Systematic Comparison of Experimental Crystallographic Geometries and Gas-Phase Computed Conformers for Torsion Preferences

Dakota L. Folmsbee,†‡ David R. Koes,¶ and Geoffrey R. Hutchison*,†§

†Department of Chemistry, University of Pittsburgh, 219 Parkman Avenue, Pittsburgh, PA 15260
‡Department of Anesthesiology & Perioperative Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, United States
¶Department of Computational & Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States
§Department of Chemical & Petroleum Engineering, University of Pittsburgh, 3700 O’Hara Street, Pittsburgh, PA 15261

E-mail: geoffh@pitt.edu

Abstract

We have performed exhaustive torsion sampling on more than 3 million compounds using the GFN2 method and performed a comparison of experimental crystallographic and gas-phase conformers. Many conformer sampling methods derive torsional angle distributions from experimental crystallographic data, limiting the torsion preferences to molecules that must be stable, synthetically accessible, and able to be crystallized. In this work, we evaluate the differences in torsional preferences of experimental crystallographic geometries and gas-phase computed conformers from a broad selection
of compounds to determine whether torsional angle distributions obtained from semi-empirical methods are suitable for conformer sampling. We find that differences in torsion preferences can be mostly attributed to a lack of available experimental crystallographic data, with some deviations derived from gas-phase geometry differences. GFN2 demonstrates the ability to provide accurate and reliable torsional preferences that can provide a basis for new methods free from limitations of experimental data collection. We provide Gaussian-based fits and sampling distributions suitable for torsion sampling, and propose an alternative to ETKDG using quantum-torsion derived distance geometry methods (QTDG).

Introduction

Most molecules exhibit some level of conformational flexibility, the existence of multiple low-energy geometries that differ mostly by changes in the torsional angles of both acyclic and ring bonds. Many methods have been developed to sample conformations, with benchmarks frequently focusing on finding one geometry close to an experimental crystal structure.\(^1\text{-}^4\) Consequently, most conformer sampling methods derive torsional angle distributions from experimental crystallographic data\(^1\text{-}^8\) — not only to provide geometries close to such benchmarks, but also as large diverse repositories of “ground truth” geometric properties such as bond lengths, angles, and dihedrals.\(^9\text{-}^{11}\)

One challenge is that experimental crystallographic data are limited by the size of the data source\(^12\) and reflect some inherent biases. In order to be collected, the molecules must be stable, synthetically accessible, and actually made and crystallized. While new cryo-electron microscopy (cryo-EM) techniques are improving dramatically and have less stringent requirements on crystals, generally growing a high quality crystal for small molecule crystallography is a time-consuming process. Moreover, it is known that compounds with experimental crystal structures are generally smaller and exhibit fewer conformers than other compounds.\(^13\) Similarly, compounds containing elements outside the common organic subset
(e.g., B, As, Se) or less common chemical motifs may be poorly represented in experimental crystallographic databases. Also, even for compounds found in an existing database, much chemistry is performed in solution and gas phases, where solid-state preferences may not directly apply. Finally, several works have noted challenges with deriving data from some crystallographic databases.\textsuperscript{12,14}

Consequently, finding unbiased alternative sources of accurate and reliable torsional angle preferences could significantly expand the use of conformational sampling to new chemical space. Typically, sampling has been performed using small molecule force fields, which have shown limited fidelity compared to density functional and other first principles quantum chemical methods.\textsuperscript{15,16} The development of efficient dispersion-corrected semiempirical methods such as GFN2,\textsuperscript{17} as well as new machine learning methods such as ANI\textsuperscript{18–21} and OrbNet,\textsuperscript{22,23} offer improved accuracy of torsional angles and non-bonded interactions with moderate computational cost. Moreover, several large-scale computational efforts including PubChemQC\textsuperscript{24} and the QCArchive torsion scans\textsuperscript{25} have provided large amounts of optimized geometries using high-quality density functional methods.

In this work, we outline an extensive effort to analyze the conformers and torsional angle preferences of over 3 million organic small molecules, using exhaustive sampling using the GFN2 method across both the experimental Crystallographic Open Database\textsuperscript{26} (COD) and multiple sets of small molecules, including PubChemQC\textsuperscript{24} We compare the potential bias between the crystallographic and gas-phase geometries and individual torsion patterns, including analysis with $\omega$B97X-D3\textsuperscript{27} with the def2-SVP basis set.\textsuperscript{28,29}

\textbf{Methods}

Molecules for this work were compiled from several sources, including 88,106 organic compounds from the Crystallographic Open Database,\textsuperscript{9,10} 3,009,591 from PubChemQC,\textsuperscript{24} 88,550 molecules from the Pitt Quantum Repository, and previous work on conformational flexi-
bility, which included 70,850 molecules from a subset of ZINC and 4,378 molecules from the Platinum ligand database. For all sources, the largest substructure was retained (i.e., removing solvent or salts from the crystallographic unit cells). For compounds without initial 3D coordinates, Open Babel 3.1 was used to generate initial coordinates. As noted above, the total set of compounds included over 3 million unique molecules.

For each molecule, conformers were generated using the CREST program to exhaustively sample the potential energy surface, using default parameters and the GFN2 method for energies and optimized geometries. In some cases, CREST produced fragments or chemical rearrangements (e.g., producing different compounds than the input, based on the InChI identifier) — these systems were excluded from analysis. While, in principle, CREST can generate many conformers per compound, we find that for the vast majority of compounds, only ca. 10-20 conformers are generated within 6 kcal/mol as calculated by GFN2 (Figure S1), consistent with our previous work.

In this work, the lowest energy conformer by GFN2 energy was analyzed. Torsional angle SMARTS patterns from the experimental torsion knowledge-based distance geometry approach, (ETKDG) were used to generate the histograms for the gas-phase data. ETKDG derives torsional preference distributions for the RDKit distance geometry coordinate generation method, from analysis of experimental crystallographic data for a set of hierarchical dihedral patterns.

These patterns are constructed from molecules with central bonds of C–C (168 acyclic patterns), C–O (56 acyclic patterns), C–S (16 acyclic patterns), N–C (131 acyclic patterns), N–S (4 acyclic patterns), and S–S (1 pattern) and create the 387 acyclic and 105 ring dihedral SMARTS patterns used to generate histograms of matching torsions with a stepsize of 5° using RDKit Python scripts (see supporting information). Figures depicting the SMARTS patterns were generated using SMARTS.plus.

For selected compounds, to compare the GFN2 geometry with density functional theory, optimization was performed with ORCA 4.2.0, using the ωB97X-D functional and
def2-SVP basis set, which has proven to produce fairly accurate conformational energetics although some errors still exist, when compared with more accurate methods.\textsuperscript{43–46}

**Results and Discussion**

Traditionally, conformer sampling is refined via classical force fields, which yield a poor correlation with energies from more accurate quantum methods.\textsuperscript{15,16} On the other hand, geometry optimization with most density functional methods requires hours per conformer, making large-scale sampling prohibitive. Recently larger data sets have emerged, including this work, as well as the ANI-1x,\textsuperscript{47} GEOM,\textsuperscript{48} QCArchive torsion scans,\textsuperscript{25} QMugs,\textsuperscript{49} and SPICE sets.\textsuperscript{50}

**Comparing Overall Geometries**

Often, coordinate generation and conformer tools are evaluated, in part, by comparing the ensemble of generated geometries against experimental crystallographic geometries. For example, the ETKDG method has proven to generate structures with small root mean square displacement (RMSD) when compared to small molecule crystal structures and bound ligand geometries.\textsuperscript{1,3,6,7}

Consequently, our first comparison is between the CREST-generated GFN2 conformer ensembles and the ETKDG conformer ensembles across the Crystallographic Open Database (COD) and Platinum data sets. Computing the smallest non-hydrogen RMSD of each generated ensemble and the experimental crystal structure geometry (Fig. S2) we consider the ratio between the CREST and ETKDG RMSDs (i.e., values lower than one indicate CREST ensembles have smaller RMSD than ETKDG and are thus closer to the experimental crystal structure).

We find that the CREST ensembles perform substantially better than ETKDG on molecules with few rotatable bonds (e.g., zero to \(\sim 3–4\) rotors) on the COD set, but comparable RMSD
across a broader range of COD molecules, and potentially slightly worse on the Platinum set (e.g. Figure 1).

Figure 1: Ratio of the smallest non-hydrogen RMSD between experimental crystallographic geometry and CREST or ETKDG conformers, as a function of the number of rotatable bonds, showing the (a) boxplots, (b) medians) from Crystallographic Open Database (COD) and (c) boxplots, (d) medians from the Platinum dataset.

As illustrated in Figure 1, for the COD small molecule crystal structures, the CREST geometries have much smaller RMSD compared to ETKDG (e.g., the median RMSD ratio is less than 1.0) up to molecules with 3-4 rotatable bonds. The improved treatment of non-bonded interactions and electrostatics in the quantum GFN2 method likely gives rise to these geometries more closely matching experimental crystallographic geometries.15

Note, however, that for the Platinum set, derived from bound PDB ligands, performance of the CREST ensembles is generally somewhat worse than ETKDG (e.g., the median RMSD ratio is greater than 1.0) particularly starting at molecules with five rotatable bonds and
growing significantly with the number of rotatable bonds. We speculate that this difference derives from comparing gas-phase GFN2 conformers from CREST with bound ligand geometries, which may be more stabilized in extended conformations due to intermolecular interactions with a binding site.\textsuperscript{51}

While there is a wide distribution of computed radius of gyration between the lowest-energy CREST/GFN2 generated conformations and the experimental crystal structures from COD and Platinum sets (Figure S3), as compiled in Figure 2, the median radius of gyration for experimental COD and CREST-generated conformers are relatively close, (e.g., the ratio of the two is usually close to 1.0, Figure S4). On the other hand, the Platinum crystallographic geometries show a notably larger radius of gyration than the CREST conformer (e.g., Figure S4), with the deviation growing as a function of the number of rotatable bonds.

![Figure 2: Calculated radius of gyration for lowest energy CREST conformer and experimental geometries as a function of the number of rotatable bonds for the (a) Crystallographic Open Database (COD) and (b) Platinum datasets.](image)

Overall, the results indicate that CREST-derived conformer ensembles perform comparable or better to ETKDG ensembles, particularly on the small molecule COD set, with the caveat of a potential bias towards compact conformations relative to the Platinum bound PDB ligands, likely due to a neglect of intermolecular interactions in the gas-phase geometries.
Individual Torsion Preferences

As mentioned above, traditional efforts to derive torsional preferences use experimental crystallographic databases.\(^{1,3,5-8}\) Given the reasonable agreement between CREST/GFN2 conformers and crystallographic geometries, particularly on the small-molecule COD set, the individual torsion preferences should be also be comparable. Thus, we compared the experimental torsions from COD to the gas-phase CREST/GFN2 generated conformers from COD, as well as to CREST/GFN2 generated conformers of all combined sets (e.g. PubChemQC, COD, etc.) across over 3 million compounds.

Comparing torsions between experimental geometries from COD and gas-phase lowest-energy conformers from CREST/GFN2 is intended to show the suitability of gas-phase torsion preferences to replace or supplement standard crystallographic analysis. Poor correlation between the individual torsion distributions could occur for a few main reasons — that the gas-phase and experimental crystal structures show distinct angles due to intermolecular interactions and packing effects, that the distributions have few points and are thus inherently noisy, or that the GFN2 semiempirical method is not sufficiently accurate to reproduce torsion angles.\(^{43,52}\)

Since the COD is significantly smaller than the Cambridge Crystallographic Database,\(^{11,53}\) however, this limits the number of torsion data points for some structural motifs. Expanding our analysis to include millions of molecules can improve the understanding of uncommon motifs, as well as potential torsional preference differences in the solid state, compared to the gas-phase geometries. This work will examine the comparison of crystal structure and calculated gas-phase torsional preferences for both ring and acyclic containing torsions.

To determine the degree of correlation, the \(r^2\) of the kernel density estimation (KDE) for both acyclic and ring torsion patterns was compiled in Figure 3a. The \(r^2\) of the KDE plot was used to smooth the histogram and avoid issues when correlating areas with few or no torsions in the COD with the regions where some but few torsions were present in the combined set. A total of 127 patterns yield \(r^2 > 0.8\), and the median acyclic \(r^2\) is 0.61,
indicating that while there are some differences, the torsional preferences show reasonable correlation.

The correlation is likely better than indicated by the median value, as shown by analyzing the 76 patterns with an $r^2$ less than 0.2. The patterns in this regime had a median of only 175 instances across the COD, which appears to be too few to form accurate distributions of dihedral angle preferences.

In short, we believe that in most cases the correlation between experimental crystallographic geometries and gas-phase conformer sampling with semiempirical quantum methods such as GFN2 are high enough to derive accurate torsional preferences. This is consistent with previous comparisons of energetic rankings of conformers between GFN2 and higher-level quantum chemical methods including density functional and coupled cluster calculations,\textsuperscript{15} as well as efforts to estimate torsional strain in crystal structures using DFT methods.\textsuperscript{52,54,55}

Ring torsional patterns, constrained by the nature of a ring, show even higher correlation, with the median $r^2$ at 0.83 (Figure 3b) and few patterns (16 out of 105) showing correlation below 0.5, again correlated with few matches.

![Figure 3: Correlation between experimental and gas-phase torsions across the COD data set for (a) acyclic patterns and (b) ring patterns.](image)

Individual acyclic torsional preferences can be further analyzed to determine the qualitative correlation between the crystal structures and gas-phase conformers. Torsion pattern
229 shown in Figure 4 demonstrates a high degree of correlation \( r^2 = 0.76 \) between the crystal structures and gas-phase CREST/GFN2 lowest energy conformers, while demonstrating the advantage of using additional data from the entire set of compounds. Pattern 229 has only 179 torsions from the COD, while the expanded data set boasts over 25 thousand instances. This increase in data clarifies torsional preferences in the range of 90° to 150°, as the conformers demonstrate a clearer peak in this region and less noise overall.

Figure 4: Histograms for pattern 229, including COD experimental torsional angles, gas-phase lowest-energy conformers from the same COD molecules, and gas-phase lowest-energy conformers across the entire dataset, indicating the strong correlations, and that increased quantity of data greatly refines the histograms.

Figure 5 demonstrates how the lack of data in the COD can impact qualitative assessments of torsional preferences. Correlation between the experimental and gas-phase COD torsions (left and middle panels) is 0.58. The experimental torsions show a mild preference around 60° and 120°, with a lot of noise between, while the calculated torsions have more defined peaks at 50°, 90°, and 132°. Much like the case above, the 10x increase in the number
of torsions in the combined set was able to provide enough to discern preferential angles.

Figure 5: Histograms for pattern 307, including COD experimental torsional angles, gas-phase lowest-energy conformers from the same COD molecules, and gas-phase lowest-energy conformers across the entire dataset, again indicating the how increased quantity of data greatly refines the torsional preferences (e.g., peaks near 50°, 90°, and 132°).

In addition to acyclic preferences, we analyzed the preferences of ring torsional patterns. Ideally, torsional patterns in rings should correlate well between experimental and calculated gas-phase geometries due to the steric constraints on the geometry. The correlation was analyzed in the same manner as above, by taking the r² of the kernel density estimation (KDE) for each ring torsion pattern and compiling them into Figure 3b. Compared to the median r² of 0.61 for the acyclic patterns, the ring patterns boasted a median r² of 0.83, indicating a significant correlation between the two sets.

Similar to acyclic patterns, the increase in available data bolsters the angular preferences of the ring torsion patterns. Figure 6 exemplifies this by demonstrating how the increase in the number of torsions for ring torsion pattern 50 affects the torsional preference. The cor-
The relation between experimental and gas-phase COD data is 0.79. Although the experimental data suggests 0°, 60° and 180° are prominent angles, the additional data from the combined sets firmly demonstrates that these are prominent angles for the torsion pattern, with a small subpopulation ~41°. The addition of more data exhibits a more complete picture of torsional preferences that better represent chemical space.

![Ring Torsion Pattern 50](image)

Figure 6: Histograms for COD experimental torsional angles, gas-phase lowest-energy conformers from the same COD molecules, and conformers across the entire dataset, again indicating that the increased quantity of data greatly refines the torsional preferences (e.g., a strong peak at 60°).

These expected similarities for the ring patterns are due to the intermolecular constraints the ring structure imposes on the geometry. There is less flexibility in ring systems. The correlation of individual ring torsions demonstrates the accuracy of the quantum torsional preferences and indicates the ability to use this information as an additional method for determining torsional preferences of structural motifs not yet examined through experimental means.
Importantly, sampling ring torsion angles in isolation is inherently challenging, since such angles are strongly correlated with other torsion angles in the same ring. In the ETKDG method, such correlation is resolved through the distance geometry methods, refining distances between all atoms in the ring (and molecule) together. As such, the ring torsion data in this work is intended either for refining such distance geometry methods or as an initial effort to refine ring puckering distributions using Cremer-Pople schemes.

The outlined work demonstrates the desire for additional methods based on quantum calculations where crystal structure constraints may not be suitable or data may not be prevalent due to experimental limitations. Using the example of ETKDG, an alternative, quantum torsion distance geometry (QTDG), could be useful for gas-phase applications. Moreover, a QTDG method would no longer be constrained to structural preferences derived from what can be synthesized and crystallized. This allows for an increase in the capability of the method because a larger amount of data would be available for a more diverse representation of chemical space.

Comparisons Between GFN2 and DFT Optimized Geometries

As noted above, some differences between the COD crystallographic torsional preferences and those calculated from gas-phase CREST/GFN2 lowest energy conformers derive from patterns with few candidates and thus somewhat noisy histograms. In other cases, the histograms suggest noticable differences involving new peaks, particularly from the combined set. For example in Figure 7, the experimental torsions show a mix of torsions in the range of 70° to 110° while the calculated torsions from the combined set also show a dominant peak at 90°.

This new peak could derive from a new subpopulation, for example, compounds not present in the COD or not synthetically accessible or amenable to crystallization. The new peaks could also derive from problems with the GFN2 method predicting incorrect torsion angles.
To test the latter hypothesis, $\omega$B97X-D/def2-SVP geometry optimizations were performed to verify the GFN2 optimized geometries. Five molecules containing torsion pattern 270 and an angle of $\sim$90° were randomly selected from the combined set. The DFT optimized geometries were found to be in strong agreement (e.g., within a mean absolute deviation of 1.54°) with the GFN2 torsion angles. Such compounds include steric constraints restricting the torsion to be $\sim$90°, as illustrated by Figure 8.

We compared ten patterns with obvious subpopulations (e.g., 90° in Fig. 7), including patterns 10, 23, 48, 53, 125, 149, 216, 253, 270, and 306. For each pattern, five molecules were selected, with the exception of pattern 23, which had only one molecule in the subpopulation. As illustrated in Figure S5, overall absolute torsion angle deviations between the GFN2-optimized and $\omega$B97X-D/def2-SVP optimized geometries were between 0-5°, with the median absolute deviation of 2.89°, and a mean absolute deviation of 4.88°.

Comparing across torsion patterns, Table S1 and Figure S6 indicate that some patterns,
Figure 8: Example of compound matching torsion pattern 270 with steric constraint forcing an angle \( \sim 90^\circ \). Figure from Avogadro2.\textsuperscript{58,59}

particularly 48 and 253 reflect larger absolute torsion angle differences of 15° and 10°, respectively, but overall \( r^2 \) correlation between GFN2 optimized and \( \omega \)B97X-D/def2-SVP torsion angles is high, even for these patterns.

Consequently, while GFN2 is an approximate semiempirical method, and correlations with more accurate quantum chemical methods are imperfect,\textsuperscript{15} we can conclude that differences between the crystallographic and gas-phase torsional preferences appear to be mainly derived from lack of data and the presence of new subpopulations in the combined set.

**Fitting and Sampling Torsion Angles**

The ETKDG methods derive torsional preferences using a potential energy term fit to a discrete cosine Fourier analysis:\textsuperscript{6,7}

\[
V(\phi) = \sum_{i=1}^{6} K_i [1 + \cos(d_i) \cos(i\phi)]
\]  

(1)

in which \( K \) is a force constant and \( d \) is a phase shift and limited to 0 and \( \pi \) arbitrarily. Consequently, we wish to compare the original published ETDKG fits to a similar cosine analysis of our data. In our work, the phase shift is also varied as a parameter using nonlinear curve fitting. Note that for both the ETKDG and cosine fits, the histogram probabilities
must be transformed into relative energies before fitting. We compare the ETKDG and cosine fits to a sum of fitted Gaussian peaks.

Compiled in Figure 9 are histograms of correlations between the ETKDG, cosine, and Gaussian fits and the underlying torsional histograms for both acyclic and ring patterns. The Pearson $r^2$ correlation is used since the magnitudes of the peak heights may vary, but the key features should be the relative intensities at each torsion angle.

Table 1 indicates the median $r^2$ for each fit across both acyclic and ring patterns. Note that since the cosine fits in this work allow the phase shift parameter to vary, the results are improved over the published ETKDG fits. Nevertheless, simple fits to Gaussian peaks perform noticeably better, with most fits yielding $r^2$ above 0.9. In part, this method enables better fitting to subpopulations as discussed above. Further refinement of torsion patterns may resolve such subpopulations, improving the accuracy of ETKDG-style fits.\(^{60}\)

Table 1: Median $r^2$ correlation between fit functions and underlying quantum torsions for both acyclic and ring patterns.

<table>
<thead>
<tr>
<th>Method</th>
<th>Median acyclic $r^2$</th>
<th>Median ring $r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETKDG</td>
<td>0.26</td>
<td>0.03</td>
</tr>
<tr>
<td>Cosine Fits</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>Gaussian Fits</td>
<td>0.91</td>
<td>0.93</td>
</tr>
</tbody>
</table>

For sampling torsion angles from the Gaussian fits, we have generated the cumulative sum across all angles 0-360°, normalized, and inverted, such that generating a uniform random number yields an appropriate torsion angle for a given torsion pattern. Such distributions can either be used for individual torsion driving, as demonstrated in the included script, Bayesian sampling,\(^{61,62}\) or as part of distance geometry coordinate generation methods such as ETKDG.
Figure 9: Histograms of correlation between (a, d) ETKDG, (b, e) cosine fits, or (c, f) Gaussian fits and derived torsional histograms for (a, b, c) acyclic and (d, e, f) ring torsional patterns.

Conclusions

Overall, comparing geometric RMSD between experimental small molecule crystal structure geometries and gas-phase GFN2 conformer ensembles indicate a high fidelity compared to the widely-used ETKDG methods. Particularly for molecules with few rotatable bonds, the CREST/GFN2 ensembles yielded smaller RMSD than ETKDG and comparable performance across larger molecules from the Crystallographic Open Database. We note somewhat worse performance across the Platinum set than ETKDG, particularly for molecules with more rotatable bonds. There is a notable difference in the radius of gyration between the Platinum and low-energy CREST/GFN2 compounds, particularly with increasing numbers of rotatable bonds, likely indicating some preference for extended conformations in the Platinum set due to intermolecular interactions of bound ligands, not present in an isolated gas-phase calculation.

Comparing individual torsion preferences between the experimental COD geometries
and the lowest-energy CREST/GFN2 conformers indicated good correlation (i.e., median $r^2$ above 0.6 for acyclic torsions and above 0.8 for ring torsions), with most cases of poor correlation deriving from few examples in the COD set.

The advantage of gas-phase conformer sampling, particularly with the semiempirical GFN2 method, is that many compounds can be analyzed beyond those available in crystallographic databases. Consequently, distributions of torsion preferences were analyzed across over 3 million compounds compiled from multiple small molecule sets.

Though some differences in torsional preferences were found, most cases of poor correlation between experimental COD geometries and lowest-energy CREST/GFN2 conformers occur with a lack of sufficient experimental crystallographic data. Some deviations occur from new subpopulations not present in the experimental databases, since not every compound can be crystallized.\textsuperscript{13}

Using density functional methods to optimize compounds in such subpopulations, we find torsion angle deviations $\sim$0-5°, suggesting CREST/GFN2 to be sufficiently accurate for generating overall torsion angle distributions, even if some errors remain.

Finally, fits to Gaussian peaks were used to generate sampling distributions and prove to be more accurate than Fourier analysis used in ETKDG, likely due to the presence of such subpopulations. Further refinement of torsion patterns will be useful,\textsuperscript{60} as well as analysis of correlated torsional preferences for both acyclic and ring torsions.\textsuperscript{36,62}

This work has demonstrated the ability of CREST/GFN2 to provide accurate and reliable torsional preferences that could provide a basis for a quantum-based ETKDG method, QTDG. Such a method could provide an alternative which does not rely on experimental crystal structure elucidation, as well as a method designed particularly for conformer sampling for gas-phase applications rather than targeting crystal structure or bound ligand geometries which include intermolecular interactions. Finally, QTDG could be used as part of automated refinement of torsion patterns,\textsuperscript{60} since additional conformer sampling can be performed to refine patterns with sparse representation in experimental databases, including
charged, novel, or hard to crystallize species.\textsuperscript{13}

**Acknowledgement**

We acknowledge the National Science Foundation (CHE-1800435 and CHE-2102474) for support, the University of Pittsburgh Center for Research Computing through the computational resources provided, as well as resources provided by the Open Science Grid,\textsuperscript{63,64} which is supported by the National Science Foundation award 2030508, and the U.S. Department of Energy’s Office of Science. We also thank Sereina Riniker and Greg Landrum for helpful discussions.

**Supporting Information Available**

Additional plots and data tables, including histograms of the rotatable bonds and conformers in the COD set, comparisons of RMSD and radius of gyration for the COD and Platinum sets, and histograms of torsional angle deviations for GFN2-optimized and \textit{omega}B97X-D3/def2-SVP optimized geometries. All raw data, Python notebooks, and torsion pattern figures can be found at [https://github.com/hutchisonlab/quantum-torsions](https://github.com/hutchisonlab/quantum-torsions).

**References**


(3) Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.;


(38) [https://smarts.plus/](https://smarts.plus/).


(43) Johansson, M. P.; Olsen, J. Torsional Barriers and Equilibrium Angle of Biphenyl:


(51) Foloppe, N.; Chen, I.-J. Energy windows for computed compound conformers: covering
artefacts or truly large reorganization energies? *Future Medicinal Chemistry* 2019, 11, 97–118.


(55) Tong, J.; Zhao, S. Large-Scale Analysis of Bioactive Ligand Conformational Strain Energy by Ab Initio Calculation. *Journal of Chemical Information and Modeling* 2021, 61, 1180–1192.


(59) Avogadro2 version 1.97. [https://two.avogadro.cc/](https://two.avogadro.cc/).


Graphical TOC Entry

Torsional Preferences: From Crystal Structures to Exhaustive Conformer Sampling

Analyzing over 3 million compounds

COD Experimental Torsions

GFN2 Torsions