# Efficient Search for Energetically Favorable Molecular Conformations against Metastable States via Gray-Box Optimization

AUTHOR(S):

Terayama, Kei; Sumita, Masato; Katouda, Michio; Tsuda, Koji; Okuno, Yasushi

CITATION:

Terayama, Kei ...[et al]. Efficient Search for Energetically Favorable Molecular Conformations against Metastable States via Gray-Box Optimization. Journal of Chemical Theory and Computation 2021, 17(8): 5419-5427

ISSUE DATE:

2021-08

URL:

http://hdl.handle.net/2433/277287

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

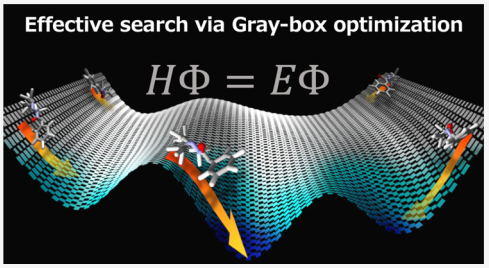
# Efficient Search for Energetically Favorable Molecular Conformations against Metastable States via Gray-Box Optimization

Kei Terayama,* Masato Sumita, Michio Katouda, Koji Tsuda, and Yasushi Okuno*

Cite This: *J. Chem. Theory Comput.* 2021, 17, 5419−5427 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** In order to accurately understand and estimate molecular properties, finding energetically favorable molecular conformations is the most fundamental task for atomistic computational research on molecules and materials. Geometry optimization based on quantum chemical calculations has enabled the conformation prediction of arbitrary molecules, including *de novo* ones. However, it is computationally expensive to perform geometry optimizations for enormous conformers. In this study, we introduce the gray-box optimization (GBO) framework, which enables optimal control over the entire geometry optimization process, among multiple conformers. Algorithms designed for GBO roughly estimate energetically preferable conformers during their geometry optimization iterations. They then preferentially compute promising conformers. To evaluate the performance of the GBO framework, we applied it to a test set consisting of seven dipeptides and mycophenolic acid to determine their stable conformations at the density functional theory level. We thus preferentially obtained energetically favorable conformations. Furthermore, the computational costs required to find the most stable conformation were significantly reduced (approximately 1% on average, compared to the naive approach for the dipeptides).




## INTRODUCTION

Finding energetically favorable molecular conformations is essential for understanding molecules' chemical and physical properties (e.g., reactivity, catalytic activity, or optical properties) and for exploiting design of novel molecules. Recent developments in various computational chemistry methods, combined with computational advances, have enabled the generation and prediction of molecular conformations.[1,2] In particular, search methods using quantum mechanics (QM) calculation-based geometry optimization and search algorithms have been developed.[3,4] In these methods, density functional theory (DFT)[5] is widely used because of its ease of use, although it is necessary to consider the effects of incorporation, such as long-range interactions, depending on the target molecule. Optimization and search algorithms include simulated annealing,[6] Monte Carlo-minimization,[7] particle swarm optimization,[8] basin hopping,[9] genetic/evolutionary algorithms,[3] Bayesian optimization,[10,11] and firefly algorithm;[12] others methods have also been utilized. These inductive approaches based on QM calculations are essentially applicable to any molecule, and they are expected to provide more detailed information than data-driven approaches,[13−15] which use conformation databases and machine learning. This property is particularly useful in molecule design and can be combined with recently developed molecule generation methods based on deep learning,[16−19] in which molecules that do not yet exist in reality are generated.

Despite various methodological developments in the conformation search method, however, it is still difficult to find energetically favorable conformations based on expensive QM calculations because there are large numbers of locally stable conformations with relatively high energies. This hinders access to the exact stable conformation. The basic strategy of the conformation search method based on QM calculations is to generate a large number of various conformation candidates (conformers) and then perform geometry optimizations of them, one by one, through QM calculations. However, compared to the number of energetically favorable conformations, the number of conformations with higher energies increases as the number of degeneracies increases. Assuming that the numbers of rotatable bonds are independent of each other, the probable energy could distribute as a canonical distribution.[20] Here, the probability of energy would distribute around the center where the number of states is the highest point because temperature is not considered. Therefore, geometry optimizations are more likely to be trapped in conformations with higher energies. As a result, most of the computational cost of geometry optimization is spent on finding energetically unfavorable conformations.

**Received:** March 27, 2021
**Published:** July 15, 2021







https://doi.org/10.1021/acs.jctc.1c00301
*J. Chem. Theory Comput.* 2021, 17, 5419−5427

To circumvent this issue, we introduce the gray-box optimization[21−23] (GBO) framework, which has been actively studied in the context of best arm identification (a special case of reinforcement learning) and has been used for hyperparameter optimization in deep learning.[24] From the viewpoint of search algorithms, finding energetically favorable conformations can be regarded as an optimization problem of a black-box function (black-box optimization)[25,26] by considering the QM-based geometry optimization as a complex function (black-box function). Then, finding energetically favorable conformations can be formulated as a problem to find input conformers that output favorable conformations for a given black-box function. However, the issue of computational cost, as described above, is inevitable. Here, we employed the GBO framework, which provides efficient algorithms for solving such black-box optimization problems while gradually evaluating the given black-box function. GBO is useful when many candidates (e.g., initial conformers) need to be evaluated and their evaluation values are gradually fixed (e.g., geometry optimization, where the final energy gradually converges according to the optimization iteration). GBO repeats two phases: (1) selection of the next candidate to be calculated and (2) slight advancement of the calculation for the selected candidate. This achieves an efficient search by gradually decreasing the number of promising candidates and concentrating their computational resources while considering all candidates.

In this study, we assess the performances of four GBO algorithms: LAQA (Look Ahead based on Quadratic Approximation),[27] sLAQA (sequential LAQA), successive rejects[28] (SR), and successive halving[21] (SH). LAQA and sLAQA were proposed by some of the authors for the efficient crystal structure prediction method.[27] While most GBO algorithms, including SR and SH, consider only the score as an indicator (energy, in this study), LAQA and sLAQA use the differential of the score (i.e., force in geometry optimization) to estimate promising candidates. To validate the effectiveness of the GBO algorithms, we prepared a test set of seven dipeptides and mycophenolic acid (MPA), which is used as a drug. Using the GBO algorithms, we have succeeded in obtaining energetically favorable conformations, including the most stable structures by reducing the total number of iterations required for geometry optimization based on DFT calculations, although the exhaustive search tended to produce many conformations with relatively high energies. In particular, the total number of iterations required for geometry optimization by LAQA to obtain the most stable structure was reduced to approximately 1% on average, compared to the naively exhaustive computation; it was reduced to approximately 10% compared to a random search. Our implementation is available on GitHub at http://github.com/inter-info-lab/chem_laqa. Users can easily set up parameters (e.g., computational resources) and quantum computation modules and can search conformations using the GBO algorithms with the standard database format (SDF) file of a target molecule.

## METHODS

Figure 1 shows an overview of a molecular conformation search based on exhaustive relaxation (ER) and GBO. First, a large number of initial conformers are generated for a target molecule. Subsequently, these conformers are relaxed at the molecular mechanics (MM) level. In the ER approach, all conformers are relaxed based on QM calculation, as shown on the left side of Figure 1. In the GBO approach, energetically favorable conformers are preferentially relaxed by controlling the geometry optimization. Details of these procedures are given below.

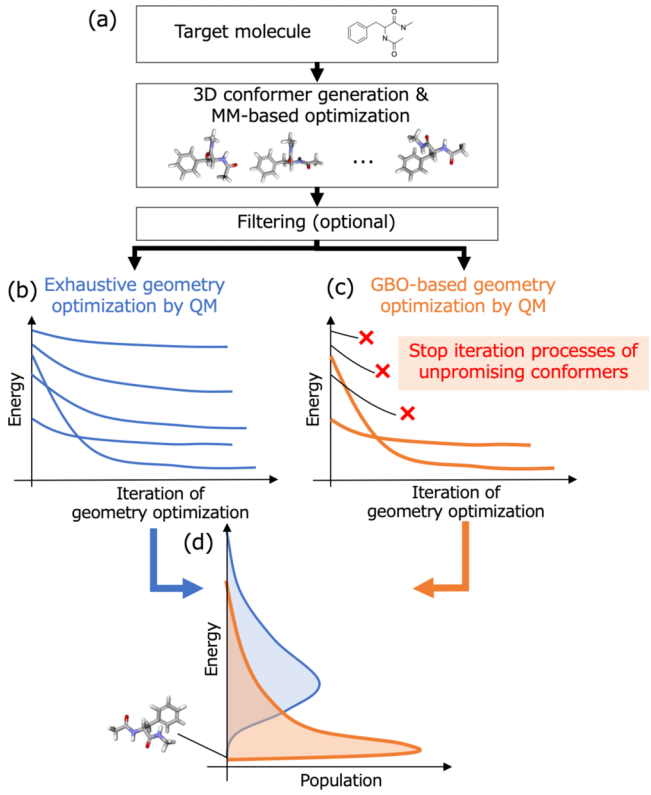



**Figure 1.** Overviews of ER and GBO-based conformation searches. (a) Number of various initial conformers are generated; these conformers are relaxed at the MM level. (b) Relaxed conformers with various energies are obtained by exhaustively relaxing all of the generated conformers at the QM level [blue distribution in (d)]. (c) GBO-based approach preferentially relaxes energetically favorable conformers by controlling the geometry optimization, and consequently, relatively low-energy conformers [orange distribution in (d)] are intensively obtained.

**Generation of Initial Conformers.** Initial conformers of a target molecule were generated from its simplified molecular-input line-entry system string (SMILES) representation or SDF file using modules in the fafoom code.[3] In the modules, first, a random three-dimensional structure was generated directly from the input SMILES using the ETKDG[29] method included in RDKit.[30] This structure was used as a template to generate the geometries of the conformers. Next, rotatable dihedral and *cis*/*trans* bonds were analyzed from SMILES and picked up as variables of the coordinates of the conformers. Geometries of the conformers are generated by assigning random values to the variables. Then, structural relaxations are performed for the generated conformers at the MM level described by the universal force field (UFF).[31] The trial of conformer structure generation is conducted until a certain number of times and added to the succeeded list. Conformationally different initial molecules often converge into the same conformer through relaxations at the MM level. Hence, we optionally performed filtered relaxed conformers and selected representative conformers based on the root-mean-square deviation (rmsd) values of heavy atoms with certain values ($d_{mm}$).


https://doi.org/10.1021/acs.jctc.1c00301
J. Chem. Theory Comput. 2021, 17, 5419−5427

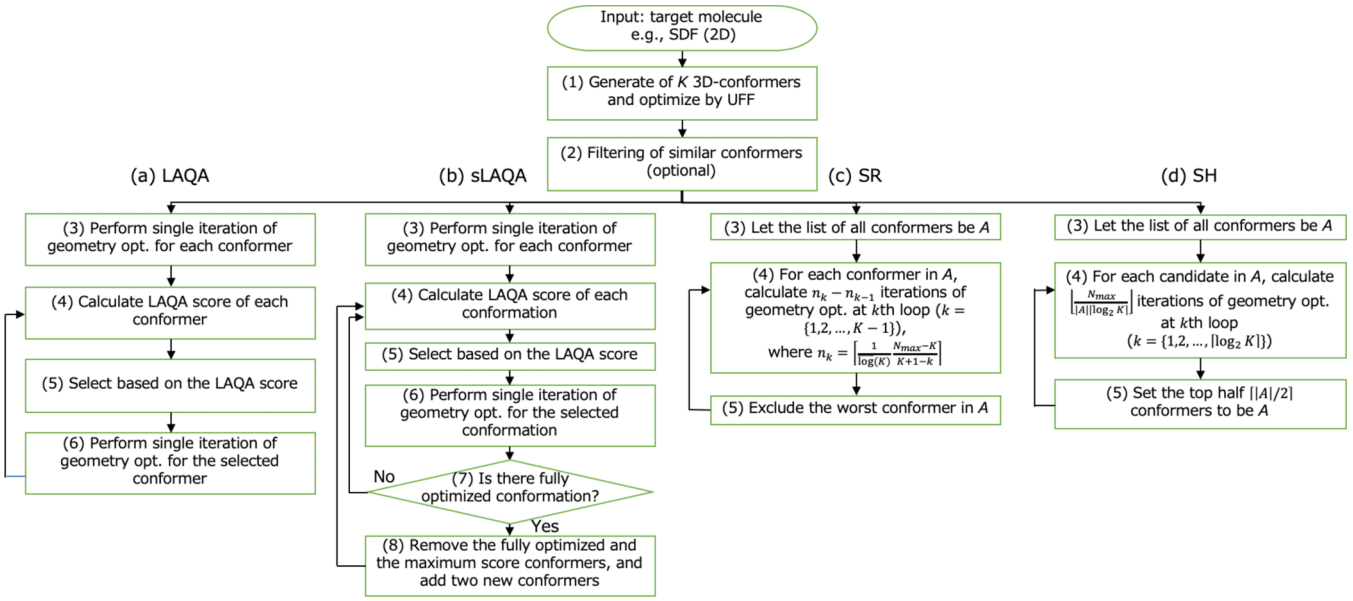



**Figure 2.** Flowcharts of four GBO-based algorithms: LAQA (a), sLAQA (b), SH (c), and SR (c). The step numbers assigned for each algorithm are shown in parentheses. In (4) of (a,b), we initially calculate the LAQA score for each conformer. Thereafter, the LAQA score is updated for the conformer on which the geometry optimization has been performed.

**Gray-Box Optimization.** Here, we introduce four GBO algorithms: LAQA, sLAQA, SH, and SR. Figure 2 shows flowcharts of these algorithms. All of these algorithms require the total number of iterations in the geometry optimization steps, $N_{max}$, as an input parameter. First, the initial conformers are generated and their relaxations are performed at the MM level, as described above (step 1). Then, the filtering based on the structural similarity is optionally conducted (step 2). In the results section, we assess the effect of this filtering. For LAQA (a), only a single iteration of QM-based geometry optimization is performed for each conformer (step 3). Then, we calculate the following score $L_{i,T}$ for each conformer $i$ (step 4), select the conformer with the best score (step 5), and perform a single iteration of geometry optimization (step 6).

$$L_{i,T} = \begin{cases} \min_{1 \le t \le T} E_{i,t} - \dfrac{{F_{i,T}}^2}{2\Delta F_{i,T}} & \text{(not optimized)} \\ \infty & \text{(fully optimized)} \end{cases} \quad (1)$$

Here, $T$ is the iteration number of the geometry optimization for candidate $i$, $E_{i,T}$ and $F_{i,T}$ are the energy and force to the nucleus at iteration number $T$, respectively, and $\Delta F_{i,T} = |F_{i,T} - F_{i,T-1}|$. $F_{i,T}$ is the mean of the norms of the force to all atoms. When $T = 1$, we fix $\Delta F_{i,T} = 1$. The score of eq 1 corresponds to a quadratic approximation of the final energy that is predicted from the current energy and force.[27] A conformer with a large force is expected to undergo a large energetic decrease in the future. By using this score, we can roughly estimate the energetic decrement of current conformers. In LAQA, steps 4−6 are repeated until sufficient optimized conformers are obtained or until the total number of iterations for geometry optimization reached the value of $N_{max}$. As LAQA runs steps 4−6 while checking all conformers, it is necessary to keep the information used to calculate all conformers, which can require considerably large storage. In sLAQA (b), some of the conformers are pooled as a conformer set for calculation ($N_{pool}$), and the calculation processes (steps 4−6) of LAQA are performed in the pooled conformers. After calculation in step 6, if the conformer is fully relaxed, sLAQA removes the conformer and the highest-energy one from the pool and adds two new conformers from the remaining ones in the prepared conformers (step 7). This operation allows us to perform geometry optimization while considering the overall conformers prepared up to step 3. LAQA and sLAQA are sequential methods, that is, they never advance the iteration of geometry optimization more than one conformer at a time.

In the case of LAQA, the score of eq 1 is calculated for all the candidate conformers, and then, the best one is selected for geometry optimization. On the other hand, SR[28] and SH[21] gradually reduce the number of candidates. In the geometry optimization step, SR and SH proceed with a designated number of iterations with all the retained candidates. The difference between SR and SH is the reduction procedure of candidate conformers: SR reduces the number of candidates one by one, and SH reduces them by half each time. The details of SR and SH are described below. Here, we assume that the number of initial conformers is $K$. In both SR and SH, let the list of all conformers be $A$ [step 3 in (c) and (d)]. For SR, the calculation of geometry optimization (step 4 in (c)) and selection of conformer (step 5 in (c)) are repeated $K - 1$ times. In the $k$th loop ($k \in \{1, 2, \ldots, K - 1\}$), for each conformer in $A$, we calculate $n_k - n_{k-1}$ iterations of geometry optimization, where $n_k = \left\lceil \frac{1}{\overline{\log}(K)} \frac{N_{max} - K}{K + 1 - k} \right\rceil$, $\overline{\log}(K) = \frac{1}{2} + \sum_{i=1}^{K} \frac{1}{i}$, and $n_0 = 0$. Then, in step 5, the candidate with the worst score in $A$ is excluded. In both SH and SR, the minimum energy is used as the score for each conformer. For SH, $\lceil \log_2 K \rceil$ iterations of steps 4 and 5 are performed. In the $k$th loop ($k \in \{0, 1, \ldots, \lceil \log_2 K \rceil\}$), for each conformer in $A$, we calculate $\left\lfloor \frac{N_{max}}{|A| \lceil \log_2 K \rceil} \right\rfloor$ iterations of geometry optimization. Then, in step 5, we set the top half $\lceil |A|/2 \rceil$ conformers in $A$ to be $A$. SR and SH are well designed such that the total number of iterations for geometry optimization



https://doi.org/10.1021/acs.jctc.1c00301
J. Chem. Theory Comput. 2021, 17, 5419−5427

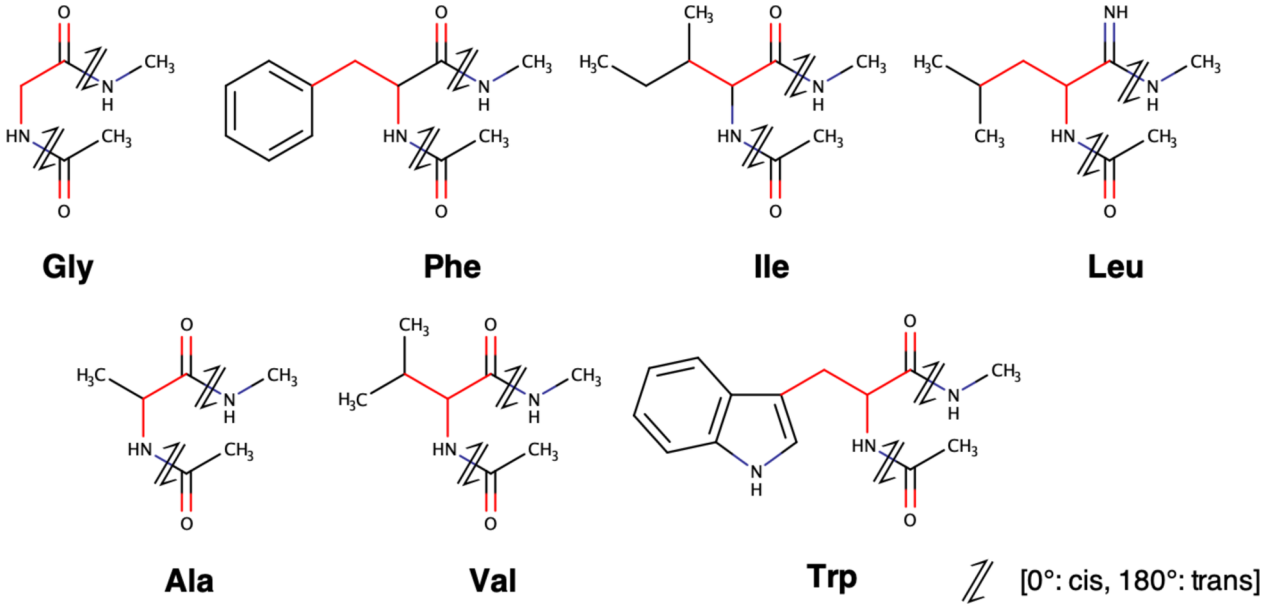



**Figure 3.** Structures of the seven amino acid dipeptides in the data set. The rotatable bonds are indicated in red. Double arrows indicate the *cis*/*trans* bonds.

**Table 1. Basic Information of the Seven Dipeptides Used to Evaluate the GBO-Based Conformation Search**

| | glycine | alanine | valine | isoleucine | leucine | phenylalanine | tryptophan |
|---|---|---|---|---|---|---|---|
| abbreviations | Gly | Ala | Val | Ile | Leu | Phe | Trp |
| num. of atoms | 19 | 22 | 28 | 31 | 31 | 32 | 36 |
| num. of rotatable bonds | 2 | 2 | 3 | 4 | 4 | 4 | 4 |
| num. of *cis*/*trans* bonds | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| initial conformers | 564 | 644 | 1358 | 1123 | 1699 | 2005 | 1875 |
| num. of succeeded DFT calculations | 556 | 644 | 1340 | 1110 | 671 | 1977 | 1875 |
| total geometry optimization steps | 25,589 | 24,304 | 56,859 | 50,448 | 74,659 | 90,365 | 92,367 |
| average geometry optimization steps | 46.02 | 37.74 | 42.43 | 45.45 | 44.68 | 45.73 | 49.26 |
| most stable structures (%) | 54 (9.71) | 23 (3.56) | 45 (3.58) | 32 (2.88) | 13 (0.78) | 40 (2.02) | 6 (0.32) |
| unique conformers (UFF optimized, dmm = 0.2) | 47 | 45 | 84 | 240 | 255 | 145 | 249 |
| unique conformers (DFT-optimized, dmm = 0.2) | 22 | 31 | 68 | 173 | 212 | 89 | 175 |

does not exceed $N_{max}$. In this study, there was often a large difference between $N_{max}$ and the total number of iterations after the calculation in steps 4−5 because a conformer may be fully relaxed in fewer steps than in step 5. In such cases, we calculated the scores in step 4 for conformers that are not fully relaxed and additionally performed geometry optimization starting from the candidate with the lowest score, as long as the total number of iterations for geometry optimization does not exceed $N_{max}$.

**Geometry Optimization by DFT.** All DFT calculations were performed with the Gaussian 16 RevA.03 program.[32] The wB97XD functional,[33] which includes empirical dispersion interactions, was used with the ultrafine numerical integration grid. The 6-31G(d)[34−36] basis functions were used as the spherical Gaussian basis functions. The hybrid geometry (Berny) optimization algorithm,[37] which is the default algorithm in Gaussian 16 employing the energy-represented direct inversion in the iterative subspace (GEDIIS) search[37] and the rational function optimization/linear search,[38] was used with the tight convergence criterion (i.e., opt = tight option). The convergence is judged by the values of the maximum force, RMS force, maximum displacement, and RMS displacement whose respective thresholds were set to less than $4.5 \times 10^{-4}$, $1 \times 10^{-5}$, $1.8 \times 10^{-3}$, and $1.2 \times 10^{-2}$ a.u. We performed the vibrational analysis for confirming that the most stable conformer lies at a stationary point.

**Data Set and Code Implementation.** To evaluate the performance of the implemented algorithm, we used two types of reference data that have been extensively investigated in previous simulation studies. The first reference data set is seven amino acid dipeptides (glycine, alanine, valine, isoleucine, leucine, phenylalanine, and tryptophan, Figure 3) extracted from a database of computational data for amino acid dipeptides.[3,39] The properties of the dipeptide data set (number of atoms, number of rotatable bonds, number of *cis*/*trans* bonds, statistical summary of initial conformer generation, and DFT geometry optimization of all generated conformers) are summarized in Table 1. The second reference data set is MPA (Figure S1), which is present in the collection of X-ray crystal structures of complexes containing ligands from the Protein Data Bank (PDB),[40] and its conformers were obtained with three different conformation search techniques:[3] (1) a genetic algorithm with fafoom, (2) random search with fafoom, and (3) systematic search with Confab.[41] MPA (target protein: 1MEH) has 43 atoms, eight rotatable bonds, and one *cis*/*trans* double bond. This molecule is very flexible and presents a challenging example for a conformation search. If a coarse systematic grid search is performed for six grids (every


https://doi.org/10.1021/acs.jctc.1c00301
J. Chem. Theory Comput. 2021, 17, 5419−5427

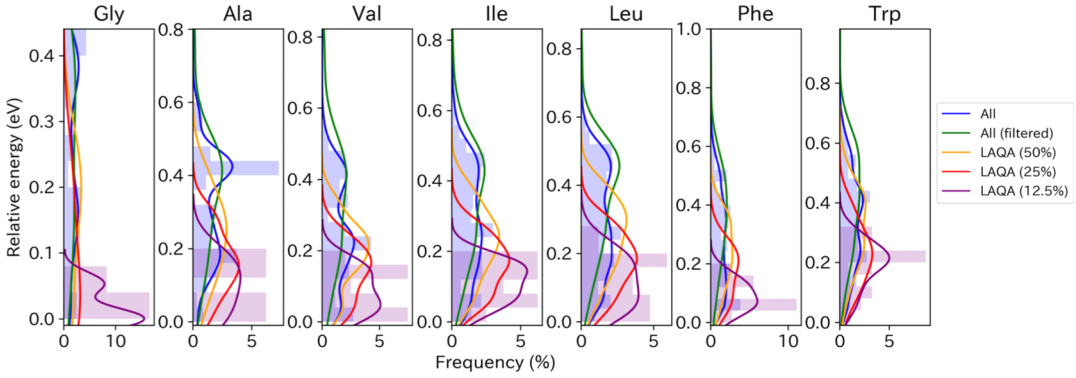



**Figure 4.** Relative energy distributions of optimized conformers determined by the LAQA algorithm and exhaustive geometry optimizations for the dipeptide data set. The relative energies to the energy of the most stable structure in each system were calculated. Each blue bar and line shows a histogram of the relative energies optimized by DFT for all generated conformers and their probability density functions estimated by KDE, respectively. The green lines show the probability density functions estimated by KDE for the filtered conformers. The yellow, red, and blue lines show the probability density functions obtained by applying LAQA for the total number of geometry optimization steps of 50, 25, and 12.5% for the total number of steps for the filtered structures, respectively. The purple histograms represent the energy distributions of the relaxed conformers using LAQA (12.5%). The total numbers of all structures and the filtered structures are shown in Figure 5.

60°) for six rotatable torsion dihedrals, two patterns of one *cis*/*trans* double bond, and two patterns of two X−X−O−H torsions, 66 × 2 × 2 × 2 = 373, 248 conformers should be tested. The properties of the MPA data set (number of atoms, number of rotatable bonds, number of *cis*/*trans* bonds, statistical summary of initial conformer generation, and DFT geometry optimization of all generated conformers) are summarized in Table S1. We also show the lowest energy conformer of each dipeptide in Figure S2 of Supporting Information. All the DFT-optimized conformers are available in Supporting Information (Optimized_conformers_SI.zip).

The GBO-based geometry optimization code was implemented using Python with the numpy and rdkit libraries. The initial random geometry generation code was implemented by modifying modules in the fafoom code using Python.

## RESULTS AND DISCUSSION

**Result of GBO-Based Conformation Search for the Dipeptide Data Set.** For the dipeptide data set (Table 1), we evaluated the effectiveness of LAQA in searching for energetically favorable conformations (Figure 4). First, we performed ERs using DFT calculations for all of the generated conformers. Details of the number of iterations required for geometry optimization and the number of unique conformers are listed in Table 1. The blue histograms and lines in Figure 4 show the energy distributions of the relaxed conformers and the probability density functions estimated by the kernel density estimation[42] (KDE) method, respectively. Moreover, we exhaustively optimized the filtered conformers (see the Methods section for details). The estimated probability density functions are shown as green lines in Figure 4. In Figure S3, we show the details of the energy distributions of the relaxed conformers as blue (all conformers) and green lines. From Figures 4 and S3, it can be seen that there were a large number of suboptimal conformers, and many energetically unfavorable structures were obtained upon exhaustively relaxing the generated structures.

The yellow, red, and blue lines in Figure 4 show the energy distributions obtained by LAQA for the total number of iterations required for geometry optimization of 50, 25, and 12.5% of the total number of iterations for the filtered conformers (green lines), respectively. The purple histograms represent the energy distributions of the relaxed conformers using LAQA (12.5%). The details of the energy distributions are shown as yellow, red, and blue lines in Figure S3. As an example, the improved relaxation process for glycine is shown in Figure S4. From Figures 4 and S3, the LAQA results show that energetically favorable conformers could be preferentially relaxed as the parameter for the total number of iterations required for geometry optimization was reduced. The probable distribution of energy could naively be the canonical distribution through nonbias sampling, such as random sampling. Hence, the probable of energy would distribute around the center where the number of states is the largest point because we did not consider the temperature. However, the probable of energy sampled by LAQA produced a distribution similar to that of the probable of states. This means that LAQA can search stable structures effectively against the distribution of a number of states. Additionally, the most stable structure of each peptide was found in most cases, except for LAQA (12.5%) for tryptophan. Tryptophan can be regarded as the most difficult case among the dipeptide data set because its population of conformers appears at much higher energy levels (the main population of conformers is located at 0.2 eV above the most stable structures) than the other six dipeptides as shown in Figure 4. Conformer searches of such kinds of examples often tend to be trapped in the local minimum conformations in the main population and make it difficult to find the most stable structure without conducting a large number of search steps. By contrast, as shown in these results, LAQA was able to dramatically reduce the number of calculations while preferentially relaxing conformers with low energy, including the most stable conformation. The vibrational analysis shows that the most stable conformers of the dipeptides obtained by LAQA lie at minima except for Ile. The most stable conformer of Ile lies at the saddle point for the rotation of a terminal methyl with a small imaginary frequency (21i).

We also examined the effect of the filtering on the final conformations obtained in detail. The numbers of unique relaxed structures for all of the generated structures are listed in Table 1; these numbers were similar to, or higher than, those reported in a previous study.[3] Figure S5 summarizes the conformers obtained when the threshold for filtering was


https://doi.org/10.1021/acs.jctc.1c00301
J. Chem. Theory Comput. 2021, 17, 5419−5427

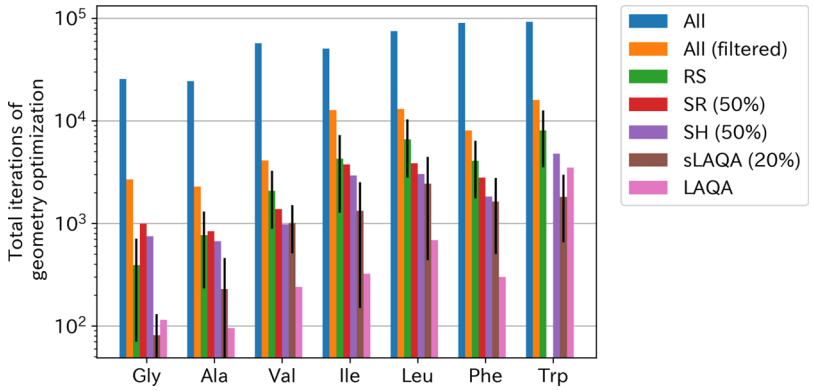



**Figure 5.** Total number of iterations required to find the most stable conformation of each test with each GBO algorithm. The blue bars represent the total number of iterations of geometry optimization required to relax all of the generated initial conformers. The orange bars indicate the total number of iterations for the filtered conformers. RS, LAQA, sLAQA, SH, and SR were applied to the conformers after filtering. The green bars show the expected values of the total number of iterations of geometry optimization required to find the most stable conformation at least one using RS. The black bars denote the standard deviations. The red, purple, and pink bars show the total number of iterations required to find at least one most stable structure by SR, SH, and LAQA, respectively. The blown bars indicate the expected values of total iteration numbers required to find at least one of the most stable conformations. For SR (50%) of Trp, the bar is not shown because the algorithm could not find the most stable structure under this parameter.

changed from 0.2 to 2.0 × $10^5$. When the threshold value was low, both the number of conformers obtained after the filtering and the final number of unique conformers increased. In particular, when the threshold value was set to 0.02, the average number of unique conformers was 96.2%, reducing the number of conformation candidates by the filter to approximately 15.1%. The green lines in Figure 4 show the distributions when the threshold was set to 0.02. This result indicates that, although there were some omissions, it is possible to obtain sufficient conformations while reducing the number of conformers by filtering the conformers relaxed by UFF.

**Performance for Finding the Most Stable Conformation.** To verify the performance of the GBO algorithms, including LAQA, we evaluated the total number of iterations of geometry optimization required to find the most stable structure. In Figure 5, the blue and orange bars show the total number of iterations required to relax all of the generated conformers and the filtered ones as baselines, respectively. Figure 5 also shows the total number of iterations required to find the most stable structures using the GBO-based approaches. We summarize the obtained conformers by the GBO algorithms in Figure S2. To evaluate the performances of the GBO algorithms, we conducted conformation exploration using RS, which is a method that repeats a procedure consisting of a random selection of a conformer and its full relaxation. We applied the GBO algorithms and RS to the filtered structures. Figure 5 shows the results of sLAQA, SH, and SR with 20, 50, and 50% of steps for the total number of iterations of the filtered structures as inputs. As sLAQA and RS are stochastic algorithms, we performed 50 trials by changing the random seed and calculated the expected values and standard deviations of the total number of iterations of geometry optimization required to find a stable conformation at least once. Figure 5 shows that the GBO algorithms achieved better performances compared to RS in most cases and drastically reduced the computational cost of finding the most stable structure compared to naive ERs. In particular, LAQA reduced the average number of steps to 0.989% compared to the ERs (blue bars). For tryptophan, which has a relatively complex structure, the reduction rate achieved by LAQA was relatively low (3.78%). The efficiency of these GBO algorithms compared to RS or exhaustive search is considered to be due to the fact that computations for energetically unfavorable structures can be omitted. In Figure S7, we also show the relationship between total number of iterations in geometry optimization and molecular weights. From the plot, the total number of iterations ($T_{it}$) scales with molecular weight (MW) as $T_{it} = 3.62 \times 10^{0.0107\times MW}$. These results show the effectiveness of the GBO algorithms, especially LAQA.

Among the GBO algorithms, as shown in Figure 5, LAQA showed the largest reductions, followed by sLAQA, SH, and SR. To analyze this tendency in detail, we have summarized the performances of sLAQA, SH, and SR under different input parameters in Figure S6. sLAQA has the advantage that the number of structures allocated in memory can be smaller than that of LAQA. The sLAQA results shown in Figure S6 reveal that, although there were some exceptions, the number of iterations required to find the most stable conformation could be reduced as the size of the $K_{pool}$ increased. As the cost (in memory) of keeping the computational information increases when $K_{pool}$ is large, it is necessary to set an appropriate number for $K_{pool}$ according to the computational environment and the target molecule. For SH and SR, it is confirmed that the required number of iterations could be reduced as the number of $N_{max}$ increases, as shown in Figure S6. On the other hand, when $N_{max}$ is small, the search for the most stable conformation may fail. For example, the case of SR (50%) for tryptophan in Figure 5 actually failed. The calculation process for this case is shown in Figure S7. As the optimization process of the most stable structure (black line) had a relatively high energy in the middle, its optimization was aborted by the algorithm shown in the dashed box in Figure S7. When $N_{max}$ of SR was 75% for tryptophan, the search was successful, as shown in S5. In addition, comparing SH and SR reveals that SH has a higher reduction rate. Overall, these results suggest that SH with a relatively large $N_{max}$ value enables an efficient search when computations can be executed in parallel.

**Result of the GBO Approaches for MPA.** We also evaluated the performances of the GBO algorithms for MPA as a more complex example. We generated 1000 initial candidate


https://doi.org/10.1021/acs.jctc.1c00301
J. Chem. Theory Comput. 2021, 17, 5419−5427

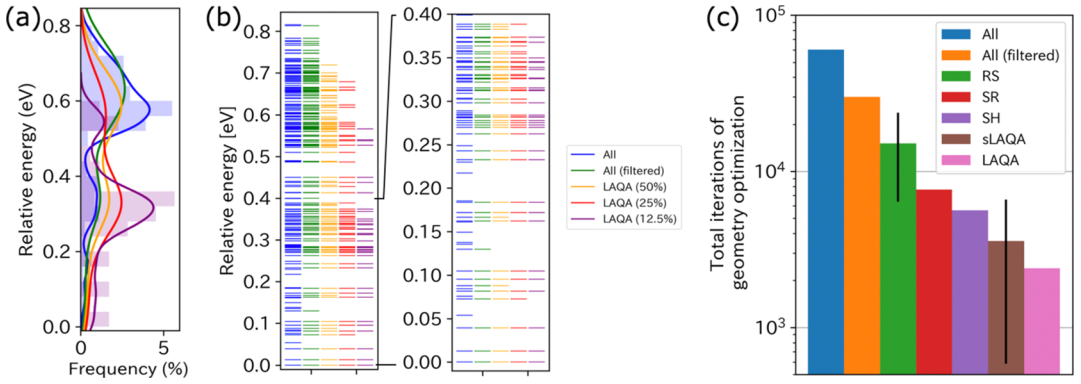



**Figure 6.** Results of the conformation search by the GBO algorithms and an exhaustive search for MPA. (a) Estimated probability density functions of the relaxed conformers' energies obtained by an exhaustive search for all of the generated conformers (blue) and filtered ones (green), LAQA with 50% (yellow), 25% (red), and 12.5% (purple) of the total number of iterations of the filtered conformers. The blue and red bars show histograms of the relaxed conformers' energies obtained by the exhaustive search for all generated conformers and the LAQA (12.5%) search. (b) Detailed energy distributions of relaxed conformers using the exhaustive search and LAQAs. The colors correspond to those shown in (a). (c) Total number of iterations required to find the most stable conformation of each test with each GBO algorithm. The black bars indicate the standard deviations.

conformers using fafoom and succeeded in relaxing 992 conformers using DFT calculations. In Figure 6, the estimated probability density function [blue line in (a)] and energy distributions [blue lines in (b)] of all of the relaxed 992 conformers are shown. A large number of conformers were included within a difference of 1 eV from the stable conformation. Figure 6a clearly shows that LAQA preferentially relaxed structures with low energy (<0.4 eV), while many conformers with energies of 0.5 eV or higher were obtained in the ERs. Figure S8 shows the relationship between the RMSD and energy values of the relaxed conformers from the reported conformation.[40] The number of unique conformers optimized by DFT and filtered conformers were 333 and 483, respectively. Here, the threshold and filtering-threshold values were set to 0.02 and 0.00002, respectively, because the structure of MPA was complicated and there were many different structures, even though the RMSD threshold was small. The yellow, red, and purple lines in Figure 6a,b show the probability density functions and energy distributions of the conformers obtained when the total number of iterations for the geometry optimization were set to 50, 25, and 12.5%, respectively, of the steps required to relax all of the filtered conformers. Although some structures could not be found in the region with relatively low energy, we succeeded in discovering three structures using LAQA and confirmed that the most stable structure of MPA lies at a minimum through the vibrational analysis.

We evaluated the performances of the GBO algorithms to find the most stable structure of MPA (Figure 6c). Initially, 60,184 iterations (60.67 iterations on average) and 30,060 iterations were required to relax all 992 initial conformers and the filtered 483 conformers, respectively. The RS and GBO algorithms were applied to the filtered structure. LAQA showed the best performance (2397 steps) as shown in Figure 6c, and the other GBO algorithms also performed better than RS. This tendency was similar to that observed for the dipeptide data set. In particular, LAQA succeeded in searching for the most stable structure using only 3.9% of the number of iterations numbers compared to when all of the initial conformers were relaxed exhaustively.

## CONCLUSIONS

To effectively search for energetically favorable conformations, we introduced GBO algorithms LAQA, sLAQA, SH, and SR, which can finely control geometry optimization among multiple conformers. To validate the performances of these GBO algorithms, we applied them to seven dipeptides and MPA. The GBO algorithms were able to avoid making computations for energetically unfavorable conformers and could preferentially compute energetically favorable conformers, including the most stable structures. Consequently, these algorithms efficiently obtained energetically favorable conformations against a large number of metastable states. In particular, when using LAQA, the most stable structure was obtained using below 1% of the number of iterations of geometry optimization in most systems, compared to the exhaustive calculation of all generated conformers. The other GBO algorithms, sLAQA, SH, and SR, performed worse than LAQA in terms of searching for the most stable conformation. However, they performed better than RS and have other advantages compared to LAQA. The sLAQA approach requires fewer conformers in memory at a time, and both SH and SR can be parallelized. Therefore, we would benefit from these GBO algorithms by their appropriate use depending on the computational environment and target molecules.

These GBO algorithms are applicable to various molecules and computational methods because they are independent of specific molecules and computational details. As shown in Figures 5 and 6, if the size and complexity of a molecule increases, the number of possible conformations will dramatically increase, and thus, the required computational cost will increase too, even if the GBO algorithms were to be employed. If possible conformers that would reach the most stable conformation are not in an initial pool of conformers, the GBO algorithms fail to find the most stable one. On the other hand, systematic methods, such as genetic algorithms[3] and Monte Carlo-minimization,[7] can find the most stable structure even if such an initial structure is not prepared. The reliability of the GBO algorithms is elevated only by increasing the diversity of conformers that are prepared in the initial pool. In this study, we prepared a large number of conformers in


https://doi.org/10.1021/acs.jctc.1c00301
J. Chem. Theory Comput. 2021, 17, 5419−5427

advance using fafoom. Combining the GBO algorithms and conformer generation with optimization methods such as evolutionary algorithms and machine learning-based conformation predictions, the computational efficiency can be improved, especially for complex molecules. In this work, we used a minimum-energy search algorithm by evaluating molecular potential energy for demonstrating the GBO algorithms. However, the GBO algorithms would be applicable to search any stationary points such as a saddle point on any surfaces such as free energy surface as the application of the basin-hopping method.[43]

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00301.

Structure of MPA, lowest energy conformer of each dipeptide and the lowest energy conformers obtained by GOB algorithms, SR (50%), SH (50%), sLAQA (20%), and LAQA, energy distribution of optimized conformers, geometry optimization process for Gly by ER and LAQA, scaling of total iterations of geometry optimization with molecular weights of the dipeptides for the GBO algorithms, effect of filtering, details of the performances of sLAQA, SH, and SR with different input parameters, relaxation process of SR (50%) for tryptophan, relationship between RMSD and energy of the relaxed conformers, and MPA data set (PDF)

Data for optimized conformers (ZIP)

## ■ AUTHOR INFORMATION

### Corresponding Authors

**Kei Terayama** − *Graduate School of Medical Life Science, Yokohama City University, Yokohama 230-0045, Japan; RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan; Medical Sciences Innovation Hub Program, RIKEN, Yokohama 230-0045, Japan; Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan;* orcid.org/0000-0003-3914-248X; Phone: +81 (0)4 55087231; Email: terayama@yokohama-cu.ac.jp; Fax: +81 (0)4 55087367

**Yasushi Okuno** − *Medical Sciences Innovation Hub Program, RIKEN, Yokohama 230-0045, Japan; Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan;* Email: okuno.yasushi.4c@kyoto-u.ac.jp

### Authors

**Masato Sumita** − *RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan; International Center for Materials Nanoarchitectonics(WPI-MANA), National Institute for Materials Science, Tsukuba 305-0044, Japan*

**Michio Katouda** − *Department of Computational Science and Technology, Research Organization for Information Science and Technology, Tokyo 105-0013, Japan; Waseda Research Institute for Science and Engineering, Waseda University, Tokyo 169-8555, Japan*

**Koji Tsuda** − *Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-8561, Japan; Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba 305-0047, Japan; RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan;* orcid.org/0000-0002-4288-1606

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.1c00301

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This research was supported by the New Energy and Industrial Technology Development Organization (NEDO), Ministry of Education, Culture, Sports, Science and Technology (MEXT), as "Program for Promoting Researches on the Supercomputer Fugaku" (MD-driven Precision Medicine), Japan Agency for Medical Research and Development (AMED) under grant number JP20nk0101111, and Japan Society for the Promotion of Science (JSPS) under KAKENHI (Grant-in-Aid for Scientific Research (C): grant number 19K05443). The computations in this work were performed on the supercomputer system(s) at the Information Technology Center, Nagoya University, and at the Research Institute for Information Technology, Kyushu University.

## ■ REFERENCES

(1) Hawkins, P. C. D. Conformation generation: the state of the art. *J. Chem. Inf. Model.* **2017**, *57*, 1747−1756.
(2) Mansimov, E.; Mahmood, O.; Kang, S.; Cho, K. Molecular geometry prediction using a deep generative graph neural network. *Sci. Rep.* **2019**, *9*, 20381.
(3) Supady, A.; Blum, V.; Baldauf, C. First-principles molecular structure search with a genetic algorithm. *J. Chem. Inf. Model.* **2015**, *55*, 2338−2348.
(4) Chandramouli, B.; Del Galdo, S.; Fusè, M.; Barone, V.; Mancini, G. Two-level stochastic search of low-energy conformers for molecular spectroscopy: implementation and validation of MM and QM models. *Phys. Chem. Chem. Phys.* **2019**, *21*, 19921−19934.
(5) Parr, R. G. *Horizons of Quantum Chemistry*; Springer, 1980; pp 5−15.
(6) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by simulated annealing. *Science* **1983**, *220*, 671−680.
(7) Li, Z.; Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611−6615.
(8) Eberhart, R.; Kennedy, J. Particle Swarm Optimization. *Proceedings of the IEEE International Conference on Neural Networks*, 1995; pp 1942−1948.
(9) Wales, D. J.; Doye, J. P.; Miller, M. A.; Mortenson, P. N.; Walsh, T. R. Energy landscapes: from clusters to biomolecules. *Adv. Chem. Phys.* **2000**, *115*, 1−111.
(10) Chan, L.; Hutchison, G. R.; Morris, G. M. Bayesian optimization for conformer generation. *J. Cheminf.* **2019**, *11*, 32.
(11) 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.
(12) Mitra, A.; Jana, G.; Agrawal, P.; Sural, S.; Chattaraj, P. K. Integrating firefly algorithm with density functional theory for global optimization of $Al_4^{-2}$ clusters. *Theor. Chem. Acc.* **2020**, *139*, 32.
(13) Andronico, A.; Randall, A.; Benz, R. W.; Baldi, P. Data-driven high-throughput prediction of the 3-D structure of small molecules: review and progress. *J. Chem. Inf. Model.* **2011**, *51*, 760−776.
(14) Yoshikawa, N.; Hutchison, G. R. Fast, efficient fragment-based coordinate generation for Open Babel. *J. Cheminf.* **2019**, *11*, 49.
(15) Xu, M.; Luo, S.; Bengio, Y.; Peng, J.; Tang, J. Learning neural generative dynamics for molecular conformation generation. **2021**, arXiv preprint arXiv:2102.10240.

(16) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360−365.
(17) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828−849.
(18) Yang, X.; Zhang, J.; Yoshizoe, K.; Terayama, K.; Tsuda, K. ChemTS: an efficient python library for de novo molecular generation. *Sci. Technol. Adv. Mater.* **2017**, *18*, 972−976.
(19) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(20) MacQuarrie, D. A. *Statistical Mechanics*; University Science Books, 2000.
(21) Karnin, Z.; Koren, T.; Somekh, O. Almost Optimal Exploration in Multi-Armed Bandits. *International Conference on Machine Learning*, 2013; pp 1238−1246.
(22) Jamieson, K.; Talwalkar, A. Non-Stochastic Best Arm Identification and Hyperparameter Optimization. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016; pp 240−248.
(23) Feurer, M.; Hutter, F. In *Automated Machine Learning: Methods, Systems, Challenges*; Hutter, F., Kotthoff, L., Vanschoren, J., Eds.; Springer International Publishing: Cham, 2019; pp 3−33.
(24) Li, L.; Jamieson, K.; DeSalvo, G.; Rostamizadeh, A.; Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.* **2017**, *18*, 6765−6816.
(25) Jones, D. R.; Schonlau, M.; Welch, W. J. Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **1998**, *13*, 455−492.
(26) Terayama, K.; Sumita, M.; Tamura, R.; Tsuda, K. Black-Box Optimization for Automated Discovery. *Acc. Chem. Res.* **2021**, *54*, 1334−1346.
(27) Terayama, K.; Yamashita, T.; Oguchi, T.; Tsuda, K. Fine-grained optimization method for crystal structure prediction. *npj Comput. Mater.* **2018**, *4*, 32.
(28) Audibert, J.-Y.; Bubeck, S. Best Arm Identification in Multi-Armed Bandits. *Conference on Learning Theory*, Haifa, Israel, 2010; p 13.
(29) Riniker, S.; Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562−2574.
(30) Landrum, G.; et al. RDKit: Open-source cheminformatics. http://www.rdkit.org/ (accessed on 1 Dec, 2020).
(31) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.
(32) Frisch, M. J.; et al. *Gaussian 16*, Revision A.03; Gaussian Inc.: Wallingford CT, 2016.
(33) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom−atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615−6620.
(34) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules. *J. Chem. Phys.* **1971**, *54*, 724−728.
(35) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* **1972**, *56*, 2257−2261.
(36) Hariharan, P. C.; Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta* **1973**, *28*, 213−222.
(37) Li, X.; Frisch, M. J. Energy-represented direct inversion in the iterative subspace within a hybrid geometry optimization method. *J. Chem. Theory Comput.* **2006**, *2*, 835−839.
(38) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. Search for stationary points on surfaces. *J. Phys. Chem.* **1985**, *89*, 52−57.
(39) Ropo, M.; Schneider, M.; Baldauf, C.; Blum, V. First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Sci. Data* **2016**, *3*, 160009.
(40) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein- ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.
(41) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab-Systematic generation of diverse low-energy conformers. *J. Cheminf.* **2011**, *3*, 8.
(42) Silverman, B. W. *Density Estimation for Statistics and Data Analysis*; CRC press, 1986; Vol. *26*.
(43) Sutherland-Cash, K. H.; Wales, D. J.; Chakrabarti, D. Free energy basin-hopping. *Chem. Phys. Lett.* **2015**, *625*, 1−4.


https://doi.org/10.1021/acs.jctc.1c00301
*J. Chem. Theory Comput.* 2021, 17, 5419−5427