ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Initial Sample Selection in Bayesian Optimization for Combinatorial Optimization of Chemical Compounds

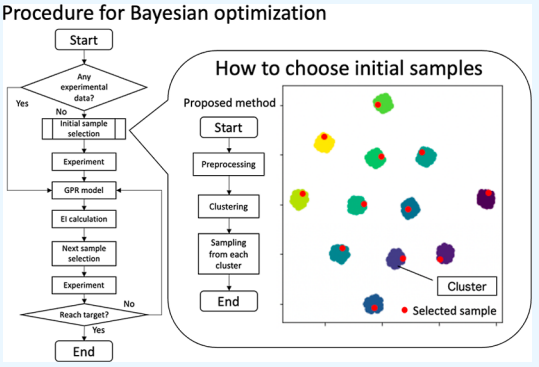
Toshiharu Morishita and Hiromasa Kaneko*

Cite This: *ACS Omega* 2023, 8, 2001−2009

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: An efficient search for optimal solutions in Bayesian optimization (BO) entails providing appropriate initial samples when building a Gaussian process regression model. For general experimental designs without compounds or molecular descriptors in explanatory variable x, selecting initial samples with a larger D-optimality allows little correlation between x in the selected samples, which leads to effective regression model building. However, in the case of experimental designs with compounds, a high correlation always exists between molecular descriptors calculated from chemical structures, and compounds with similar structures form clusters in the chemical space. Therefore, selecting the initial samples uniformly from each cluster is desirable for obtaining initial samples with maximum information on experimental conditions. As D-optimality does not work well with highly correlated molecular descriptors and does not consider information on clusters in sample selection, we propose an initial sample selection method based on clustering and apply it to the optimization of coupling reaction conditions with BO. We confirm that the proposed method reaches the optimal solution with up to 5% fewer experiments than random sampling or sampling based on D-optimality. This study makes a contribution to the initial sample selection method for BO, and we are convinced that the proposed method improves the search performance of BO in various fields of science and technology if initial samples can be determined using cluster information appropriately formed by utilizing domain knowledge.



INTRODUCTION

The search for the optimal solution is one of the most important issues in various fields of science and technology, such as machine learning, neural networks, robotics, aerospace engineering, and the design of experiments (DoE). In this context, Bayesian optimization (BO)[1,2] has been widely studied and utilized as one of the solutions. For example, computational fluid dynamics (CFD)-based optimal design for chemical reactors, in which MO BO was utilized to reduce the number of required CFD runs. The developed optimizer was applied to minimize the power consumption and maximize the gas holdup in a gas-sparged stirred tank reactor, which had six design variables. The saturated Pareto front was obtained after 100 iterations and comprised many near-optimal designs with significantly enhanced performances compared to conventional reactors reported in the literature.[14] Further, BO has been utilized in generating novel molecules with optimized properties. The original scheme, featuring BO over the latent space of a variational autoencoder (VAE), had a problem as it tended to produce invalid molecular structures. Constrained BO was demonstrated as a good approach for solving this type of training set mismatch in many generative tasks involving BO over the latent space of VAE.[4] Additionally, BO has also been applied to search for the most stable molecular conformers. Finding low-energy molecular conformers was challenging because of the high dimensionality of the search space and the computational cost of accurate quantum chemical methods for determining the conformer structures and energies. BO algorithms were combined with quantum chemistry methods to address this challenge. After only 1000 single-point calculations and approximately 80 structure relaxations, which was less than 10% of the computational cost of the current fastest method, the low-energy conformers were found to be congruent with experimental measurements and reference calculations.[5] By contrast, although examples of application to the optimization of reaction conditions were not many, in 2021, the development of a framework for Bayesian reaction optimization and an open-source software tool that allows chemists to easily integrate the optimization algorithms into their laboratory practices was reported.[3] In the study, a large benchmark data set for a palladium-catalyzed direct arylation reaction was collected by using high-throughput experimentation (HTE). BO was applied

Received: August 11, 2022
Accepted: December 20, 2022
Published: December 30, 2022

https://doi.org/10.1021/acsomega.2c05145
ACS Omega 2023, 8, 2001−2009

to two real-world optimization efforts (Mitsunobu and deoxyfluorination reactions) and found to outperform human experts' decision-making in both average optimization efficiency and consistency.

Thus, adaptive DoE using BO has been examined and used in various fields. An efficient search for optimal solutions necessitates providing appropriate initial samples when constructing a GPR model for BO. Generally, as information on the data set of objective variables y and the relationship between y and explanatory variables x are unknown before performing experiments, selecting initial samples using only information on x is necessary. As the optimal solution condition is unknown, selecting initial conditions that are scattered as widely as possible in the chemical space is commonly desirable. Methods such as random sampling, sample selection based on D-optimality, and Latin hypercube sampling are often employed for initial sample selection.[1,9,12,15,23−25]

In the case of experimental designs without compounds as explanatory variables, samples with low similarity, for example, in samples selected by D-optimality-based sampling or Kennard−Stone sampling,[12] initial samples with little correlation among them can be selected and initial experimental samples for model construction can be efficiently obtained. D-optimality is a commonly used initial sample determination method. By contrast, in the case of experimental designs including chemical compounds as one of the explanatory variables, selecting an initial sample corresponds to choosing a combination of compounds. As the explanatory variables for compounds are often molecular descriptors calculated from the chemical structure, the explanatory variables are always highly correlated. Therefore, the method of selecting initial samples based on samples with low similarity such as D-optimality with little correlation among the explanatory variables may not be appropriate. Furthermore, in experimental conditions including compounds as explanatory variables, clusters are often formed for each set of structurally similar compounds. Although uniform selection from each cluster to obtain initial samples that do not have similar experimental conditions is desirable, D-optimality does not consider any information on clusters in sample selection. Given that initial samples can be universally selected from the chemical space, we believe that selecting initial conditions from all clusters, rather than considering correlations among experimental conditions, would improve search efficiency in BO.

Therefore, in this study, we propose a method to select initial samples based on clustering information, considering the characteristics that form clusters for each set of structurally similar compounds. Specifically, clustering is performed as a preprocessing step, based only on the information on explanatory variables, and at least one sample is selected from each cluster after confirming that the samples have formed clusters for each compound. Some clustering-based methods have been proposed before. For example, three initial training data selection methods based on fuzzy clustering were proposed to improve the performance of active learning.[26] On the other hand, there are few cases that confirm the effectiveness of the initial sampling method based on clustering in BO for compound combinations. In this study, the optimization of coupling reactions that include compound structures as explanatory variables was selected as the target of BO application, and after preprocessing,initial samples were determined by random sampling, sample selection based on D-optimality, and sample selection based on clustering, and the number of experiments required to reach an optimal solution in BO. The proposed method was validated by checking the number of experiments required to reach the optimal solution in BO.

METHODS

Bayesian Optimization. The GPR model was used in BO. The Gaussian process GP($\mu(\boldsymbol{x})$, $k(\boldsymbol{x},\boldsymbol{x}')$) represents a distribution over functions characterized by a prior mean $\mu(\boldsymbol{x})$ and a kernel function $k(\boldsymbol{x},\boldsymbol{x}')$. The kernel function $k(\boldsymbol{x},\boldsymbol{x}')$ Matern52 was selected. Since the hyperparameters had already been optimized in a reference paper,[3] the same values were used for the GPR and BO applied in this study.

$$f \sim \mathrm{GP}(\mu(\boldsymbol{x}),\ k(\boldsymbol{x},\ \boldsymbol{x}')) \tag{1}$$

$$k(r) = \alpha\left(1 + \frac{\sqrt{5}\,r}{l} + \frac{5r^2}{3l^2}\right)\mathrm{e}^{-\sqrt{5}\,r/l} \tag{2}$$

$$r = |\boldsymbol{x} - \boldsymbol{x}'| \tag{3}$$

where r is the distance between experimental conditions, α is the output scale parameter, and l is the length scale parameter. The Gaussian process posterior distribution mean μ under experimental condition $\boldsymbol{x}$ is given by

$$\mu(\boldsymbol{x}) = k(\boldsymbol{x})^T(K_\theta + \sigma_n^2 I)^{-1}\boldsymbol{y} \tag{4}$$

where $k(\boldsymbol{x})$ is the covariance vector between the experimental condition $\boldsymbol{x}$ and the training conditions, K is the covariance matrix between all training conditions, σn^2 is the variance of the estimated noise, I is the identity matrix, and $\boldsymbol{y}$ is a vector of responses corresponding to the training data. The variance in the posterior distribution of the Gaussian process at experimental condition $\boldsymbol{x}$ is given by

$$\sigma^2(\boldsymbol{x}) = k(\boldsymbol{x},\ \boldsymbol{x}) - k(\boldsymbol{x})^T(K_\theta + \sigma_n^2 I)^{-1}k(\boldsymbol{x}) \tag{5}$$

In GPR, the predictions are normally distributed; therefore, the hyperparameters can be calculated by the maximum likelihood estimation. The hyperparameters are determined so that the following log-likelihood function is maximized.

$$\log p(\boldsymbol{y}|\theta) = -\frac{1}{2}\boldsymbol{y}^T(K_\theta + \sigma_n^2 I)^{-1}\boldsymbol{y} - \frac{1}{2}\log|K_\theta + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \tag{6}$$

Expected improvement (EI), which is the expected value of $I(\boldsymbol{x})$, is generally selected as the acquisition function in BO. The improvement $I(\boldsymbol{x})$ represents the increase in the objective function $f(\boldsymbol{x})$ relative to the currently best observed outcome f^+.

$$I(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) - f^+ - \delta & f(\boldsymbol{x}) > f^+ \\ 0 & f(\boldsymbol{x}) < f^+ \end{cases} \tag{7}$$

The expectation value of $I(\boldsymbol{x})$, EI($\boldsymbol{x}$), for a given experimental condition $\boldsymbol{x}$ has the form

$$\mathrm{EI}(\boldsymbol{x}) = \begin{cases} I(\boldsymbol{x})\Phi\left(\frac{I(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right) + \sigma(\boldsymbol{x})\varphi\left(\frac{I(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right) & \sigma(\boldsymbol{x}) > \delta \\ 0 & \sigma(\boldsymbol{x}) < \delta \end{cases} \tag{8}$$

where $I(\boldsymbol{x})$ is the improvement of the surrogate mean prediction $\mu(\boldsymbol{x})$ diminished by δ, an empirical exploration parameter, $\sigma(\boldsymbol{x})$

https://doi.org/10.1021/acsomega.2c05145
ACS Omega 2023, 8, 2001−2009

is the surrogate standard deviation, and Φ and φ are the cumulative distribution function and probability density function of the standard normal distribution, respectively. The parameter δ was set to the commonly used value of 0.01.

The next experimental condition is selected by the value of EI($\boldsymbol{x}$). Given that the experimental conditions are expressed as a combination of various compounds, the chemical space is finite, and $\boldsymbol{x}$ can be selected such that the expected value of EI($\boldsymbol{x}$) is the highest.

$$\arg\max_{\boldsymbol{x}\in X} \mathrm{EI}(\boldsymbol{x}) \tag{9}$$

If the experiments are conducted parallelly, the Kriging believer algorithm[21] is used to iteratively compute $\boldsymbol{x}$ for which EI($\boldsymbol{x}$) is the maximum. This is achieved by adding the Gaussian process posterior mean $\mu(\boldsymbol{x})$ to the known data and updating the GPR model. We pursued the following procedure for calculating BO (Figure 1).

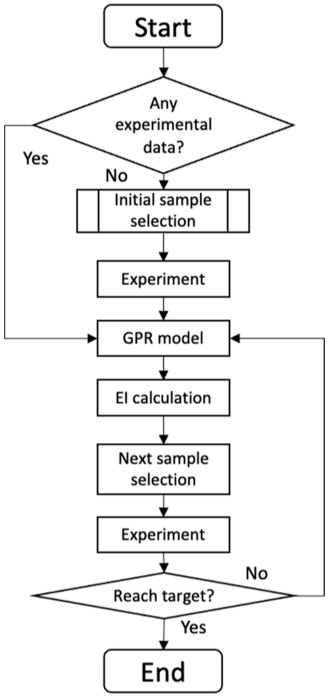


Figure 1. Procedure for Bayesian optimization.

1. Define the experimental space (e.g., solvent, ligand, and temperature) and select initial samples $\boldsymbol{x}$. If the information on samples $\boldsymbol{x}$ and corresponding $\boldsymbol{y}$ is already available, proceed to step 3.
2. Perform the experiments based on the selected samples $\boldsymbol{x}$.
3. Build a GPR model using the information on samples $\boldsymbol{x}$ and the corresponding $\boldsymbol{y}$ and calculate the EI of all samples. All $\boldsymbol{x}$ and $\boldsymbol{y}$ are normalized and used in the calculation.
4. Select the sample with the largest EI as the next experimental condition. If the experiments are conducted parallelly, repeat the calculations.
5. Conduct the experiment based on the selected samples.
6. Repeat steps 3−5 until $\boldsymbol{y}$ reaches the target value.

Initial Sample Selection Based on D-Optimality. D-optimality is commonly used when selecting initial experimental conditions, such as in the DoE. A general linear regression model is expressed by the following equation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} \tag{10}$$

where $\boldsymbol{y}$ is the objective variable vector, $\boldsymbol{X}$ is the explanatory variable matrix (design matrix), and $\boldsymbol{w}$ is the regression coefficient vector, which can be calculated using the following equation

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{11}$$

For the same number of experiments, selecting samples in which the variance of the estimated value is minimum and the covariance is close to zero is desirable. Therefore, the samples should be selected to minimize the elements of the covariance matrix $\sigma 2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

When designing an experiment, after defining the experimental factors, such as the range of operation and type of compounds, initial samples are generated by combining these factors. Obtaining initial samples that do not have similar experimental conditions entails that the initial samples are uniformly selected from the chemical space. D-optimality (determinant of $\boldsymbol{X}^T\boldsymbol{X}$) is repeatedly calculated for the selected samples, and the combination with the largest value is adopted as the initial sample. We performed the following calculation procedure (Figure 2a).

1. Create multiple samples by combining exploratory variables within the defined chemical space.
2. Randomly select initial samples from the obtained samples.
3. Calculate the value of D-optimality for the selected samples and save the sample set with the maximum value of the D-optimality.
4. Repeat steps 2 and 3 until the maximum value of D-optimality is not updated.

There are other similar methods using similarity in experimental space, such as Kennard−Stone sampling;[12] however, in this study, sampling based on D-optimality was used as a representative of those methods.

Clustering. In this study, clustering was utilized as a preprocessing step during the initial sample selection. Partitioning around medoids (PAM)[19] and density-based spatial clustering of applications with noise (DBSCAN)[20] were applied as typical clustering methods.

In PAM, a cluster is represented by medoids, which are data points in the cluster that minimize the sum of distances among all other points in the cluster. When the clusters are $Xi = \{x\}$ and the distance between data is d(x,y), the medoids are described by the following equation

$$\arg\min_{x\in X_i} \sum_{y\in(X_i-\{x\})} \mathrm{d}(x, y) \tag{12}$$

Initially, k medoids are randomly selected, and each medoid is exchanged with the other data point repeatedly to improve the evaluation value. The process is terminated when the evaluation value is no longer improved. The number of clusters k must be given in advance.

DBSCAN is a density-based clustering in which an algorithm assigns labels to each data point based on the following conditions:

- Points with at least *MinPts* of neighbors within a radius ϵ are considered as core points.

https://doi.org/10.1021/acsomega.2c05145
ACS Omega 2023, 8, 2001−2009

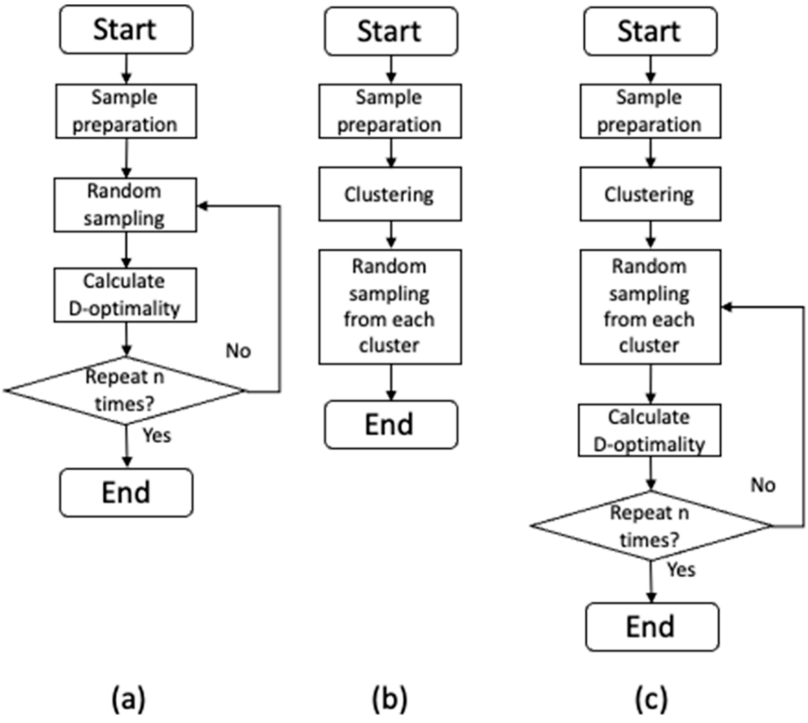


Figure 2. Procedure for initial sampling: (a) D-optimality-based sampling, (b) random sampling with clustering, and (c) D-optimality-based sampling with clustering.

- Points such that the number of adjacent points within radius ϵ is less than *MinPts* but located within radius ϵ of the core point are considered border points.
- All other points are considered noise.

Clustering is performed by forming clusters for each core point and assigning each border point to a cluster of the nearest-neighbor core points. Unlike *k*-means and *k*-medoids, clustering can be performed without specifying the number of clusters; however, the radius distance ϵ used to determine clusters and the threshold MinPts for the number of neighboring points considered as core points must be given as hyperparameters in advance.

Initial Sample Selection Based on Clustering Information. In this study, we propose a method for selecting initial samples based on clustering information. When designing experiments including compounds, after defining the experimental factors, such as the range of operation and type of compound, initial samples are generated by combining these factors. Obtaining initial samples that do not have similar experimental conditions requires the initial samples to be uniformly selected from the chemical space. After clustering based on explanatory variable information and confirming that the samples have formed clusters for each compound group, at least one sample from each cluster is randomly selected to avoid similar experimental conditions in the initial sample selection. Domain knowledge should be utilized to perform clustering on factors that have a large impact on the objective variable. The number of clusters must be less than the number of initial samples. As selecting at least one or more initial samples from each cluster is desirable, the number of initial samples should be n times the number of clusters (n = 1, 2, ...). Additionally, the number of clusters can be automatically determined by applying DBSCAN with appropriate hyperparameters. Methods such as *k*-means and *k*-medoids are not suitable for this approach because they may bias the number of samples belonging to the clusters even if hyperparameters are appropriate. The type of clusters depends on the system; however, in our experience, clusters are often formed by one or two factors (e.g., solvent and ligand). We employ the following calculation procedure (Figure 2b).

1. Create multiple samples by combining exploratory variables within the defined chemical space.
2. Perform clustering method on the obtained samples.
3. Select at least one sample from each cluster to form the initial sample set.

Instead of random sampling, a sampling method based on D-optimality can also be selected. In that case, we perform the following steps in addition to the above procedure (Figure 2c).

4. Calculate the value of D-optimality for the selected sample and save the sample set with the maximum value of D-optimality.
5. Repeat steps 3 and 4 until the maximum value of D-optimality is no longer updated.

There are other similar clustering methods, such as k-means, hierarchical clustering,[13] DBSCAN, and PAM, which were used as representatives of those methods in this study.

RESULTS AND DISCUSSION

Data Set. In this study, we utilized the experimental data reported in a previous paper[3] that employed HTE to collect a large benchmark data set for a palladium-catalyzed direct arylation coupling reaction. The experimental conditions were fixed for the reaction substance, catalyst, and ligand equivalents, and a total of 1728 combinations of 3 reaction temperatures, 3 substance concentrations, 12 ligands, 4 solvents, and 3 bases were used. Only 10 conditions had yields of 95% or higher, accounting for 0.58% of 1728 conditions in the data set. As ligands, solvents, and bases are categorical data, not quantitative variables, we utilized Mordred[7,8] to convert the molecular

https://doi.org/10.1021/acsomega.2c05145
ACS Omega 2023, 8, 2001−2009

structures of the mol files into zero-, one-, and two-dimensional molecule descriptors,[10] and all resulting descriptors were used for calculation. When dealing with compounds, the number of explanatory variables is generally very large. As explanatory variables comprising approximately 5800 variables are high-dimensional data and difficult to interpret, thereby rendering the confirmation of the validity of clustering complex, all values were standardized to mean 0 and standard deviation 1 for preprocessing, and latent variables were calculated by principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) methods.[16] In PCA, the number of variables was reduced to 20 principal components, whose cumulative contribution ratio was approximately 1. Although the perplexity in t-SNE was varied from 10 to 1000, the clusters did not change significantly. Therefore, the result of perplexity 85 was used in this study for validation.

In Bayesian reaction optimization, as a combination of discrete values such as compound species, temperature, and concentration is optimized, experimental condition samples are spatially discretely distributed, and each compound forms clusters with an almost uniform number of samples. In this case, 12 clusters were constructed for each ligand, and samples with high yields were concentrated in certain clusters (Figure 3).

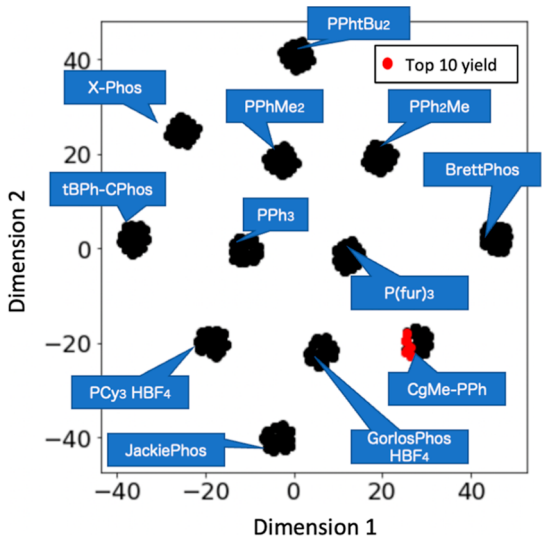


Figure 3. Meaning of each cluster on t-SNE (red dot: top 10 yield conditions).

Clusters formed by ligands are reasonable as they are one of the most important factors in coupling reactions. The t-SNE method maps high-dimensional data to points in low-dimensional space. Learning is performed so that the similarity of samples in the high-dimensional space is reflected in the similarity of samples in the low-dimensional space. In this case, the ligands were represented as relatively large distances between compounds in the higher-dimensional space, which may have led to the formation of clusters for each ligand.

Comparisons of Initial Sample Selection Methods. To confirm the impact of the initial sample selection and clustering methods on the search results in BO, we compared the search performance when the initial samples were determined using the following six sampling methods.

- Random sampling (Random)
- Sampling based on D-optimality (D-optimal)
- Random sampling from each cluster after clustering by DBSCAN (DBSCAN + Random)
- Sampling based on D-optimality from each cluster after clustering by DBSCAN (DBSCAN + D-optimal)
- Random sampling from each cluster after clustering by PAM (PAM + Random)
- Sampling based on D-optimality from each cluster after clustering by PAM (PAM + D-optimal)

The calculation procedures are illustrated in Figure 4. In all cases, the number of experiments per round and the number of samples from clusters were varied to check the effect on the mean and standard deviation of the number of rounds required to reach samples with a yield of 95% or higher. To obtain reliable results, the calculations were repeated more than 10,000 times until the mean and standard deviation for each condition converged. The number of clusters for sampling using clustering information was set to 12, the same number of clusters formed by t-SNE.

To observe the differences among the initial samples determined by each initial sample selecting method, the clustering results were color-coded and visualized by t-SNE. The color coding in Figure 5a−c describe different clusters, with red dots indicating selected initial samples. As mentioned in the Data Set section, t-SNE was performed multiple times with different perplexities; however, the results did not change significantly. Therefore, one of the multiple visualization results was used to plot changes in the initial sample position. When utilizing D-optimal, in some of the cases, two initial samples were selected from one cluster, and in other cases, no initial samples were selected from a cluster (Figure 5a). In the case of DBSCAN + Random, initial samples were uniformly selected from all clusters (Figure 5b). When using PAM + Random, the clustering results were different from the visualization results in t-SNE and included clusters where no initial sample was selected at all and clusters where two samples were selected. In addition, the number of samples in each cluster was not uniform and biased (Figure 5c).

The number of experiments per round was varied to check the mean and standard deviation of the number of rounds required to reach experimental conditions with a yield of 95% or higher. The results are depicted in Figure 6a−c. In the graph, the vertical axis shows that the average number of rounds required to reach 95% yield and the horizontal axis shows the number of experiments per round. Error bars in each figure represent 95% confidence intervals of standard error of the means. In all cases, the differences in means were statistically significant. DBSCAN + Random needed fewer rounds than Random and D-optimal to reach a yield of 95% or higher because we could select at least one compound from all ligands, which is one of the most important factors in coupling reactions, as the initial samples. We could cover all ligands as the initial conditions allowed us to search for conditions with high yields at an early stage. This effect was particularly large when the number of experiments per round and the number of initial samplings were small in all cases. The larger the number of rounds and the initial sampling, the smaller the difference among sampling methods. We speculate that this was because the larger the number of experiments per round or the initial sampling, the fewer the clusters for which the experimental condition was not selected. We confirm that the proposed method reached the optimal solution with up to 5% fewer experiments than random sampling or sampling based on D-optimality. DBSCAN + D-optimal was not significantly

https://doi.org/10.1021/acsomega.2c05145
ACS Omega 2023, 8, 2001−2009

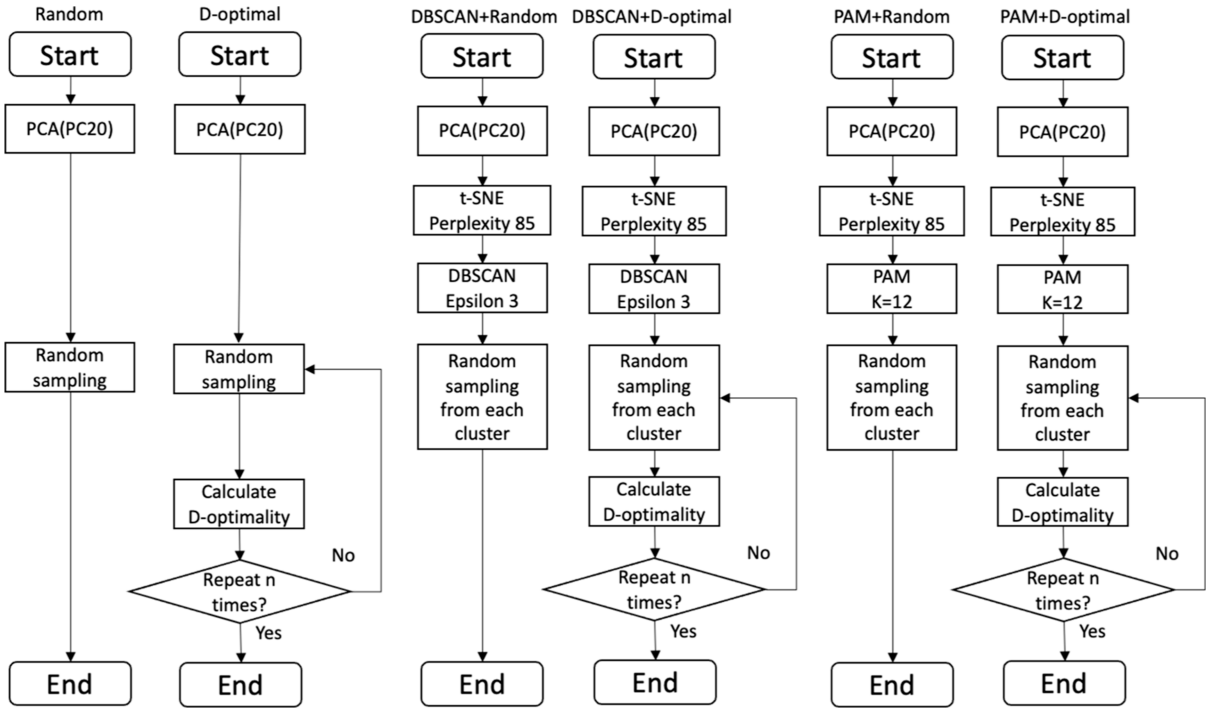


Figure 4. Procedures for each initial sample selecting method.

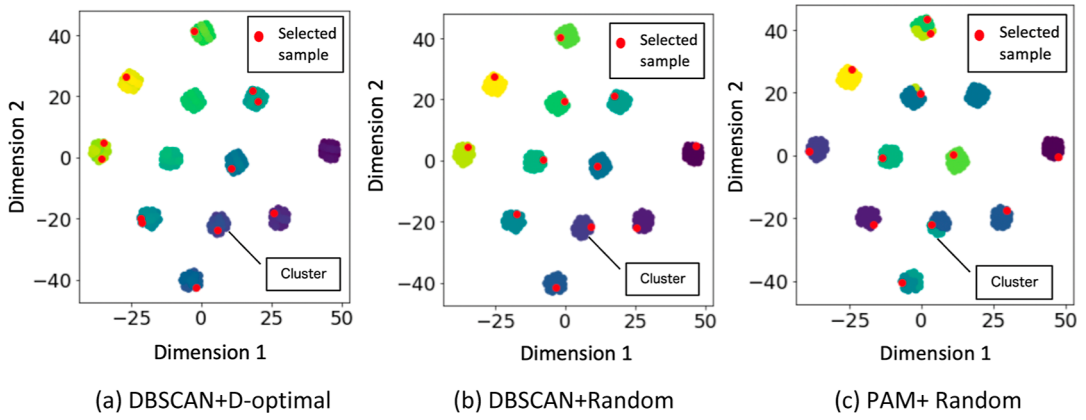


Figure 5. Example of initial samples for different sampling methods on t-SNE. Different colors (blobs) mean different clusters. (a) DBSCAN + D-optimal, (b) DBSCAN + Random, and (c) PAM + Random.

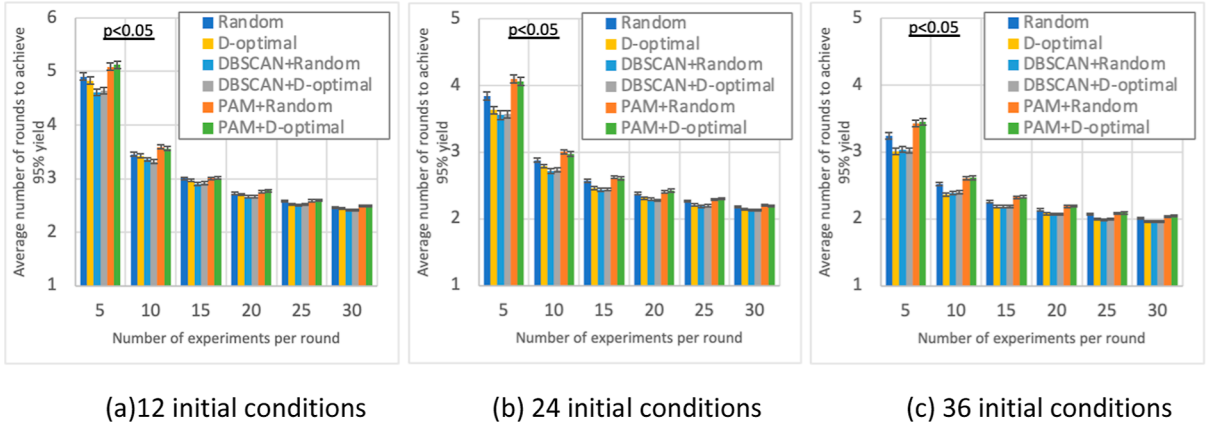


Figure 6. Number of experiments per round vs average number of rounds to achieve 95% yield. (a) 12 initial conditions, (b) 24 initial conditions, and (c) 36 initial conditions.

https://doi.org/10.1021/acsomega.2c05145
ACS Omega 2023, 8, 2001−2009

different from DBSCAN + Random. We suppose that the effect of sampling based on D-optimality was relatively small in this case because the initial conditions were fully selected with DBSCAN clustering before utilizing sampling based on D-optimality. D-optimal, DBSCAN + Random, and DBSCAN + D-optimal provided better results than Random. By contrast, the mean and standard deviation of the number of experiments required for the search for optimal solutions using PAM + Random and PAM + D-optimal was around 3−5% larger than that using Random. This is mainly due to the random selection of initial values, which causes bias among clusters in the experimental conditions and number of experiments, so that clusters are not formed for each ligand species and selecting one sample from each cluster did not cover all ligand species. Usually, the experimental budget is fixed, and the initial number of experiments (number of clusters) tends to be fixed. However, *k*-means-/*k*-medoids-based clustering have the issue described above; therefore, these methods that must specify the number of clusters beforehand should not be used in the clustering-based initial sampling method.

Thus, even if the experimental conditions could be sampled from the clusters with optimal conditions, the search performance could be better or worse depending on the clustering results (e.g., number of samples belonging to a cluster and sample species). In actual optimization, intentionally increasing the probability of sampling experimental conditions that are close to the optimal solution in the reaction space is difficult because prior knowledge of the cluster in which the optimal solution is located is not possible. However, by maintaining the number of samples belonging to a cluster, the variation in the number of experiments required for the search for an optimal solution can be reduced.

In this study, 12 clusters were constructed for each ligand, and the experimental conditions with high yields were concentrated in specific clusters. However, clusters may be formed by other factors (e.g., temperature, solvent, base, and catalyst) depending on compound information, clustering method, and hyperparameters. In such cases, samples with high yields may not be concentrated in a particular cluster and may be scattered over several clusters, and search performance may be poorer than that in the condition where they are clustered by a factor with a large impact on the target variable. For example, in the present case, assuming clustering by solvent rather than ligand, one sample from each cluster can be selected from all solvent species; however, covering the ligand that is the most important factor in the reaction may not be possible. An example of clustering without ligand information is illustrated in Figure 7. Clusters are formed by combinations of solvents and bases, and the experimental conditions for top 10 yields are not concentrated in a single cluster. In such a case, sampling the initial conditions from each cluster would cover all base−solvent combinations, but not all ligands, and the search efficiency improvement effect could not be achieved.

The initial sample selection method proposed in this study, which utilizes clustering information, can reduce the number of experiments required to reach the optimal solution compared to random sampling or sampling based on D-optimality if the initial samples can be clustered appropriately. Appropriate clustering is defined as a case in which the number of samples belonging to the cluster is almost uniform, and clusters can be formed with factors that have a large contribution to the target variable. In the case of compound combination optimization such as reaction condition optimization, experimental conditions are spatially

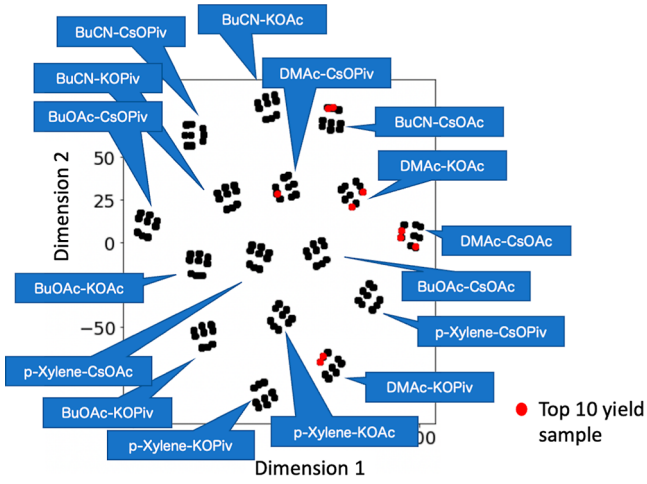


Figure 7. Example of DBSCAN clustering without information for ligand on t-SNE (perplexity 20).

discretely distributed, forming clusters with an almost uniform number of conditions. By contrast, to form clusters with factors that contribute significantly to the target variable, knowledge of organic synthetic chemistry is required. If the initial samples can be determined using appropriately formed cluster information by utilizing domain knowledge, the search performance of BO can be improved further, although it is not easy because the experimental results are unknown before the experiment. In addition, as the number of explanatory variables in the reaction condition optimization in this study was extremely large, the clustering validity was considered to be difficult to confirm, given the difficulty of interpretation with high-dimensional data. Therefore, visualization using t-SNE and clustering using DBSCAN were conducted. Although the hyper-parameters for clustering and visualization methods seemed to need to be determined through trial and error, the result of this study indicated that these are not so significant issues. Appropriate initial samples need to consist of at least one compound from the factors that have a significant impact on the object variable. If one only selects samples one by one from a particular factor, this can be accomplished without using specific machine learning methods (*k*-means, PAM, etc.). The method of determining experimental conditions using clustering by factors is a conventional method and is probably the most accessible to experimenters. When BO is applied to experimental design with composite descriptors, if we can utilize the wisdom of experts, we can improve the search performance of BO by determining initial conditions by sampling at least one factor that is likely to have a significant impact on the objective variable, instead of determining initial samples using the commonly used uncorrelation approach.

CONCLUSIONS

An efficient search for optimal solutions in BO necessitates providing appropriate initial samples when building a GPR model. In the case of experimental designs with compounds, a high correlation always exists between molecular descriptors calculated from chemical structures and compounds with similar structures form clusters in the chemical space. Therefore, selecting the initial samples uniformly from each cluster is desirable for obtaining initial samples with maximum information on experimental conditions. As sampling methods using similarity in experimental space, such as D-optimality,

https://doi.org/10.1021/acsomega.2c05145
ACS Omega 2023, 8, 2001−2009

does not work well with highly correlated molecular descriptors and does not consider information on clusters in sample selection, in this study, we proposed an initial sample selection method based on clustering information that covers the factors with a large impact on the objective variable and applied it to the optimization of coupling reaction conditions with BO. We confirmed that when clusters were appropriately formed and initial samples were selected from each cluster, the proposed method reached the optimal solution with fewer experiments than random sampling or sampling based on D-optimality. Additionally, we also found that when the number of experiments per round was small, the effect of the proposed method was greater than that of other methods not including cluster information, and the number of rounds required for the search could be reduced. Appropriate clustering is defined as a case in which the number of samples belonging to the clusters is almost uniform and clusters can be formed by factors that contribute considerably to the target variable. Clustering is unsupervised learning, and we cannot know information about the objective variable (experimental results) before experiments. Additional information (knowledge of experts) is needed to connect these pieces of information. In the laboratory, there are many requests to use BO in combination with the knowledge of experts. If we can form appropriate clusters and determine initial conditions by utilizing domain knowledge, we can further improve the search performance of BO in the case of compound combinations such as reaction condition optimization. This study makes a contribution to the initial sample selection method for BO, and we are convinced that the proposed method improves the BO search performance in various fields of science and technology if initial samples can be determined using cluster information appropriately formed by utilizing domain knowledge. In the future, we plan to study the handling of cases in which the number of factors is large (when the number of combinations is huge), when multiple factors have a large impact on the objective variable, to determine optimal parameters by quantifying the diversity[22] in clustering results, and in the case of multi-objective optimization.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c05145.

Reaction scheme, compound list, experiment result information, result of principal component analysis, and results of other numerical experiments (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Hiromasa Kaneko − *Department of Applied Chemistry, School of Science and Technology, Meiji University, Kawasaki, Kanagawa 214-8571, Japan;* orcid.org/0000-0001-8367-6476; Email: hkaneko@meiji.ac.jp

Author

Toshiharu Morishita − *Department of Applied Chemistry, School of Science and Technology, Meiji University, Kawasaki, Kanagawa 214-8571, Japan;* orcid.org/0000-0003-4150-2675

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c05145

Notes

The authors declare no competing financial interest.

DATA AND SOFTWARE AVAILABILITY: We used experimental data reported in a paper[3] that applied BO to palladium-catalyzed direct arylation reactions. All programs used in this study were written in Python 3.7.11. EDBO 0.1.0,[11] a Python library for Bayesian reaction optimization, experimental condition, and results can be downloaded at the GitHub site.[6] In EDBO, Scikit-learn 1.0.2[17] and PyClustering 0.9.3.1[18] are applied to calculate DBSCAN, PAM, and t-SNE algorithms. To calculate descriptors[10] from molecular structures, Mordred 1.2.0[7] was used. All data underlying the results are published as part of the paper or on the referenced site and no additional source data are required.

■ ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Scientific Research (KAKENHI) (grant number 19K15352) from the Japan Society for the Promotion of Science.

■ ABBREVIATIONS

BO, Bayesian optimization; CFD, computational fluid dynamics; DBSCAN, density-based spatial clustering of applications with noise; DoE, design of experiments; EI, expected improvement; GPR, Gaussian process regression; HTE, high throughput experimentation; MO, multi-objective; PAM, partitioning around medoids; PCA, principal component analysis; VAE, variational autoencoder

■ REFERENCES

(1) Greenhill, S.; Rana, S.; Gupta, S.; Vellanki, P.; Venkatesh, S. Bayesian Optimization for Adaptive Experimental Design: A Review. *IEEE Access* **2020**, *8*, 13937−13948.

(2) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104*, 148−175.

(3) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590*, 89−96.

(4) Griffiths, R. R.; Hernández-Lobato, J. M. Constrained Bayesian Optimization for Automatic Chemical Design Using Variational Autoencoders. *Chem. Sci.* **2020**, *11*, 577−586.

(5) Fang, L.; Makkonen, E.; Todorović, M.; Rinke, P.; Chen, X. Efficient Amino Acid Conformer Search with Bayesian Optimization. *J. Chem. Theory Comput.* **2021**, *17*, 1955−1966.

(6) b-Shields. https://github.com/b-shields/edbo (accessed July 9th, 2022).

(7) Mordred-Descriptor. https://github.com/mordred-descriptor/mordred (accessed July 9th, 2022).

(8) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminf.* **2018**, *10*, 4.

(9) Nakayama, R.; Shimizu, R.; Haga, T.; Kimura, T.; Ando, Y.; Kobayashi, S.; Yasuo, N.; Sekijima, M.; Hitosugi, T. Tuning of Bayesian Optimization for Materials Synthesis: Simulation of the One-dimensional Case. *Sci. Technol. Adv. Mater.* **2022**, *2*, 119−128.

(10) https://mordred-descriptor.github.io/documentation/master/descriptors.html (accessed Sept 2nd, 2022)

(11) https://b-shields.github.io/edbo/index.html (accessed July 9th, 2022).

(12) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137−148.

(13) Rännar, S.; Andersson, P. L. A Novel Approach Using Hierarchical Clustering To Select Industrial Chemicals for Environmental Impact Assessment. *J. Chem. Inf. Model.* **2010**, *50*, 30−36.

(14) Park, S.; Na, J.; Kim, M.; Lee, J. M. Multi-objective Bayesian Optimization of Chemical Reactor Design Using Computational Fluid Dynamics. *Comput. Chem. Eng.* **2018**, *119*, 25−37.
(15) Klein, A.; Falkner, S.; Bartels, S.; Hennig, P.; Hutter, F.Fast Bayesian Optimization of Machine Learning Hyperparameters on Large Datasets. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*; JMLR: Fort Lauderdale, Florida, USA, 2017; pp 528−536.
(16) Van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(17) scikit-learn Machine Learning in Python. https://scikit-learn.org/stable/ (accessed July 9th, 2022).
(18) https://pyclustering.github.io (accessed July 9th, 2022).
(19) k-Medoids. https://en.wikipedia.org/wiki/K-medoids (accessed July 9th, 2022).
(20) DBSCAN. https://en.wikipedia.org/wiki/DBSCAN (accessed July 9th, 2022).
(21) Yang, K.; Palar, P. S.; Emmerich, M.; Shimoyama, K.; Bäck, T.A Multi-point Mechanism of Expected Hypervolume Improvement for Parallel Multi-objective Bayesian Global Optimization. *Proceedings of the Genetic and Evolutionary Computation Conference*; ACM, 2019; pp 656−663.
(22) van Dam, A. Diversity and its decomposition into variety, balance and disparity. *R. Soc. Open Sci.* **2019**, *6*, 190452.
(23) Zhang, Y.; Bahadori, M. T.; Su, H.; Sun, J.FLASH: Fast Bayesian Optimization for Data Analytic Pipelines. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM, 2016; pp 2065−2074.
(24) Nandiwale, K. Y.; Hart, T.; Zahrt, A. F.; Nambiar, A. M.; Mahesh, P. T.; Mo, Y.; Nieves-Remacha, M. J.; Johnson, M. D.; García-Losada, P.; Mateos, C.; Rincón, J. A.; Jensen, K. F. Continuous stirred-tank reactor cascade platform for self-optimization of reactions involving solids. *React. Chem. Eng.* **2022**, *7*, 1315−1327.
(25) Iwama, R.; Kaneko, H. Design of ethylene oxide production process based on adaptive design of experiments and Bayesian optimization. *J. Adv. Manuf. Process.* **2021**, *3*, 1.
(26) Yuan, W.; Han, Y.; Guan, D.; Lee, S.; Lee, Y.Initial Training Data Selection for Active Learning. *ICUIMC '11: Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*; ACM, 2011; pp 1−7. Article No. 5.

https://doi.org/10.1021/acsomega.2c05145
ACS Omega 2023, 8, 2001−2009