BOKEI: Bayesian Optimization Using Knowledge of Correlated Torsions and Expected Improvement for Conformer Generation

Lucian Chan,[†] Geoffrey R. Hutchison,[*,‡] and Garrett M. Morris[*,†]

†*Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, UK*

‡*Department of Chemistry and Chemical Engineering, University of Pittsburgh, 219 Parkman Avenue, Pittsburgh, PA 15260, USA*

E-mail: geoffh@pitt.edu; morris@stats.ox.ac.uk

Abstract

A key challenge in conformer sampling is to find low-energy conformations with a small number of energy evaluations. We recently demonstrated the Bayesian Optimization Algorithm (BOA) is an effective method for finding the lowest energy conformation of a small molecule. Our approach balances between *exploitation* and *exploration*, and is more efficient than exhaustive or random search methods. Here, we extend strategies on proteins and oligopeptides (*e.g.* Ramachandran plots of secondary structure) and study the correlated torsions in small molecules. We use bivariate von Mises distributions to capture correlations, and use them to constrain the search space. We validate the performance of our new method, Bayesian Optimization with Knowledge-based Expected Improvement (BOKEI), on a dataset consisting of 533 diverse small molecules, using (i) a force field (MMFF94); and (ii) a semi-empirical method (GFN2), as the

objective function. We compare the search performance of BOKEI, the BOA with Expected Improvement (BOA-EI), and a genetic algorithm (GA), using a fixed number of energy evaluations. In more than 60% of the cases examined, BOKEI finds lower energy conformations than global optimization with BOA-EI or GA. More importantly, we find correlated torsions in up to 15% of small molecules in larger data sets, up to 8 times more often than previously reported. The BOKEI patterns not only describe steric clashes, but also reflect favorable intramolecular interactions such as hydrogen bonds and π-π stacking. Increasing our understanding of the conformational preferences of molecules will help improve our ability to find low energy conformers efficiently, which will have impact in a wide range of computational modeling applications.

1 Introduction

Many molecules can adopt multiple geometrically-distinct conformers. Considerable effort has been devoted to understanding the influence of structure on function, notably in the fields of protein folding, and protein-ligand binding.[1–3] Finding the energetically-lowest conformation of a small molecule is a common task in computational chemistry.[4,5] Here, we introduce a new search method that extends our previously proposed search strategy, Bayesian Optimization Algorithm, BOA,[6] by incorporating prior knowledge of correlated adjacent pairs of torsional angles.

We recently showed BOA tends to find the lowest energy conformation of small to medium organic molecules more efficiently than both an exhaustive systematic search, Confab,[7] and a uniform random search, when evaluated using a "fixed-rotor approximation" and the MMFF94[6] force field. BOA required an order of magnitude fewer energy evaluations than these methods to find the lowest energy conformation. Drawing from statistical mechanics, BOA begins with an initial estimate of the probability of likely dihedral angles (*e.g.* 60°, 120°, 180°, *etc.*) and updates these probabilities by evaluating a new point on the potential

energy surface (PES). In this way, BOA "learns" the most likely dihedral angles of a molecule from all previous observations, and determines the next query point based on the model's uncertainty. This approach balances *exploration* and *exploitation*, and prevents the search from being trapped in local minima. However, BOA suffers from two limitations: one is that it tends to unnecessarily sample high energy regions of the potential energy surface. A second is that the Gaussian Process (GP) regression that is used as a surrogate model, which is well-known to have high computational complexity, although recent studies[8] have shown the asymptotic complexity of the exact GP inference can be reduced to $O(N^2)$, where N is the total number of energy evaluations. Here, we introduce a new *knowledge-based* acquisition function to address the first issue. Although GP has high computational complexity, we still use it as the surrogate model in this work. Its well-calibrated model of uncertainty is heavily used to guide the sampling on the PES in this work.

Various knowledge-based methods[9–12] have been proposed for (diverse) conformer generation. These methods utilise the torsional preferences in guiding conformer sampling and the torsion rules are typically derived from databases of experimental X-ray crystal structures, such as the Protein Data Bank[13] and the Cambridge Structural Database,[14] although they can also be derived from computed structures. Context and nearest neighbour effects are generally ignored in the torsion rules: each dihedral is treated as an independent free rotor.[1]

Structural information about adjacent and proximal rotatable bonds can, however, be crucial for conformer generation, as these torsion angles are naturally constrained. This can help to reduce steric clashes, retain π-conjugation, align intramolecular hydrogen bonds, or preserve other similar non-covalent interactions. For instance, Figure 1b shows the computed MMFF94 potential energy surface for 5-phenylthioquinazoline-2,4-diamine, with light blue indicating conformations with low energies. Due to steric clashes, the neighboring dihedral angles are clearly correlated and thus the conformational search can be greatly focused on the most favorable regions of the state space by incorporating this information. Even in a simpler

molecule such as *ortho*- 1,1′:2′,1″-terphenyl, there are correlations between non-neighboring dihedral angles due to steric clashes (Figure 1d).

In this work, we examine the distributions of correlated torsions in (i) X-ray crystal structures; (ii) the lowest-energy conformations from MMFF94; and (iii) an approximate quantum method, GFN2. We analyse the distributions of the correlated torsions in lowest-energy conformations computed first by MMFF94, and then by GFN2, and use these distributions to constrain the search space in Bayesian optimization using a modified acquisition function. We show that this significantly improves the efficiency of the search for low-energy conformations. We also show that correlated dihedral angles are common in organic small molecules.

2 Material and Methods

2.1 Knowledge-based Method

Guba et al.[15] derived a set of rotatable bond SMARTS patterns that are used in RDKit's[16] Experimental-Torsion Distance Geometry with basic Knowledge (ETKDG) algorithm,[17] and also in BOA,[6] the predecessor of our new method, BOKEI. This library only considered substructures consisting of a single rotatable bond. We expanded this set of patterns to consider correlated torsions in substructures consisting of two adjacent rotatable bonds.

2.1.1 Correlated Torsion Rules

We enumerated all possible pairs of rotatable bond SMARTS patterns as defined in the library of Guba et al.,[15] and counted the frequencies of the corresponding pairs of torsion angles observed in ligands having five or fewer rotatable bonds in the Crystallography Open Database (COD).[18,19] We excluded pair patterns with fewer than 100 observations, resulting

in 19 correlated torsion patterns, including one SMARTS pattern suggested by Cole et al.[9] (see Appendix 1 Table S1). The remaining substructures defined by Cole et al. were discarded due to insufficient observations in our dataset. Future work will expand on this initial set as discussed below.

For molecules that contain any matches of the derived SMARTS patterns, we calculated the lowest energy conformation and recorded the corresponding torsion angles. We performed the calculations with two energy functions: (i) a force field, MMFF94,[20] and (ii) a semi-empirical method, GFN2.[21,22] We then modeled the observed torsion distribution with bivariate von Mises mixture models for all three sources of structure (X-ray crystal structure, MMFF94, and GFN2) separately. Details of the calculations and models can be found in Appendix 2. The bivariate von Mises distribution has been used to model torsion angles in protein structures.[23–25] There are two advantages of using this distribution for correlated torsions: (i) it can model correlated torsions that cannot be described by a simple clash term; and (ii) it can be easily integrated with existing conformer sampling schemes, such as distance geometry.[26]

2.2 Bivariate von Mises Distribution and Mixture Models

The bivariate von Mises distribution is a probability distribution that can be used to jointly model two angular variables (θ_1 and θ_2). It can be thought of as an analogue of the bivariate normal distribution on a torus. In particular, we used the cosine variant of the von Mises distribution, which is as follows:

$$f(\theta_1, \theta_2) \propto \exp\{\kappa_1 \cos(\theta_1 - \mu) + \kappa_2 \cos(\theta_2 - \nu) - \kappa_3 \cos(\theta_1 - \mu - \theta_2 + \nu)\} \qquad (1)$$

where the parameters (μ, ν) and (κ_1, κ_2) in the model represent the mean, and concentrations respectively. (κ_3) is a parameter controlling the correlation.

The bivariate von Mises distribution can be unimodal or bimodal under different conditions (see Appendix 2). A single bivarate von Mises distribution is not sufficient to capture the multimodality of the torsion preferences. We used mixture models (Eq. 2) to describe the torsion preference entirely; the probability, $P(\theta_1, \theta_2)$ of observing the pair of torsion angles, θ_1 and θ_2, is given by:

$$P(\theta_1, \theta_2) = \sum_{i=1}^{K} w_i f_i(\theta_1, \theta_2) \tag{2}$$

where K is the number of components in the model, and w_i is the weight of each component. The estimation procedure of the parameters of the mixture models are explained in Appendix 2.

We should note that the bivariate von Mises mixture model requires sufficient data to accurately describe torsion preferences, so cases with small numbers of observations were excluded in this work. This limits the current performance of our algorithm, and we discuss some potential solutions in Section 3.7.

2.3 Bayesian Optimization

The general idea of Bayesian optimization is to construct a surrogate model that approximates a black box objective function, $f(x)$, and exploit this model to decide which points to evaluate next. The sampling strategy is determined by the choice of acquisition function, such as Expected Improvement (EI) and Lower Confidence Bound (LCB). Gaussian Process is commonly used as the surrogate model. For more detail about the Gaussian Process and a review of Bayesian optimization, see Rasmussen and Williams,[27] Brochu et al.[28] and Shahriari et al.[29]

$$\mathrm{EI}(\theta) = \sigma(\theta)(z(\theta)\Phi(z(\theta)) + \phi(z(\theta)) \tag{3}$$

$$\mathrm{LCB}(\theta) = \mu - \gamma\sigma(\theta) \tag{4}$$

Here, $z(\theta) = \frac{f(\theta_{best})-\mu(\theta)}{\sigma(\theta)}$, where θ_{best}, $\mu(\theta)$, and $\sigma^2(\theta)$ are the best current value, predictive mean, and predictive variance respectively; while $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and probability density function, respectively; and γ is a parameter to balance exploration against exploitation. We should note that the acquisition functions (Eq. 3 and 4) takes the model uncertainty into account when selecting next query points, but it is still possible to select points in regions with steric clashes. In this work, we therefore define a new acquisition function that makes use of our domain knowledge, namely Knowledge-based Expected Improvement (KEI), to address this problem.

2.3.1 Knowledge-based Expected Improvement

Knowledge-based Expected Improvement (KEI) can be considered as a modified expected improvement (EI) acquisition function that offers improvement only when a set of torsion constraints are satisfied:

$$a_{KEI}(\boldsymbol{\theta}) = EI(\boldsymbol{\theta}) \prod_{m=1}^{M} P_m(\theta_{m,1}, \theta_{m,2}) \tag{5}$$

where M is the total number of correlated torsions found in the molecule, $P_m(\theta_{m,1}, \theta_{m,2})$ is the mixture model of the torsion angle pairs in pattern, m. We assume independence between each pair of correlated torsions. The idea of KEI is similar to the method of expected improvement with Boolean constraints suggested by Griffiths and Hernández-Lobato, and Gelbart et al., with a user-specified minimum confidence of the constraints.[30,31] Instead of Boolean constraints, we derived separate distributions of the correlated torsions from the lowest energy conformations found by MMFF94, and GFN2. These were encapsulated by bivariate von Mises mixture models, which are used to constrain the search.

2.3.2 Covariance Function

Since potential energy is known to be periodic with respect to dihedral angle, a locally periodic kernel, k_{LP}, which is a product of a periodic kernel and a squared exponential kernel, was used:

$$k_{LP}(\theta, \theta') = \sigma^2 \exp(\frac{-\|\theta - \theta'\|^2}{2l^2}) \exp(\frac{-2\sin^2(\pi|\theta - \theta'|/p)}{l^2}) \quad (6)$$

where l, p, and σ^2 are the length-scale, periodicity, and variance, respectively. The periodicity is determined by torsional potentials corresponding to the rotatable bond SMARTS patterns. Note that for missing patterns did not match a specific type of rotatable bond, *i.e.* did not match a SMARTS pattern, we assigned general values for the periodicity based on the atomic hybridization of the two atoms in the central rotatable bond, *i.e.* $sp^2 - sp^2$, $sp^2 - sp^3$, or $sp^3 - sp^3$. Boundary constraints were added to the length-scale in the kernel for numerical stability.

2.4 Data

We used two datasets, the Platinum dataset,[3] and a dataset assembled by Ebejer et al.,[32] to benchmark the performance of the search algorithms. Duplicated molecules in the two datasets were removed based on their InChI Key. Molecules with 2 to 18 rotatable bonds, and containing two adjacent rotatable bonds matching the set of rotatable bond-SMARTS patterns, were selected for the study, giving a set of 533 unique molecules.

We extracted small molecules with matching rotatable bonds from the Crystallography Open Database (COD), and removed duplicate molecules from the COD set that were present in both the validation set and COD based on their InChI Key. We recorded the torsional preferences in these crystal structures. In addition, we calculated the lowest energy conformations of these molecules using MMFF94 and GFN2, and recorded the resultant calculated

torsional for each. We then derived bivariate von Mises mixture models from the torsion preferences found in X-ray crystal structures, and calculated by the MMFF94 and GFN2 methods.

2.5 Evaluation

2.5.1 Energy function

Previously,[6] we computed a single-point MMFF94 force field energy while keeping the small molecule's bond lengths and bond angles fixed. Here, we relaxed the framework and evaluate the energy of the fully-optimized molecule, *i.e.* the energy value at a given set of dihedral angles was a result of a short (50 steps) MMFF94 energy minimization, with a concomittant change in bond lengths and bond angles, while the torsion angles remain fixed in geometry optimization. We used the MMFF94 implementation in Open Babel 2.4.1.[33] The configurations in the benchmark datasets were used as the initial structures.

In addition, we performed single-point energy calculations with a more accurate semi-empirical method, namely GFN2.[21,22] Similarly, the configurations in the benchmark datasets were used as initial structures. The bond length, bond angles and other parameters were inherited from the input structures.

2.5.2 Comparison

Three global optimization conformational search methods were compared: (i) a Genetic Algorithm (GA), and our Bayesian Optimization Algorithm, BOA, with two different acquisition functions: (ii) our previous expected improvement (EI) method,,[6] and (iii) the new knowledge-based expected improvement (KEI) strategy. The implementations are described below.

(i) Bayesian Optimization with EI (BOA-EI).[6] GPyOpt[34] was used for the Bayesian optimization and Pybel[35] was used to drive the torsion angles of the molecules and energy minimization. Note that torsion constraints were used in the geometry optimization step (i.e., minimization of all other degrees of freedom, bonds, angles, etc. with fixed dihedral angles). Expected improvement was used as the acquisition function. Five initial random samples were generated to begin a Gaussian Process regression. We also added boundary constraints to the length-scale parameter for the sake of numerical stability. The lowest energy conformation calculated from all iterations returned as the final output structure.

(ii) Bayesian Optimization with KEI (BOKEI). The implementation was the same as the standard Bayesian optimization in (i), except a knowledge-based acquisition function KEI was used.

(iii) Genetic Algorithm (GA). The implementation in Open Babel[33] for GA was used. Termination criterion, *i.e.* convergence was reached when three identical generations were observed, was applied in GA; all other GA parameters were left as their default values. Note that no torsion constraints were added in the energy minimization step as surrogate model was not required.

2.6 Search Space

The search space for each molecule is determined by the set of freely-rotatable bonds in each. The search space for the Bayesian optimization and its variant was defined by hypercube $[-\pi, \pi)^d$, where d is the number of rotatable bonds. A discrete grid space was used for GA.

2.6.1 Search Budget

The number of energy evaluations was determined by the number of rotatable bonds in the molecule (see Table 1). Since five initial samples were used to fit a Gaussian process in

Bayesian optimization, only $K-5$ conformations were sampled after initial sampling. For accurate statistical comparisons of these stochastic methods, five runs were performed for each algorithm.

Table 1: Sample size versus number of rotatable bonds

Number of rotatable bonds	Number of conformers (K)
2-3	25
4-5	50
6-7	100
$\geq$8	200

2.6.2 Analysis

Energy difference (ΔE) between the lowest energy conformation obtained by other methods and that from BOA-EI was calculated. The average energy difference was used to evaluate the performance of the search methods. Note that the energy. The average energy difference is calculated as follows:

$$\overline{\Delta E^{KEI}} = \frac{1}{N}\sum_{i=1}^{N}(E_{KEI,i} - E_{EI,i}) \tag{7}$$

$$\overline{\Delta E^{GA}} = \frac{1}{N}\sum_{i=1}^{N}(E_{GA} - E_{EI,i}) \tag{8}$$

where N is the number of runs. $E_{EI,i}$ and $E_{KEI,i}$ are the lowest energy found by EI and KEI in i-th run respectively. E_{GA} is the lowest energy conformation found in all runs (do not depend on i-th run). The lowest energy conformation found by BOA-EI was used as reference in both cases. Since same initialization was used in both BOKEI and BOA-EI, *i.e.* five initial samples were the same, we could directly compare the performance between them in each run. In GA, We compared its best performance to each run in BOA-EI. Two different energy functions were used in the context and we denote $\overline{\Delta E_{MMFF94}}$ and $\overline{\Delta E_{GFN2}}$ to be the average energy difference in MMFF94 and GFN2 respectively.

Wilcoxon signed-rank test was used to test whether the proposed method, BOKEI, finds

lower average energy conformations than BOA-EI and GA, *i.e.* one-sided test. We tested it across number of rotatable bonds with both MMFF94 and GFN2 energy functions.

Furthermore, we calculated the frequency of BOKEI in finding lower energy conformations than BOA-EI and GA. We also computed the pattern frequency of our derived torsions pattern across three datasets, namely Platinum dataset,[3] COD, and ChEMBL 25,[36] and compared to that of the correlated torsion patterns defined in CSD Conformer Generator.

Lastly, we performed run time analysis on BOA-EI and BOKEI, using a desktop running Fedora 30 with an Intel Core i7-6700 operating at 3.40 GHz, and 32 GB of RAM. Single core was used to read molecule, drive torsions and write conformers to disks. All cores were used in the GPyOpt operations. The time included reading input molecules and writing the conformers to disk. Fifteen molecules with two to six rotatable bonds, three for each, were sampled. We repeated the search with four times each molecule.

3 Results and Discussion

3.1 Example

Two molecules were sampled to illustrate the strength and the weakness of the BOKEI algorithm, using the geometry-optimized MMFF94 energy (see Appendix 3 for more examples with GFN2). Ten runs were performed for each molecule. From Figure 2, we observed the BOKEI was able to find lower energy conformation consistently with same number of iterations. The energy gap between BOKEI and BOA-EI decreased as the number of iterations increased, since both methods should converge to the same global minimum. We should note that the performance of BOKEI was worse than BOA-EI in this example 2b, which was a result of under-estimation of the correlated torsion distribution. Insufficient sampling or biased selection of molecules gave rise to the incomplete prior information, and led to the

inferior performance. Using additional data to improve the correlated distributions, even in this case, the performances became similar, as discussed in Section 3.7.

We revisited the example, 5-Phenylthioquinazoline-2,4-diamine, and studied the effect of the new acquisition function on the posterior. Figure S5 in Appendix 2 showed that the posterior of BOA-EI and BOKEI after twenty iterations (25 observations in total, with five initial samples). GFN2 energy function was used in this example. The posterior mean showed a coarse boundary between high and low energy regions. The observations in BOKEI were more closely packed in the low energy region, comparing to that in BOA-EI. Only a small number of observations were sampled in medium or high energy region, the posterior standard deviations thus remained high.

3.2 Dataset Summary

For a broader comparison, 533 molecules with sizes from 2–18 rotatable bonds were used to validate the performance. In GFN2, we benchmarked the performance with molecules up to 13 rotatable bonds only. The number of matches for each pattern in the validation set is summarized in Appendix 3 Figure S7. Five correlated SMARTS patterns are frequently found, with frequency greater than 100. Note that there were nine molecules (five in MMFF94 and four in GFN2) excluded from analysis due to early stopping in one of the five runs (see Appendix 4 Table S8 and S9). This was manifested by a non-positive definite kernel error.

3.3 MMFF94

Figure 3a showed that BOKEI consistently found lower energy conformation than BOA-EI and GA. A Wilcoxon signed rank test showed that energy difference between BOKEI and BOA-EI was statistically significant ($p << 0.01$, see Appendix 3 Table S4) across all

rotatable bonds. On the other hand, we observed that the GA outperformed BOKEI and BOA-EI for molecules with more than twelve rotatable bonds. This was because the small number of samples (200 energy evaluations) may not be sufficient for the BOA-EI or BOKEI models to learn the most likely dihedral angles in high dimensional problems. Figure S6a in Appendix 3 showed that BOKEI frequently ($> 63\%$ and $> 70\%$) found lower energy conformations than BOA-EI and GA for molecules with fewer than eleven rotatable bonds respectively

Furthermore, Figure 2 highlighted that BOKEI showed greater benefit earlier on in the evaluations. Comparing the mean energy difference between BOKEI and BOA-EI at different stages (40%, 60% and 100% of the maximum number of energy evaluations), the energy gap was indeed greater, and in favor of BOKEI, in the early stages (see Appendix 3, Table S6). The gap diminished when more evaluations were used, since both methods converged to the same global optimum. These results suggest that the information about correlated torsions greatly helped the search in the earlier stages, pointing the search towards favorable regions of the potential energy landscape.

3.4 GFN2

In GFN2, we used a single-point energy calculation, and excluded GA in the analysis. Figure 3b showed that BOKEI consistently found lower energy conformations than BOA-EI. Similarly, Wilcoxon signed rank test showed the average energy difference between BOKEI and BOA-EI was statistically significant ($p << 0.01$, see Appendix 3 Table S5). The energy gap was greater in the early stage and the gap diminished as more energy evaluations were used (see Appendix 3 Table S7). In addition, Figure S6b in Appendix 3 showed that BOKEI frequently ($> 60\%$) found lower energy conformations than BOA-EI across all rotatable bonds.

3.5 Correlated Torsion

When adjacent rotatable bonds have correlated dihedral angles, this is typically caused by unfavorable steric interactions — but not always. Four of the nineteen patterns in our library arise because of favorable intramolecular interactions, such as hydrogen bonds and π-π stacking. For example, in pattern 15, the lowest energy conformations found by GFN2 often form intramolecular hydrogen bonds between the N-H or O-H groups and the adjacent carbonyl oxygen atoms in the esters (see Figure 4a).

The thioamide functional group is a key part of patterns 2 and 16 (see Figure 4b). The delocalization of the nitrogen lone pairs in this group contributes to its overall planarity, but it could exist in either the *cis* or *trans* form. The orientation of the aromatic ring in pattern 2 is thus highly constrained. The *cis*/*trans* preference is easily revealed by examining higher-order correlated torsions, *i.e.* between three adjacent rotatable bonds. In particular, we considered a specific thiourea derivative that is bonded to a carbonyl group (see Appendix 1 Table S2). This is always observed to adopt the following conformation: (i) the C=S and C=O are oriented in 'opposite' directions, while (ii) the thiourea adopts the syn-anti[37] conformation. This results in the formation of a pseudo six-membered ring that is stabilized by a C=O - - H-N intramolecular hydrogen bond (see for example, Figure 4b). Figure S3 in Appendix 3 shows these three torsion angles are highly concentrated around 0° in COD.

For pattern 17, π-π stacking is evident: when aromatic rings are attached to both ends of the pattern, both rings prefer to interact with one another (see Fig 4c).

It should be noted that the CSD Conformer Generator[9] also considers 11 correlated torsions, but a simple clash term are used for all other interactions. Here, we use a more flexible approach that employs bivariate von Mises mixture models to fully describe the correlated torsions. Both favorable intramolecular interactions and unfavorable steric clashes can be described. It would also possible to expand this to a multivariate case,[38] in order to capture higher-order correlations as mentioned earlier. Additionally, the framework is easy to inte-

grate with other conformer sampling schemes, such as distance geometry.[17,26] We intend to integrate with the ETKDG in RDKit in the future as the current implementation does not consider the correlated torsion.

In addition, we observed the torsion preference of MMFF94 differed from the one in GFN2 and crystal structures in pattern 1, 3 and 12. It suggested some of the molecular interactions could not be represented in the classical force field. In pattern 2, we notice a discrepancy between the crystal conformation and the lowest energy conformation in GFN2 – the second torsion angles (*i.e.* the torsion angles measured from the thioamide group) are highly concentrated around 180°, *i.e.* the trans-form. This could be explained by the separate hydrogen bond interaction of the N-H group and the C=S in the crystal.

Furthermore, we compiled the number of the molecules with the presence of correlated torsions in three different datasets: (i) Platinum, (ii) COD and (iii) ChEMBL 25 (see Table 2). We showed that our SMARTS patterns library surprisingly matched $10 - 15\%$ organic molecules, which was noticeably higher than the patterns defined in CSD Conformer Generator ($1 - 4\%$). These results suggest that broader investigation of correlated torsions is warranted, despite the conventional assumption of each rotatable bond as an independent free rotor.

Table 2: Frequency of molecules with the presence of correlated torsion patterns, comparing this work to previous steric constraints[9] across various databases, including the Open Crystallographic Database (COD).

Dataset	Number of Molecules	% Matches (New)	% Matches (CSD)
Platinum	4,548	9.2	2.5
COD	110,623	13.5	1.6
ChEMBL 25	1,870,461	14.6	3.6

3.6 Computational Time

Figure 5 showed the average run time of BOA-EI and BOKEI with GFN2 energy function, varying number of iterations (50, 100) and number of rotatable bonds (two to six rotatable bonds). Both computational cost increased as the number of rotatable bonds increased. The computational time also increase when BOKEI was used, but was primarily dominated by the number of conformers generated. Note that the current implementation can be further optimized by providing the gradient information of the acquisition function.

In theory, extra multiplication in the BOKEI acquisition function increases the computational complexity to $O(mn)$, where m is the number of correlated torsions found in a molecule, and n is the number of samples used to evaluate the acquisition function. We should note that the relative contribution to the computational time of the new acquisition function will be small, when a more accurate and computational expensive method, such as density functional theory (DFT), is used for energy evaluation. Our new algorithm will be more cost-effective than the old version in such settings.

3.7 Limitation and Future work

We only calculated the lowest energy conformation of the molecules with up to five rotatable bonds in the COD set, and the low energy region of the correlated dihedral could be incomplete. This limited the performance of the algorithm. This issue can be easily solved by increase sampling of the corresponding substructures in a larger database, for instance ChEMBL[36] and PubChem,[39] and re-estimate the distribution. Figure 6a showed the original prior used in Example 2b. We observed that the cluster centroids shifted when observations from ChEMBL were used (see Figure 6b), and Figure 6c showed a great improvement in convergence rate comparing to the original case.

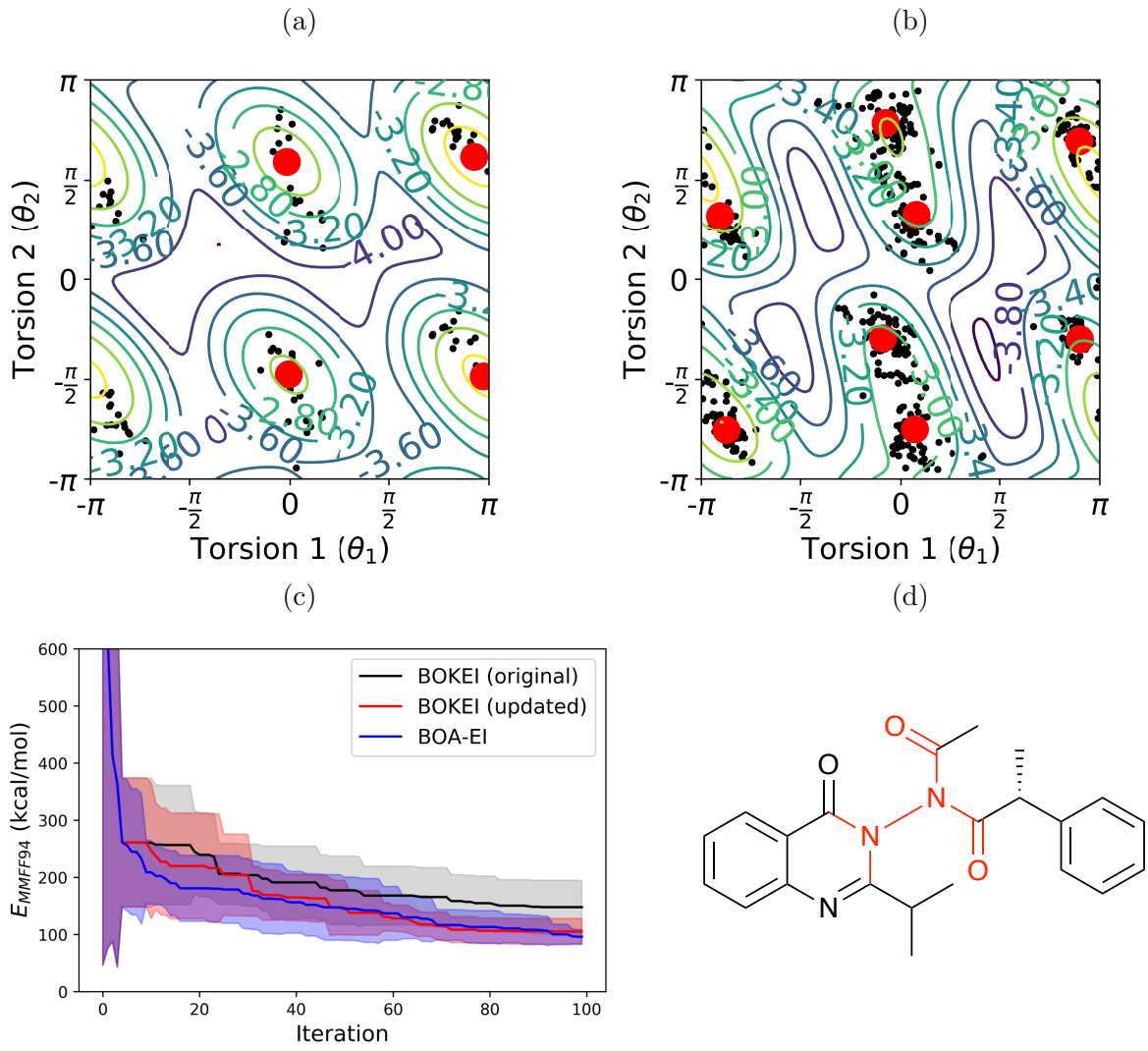


Figure 6: **(a)** Mixture model derived from the COD dataset. **(b)** Mixture model derived from COD and ChEMBL database. The contour plot indicates the log density of a mixture model and the points (in red) mark the mean location for the components. **(c)** Convergence plot. **(d)** Molecule that used to validate the performance of the updated prior.

In addition, hundreds of observations were typically required to fit the mixture models. Cole et al. derived a set of SMARTS patterns, and we did not use all of them due to insufficient observations. In order to overcome this obstacle, we could apply a meta-learning approach proposed by Ton et al.,[40] which attempt to learn the conditional distribution of the correlated torsion. They showed that the approach could generalise the density of the correlated torsions with few observations. This meta-learning setting could greatly reduce the computational

cost in learning torsion rules, and potentially help discover more unexpected torsion patterns for the sampling scheme.

4 Conclusions

By reformulating the search of the lowest energy conformation of a given molecule as a constrained Bayesian optimization problem, we have shown concrete improvements. Prior knowledge of correlated torsions were used to confine the exploration to regions of low energy. We compared the Bayesian optimization with two different acquisition functions: standard expected improvement (EI) and knowledge-based expected improvement (KEI), and genetic algorithm (GA), using two energy functions: MMFF94 and GFN2. We showed that with the same number of energy evaluations, the Bayesian optimization with KEI (BOKEI) frequently ($> 60\%$) found lower energy conformations (median energy difference 1.95 kcal/mol in MMFF94 and 1.54 kcal/mol in GFN2) than the Bayesian optimization with EI (BOA-EI) in both cases, across all rotatable bonds.

Importantly, using bivarivate von Mises mixture models to describe the correlated dihedral allowed us to capture correlation that could not be explained by simple clash terms, and it could be integrated into other conformer sampling frameworks easily. Furthermore, we showed that the correlated torsion not only reflect steric clashes, but also favorable intramolecular interactions such as hydrogen bonds and π-π stacking.

Future work should focus on expanding data sources, to ensure sufficient sampling across a wide range of correlated dihedrals including other types of neighbors, non-nearest neighbors. Moreover, ring torsions were not investigated, which are well-known to involve correlations torsional motion (e.g. Cremer-Pople angles and ring pucker).[41,42] Such efforts will improve the efficiency in sampling low-energy conformers for applications in property-driven drug design, materials screening, and crystal structure prediction.

Availability of data and materials

All the data and code will be available online and GitHub `https://github.com/lucianlschan/Conformer-Geometry-v2`. See DOI: 10.26434/chemrxiv.9209213.

Conflicts of interest

There are no conflicts to declare.

Acknowledgement

GRH thanks the National Science Foundation (CHE-1800435) for support. GMM thanks the EPSRC and MRC for financial support under grant number EP/L016044/1. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work, and supported in part by the University of Pittsburgh Center for Research Computing through the computational resources provided. We also thank Jotun Hein, Susan Leung, Carlos Outeiral for helpful discussion.

Supporting Information Available

- Appendices 1: List of correlated torsion SMARTS patterns
- Appendices 2: Bivariate von Mises Distribution and EM Algorithm
- Appendices 3: Tables and Figures
- Appendices 4: Benchmark Set and Molecules Excluded from Analysis

References

(1) Hawkins, P. C. D. Conformation Generation: The State of the Art. *Journal of Chemical Information and Modeling* **2017**, *57*, 1747–1756.

(2) Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *Journal of Chemical Information and Modeling* **2017**, *57*, 2719–2728.

(3) 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. *Journal of Chemical Information and Modeling* **2017**, *57*, 529–539.

(4) Grimme, S.; Bannwarth, C.; Dohm, S.; Hansen, A.; Pisarek, J.; Pracht, P.; Seibert, J.; Neese, F. Fully Automated Quantum-Chemistry-Based Computation of Spin-Spin-Coupled Nuclear Magnetic Resonance Spectra. *Angewandte Chemie International Edition* **2017**, *56*, 14763–14769.

(5) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *Journal of Chemical Theory and Computation* **2019**, *15*, 2847–2862.

(6) Chan, L.; Hutchison, G. R.; Morris, G. M. Bayesian optimization for conformer generation. *Journal of Cheminformatics* **2019**, *11*, 32.

(7) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic generation of diverse low-energy conformers. *Journal of Cheminformatics* **2011**, *3*, 8.

(8) Gardner, J. R.; Pleiss, G.; Bindel, D.; Weinberger, K. Q.; Wilson, A. G. GPyTorch:

Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. Advances in Neural Information Processing Systems. 2018.

(9) Cole, J. C.; Korb, O.; McCabe, P.; Read, M. G.; Taylor, R. Knowledge-Based Conformer Generation Using the Cambridge Structural Database. *Journal of Chemical Information and Modeling* **2018**, *58*, 615–629.

(10) Friedrich, N.-O.; Flachsenberg, F.; Meyder, A.; Sommer, K.; Kirchmair, J.; Rarey, M. Conformator: A Novel Method for the Generation of Conformer Ensembles. *Journal of Chemical Information and Modeling* **2019**, *59*, 731–742.

(11) Kothiwale, S.; Mendenhall, J. L.; Meiler, J. BCL::Conf: small molecule conformational sampling using a knowledge based rotamer library. *Journal of Cheminformatics* **2015**, *7*, 47.

(12) Hawkins, P. C.; 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*, 572–84.

(13) 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 Research* **2000**, *28*, 235–242.

(14) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallographica Section B* **2016**, *72*, 171–179.

(15) Guba, W.; Meyder, A.; Rarey, M.; Hert, J. Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations of Small Molecules. *Journal of Chemical Information and Modeling* **2016**, *56*, 1–5.

(16) Landrum, G. RDKit: Open-Source Cheminformatics. Available at `http://www.rdkit.org`, 2018; `http://www.rdkit.org`.

(17) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling* **2015**, *55*, 2562–2574.

(18) Gražulis, S.; Chateigner, D.; Downs, R. T.; Yokochi, A. F. T.; Quirós, M.; Lutterotti, L.; Manakova, E.; Butkus, J.; Moeck, P.; Le Bail, A. Crystallography Open Database – an open-access collection of crystal structures. *Journal of Applied Crystallography* **2009**, *42*, 726–729.

(19) Gražulis, S.; Daškevič, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quirós, M.; Serebryanaya, N. R.; Moeck, P.; Downs, R. T.; Le Bail, A. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research* **2012**, *40*, D420–D427.

(20) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry 17*, 490–519.

(21) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *Journal of Chemical Theory and Computation* **2017**, *13*, 1989–2009.

(22) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019**, *15*, 1652–1671.

(23) Mardia, K. V.; Taylor, C. C.; Subramaniam, G. K. Protein Bioinformatics and Mixtures of Bivariate von Mises Distributions for Angular Data. *Biometrics* **2007**, *63*, 505–512.

(24) Boomsma, W.; Mardia, K. V.; Taylor, C. C.; Ferkinghoff-Borg, J.; Krogh, A.; Hamelryck, T. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, *105 26*, 8932–7.

(25) Mardia, K. V.; Frellsen, J. In *Bayesian Methods in Structural Bioinformatics*; Hamelryck, T., Mardia, K., Ferkinghoff-Borg, J., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; pp 159–178.

(26) Spellmeyer, D. C.; Wong, A. K.; Bower, M. J.; Blaney, J. M. Conformational analysis using distance geometry methods. *Journal of Molecular Graphics and Modelling* **1997**, *15*, 18 – 36.

(27) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, Massachusetts,, 2006.

(28) 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. *CoRR* **2010**, *abs/1012.2599*.

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

(30) Griffiths, R.-R.; Hernández-Lobato, J. M. Constrained Bayesian Optimization for Automatic Chemical Design. *arXiv e-prints* **2017**, arXiv:1709.05501.

(31) Gelbart, M. A.; Snoek, J.; Adams, R. P. Bayesian Optimization with Unknown Constraints. Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence. 2014; pp 250–259.

(32) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation

Methods: How Good Are They? *Journal of Chemical Information and Modeling* **2012**, *52*, 1146–1158.

(33) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*, 33.

(34) Authors, G. GPyOpt: A Bayesian Optimization framework in Python. `http://github.com/SheffieldML/GPyOpt`, 2016.

(35) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal* **2008**, *2*, 5.

(36) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Research* **2016**, *45*, D945–D954.

(37) Sahu, S.; Rani Sahoo, P.; Patel, S.; Mishra, B. Oxidation of thiourea and substituted thioureas: a review. *Journal of Sulfur Chemistry* **2011**, *32*, 171–197.

(38) Mardia, K. V.; Hughes, G.; Taylor, C. C.; Singh, H. A Multivariate Von Mises Distribution with Applications to Bioinformatics. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **2008**, *36*, 99–109.

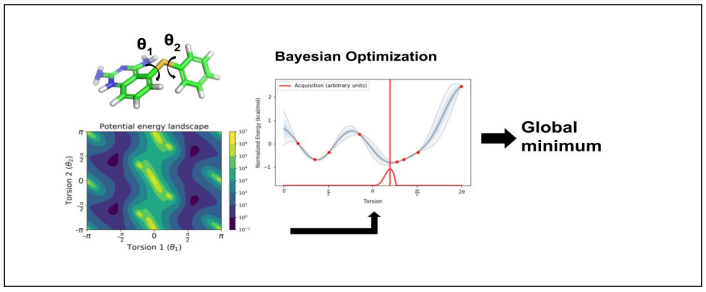
(39) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research* **2018**, *47*, D1102–D1109.

(40) Ton, J.-F.; Chan, L.; Teh, Y. W.; Sejdinovic, D. Noise Contrastive Meta-Learning for Conditional Density Estimation using Kernel Mean Embeddings. *arXiv e-prints* **2019**, arXiv:1906.02236.

(41) Cremer, D.; Pople, J. A. General definition of ring puckering coordinates. *Journal of the American Chemical Society* **1975**, *97*, 1354–1358.

(42) Hill, A. D.; Reilly, P. J. Puckering Coordinates of Monocyclic Rings by Triangular Decomposition. *Journal of Chemical Information and Modeling* **2007**, *47*, 1031–1035, doi: 10.1021/ci600492e.

Graphical TOC Entry



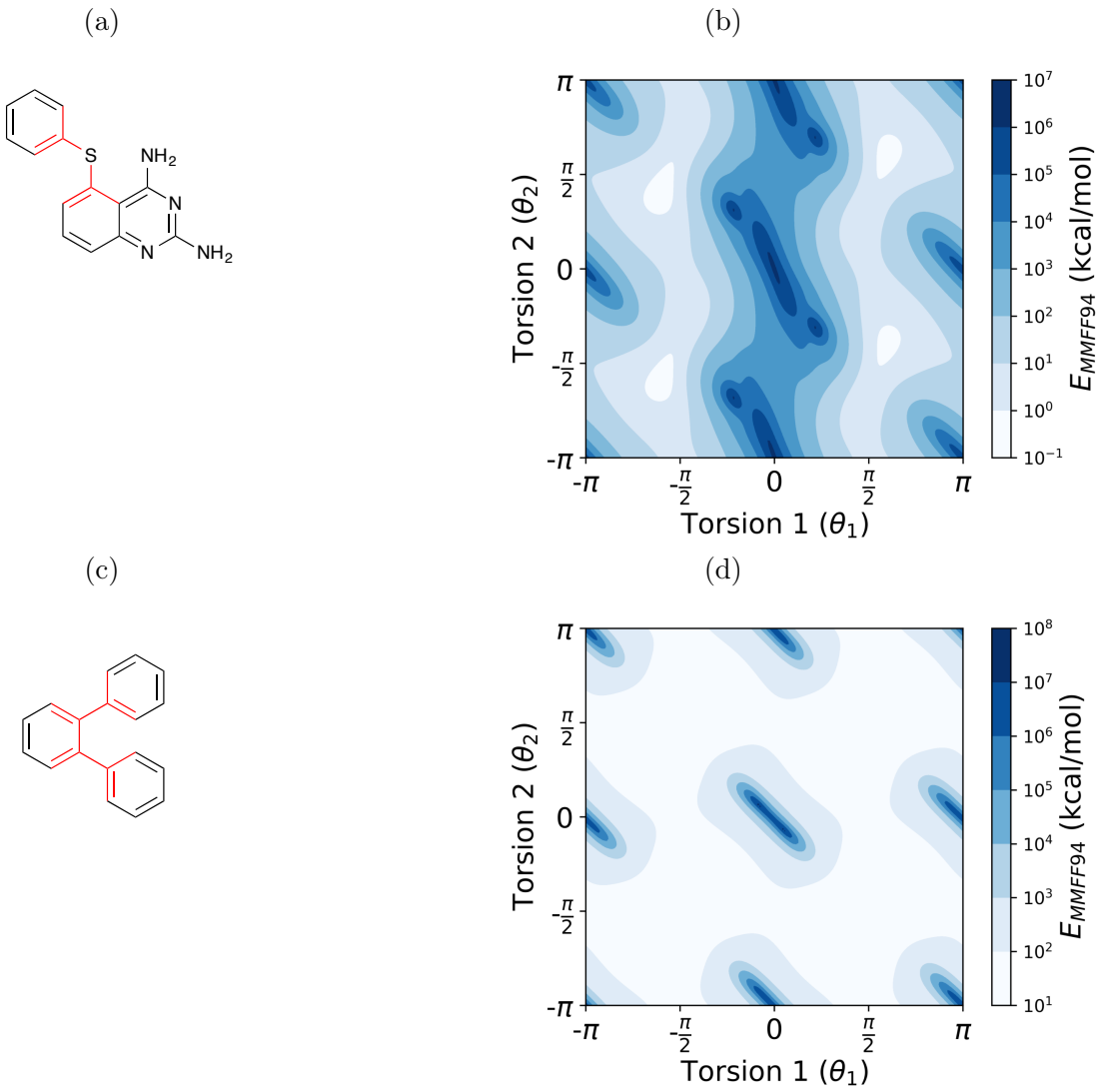


Figure 1: **(a)** 5-Phenylthioquinazoline-2,4-diamine. **(b)** MMFF94 potential energy landscape for 5-Phenylthioquinazoline-2,4-diamine. **(c)** *ortho*- 1,1′:2′,1″-terphenyl. **(d)** MMFF94 potential energy landscape for *ortho*- 1,1′:2′,1″-terphenyl. The areas in light blue show the lowest-energy regions. Torsion angles are measured in radians. The correlated torsions in the molecules are highlighted in red.

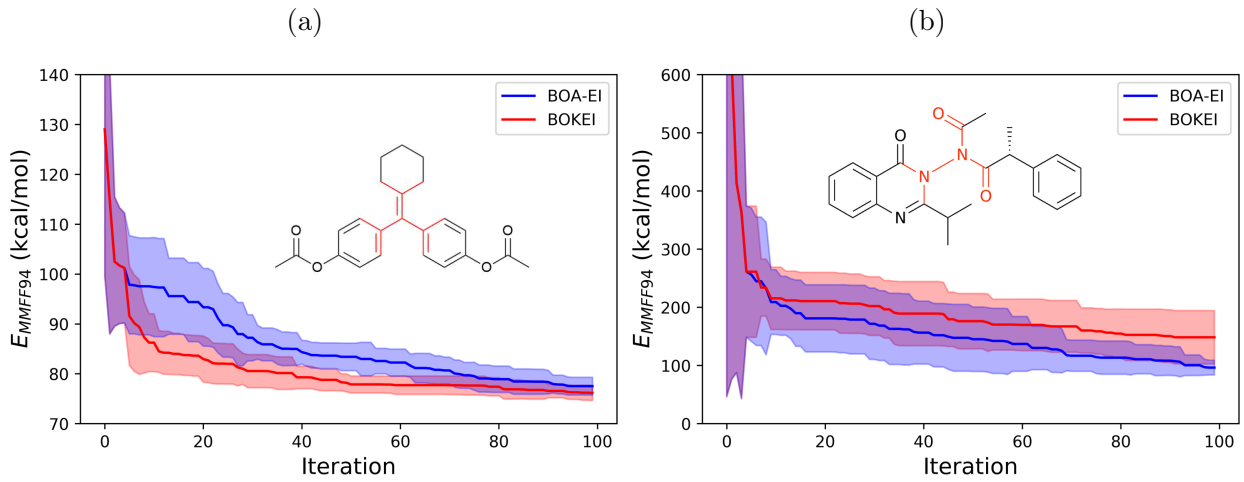


Figure 2: The red line and the blue line in the convergence plots represented average rate of the BOKEI and BOA-EI in finding lower energy conformations respectively, with ±1 sample standard deviation (shaded region). The corresponding molecule and the correlated torsion is highlighted in red. Geometry-optimized MMFF94 energy function was used in both cases. More examples with GFN2 energy function could be found in Appendix 3 Figure S4. **(a)** BOKEI consistently found lower energy conformations than BOA-EI in early stage and the energy gap reduced as the number of iterations increased. **(b)** Contrarily, BOKEI performed worse than BOA-EI, which was a result of under-estimation of the correlated torsion.

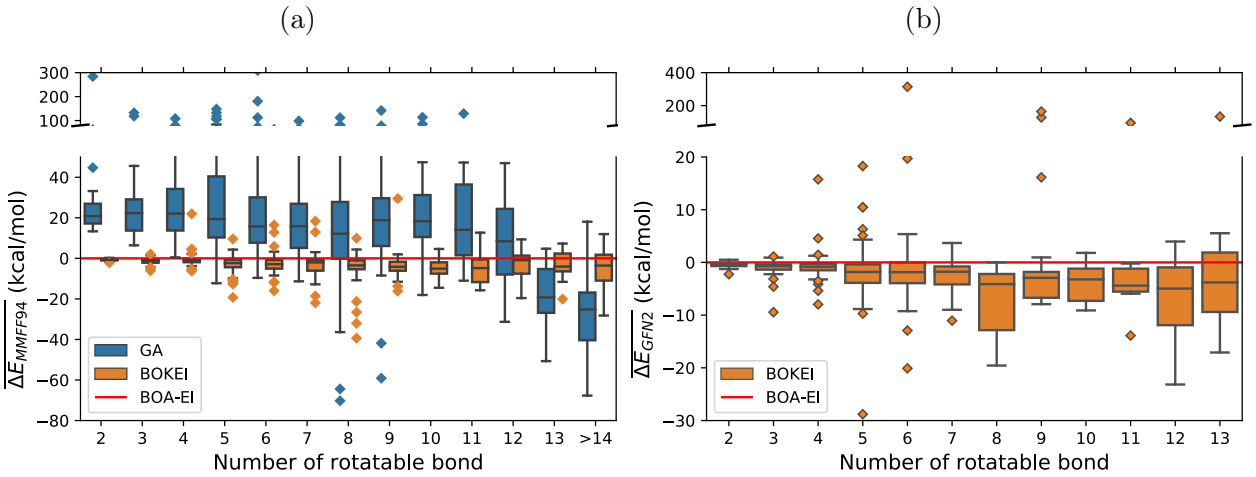


Figure 3: **(a)**: MMFF94 average energy difference from five runs. **(b)** GFN2 average energy difference from five runs. The average energy of the outputs from all runs found by BOA-EI was used as the reference point (red line) in (a) and (b). The BOKEI often found lower energy conformations than BOA-EI in both cases. The GA in MMFF94 outperformed BOKEI and BOA-EI for molecules with eleven or more rotatable bonds.

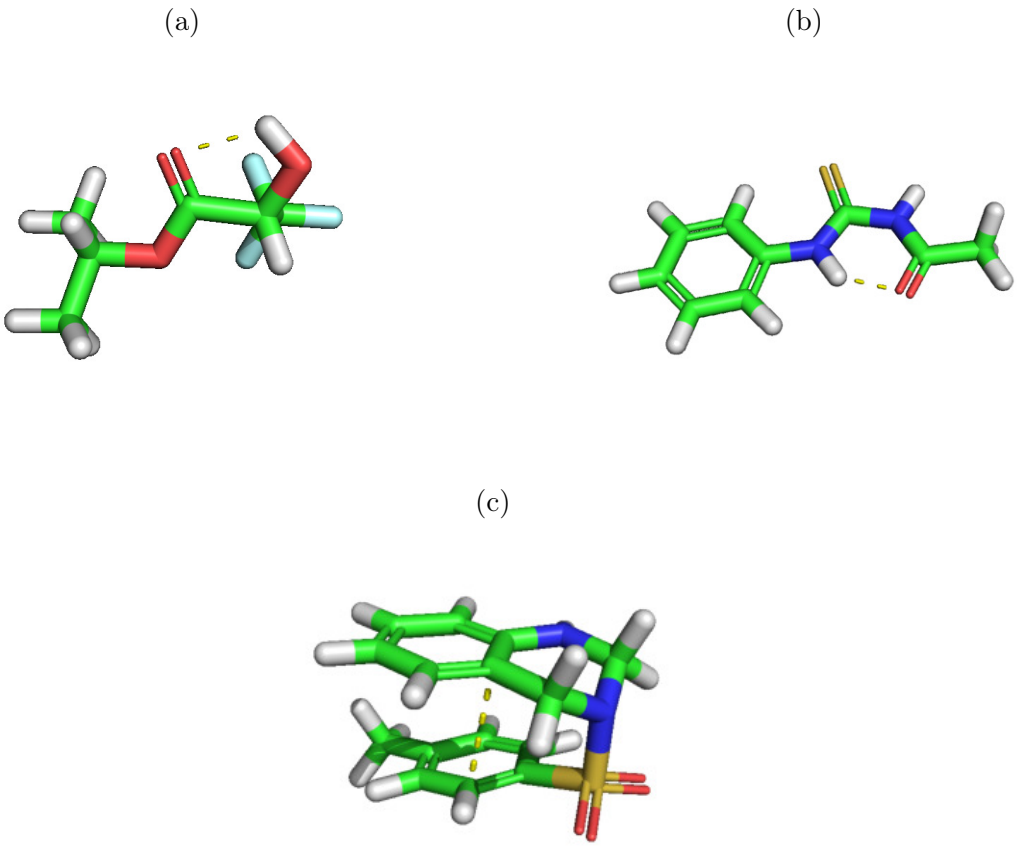


Figure 4: Intramolecular H-bonds are observed in pattern 15 (a) and patterns 2 and 16 (b); while intramolecular π-π stacking is evident in pattern 17 (c).

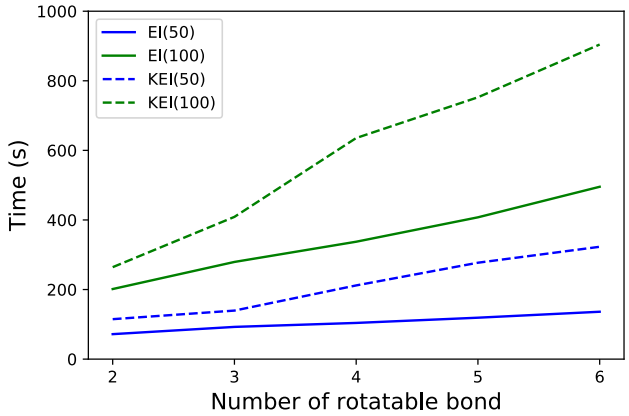


Figure 5: Average computational time for BOA-EI and BOKEI with GFN2 energy function, using different number of energy evaluations (50, 100). The computational time increased as the number of rotatable bonds increase, but was dominated by the number of conformers generated.