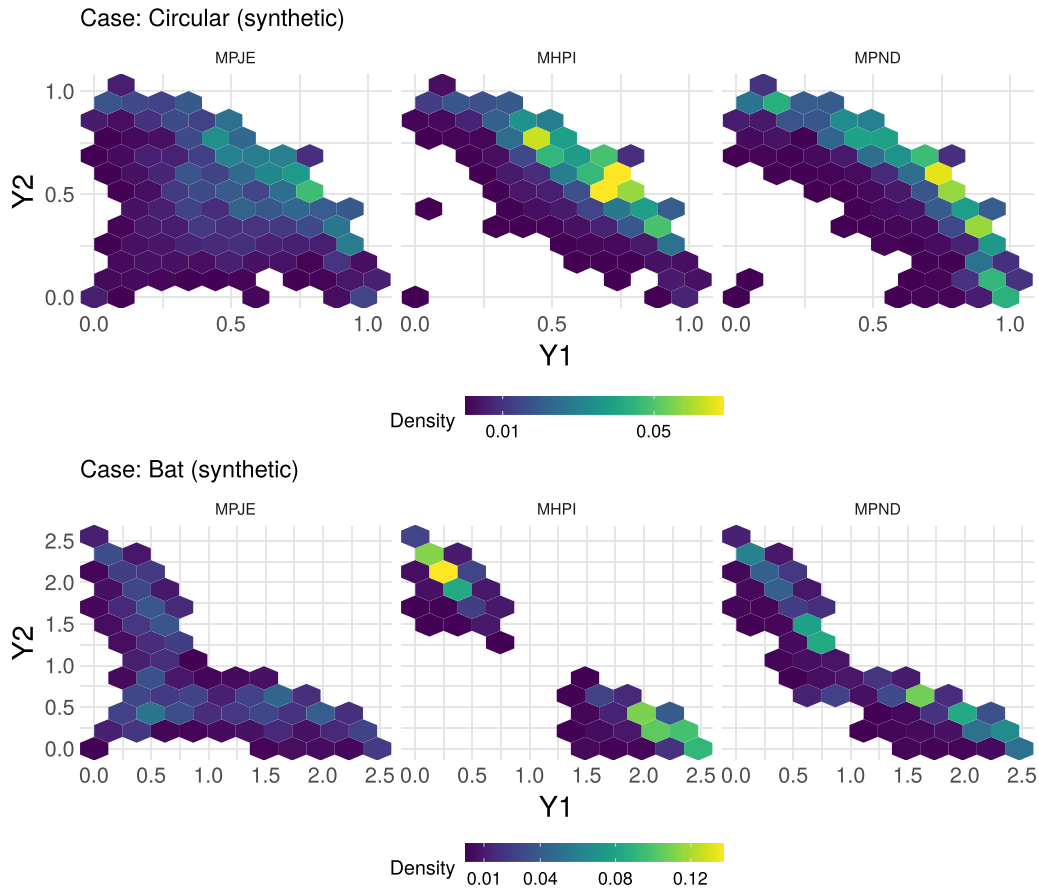

FIG. 10. *Higher-fidelity acquisition functions explore the frontier more thoroughly.* Density of selected points across all AL runs on the `Circular` and `Bat` cases, faceted by acquisition functions; higher density indicates that the acquisition strategy "visits" that region more frequently over independent replications. The selections made according to MPJE are scattered among the entire domain, as compared with MHPI and MPND. MPND tends to distribute its selections along the entire Pareto frontier, as opposed to MHPI, which concentrates on "hot spots" within the frontier. MPND also has wider support than MHPI, indicating that the former criteria tend to (occasionally) select more strongly dominated candidates than the latter.

E. Qualitative performance

Here, we report how the acquisition functions of Sub section III C behave in terms of qualitative performance, where their selections tend to lie in output space. Figure 10 reports empirical densities for selected candidates in output space. Broadly, MPJE tends to select candidates throughout the output space, MHPI focuses on "hotspots" along the frontier, and MPND thoroughly explores the frontier but infrequently selects strongly dominated candidates. This qualitative analysis helps to explain the difference in NNDP (Fig. 7) and mean stratum (Fig. 8) performance: MPND is able to explore the frontier thoroughly, selecting many non-dominated points. However, MPND is also more liable than MHPI to select strongly non-dominated candidates, leading to a higher mean shell number. Conversely, MHPI tends to concentrate its selections closer to a limited region of the Pareto frontier, leading to fewer possible non-dominated selections but a lower mean stratum.

V. CONCLUSIONS

In this work, we explored the relationship between different aspects of machine learning accuracy and candidate acquisition in multi-objective candidate discovery. We showed that AL schemes which optimize for the usual notion of model accuracy—global error—do not guarantee optimal candidate discovery. To ameliorate the situation, we introduced Pareto shell error, which we found to be more closely associated with discovering improved candidates.

We also studied the relationship between the fidelity with which acquisition functions represent the Pareto frontier and AL performance. We demonstrated that dimensionally inhomogeneous acquisition functions can lead to non-robust decision making and so limited our attention to dimensionally homogeneous acquisition functions. We found that the long-run discovery of non-dominated candidates was improved by modeling the Pareto frontier with greater fidelity. However, an acquisition function, which rendered the Pareto frontier with lesser fidelity (MHPI), uncovered more candidates at or near the Pareto frontier in the early stages of AL, leading to a lower mean stratum number than the highest fidelity acquisition function (MPND). We found that these acquisition functions tended to select material candidates at varying locations along the Pareto frontier with different frequency, leading to preferred "hot-spots" in material property space.

These results have ramifications for scientists seeking to use AL for candidate discovery. Since global error is not always predictive of optimal candidate discovery, an analyst should check both global and shell error when deciding whether a model is sufficiently accurate to be used to rank candidates. Furthermore, an analyst can use these insights prescriptively, choosing hyperparameters to optimize shell error rather than global error.

The selection of an appropriate acquisition function is highly dependent on the analyst's goals. Based on our results, accurately modeling the Pareto frontier in an acquisition function is not critical for early stage AL, that is, if the problem has a very large set of uncharacterized material candidates. Accurately capturing the Pareto frontier in the acquisition function logic becomes important when the design space is more thoroughly explored. In the absence of more specific preference criteria than non-dominated, the maximum probability non-dominated (MPND) strategy is the most performant among the acquisition functions tested here.

There are a number of remaining questions related to the present topic. By the simulation nature of our experiments, we were able to evaluate the true Pareto shell error; in practice, one must devise an estimation technique based on the available data. Given the suitability of different acquisition functions for different phases of active learning, a *homotopy* approach that weights different objectives $\lambda f_1 + (1 - \lambda) f_2$ with λ varying between iterations may be a useful framework. In this work, we fit independent models to all output quantities: Prior work suggests that modeling improvements can be made by accurately capturing the output dependence structure among output quantities.[39,40] It would be interesting to study whether similar improvements can be found in AL performance. A conceptually different way to treat the multi-objective setting is to turn some objectives into constraints; it would be interesting to compare the constrained approach against a multi-objective acquisition function in terms of the mean stratum number (cf. Fig. 8) and acquisition densities (cf. Fig. 10). Similarly, it would be interesting to compare performance across AL strategies using different scalarizations of the multi-objective space, e.g., the thermoelectric figure of merit zT.

SUPPLEMENTARY MATERIAL

See the supplementary material for full code to reproduce the datasets used in this study.

AUTHORS' CONTRIBUTIONS

Z.R., Y.K., E.A., and J.L. designed the research. Z.R. performed the research and analyzed the data. Z.R. and M.R. wrote the paper. All the authors read the manuscript, commented on the contents, and agreed with the publication of the results. M.R. contributed to the present work while employed by Citrine Informatics.

ACKNOWLEDGMENTS

Z.R. was funded by the Diversifying Academia, Recruiting Excellence program at Stanford University. Part of this work was performed while M.R. was visiting the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation (Grant No. DMS-1440415).

APPENDIX: TEST CASE DETAILS

The following are the underlying distributions and function used to generate the synthetic data cases. These test cases were designed to provide a variety of Pareto frontier geometries. For all four synthetic test cases, the output space is two dimensional $\boldsymbol{Y} \in \mathbb{R}^2$, and both outputs are to be maximized. Code for reproducing these datasets is provided in the supplementary material.

1. `Linear` test case

A simple linear frontier geometry, generated by rotating (and stretching) a uniform distribution,

$$\begin{aligned} \boldsymbol{X}_i &\sim \mathscr{U}[0,1]^2, \\ \boldsymbol{Y} &= [X_{i,1} - X_{i,2}, X_{i,1} + X_{i,2}]^\top. \end{aligned} \tag{A1}$$

2. `Circular` test case

A circular frontier geometry, generated via trigonometric functions,

$$\begin{aligned} X_{i,1} &\sim \mathscr{U}[0,1], \\ X_{i,2} &\sim \mathscr{U}[0,\pi/2], \\ \boldsymbol{Y}_i &= [X_{i,1}\cos(X_{i,2}), X_{i,1}\cos(X_{i,2})]^\top. \end{aligned} \tag{A2}$$

3. `Hyperbolic` test case

The hyperbolic responses lead to a far greater density of points near the origin. This results in a "sparse" Pareto frontier, which often has very few non-dominated candidates,

$$\begin{aligned} \boldsymbol{X}_i &\sim \mathscr{U}[0,10]^2, \\ \boldsymbol{Y}_i &= [1/X_{i,1}, 1/X_{i,2}]^\top. \end{aligned} \tag{A3}$$

4. `Bat` test case

A non-convex Pareto frontier generated by perturbing the radius of the circular test-case,

$$\begin{aligned} X_{i,1} &\sim \mathscr{U}[0,1], \\ X_{i,2} &\sim \mathscr{U}[0,\pi/2], \\ \boldsymbol{Y}_i &= [(X_{i,1} + 2|X_{i,2} - \pi/4|)\cos(X_{i,2}), \\ &\quad (X_{i,1} + 2|X_{i,2} - \pi/4|)\sin(X_{i,2})]^\top. \end{aligned} \tag{A4}$$

DATA AVAILABILITY

The data that support the findings of this study are available within the article (and its supplementary material).

REFERENCES

[1] B. Settles, *Active Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning Vol. 18 (Morgan & Claypool, 2012).

[2] R. Aggarwal, M. J. Demkowicz, and Y. M. Marzouk, "Information-driven experimental design in materials science," in *Information Science for Materials Discovery and Design* (Springer, 2016), pp. 1–44.

[3] E. Garijo del Río, J. Jørgen Mortensen, and K. W. Jacobsen, "Local Bayesian optimizer for atomic structures," Phys. Rev. B **100**(10), 104103 (2019).

[4] K. Gubaev, E. V. Podryabinkin, G. L. W. Hart, and A. V. Shapeev, "Accelerating high-throughput searches for new alloys with active learning of interatomic potentials," Comput. Mater. Sci. **156**, 148–156 (2019).

[5] S. Ju, T. Shiga, L. Feng, Z. Hou, K. Tsuda, and J. Shiomi, "Designing nanostructures for phonon transport via Bayesian optimization," Phys. Rev. X **7**(2), 021024 (2017).

[6] J. Ling, M. Hutchinson, E. Antono, S. Paradiso, and B. Meredig, "High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates," Integr. Mater. Manuf. Innovation **6**(3), 207–217 (2017).

[7] M. Nakayama, K. Kanamori, K. Nakano, R. Jalem, I. Takeuchi, and H. Yamasaki, "Data-driven materials exploration for Li-ion conductive ceramics by exhaustive and informatics-aided computations," Chem. Rec. **19**(4), 771–778 (2019).

[8] D. Packwood, *Bayesian Optimization for Materials Science*, Springer Briefs in the Mathematics of Materials Vol. 3 (Springer Nature, Singapore, 2017).

[9] D. M. Packwood and T. Hitosugi, "Rapid prediction of molecule arrangements on metal surfaces via Bayesian optimization," Appl. Phys. Express **10**(6), 065502 (2017).

[10] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, "Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization," Phys. Rev. Lett. **115**(20), 205901 (2015).

[11] A. Talapatra, S. Boluki, T. Duong, X. Qian, E. Dougherty, and R. Arróyave, "Autonomous efficient experiment design for materials discovery with Bayesian model averaging," Phys. Rev. Mater. **2**(11), 113803 (2018).

[12] M. Todorović, M. U. Gutmann, J. Corander, and P. Rinke, "Bayesian inference of atomistic structure in functional materials," npj Comput. Mater. **5**, 35 (2019).

[13] K. Tran and Z. W. Ulissi, "Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution," Nat. Catal. **1**(9), 696–703 (2018).

[14] T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, and K. Tsuda, "COMBO: An efficient Bayesian optimization library for materials science," Mater. Discovery **4**, 18–21 (2016).

[15] Y. Wang, K. G. Reyes, K. A. Brown, C. A. Mirkin, and W. B. Powell, "Nested-batch-mode learning and stochastic optimization with an application to sequential multistage testing in materials science," SIAM J. Sci. Comput. **37**(3), B361–B381 (2015).

[16] D. Xue, D. Xue, R. Yuan, Y. Zhou, P. V. Balachandran, X. Ding, J. Sun, and T. Lookman, "An informatics approach to transformation temperatures of NiTi-based shape memory alloys," Acta Mater. **125**, 532–541 (2017).

[17] T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, and T. Oguchi, "Crystal structure prediction accelerated by Bayesian optimization," Phys. Rev. Mater. **2**, 013803 (2018).

[18] B. Settles, "Active learning literature survey," Technical Report No. 1648, University of Wisconsin-Madison, Department of Computer Sciences, 2009.

[19] M. J. Kochenderfer and T. A. Wheeler, *Algorithms for Optimization* (MIT Press, 2019).

[20] P. Roocks, "Computing Pareto frontiers and database preferences with the rPref package," R J. **8**(2), 393–404 (2016).

[21] D. C. Montgomery, *Design and Analysis of Experiments*, 9th ed. (Wiley, 2017).

[22] M. S. Jørgensen, U. F. Larsen, K. W. Jacobsen, and B. Hammer, "Exploration versus exploitation in global atomistic structure optimization," J. Phys. Chem. A **122**(5), 1504–1509 (2018).

[23] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, "Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids," Phys. Rev. B **89**(5), 054303 (2014).

[24] R. O. Ritchie, "The conflicts between strength and toughness," Nat. Mater. **10**(11), 817 (2011).

[25] M. Ashby, *Materials Selection in Mechanical Design*, 4th ed. (Butterworth-Heinemann, Oxford, 2011).

[26] A. Solomou, G. Zhao, S. Boluki, J. K. Joy, X. Qian, I. Karaman, R. Arróyave, and D. C. Lagoudas, "Multi-objective Bayesian materials discovery: Application on the discovery of precipitation strengthened NiTi shape memory alloys through micromechanical modeling," Mater. Des. **160**, 810–827 (2018).

[27] S. Weisberg, *Applied Linear Regression* (John Wiley & Sons, 2005), Vol. 528.

[28] C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken, and M. Scheffler, "Identifying domains of applicability of machine learning models for materials science," Nat. Commun. (published online 2020).

[29] A. J. Keane, "Statistical improvement criteria for use in multiobjective design optimization," AIAA J. **44**(4), 879–891 (2006).

[30] J. Svenson and T. Santner, "Multiobjective optimization of expensive-to-evaluate deterministic computer simulator models," Comput. Stat. Data Anal. **94**, 250–264 (2016).

[31] A. Owen, Stats 305 notes, 2013.
[32] M. W. Gaultois, T. D. Sparks, C. K. H. Borg, R. Seshadri, W. D. Bonificio, and D. R. Clarke, "Data-driven review of thermoelectric materials: Performance and resource considerations," Chem. Mater. **25**(15), 2911–2920 (2013).
[33] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Comput. Mater. **2**, 16028 (2016).
[34] L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Qi Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. Jeffrey Snyder, I. Foster, and A. Jain, "Matminer: An open source toolkit for materials data mining," Comput. Mater. Sci. **152**, 60 (2018).
[35] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Sci. Data **1**, 140022 (2014).
[36] E. L. Willighagen, J. W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Chertó, O. Spjuth, G. Torrance, C. T. Evelo, R. Guha, and C. Steinbeck, "The chemistry development kit (CDK) v2.0: Atom typing, depiction, molecular formulas, and substructure searching," J. Cheminf. **9**(1), 33 (2017).
[37] M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger, and C. J. Brabec, "Design rules for donors in bulk-heterojunction solar cells—Towards 10% energy-conversion efficiency," Adv. Mater. **18**(6), 789–794 (2006).
[38] M. Hutchinson, Lolo, 2016; accessed 1 July 2019.
[39] *Predicting Structured Data*, edited by G. Bakir, T. Hofmann, B. Schölkopf, S. Alexander, B. Taskar, and V. Vishwanathan (MIT Press, Cambridge, 2007).
[40] T. E. Fricker, J. E. Oakley, and N. M. Urban, "Multivariate Gaussian process emulators with nonseparable covariance structures," Technometrics **55**(1), 47–56 (2013).