JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

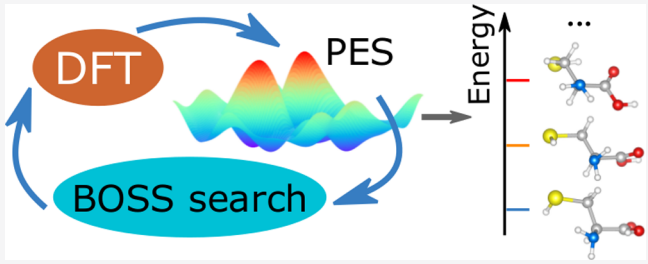
Efficient Amino Acid Conformer Search with Bayesian Optimization

Lincan Fang, Esko Makkonen, Milica Todorović, Patrick Rinke,* and Xi Chen*

Cite This: J. Chem. Theory Comput. 2021, 17, 1955−1966 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Finding low-energy molecular conformers is challenging due to the high dimensionality of the search space and the computational cost of accurate quantum chemical methods for determining conformer structures and energies. Here, we combine active-learning Bayesian optimization (BO) algorithms with quantum chemistry methods to address this challenge. Using cysteine as an example, we show that our procedure is both efficient and accurate. After only 1000 single-point calculations and approximately 80 structure relaxations, which is less than 10% computational cost of the current fastest method, we have found the low-energy conformers in good agreement with experimental measurements and reference calculations. To test the transferability of our method, we also repeated the conformer search of serine, tryptophan, and aspartic acid. The results agree well with previous conformer search studies.



INTRODUCTION

A molecular conformer is a distinct conformation corresponding to a minimum on the molecule's potential energy surface (PES). Any molecule with rotatable bonds has several stable conformer structures, each associated with different chemical and electronic properties. At ambient temperatures, all the properties of that molecule are the combination of the properties of its conformers accessible at the temperature of the study.[1−3] Therefore, identifying the low-energy conformers and determining their energy ranking continues to be a topic of great interest in computational chemistry,[4] cheminformatics,[5,6] computational drug design,[7] and structure-based virtual screening.[8] While one configuration of a small molecule can be simulated routinely by *ab initio* methods, the large size of configurational phase space and the considerable number of local minima in typical energy landscapes make conformer searches one of the persistent challenges in molecular modeling.[1,5,6]

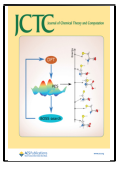
The first challenge in conformer search is sufficient sampling of the configurational space. The conformational space (bond lengths, bond angles, and torsions) for even relatively small molecules is enormous.[9,10] For this reason, dimensionality reduction is commonly applied to make the problem more tractable. Since the bond lengths and angles are relatively rigid in molecules and the different conformers originate from the flexible rotational groups, most search methods focus on sampling the torsion angles in molecules while keeping bond length and angles fixed.[1] A variety of methods and tools have been developed to generate diverse conformer structures.[11−16] These methods can be broadly classified to be either systematic or stochastic.

A systematic method relies on a grid to sample all the possible torsion angles in the molecule. This approach is deterministic but limited to small molecules because it scales poorly with increasing numbers of relevant torsion angles, i.e., search dimensions. Stochastic methods randomly sample the phase space of torsion angles (sometimes restricted to predefined, most relevant ranges) based on different algorithms such as Monte Carlo annealing,[17,18] minima hopping,[19] basin hopping,[20,21] distance geometry,[22] and genetic algorithms.[11,23] Stochastic methods can be applied to larger molecules with high-dimensional conformer spaces, but the predicted conformers may vary. Extensive sampling is required, and the results may be affected by the random nature of the process.

Knowledge-based methods have also been developed[24,25] to achieve more consistent results. They use a predefined library for torsion angles and ring conformations. The library is typically based on experimental structures in databases such as the Cambridge Structure Database (CSD)[26] or the Protein Data Bank (PDB).[27] To search the conformers, knowledge-based methods usually need to be combined with the different systematic or stochastic algorithms mentioned before.

The second challenge in conformer searches is the sufficiently accurate mapping of energies and structures. Two classes of total energy approaches are commonly used: force field-based methods and quantum chemistry methods such as the density functional theory (DFT) and coupled cluster (CC) theory. Quantum chemistry methods achieve higher accuracy in the

Received: June 24, 2020
Published: February 12, 2021



https://dx.doi.org/10.1021/acs.jctc.0c00648
J. Chem. Theory Comput. 2021, 17, 1955−1966

estimation of molecular energies than force fields because they describe the interactions and polarization in molecules more accurately. However, they are computationally costly. More often than not, quantum chemistry methods are too expensive to provide energies for all configurations produced in the search.

To balance efficiency and accuracy, hierarchical methods have been developed. Fast computational methods with lower accuracy are employed to to scan the configurational space. Promising candidate structures are then funneled through more costly methods with higher accuracy to refine the conformer structures and energies (such as force fields → DFT[28,29] or HF → MP2 → CCSD(T)[30]). Methods at different levels predict different PESs. To avoid missing the true low-energy conformers, a large portion of configurational space has to be sampled at a lower accuracy method level, and many structures need to be optimized at a higher level.

In recent years, artificial intelligence (AI) and machine learning (ML) techniques such as genetic algorithms,[31,32] artificial neural network,[33,34] Gaussian process regression (GPR),[35−37] and machine-learned force fields[38] were used to accelerate structure-to-energy predictions and geometry optimization for molecules. The majority of these schemes requires a large number of data points, which may be costly to compute with *ab initio* methods. To reduce the amount of required data, Bayesian optimization was introduced in the structure search.[39−41] Bayesian optimization search schemes belong to the active learning family of methods, which generate data on the fly for optimal knowledge gain.

In this article, we present a new procedure for molecular conformer identification and ranking. We combined the Bayesian optimization structure search (BOSS) approach[40] and quantum chemistry simulations to find the conformers of small molecules and accurately predict their relative stability. BOSS is a python-based tool for global phase space exploration based on Bayesian optimization.[42] Beyond the Bayesian active learning method for the global minimum conformer search in ref 39, our procedure aims to find all the relevant conformers in one run. We use cysteine as a model system to demonstrate our methodology and then later generalize to other amino acids.

Cysteine was chosen for several reasons. First, it is an amino acid with critical biological functions. Second, it is the only amino acid that has a −SH group. The strong S−Ag and S−Au bonds make it interesting for use in hybrid nanomaterials.[43,44] Third, cysteine has five rotational groups, as shown in Figure 1. Therefore it is an interesting and accessible five-dimensional (5D) system for Bayesian optimization. Last, the structures and the energy order of cysteine's conformers have been calculated by several groups using the grid sample method[30,45,46] and characterized by Fourier transform microwave spectroscopy experiments[47] so that we can compare the accuracy and efficiency of our new procedure with other computational and experimental results.

In brief, using cysteine as an example, we present an efficient and reliable procedure to predict the structures and energies of molecular conformers. BOSS ensures sufficient sampling of the configurational phase space and outputs the structures associated with local energy minima. We post-processed the machine-learned conformer candidates with geometry optimization and then added free energy corrections to obtain the final ranking. We tested the effect of different exchange-correlation functionals and van de Waals interactions on the ranking order. Finally, we applied coupled cluster corrections to the lowest-energy conformers.

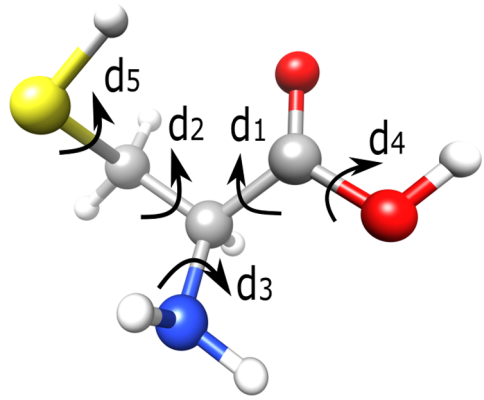


Figure 1. Ball-and-stick model of the cysteine molecule. Red is used for oxygen, white for hydrogen, gray for carbon, blue for nitrogen, and yellow for sulfur. d_1, d_2, d_3, d_4, and d_5 label the five dihedral angles of cysteine that we use to define our search space.

To test the generalizability and transferability of our method, we also studied the conformers of three other amino acids: tryptophan, serine, and aspartic acid. Serine and tryptophan have a five-dimensional phase space for our purposes, and aspartic acid has 6 rotational degrees of freedom. The methods and results will be presented in the following sections.

METHODS

BOSS-based Molecular Conformer Search. Our BOSS-based procedure for molecular conformer search contains four steps: (i) System preparation, (ii) Bayesian optimization conformer search, (iii) refinement, and (iv) validation, as illustrated in Figure 2a.

In step (i), we first obtain an xyz-file of our molecule of interest from a database and then perform a single geometry optimization with a quantum chemistry method. Then, we calculate the *z*-matrix to find the dihedral angles. We chose the dihedral angles d_n to describe the different conformers, as they are typically the most informative degrees of freedom for conformers. We keep all bond lengths and angles fixed at their optimized values. Such dimensionality reduction is standard practice to expedite the molecular conformer search, as mentioned in the Introduction.

In step (ii), BOSS actively learns the PES of the molecule by Bayesian optimization iterative data sampling. Each data "point" consists of the set of dihedral angles d_n for a molecular configuration and its corresponding total energy E. In this step, we use DFT as the calculator. E therefore corresponds to the DFT total energy of a molecular configuration.

BOSS employs Gaussian process (GP) models[48] to fit a surrogate PES to the data points, and then refines it by acquiring more data points at locations that minimize the exploratory lower confidence bound (eLCB) acquisition function.[42] The most-likely PES model for the given data is the GP posterior mean. The lack of confidence in the model is reflected by the GP posterior variance, which vanishes at the data points and rises in unexplored areas of phase space. The key concepts of this active learning approach are illustrated in Figure 2b, in which BOSS iteratively infers a one-dimensional PES of the d_1 dihedral angle of cysteine. The global minimum location and the entire PES are learned in 10 data acquisitions. In analogy with the 1D example, BOSS actively learns the PES in N dimensions until convergence is achieved. The advantage of BOSS is not only its efficiency but also the fact that it explores both the global minimum and local

https://dx.doi.org/10.1021/acs.jctc.0c00648
J. Chem. Theory Comput. 2021, 17, 1955−1966

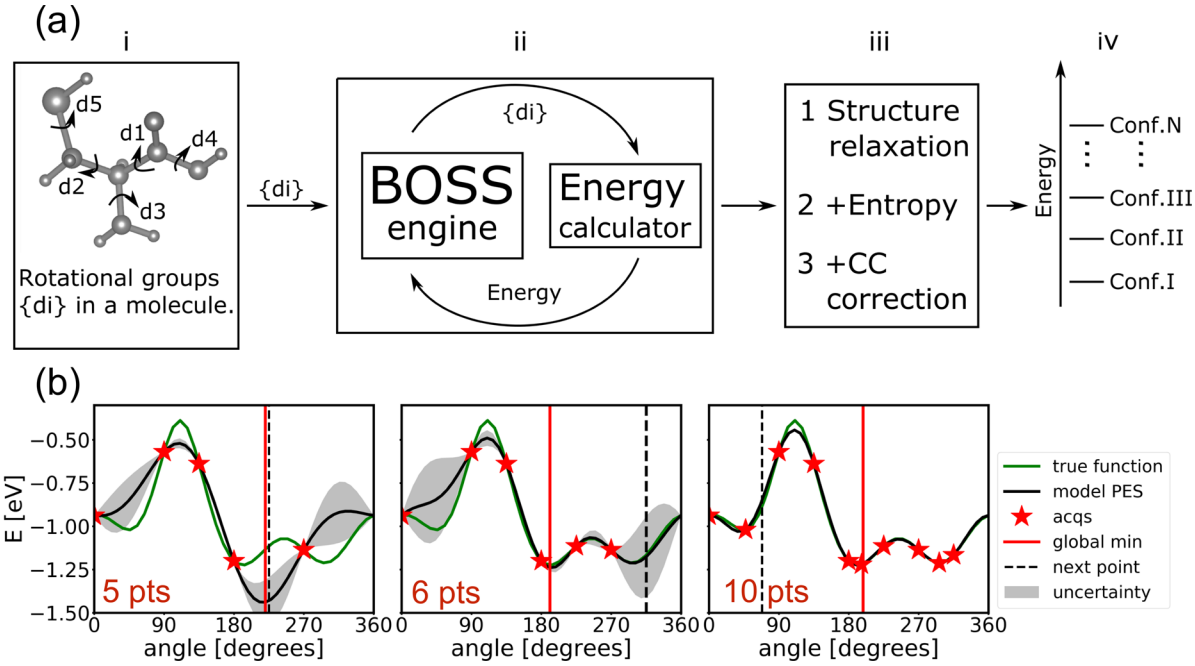


Figure 2. (a) Overview of our BOSS-based procedure for molecular conformer search, featuring (i) system preparation, (ii) Bayesian optimization conformer search, (iii) refinement, and (iv) validation. (b) BOSS iterative inference of a one-dimensional (1D) PES of the d_1 dihedral angle of cysteine. The GP's native uncertainty (gray areas) facilitates exploratory data sampling. The global minimum location and the entire PES are learned in 10 data acquisitions.

minima of the PES during the search. We exploit this feature to find conformers beyond the global minimum, which we associate with the local minima of the PES. A more detailed introduction of the BOSS approach can be found in refs 40, 49, and.[50]

The current BOSS implementation does not restrict the search space, which for rotational degrees of freedom may result in steric clashes. For example, for aspartic acid and tryptophan, BOSS occasionally samples physically non-meaningful structures with very high energies. Such energy spikes can obstruct model fitting and should be avoided. In this work, we refrain from restricting the search space and instead apply an energy transformation: $E_{new} = E_{cut} + \log_{10}(E)$. If the DFT energy E of a given structure is higher than E_{cut}, we damp it down by taking the logarithm. We tested E_{cut} = 1.0, 2.0, 3.0, and 4.0 eV for aspartic acid and found 2.0 eV to be optimal. BOSS hyper-parameters converge fastest for this E_{cut} value, and only 0.017% acquisitions needed to be transformed (Figure S7). We therefore adopted E_{cut} = 2 also for the other amino acids.

After the BOSS-predicted PES has converged, in step (iii), we analyzed the PES to extract the local minima locations and related structures. Since the PES and its gradients can be computed efficiently at any location in the N-dimensional phase space from the GP model, BOSS post-processing routines perform multiple L-BFGS (limited-memory Broyden−Fletcher−Goldfarb−Shanno algorithm) minimizations, using the locations of the data acquisitions as starting points. Because models built with more datapoints tend to be more complex and feature more minima, starting numerous minimisers from different points allows us to identify as many different minimum basins as possible in the PES surrogate model. This procedure potentially reports the same minima multiple times. For this reason, we developed automated duplicate purging routines to output only different minima after postprocessing (typically about 10% of all minima found). The resulting shortlist of minima may still contain similar structures, and the final pruning is left to the user, as required by the application.

Next, we refine the local minima output by BOSS by geometry optimization and entropy corrections. First, all degrees of freedom (including bond lengths and angles) are relaxed to obtain optimized structures and energies. Next, we add vibrational entropy corrections following previous studies.[51,52] We compute and add the zero-point energy and the vibrational free energy at 300 K to the energies of optimized conformers. Since most experiments are performed at room temperature, we picked a temperature of 300 K for the vibrational corrections.

In step (iii), we also go beyond DFT. We perform coupled cluster calculations for the DFT-optimized conformer structures in a relevant energy window. Coupled cluster (CC) theory is an approximate infinite-order perturbation theory, in the form of exponential cluster operators describing the quantum many-body effects of the electronic wave function. Despite being significantly more expensive than DFT and scaling polynomially with system size, CC theory provides a systematically improvable hierarchy of approximations for accurate energy predictions. Due to the high computational cost, we only apply the CC method to the low-energy conformers we are interested in. The difference between the coupled cluster and DFT total energy, here called CC correction, is then added to the entropy corrections we added earlier in step (iii).

In step (iv), we validate our results by comparing the low-energy conformers we found to experimental and other computational results. System preparation and final validation require human input, but procedures featuring structure search and refinement can be made fully automated into a computational workflow.

Computational Methods. In this work, we employed DFT as the predominant energy calculator and employed the all-electron code FHI-aims[53−55] for all DFT calculations. "Tight" numerical settings and "tier 2" basis sets were used throughout.

https://dx.doi.org/10.1021/acs.jctc.0c00648
J. Chem. Theory Comput. 2021, 17, 1955−1966

To investigate the influence of the exchange-correlation functional and the level of dispersion correction on the final results, we performed our conformer search with the PBE + TS,[56,57] PBE + MBD,[56,58] PBE0 + TS,[57,59] and the PBE0 + MBD[58,59] functionals. For geometry optimizations, the geometry was considered to be converged when the maximum residual force (fmax) was below 0.01 eV/Å. To ensure that this fmax setting is tight enough, we have performed test calculations with fmax = 0.0001 eV/Å. The root mean square (RMS) difference of all atomic coordinates is 0.00036 Å, and the energy difference is 0.000003 eV.

Vibrational free energies were computed using the finite-difference method within the harmonic approximation. We used a finite-difference displacement length of δ = 0.0025 Å. The vibrational free energy F_{vib} was then calculated as follows

$$F_{vib}(T) = \int d\omega g(\omega)\frac{\hbar\omega}{2} + \int d\omega g(\omega) k_B T \ln\left[1 - \exp\left(-\frac{\hbar\omega}{k_B T}\right)\right] \quad (1)$$

where $g(\omega)$ is the phonon density of states and T, ω and k_B are the temperature, frequency, and Boltzmann constant, respectively.

Going beyond DFT, we performed CC calculations with single, double, and perturbative triple excitations (CCSD(T)). These were done as single-point calculations using the structures from the PBE0 + MBD calculation with aug-cc-pVTZ basis sets. For validation purposes, we also performed MP4 and MP2 single-point calculations for selected conformers in their PBE0 + MBD geometries with 6-311++G(d,p), 6-311++G**, or aug-cc-pVTZ basis sets. We used the Gaussian16 code[60] for the CCSD(T), MP4, and MP2 simulations.

To support open data-driven chemistry and materials science,[61] we uploaded all calculations of this work to the Novel Materials Discovery (NOMAD) laboratory.[62]

2D Test. To test the accuracy and efficiency of step (ii) in our procedure, we started with a 2D search case in cysteine (Figure 3). First, we rotated the d_1 and d_2 dihedral angles to generate a reference map on a fine grid (30 × 30 points, Figure 3a). Then, d_1 and d_2 were sampled by BOSS. In both approaches, the bond lengths, bond angles, and other dihedral angles (d_3 =180.03, d_4 = 145.59, d_5 = 180.03) were fixed in their DFT-optimized values. We obtained the energy of each structure with single-point PBE0 + MBD calculations and then plot the energy relative to the global minimum.

The 2D PES maps after 60 and 120 data acquisitions are shown in Figure 3b,c. Looking at Figure 3, we find that BOSS captures correct minima and maxima already after 60 data acquisitions (6% of the computational cost of the grid method), while after 120 data acquisitions, the BOSS PES resembles the reference map very well. This 2D PES features 6 energy minima of similar depth, suggesting considerable complexity of cysteine conformational phase space and many competing minima. We apply abundant sampling in high-dimensional problems so that we can recover all relevant conformer solutions.

Cysteine Conformer Search in 5D. After demonstrating the BOSS rationale in 1D and 2D, we proceed to five dimensions. The five dihedral angles (d_1–d_5) in cysteine were sampled simultaneously by BOSS, and the energies of the corresponding configurations were evaluated with the PBE0 + MBD functional.

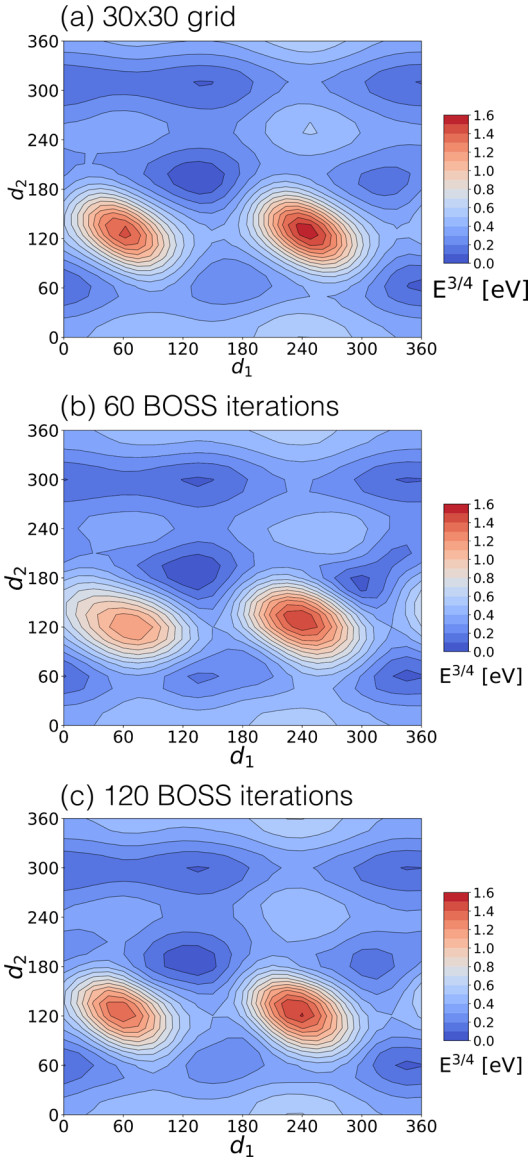


Figure 3. 2D (d_1, d_2) PES map of cysteine generated by (a) 30 × 30 = 900 DFT single-point energy calculations, (b) 60 BOSS iterations, and (c) 120 BOSS iterations. To increase the PES contrast, $E^{3/4}$ instead of E is plotted.

Figure 4 illustrates the refinement of the predicted global minimum with iterative configurational sampling. The lowest observed energy (calculated from the BOSS-predicted global minimum conformer) is shown in Figure 4a, and the values of the corresponding dihedral angles d_n is shown in Figure 4b. The lowest energy observed decreases continuously. Throughout the procedure, the geometry of the global minimum conformer changes, as Figure 4b illustrates. The global minimum undergoes several refinements until, at iteration 830, both the energy and the dihedral angles are converged and only have negligible changes (ΔE< 0.025 eV and Δd< 10°).

Improvements of the global minimum prediction is due to instances of visiting low energy configurations chosen smartly form a vast 5D space. However, most model refinements proceeded with higher energy conformers and explores local minima of the PES, on average in the region 0.4 eV above the predicted global minimum, as shown by the red dashed line in Figure 4a.

https://dx.doi.org/10.1021/acs.jctc.0c00648
J. Chem. Theory Comput. 2021, 17, 1955−1966

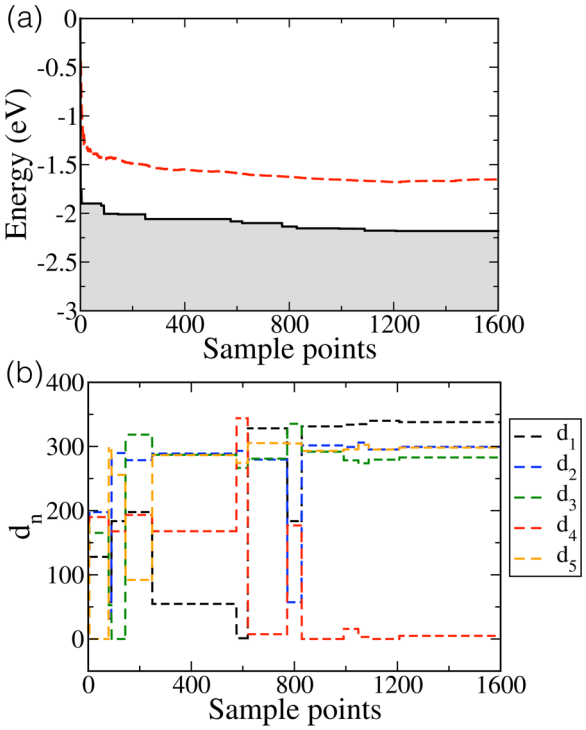


Figure 4. (a) Convergence of the global minimum energy computed from the BOSS-predicted global minimum configuration (black line). The average computed energy of the sampled conformers is shown with a red dashed line. (b) Value of the dihedral angles d_n of the BOSS-predicted global minimum as a function of the number of sampled points.

Next, we address the convergence of the low energy part of the PES. This is not a trivial task, as we cannot monitor the PES in every point of the 5D space. It also turns out to be impractical to track the dihedral angles of several low energy conformers and monitor convergence as we did for the global minimum. The reason is that many conformers are very close in energy and switch order as the iterations progress. We therefore decided to take the energy-versus-conformer-number curve as the convergence indicator.

Figure 5a shows the relative energy of all local minima after 400, 600, 800, 1000, 1200, 1400, and 1600 BOSS iterations. BOSS uses the acquisition locations as starting points for local energy minimizations on the PES, so the number of minima found tends to increase as the iteration steps increase. In the figure, 0 eV is set to be the lowest energy found in the 1000th iteration. The curves after only 400 and 600 iterations still rise steeply and feature the wrong global minimum (i.e., do not start at 0 eV). With increasing number of iterations, the curves gradually approach the curve for 1200 iterations. At 1000 iterations, the curve is very similar to that of 1200 iterations in the low energy region (<0.25 eV), which suggests that not only the global but also the low-energy local minima conformers are converged. When the BOSS iterations increase to 1400 and 1600, more local minima were found in the higher energy region (>0.25 eV), but few changes are observed below 0.25 eV. Further proof of this is presented in the Supporting Information, where we show the 2D (d_1, d_2)-projected and (d_3, d_4)-projected BOSS-predicted PESs in Figures S1 and S2. The similarity of the 2D PESs at 1000 and 1200 iterations again suggests that the model is sufficiently converged in the low energy region at 1000 iterations.

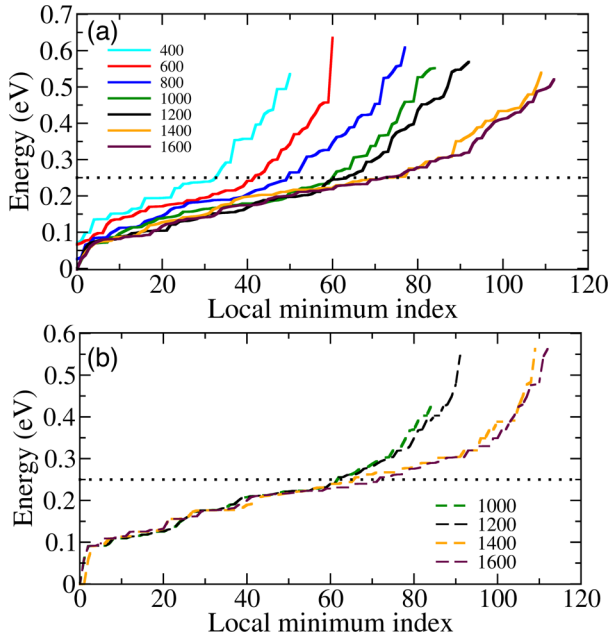


Figure 5. (a) Progression of the relative energy of predicted local minima for a PBE0 + MBD BOSS run with a total number of 1600 iterations. Shown are intermediate curves at 400, 600, 800, 1000, 1200, 1400, and 1600 iterations. (b)We took the conformers from 1000, 1200, 1400, and 1600 iterations and did the DFT structure optimization with PBE0 + MBD. The conformers are reordered from the lowest to the highest energy.

We then extracted all conformers from runs up to 1000, 1200, 1400, and 1600 iterations and performed DFT geometry optimizations for all structures. The corresponding energy vs conformer index curves are shown in Figure 5b. Now the different lines lie almost on top of each other below 0.25 eV, confirming our PES in the low-energy region is sufficiently converged for 1000 iterations.

Next, we use the optimized structures at 1000 BOSS iterations and include the vibration energy as described in BOSS-based Molecular Conformer Search. Finally, we apply CCSD(T) single-point corrections to the 15 lowest energy conformers obtained from the PBE0 + MBD calculations.

RESULTS AND DISCUSSION

Using the methodology introduced in the previous sections, we performed four independent conformer searches with the PBE + TS, PBE + MBD, PBE0 + TS, and the PBE0 + MBD functionals for cysteine. In this section, we systemically assess how the different exchange-correlation functionals and van de Waals corrections affect the results and discuss how the different steps improve accuracy. We also compare our predictions with the experimental results and reference calculations.[30,47]

We chose two references to make the comparison and validate our results. Reference[47] reports both experimental and computational results. The computational energy ordering is obtained from single-point MP4 calculation on MP2 optimized structures using 6-311++G(d,p) basis sets. In the reference, six experimental conformers were found by rotational spectroscopy (labeled in red in Figure 6); five other low-energy conformers were predicted from the MP4 simulations but were not detected in the experiment (labeled in black in Figure 6). The authors of ref 30 did a systematic scan of 11,644 initial structures at the HF/3-21G level, located 71 unique conformers of cysteine using

https://dx.doi.org/10.1021/acs.jctc.0c00648
J. Chem. Theory Comput. 2021, 17, 1955−1966

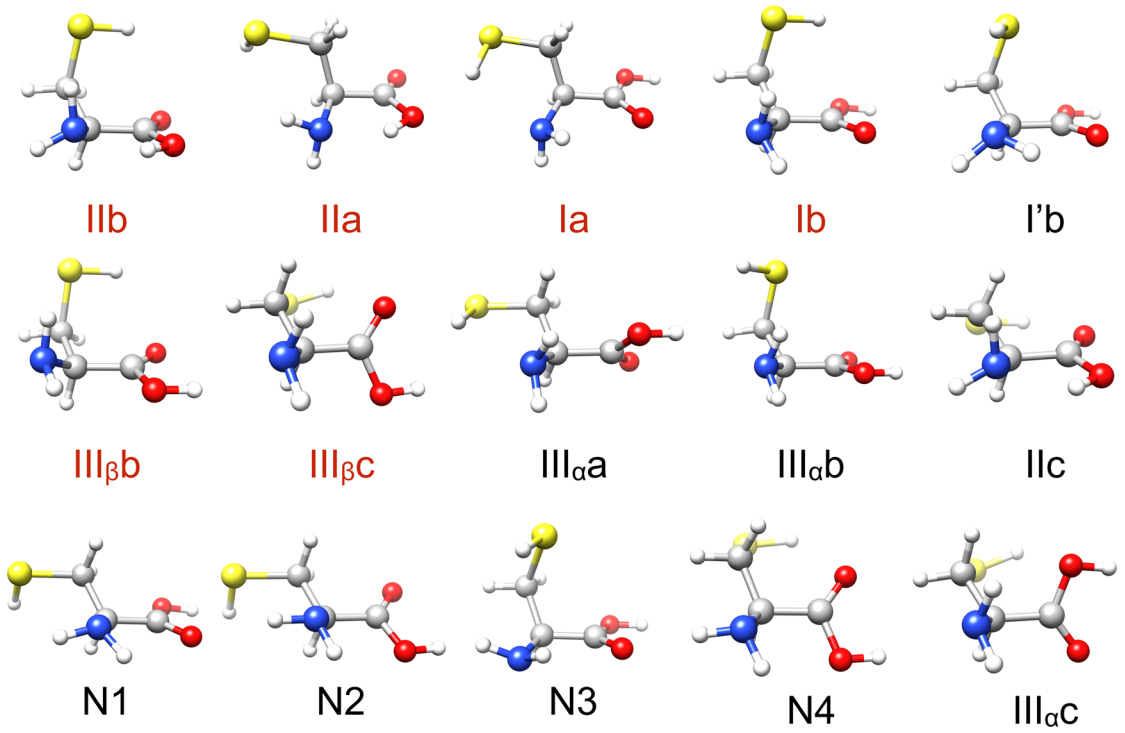


Figure 6. Predicted low energy conformers of cysteine from the PBE0 + MBD search. Conformers are named as I (NH−O═C), II (OH−N), and III (NH−OH) depending on the type of the hydrogen bonds, and as a, b, or c depending on the configuration of the $-CH_2SH$ side chain, following ref 47. The experimentally detected conformers are marked in red and other conformers marked in black. The colour scheme of the atoms is the same as in Figure 1.

the MP2(FC)/cc-pVTZ method, and finally determined the relative energies of the 11 lowest-energy conformers with CCSD(T). Reference[30] also provides *xyz*-coordinates for the observed conformers.

Conformational Energy Hierarchy of Cysteine. The predicted 15 lowest energy conformer structures of cysteine with the PBE0 + MBD functional are shown in Figure 6. The atomic coordinates of the conformers can been found in the Supporting Information. To directly compare our results with those reported in ref 47, we assign our structures the same labels as ref 47. All the 11 conformers in ref 47 have been identified in our simulations within an energy window of 0.2 eV from the global minimum. In addition, BOSS predicted new conformers, which we named N1, N2, ... Some of them are shown in Figure 6. The new conformers BOSS predicted generally have a higher energy.

The relative stability of the PBE0 + MBD conformers is shown in Figure 7a. Corresponding plots for the PBE + TS, PBE + MBD and the PBE0 + TS functionals are presented in Figures S4−S6 of the Supporting Information. To illustrate the importance of different contributions to the energy hierarchy, Figure 7a and Figures S4−S6 show not only the final energy order but also intermediate steps.

The hierarchy figures show that once the conformers are extracted, geometry optimization plays a major role in refining their energy ranking. The largest energy changes and reordering happens in this step. This is expected because BOSS models rely on the fixed bond lengths and angles (building block approximation). In the PBE0 + MBD simulation, the average energy change of the most stable 15 conformers during the geometry optimization is 0.095 eV, while the dihedral angles of the corresponding structures change on average by $\Delta d_1 = 16.9°$, $\Delta d_2 = 20.9°$, $\Delta d_3 = 8.9°$, $\Delta d_4 = 26.1°$, and $\Delta d_5 = 11.9°$. How the geometry optimization changes the total energy of individual conformers can be seen in Figure S3.

The entropy corrections have a smaller effect on the conformer ordering. The zero-point energy contributions (+VE (0 K) column) does not trigger any conformer reordering. It does, however, compress the energy spectrum as corrections for higher-energy conformers are larger than for the global minimum. The finite temperature corrections (+VE (300 K) column) leads to a further compression of the energy spectrum. Now a couple of conformers above 0.1 eV switch orders as their vibrational entropy contributions differ.

The final column in Figure 7a shows our most accurate conformer energy hierarchy, which now includes also the CCSD(T) corrections. We observe that the CCSD(T) corrections are sensitive to the conformer geometry. They generally shift conformers down in energy toward the global minimum conformer. This reduces the energy spacing between the conformers. Conformers IIa and IIc are an exception. They remain at roughly the same relative energy to the global minimum, which is also of conformer type II. They subsequently trade places with other conformers in the hierarchy.

To validate our optimized conformer structures, we start with ref 30. The geometries reported in ref 30 were obtained at the MP2(FC)/aug-cc-pV(T^+d)Z level, and we compare them against our PBE0 + MBD geometries. To standardize the comparison, we use the same conformer naming convention as in ref 47.

Among the top 10 most stable structures, ref 30 reports eight structures that we and ref 47 also found (see Table 2). These are IIb, IIa, Ib, I'b, Ia, III$_\beta$b, III$_\beta$c and III$_\alpha$b.[1] The average differences of the dihedral angles between our and ref 30's geometries are $\Delta d_1 = 4.6°$, $\Delta d_2 = 1.4°$, $\Delta d_3 = 2.8°$, $\Delta d_4 = 0.7°$ and $\Delta d_5 = 3.0°$. These small differences indicate that we indeed found the right

https://dx.doi.org/10.1021/acs.jctc.0c00648
J. Chem. Theory Comput. 2021, 17, 1955−1966

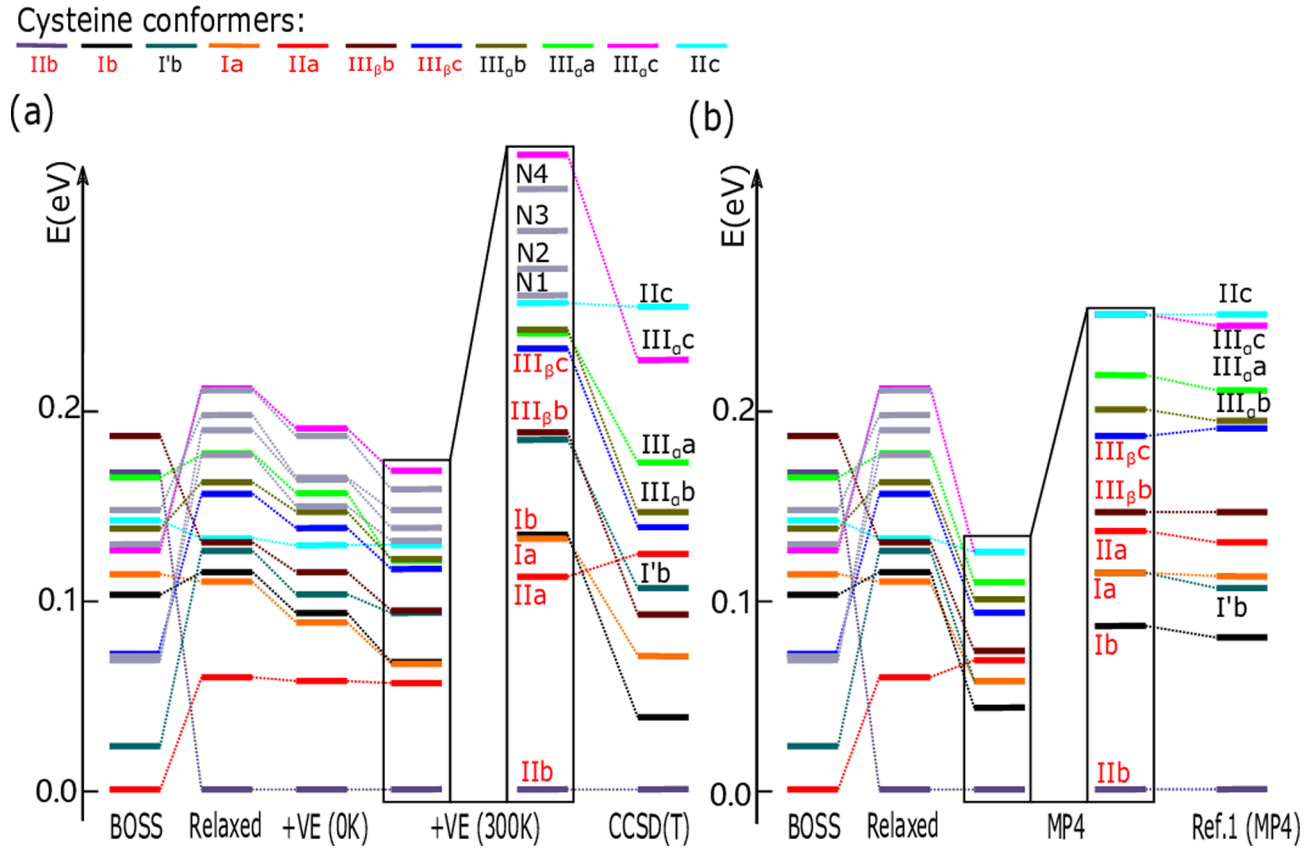


Figure 7. Relative stability for all steps of the PBE0 + MBD-based search. (a) From left to right: BOSS prediction, after structure optimization, after adding the vibrational energy at 0 K (+VE (0 K)), and adding the vibration energy at 300 K (+VE (300 K)). The two most right ones are +VE (300 K) and the energy order of CCSD(T) result but enlarged two times. For each step, the energy of the most stable structure defines the zero of energy for that column. (b) From the left to right: BOSS prediction, after optimization, and MP4 energy calculations. The last two columns show an enlarged version of the MP4 results in comparison with the MP4 results of Ref 47.

Table 1. Energy Order of the 10 most Stable Conformers of Cysteine from our DFT, MP4, and CCSD(T) Computations, Ref 47 and Ref 30[a]

	Energy order									
PBE+TS	IIb	IIa	Ib	Ia	III$_\beta$b	I′b	III$_\beta$c	IIc	III$_\alpha$a	III$_\alpha$b
PBE+MBD	IIb	IIa	Ia	Ib	III$_\beta$b	I′b	IIc	III$_\beta$c	III$_\alpha$a	III$_\alpha$b
PBE0+TS	IIb	IIa	Ia	Ib	I′b	III$_\beta$b	III$_\beta$c	III$_\alpha$a	N1	N2
PBE0+MBD	IIb	IIa	Ia	Ib	I′b	III$_\beta$b	III$_\beta$c	III$_\alpha$a	III$_\alpha$b	IIc
MP4 (b1)*	IIb	Ib	Ia	I′b	IIa	III$_\beta$b	III$_\beta$c	III$_\alpha$b	III$_\alpha$a	III$_\alpha$c
MP4 (b1)*[47]	IIb	Ib	I′b	Ia	IIa	III$_\beta$b	III$_\beta$c	III$_\alpha$b	III$_\alpha$a	III$_\alpha$c
MP4 (b2)	IIb	Ib	Ia	III$_\beta$b	I′b	IIa	III$_\beta$c	III$_\alpha$b	III$_\alpha$a	III$_\alpha$c
CCSD(T) (b2)	IIb	Ib	Ia	III$_\beta$b	I′b	IIa	III$_\beta$c	III$_\alpha$b	III$_\alpha$a	III$_\alpha$c
CCSD(T)[30]	IIb	Ib	Ia	I′b	IIa	III$_\beta$b	n/a	n/a	III$_\beta$c	III$_\alpha$b
Exp[47]	IIb	Ia	Ib	IIa	III$_\beta$c	III$_\beta$b				
Abundance ratio[47]	10	10	8	3	3	2				

[a]Our CCSD(T) and MP4 results are based on PBE0 + MBD structures. b1: 6-311++G(d,p) basis set, b2: aug-cc-pvtz basis set, *: vibrational energy corrections not included.

Table 2. Predicted Low-Energy Conformers of Cysteine and Relative Energies with Respect to the Global Minimum in eV[a]

Conformer	IIb	Ib	Ia	III$_\beta$b	I′b	IIa	III$_\beta$c	III$_\alpha$b	III$_\alpha$a	III$_\alpha$c	IIc
MP4 (b1)*	0.000	0.043	0.057	0.073	0.057	0.068	0.093	0.100	0.109	0.125	0.125
MP4 (b1)*[47]	0.000	0.040	0.056	0.073	0.053	0.065	0.095	0.097	0.105	0.122	0.125
MP4 (b2)	0.000	0.025	0.046	0.054	0.056	0.065	0.075	0.079	0.092	0.116	0.129
CCSD(T) (b2)	0.000	0.019	0.035	0.046	0.053	0.062	0.069	0.073	0.086	0.113	0.127
CCSD (T)[30]	0.000	0.050	0.060	0.069	0.062	0.068	0.099	0.100			

[a]b1: 6-311++G(d,p) basis set, b2: aug-cc-pvtz basis set,*: vibrational energy corrections not included.

https://dx.doi.org/10.1021/acs.jctc.0c00648
J. Chem. Theory Comput. 2021, 17, 1955−1966

conformers and that the PBE0 + MBD and MP2 geometries agree closely.

Reference 47 unfortunately, does not provide atomic coordinates for the reported conformers. To validate our optimized conformer structures against those of ref 47, we therefore performed MP4 single-point energy calculations with the same basis set 6−311++G(d,p) as in Ref 47, but for our PBE0 + MBD geometries. The results are reported in Figure 7b and Table 1.

Figure 7b and Table 2 show that the energies of the two MP4 calculations (MP4(b1) and MP4(b1)[47]) agree within 4 meV for each conformer. This close match indicates that our conformer geometries agree very well with those of ref 47, validating our BOSS-based conformer search procedure.

Table 1 shows the final energy ranking of the top 10 most stable conformers in ref 47, ref 30 and our computational predictions. A more complete list of the low-energy conformers and their relative energy can be found in Table S1.

In our simulations, PBE + TS, PBE + MBD, PBE0 + TS, and PBE0 + MBD all found the correct global minimum structure IIb. PBE + TS, PBE0 + TS and PBE0 + MBD predicted the six experimental identified conformers among the top seven most stable structures, while PBE + MBD locates the six conformers among the top eight most stable ones.

In Figure 8, we summarize the comparison across the four different exchange-correlation functionals we tested. Our

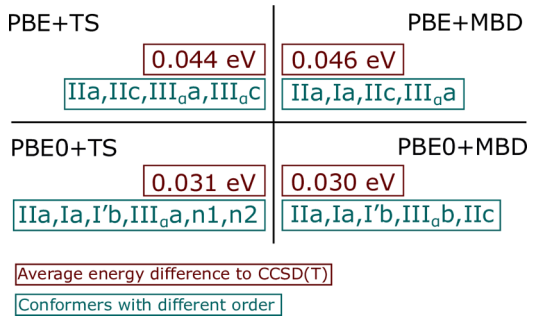


Figure 8. Summary of DFT results: each panel shows the average energy difference between the respective DFT functional and the CCSD(T) reference energies for the 10 lowest conformers. In addition, each panel lists the conformers that have a different order than in CCSD(T).

reference are the CCSD(T) energies at the PBE0 + MBD geometries. In Figure 8, we list the conformers that have a different energy ordering in the DFT and CCSD(T). The energy differences between the cysteine conformers are extremely small. Therefore, it is no surprise that the DFT energy rankings differ from the CCSD(T) results. The accuracy of the different DFT functional are then evaluated by the energy differences comparing to CCSD(T), using the 10 lowest energy conformers in CCSD(T)). Comparing to CCSD(T), the average energy difference is 0.044 eV for PBE + TS, 0.046 eV for PBE + MBD, 0.031 eV for PBE0 + TS, and 0.030 eV for PBE0 + MBD (Figure 8). PBE0 is on average 0.01 eV more accurate than PBE. The difference between the different van der Waals treatments (TS or MBD) is an order of magnitude smaller (1 or 2 meV on average), but MBD is more than 10^2 times more expensive than TS for cysteine. The influence of the different vdW treatments is negligible for a small molecule like cysteine; however, MBD may become important for accurate treatments of larger molecules, e.g., biomolecules. For cysteine, we can conclude that PBE + TS is sufficient for the conformer search.

Since BOSS is able to sample the configurational space very efficiently, we performed the whole conformer search at the PBE0 + MBD level. For larger molecules, it might become more economical to perform an initial BOSS-based conformer search at the PBE + TS level and to post-relax only a certain number of low-energy conformers with PBE0 + MBD.

Our CCSD(T) calculations produce a very similar energy ranking as the MP4 results in ref 47, as shown in Table 1. The only difference with ref 47 is the placement of I′b, III$_\beta$b. If we use the same aug-cc-pvtz basis set, same geometries and same vibrational energy correction from our PBE0 + MBD simulations in both CCSD(T) and MP4, we get the same energy order. Therefore, the differences are not caused by the choice of CCSD(T) or MP4. Since we have validated that we have found very similar structures as ref 47 (Figure 7b), the difference may due to the fact that ref 47 did not include the entropy correction and used different basis sets.

Reference 30 reports two structures that are similar to IIa but do not appear in ref 47 or our conformer search. Except for these two new structures, the only difference between our CCSD(T) and the CCSD(T) results in ref 30 is the ordering of I′b and III$_\beta$b. Again, the energy differences between the conformers in this range are extremely small, and ordering differences in our results and the reference can be ascribed to the slight difference of the conformer structures and computational settings. Reference[30] used a different vibrational correction method and included the focal-point analysis to extrapolate the energies to the complete basis set limit.

Comparing our CCSD(T) results to the experiment, we note that the CCSD(T) ordering of IIb, Ib, and Ia as the three lowest energy conformers agrees with the experimental ordering derived from the relative abundance of the detected conformers. However, the order of Ia and Ib is switched, which is the same as the computational ranking in refs 47 and 30. For the next three conformers, the experiment finds IIa, III$_\beta$b, and III$_\beta$c, however, with much lower overall abundance than the first three conformers. The coupled cluster order is different with III$_\beta$b, I′b, IIa, and III$_\beta$c. These differences can be ascribed to the low experimental abundance, which might make an unambiguous classification difficult, or to additional experimental factors that are not taken into account in our simulations.

Conformational Energy Hierarchy of Serine, Aspartic Acid, and Tryptophan. In this section, we applied our conformer search procedure with the PBE0 + MBD functional to serine, aspartic acid, and tryptophan. For comparison, we label their conformers in accordance with the corresponding reference.[63−65]

The BOSS convergence of serine, aspartic acid, and tryptophan is similar to that of cysteine. Serine and aspartic acid converged in 1200 and tryptophan in 1000 iterations (Figure S8). We then followed the same procedure as for cysteine, i.e., we extracted and relaxed the local minima structures and included entropy corrections at 300 K. Finally, we added CC corrections to the 15 lowest energy conformers. The global minimum structures of the three molecules are shown in Figure 9, and the relative energy of the 10 lowest energy conformers are listed in Table 3.

For serine, we found the seven experimental detected conformers among the top nine most stable structures.[63] The CCSD(T) energy ranking agrees well with the experimental population order, which is Ia > IIb > I′b > IIc > III$_\beta$b ≈ III$_\beta$c ≈

https://dx.doi.org/10.1021/acs.jctc.0c00648
J. Chem. Theory Comput. 2021, 17, 1955−1966

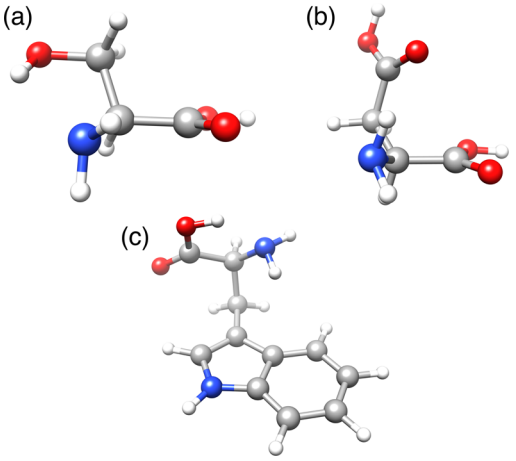


Figure 9. Global minimum structure of (a) serine, (b) aspartic acid, and (c) tryptophan.

IIa.[63] For aspartic acid, we found the six experimental reported conformers among the top eight most stable structures. Our order is close to the MP2-calculated conformer order in ref 64 (Table S3). Tryptophan has a more complicated structure and can form eight types of hydrogen bonds (A−H).[65] Experiments and previous simulations have confirmed that the most stable one is A-type, which dominated the tryptophan population, followed by two B-type conformers.[65,66] We also got the same order from our CCSD(T) energies.

To compare with reported computational results, we calculated the MP2 or MP4 energies of the three molecular conformers, using our PBE0 + MBD optimized structures and the same basis sets as in refs 63, 65. The relative energy and ranking of the 10 most stable molecular conformers are shown in Tables S2−S4.

For Serine, our MP4 results are very similar to ref 63. The only difference is that the order of IIc and I′b is switched. The average energy difference is 0.006 eV for all the conformers in Table S2.

For aspartic acid, the orders of IIb-I, III_{β}c, Ia-nπ2, and Ib-III are different between our study and ref 64. However, the average energy difference is only 0.012 eV for all the conformers in Table S3, which reduces to 0.003 eV if we only consider the experimentally detected ones.

Reference 65 and we both found that the A and B types are more stable than other types for tryptophan. The average energy difference of A and B type conformers in Table S4 is 0.010 eV. These results proved that we had found similar conformer structures as the previous computational studies.

Computational Efficiency. We close with a note on the efficiency of our new conformer search procedure without explicitly performing other search methods in this work. BOSS predicts a physically meaningful PES for the four amino acids with 5−6 degrees of freedom with only ∼1000 single-point DFT calculations. We can put this number of single-point calculations into perspective by considering that FHI-aims requires on average 30 geometry optimization steps to relax the structure of an organic molecule. The computational cost of 1000 single-point DFT calculations is therefore equivalent to approximately 30 DFT geometry optimizations.

From the PES, we extract all relevant low-energy conformers with the BOSS postprocessing minima search tool at a small computational expense. In this work, we consider approximately 80 local minima, each of which is geometry optimized with DFT. This amounts to 80 geometry optimizations, which is equivalent to approximately 2400 DFT single-point calculations.

Our total computational expense per DFT functional for a complete conformer search of cysteine is therefore 3400 DFT single-point calculations or equivalently about 100 geometry optimizations. Similar DFT steps were used to search the conformer of serine, aspartic acid, and tryptophan. This is a very small computational budget, compared to systematic[30] or stochastic[32] conformer search methods that need to relax thousands of structures. Supady *et al.* provided detailed numbers for a genetic algorithm (GA)-based conformer search of dipeptides.[32] Their search encompasses between 4 and 6 degrees of freedom and is therefore similar to ours, as is the size of the molecules. The GA search requires between 20,000 and 60,000 single-point DFT calculations (referred to as force evaluations in ref 32) depending on the size of the search space and the density of conformers in the energy hierarchy. Our BOSS-based procedure is a factor of 10 more efficient. A similar speed up was recently observed in a Gaussian-process-based structure search of oxidized graphene on the Ir(111).[67] It is important to mention that different systems have different funneled PES, so the number of degrees of freedom is not the only important fact for conformer search. The comparison to the previous GA study[32] is informative rather than quantitative.

CONCLUSIONS

In summary, we propose a new conformer search procedure that combines the Bayesian optimization active learning with

Table 3. Predicted Low-Energy Conformers of Serine, Tryptophan, and Aspartic Acid and Relative Energies with Respect to the Global Minimum in eV[a]

Serine	Ia	IIb	I′b	IIc	III_{β}b	Ib	III_{α}a	IIa	III_{β}c	Ic
PBE0+MBD	0.036	0.000	0.062	0.005	0.098	0.147	0.112	0.078	0.131	0.169
+VE (300K)	0.012	0.000	0.049	0.019	0.088	0.096	0.078	0.069	0.103	0.118
CCSD(T)	0.000	0.001	0.028	0.044	0.053	0.057	0.061	0.068	0.070	0.089
Aspartic acid	Ib-I	IIb-I	IIa-I	III_{β}b-I	Ia-I	III_{α}a-I	Ia-nπ2	Ia-II	Ic	III_{α}c
PBE0+MBD	0.082	0.000	0.003	0.088	0.087	0.110	0.150	0.030	0.173	0.179
+VE (300K)	0.062	0.009	0.000	0.070	0.059	0.077	0.101	0.040	0.103	0.142
CCSD (T)	0.000	0.001	0.008	0.010	0.019	0.031	0.046	0.047	0.057	0.081
Tryptophan	A	B	B	A	A	D	D	C	C	C
PBE0+MBD	0.000	0.125	0.132	0.078	0.044	0.149	0.148	0.173	0.184	0.149
+VE (300K)	0.000	0.089	0.096	0.027	0.029	0.111	0.124	0.137	0.105	0.098
CCSD(T)	0.000	0.018	0.020	0.042	0.044	0.049	0.053	0.063	0.073	0.075

[a]Aug-cc-pvtz basis set was used for the CCSD(T) calculations for serine and aspartic acid; 6-311++G(d,p) basis set was used for the CCSD(T) calculations for tryptophan.

https://dx.doi.org/10.1021/acs.jctc.0c00648
J. Chem. Theory Comput. 2021, 17, 1955−1966

quantum chemistry methods. BOSS performs a global phase space search and finds all the relevant conformers in one run. Then, we refine the low-energy conformers by DFT structure relaxation, vibrational energy, and coupled cluster correction. We conclude that the DFT structure relaxation plays a major role in the refinement of the energy order. We also find that PBE0 gives slightly better results than PBE, but the difference between the TS and MBD van der Waals interactions are tiny for our system.

Unlike traditional conformer search methods, our approach is computationally tractable while retaining the accuracy of the chosen quantum chemical method throughout. This approach is most suitable for small molecules that require highly accurate and expensive quantum chemistry methods for conformer ranking. Extending the method to larger molecules with a much larger search space will require reliable dimension reduction strategy, either based on previous knowledge or computational techniques.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.0c00648.

xyz coordinates of the low-energy conformers of cysteine, serine, aspartic acid, and tryptophan (ZIP)

(Table S1) Energy order of the low-energy conformers of cysteine; (Tables S2−S4) Predicted low-energy conformers and relative energies of serine (MP4), aspartic acid (MP2), tryptophan (MP2); (Figures S1 and S2) 2D-projected PES maps predicted by BOSS in the 5D case of cysteine; (Figure S3) total energy of cysteine conformers from BOSS, DFT single-point calculation, and after DFT optimization; (Figures S4−S6) Relative stability for all steps of the PBE + TS-based, PBE + MBD-based, and PBE0 + TS-based search; (Figure S7) DFT-calculated and transferred energy of the new structure each BOSS iteration predicted (aspartic acid); and (Figure S8) progression of the relative energy of predicted local-minima for a PBE0 + MBD BOSS run for serine, aspartic acid, and tryptophan (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Patrick Rinke − *Department of Applied Physics, Aalto University, AALTO 00076, Finland;* orcid.org/0000-0003-1898-723X; Email: patrick.rinke@aalto.fi

Xi Chen − *Department of Applied Physics, Aalto University, AALTO 00076, Finland;* orcid.org/0000-0001-6149-2270; Email: xi.6.chen@aalto.f

Authors

Lincan Fang − *Department of Applied Physics, Aalto University, AALTO 00076, Finland*

Esko Makkonen − *Department of Applied Physics, Aalto University, AALTO 00076, Finland*

Milica Todorović − *Department of Applied Physics, Aalto University, AALTO 00076, Finland;* orcid.org/0000-0003-0028-0105

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.0c00648

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Academy of Finland (project numbers 308647, 314298, and 316601) and through their Flagship program: Finnish Center for Artificial Intelligence FCAI. We thank CSC, the Finnish IT Center for Science, and Aalto Science IT for computational resources. This work is supported by COST (European Cooperation in Science and Technology) Action 18234. L.F. thanks Guoxu Zhang, Marc Dvorak, Jingrui Li, and Annika Stuke for the help with FHI-Aims. He also acknowledges financial support from the Chinese Scholarship Council (grant no. [2017]3109).

■ ADDITIONAL NOTE

[1]Using our classification standard, we classified the VI(n/a) conformer in Ref 30. to be IIb.

■ REFERENCES

(1) Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747−1756.

(2) Wales, D. J.; Miller, M. A.; Walsh, T. R. Archetypal energy landscapes. *Nature* **1998**, *394*, 758−760.

(3) Wales, D. J.; Bogdan, T. V. Potential Energy and Free Energy Landscapes. *J. Phys. Chem. B* **2006**, *110*, 20765−20776.

(4) Wlodek, S.; Skillman, A. G.; Nicholls, A. Ligand Entropy in Gas-Phase, Upon Solvation and Protein Complexation. Fast Estimation with Quasi-Newton Hessian. *J. Chem. Theory Comput.* **2010**, *6*, 2140−2152.

(5) Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 2719−2728.

(6) Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 529−539.

(7) Schwab, C. H. Conformations and 3D pharmacophore searching. *Drug Discovery Today: Technol.* **2010**, *7*, e245−e253.

(8) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74−82.

(9) Stillinger, F. H.; Weber, T. A. Packing Structures and Transitions in Liquids and Solids. *Science* **1984**, *225*, 983−989.

(10) Stillinger, F. H.; Weber, T. A. Hidden structure in liquids. *Phys. Rev. A* **1982**, *25*, 978.

(11) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462−2474.

(12) Puranen, J. S.; Vainio, M. J.; Johnson, M. S. Accurate conformation-dependent molecular electrostatic potentials for high-throughput *in silico* drug discovery. *J. Comput. Chem.* **2010**, *31*, 1722−1732.

(13) Miteva, M. A.; Guyon, F.; Tufféry, P. Frog2: Efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Res.* **2010**, *38*, W622−W627.

(14) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.

(15) Landrum, G. *RDKit: open-source cheminformatics*; 2011.

(16) Chemical Computing Group.

(17) Chang, G.; Guida, W. C.; Still, W. C. An internal-coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **1989**, *111*, 4379−4386.

(18) Wilson, S. R.; Cui, W.; Moskowitz, J. W.; Schmidt, K. E. Applications of simulated annealing to the conformational analysis of flexible molecules. *J. Comput. Chem.* **1991**, *12*, 342−349.
(19) Goedecker, S. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **2004**, *120*, 9911.
(20) Sutherland-Cash, K. H.; Wales, D. J.; Chakrabarti, D. Free energy basin-hopping. *Chem. Phys. Lett.* **2015**, *625*, 1−4.
(21) Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A* **1997**, *101*, 5111−5116.
(22) Spellmeyer, D. C.; Wong, A. K.; Bower, M. J.; Blaney, J. M. Conformational analysis using distance geometry methods. *J. Mol. Graphics Modell.* **1997**, *15*, 18−36.
(23) Mekenyan, O.; Dimitrov, D.; Nikolova, N.; Karabunarliev, S. Conformational Coverage by a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 997−1016.
(24) Kothiwale, S.; Mendenhall, J. L.; Meiler, J. BCL::Conf: small molecule conformational sampling using a knowledge based rotamer library. *Aust. J. Chem.* **2015**, *7*, 47.
(25) Cole, J. C.; Korb, O.; McCabe, P.; Read, M. G.; Taylor, R. Knowledge-Based Conformer Generation Using the Cambridge Structural Database. *J. Chem. Inf. Model.* **2018**, *58*, 615−629.
(26) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380−388.
(27) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(28) Rossi, M.; Scheffler, M.; Blum, V. Impact of Vibrational Entropy on the Stability of Unsolvated Peptide Helices with Increasing Length. *J. Phys. Chem. B* **2013**, *117*, 5574−5584.
(29) Chutia, S.; Rossi, M.; Blum, V. Water Adsorption at Two Unsolvated Peptides with a Protonated Lysine Residue: From Self-Solvation to Solvation. *J. Phys. Chem. B* **2012**, *116*, 14788−14804.
(30) Wilke, J. J.; Lind, M. C.; Schaefer, H. F., III; Császár, A. G.; Allen, W. D. Conformers of Gaseous Cysteine. *J. Chem. Theory Comput.* **2009**, *5*, 1511−1523.
(31) Nair, N.; Goodman, J. M. Genetic Algorithms in Conformational Analysis. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 317−320.
(32) Supady, A.; Blum, V.; Baldauf, C. First-Principles Molecular Structure Search with a Genetic Algorithm. *J. Chem. Inf. Model.* **2015**, *55*, 2338−2348.
(33) Chen, X.; Jørgensen, M. S.; Li, J.; Hammer, B. Atomic Energies from a Convolutional Neural Network. *J. Chem. Theory Comput.* **2018**, *14*, 3933−3942.
(34) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 2903−2910.
(35) Schmitz, G.; Godtliebsen, I. H.; Christiansen, O. Machine learning for potential energy surfaces: An extensive database and assessment of methods. *J. Chem. Phys.* **2019**, *150*, 244113.
(36) Denzel, A.; Kästner, J. Gaussian process regression for geometry optimization. *J. Chem. Phys.* **2018**, *148*, No. 094114.
(37) Meyer, R.; Hauser, A. W. Geometry optimization using Gaussian process regression in internal coordinate systems. *J. Chem. Phys.* **2020**, *152*, No. 084112.
(38) Sauceda, H. E.; Chmiela, S.; Poltavsky, I.; Müller, K.-R.; Tkatchenko, A. Construction of Machine Learned Force Fields with Quantum Chemical Accuracy: Applications and Chemical Insights. *arXiv:1909.08565* .
(39) Chan, L.; Hutchison, G. R.; Morris, G. M. Bayesian optimization for conformer generation. *Aust. J. Chem.* **2019**, *11*, 32.
(40) Todorović, M.; Gutmann, M. U.; Corander, J.; Rinke, P. Bayesian inference of atomistic structure in functional materials. *npj Comput. Mater.* **2019**, *5*, 35.
(41) Garijo del Río, E.; Mortensen, J. J.; Jacobsen, K. W. Local Bayesian optimizer for atomic structures. *Phys. Rev. B* **2019**, *100*, 104103.
(42) Gitlab *Bayesian Optimization Structure Search (BOSS)*. https://gitlab.com/cest-group/boss.
(43) Lee, H.-E.; Ahn, H.-Y.; Mun, J.; Lee, Y. Y.; Kim, M.; Cho, N. H.; Chang, K.; Kim, W. S.; Rho, J.; Nam, K. T. Amino-acid- and peptide-directed synthesis of chiral plasmonic gold nanoparticles. *Nature* **2018**, *556*, 360−365.
(44) Häkkinen, H. The goldsulfur interface at the nanoscale. *Nat. Chem.* **2012**, *4*, 443−455.
(45) Gronert, S.; O'Hair, R. A. J. Ab Initio Studies of Amino Acid Conformations. 1. The Conformers of Alanine, Serine, and Cysteine. 1995, *117*, 2071−2081, DOI: 10.1021/ja00112a022.
(46) Dobrowolski, J. C.; Rode, J. E.; Sadlej, J. Cysteine conformations revisited. *J. Mol. Struct.: THEOCHEM* **2007**, *810*, 129−134.
(47) Sanz, M. E.; Blanco, S.; López, J. C.; Alonso, J. L. Rotational Probes of Six Conformers of Neutral Cysteine. *Angew. Chem., Int. Ed.* **2008**, *47*, 6216−6220.
(48) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; the MIT Press: 2006.
(49) Järvi, J.; Rinke, P.; Todorović, M. Detecting stable adsorbates of (1S)-camphor on Cu(111) with Bayesian optimization. *Beilstein J. Nanotechnol.* **2020**, *11*, 1577−1589.
(50) Brochu, E.; Cora, V. M.; de Freitas, N. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv:1012.2599v1* .
(51) Rossi, M.; Chutia, S.; Scheffler, M.; Blum, V. Validation Challenge of Density-Functional Theory for Peptides−Example of Ac-Phe-Ala_5-$LysH^+$. *J. Phys. Chem. A* **2014**, *118*, 7349−7359.
(52) Hoja, J.; Ko, H.-Y.; Neumann, M. A.; Car, R.; DiStasio, R. A., Jr.; Tkatchenko, A. Reliable and practical computational description of molecular crystal polymorphs. *Sci. Adv.* **2019**, *5*, eaau3338.
(53) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. *Ab initio* molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **2009**, *180*, 2175−2196.
(54) Havu, V.; Blum, V.; Havu, P.; Scheffler, M. Efficient $O(N)$ integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.* **2009**, *228*, 8367−8379.
(55) Ren, X.; Rinke, P.; Blum, V.; Wieferink, J.; Tkatchenko, A.; Sanfilippo, A.; Reuter, K.; Scheffler, M. Resolution-of-identity approach to Hartree−Fock, hybrid density functionals, RPA, MP2 and *GW* with numeric atom-centered orbital basis functions. *New J. Phys.* **2012**, *14*, No. 053020.
(56) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.
(57) Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, No. 073005.
(58) Tkatchenko, A.; DiStasio, R. A., Jr.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.
(59) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158.
(60) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; et al. *Gaussian16*, Revision C.01; Gaussian Inc.: Wallingford CT, 2016.
(61) Himanen, L.; Geurts, A.; Foster, A. S.; Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Adv. Sci.* **2019**, *6*, 1900808.
(62) *The Novel Materials Discovery (NOMAD) Laboratory*. https://nomad-coe.eu/.
(63) Blanco, S.; Sanz, M. E.; López, J. C.; Alonso, J. L. Revealing the multiple structures of serine. *Proc. Natl. Acad. Sci.* **2007**, *104*, 20183−20188.

https://dx.doi.org/10.1021/acs.jctc.0c00648
J. Chem. Theory Comput. 2021, 17, 1955−1966

(64) Sanz, M. E.; López, J. C.; Alonso, J. L. Six conformers of neutral aspartic acid identified in the gas phase. *Phys. Chem. Chem. Phys.* **2010**, *12*, 3573−3578.
(65) Huang, Z.; Lin, Z. Detailed Ab Initio Studies of the Conformers and Conformational Distributions of Gaseous Tryptophan. *J. Phys. Chem. A* **2005**, *109*, 2656−2659.
(66) Compagnon, I.; Hagemeister, F. C.; Antoine, R.; Rayane, D.; Broyer, M.; Dugourd, P.; Hudgins, R. R.; Jarrold, M. F. Permanent Electric Dipole and Conformation of Unsolvated Tryptophan. *J. Am. Chem. Soc.* **2001**, *123*, 8440−8441.
(67) Bisbo, M. K.; Hammer, B. Efficient Global Structure Optimization with a Machine-Learned Surrogate Model. *Phys. Rev. Lett.* **2020**, *124*, No. 086102.

https://dx.doi.org/10.1021/acs.jctc.0c00648
J. Chem. Theory Comput. 2021, 17, 1955−1966