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

Lucian Chan[1], Geoffrey R. Hutchison[2], and Garrett M. Morris[1]

[1]Department of Statistics, University of Oxford, 24-29 St Giles, Oxford, OX1 3LB, UK
[2]Department of Chemistry and Chemical Engineering, University of Pittsburgh, 219 Parkman Avenue, Pittsburgh, PA 15260, USA

Appendices 1: List of correlated torsion SMARTS patterns

The correlated torsion SMARTS patterns and the corresponding atom numbers that define the torsion angles are listed in Table S1. Higher order correlated torsion, *i.e.* three adjacent rotatable bonds, and the atom numbers that define the torsion angles are listed in S2.

Table S1: SMARTS patterns and the atom numbers that define the torsion angles. The atom numbers (1-4) and (5-8) define the first and the second torsion angles respectively. The SMARTS pattern in bold is the pattern defined by Cole et al. [1]

Pattern Number	SMARTS	1	2	3	4	5	6	7	8
1	[a][a]!@;-[CX3](=[CX3])!@;-[a][a]	0	1	2	4	1	2	4	5
2	[a][a]!@;-[NX3H1]!@;-[CX3](=S)[!#1]	0	1	2	3	1	2	3	5
3	[a][c]!@;-[CX4H0]([CX3,N])!@;-[c][a]	0	1	2	4	1	2	4	5
4	[a][c]!@;-[CX4H1]([N,O,H])!@-[c][a]	0	1	2	4	1	2	4	5
5	[a][c]!@;-[CH2]!@;-[n][a]	0	1	2	3	1	2	3	4
6	[a][c]!@;-[CX4H2]!@;-[OX2][C]	0	1	2	3	1	2	3	4
7	[a][c]!@;-[CX4H1]!@;-[OX2][!#1]	0	1	2	3	1	2	3	4
8	[c][c]!@;-[CX4]!@;-[c][c]	0	1	2	3	1	2	3	4
9	[cH0][c]!@;-[CX4H2]!@;-[a][a]	0	1	2	3	1	2	3	4
10	[cH0][c]([cH0])!@;-[NX3]!@;-[a][a]	0	1	3	4	1	3	4	5
11	[cH1][c]([cH1])!@;-[NX3]([CX4])!@;-[a][a]	0	1	3	5	1	3	5	6
12	[!#1][c]!@;-[SX2]!@;-[c][aH1,aH0]	0	1	2	3	1	2	3	4
13	[!#1][NX3H0]!@;-[C](=O)!@;-[O][CH0]	0	1	2	4	1	2	4	5
14	[CX3,CX4][CX4H2]!@;-[C](=O)!@;-[O]~[C]	0	1	2	4	1	2	4	5
15	[N,O,NH1,OH1][CX4]!@;-[C](=O)!@;-[O]~[C]	0	1	2	4	1	2	4	5
16	**[C,c][NH]!@;-[C](=S)!@;-[NH][C,c]**	0	1	2	4	1	2	4	5
17	[aH1][c]([aH1])!@;-[$(S(=O)=O)]!@;-[NX3H0][*]	0	1	3	4	1	3	4	5
18	[O]=[C]!@;-[O]!@;-[CX4H0][!#1]	0	1	2	3	1	2	3	4

Continued on next page

Table S1: SMARTS patterns and the atom numbers that define the torsion angles. The atom numbers (1-4) and (5-8) define the first and the second torsion angles respectively. The SMARTS pattern in bold is the pattern defined by Cole et al. [1]

Pattern Number	SMARTS	1	2	3	4	5	6	7	8
19	[O]=[C]!@;-[NX3H0](A)!@;-[a][cH0]	0	1	2	4	1	2	4	5

Table S2: Higher order correlated torsion SMARTS pattern and the atom numbers that define the torsion angles. Atoms (1-4), (5-8) and (9-12) defines the first, second and third torsion angles respectively.

SMARTS	1	2	3	4	5	6	7	8	9	10	11	12
[#1][N](c(c)c)!@;-[C](=S)!@;-[NH1]!@;-[C](=O)	0	1	5	7	1	5	7	8	5	7	8	9

Appendices 2: Bivariate von Mises distribution and EM algorithm

Bivariate von Mises distribution and EM algorithm

The (univariate) von Mises distribution (Eq. 1) is a continuous probability distribution on the circle, and it is the circular analogue of the normal distribution. This distribution has been used to model angular data, such as torsion angles and bond angles in molecules [2].

$$f(\theta) = \frac{\exp(\kappa \cos(\theta - \mu))}{2\pi I_0(\kappa)} \tag{1}$$

The parameters μ, κ, $I_0(\kappa)$ are the mean, concentration parameters and the modified Bessel function of order 0 respectively. This distribution is unimodal and symmetrical around the mode (or mean) μ. Large value of κ indicates the high concentration around the mode, while the distribution is reduced to uniform when $\kappa = 0$.

In order to study the correlated torsion, we extend the univariate von Mises distribution and jointly model the correlated conformational angles (θ_1, θ_2) with a bivariate von Mises distribution. Note that there are various (simplified) versions of bivariate von Mises distribution (see [3, 4]), namely Sine model and Cosine model. The Cosine model (Eq. 2 and 3) was used in our implementation.

Cosine density with positive interaction

$$f(\theta_1, \theta_2) = c(\kappa_1, \kappa_2, \kappa_3) \exp\{\kappa_1 \cos(\theta_1 - \mu) + \kappa_2 \cos(\theta_2 - \nu) - \kappa_3 \cos(\theta_1 - \mu - \theta_2 + \nu)\} \tag{2}$$

Cosine density with negative interaction

$$f(\theta_1, \theta_2) = c(\kappa_1, \kappa_2, \kappa_3) \exp\{\kappa_1 \cos(\theta_1 - \mu) + \kappa_2 \cos(\theta_2 - \nu) - \kappa_3 \cos(\theta_1 - \mu + \theta_2 - \nu)\} \tag{3}$$

where $c(\kappa_1, \kappa_2, \kappa_3)^{-1}$ is the normalizing constant with the following form:

$$c(\kappa_1, \kappa_2, \kappa_3)^{-1} = (2\pi)^2 \{I_0(\kappa_1) I_0(\kappa_2) I_0(\kappa_3) + 2 \textstyle\sum_{p=1}^{\infty} I_p(\kappa_1) I_p(\kappa_2) I_p(\kappa_3)\}$$

$I_r(\cdot)$ denotes the modified Bessels function of the first kind and order r. 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 Cosine densities are unimodal if $\kappa_3 < \frac{\kappa_1 \kappa_2}{\kappa_1 + \kappa_2}$ and is bimodal if $\kappa_3 > \frac{\kappa_1 \kappa_2}{\kappa_1 + \kappa_2}$. The random variables (θ_1, θ_2) are approximately bivariate normally distributed if and only if $\kappa_3 < \frac{\kappa_1 \kappa_2}{\kappa_1 + \kappa_2}$. Furthermore, the cosine density is flexible which allows us to consider transformations of all the data when estimating the model parameters. In particular, the cosine density with negative interaction can be obtained by transforming $(\theta_1, \theta_2) \mapsto (\theta_1, -\theta_2)$ in the model of cosine density with positive interaction (see [4]). In practice, we check the likelihood for original data and the transformed data, and select the one with larger value.

Typically, there are multiple modes in the torsional space (see Figure S1) and a single bivariate von Mises distribution is not sufficient to describe the correlated torsion. Therefore, a mixture model (Eq. 4) is commonly used.

$$f_M = \sum_{j=1}^{K} \pi_j f_j(x, y) \tag{4}$$

where K is the number of components, f_j denotes a cosine density with parameters, and π_j is the weight of each component (with $\sum_i \pi_i = 1$).

We used the Expectation Maximisation (EM)[5] to fit the Eq.4. It is well-known that the EM algorithm can easily get stuck in the local optimal. Hence we performed the EM algorithm multiple times with different initialisation, and chose the best final solution. We also excluded any solutions with extremely high concentration. In the M-step, gradient ascent algorithm was used to update the model parameters.

In this analysis, we considered the correlated torsion in three cases: (i) the conformations that observed in crystal structure, (ii) the lowest energy conformation from Merck Molecular Force Field (MMFF94)[6], and (iii) the lowest energy conformation from semi-empirical method, GFN2[7]. The calculation of the lowest energy conformation, the resulting parameters of the mixture models and the corresponding contour plots are summarised below.

Simulation of the lowest energy conformation

We calculated the lowest energy conformation of the molecules from the COD set, and used it to derive correlated torsion distribution. We only considered molecules with five or fewer rotatable bonds in this calculation. Under this setting, we could find the lowest energy conformation for both GFN2 and MMFF94 with high probability. The sampling schemes are described below.

MMFF94

We simulated diverse conformers by ETKDG[8] followed by energy minimization, and calculated the lowest energy conformation. The implementation in RDKit[9] was used. Note that this is a basin-hopping style optimization [10]. Here we sampled a large set of conformers (see Table S3) for molecules in the COD set.

Table S3: Number of simulated conformations versus number of rotatable bonds

Rotor Size	Number of conformers
2	50
3	100
4	250
5	500

GFN2

We calculated the lowest energy conformation using the Conformer-Rotamer Ensemble Sampling Tool[11] (CREST) based on GFN-xtb method. iMTD-GC workflow was used in the search. Note that we failed to simulate some of the molecules with CREST due to some computational issues. Hence, we may observe fewer observations than that in MMFF94.

Correlated Torsion Plots

There are nineteen correlated torsion SMARTS patterns. Three plots are given for each SMARTS pattern: torsion preferences in crystal structures, the lowest energy conformation in MMFF94 and GFN2 respectively. All torsion angles are measured in radian. For each figure, the contour plot indicates the log density of a mixture model and the points (in red) mark the mean location for each component. All parameters of each mixture are listed below the plots. μ, ν denote the mean location of the pair of torsion respectively. ω represents the weight of the mixture. κ_1, κ_2, are the concentrations. κ_3 is the parameter controlling the correlation respectively.

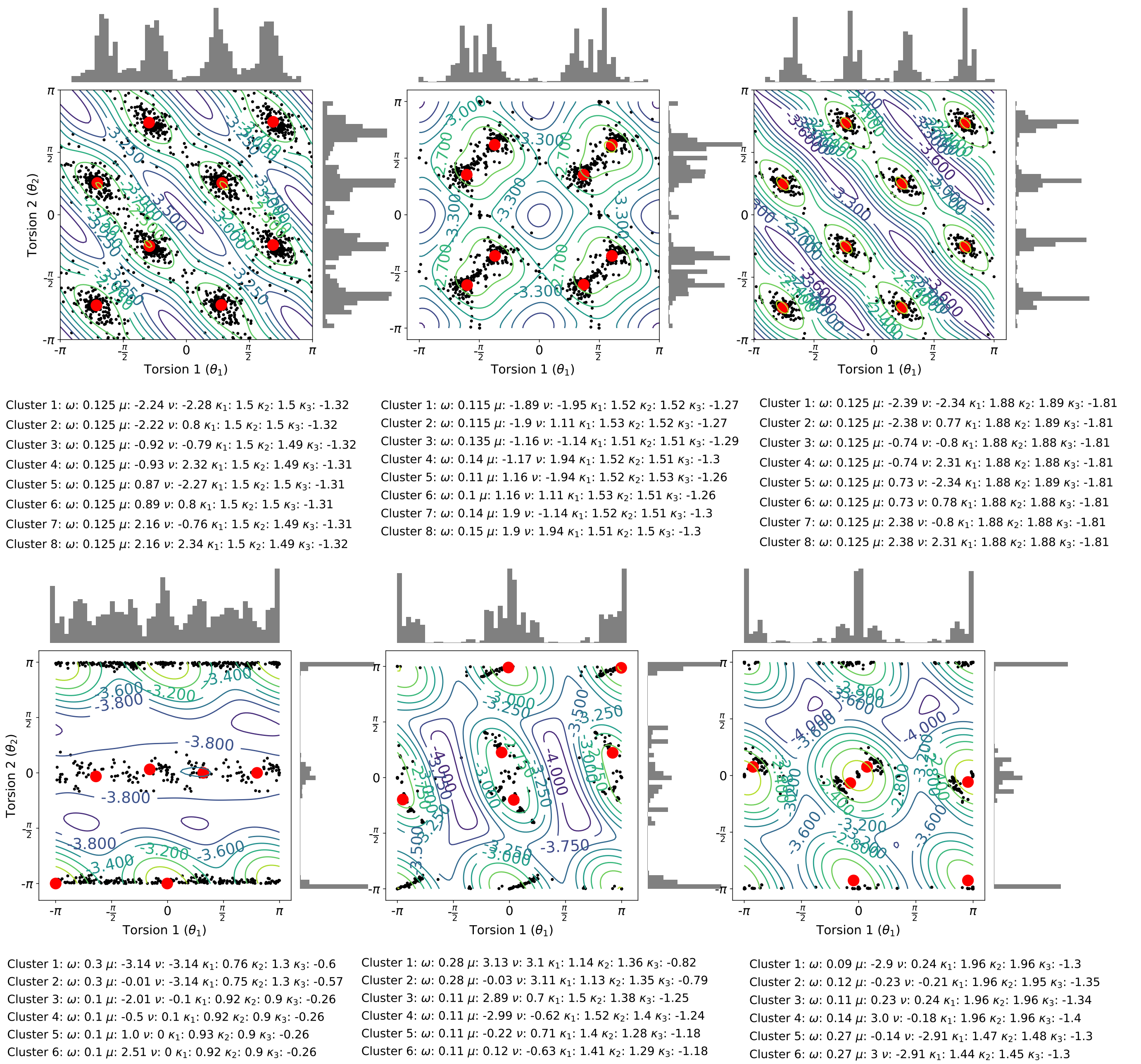
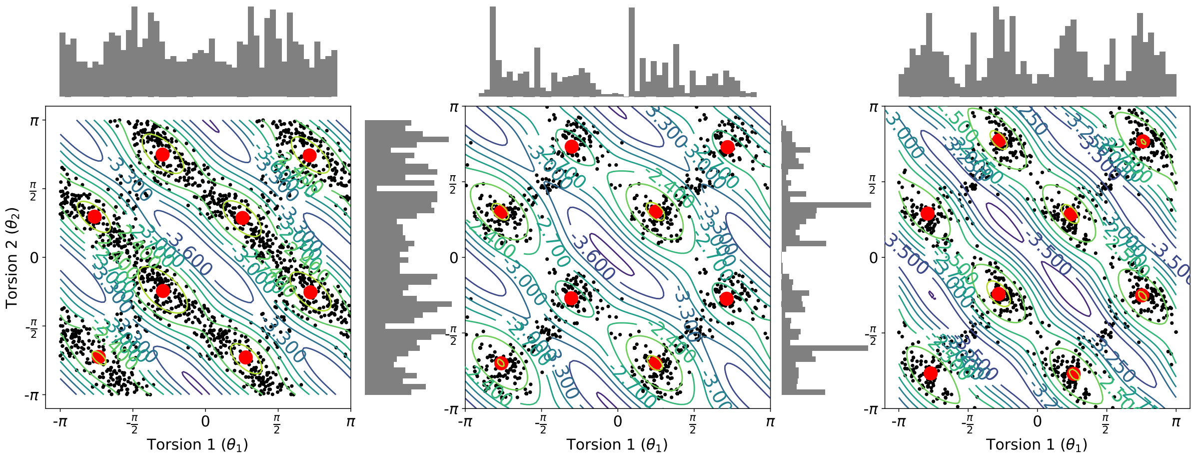


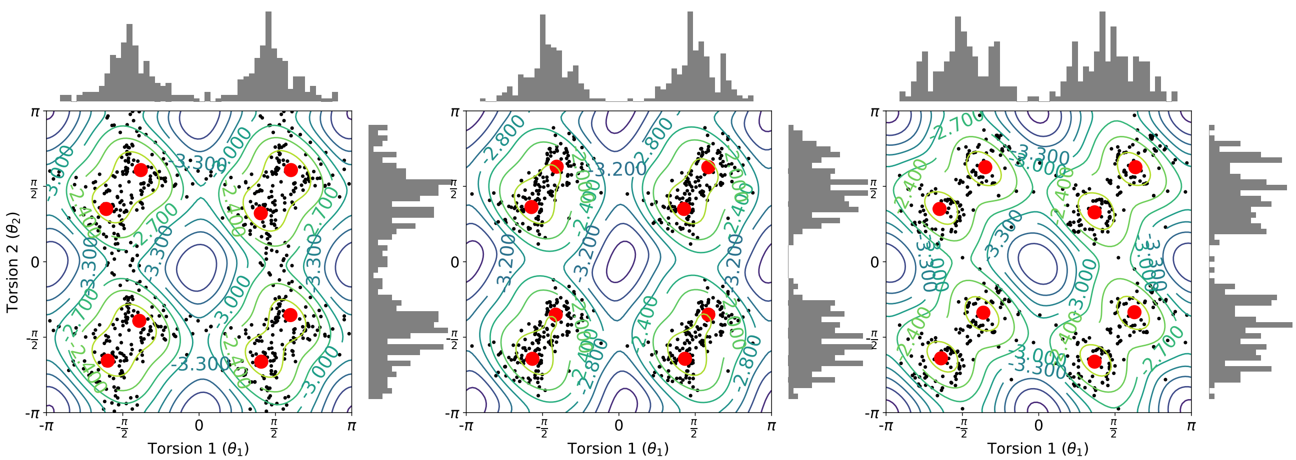
Figure S1: Mixture models for correlated torsion. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components.



Cluster 1: ω: 0.11 μ: -2.31 ν: -2.28 κ_1: 1.63 κ_2: 1.63 κ_3: -1.51
Cluster 2: ω: 0.12 μ: -2.4 ν: 0.93 κ_1: 1.62 κ_2: 1.63 κ_3: -1.52
Cluster 3: ω: 0.14 μ: -0.92 ν: -0.77 κ_1: 1.62 κ_2: 1.61 κ_3: -1.58
Cluster 4: ω: 0.13 μ: -0.93 ν: 2.35 κ_1: 1.63 κ_2: 1.62 κ_3: -1.53
Cluster 5: ω: 0.12 μ: 0.87 ν: -2.29 κ_1: 1.63 κ_2: 1.62 κ_3: -1.53
Cluster 6: ω: 0.125 μ: 0.8 ν: 0.9 κ_1: 1.62 κ_2: 1.63 κ_3: -1.54
Cluster 7: ω: 0.13 μ: 2.27 ν: -0.8 κ_1: 1.61 κ_2: 1.62 κ_3: -1.56
Cluster 8: ω: 0.125 μ: 2.25 ν: 2.33 κ_1: 1.63 κ_2: 1.63 κ_3: -1.51

Cluster 1: ω: 0.16 μ: -2.4 ν: -2.2 κ_1: 1.63 κ_2: 1.65 κ_3: -1.45
Cluster 2: ω: 0.16 μ: -2.41 ν: 0.94 κ_1: 1.65 κ_2: 1.66 κ_3: -1.46
Cluster 3: ω: 0.09 μ: -0.96 ν: -0.85 κ_1: 1.74 κ_2: 1.73 κ_3: -1.28
Cluster 4: ω: 0.09 μ: -0.95 ν: 2.3 κ_1: 1.74 κ_2: 1.73 κ_3: -1.28
Cluster 5: ω: 0.16 μ: 0.77 ν: -2.2 κ_1: 1.66 κ_2: 1.65 κ_3: -1.46
Cluster 6: ω: 0.16 μ: 0.78 ν: 0.95 κ_1: 1.65 κ_2: 1.67 κ_3: -1.47
Cluster 7: ω: 0.09 μ: 2.24 ν: -0.86 κ_1: 1.73 κ_2: 1.74 κ_3: -1.26
Cluster 8: ω: 0.09 μ: 2.26 ν: 2.29 κ_1: 1.73 κ_2: 1.74 κ_3: -1.27

Cluster 1: ω: 0.12 μ: -2.42 ν: -2.41 κ_1: 1.68 κ_2: 1.68 κ_3: -1.41
Cluster 2: ω: 0.12 μ: -2.49 ν: 0.91 κ_1: 1.68 κ_2: 1.67 κ_3: -1.42
Cluster 3: ω: 0.135 μ: -0.88 ν: -0.76 κ_1: 1.67 κ_2: 1.66 κ_3: -1.44
Cluster 4: ω: 0.125 μ: -0.87 ν: 2.43 κ_1: 1.69 κ_2: 1.67 κ_3: -1.43
Cluster 5: ω: 0.125 μ: 0.82 ν: -2.42 κ_1: 1.67 κ_2: 1.67 κ_3: -1.43
Cluster 6: ω: 0.13 μ: 0.75 ν: 0.893 κ_1: 1.66 κ_2: 1.67 κ_3: -1.43
Cluster 7: ω: 0.125 μ: 2.37 ν: -0.79 κ_1: 1.67 κ_2: 1.67 κ_3: -1.43
Cluster 8: ω: 0.12 μ: 2.39 ν: 2.41 κ_1: 1.69 κ_2: 1.68 κ_3: -1.41

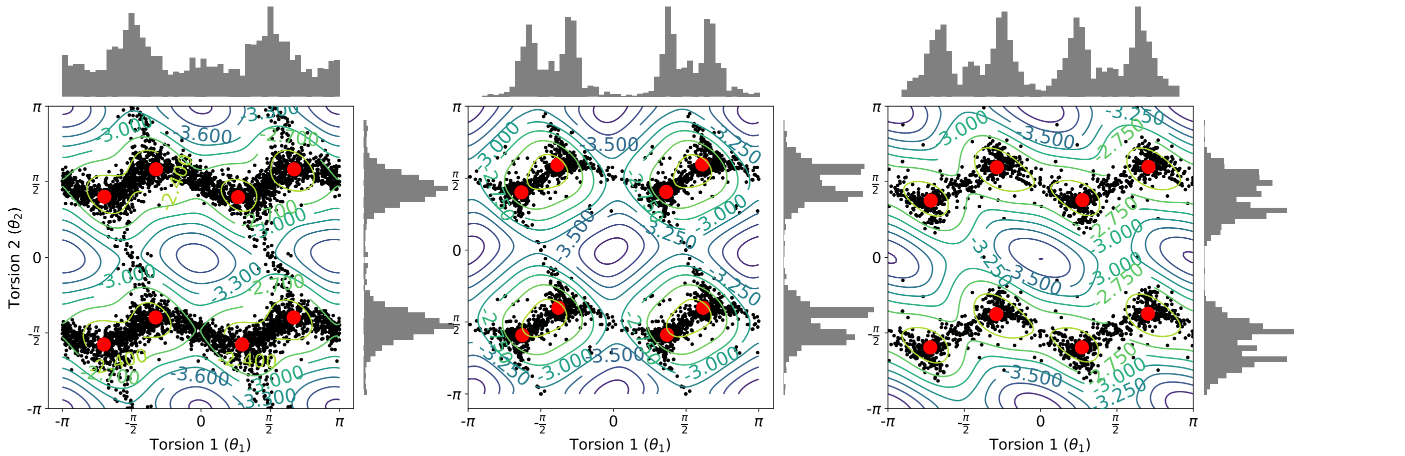


Cluster 1: ω: 0.125 μ: -1.88 ν: -2.06 κ_1: 1.66 κ_2: 1.64 κ_3: -1.43
Cluster 2: ω: 0.12 μ: -1.91 ν: 1.1 κ_1: 1.67 κ_2: 1.66 κ_3: -1.43
Cluster 3: ω: 0.13 μ: -1.2 ν: 1.91 κ_1: 1.67 κ_2: 1.66 κ_3: -1.44
Cluster 4: ω: 0.125 μ: -1.23 ν: -1.23 κ_1: 1.67 κ_2: 1.65 κ_3: -1.45
Cluster 5: ω: 0.125 μ: 1.28 ν: -2.08 κ_1: 1.68 κ_2: 1.65 κ_3: -1.44
Cluster 6: ω: 0.125 μ: 1.27 ν: 1.01 κ_1: 1.69 κ_2: 1.67 κ_3: -1.43
Cluster 7: ω: 0.12 μ: 1.88 ν: -1.11 κ_1: 1.69 κ_2: 1.68 κ_3: -1.44
Cluster 8: ω: 0.13 μ: 1.89 ν: 1.91 κ_1: 1.69 κ_2: 1.68 κ_3: -1.44

Cluster 1: ω: 0.14 μ: -1.78 ν: -2.02 κ_1: 1.86 κ_2: 1.85 κ_3: -1.71
Cluster 2: ω: 0.14 μ: -1.8 ν: 1.14 κ_1: 1.86 κ_2: 1.86 κ_3: -1.71
Cluster 3: ω: 0.11 μ: -1.28 ν: 1.98 κ_1: 1.89 κ_2: 1.87 κ_3: -1.68
Cluster 4: ω: 0.11 μ: -1.3 ν: -1.1 κ_1: 1.9 κ_2: 1.86 κ_3: -1.68
Cluster 5: ω: 0.14 μ: 1.36 ν: -2.02 κ_1: 1.87 κ_2: 1.86 κ_3: -1.72
Cluster 6: ω: 0.13 μ: 1.34 ν: 1.1 κ_1: 1.89 κ_2: 1.89 κ_3: -1.71
Cluster 7: ω: 0.11 μ: 1.84 ν: -1.1 κ_1: 1.9 κ_2: 1.88 κ_3: -1.68
Cluster 8: ω: 0.12 μ: 1.84 ν: 1.97 κ_1: 1.88 κ_2: 1.87 κ_3: -1.69

Cluster 1: ω: 0.125 μ: -2.01 ν: -2.01 κ_1: 1.67 κ_2: 1.67 κ_3: -1.42
Cluster 2: ω: 0.125 μ: -2.04 ν: 1.1 κ_1: 1.67 κ_2: 1.68 κ_3: -1.42
Cluster 3: ω: 0.125 μ: -1.1 ν: 1.97 κ_1: 1.69 κ_2: 1.68 κ_3: -1.43
Cluster 4: ω: 0.125 μ: -1.14 ν: -1.06 κ_1: 1.69 κ_2: 1.67 κ_3: -1.43
Cluster 5: ω: 0.125 μ: 1.15 ν: -2.08 κ_1: 1.68 κ_2: 1.67 κ_3: -1.42
Cluster 6: ω: 0.125 μ: 1.15 ν: 1.03 κ_1: 1.69 κ_2: 1.69 κ_3: -1.42
Cluster 7: ω: 0.125 μ: 1.97 ν: -1.05 κ_1: 1.69 κ_2: 1.68 κ_3: -1.41
Cluster 8: ω: 0.125 μ: 1.99 ν: 1.97 κ_1: 1.68 κ_2: 1.68 κ_3: -1.41

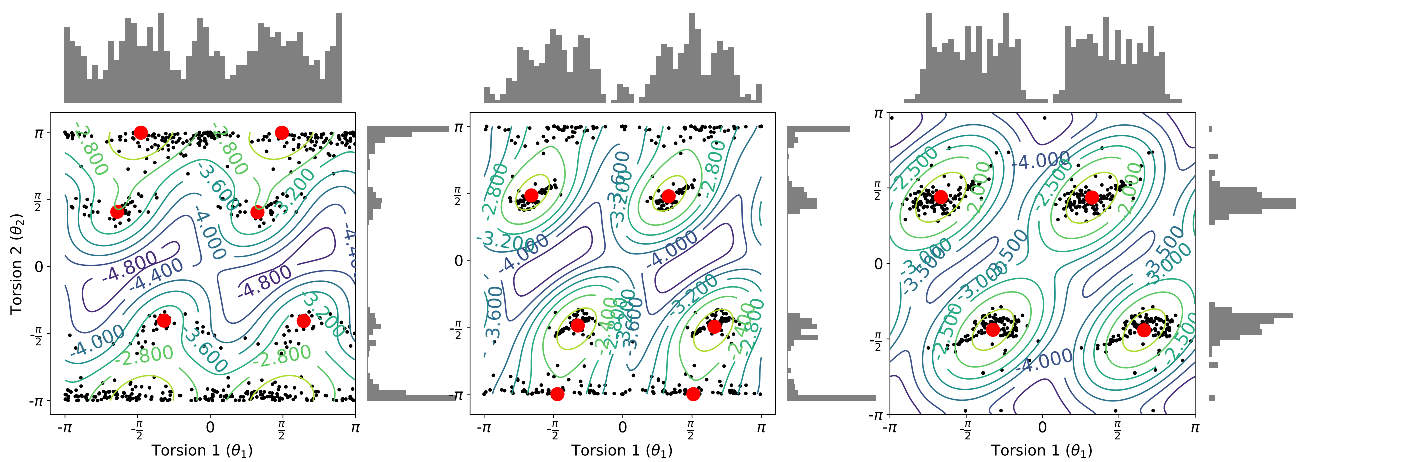
Figure S1: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. (Continued)



Cluster 1: ω: 0.13 μ: -2.2 ν: -1.81 κ_1: 1.66 κ_2: 1.64 κ_3: -1.11
Cluster 2: ω: 0.12 μ: -2.19 ν: 1.26 κ_1: 1.67 κ_2: 1.66 κ_3: -1.11
Cluster 3: ω: 0.12 μ: -1.03 ν: -1.25 κ_1: 1.67 κ_2: 1.66 κ_3: -1.07
Cluster 4: ω: 0.13 μ: -1.02 ν: 1.83 κ_1: 1.67 κ_2: 1.65 κ_3: -1.08
Cluster 5: ω: 0.125 μ: 0.93 ν: -1.81 κ_1: 1.68 κ_2: 1.65 κ_3: -1.08
Cluster 6: ω: 0.12 μ: 0.84 ν: 1.25 κ_1: 1.69 κ_2: 1.67 κ_3: -1.05
Cluster 7: ω: 0.125 μ: 2.1 ν: -1.25 κ_1: 1.69 κ_2: 1.68 κ_3: -1.06
Cluster 8: ω: 0.13 μ: 2.11 ν: 1.83 κ_1: 1.69 κ_2: 1.68 κ_3: -1.07

Cluster 1: ω: 0.12 μ: -1.97 ν: -1.86 κ_1: 1.44 κ_2: 1.56 κ_3: -1.19
Cluster 2: ω: 0.11 μ: -1.99 ν: 1.26 κ_1: 1.47 κ_2: 1.59 κ_3: -1.16
Cluster 3: ω: 0.13 μ: -1.19 ν: -1.26 κ_1: 1.46 κ_2: 1.56 κ_3: -1.21
Cluster 4: ω: 0.14 μ: -1.21 ν: 1.86 κ_1: 1.43 κ_2: 1.52 κ_3: -1.22
Cluster 5: ω: 0.12 μ: 1.16 ν: -1.85 κ_1: 1.45 κ_2: 1.56 κ_3: -1.18
Cluster 6: ω: 0.11 μ: 1.14 ν: 1.27 κ_1: 1.48 κ_2: 1.59 κ_3: -1.14
Cluster 7: ω: 0.13 μ: 1.94 ν: -1.26 κ_1: 1.46 κ_2: 1.55 κ_3: -1.21
Cluster 8: ω: 0.14 μ: 1.92 ν: 1.87 κ_1: 1.42 κ_2: 1.51 κ_3: -1.22

Cluster 1: ω: 0.12 μ: -2.27 ν: -1.87 κ_1: 1.42 κ_2: 1.57 κ_3: -1.2
Cluster 2: ω: 0.12 μ: -2.26 ν: 1.18 κ_1: 1.42 κ_2: 1.56 κ_3: -1.19
Cluster 3: ω: 0.13 μ: -0.91 ν: -1.18 κ_1: 1.42 κ_2: 1.54 κ_3: -1.21
Cluster 4: ω: 0.13 μ: -0.9 ν: 1.87 κ_1: 1.41 κ_2: 1.54 κ_3: -1.22
Cluster 5: ω: 0.12 μ: 0.85 ν: -1.87 κ_1: 1.41 κ_2: 1.56 κ_3: -1.19
Cluster 6: ω: 0.12 μ: 0.86 ν: 1.19 κ_1: 1.41 κ_2: 1.56 κ_3: -1.19
Cluster 7: ω: 0.13 μ: 2.21 ν: -1.17 κ_1: 1.41 κ_2: 1.54 κ_3: -1.21
Cluster 8: ω: 0.13 μ: 2.22 ν: 1.88 κ_1: 1.41 κ_2: 1.54 κ_3: -1.21

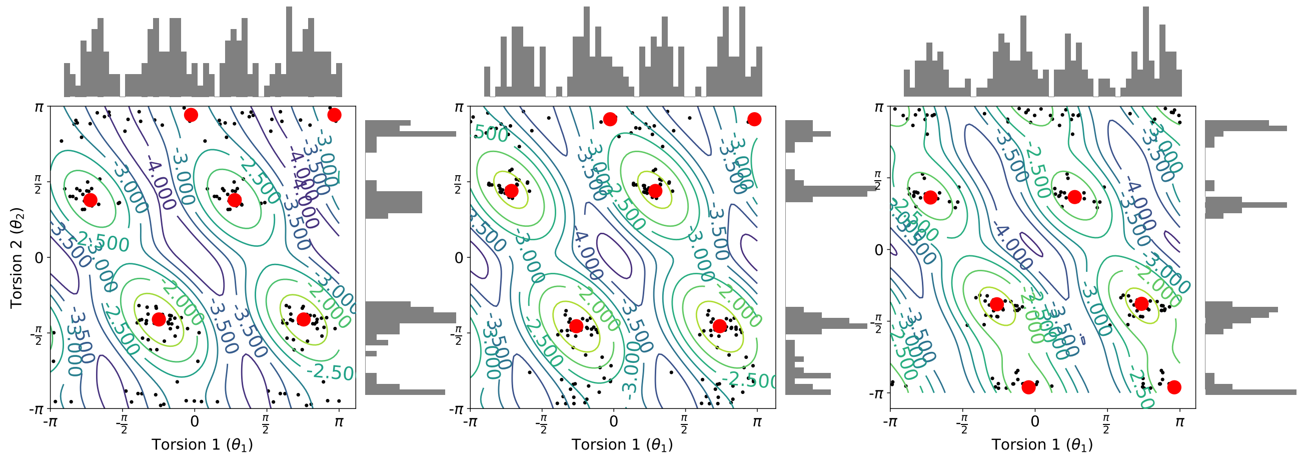


Cluster 1: ω: 0.06 μ: -2.01 ν: 1.28 κ_1: 1.78 κ_2: 1.77 κ_3: -1.08
Cluster 2: ω: 0.05 μ: -1 ν: -1.27 κ_1: 1.78 κ_2: 1.77 κ_3: -1.06
Cluster 3: ω: 0.06 μ: 1.02 ν: 1.27 κ_1: 1.78 κ_2: 1.77 κ_3: -1.08
Cluster 4: ω: 0.05 μ: 2.02 ν: -1.28 κ_1: 1.78 κ_2: 1.77 κ_3: -1.06
Cluster 5: ω: 0.39 μ: -1.5 ν: 3.13 κ_1: 1.02 κ_2: 1.71 κ_3: -0.49
Cluster 6: ω: 0.39 μ: 1.55 ν: 3.13 κ_1: 1.02 κ_2: 1.71 κ_3: -0.49

Cluster 1: ω: 0.145 μ: -2.07 ν: 1.52 κ_1: 1.74 κ_2: 1.7 κ_3: -1.37
Cluster 2: ω: 0.155 μ: -1.02 ν: -1.53 κ_1: 1.74 κ_2: 1.7 κ_3: -1.37
Cluster 3: ω: 0.145 μ: 1.04 ν: 1.5 κ_1: 1.74 κ_2: 1.71 κ_3: -1.37
Cluster 4: ω: 0.155 μ: 2.08 ν: -1.56 κ_1: 1.73 κ_2: 1.7 κ_3: -1.37
Cluster 5: ω: 0.2 μ: -1.48 ν: -3.14 κ_1: 1.0 κ_2: 1.45 κ_3: -0.74
Cluster 6: ω: 0.2 μ: 1.6 ν: -3.14 κ_1: 0.99 κ_2: 1.45 κ_3: -0.74

Cluster 1: ω: 0.265 μ: -2.09 ν: 1.38 κ_1: 1.67 κ_2: 1.71 κ_3: -1.58
Cluster 2: ω: 0.235 μ: -1.02 ν: -1.38 κ_1: 1.7 κ_2: 1.73 κ_3: -1.56
Cluster 3: ω: 0.265 μ: 1.02 ν: 1.37 κ_1: 1.68 κ_2: 1.71 κ_3: -1.59
Cluster 4: ω: 0.235 μ: 2.09 ν: -1.39 κ_1: 1.7 κ_2: 1.73 κ_3: -1.58

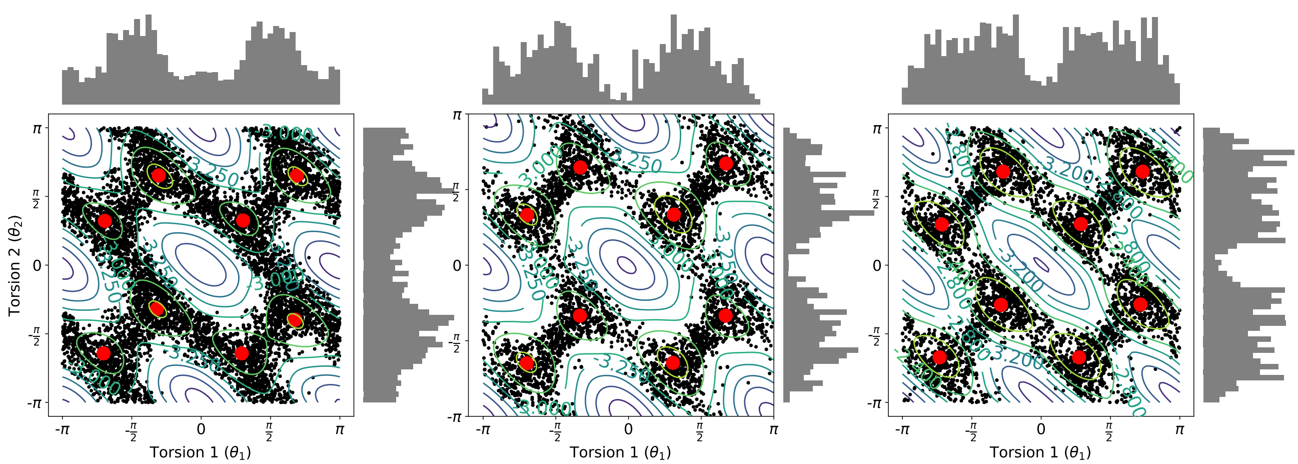
Figure S1: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components.(Continued)



Cluster 1: ω: 0.135 μ: 3.04 ν: 2.96 κ_1: 1.05 κ_2: 1.3 κ_3: -0.92
Cluster 2: ω: 0.135 μ: -0.08 ν: 2.96 κ_1: 1.04 κ_2: 1.28 κ_3: -0.92
Cluster 3: ω: 0.145 μ: -2.27 ν: 1.19 κ_1: 1.92 κ_2: 1.89 κ_3: -1.43
Cluster 4: ω: 0.22 μ: -0.78 ν: -1.29 κ_1: 1.87 κ_2: 1.84 κ_3: -1.61
Cluster 5: ω: 0.15 μ: 0.87 ν: 1.19 κ_1: 1.92 κ_2: 1.89 κ_3: -1.43
Cluster 6: ω: 0.215 μ: 2.37 ν: -1.29 κ_1: 1.86 κ_2: 1.84 κ_3: -1.6

Cluster 1: ω: 0.1 μ: 3.08 ν: 2.87 κ_1: 1.11 κ_2: 1.27 κ_3: -1.04
Cluster 2: ω: 0.1 μ: -0.08 ν: 2.87 κ_1: 1.11 κ_2: 1.27 κ_3: -1.04
Cluster 3: ω: 0.18 μ: -2.24 ν: 1.38 κ_1: 1.91 κ_2: 1.84 κ_3: -1.56
Cluster 4: ω: 0.22 μ: -0.82 ν: -1.43 κ_1: 1.87 κ_2: 1.79 κ_3: -1.61
Cluster 5: ω: 0.18 μ: 0.91 ν: 1.38 κ_1: 1.91 κ_2: 1.84 κ_3: -1.56
Cluster 6: ω: 0.22 μ: 2.32 ν: -1.43 κ_1: 1.87 κ_2: 1.79 κ_3: -1.61

Cluster 1: ω: 0.21 μ: 3.02 ν: -3.02 κ_1: 1.48 κ_2: 1.55 κ_3: -1.43
Cluster 2: ω: 0.21 μ: -0.14 ν: -3.02 κ_1: 1.47 κ_2: 1.54 κ_3: -1.43
Cluster 3: ω: 0.125 μ: -2.27 ν: 1.14 κ_1: 1.91 κ_2: 1.9 κ_3: -1.5
Cluster 4: ω: 0.165 μ: -0.83 ν: -1.2 κ_1: 1.9 κ_2: 1.92 κ_3: -1.64
Cluster 5: ω: 0.125 μ: 0.86 ν: 1.15 κ_1: 1.92 κ_2: 1.92 κ_3: -1.5
Cluster 6: ω: 0.165 μ: 2.31 ν: -1.19 κ_1: 1.89 κ_2: 1.89 κ_3: -1.63

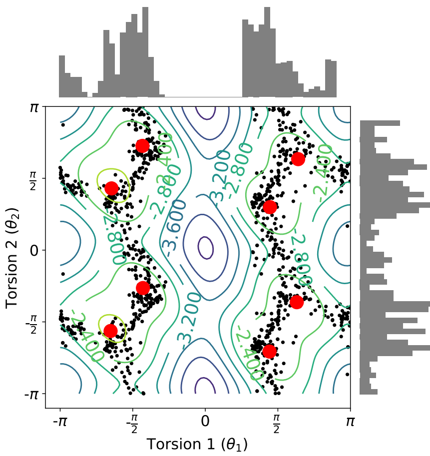


Cluster 1: ω: 0.12 μ: -2.21 ν: -2.02 κ_1: 1.39 κ_2: 1.39 κ_3: -1.08
Cluster 2: ω: 0.11 μ: -2.18 ν: 1.01 κ_1: 1.4 κ_2: 1.41 κ_3: -1.07
Cluster 3: ω: 0.13 μ: -1.0 ν: -1 κ_1: 1.41 κ_2: 1.4 κ_3: -1.12
Cluster 4: ω: 0.14 μ: -0.96 ν: 2.05 κ_1: 1.4 κ_2: 1.38 κ_3: -1.12
Cluster 5: ω: 0.12 μ: 0.92 ν: -2.02 κ_1: 1.39 κ_2: 1.4 κ_3: -1.07
Cluster 6: ω: 0.11 μ: 0.95 ν: 1.02 κ_1: 1.41 κ_2: 1.41 κ_3: -1.07
Cluster 7: ω: 0.13 μ: 2.13 ν: -1.25 κ_1: 1.41 κ_2: 1.39 κ_3: -1.11
Cluster 8: ω: 0.14 μ: 2.17 ν: 2.05 κ_1: 1.38 κ_2: 1.37 κ_3: -1.11

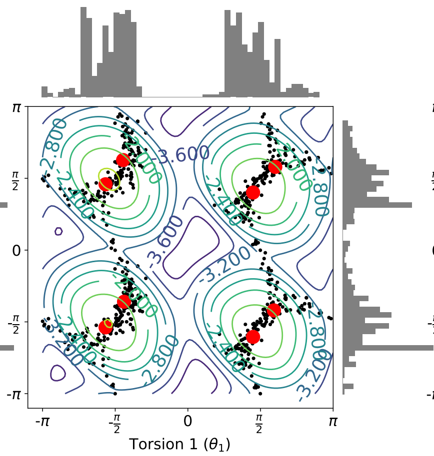
Cluster 1: ω: 0.135 μ: -2.2 ν: -2.04 κ_1: 1.39 κ_2: 1.41 κ_3: -1.11
Cluster 2: ω: 0.13 μ: -2.19 ν: 1.04 κ_1: 1.42 κ_2: 1.43 κ_3: -1.11
Cluster 3: ω: 0.115 μ: -1.05 ν: -1.05 κ_1: 1.44 κ_2: 1.44 κ_3: -1.07
Cluster 4: ω: 0.12 μ: -1.04 ν: 2.03 κ_1: 1.43 κ_2: 1.42 κ_3: -1.08
Cluster 5: ω: 0.14 μ: 0.96 ν: -2.03 κ_1: 1.4 κ_2: 1.41 κ_3: -1.12
Cluster 6: ω: 0.135 μ: 0.98 ν: 1.05 κ_1: 1.42 κ_2: 1.53 κ_3: -1.12
Cluster 7: ω: 0.11 μ: 2.1 ν: -1.05 κ_1: 1.44 κ_2: 1.43 κ_3: -1.06
Cluster 8: ω: 0.115 μ: 2.11 ν: 2.11 κ_1: 1.42 κ_2: 1.42 κ_3: -1.06

Cluster 1: ω: 0.12 μ: -2.3 ν: -2.11 κ_1: 1.84 κ_2: 1.85 κ_3: -1.54
Cluster 2: ω: 0.12 μ: -2.25 ν: 0.93 κ_1: 1.86 κ_2: 1.86 κ_3: -1.56
Cluster 3: ω: 0.13 μ: -0.91 ν: -0.91 κ_1: 1.86 κ_2: 1.86 κ_3: -1.6
Cluster 4: ω: 0.13 μ: -0.86 ν: 2.14 κ_1: 1.85 κ_2: 1.84 κ_3: -1.58
Cluster 5: ω: 0.125 μ: 0.86 ν: -2.11 κ_1: 1.85 κ_2: 1.85 κ_3: -1.56
Cluster 6: ω: 0.12 μ: 0.9 ν: 0.94 κ_1: 1.86 κ_2: 1.86 κ_3: -1.58
Cluster 7: ω: 0.125 μ: 2.25 ν: -0.9 κ_1: 1.85 κ_2: 1.85 κ_3: -1.58
Cluster 8: ω: 0.13 μ: 2.3 ν: 2.14 κ_1: 1.85 κ_2: 1.84 κ_3: -1.57

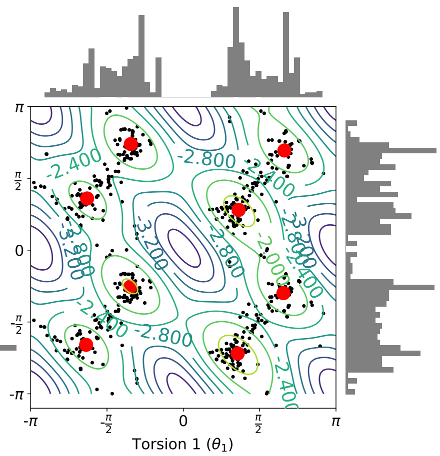
Figure S1: Mixture models for correlated torsion. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. (Continued)



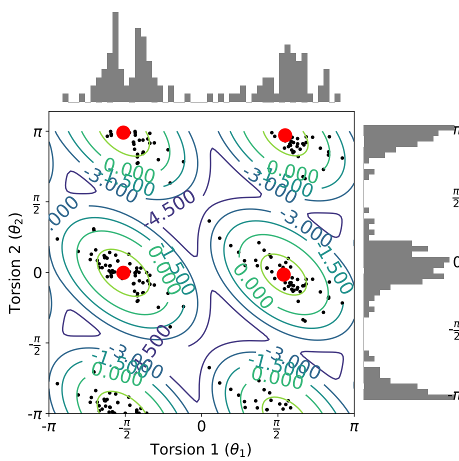
Cluster 1: ω: 0.15 μ: -2.05 ν: -1.77 κ_1: 1.59 κ_2: 1.61 κ_3: -1.38
Cluster 2: ω: 0.15 μ: -2.03 ν: 1.35 κ_1: 1.62 κ_2: 1.63 κ_3: -1.38
Cluster 3: ω: 0.11 μ: -1.35 ν: -0.83 κ_1: 1.71 κ_2: 1.68 κ_3: -1.37
Cluster 4: ω: 0.11 μ: -1.36 ν: 2.28 κ_1: 1.71 κ_2: 1.68 κ_3: -1.38
Cluster 5: ω: 0.12 μ: 1.39 ν: -2.22 κ_1: 1.7 κ_2: 1.66 κ_3: -1.39
Cluster 6: ω: 0.12 μ: 1.4 ν: 0.94 κ_1: 1.7 κ_2: 1.65 κ_3: -1.35
Cluster 7: ω: 0.12 μ: 1.98 ν: -1.14 κ_1: 1.66 κ_2: 1.65 κ_3: -1.39
Cluster 8: ω: 0.12 μ: 2.01 ν: 1.99 κ_1: 1.65 κ_2: 1.64 κ_3: -1.35



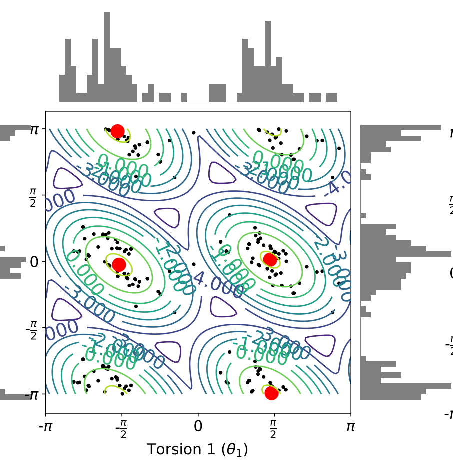
Cluster 1: ω: 0.17 μ: -1.78 ν: -1.69 κ_1: 1.81 κ_2: 1.78 κ_3: -1.77
Cluster 2: ω: 0.17 μ: -1.78 ν: 1.44 κ_1: 1.83 κ_2: 1.78 κ_3: -1.77
Cluster 3: ω: 0.08 μ: -1.38 ν: -1.13 κ_1: 1.91 κ_2: 1.78 κ_3: -1.63
Cluster 4: ω: 0.09 μ: -1.4 ν: 1.97 κ_1: 1.9 κ_2: 1.78 κ_3: -1.67
Cluster 5: ω: 0.15 μ: 1.41 ν: -1.9 κ_1: 1.87 κ_2: 1.8 κ_3: -1.76
Cluster 6: ω: 0.15 μ: 1.41 ν: 1.26 κ_1: 1.88 κ_2: 1.8 κ_3: -1.76
Cluster 7: ω: 0.09 μ: 1.86 ν: -1.32 κ_1: 1.87 κ_2: 1.79 κ_3: -1.63
Cluster 8: ω: 0.1 μ: 1.89 ν: 1.82 κ_1: 1.85 κ_2: 1.79 κ_3: -1.65



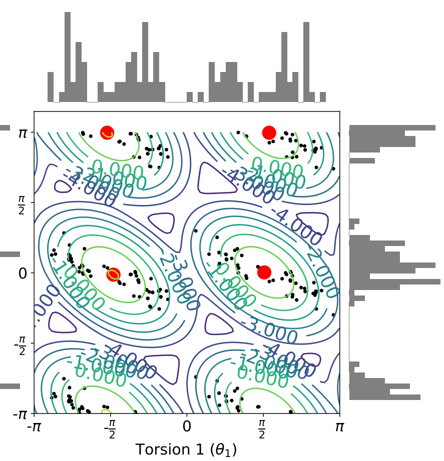
Cluster 1: ω: 0.11 μ: -2.0 ν: -2.07 κ_1: 1.85 κ_2: 1.85 κ_3: -1.65
Cluster 2: ω: 0.11 μ: -1.98 ν: 1.13 κ_1: 1.83 κ_2: 1.78 κ_3: -1.63
Cluster 3: ω: 0.13 μ: -1.09 ν: -0.8 κ_1: 1.89 κ_2: 1.81 κ_3: -1.7
Cluster 4: ω: 0.12 μ: -1.08 ν: 2.32 κ_1: 1.9 κ_2: 1.82 κ_3: -1.69
Cluster 5: ω: 0.15 μ: 1.12 ν: -2.26 κ_1: 1.89 κ_2: 1.82 κ_3: -1.74
Cluster 6: ω: 0.14 μ: 1.14 ν: 0.88 κ_1: 1.88 κ_2: 1.82 κ_3: -1.73
Cluster 7: ω: 0.12 μ: 2.06 ν: -0.94 κ_1: 1.85 κ_2: 1.79 κ_3: -1.66
Cluster 8: ω: 0.12 μ: 2.08 ν: 2.18 κ_1: 1.85 κ_2: 1.79 κ_3: -1.66



Cluster 1: ω: 0.27 μ: -1.61 ν: -0.01 κ_1: 3.45 κ_2: 3.8 κ_3: -3.2
Cluster 2: ω: 0.27 μ: -1.61 ν: 3.11 κ_1: 3.45 κ_2: 3.79 κ_3: -3.21
Cluster 3: ω: 0.23 μ: 1.69 ν: -0.05 κ_1: 3.44 κ_2: 3.81 κ_3: -3.18
Cluster 4: ω: 0.23 μ: 1.72 ν: 3.05 κ_1: 3.44 κ_2: 3.82 κ_3: -3.18



Cluster 1: ω: 0.26 μ: -1.63 ν: -0.08 κ_1: 3.46 κ_2: 3.86 κ_3: -3.14
Cluster 2: ω: 0.26 μ: -1.66 ν: 3.08 κ_1: 3.46 κ_2: 3.86 κ_3: -3.13
Cluster 3: ω: 0.24 μ: 1.48 ν: 0.04 κ_1: 3.47 κ_2: 3.88 κ_3: -3.12
Cluster 4: ω: 0.24 μ: 1.51 ν: -3.12 κ_1: 3.46 κ_2: 3.88 κ_3: -3.12



Cluster 1: ω: 0.27 μ: -1.51 ν: -0.04 κ_1: 3.25 κ_2: 3.88 κ_3: -3.04
Cluster 2: ω: 0.27 μ: -1.64 ν: 3.13 κ_1: 3.25 κ_2: 3.88 κ_3: -3.04
Cluster 3: ω: 0.23 μ: 1.59 ν: 0.01 κ_1: 3.26 κ_2: 3.91 κ_3: -3.04
Cluster 4: ω: 0.23 μ: 1.69 ν: 3.13 κ_1: 3.26 κ_2: 3.9 κ_3: -3.04

Figure S1: Mixture models for correlated torsion. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. (Continued)

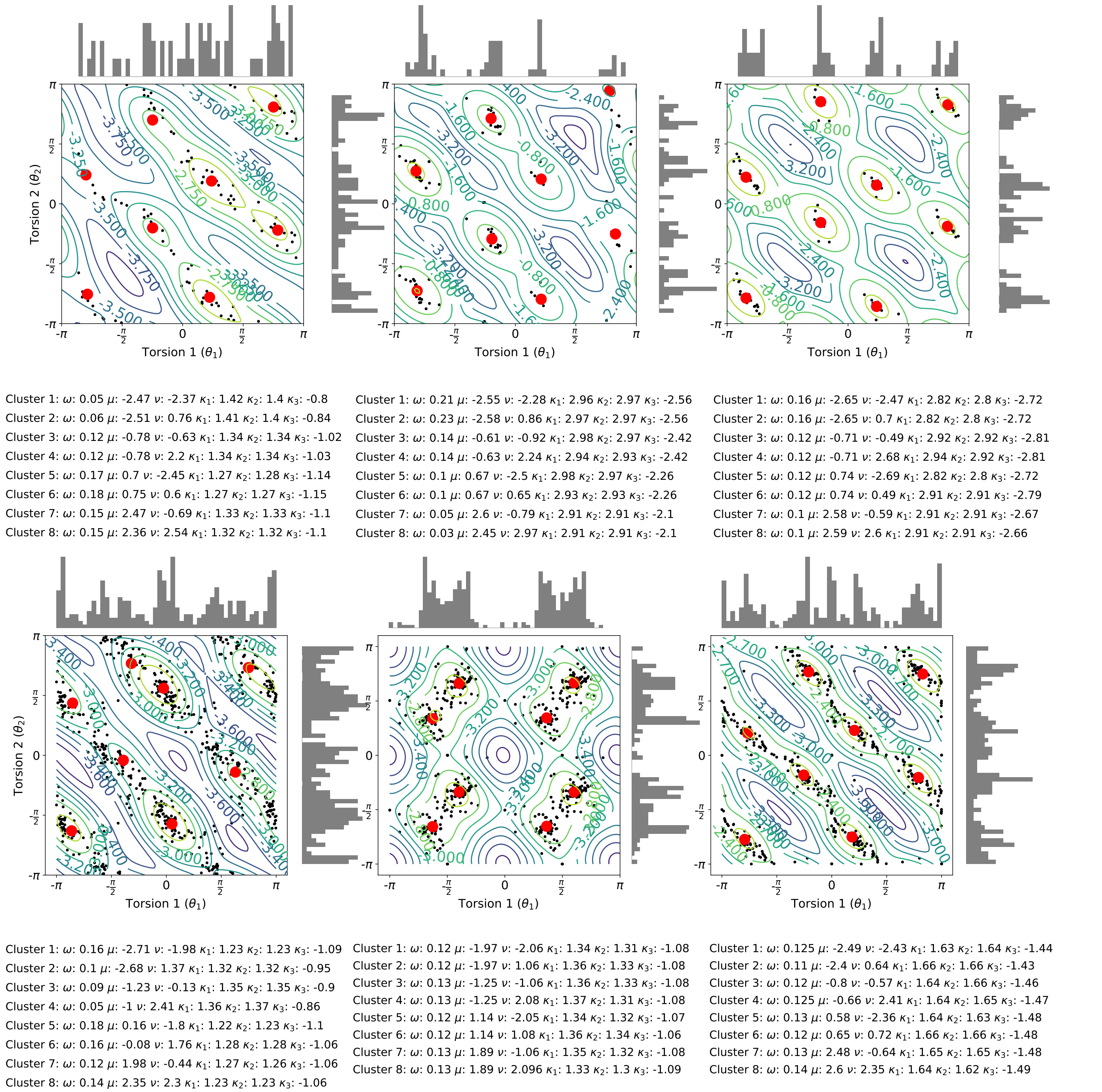


Figure S1: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components.(Continued)

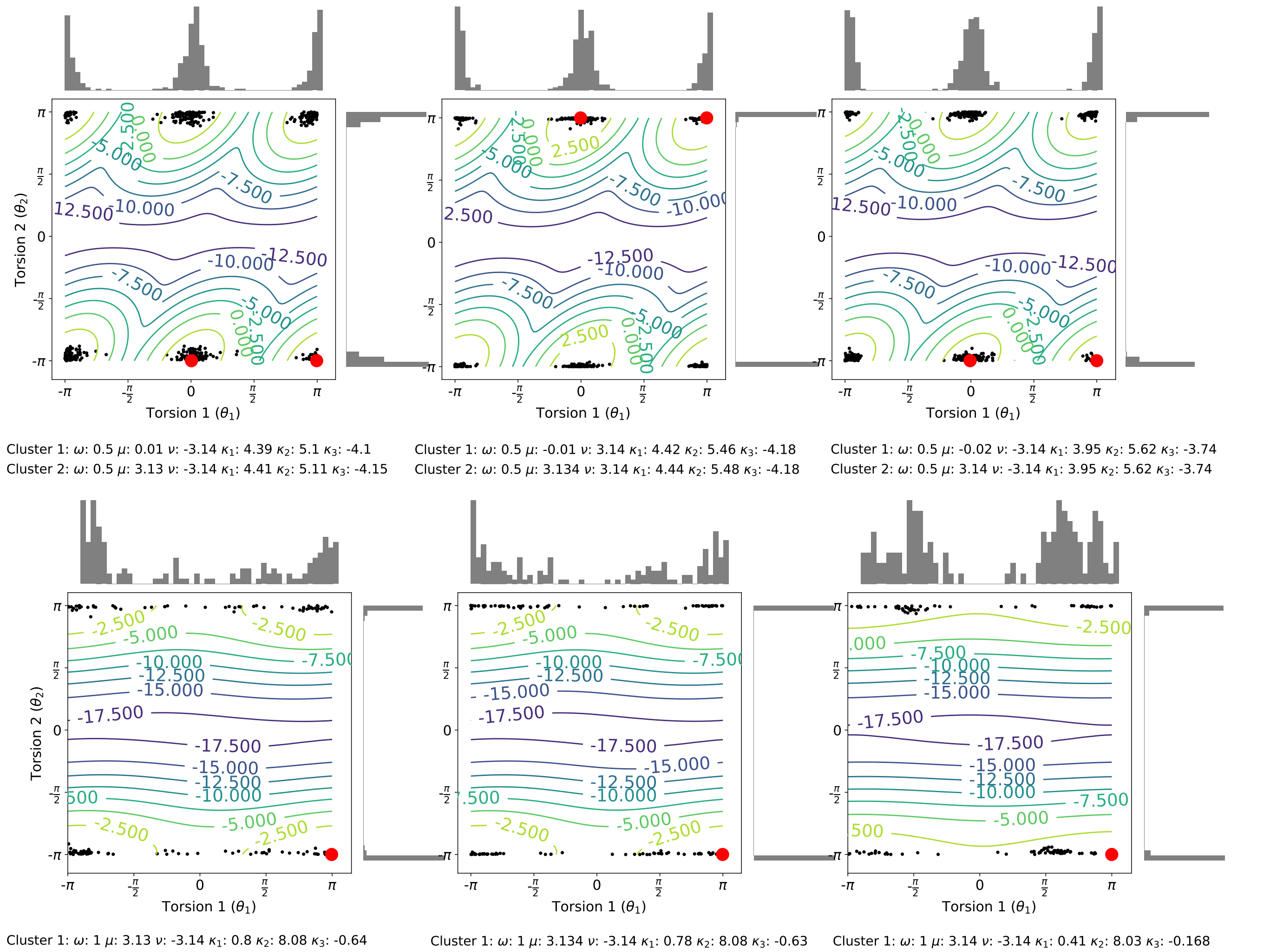


Figure S1: Mixture models for correlated torsion. The contour plot indicates the log density of a mixture model and the points (in red) mark the mean location for the components.(Continued)

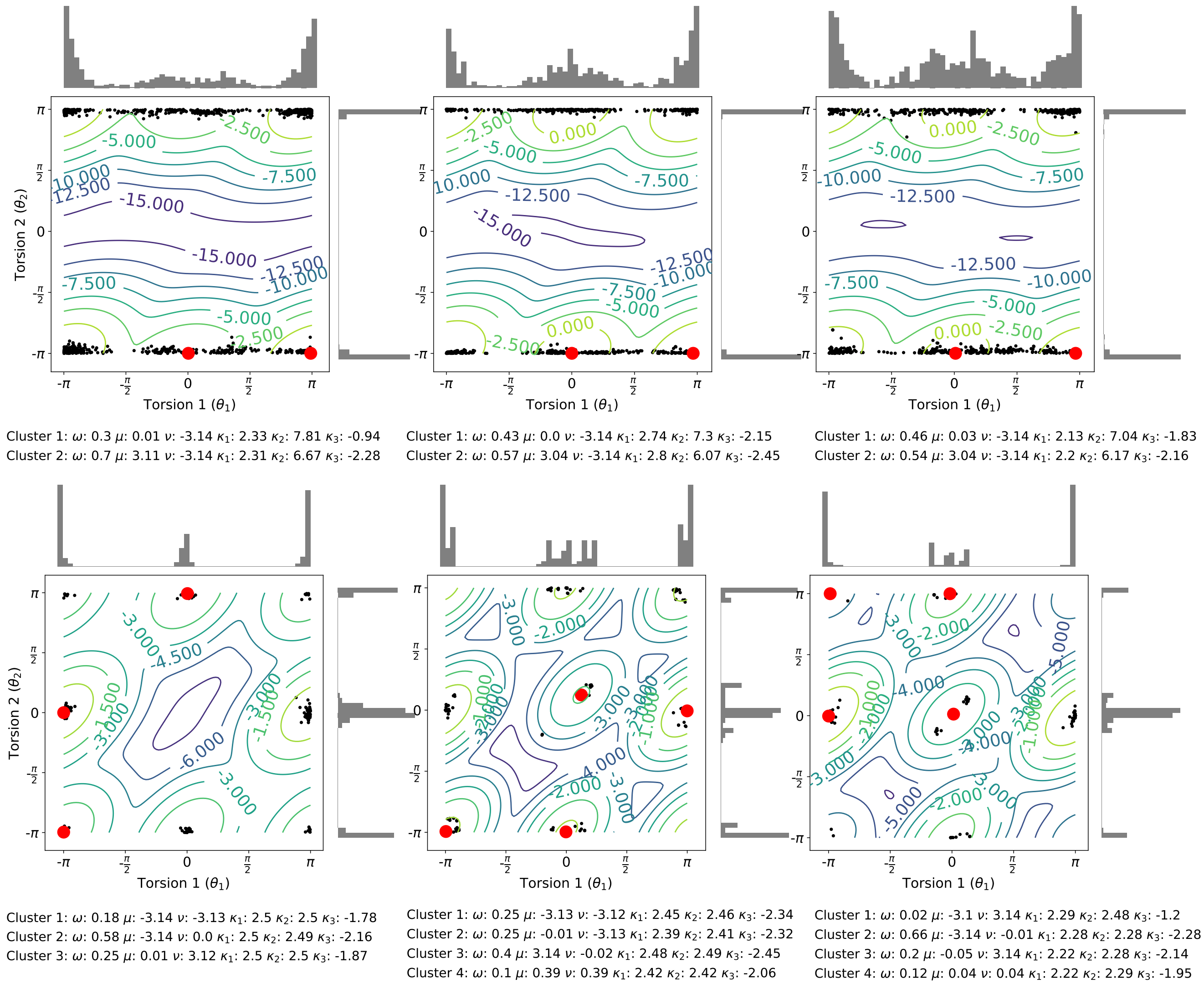


Figure S1: Mixture models for correlated torsion. The contour plot indicates the log density of a mixture model and the points (in red) mark the mean location for the components.(Continued)

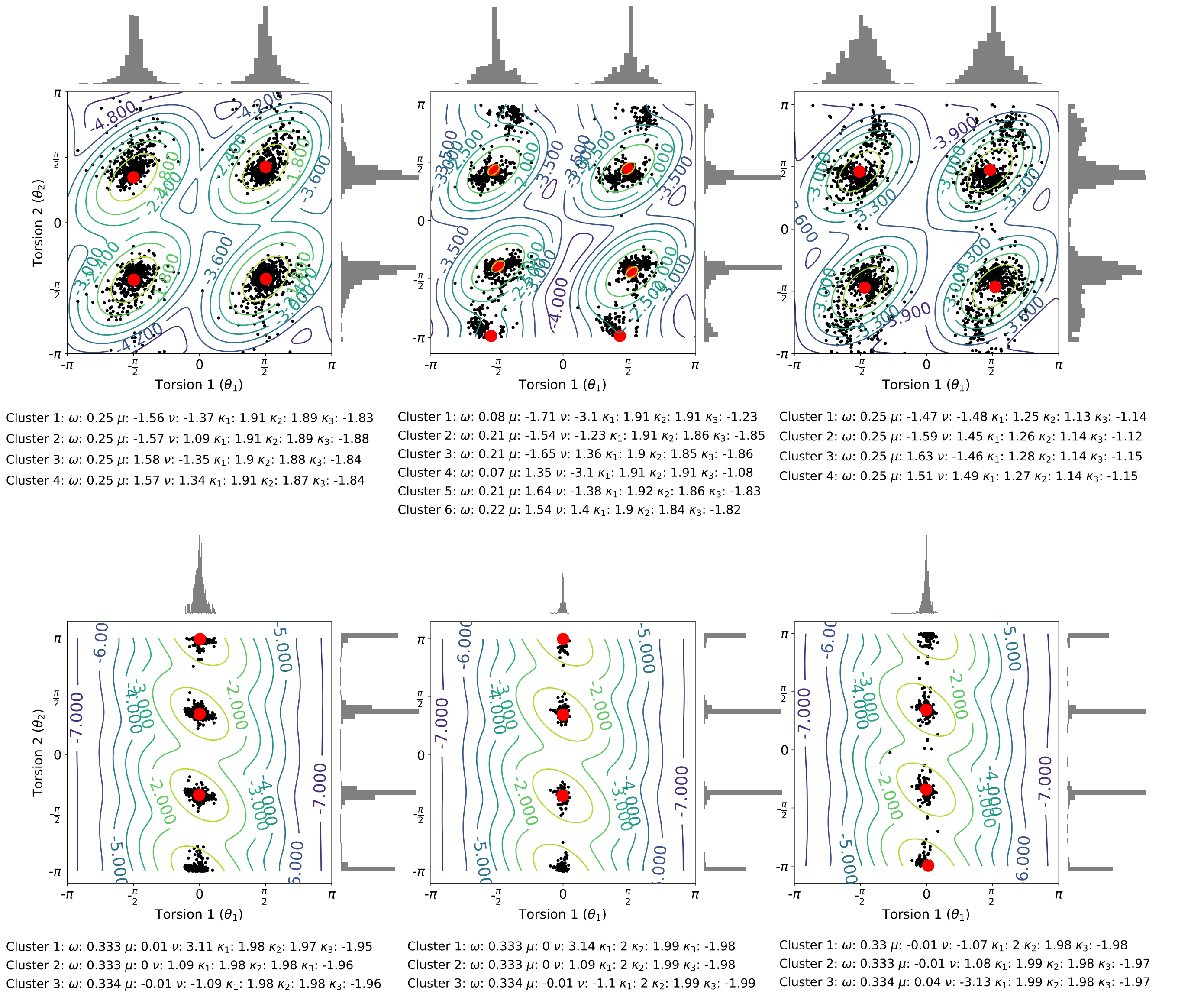
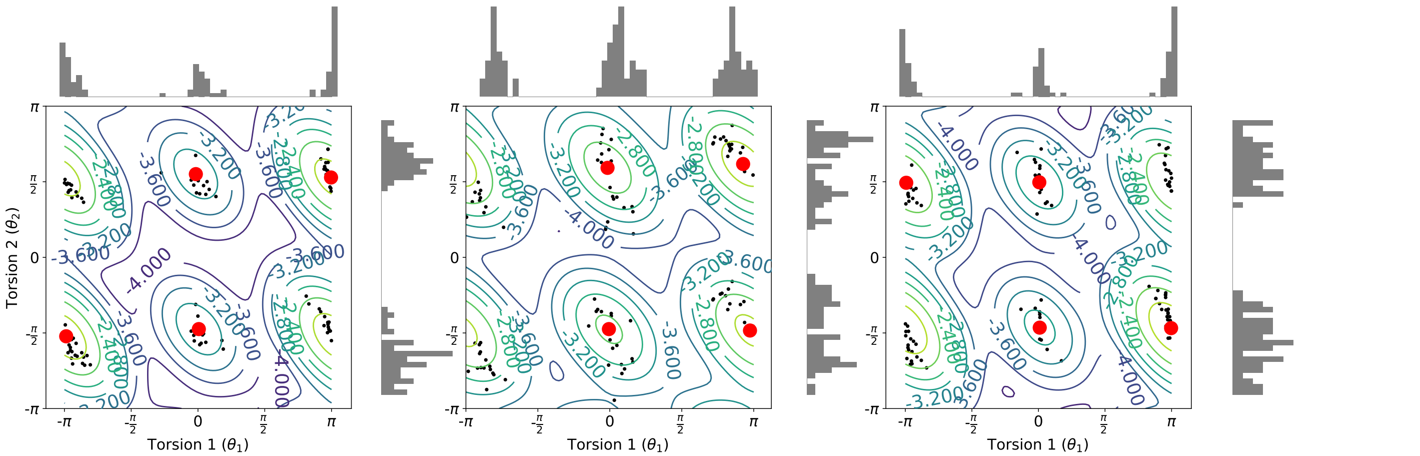


Figure S1: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. (Continued)

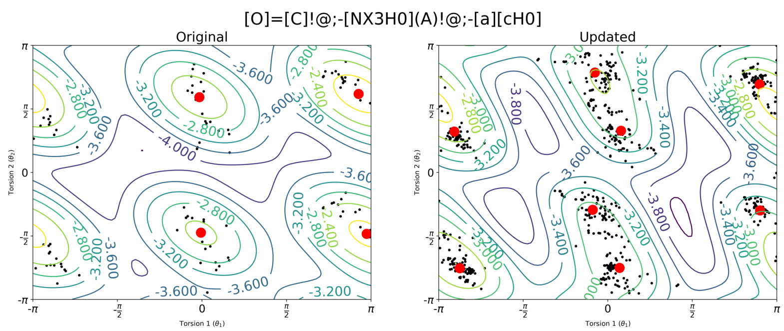


Cluster 1: ω: 0.38 μ: -3.1 ν: -1.64 κ_1: 1.4 κ_2: 1.38 κ_3: -1.33
Cluster 2: ω: 0.34 μ: 3.13 ν: 1.66 κ_1: 1.44 κ_2: 1.42 κ_3: -1.31
Cluster 3: ω: 0.14 μ: 0.02 ν: -1.49 κ_1: 1.47 κ_2: 1.46 κ_3: -0.99
Cluster 4: ω: 0.14 μ: -0.05 ν: 1.73 κ_1: 1.47 κ_2: 1.45 κ_3: -0.97

Cluster 1: ω: 0.29 μ: 3.06 ν: -1.52 κ_1: 1.31 κ_2: 1.25 κ_3: -1.17
Cluster 2: ω: 0.31 μ: 2.91 ν: 1.94 κ_1: 1.32 κ_2: 1.27 κ_3: -1.19
Cluster 3: ω: 0.18 μ: -0.02 ν: -1.49 κ_1: 1.44 κ_2: 1.36 κ_3: -1.03
Cluster 4: ω: 0.22 μ: -0.05 ν: 1.86 κ_1: 1.44 κ_2: 1.3 κ_3: -1.05

Cluster 1: ω: 0.41 μ: 3.13 ν: -1.47 κ_1: 1.45 κ_2: 1.39 κ_3: -1.25
Cluster 2: ω: 0.31 μ: -3.12 ν: 1.55 κ_1: 1.47 κ_2: 1.44 κ_3: -1.16
Cluster 3: ω: 0.13 μ: 0.03 ν: -1.46 κ_1: 1.48 κ_2: 1.47 κ_3: -0.89
Cluster 4: ω: 0.15 μ: 0.02 ν: 1.56 κ_1: 1.48 κ_2: 1.46 κ_3: -0.92

Figure S1: Mixture models for correlated torsions. The contour plot indicates the log density of the mixture model and the points (in red) mark the mean location for the components. (Continued)



Cluster 1: ω: 0.29 μ: 3.06 ν: -1.52 κ_1: 1.31 κ_2: 1.25 κ_3: -1.17
Cluster 2: ω: 0.31 μ: 2.91 ν: 1.94 κ_1: 1.32 κ_2: 1.27 κ_3: -1.19
Cluster 3: ω: 0.18 μ: -0.02 ν: -1.49 κ_1: 1.44 κ_2: 1.36 κ_3: -1.03
Cluster 4: ω: 0.22 μ: -0.05 ν: 1.86 κ_1: 1.44 κ_2: 1.3 κ_3: -1.05

Cluster 1: ω: 0.13 μ: -0.24 ν: 2.47 κ_1: 1.24 κ_2: 1.19 κ_3: -0.8
Cluster 2: ω: 0.11 μ: 0.245 ν: 1.03 κ_1: 1.26 κ_2: 1.22 κ_3: -0.79
Cluster 3: ω: 0.11 μ: -0.281 ν: -0.925 κ_1: 1.26 κ_2: 1.22 κ_3: -0.8
Cluster 4: ω: 0.12 μ: 0.22 ν: -2.36 κ_1: 1.25 κ_2: 1.2 κ_3: -0.81
Cluster 5: ω: 0.17 μ: -2.75 ν: -2.37 κ_1: 1.22 κ_2: 1.18 κ_3: -0.88
Cluster 6: ω: 0.13 μ: -2.85 ν: 1.01 κ_1: 1.24 κ_2: 1.21 κ_3: -0.85
Cluster 7: ω: 0.09 μ: 2.83 ν: -0.94 κ_1: 1.26 κ_2: 1.23 κ_3: -0.77
Cluster 8: ω: 0.14 μ: 2.82 ν: 2.18 κ_1: 1.23 κ_2: 1.19 κ_3: -0.87

Figure S2: Mixture model for correlated torsions. The contour plot indicates the log density of a mixture model and the points (in red) mark the mean location for the components. **Left**: original correlated torsion that derived from filtered COD set. **Right**: updated correlated torsion by adding more observations from ChEMBL[12] database.

(a)

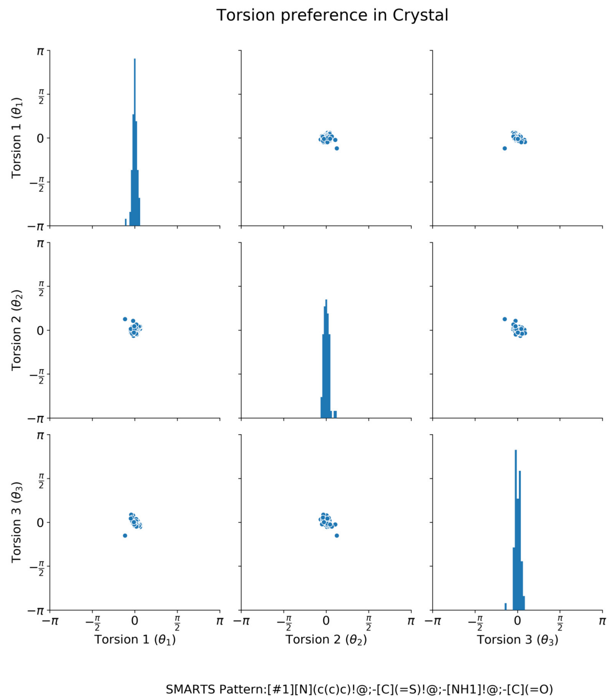
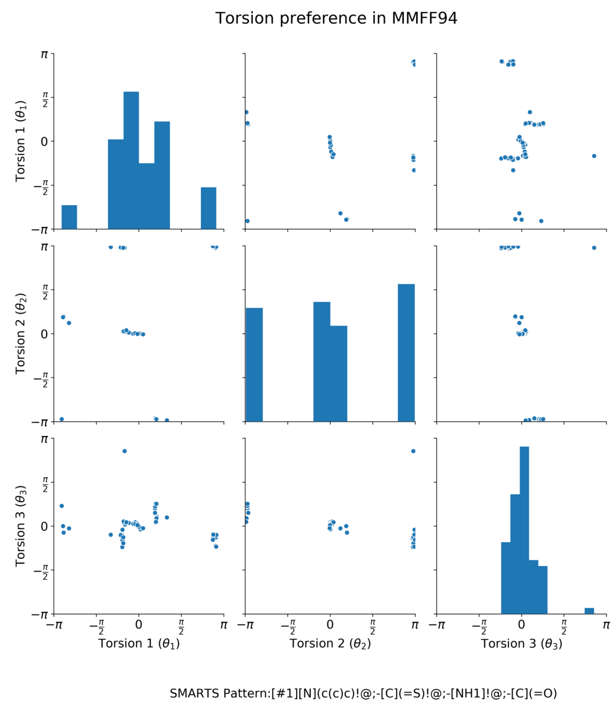


Figure S3: Higher order correlated torsions in (a) Crystal (b) MMFF94 and (c) GFN2. Torsion angles are measured in radian. We can see that all torsion angles are highly concentrated around 0° in crystal and GFN2. It suggests that the C=O and C=S are oriented in opposite directions and form a pseudo six-membered ring. This conformation promotes C=O – H-N intramoleccular hydrogen bond. The torsion angles in MMFF94 form a bimodal distribution (around 0° and 180°).

(b)



(c)

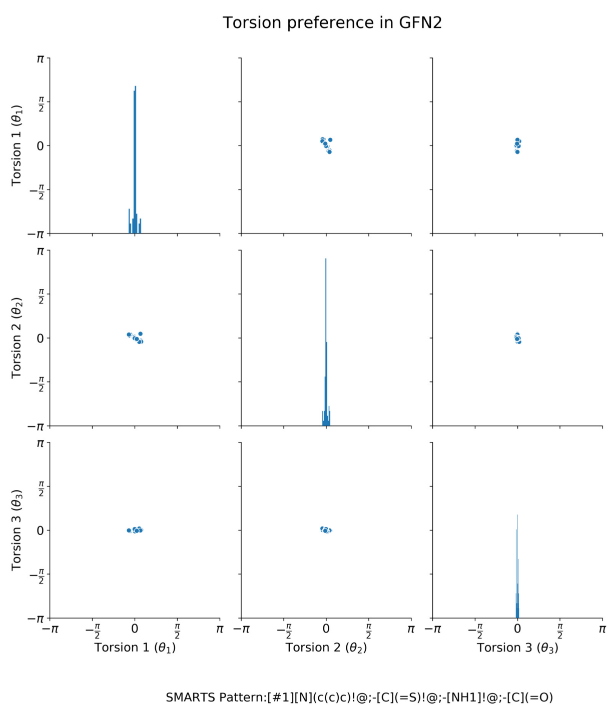


Figure S3: Higher order correlated torsions in (a) Crystal (b) MMFF94 and (c) GFN2. Torsion angles are measured in radian. We can see that all torsion angles are highly concentrated around 0° in crystal and GFN2. It suggests that the C=O and C=S are oriented in opposite directions and form a pseudo six-membered ring. This conformation promotes C=O – H-N intramoleccular hydrogen bond. The torsion angles in MMFF94 form a bimodal distribution (around 0° and 180°).

Appendices 3: Tables and Figures

Tables

Table S4: Wilcoxon signed rank test versus number of rotatable bonds with MMFF94 as energy function. BOA-EI and BOKEI are the Bayesian optimization with standard expected improvement and knowledge-based expected improvement respectively. GA represents the Genetic algorithm. We tested whether BOKEI found lower energy conformations than the two other methods, GA and BOA-EI, in the search (*i.e.* one-sided test). R (3.4.4)[13] was used to perform the hypothesis test. As the number of molecules with more than 10 rotatable bonds were small, we grouped the molecules with 11-13 and 14-18 rotatable bonds together when performing statistical test. The test showed the energy difference (between BOKEI and BOA-EI) was statistically significant ($p < 0.01$) across all rotatable bonds. The energy difference (between BOKEI and GA) were statistically significant up to ten rotatable bonds.

Number of rotatable bonds	BOKEI	BOA-EI	GA
2	NA	**3.82e-05**	**3.82e-06**
3	NA	**1.50e-08**	**1.23e-10**
4	NA	**3.04e-06**	**2.16e-13**
5	NA	**3.42e-11**	**3.54e-16**
6	NA	**9.84e-09**	**3.66e-13**
7	NA	**8.15e-06**	**1.05e-11**
8	NA	**1.14e-07**	**1.66e-05**
9	NA	**1.25e-06**	**1.75e-05**
10	NA	**2.21e-05**	**4.17e-07**
11-13	NA	**4.00e-03**	0.06
14-18	NA	**1.00e-03**	1

Table S5: Wilcoxon signed rank test across number of rotatable bonds on all stochastic search algorithms with GFN2 as energy function. BOKEI and BOA-EI stand for Bayesian optimization with standard expected improvement and knowledge-based expected improvement respectively. We would like to test whether BOKEI found lower energy conformation than BOA-EI in the search (*i.e.* one-sided test). We performed the statistical test in R (3.4.4)[13]. As the number of molecules with more than 10 rotatable bonds were small, we grouped the molecules with 11-13 together when performing statistical test. The test showed the energy difference (between BOKEI and BOA-EI) was statistically significant ($p < 0.01$) across all rotatable bonds.

Number of rotatable bonds	BOKEI	BOA-EI
2	NA	**7.90e-04**
3	NA	**1.41e-06**
4	NA	**3.41e-08**
5	NA	**9.31e-07**
6	NA	**1.21e-04**
7	NA	**2.02e-06**
8	NA	**9.84e-07**
9	NA	**1.34e-04**
10	NA	**2.34e-04**
11-13	NA	**6.14e-03**

Table S6: Average MMFF94 energy difference in different stages: 40%, 60% and 100% of the maximum number of energy evaluations. The value found by BOA-EI was used as reference. Negative value indicated BOKEI found lower energy conformations than BOA-EI. The median and the interquartile range were reported. In general, BOKEI outperformed BOA-EI in the early stage, and the energy difference diminished as more evaluations were used.

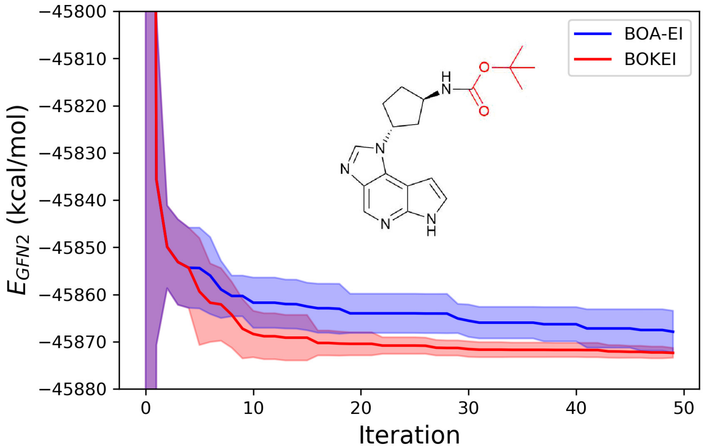
No. of rotatable bonds	$\text{Median}_{40\%}$	$\text{Median}_{60\%}$	$\text{Median}_{100\%}$	$\text{IQR}_{40\%}$	$\text{IQR}_{60\%}$	$\text{IQR}_{100\%}$
2	-12.567	-8.482	-0.736	10.445	7.708	0.673
3	-17.910	-11.980	-1.151	22.897	14.648	1.454
4	-12.305	-8.897	-0.962	25.711	21.753	1.721
5	-24.361	-12.811	-2.324	55.886	44.117	3.666
6	-24.204	-15.089	-2.883	66.497	43.808	4.267
7	-16.496	-20.423	-1.995	87.715	72.510	5.328
8	-10.831	-15.354	-3.472	75.566	65.496	4.443
9	-30.170	-12.668	-4.177	174.034	117.217	4.443
10	-83.867	-18.585	-5.077	166.109	122.901	5.592
11	-87.031	-18.255	-4.801	585.249	133.746	11.216
12	-7.456	-66.339	-0.901	1072.109	410.378	8.915
13	24.885	284.054	-4.037	903.960	742.068	8.823
14-18	68.935	58.545	-3.569	1118.353	1020.377	12.773

Table S7: Average GFN2 energy difference in different stages: 40%, 60% and 100% of the maximum number of energy evaluations. The value found by BOA-EI was used as reference. Negative value indicated BOKEI found a lower energy conformation than BOA-EI. The median and the interquartile range were reported. In general, the BOKEI found lower energy conformations than BOA-EI in the early stage for molecules with seven or fewer rotatable bonds. The energy gap between BOKEI and BOA-EI diminished as more evaluations were used. The molecules with eight or more rotatable bonds had positive energy difference in the early stage, which suggested that some of the correlated torsions could be under-estimated.

No. of rotatable bonds	$\text{Median}_{40\%}$	$\text{Median}_{60\%}$	$\text{Median}_{100\%}$	$\text{IQR}_{40\%}$	$\text{IQR}_{60\%}$	$\text{IQR}_{100\%}$
2	-14.143	-8.438	-0.276	19.0	24.0	0.0
3	-33.585	-13.870	-0.658	79.0	68.0	1.0
4	-8.817	-11.639	-0.856	62.0	58.0	1.0
5	-53.612	-15.612	-1.815	371.0	221.0	4.0
6	-14.688	-22.946	-1.857	210.0	134.0	4.0
7	-71.926	-50.814	-1.772	391.0	367.0	3.0
8	54.914	4.356	-4.117	839.0	889.0	11.0
9	119.435	11.961	-2.927	1415.0	670.0	5.0
10	-57.236	-95.970	-3.228	1866.0	948.0	6.0
11	88.074	-950.314	-4.405	1728.0	3083.0	4.0
12	-1075.099	-921.136	-4.961	2683.0	3248.0	11.0
13	1148.000	1937.425	-3.796	7959.0	5380.0	11.0

Figures

(a)



(b)

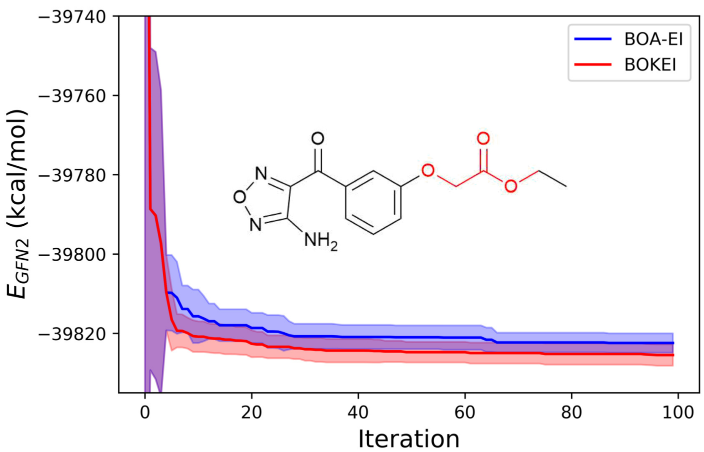
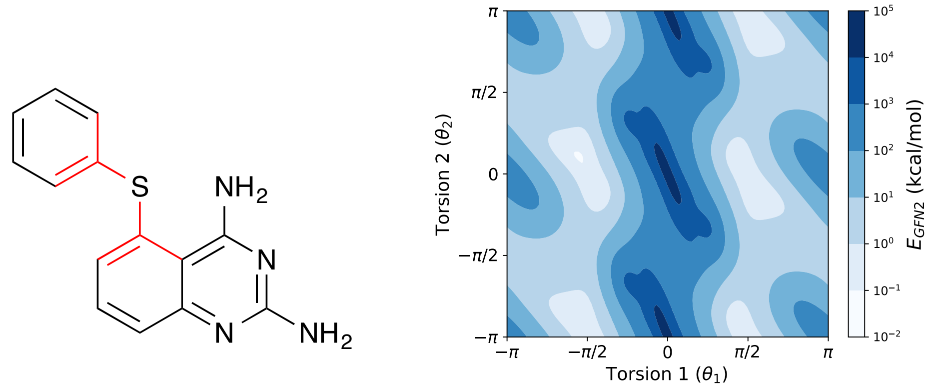
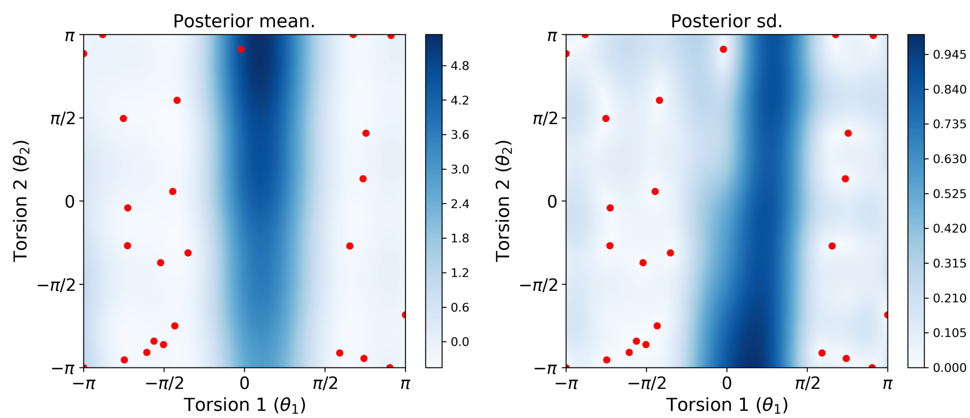


Figure S4: **(a)** Example S1. **(b)** Example S2. The red line and the blue line in the plot represented average convergence rate of the BOKEI and BOA-EI in finding lower energy conformations respectively, with ±1 sample standard deviation. The corresponding molecule and the correlated torsion is highlighted in red. GFN2 energy function was used. BOKEI consistently found lower energy conformations than BOA-EI in early stage and the energy gap reduced as the number of iterations increased.

(a)



(b)



(c)

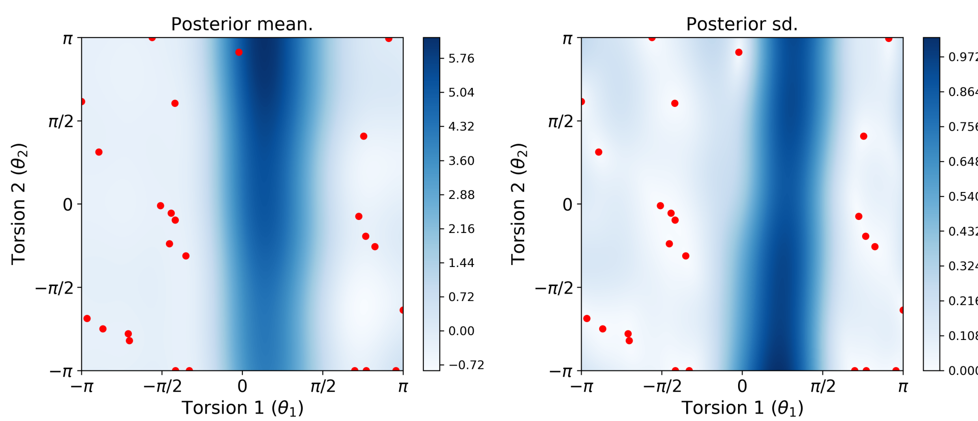
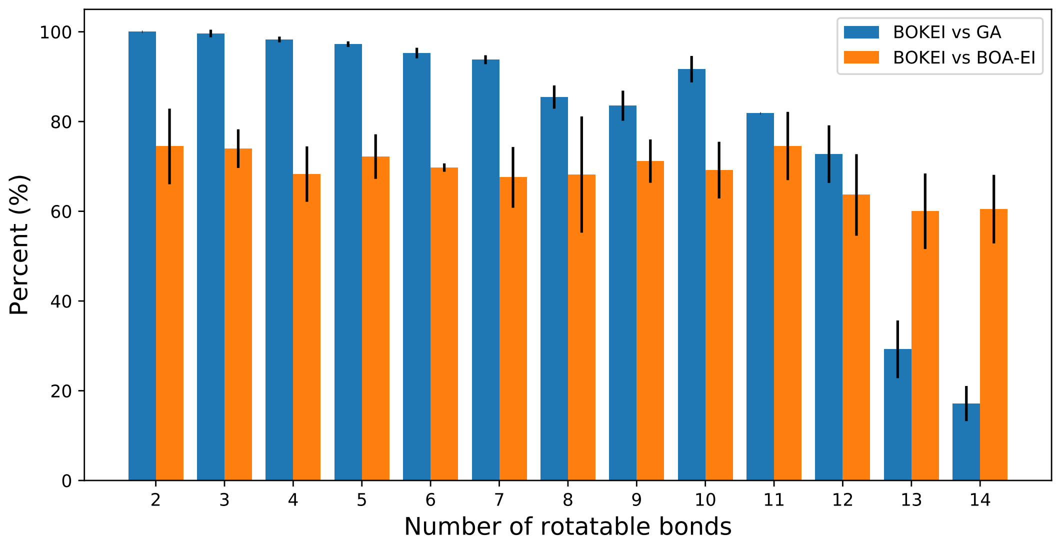


Figure S5: **(a)** 5-phenylthioquinazoline-24diamine and its GFN2 potential energy surface. Posterior mean (normalized) energy landscape and posterior standard deviation. **(b)** BOA-EI **(c)** BOKEI. The samples were more concentrated in the low energy regions for BOKEI than that in BOA-EI. The BOA-EI had higher chance to sample in high energy region (around the corners). The uncertainty is high in the high energy regions, due to small number of observations.

(a)



(b)

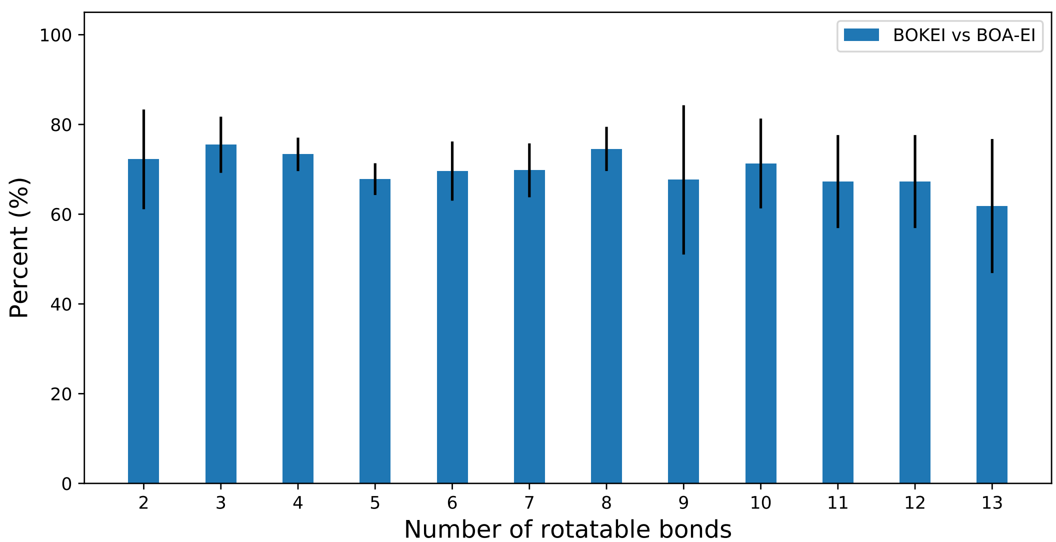


Figure S6: **(a)** Percentage frequency of BOKEI found lower energy conformations than BOA-EI and GA, with geometry-optimized MMFF94 energy. **(b)** Percentage frequency of BOKEI found lower energy conformations than BOA-EI, with GFN2 energy.

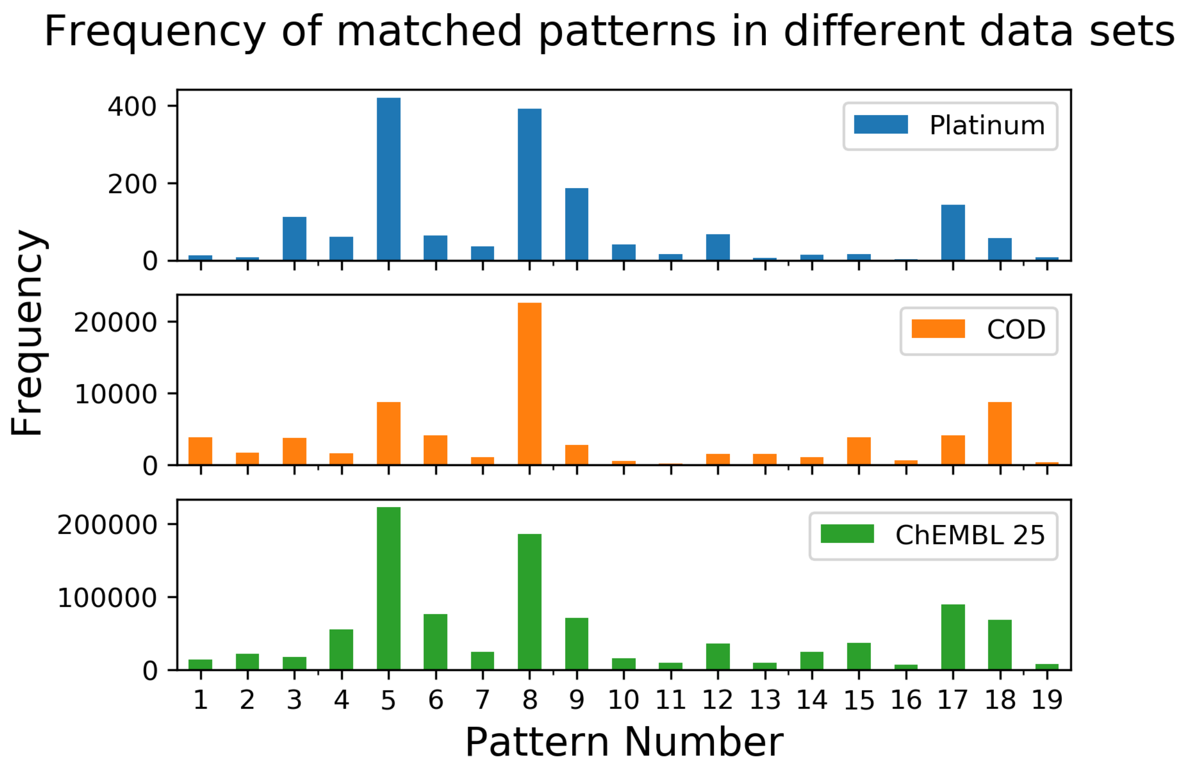
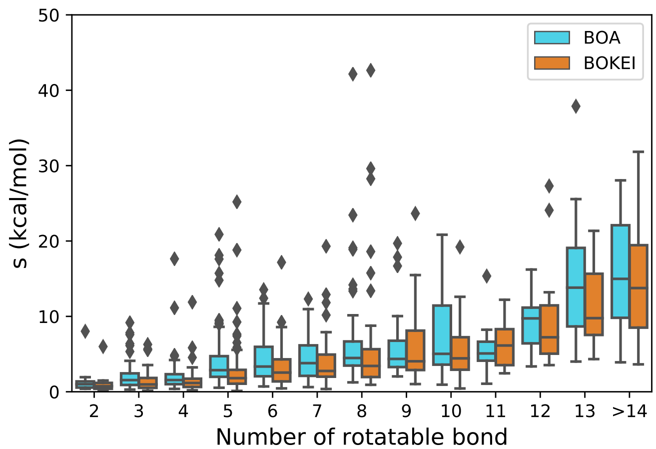


Figure S7: Frequency of matched patterns in three data sets: Platinum dataset, Crystallography Open Database (COD) and ChEMBL 25. The pattern 5 and 8 are frequently observed, followed by pattern 17 and 18.

(a)



(b)

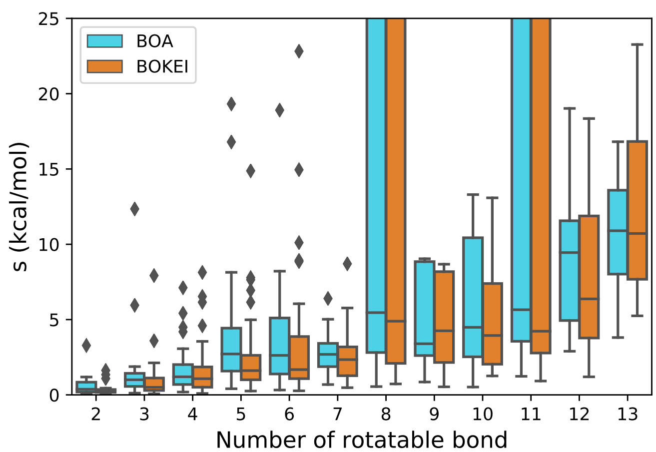


Figure S8: Sample standard deviation, s, of the energy of the output conformations in five independent runs: (a) MMFF94, and (b) GFN2. The sample standard deviation increases as the number of rotatable bonds increases in both cases. BOKEI has a lower median sample standard deviation than BOA in almost all cases. In (b), for a few molecules with eight and eleven rotatable bonds, one in five runs terminated with a high energy, which resulted in high sample standard deviations.

Appendices 4: Molecules in benchmark set and the molecules excluded from analysis

Name			
omegacsd_PMBSAN10	0C1_3UPH_A	7X7_2VX0_A	LZ0_2VUD_B
omegapdb_1eby	0CA_3UPE_A	7X8_2VX1_A	M0A_5AWB_A
omegacsd_PEXJEH	0CZ_3UUA_A	7XY_2OZ5_A	M72_4M73_A
omegacsd_VORDAH	0HL_3VFE_A	85Z_5AK2_A	M73_4DJP_A
omegacsd_DIGSIV	0HM_4D8N_A	861_4G50_A	M86_4DJQ_A
omegacsd_CIJVAS	0KJ_4DJR_A	879_3L38_A	MG0_4MRO_A
omegacsd_DCTXAN	0LQ_4GB2_B	8PC_3FNE_A	MGV_2V8Y_A
omegapdb_1s63	0LR_4GJ5_A	8Y6_2YOJ_A	MJU_4KTN_A
omegacsd_HEXYIS	0LX_4DV8_A	94M_3O6O_A	ML0_3HF8_A
omegacsd_BEKDIE	0NL_4E4N_A	965_3IPQ_A	MP7_3QC4_A
omegacsd_PMPAIN	0NR_4E73_A	988_3G58_A	MPX_4ITO_A
omegacsd_MVERIQ	0OG_4E92_A	993_2P3T_B	MS4_2WD3_A
omegacsd_NADYIA	0QX_4EPV_A	9RA_4K6I_A	MUJ_4AT5_A
omegacsd_LINLAV	0S2_4EO6_A	A18_2Q96_A	MUU_2PSU_A
omegacsd_SAWZUL	0S3_4EO8_A	A60_3O9E_A	MW5_2YC3_A
omegapdb_1w5v	0S6_4F3I_A	A68_3SA9_A	MZ3_2QI0_B
omegacsd_AOPCHY	0SL_4FA3_A	ANH_3PLK_A	MZ4_2QI1_A
omegapdb_2f4j	0SZ_4FAM_A	AVD_3CEM_A	MZ5_2QI3_B
omegapdb_2h42	0YN_4MHW_A	AWF_4CCU_A	MZ6_2QI4_A
omegacsd_FABSOQ10	16T_4HLD_A	AX4_3BMO_A	MZ7_2QI5_B
omegacsd_HEVXEL	17D_3QS8_A	AXI_4AG8_A	MZ8_2QI6_A
omegacsd_HEPGAK	18F_4JV6_B	B84_3IMX_A	N4I_3QPJ_B
omegacsd_MBZTZT10	1A0_4HXW_A	BBE_4B5B_A	N7F_4A5S_A
omegacsd_SADXIE	1AZ_2WHI_A	BK1_3I7B_A	NH7_2ZIR_B
omegapdb_1xp0	1BN_3PB9_X	BK3_3MWU_A	NHK_3NX7_A
omegacsd_FEYKUP	1C2_2F6T_A	BL2_4IK2_A	NJQ_3NJQ_A
1v4s	1GJ_4IVW_A	BLF_4DIJ_A	NK8_3QRS_B
omegacsd_LIKGIV	1H2_4J0R_A	BLR_4I3B_A	NN3_2GIR_A
omegacsd_FAHXIV	1JT_4J3F_A	BPQ_3BLL_A	NPM_4JIY_A
omegacsd_KAVLIC	1KG_4JCH_A	BPZ_2ZKC_A	NSI_2HFP_A
omegacsd_ANTZOA	1M9_4JJS_A	BQM_3BQM_B	NWL_4B3U_A
1lpz	1MB_4JJU_A	BQN_3BQN_B	NX6_3QH1_A
omegacsd_FURACM	1ML_4JMU_A	BSI_1I76_A	NZA_2Q5S_A
omegacsd_YOWYAK	1OQ_4K2G_B	C03_2P4Y_A	O2N_3ZSO_A
omegacsd_JOTDAX	1P9_4K69_A	C9A_3PHE_A	OAP_1KVO_A
omegacsd_LADTOZ	1QR_4KFN_A	CA4_2OZ7_A	OCV_1HB1_A
omegacsd_VUSKID	1U0_4KXY_A	CB6_2Y71_A	OFG_4CCB_A
omegacsd_TAZXAT	1U4_3FV4_A	CBO_2BEL_A	OHT_2GPU_A
omegacsd_LIKBUC	1V6_4L2L_A	CEI_4MRY_A	OYP_3RDE_A
omegacsd_GEKWAU	1VJ_4OYB_A	CGS_2W0D_A	P36_3HP2_A
omegacsd_HASCOT	208_2GTK_A	CJC_4GGL_A	P74_4AGO_A

Continued on next page

Name			
omegacsd_CLPHAC	218_4MUU_A	CL6_2XFH_A	P9L_4A9L_A
omegacsd_PXBVCP10	21L_3V9V_A	CVI_3VW1_B	PBD_4MHY_A
omegacsd_FOPMIG	23U_3DHK_H	CZE_5DQF_A	PF6_3HQY_A
omegacsd_SURREC	240_2Q59_A	D2J_3NZ6_X	PI6_1B6M_B
omegapdb_1h6h	24P_3FH5_A	D2R_3TD8_A	PN3_3PN3_A
omegacsd_TAZPUF	26G_4MES_A	D32_3GZ9_A	PT6_4CV2_A
omegacsd_FAJVAN	274_2J7T_A	D71_3G4I_A	PU8_1UYD_A
omegacsd_HEHVOF	28P_3U9W_A	D99_3ASX_A	PU9_1UYE_A
omegacsd_SEMXEN	2AZ_2GM1_A	DB4_4MXP_A	PUZ_1UYI_A
omegacsd_LACPAG	2EY_4MYA_A	DB8_4MXO_A	PYI_3D2V_A
omegapdb_2fwz	2G6_4OI5_E	DF6_3VW8_A	PZX_4IE9_A
omegapdb_1d4i	2KD_4NJ3_A	DFH_4F6O_A	Q13_4LWA_A
omegacsd_KOFLIA	2LK_4NR4_A	DG7_3PG3_A	Q16_4UGF_A
omegacsd_PACVUK	2N0_4PCE_A	DH1_3F8Y_A	QGF_3Q2G_A
omegacsd_KOFKUL	2OH_2E2R_A	DKI_4AAA_A	QLE_4B7S_A
omegacsd_YOXDEU	2PJ_4PES_A	DS4_3W0Y_A	QN7_4ACU_A
omegacsd_ACPIXZ	2QE_4NWC_A	DS5_3W0A_A	R11_1G32_B
omegacsd_FUHLID	2SQ_4OJR_A	DS6_3AZ3_A	RF1_3G1O_A
omegacsd_KOFLOG	2T2_4OK5_A	DU4_3T70_A	RKD_4AXA_A
omegacsd_SIZDOU	2T9_4OKS_A	DUQ_2Y8C_A	RO0_2FVJ_A
omegacsd_GIKJOZ	2TP_2C3U_A	DWA_4FRI_A	RX5_3Q4B_A
omegacsd_CODYUP10	2WZ_4PX5_A	E41_3ZXH_A	S5B_4AKN_A
omegapdb_1g4s	2YK_4PY1_A	E5S_3WCJ_A	S6L_5AJW_A
omegapdb_2aj8	2ZL_4PKR_A	EA4_3RDD_A	S8Z_4ACX_A
omegacsd_CYCLIZ10	30H_4PKS_A	EFU_4EFU_A	SF1_2Q61_A
omegacsd_COYPIP	33M_4QJZ_D	ENM_2PNU_A	SK2_2F6V_A
omegacsd_PMCPRC10	33U_2ZO3_H	ER4_3WCM_A	SK4_3FRG_A
omegacsd_BZAPUC20	355_4TOS_A	F21_2G72_A	SP6_2OW9_A
omegacsd_NADZIB	370_4CAF_A	F53_3O9G_A	SUY_4GFN_A
omegacsd_FOYLIO	389_3F7G_E	F72_3SA4_B	SX8_3DKG_A
omegacsd_HALSES	3BM_3EQC_A	FPW_3TUC_A	SZ8_3RXK_A
omegapdb_2h03	3CZ_3CZR_A	FRG_1M48_A	T05_4HLC_A
omegapdb_1v2k	3EJ_3KEJ_A	FTP_1G4T_A	T08_3W0J_A
1ig3	3KE_3KEC_A	G08_4NJS_B	T27_4KFB_A
omegacsd_DIAVER	3PZ_3PZ1_B	G0G_2PQZ_A	T74_3EQR_A
omegapdb_1sa4	3Q5_3QD3_A	G3G_2R3W_A	TDK_2EZ9_A
1u1c	3Q6_3QD4_A	G73_4I2P_A	TDL_2EZ8_A
omegacsd_LEZGOM	3QO_3QOA_A	GHW_3GI2_A	TDP_5DGD_A
omegacsd_DAFVUB	3QT_3QTI_A	GMF_4B0Q_A	TH2_2CBO_A
omegacsd_BEXVOP	3U9_4CNH_A	GRL_2HB3_B	TM3_3KPE_A
omegacsd_CIDSIR	3UE_4RPU_A	GVJ_2UVZ_A	TMI_2BDM_A
omegacsd_JUVPAR	419_3F7H_A	GVK_2UW0_A	TOP_4KM2_A
omegapdb_1yvx	43N_4XWA_A	GVN_2UW5_A	TPP_2PGO_A
1ia1	44U_3DA9_B	GVP_2UW9_A	TPS_5DD7_A

Continued on next page

Name			
omegacsd_COTXUE	454_3IGB_A	GVQ_2UW8_A	TPU_2WVA_A
omegacsd_FAVYEG	46C_2ZDT_A	H89_3U16_A	TPW_2JI6_A
omegacsd_FUMBOE	472_3IN3_A	H90_3U17_A	TQ2_4E1M_A
omegacsd_KUBHEU	478_3NU6_B	HBH_1Z1R_A	TQX_4E1N_A
omegacsd_PEBHAF	47D_2PDH_A	HDI_1LJT_A	TR7_4B2L_A
1mzc	4K0_4AOI_A	HI3_3TYV_A	TZD_2BFF_A
omegacsd_FAXPUP	4L3_4Z6I_A	HM4_2NQ6_A	TZM_4G44_A
omegacsd_SINKUV	4NY_4ZG6_A	HNR_4ARW_A	UA1_2I4G_A
omegacsd_HAVLUL	4O0_4ZG7_A	HTJ_3O84_A	UB2_3FVP_A
omegacsd_BZAPCX10	4O1_4Y76_A	HTL_3AHD_A	UBA_2F34_A
omegacsd_DERZAB	4Q3_4ZOM_A	I24_2VVT_A	UBC_2F35_A
omegacsd_VIPHOR	4V8_5BUE_A	I5S_2JDO_A	UBE_2QS3_A
omegacsd_PEKCEO	4VT_5BWR_A	I63_4DRK_A	UBF_2QS2_A
omegacsd_GASPUL	4XB_4B14_A	I6X_4B00_A	UBW_3T8C_A
omegacsd_ANTZOB	52U_5COK_A	IC8_2X38_A	V10_2VBD_A
omegacsd_CUNTIO	52W_5CON_B	IDQ_5AWD_A	V4E_5A64_A
omegacsd_MICONZ	532_3S7M_A	IMA_1LPG_B	VDN_3B2R_A
omegacsd_CALLEG	538_3KMG_A	IWH_4UCS_A	VGF_2WD8_A
omegacsd_DEBBOB	54M_4B77_A	IXE_4IXE_D	VGG_2WEI_A
omegacsd_VEXROF	569_3INH_A	IXF_4QHV_A	VHI_3VHI_A
omegapdb_1d4l	56M_5D25_A	IXG_4IXG_X	VJP_4URZ_R
omegacsd_FUHHAR	56R_5D1T_A	JHG_3QBF_A	VMY_4BQP_A
omegacsd_NBPENC	578_2PJ1_A	JJ3_2JJ3_A	VX6_2F4J_A
omegacsd_VOPDIN	57G_5D3H_A	JKF_3TIK_A	VYI_5APH_A
omegacsd_TPHPRO	591_3S7L_A	JPC_1YVZ_A	W07_3W0G_A
omegacsd_PILCOC	59D_5DE1_A	JQ1_3S92_A	W7Q_3W7Q_A
omegacsd_JOXFIL	59E_5DFD_A	JR9_4R3M_A	WCX_2XNP_A
omegacsd_DCLPET	5BO_3SIE_A	JWT_4B4M_A	WSH_2YEL_A
omegacsd_HEVJAT	5H8_5DPE_E	K13_3O99_A	X2O_2X2R_A
omegapdb_1ec0	5H9_5DPF_E	K14_3O9A_A	X45_3INF_A
omegapdb_2f34	5J9_5E13_A	K20_3O9C_A	XX8_4N4D_A
omegacsd_SIHGIZ	5JN_5E29_A	K23_3LBK_A	Y01_4XNV_A
1jla	5N2_5ECE_A	K2A_3O9B_A	Y38_4H84_A
omegapdb_2f14	5OU_5EIS_A	K2D_3O9F_A	Y46_3ZLK_A
omegapdb_2f6v	5Y0_4AN3_A	K88_2WEL_A	YNE_2YNE_A
omegapdb_2i0a	663_2P3U_B	KGG_4B78_A	YPW_4CMO_A
omegacsd_KOFMIB	711_1QBO_A	KIM_3CJG_A	YR4_3AUN_A
003_2JFZ_A	751_4ANV_A	KLI_3UDL_A	Z81_3LHG_A
017_4DQE_B	76E_5APK_A	KR1_3LAK_A	ZAA_3BC5_A
053_3SKA_A	778_1S63_B	KRH_4B1F_A	ZAH_2BKL_A
054_3SKE_A	77F_3SAA_B	KRW_3ZXZ_A	ZEN_1V2K_T
065_2QD7_B	78M_4AFK_A	L23_2W71_A	ZOO_3OOZ_A
082_3TY0_A	79M_4UYO_B	L28_5D26_A	ZST_2FZ8_A
08E_3U4O_A	7PP_3NMQ_A	L33_5D3J_A	ZZD_2WOG_A

Continued on next page

Name			
09A_3UDM_A	7X1_2VWU_A	LBY_2ZIN_A	ZZN_2WXG_A
09D_3UDP_A	7X2_2VWW_A	LF0_4ID1_A	ZZO_2WXH_A
09F_3UDR_A	7X4_2VWX_A	LF2_4GVM_A	
09G_3UDY_A	7X5_2VWY_A	LID_2GTM_A	
09L_4TKG_A	7X6_2VWZ_A	LUR_4OTY_A	

There were in total nine molecules (five in MMFF94 and four in GFN2) excluded from the analysis, due to early stopping in any one of the five runs. The molecules are listed in Table S9 and S10.

Table S9: Molecules that excluded from the analysis (MMFF94).

SMILES	Name
O(c1ccc(Nc2c3ccccc3[nH]c2c2ccccc2)cc1)C	OMEGACSD_PMPAIN
S(c1ccccc1)c1ncc(n1C)[C@H](O)c1ccccc1	OMEGACSD_VUSKID
N(C(C)C)(C(=O)c1ccc(cc1)C)c1c(sc(c1)c1ccccc1)C(=O)O	OMEGAPDB_1yvx
N1C(=O)/C(=C(\CCC(=O)OC)/c2ccccc2)/C=C1c1ccccc1	OMEGACSD_FAXPUP
O=C(CCC)OC[C@@H](OC(=O)CCC)CO[P@@](=O)([O-])O [C@@H]1 [C@H](O)[C@H](O)[C@@H] (OP(=O)([O-])[O-])[C@H](O)[C@H]1O	DB4_4MXP_A

Table S10: Molecules that excluded from the analysis (GFN2).

SMILES	Name
COc1c(ccc(c1)Cc1c2cc(c(cc2ccn1)OC)OC)OC	OMEGACSD_MVERIQ
O=P([O-])([O-])O[P@](=O)([O-])OCCc1c(C)c(c(s1) [C@H](O)CO)Cc1cnc(C)nc1N	1U0_4KXY_A
O=C(N[O-])[C@@H](CCCC[NH2+]Cc1ccc(cc1)F) C[C@@H](OC)c1ccc(cc1)F	0LX_4DV8_A
c1cc(c(cc1C(c1ccc(c(c1)C)OC[C@@H](C(C)(C)C)O) (CC)CC)C)O[C@H](CCO)CO	YR4_3AUN_A

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. This research was supported in part by the University of Pittsburgh Center for Research Computing through the computational resources provided.

References

(1) J. C. Cole, O. Korb, P. McCabe, M. G. Read and R. Taylor, *Journal of Chemical Information and Modeling*, 2018, **58**, PMID: 29425456, 615–629.

(2) P. McCabe, O. Korb and J. Cole, *Journal of chemical information and modeling*, 2014, **54**, DOI: `10.1021/ci500156d`.
(3) K. V. Mardia, C. C. Taylor and G. K. Subramaniam, *Biometrics*, 2007, **63**, 505–512.
(4) K. V. Mardia and J. Frellsen, in *Bayesian Methods in Structural Bioinformatics*, ed. T. Hamelryck, K. Mardia and J. Ferkinghoff-Borg, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 159–178.
(5) A. P. Dempster, N. M. Laird and D. B. Rubin, *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, **39**, 1–38.
(6) T. A. Halgren, *Journal of Computational Chemistry*, **17**, 490–519.
(7) C. Bannwarth, S. Ehlert and S. Grimme, *Journal of Chemical Theory and Computation*, 2019, **15**, PMID: 30741547, 1652–1671.
(8) S. Riniker and G. A. Landrum, *Journal of Chemical Information and Modeling*, 2015, **55**, PMID: 26575315, 2562–2574.
(9) G. Landrum, *RDKit: Open-Source Cheminformatics*, Available at `http://www.rdkit.org`, 2018.
(10) D. Wales and J. Doye, *The Journal of Physical Chemistry A*, 1998, **101**, DOI: `10.1021/jp970984n`.
(11) S. Grimme, C. Bannwarth, S. Dohm, A. Hansen, J. Pisarek, P. Pracht, J. Seibert and F. Neese, *Angewandte Chemie International Edition*, 2017, **56**, 14763–14769.
(12) A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit and A. R. Leach, *Nucleic Acids Research*, 2016, **45**, D945–D954.
(13) R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018.