The Journal of Chemical Physics

RESEARCH ARTICLE | JULY 03 2024

Multi-armed bandit algorithm for sequential experiments of molecular properties with dynamic feature selection

Md. Menhazul Abedin ; Koji Tabata ; Yoshihiro Matsumura ; Tamiki Komatsuzaki

J. Chem. Phys. 161, 014115 (2024)
https://doi.org/10.1063/5.0206042
CHORUS

AIP Publishing

The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

Multi-armed bandit algorithm for sequential experiments of molecular properties with dynamic feature selection

Cite as: J. Chem. Phys. 161, 014115 (2024); doi: 10.1063/5.0206042
Submitted: 28 February 2024 • Accepted: 16 June 2024 •
Published Online: 3 July 2024

Md. Menhazul Abedin,[1,2] Koji Tabata,[3,4,5,a)] Yoshihiro Matsumura,[5] and Tamiki Komatsuzaki[1,3,5,6,7,b)]

AFFILIATIONS

[1] Graduate School of Chemical Sciences and Engineering, Hokkaido University, Sapporo 060-8628, Japan
[2] Khulna University, Khulna 9208, Bangladesh
[3] Research Institute for Electronic Science, Hokkaido University, Sapporo 001-0020, Japan
[4] Department of Mathematics, Hokkaido University, Sapporo 060-0810, Japan
[5] Institute for Chemical Reaction Design and Discovery (ICReDD), Hokkaido University, Sapporo 001-0020, Japan
[6] Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Yamadaoka, Suita 565-0871, Osaka, Japan
[7] The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki 567-0047, Osaka, Japan

[a)] **Electronic mail:** ktabata@es.hokudai.ac.jp
[b)] **Author to whom correspondence should be addressed:** tamiki@es.hokudai.ac.jp

ABSTRACT

Sequential optimization is one of the promising approaches in identifying the optimal candidate(s) (molecules, reactants, drugs, etc.) with desired properties (reaction yield, selectivity, efficacy, etc.) from a large set of potential candidates, while minimizing the number of experiments required. However, the high dimensionality of the feature space (e.g., molecular descriptors) makes it often difficult to utilize the relevant features during the process of updating the set of candidates to be examined. In this article, we developed a new sequential optimization algorithm for molecular problems based on reinforcement learning, multi-armed linear bandit framework, and online, dynamic feature selections in which relevant molecular descriptors are updated along with the experiments. We also designed a stopping condition aimed to guarantee the reliability of the chosen candidate from the dataset pool. The developed algorithm was examined by comparing with Bayesian optimization (BO), using two synthetic datasets and two real datasets in which one dataset includes hydration free energy of molecules and another one includes a free energy difference between enantiomer products in chemical reaction. We found that the dynamic feature selection in representing the desired properties along the experiments provides a better performance (e.g., time required to find the best candidate and stop the experiment) as the overall trend and that our multi-armed linear bandit approach with a dynamic feature selection scheme outperforms the standard BO with fixed feature variables. The comparison of our algorithm to BO with dynamic feature selection is also addressed.

I. INTRODUCTION

One of the most formidable challenges in chemical science is the search for the optimal molecules that possess desired properties, particularly in drug, material, and catalyst discoveries.[1] The process typically involves a large set of potential candidate molecules, making it a daunting task.[2,3] Sequential optimization is one of the versatile approaches to this challenge, whereby candidate molecules are iteratively tested for synthesis based on the predictions of an adaptively updated machine learning model.[1,4,5] The primary objective of this approach is to identify optimal molecules with desired properties while minimizing the number of experiments required. The recent advancements of robotics in chemistry have further stimulated research in sequential optimization algorithms, making it a promising direction for future study.[6–8]

One of the most widely used methods for solving sequential optimization problems in chemistry is Bayesian optimization

(BO).[9–15] This method involves constructing a probabilistic model that maps chemical conditions, such as reaction conditions, catalysts, and solvents, to the property of interest. Importantly, the sequential strategy used to propose the next candidate balancing the trade-off between exploring new chemical conditions and exploiting known high-performing conditions.[16] The probabilistic model is updated with new data as it becomes available, allowing the algorithm to iteratively search for the optimal molecule. This approach can reduce the number of experiments required, thereby saving time and resources, particularly for the exploration of complex conditions that are difficult to navigate using traditional trial-and-error approaches.[17–19] However, the high dimensionality of features required to describe molecules causes a significant challenge for the application of this method in chemistry, as in the Gaussian process regression model. This challenge can be addressed through the use of various techniques, such as feature selection or introduction of regularization parameter(s). For example, the feature selection based on the Automatic Relevance Determination (ARD) such as SAASBO[20] can be utilized for solving the high dimensionality problem. However, the regression can easily lead to over-fitting in a small dataset, which indicates the difficulty in applying to chemical problems.[14] In addition, the construction of stopping conditions for identifying optimal molecules remains an important issue, as the optimal point is often unknown until all candidate molecules have been examined.

Another approach for solving sequential optimization problems is multi-armed bandit (MAB) algorithm, which could be regarded as a discrete version of BO. BO uses Gaussian processes as likelihood functions, including traditional homoscedastic models and more sophisticated heteroscedastic models where the noise level varies with the input.[21–23] In addition, to handle complex data distributions, transformations such as the Box-Cox transformation[24] or normalizing flows[25] are also employed. In contrast to these BO approaches, the most general class of MAB algorithm does not necessarily require defining either a model or smooth continuity along underlying feature variable(s) *a priori*, but solely requires assuming sampling from an identically independent distribution (i.i.d.) whose shape is unknown.[16,26–30] For example, the best arm identification problem of MAB algorithm[31–35] is aimed at choosing the arm (attached to a slot machine among many slot machines in front of a person) whose expected reward (e.g., number of coins) is the largest with as fewer number of pulling arms of slot machines as possible. Note that the output of rewards is probabilistic for each arm and only at the limit of infinite number of pulling arms, one can precisely know what expected reward (i.e., true mean) is for each arm. The mathematical aspects of the bandit algorithms have been well studied, giving them an advantage in the construction of stopping conditions for finding the best slot machine. In recent years, this algorithm has been applied to chemical problems, such as drug discovery, where each drug candidate is considered as an arm attached to each slot machine, and its efficacy as a drug is evaluated.[4,5,36,37] However, the direct application of the good arm identification bandit problem to chemical problems is not necessarily the most appropriate, as chemical properties are observed experimentally at once or few for each condition in most cases, that is, not so often like several hundreds to more for each experimental condition. This is because macroscopic experiments are considered as less stochastic (or rather deterministic) than single molecule experiments. Note, however, that the experimental values include the measurement (aleatoric) noise of laboratory apparatus and depend on the level of skill of experimentalists, which are considered to be compensated by introducing a noise term into the regression models.

There exist several variants of good arm identification problems in MAB algorithms, such as "Gaussian process-bandits[38,39] (corresponding to BO)," "threshold bandits,[40,41]" "bad arm identification problem,[28]" and "classification bandits.[42,43]" In the linear bandit, instead of regarding "molecule entity" as an arm to generate a reward (i.e., observed property associated with each molecule), the reward is represented by a multilinear regression that is a linear summation of some functions of feature variables.[29] The linear model derived from the observed rewards is subsequently applied to predict the rewards of unobserved candidate molecules. This prediction aids in suggesting the next molecule to be tested. Linear bandit algorithms offer several advantages, including low computational costs and ease of implementation on large datasets. However, the performance of linear bandit algorithms requires further investigation in chemical problems where the dimensionality of chemical conditions is often larger than the available data points.

To overcome the challenges associated with the high dimensionality of features of molecules, we have developed a linear bandit algorithm that incorporates a dynamic feature selection scheme, i.e., the relevant feature variables are dynamically updated along the accumulation of the data to examine. In particular, we have theoretically constructed a stopping condition within this framework. To evaluate the performance of our approach, we compared linear bandit algorithms, specifically Optimism in the Face of Uncertainty Linear (OFUL) bandit with and without dynamic feature selection, Bayesian optimization (BO) with and without dynamic feature selection, and random search (RS) using synthetic data and realistic chemical optimization problems of solubility[44] and enantioselectivity.[45]

II. METHODS AND MATERIALS

We briefly summarize our protocol of linear bandit algorithm for the selection of potential candidate molecules, with the aim of optimizing specific molecular properties (see also Fig. 1). Prior to beginning of the algorithm, a set of potential candidates (=molecules) is prepared in advance. In addition, molecular features for the set of candidates are extracted using standard chemoinformatics tools.[45–47] If a molecule in the primary set of candidates has only less than or equal to five nonzero feature values, such a molecule is removed from the consequent analysis, and all feature variables are standardized as mean zero and variance one.

The algorithm begins with randomly selecting an initial set of molecules (say, ten molecules) from the pool of candidates and includes them in the training dataset. Note here that we only utilize the molecular property associated with the training dataset and suppose that the actual values of the target molecular property of the remaining candidates in the pool are unknown (just similar to real experiments and assigned as "unobserved molecule"). With removing these initial sets of molecules from the candidate pool, the first step of the algorithm is to construct a linear regression model from the chosen training dataset of molecules, to predict the molecular property for the remaining data in the pool. Here, compared to the total number of molecular features and the number of molecules to

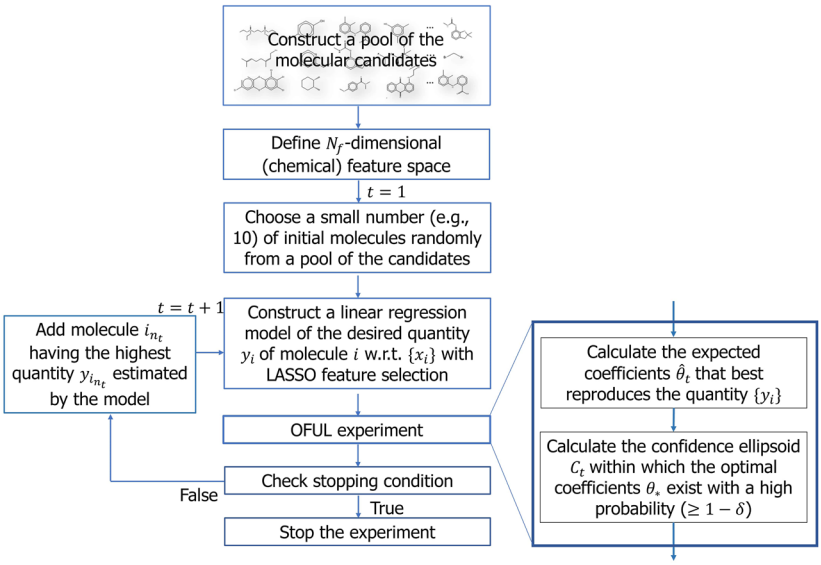

FIG. 1. An overview of our OFUL bandit algorithm combined with dynamic feature selection.

be examined, the former is usually much larger than the latter. Thus, LASSO feature selection[48] is applied to the (initial) set of molecular features, and the selected features are used for modeling. Next, the algorithm suggests a new molecule from the remaining candidates based on the estimated molecular property in question and its uncertainty corresponding to its approximated confidence bound, evaluated using the Optimism in the Face of Uncertainty Linear (OFUL) strategy.[29] Here, if LASSO does not find good feature(s) in the process of the analysis, which sometimes happens especially at very early stage of the process, the algorithm cannot infer the value of the molecular property in question based on the underlying molecular features. Then, the algorithm suggests a molecule randomly from the pool of the (unobserved) candidates. The molecule that is predicted to have a larger value of the molecular property than the largest "observed" among the training set is then "tested." In our study, "testing" corresponds to utilizing the result of the molecular property from the pool of candidates, as just equivalent to a real experiment. With removing the test molecule from the candidate pool, these processes are repeated until a condition is met by using the uncertainty of the predicted molecular property in the model to guarantee with a large probability that the largest value obtained insofar is also the largest in all candidates involving those that has not been examined (the condition is termed as a stopping condition). In this paper, if the stopping condition is met and the largest molecular property in question is actually found (in these simulation experiments, we know the ground truth, while in real experiments, we do not), such a run is considered as success; otherwise, failure. To validate the performance of our algorithm, we implement five algorithms: OFUL-AF and OFUL-SF, which skips and adopts online feature selection, respectively, and Bayesian optimization without online feature selection (BO-AF), with online feature selection (BO-SF), and random search (RS), which randomly suggests the next candidate. We have repeated the simulation several times, and each of the repetitions is termed as run. In the case of synthetic data analysis, every run uses different data of the same size, which are generated randomly as explained before in this section.

In what follows, we present the details on how to select the candidate along the run, the stopping condition, online feature selection, and data.

A. Candidate selection policy

1. Optimism in the face of uncertainty linear (OFUL) bandit

The OFUL bandit[29] algorithm assumes the following linear regression for the molecular property in question (called "reward" hereinafter) y_i of the ith molecule:

$$y_i = \theta_*^T x_i + \eta_i, \tag{1}$$

where x_i denotes the vector representing a set of molecular features associated with the ith molecule, that is, $x_i := \{x_{i,k}\}(i = 1, \ldots, N_{\text{tot}}, k = 1, \ldots, N_f)$. Here, $x_{i,k}$ denotes the kth feature variable of the ith molecule, and N_{tot} and N_f are the arbitrarily large total number of candidate molecules and the total number of feature variables, respectively. θ_* corresponds to the N_f-dimensional vector of weight parameters θ_{*k} assigned to each x_k, which can be acquired only using the full N_{tot} set of candidates, and is unknown in general when a smaller set of molecules is only available. η_i characterizes some deviation of the regression model from the actual reward, modeled by R-sub-Gaussian noise. The R-sub-Gaussian distribution[49] corresponds to a Gaussian distribution with zero mean, $\mathbb{E}[\eta_i] = 0$, and variance bounded by R^2, $\mathbb{E}[\eta_i^2] \leq R^2$.

In this paper, we estimate the values of the weight parameters θ along the iteration step t of the algorithm, denoted by $\hat{\theta}_t$, which can be different from the underlying unknown, true θ_*. An l^2-regularized least-squares estimate of θ_* at iteration step t is

given by[29]

$$\hat{\theta}_t = (X_{1:n_t}^T X_{1:n_t} + \lambda I)^{-1} X_{1:n_t}^T y_{1:n_t}, \tag{2}$$

where $n_t = t + 10$ because we begin with ten randomly chosen molecules at the initial step $t = 0$, $\lambda(> 0)$ is the regularization parameter, and I is the identity matrix of size $n_t \times n_t$. $X_{1:n_t}$ is the matrix represented by $\left(x_{i_1}^T, x_{i_2}^T, \ldots, x_{i_{n_t}}^T\right)$, where the indices from i_1 to i_{10} represent the indices of molecules randomly chosen initially and i_{11} to i_{n_t} represent those of molecules "observed" up to iteration step t, and the vector $y_{1:n_t} = (y_{i_1}, y_{i_2}, \ldots, y_{i_{n_t}})^T$ whose elements are observed rewards up to step t.

The confidence region (termed as confidence ellipsoid[29]) that includes true weight parameters θ_* with high probability at least $1 - \delta$ can be constructed using the estimated parameters $\hat{\theta}_t$ as follows:[29]

$$C_t = \left\{ \theta \in \mathbb{R}^{N_f} : \|\theta - \hat{\theta}_t\|_{\overline{V}_t} \leq R\sqrt{2\log\left(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{\frac{1}{2}} S \right\},$$

where

$$\|\theta - \hat{\theta}_t\|_{\overline{V}_t} = \sqrt{(\theta - \hat{\theta}_t)^T \overline{V}_t (\theta - \hat{\theta}_t)} \tag{3}$$

for all $t \geq 0$. Here, δ is a user-controllable parameter for failure probability and another parameter S is chosen so that the upper bound of θ_* is lower than or equal to S. Note that R is a parameter to express the extent of variance of R-sub-Gaussian noise introduced in Eq. (1) and approximately evaluated by Eq. (A4) in the Appendix. The matrix $\overline{V}_t$ is defined as $\lambda I + \sum_{m=1}^{n_t} x_{i_m} x_{i_m}^T$. Here, we use 0.05 as a typical value of the allowance-error rate δ.

The optimism in the face of uncertainty (OFU) principle suggests the next candidate (or molecule), denoted by $x_{i_{t+1}}$, using the following equation:

$$(x_{i_{t+1}}, \tilde{\theta}_{i_{t+1}}) = \underset{(x,\theta)\in D_{t+1}\times C_t}{\arg\max}\ \theta^T x. \tag{4}$$

These two vectors, $x_{i_{t+1}}$ and $\tilde{\theta}_{i_{t+1}}$, are selected to maximize the expected reward $\theta^T x$, among the unobserved remaining candidates at iteration step t (D_{t+1}), each of which is characterized by its molecular feature vector x (denoted by $x \in D_{t+1}$) with the confidence region (C_t) of the weight parameters θ [Eq. (3)]. In other words, this algorithm estimates the confidence intervals of the coefficients θ_*^T by using the data up to time step t, and then, from all unobserved data by that time in D_{t+1}, it selects the molecule candidate i_{t+1} with the potentially highest reward taking into account the uncertainty quantified by C_t (see also Fig. S1 for the illustration of geometrical relationship between C_t and selected $\tilde{\theta}_{i_{t+1}}$). This strategy can take a balance between exploitation and exploration, such as in upper confidence bound strategy, which is known to be useful to efficiently find the maximum reward.[9,10,50] Note that $\tilde{\theta}_{i_{t+1}}$ is different from the evaluated one $\hat{\theta}_{i_{t+1}}$.

2. Bayesian optimization and random search

Bayesian optimization uses a Gaussian process regression as a surrogate function with a radial basis function (RBF) kernel (see Sec. S5 for the confirmation of almost same results with the Matern 5/2 kernel, which is another common kernel). The candidate $x_{i_{t+1}}$ is recommended at iteration step t by the upper confidence bound acquisition function[51,52] defined as follows:

$$x_{i_{t+1}} = \underset{x\in D_{t+1}}{\arg\max}\left(\hat{y}(x) + \beta_t^{\frac{1}{2}}\sigma(x)\right), \tag{5}$$

where $\hat{y}(x)$ and $\sigma(x)$ are the estimated expected value and the standard deviation for the target value of molecule whose feature vector is represented by x in the unobserved set D_{t+1}, respectively, and $\beta_t = 2\log\left(\frac{|D_{t+1}|\pi^2 t^2}{6\delta}\right)$, where δ (typically used 0.05) corresponds to the allowance error rate, as applied in the context of OFUL bandit algorithms, and $|D_{t+1}|$ denotes the number of unobserved molecules remaining at time t in the pool of candidates. Here, we used the isotropic kernel, i.e., a single length scale parameter for all the features standardized by the z-score. The length scale parameter, signal amplitude, and the noise level of kernel are optimized under marginal likelihood. We employed a data-driven approach to determine the searching range of the length scale hyperparameter used in the RBF kernel, which is chosen as the minimum to the maximum of the Euclidean distances between the observed data $\{x_i\}$ up to each iteration time. In addition to OFUL and BO, we also implemented a random search (RS) policy for selecting the sequence of the candidates. While this policy randomly selects each candidate along the process from the entire database of molecules in a uniform manner, RS uses Eqs. (1) and (2) like OFUL for defining its stopping condition presented below.

B. Stopping condition

The construction of stopping condition is crucial to terminate the method automatically. Our algorithm stops testing all the N_{tot} candidates or satisfying the following conditions. Obviously, when we would test all candidates, we could identify the best candidate to have the maximum reward. However, it should be desired to identify the best one more efficiently with fewer number of experiments with ensuring accuracy more than $1 - \delta$ within the framework of Eq. (1). The stopping condition for OFUL at iteration step t is given by

$$y_{\max}(t) > \max_{(x,\theta)\in D_{t+1}\times C_t} \theta^T x, \tag{6}$$

where $y_{\max}(t)$ denotes the maximum value among the observed rewards up to iteration step t and the right-hand side of the equation expresses the upper confidence bound of expected reward at iteration step $t + 1$ within the confidence region of weight parameter constructed at iteration step t. This means that the iteration should be stopped at time t when there are no potential candidates that might become larger than the present maximum. Note that the random search (RS) method also uses this stopping condition because it shares the same prediction model of OFUL.

The stopping condition for Bayesian optimization (BO) is constructed by the following equation derived from Eq. (5):

$$y_{\max}(t) > \max_{x\in D_{t+1}}\left(\hat{y}(x) + \beta_t^{\frac{1}{2}}\sigma(x)\right), \tag{7}$$

where the right-hand side represents the maximum of the upper confidence bound among the unobserved data in the BO framework.

C. Feature selection

The data typically consist of a couple of hundred features; among them, some may not be relevant to the molecular property to predict. Consequently, removing the redundant and irrelevant features is an important step for optimization task in the huge dimensional feature space. The LASSO (Least Absolute Shrinkage and Selection Operation) regression analysis by Hastie *et al.*[48] was applied for this purpose. LASSO is a penalized regression analysis technique that uses l^1 regularization, which can efficiently reduce the number of features for the linear regression model. To decide the regularization parameter of the LASSO model, we apply grid search with fivefold cross-validation for samples observed so far. We examine the OFUL bandit algorithm with all features (AF) that is invariant along the process of the algorithm and with selected features (SF) that is updated dynamically along it, named as OFUL-AF and OFUL-SF, respectively. For comparison, we also examine two versions of Bayesian optimization (BO), BO-AF and BO-SF. Moreover, note that RS does not use any feature selection.

Because each feature has different ranges of values with different physical units, we employ the standardization method that subtracts the mean ($\bar{x}_k$) averaged over all molecules in the pool of the candidates from the kth feature variable of the ith molecule ($x_{i,k}$) with division by the standard deviation (σ_k), resulting in a standardized feature variable, i.e., $x_{i,k} \leftarrow (x_{i,k} - \bar{x}_k)/\sigma_k$.

D. Data description

In this paper, we evaluate our algorithm using four benchmark datasets, two synthetic data and two real datasets.[44,45]

Two synthetic datasets, named Syn-I and Syn-II, are constructed using a linear regression model with Gaussian noise. We define a set of N candidates $y = \{y_i\}$ approximately characterized by a set of N_f feature variables $x = \{x_{i,k}\}$, expressed as $y_i = \sum_{k=1}^{N_f} \theta_k x_{i,k} + \eta_i$, where $i = 1, \ldots, N_{tot}$ and $k = 1, \ldots, N_f$. We set a problem to find the candidate i_m whose y_{i_m} value is the largest as fewer number of experiments as possible. The feature variables intrinsic to each candidate are drawn from a Gaussian distribution with mean 0 and variance 1, $N(0, 1)$. In both datasets, the weight parameter θ_k and Gaussian noise η_i are also sampled from $N(0, 1)$. The first dataset, Syn-I, and the second dataset, Syn-II, consist of $N_{tot} = 300$ candidates and $N_f = 30$ feature variables and $N_{tot} = 300$ candidates and $N_f = 500$ feature variables, respectively. In the former, among the 30 feature variables, randomly chosen 10 are considered to be relevant and important (i.e., the corresponding weight parameters are nonzero otherwise zero). In the latter, only 30 out of 500 feature variables are assumed to be relevant.

The real datasets include hydration free energies of 642 molecules (FreeSolv[44] dataset) and a free energy barrier difference between enantiomer products ($\Delta\Delta G$) of 70 reaction systems (enantioselectivity[45] dataset). The problem is set to find the lowest hydration free energy or the maximum absolute values of $\Delta\Delta G$. The total numbers of features are 447 and 148, respectively, for FreeSolv and enantioselectivity datasets, which are ISIDA (In Silico Design and Data Analysis) fragment descriptors[46,53] based on SMILES[54] (Simplified Molecular Input Line Entry System) codes of molecules. (As a comparison, we added the results performed using Morgan fingerprints in Sec. S4 of the supplementary material.)

III. RESULTS AND DISCUSSION

A. Synthetic data analysis

Figure 2(a) presents the box-and-whisker plot (boxplot) of the time steps required to find the optimal candidate (ground truth), which we called the best finding time, and the stopping time (=the time satisfying the stopping condition) for Syn-I. The best finding time can be computed if and only if the ground truth is known *a priori* (of course this is not the case in real applications). This quantity is only referred in addressing each algorithm performance. For the best finding time, the median values are 135.0, 11.0, 6.0, 13.0, and 10.0 iteration steps for RS, BO-AF, BO-SF, OFUL-AF, and OFUL-SF, respectively. The results show that RS is significantly slower than the other four methods in finding the best candidate, which confirms the importance of model-based sequential optimizations.

Moreover, the variance of best finding time for RS is much larger than that of the other four methods. Regarding the stopping time, the median values are 286.0, 59.0, 41.0, 85.0, and 39.0 iteration steps for RS, BO-AF, BO-SF, OFUL-AF, and OFUL-SF, respectively. Although RS cannot judge to stop in most cases within the maximum number of iterations "290 (=300 candidates - 10 randomly chosen initial seeds)," the other four methods were successful in finding the best candidate and automatically stopping at an early stage of iterations, i.e., the success rates of BO-AF, BO-SF, OFUL-AF, and OFUL-SF are 100%, 100%, 94%, and 98%, respectively. In addition, it is observed that BO and OFUL improve by incorporating dynamic feature selection and the OFUL-SF and BO-SF show a comparable performance both in best finding time and in stopping time. Here, the existence of failure cases by any algorithm (e.g., 2% by OFUL-SF) means that it could not find the underlying ground truth, but before reaching the maximum number of iterations, the algorithm stopped with satisfying the stopping condition. Note that, in this case, one cannot define the best finding time, i.e., the time to find the ground truth in the iteration process, and, hence, such cases were not included in the figure [e.g., the upper figure in Fig. 2(a)].

Figure 2(b) presents the boxplots of the best finding time and the stopping time for Syn-II. The median values of the best finding time are 140.5, 47.0, 31.5, 96.5, and 35.5 iteration steps for RS, BO-AF, BO-SF, OFUL-AF, and OFUL-SF, respectively. The values are larger than those in Syn-I, which indicates the difficulty in Syn-II where many irrelevant feature variables are included. It is found that dynamic feature selection can improve the best finding time both in BO and in OFUL, and the variance of the best finding time is also reduced. For the stopping time, the median values are 290.0, 289.5, 130.0, 290.0, and 172.5 iteration steps for RS, BO-AF, BO-SF, OFUL-AF, and OFUL-SF, respectively. Notably, RS, BO-AF, and OFUL-AF reached the maximum of iteration steps "290," and only BO-SF and OFUL-SF were successful in finding the best candidate and stopping within the limit, i.e., the success rates of BO-SF and OFUL-SF are both 100%. The results show that dynamic feature selection is essential in Syn-II, to which many realistic chemical problems are expected to belong, as in the enantioselectivity dataset below. Although the best finding times of BO-SF and OFUL-SF are comparable, the stopping times of BO-SF tend to be faster in this case. It should be noted in the comparison between Syn-I and Syn-II that the effect of dynamic feature selection is much more significant in Syn-II than in Syn-I. Figures 3(a) and 3(b) present the accuracy of each model of RS, BO-AF, BO-SF, OFUL-AF, and OFUL-SF,

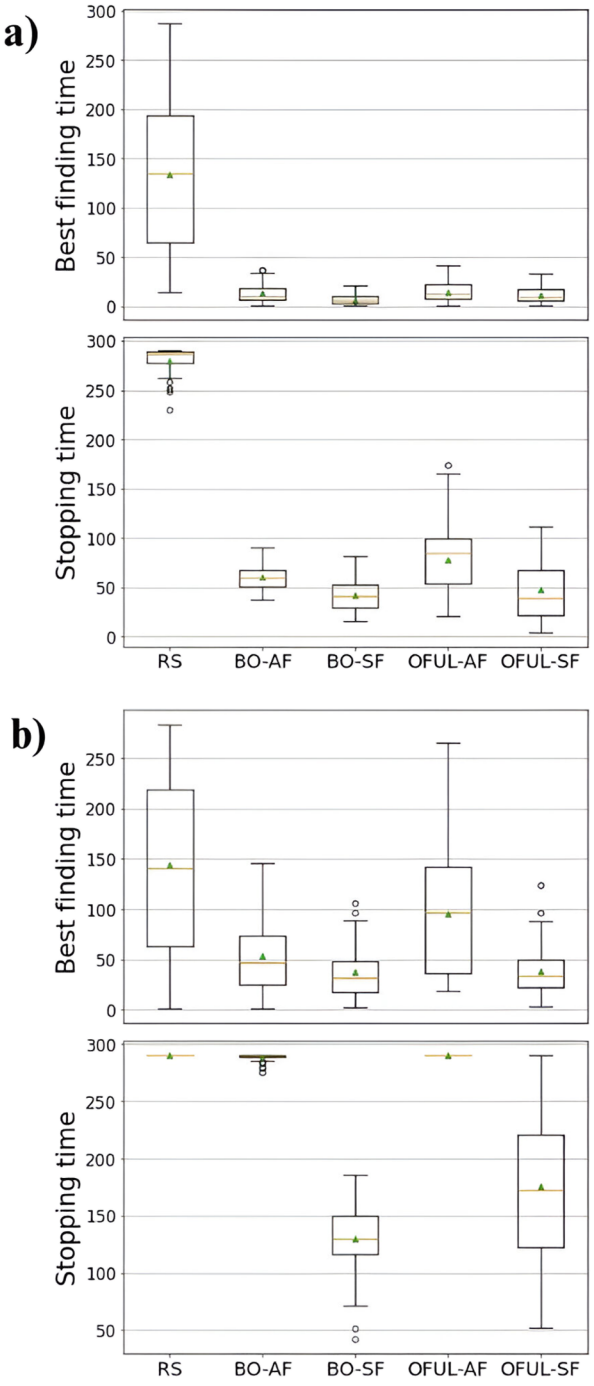

FIG. 2. Boxplots of best finding time (upper) and stopping time (lower) for (a) Syn-I and (b) Syn-II. The orange lines and green triangles indicate the median and mean, respectively. The box indicates the range between the first quartile (Q1) and the third quartile (Q3). The upper whisker indicates the largest value among less than $1.5(\mathrm{Q3}-\mathrm{Q1})+\mathrm{Q3}$. The lower whisker is defined similar to the upper whisker. Outlier points, which are not between whiskers, are shown as the circle points. The total numbers of elements in Syn-I and Syn-II datasets are 300 both, and the experiments are initiated with 10 randomly chosen seeds. Syn-I has 10 relevant feature variables out of 30 variables, and Syn-II has 30 out of 500.

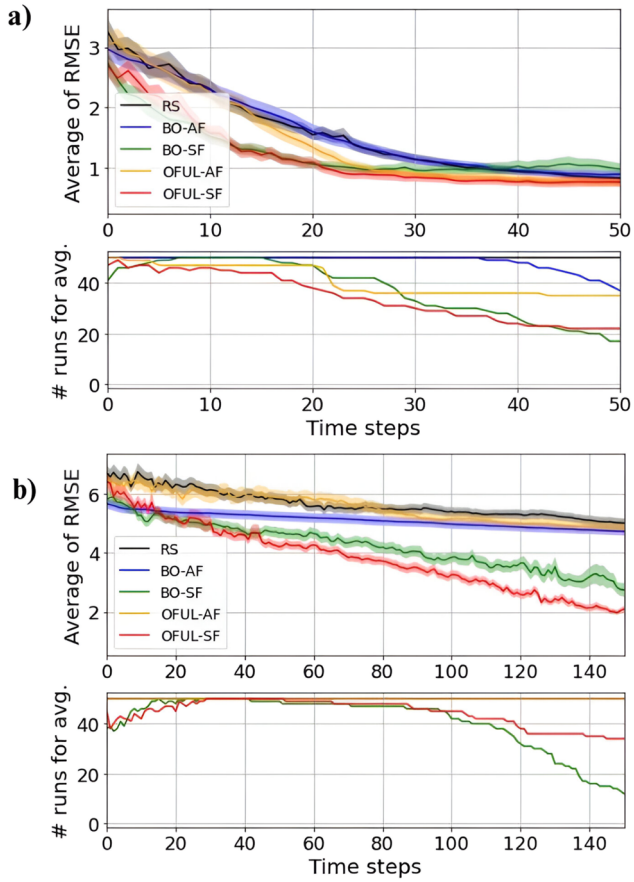

FIG. 3. RMSE analysis for unobserved candidates at each time step, (a) Syn-I and (b) Syn-II. 50 different runs with different initial seeds, and each colored shaded area corresponds to $\pm\sigma$ of the statistics. The number of runs used for statistics is also shown in the lower part. The RS, OFUL-AF, and BO-AF lines for Syn-II are overlapped completely until time step 150.

updated during the sequential optimization, for Syn-I and Syn-II data, respectively. The accuracy is evaluated by using the root-mean-square error (RMSE) for the difference between the predicted values of unobserved candidates at each time step and their true values. The averages and standard errors of RMSE at each time step are calculated for the valid runs that are unstopped yet, and at least one feature is selected by feature selection at that time step. In addition, note that RS utilized the same regression model and stopping condition as OFUL except the candidate selection policy. It is found that RMSEs of all the methods decay gradually as the time step increases, and especially the decays of OFUL-SF and BO-SF are faster than those of the other methods in both datasets. The decreases in RMSEs of these two methods with feature selection are more pronounced at longer iteration steps for the Syn-II case, as being consistent with the larger effect of dynamic feature selection on best finding and stopping times in Syn-II rather than in Syn-I. The comparison of RMSEs between BO-SF and OFUL-SF reveals that OFUL-SF predicts the unobserved candidates y more accurately than BO-SF does (on average) in both Syn-I and Syn-II. This makes us expect that

OFUL-SF should exhibit a better performance in best finding and stopping time analysis. However, the distinction between the performance of BO and OFUL methods in best finding and stopping times is nontrivial, attributed to the different construction of confidence intervals or uncertainties in the predicted values. Within the OFUL framework, the reduction of feature variables in regression is crucial for reducing the confidence region to stop properly as known from Eq. (3) in Sec. II. Similarly, in the BO framework, the variance assessed through Gaussian process regression is employed, with the variances significantly reduced by feature selection. These factors are found to be essential in comprehending the substantial differences in stopping times between the methods with and without dynamic feature selection, observed in the Syn-II results. The different construction of confidence intervals is also related to the observation that the stopping times of BO-SF are faster than those of OFUL-SF in Syn-II, despite the RMSE of BO-SF being larger than that of OFUL-SF at iteration steps beyond ~40. In addition, the stopping times of RS in Syn-I are slightly smaller than those in Syn-II, where the confidence intervals are influenced by the difference in the total number of features.

B. Analysis on experimental hydration free energies and enantioselectivity

Figure 4(a) presents the boxplots of the best finding time and the stopping time for FreeSolv data, where the total number of molecules is 642 with 447 feature variables in total. The median values for the best finding time are 370.0, 7.5, 11.0, 7.0, and 19.0 iteration steps for RS, BO-AF, BO-SF, OFUL-AF, and OFUL-SF, respectively. As for the stopping time, the median values are 632.0, 97.0, 34.0, 499.5, and 68.0 iteration steps for RS, BO-AF, BO-SF, OFUL-AF, and OFUL-SF, respectively. Note that the observed trends in best finding times and stopping times align notably with those of Syn-I rather than Syn-II, except for the OFUL-AF result. The discernible dissimilarity between BO-AF and OFUL-AF in stopping times is more pronounced compared to synthetic datasets, with OFUL-AF exhibiting difficulty in achieving a reduction in confidence intervals due to a large number of total feature variables in FreeSolv case, akin to the Syn-II scenario. In contrast, differences in variances between BO-AF and BO-SF are smaller than those in Syn-II case and akin to Syn-I. The RS reached the maximum iteration steps of "632 (=642 candidates-10 randomly chosen initial seeds)," while the other methods stopped within the maximum limit with identifying the best candidate with a high success rate; the success rates of BO-AF, OFUL-AF, BO-SF, and OFUL-SF were 100%, 100%, 90%, and 94%, respectively. These outcomes underscore that success rates can be diminished by feature selection dependent on the initial seeds, although feature selection accelerates the stopping process with a contraction of confidence intervals. This is interpreted that, if a region of the feature space spanned by initial seeds and those of consecutively chosen candidates is significantly different and isolated from that by the best candidate, the chosen feature variables do not necessarily predict the expected reward of the best candidate. However, note that the 90% or 94% success rate does not necessarily mean "significantly low performance" because the allowance error rate $\delta = 0.05$ allows 5% failure prediction at most within the framework of the Gaussian process (BO) and linear regression model (OFUL). Figure 4(b) presents the boxplots of the

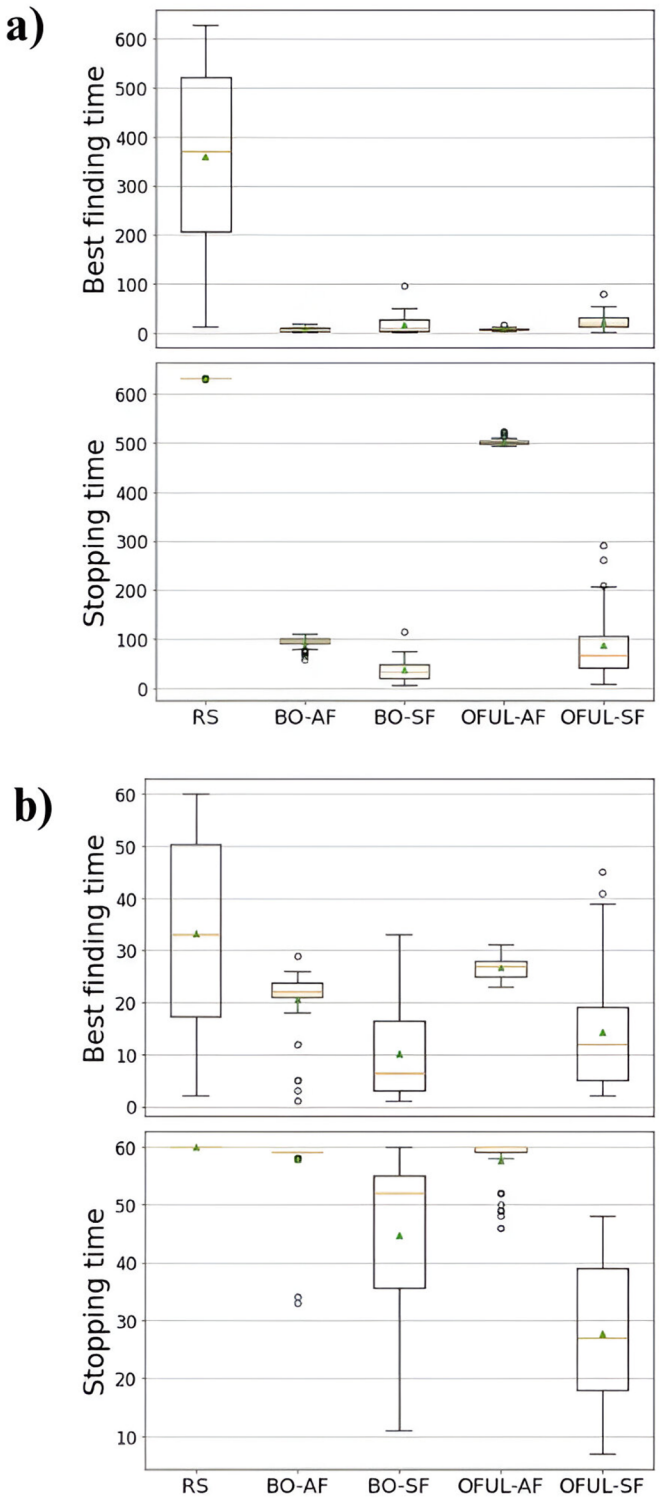

FIG. 4. Boxplots of best finding time (upper) and stopping time (lower) for (a) FreeSolv and (b) enantioselectivity datasets. The total number of molecules in the FreeSolv dataset is 642 and that of enantioselectivity dataset is 70, and the experiments are initiated with 10 randomly chosen seeds.

best finding time and stopping time for enantioselectivity data. The median values for the best finding times are 33.0, 22.0, 6.5, 27.0, and 12.0 for RS, BO-AF, BO-SF, OFUL-AF, and OFUL-SF, respectively. As for stopping times, the median values are 60.0, 59.0, 52.0, 60.0, and 27.0 for RS, BO-AF, BO-SF, OFUL-AF, and OFUL-SF, respectively. The trend in best finding times and stopping times aligns more closely with Syn-II, as opposed to the FreeSolv case. Hence, a feature selection analysis using the full candidate data was conducted for both FreeSolv and enantioselectivity datasets, revealing that FreeSolv has 95 relevant feature variables, constituting 21% of the total, while the enantioselectivity dataset has 15 relevant feature variables, representing 10%. Each value of percentage reasonably corresponds to 33% of Syn-I and 6% of Syn-II cases, respectively, which could be one factor in understanding the characteristics of the results because the ratio of relevant features is related to both the regression model and the construction of the confidence interval. Only OFUL-SF achieved successful stopping with a 98% success rate, while all other methods reached the maximum iteration limit of "60 (=70 candidates - 10 randomly chosen initial seeds)." The

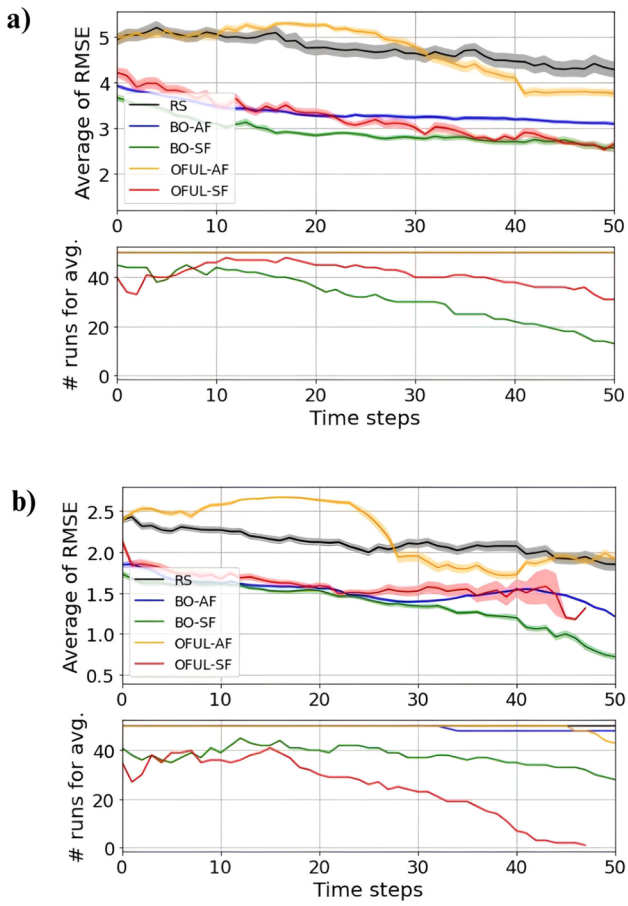

FIG. 5. RMSE for unobserved candidates at each time step, (a) FreeSolv and (b) enantioselectivity dataset. The number of runs used for statistics is also shown in the lower part. The lines corresponding to RS, BO-AF, and OFUL-AF for FreeSolv data are overlapped due to the unstopped runs until time steps 50.

OFUL-SF demonstrated an average reduction of experimental cost by two-thirds, significantly contributing to the exploration of new catalysis for desired enantioselective reactions, given the substantial human cost associated with catalyst preparation. In contrast to the Syn-II case, where BO-SF tends to stop faster than OFUL-SF, in the present case, the OFUL-SF can stop significantly faster than BO-SF, suggesting a complementary property between BO and OFUL depending on the target system.

We analyzed the corresponding RMSE for the FreeSolv and enantioselectivity datasets, presented in Figs. 5(a) and 5(b), respectively. In both datasets, the RMSEs of BO-AF, BO-SF, and OFUL-SF are significantly smaller than those of RS and OFUL-AF. The challenge encountered in regression with OFUL-AF is likely attributed to the non-linearity present in real data, which seems to be mitigated by the dynamic feature selection in OFUL-SF. Here, note that BO-AF and BO-SF utilize a non-linear kernel, achieving a better performance in RMSE. The accuracy expected from RMSE aligns well with the performances in stopping times in the FreeSolv case, resembling the Syn-I case, but not in Syn-II. In the enantioselectivity dataset, the stopping times of BO-AF and OFUL-AF are notably large due to the substantial variances in predictions and the high dimensionality in the feature space, resulting in large confidence intervals or uncertainty. These findings again underscore the correspondence of the FreeSolv and enantioselectivity datasets to Syn-I and Syn-II, respectively. Similar results were observed for Photoswitch data[55] shown in Sec. S6 of the supplementary material.

IV. CONCLUSION

In this study, we developed a novel sequential optimization method, leveraging a linear bandit framework coupled with a dynamic feature selection scheme, for finding the optimal molecule from the candidate library. A distinctive feature of our algorithm is its ability not only to identify the optimal candidate molecule but also to automatically determine when to stop the search based on the predictions for candidate molecules and their confidence intervals evaluated by the regression models. The method underwent validation, starting with synthetic datasets based on linear models and extending to real datasets, including benchmark hydration free energy data and experimental data of enantioselective reactions. Comparative analyses against the linear bandit algorithm without feature selection, Bayesian optimization with and without feature selection, and random search affirmed the essential role of dynamic feature selection, particularly in datasets with numerous irrelevant feature variables included in the total feature variables. We found the critical role of feature selection in enhancing predictive performance and reducing confidence intervals during sequential optimization, thereby accelerating the automatic judgment capabilities of the algorithm in stopping the search.

In this article, we compared the stopping time statistics of OFUL with those of BO with/without feature selection. The most widely used BO does not include dynamic change of variables to express the cost function in its learning process. The stopping conditions of OFUL and BO are not strictly the same, in relevance to how tight the confidence intervals are estimated, and we regard the comparison of the stopping time between BO and OFUL solely as a reference. Note that, in OFUL framework, to incorporate non-linearity in the feature space is straight-forwards by replacing x_i, to a

function composed of $\{x_i\}$, which is expected to extend the applicability of OFUL algorithm. In addition, because OFUL-SF performs feature selection at every time step and deviates from the martingale property originally assumed by OFUL,[29] some derivations solved for OFUL-AF are regarded as a heuristic in the extended framework of OFUL-SF.

SUPPLEMENTARY MATERIAL

See the supplementary material for derivations of Eqs. (A1), (A4), and (A6).

ACKNOWLEDGMENTS

The authors would like to thank Yuuya Nagata, Nobuya Tsuji, and Pavel Sidorov for helpful discussions. This research was partially supported by the Hokkaido University Special Grant Program and DX Fellowship provided by Japan Science and Technology Agency (JST) SPRING (Grant No. JPMJSP2119), Japan Society for the Promotion of Science (JSPS) KAKENHI (Grant No. 21H01924), JST START (Grant No. JPMJST2115), and Japan Agency for Medical Research and Development (Grant No. 22ak0101163h0002).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Md. Menhazul Abedin: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Methodology (lead); Software (lead); Visualization (equal); Writing – original draft (lead); Writing – review & editing (equal). **Koji Tabata**: Conceptualization (equal); Data curation (equal); Methodology (equal); Software (equal); Supervision (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Yoshihiro Matsumura**: Conceptualization (supporting); Data curation (supporting); Formal analysis (supporting); Methodology (supporting); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Tamiki Komatsuzaki**: Conceptualization (equal); Funding acquisition (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The code developed in this study is available at https://github.com/menhaz70/MultiArmedBanditSeqExpChem. In addition, the FreeSolv and enantioselectivity data can be accessed at https://moleculenet.org/datasets-1 and https://github.com/POSidorov/ChemInfoTools, respectively.

APPENDIX: COMPUTATIONAL DETAILS

In this study, the analytical solution of the upper bound of expected reward for each candidate whose molecular feature vector is represented by $x(=\{x_k\})$ $(k = 1, \ldots, N_f)$ is computed using Eqs. (3) and (4). The solution is expressed as

$$\max_{\theta \in C_t} \theta^T x = \hat{\theta}_t^T x + \tilde{B} \|x\|_{\overline{V}_t^{-1}},$$

where

$$\tilde{B} := R\sqrt{2\log\left(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{\frac{1}{2}} S. \tag{A1}$$

This formula can be derived by the Lagrangian multiplier method for maximizing $\theta^T x$ under the constraint $\|\theta - \hat{\theta}_t\|_{\overline{V}_t} = \tilde{B}$. (The detailed derivation is given in Sec. S1 of the supplementary material.)

Equation (3) reveals that the value of the parameter R of R-sub-Gaussian noise that models some deviations of the predicted reward value against the observed reward reflects directly the size of the confidence bound. The larger the value of R is, the more the size of the confidence bound increases. In this article, the value of R is updated during the process and substituted by the upper bound of root mean square error (RMSE) on (updating) training data at each iteration step. The true RMSE is defined as follows:

$$\mathrm{RMSE}_{\mathrm{true}}(t) = \sqrt{\frac{\sum_{t=1}^{n_t}(y_{i_t} - \theta_* x_{i_t})^2}{n_t}}. \tag{A2}$$

In addition, the sample estimate of the RMSE is calculated as follows:

$$\mathrm{RMSE}(t) = \sqrt{\frac{\sum_{t=1}^{n_t}(y_{i_t} - \hat{\theta}_t x_{i_t})^2}{n_t}}, \tag{A3}$$

where i_t denotes the index of the molecule observed at iteration step t and $n_t = t + 10$ in this article. Then, the upper bound of the RMSE (denoted by $\overline{R}_t$) is defined as

$$\overline{R}_t = \sqrt{\frac{n_t}{\chi^2_{1-\frac{\delta}{2},n_t-1}}}\mathrm{RMSE}(t). \tag{A4}$$

Here, $\chi^2_{1-\frac{\delta}{2},n_t-1}$ represents the value that corresponds to the lower end of $[1-\delta/2, \delta/2]$ confidence interval of chi-squared distribution with $(n_t - 1)$ degrees of freedom (see Sec. S2 of the supplementary material for details).

The parameter S in Eq. (3) is also dynamically estimated at each time step t from the l^2-norm of the upper bound of the estimated true parameter θ_*, denoted by $\overline{\theta}_t$, as follows:

$$S = S_t = \|\overline{\theta}_t\|_2, \tag{A5}$$

where the kth element of $\overline{\theta}_t$ is expressed by

$$\overline{\theta}_{t,k} = \hat{\theta}_{t,k} + Z\overline{R}_t\sqrt{(F_t F_t^T)_{kk}}, \tag{A6}$$

where $F_t := (X_{1:n_t}^T X_{1:n_t} + \lambda I)^{-1} X_{1:n_t}^T$, $(F_t F_t^T)_{kk}$ represents the kth diagonal element of $(F_t F_t^T)$, and $Z = 1.96$ so as to meet the condition of

maximum error rate $\delta = 0.05$, i.e., above $\hat{\theta}$, the underlying true θ_* may exist 5% at most (see Sec. S3 of the supplementary material for details).

There are two regularization parameters in our method. One is λ in Eq. (2), and the other is in LASSO for feature selection. Both parameters are tuned using the grid search method, where fivefold cross-validation is adopted on training data.

REFERENCES

[1] S. Gow, M. Niranjan, S. Kanza, and J. G. Frey, "A review of reinforcement learning in chemistry," Digital Discovery **1**, 551–567 (2022).

[2] T. Nakamura, S. Sakaue, K. Fujii, Y. Harabuchi, S. Maeda, and S. Iwata, "Selecting molecules with diverse structures and properties by maximizing submodular functions of descriptors learned with graph neural networks," Sci. Rep. **12**, 1124 (2022).

[3] P. Kirkpatrick and C. Ellis, "Chemical space," Nature **432**, 823–824 (2004).

[4] N. Kikkawa and H. Ohno, "Materials discovery using max K-armed bandit," J. Mach. Learn. Res. **25**(100), 1–40 (2024).

[5] A. Pérez, P. Herrera-Nieto, S. Doerr, and G. De Fabritiis, "AdaptiveBandit: A multi-armed bandit framework for adaptive sampling in molecular simulations," J. Chem. Theory Comput. **16**, 4685–4693 (2020).

[6] D. Caramelli, J. M. Granda, S. H. M. Mehr, D. Cambié, A. B. Henson, and L. Cronin, "Discovering new chemistry with an autonomous robotic platform driven by a reactivity-seeking neural network," ACS Cent. Sci. **7**, 1821–1830 (2021).

[7] G. Lisca, D. Nyga, F. Bálint-Benczédi, H. Langer, and M. Beetz, "Towards robots conducting chemical experiments," in *2015 IEEE/RSJ International Conference on IROS* (IEEE, 2015), pp. 5202–5208.

[8] C. W. Coley, D. A. Thomas III, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao *et al.*, "A robotic platform for flow synthesis of organic compounds informed by AI planning," Science **365**, eaax1566 (2019).

[9] K. Wang and A. W. Dowling, "Bayesian optimization for chemical products and functional materials," Curr. Opin. Chem. Eng. **36**, 100728 (2022).

[10] E. O. Pyzer-Knapp, "Bayesian optimization for accelerated drug discovery," IBM J. Res. Dev. **62**, 2:1–2:7 (2018).

[11] 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).

[12] J. Yan, H. Wei, H. Xie, X. Gu, and H. Bao, "Seeking for low thermal conductivity atomic configurations in SiGe alloys with Bayesian optimization," ES Energy Environ. **8**, 56–64 (2020).

[13] M. Imani and S. F. Ghoreishi, "Bayesian optimization objective-based experimental design," in *ACC* (IEEE, 2020), pp. 3405–3411.

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

[15] J. A. G. Torres, S. H. Lau, P. Anchuri, J. M. Stevens, J. E. Tabora, J. Li, A. Borovika, R. P. Adams, and A. G. Doyle, "A multi-objective active learning platform and web app for reaction optimization," J. Am. Chem. Soc. **144**, 19999–20007 (2022).

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 2018).

[17] P. S. Gromski, A. B. Henson, J. M. Granda, and L. Cronin, "How to explore chemical space using algorithms and automation," Nat. Rev. Chem **3**, 119–128 (2019).

[18] D. Balamurugan, W. Yang, and D. N. Beratan, "Exploring chemical space with discrete, gradient, and hybrid optimization methods," J. Chem. Phys. **129**, 174105 (2008).

[19] J.-L. Reymond and M. Awale, "Exploring chemical space for drug discovery using the chemical universe database," ACS Chem. Neurosci. **3**, 649–657 (2012).

[20] D. Eriksson and M. Jankowiak, "High-dimensional Bayesian optimization with sparse axis-aligned subspaces," in *Uncertainty in Artificial Intelligence* (PMLR, 2021), pp. 493–503.

[21] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard, "Most likely heteroscedastic Gaussian process regression," in *ICML* (ACM Digital Library, 2007), pp. 393–400.

[22] R.-R. Griffiths, A. A. Aldrick, M. Garcia-Ortegon, V. Lalchand *et al.*, "Achieving robustness to aleatoric uncertainty with heteroscedastic Bayesian optimisation," Mach. Learn.: Sci. Technol. **3**, 015004 (2021).

[23] A. Makarova, I. Usmanova, I. Bogunovic, and A. Krause, "Risk-averse heteroscedastic Bayesian optimization," in *NIPS 34* [Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2021], pp. 17235–17245.

[24] A. I. Cowen-Rivers, W. Lyu, R. Tutunov, Z. Wang, A. Grosnit, R. R. Griffiths, A. M. Maraval, H. Jianye, J. Wang, J. Peters, and H. Bou-Ammar, "HEBO: Pushing the limits of sample-efficient hyper-parameter optimisation," J. Artif. Intell. Res. **74**, 1269–1349 (2022).

[25] J. Maroñas, O. Hamelijnck, J. Knoblauch, and T. Damoulas, "Transforming Gaussian processes with normalizing flows," in *International Conference on Artificial Intelligence and Statistics* (PMLR, 2021), pp. 1081–1089.

[26] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," J. Mach. Learn. Res. **7**, 1079–1105 (2006).

[27] J.-Y. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits," in *23rd COLT* (HAL Science, 2010), pp. 41–53; ID: hal-00654404.

[28] K. Tabata, A. Nakamura, J. Honda, and T. Komatsuzaki, "A bad arm existence checking problem: How to utilize asymmetric problem structure?," Mach. Learn. **109**, 327–372 (2020).

[29] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in NIPS 24, 2011.

[30] H. Kano, J. Honda, K. Sakamaki, K. Matsuura, A. Nakamura, and M. Sugiyama, "Good arm identification via bandit feedback," Mach. Learn. **108**, 721–745 (2019).

[31] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *48th CISS* (IEEE, 2014), pp. 1–6.

[32] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," J. Mach. Learn. Res. **17**, 1–42 (2016).

[33] S. Shahrampour, M. Noshad, and V. Tarokh, "On sequential elimination algorithms for best-arm identification in multi-armed bandits," IEEE Trans. Signal Process. **65**, 4281–4292 (2017).

[34] V. Gabillon, M. Ghavamzadeh, A. Lazaric, and S. Bubeck, "Multi-bandit best arm identification," in NIPS 24, 2011.

[35] M. Soare, A. Lazaric, and R. Munos, "Best-arm identification in linear bandits," in NIPS 27, 2014.

[36] K. G. Jamieson and L. Jain, "A bandit approach to sequential experimental design with false discovery control," in NIPS 31, 2018.

[37] H. G. Svensson, E. J. Bjerrum, C. Tyrchan, O. Engkvist, and M. H. Chehreghani, "Autonomous drug design with multi-armed bandits," in *Conference on Big Data* (IEEE, 2022), pp. 5584–5592.

[38] M. Zhang, R. Tsuchida, and C. S. Ong, "Gaussian process bandits with aggregated feedback," in *AAAI Conference on Artificial Intelligence* (AAAI Press, 2022), Vol. 36, pp. 9074–9081.

[39] S. Vakili, N. Bouziani, S. Jalali, A. Bernacchia, and D.-s. Shiu, "Optimal order simple regret for Gaussian process bandits," in *NIPS 34* [Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2021], pp. 21202–21215.

[40] J. D. Abernethy, K. Amin, and R. Zhu, "Threshold bandits, with and without censored feedback," in NIPS 29, 2016.

[41] J. Cheshire, P. Ménard, and A. Carpentier, "Problem dependent view on structured thresholding bandit problems," in *ICML* (PMLR, 2021), pp. 1846–1854.

[42] K. Tabata, A. Nakumura, and T. Komatsuzaki, "Classification bandits: Classification using expected rewards as imperfect discriminators," in *PAKDD Workshop on MLMEIN* (Springer, 2021), pp. 57–69.

[43] T. Hayashi, N. Ito, K. Tabata, A. Nakamura, K. Fujita, Y. Harada, and T. Komatsuzaki, "Gaussian process classification bandits," Pattern Recognit. **149**, 110224 (2024).

[44] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," Chem. Sci. **9**, 513–530 (2018).

[45]N. Tsuji, P. Sidorov, C. Zhu, Y. Nagata, T. Gimadiev, A. Varnek, and B. List, "Predicting highly enantioselective catalysts using tunable fragment descriptors," Angew. Chem., Int. Ed. **62**, e202218659 (2023).
[46]R. I. Nugmanov, R. N. Mukhametgaleev, T. Akhmetshin, T. R. Gimadiev, V. A. Afonina, T. I. Madzhidov, and A. Varnek, "CGRtools: Python library for molecule, reaction, and condensed graph of reaction processing," J. Chem. Inf. Model. **59**, 2516–2521 (2019).
[47]G. Landrum, Rdkit: Open-source cheminformatics, 2006.
[48]T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations* (CRC Press, 2015).
[49]P. Rigollet and J.-C. Hütter, "High-dimensional statistics," arXiv:2310.19244 (2023).
[50]N. S. Madhukar, P. K. Khade, L. Huang, K. Gayvert, G. Galletti, M. Stogniew, J. E. Allen, P. Giannakakou, and O. Elemento, "A Bayesian machine learning approach for drug target identification using diverse data types," Nat. Commun. **10**, 5221 (2019).
[51]A. Gotovos, "Active learning for level set estimation," M.S. thesis, Eidgenössische Technische Hochschule Zürich, Department of Computer Science, 2013.
[52]N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting," IEEE Trans. Inf. Theory **58**(5), 3250–3265 (2012).
[53]A. Varnek, D. Fourches, F. Hoonakker, and V. P. Solov'ev, "Substructural fragments: An universal language to encode reactions, molecular and supramolecular structures," J. Comput.-Aided Mol. Des. **19**, 693–703 (2005).
[54]D. Weininger, "Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Comput. Sci. **28**, 31–36 (1988).
[55]A. R. Thawani, R.-R. Griffiths, A. Jamasb, A. Bourached, P. Jones, W. McCorkindale, A. Aldrick *et al.*, "The photoswitch dataset: A molecular machine learning benchmark for the advancement of synthetic chemistry," chemrxiv.12609899.v1 (2020).