Embracing chemical and structural diversity with UCONGA: a universal conformer generation and analysis program

Nathaniel R. Gunby, Sarah L. Masters, Deborah L. Crittenden *

Department of Chemistry, University of Canterbury, Private Bag 4800, Christchurch 8041, New Zealand

KEYWORDS: Software, molecular conformation, computer simulation, drug design, structural chemistry, inorganic chemistry

ABSTRACT

Molecular properties depend on molecular structure, so the first step in any computational chemistry investigation is to generate all thermally accessible conformers. Typically it is necessary to make a trade-off between the number of conformers to be explored and the accuracy of the method used to calculate their energies. *Ab initio* potential energy surface scans can, in principle, be applied to any molecule, but their conformational cost scales poorly with both molecular size and dimensionality of the search space. Specialized conformer generation techniques rely on parameterized force fields and may also use knowledge-based rules for generating conformers, and are typically only available for drug-like organic molecules. Neither approach is well-suited to generating or identifying chemically sensible conformers for larger non-organic molecules. The Universal CONformer Generation and Analysis (UCONGA) program package fills this niche. It requires no parameters other than built-in atomic van der Waals radii to generate comprehensive ensembles of sterically-allowed conformers, for molecules of arbitrary composition and connectivity. Analysis scripts are provided to identify representative structures from clusters of similar conformers, which may be further refined by subsequent geometry optimization. This approach is particularly useful for molecules not described by parameterized force fields, as it minimizes the number of computationally intensive *ab initio* calculations required to characterize the conformer ensemble. We anticipate that UCONGA will be particularly useful for computational and structural chemists studying flexible non-drug-like molecules.

INTRODUCTION

Generating an ensemble of low-energy conformers is the first step in most computational chemistry studies. For example, in computational drug design, a collection of possible ligand conformers must be generated before attempting *in silico* docking to proteins.[1] A range of programs have been developed for this purpose, including ALFA,[2] Balloon,[3] Confab,[4] ConfGen,[5] and OMEGA.[6]

However, these methods lack generality, because they rely on force fields that are only parameterized to describe atoms and bonds within common organic and biological molecules. This limits the range of chemical space that can be explored in designing new drugs,[3, 7] excluding known drugs that contain 'uncommon' elements such as silicon,[8-11] arsenic[12] and mercury[13]. It is a more significant problem for inorganic and organometallic chemistry,[14,15] where 'uncommon' elements and 'unusual' bonding patterns are, in fact, common.

The Universal CONformer Generation and Analysis (UCONGA) program package is designed to meet this need for a fully general conformer generation technique. UCONGA does not rely on forcefields to find minimum energy conformers, but instead generates comprehensive sets of sterically-allowed conformers. Using only fixed literature values of van der Waals radii,[16] conformer ensembles can be generated for any molecule, including molecules that contain atoms and structural features not well-described by common forcefields. A particular feature is the ability to generate ring conformers directly, without relying on a library of known structures. The UCONGA package also includes clustering algorithms to analyze conformer ensembles and identify structurally similar conformers.

In this paper, we describe the algorithms implemented within UCONGA, and characterize its performance in terms of; ability to generate experimentally-observed conformers, computational efficiency, and ability to extract representative conformers from sterically-allowed conformer ensembles.

ALGORITHM AND IMPLEMENTATION

Overview

There are four main steps to the UCONGA process: pre-analysis of the molecular structure, generation of trial conformers, screening of trial conformers, and analysis of the screened conformer ensemble. The overall procedure is illustrated in Figure 1, and the algorithmic details of each step are described below.

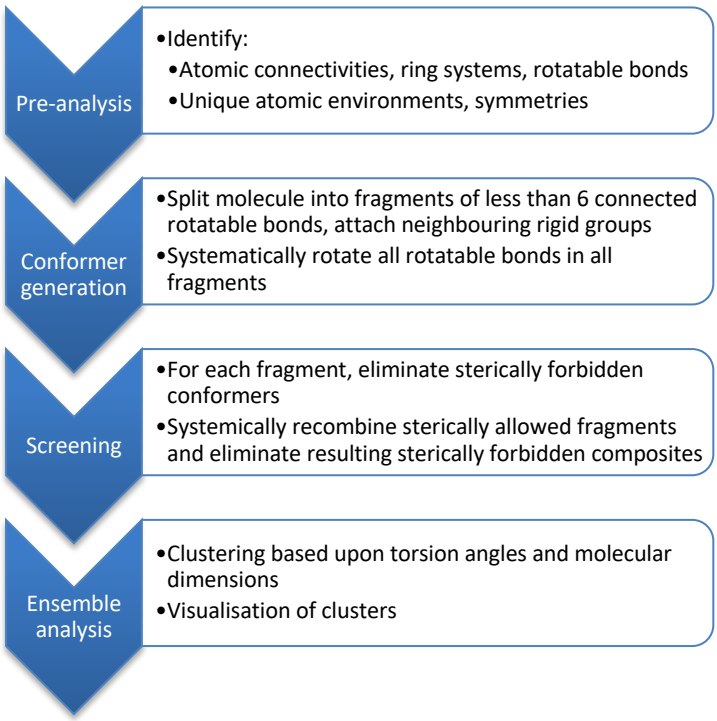
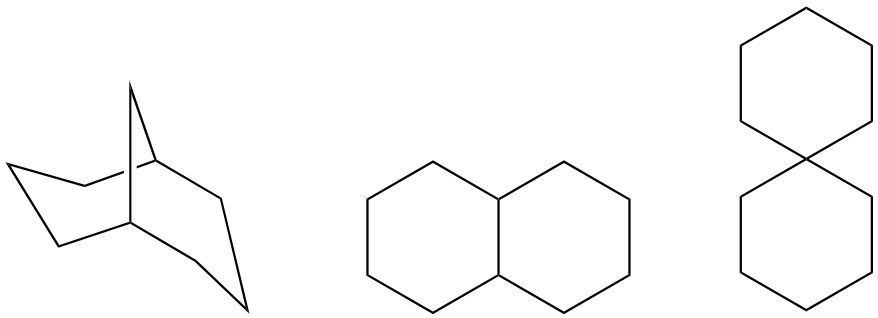


Figure 1. Blue arrows indicate key stages in the UCONGA conformer generation process, broken down into bullet-pointed steps.

Pre-analysis

Pre-analysis of each molecule has three aims: finding ring systems, identifying rotatable bonds, and determining nuclear permutational symmetry *i.e.* the ways in which atoms can have their labels interchanged without changing the identity of the molecule.

Rings are identified by analyzing atomic connectivities. During subsequent ring conformer generation, bridged and fused rings (Figure 2a and 2b) are treated as a single unit because their conformational changes are coupled. Spiro rings (Figure 2c), however, are treated independently as their conformational changes are not necessarily coupled.



(a) (b) (c)

Figure 2. Examples of (a) bridged, (b) fused and (c) spiro ring systems

Rotatable bonds are defined as single bonds that are not part of a ring system and have at least one non-hydrogen substituent non-linearly attached to the end of each bond.

Nuclear permutational symmetry is identified using the modified Morgan algorithm[17] to find atoms in identical chemical environments. The basic Morgan algorithm finds the chemical environments of atoms by assigning heavy valence identifiers to each atom where all atoms except hydrogen are considered heavy. Initially, the heavy valence of each atom is assigned as

the number of bonded non-hydrogen atoms. These identifiers are iteratively updated by adding all neighboring values at each step until the number of different values stays constant. Each atom with a different identifier at the end of this process is in a different chemical environment. The modified Morgan algorithm also checks that R/S or E/Z stereocentres are not being treated identically and, in this work, it has been further modified to account for parastereocenters, atoms which can generate stereoisomers without being traditional stereocenters e.g. the tertiary carbon atoms in 1,4-dimethylcyclohexane (Figure 3).[18]

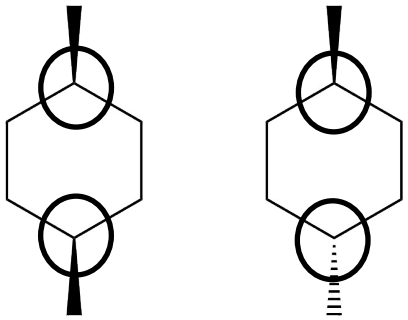


Figure 3. Despite possessing no traditional stereocenters, 1,4-dimethylcyclohexane has two stereoisomers. This is because the tertiary carbon atoms (circled) are parastereocenters.

Trial conformer generation

Maintaining universality while efficiently generating trial conformers is a challenge. While many conformer ensemble generation methods rely on a rules-driven approach using a list of preferred torsion angles, UCONGA cannot use this approach, as it will not find conformers for molecules containing high levels of steric crowding or unusual bonding patterns. UCONGA instead systematically rotates all rotatable bonds, as identified during pre-analysis, in a stepwise fashion.

However, the number of trial conformers generated scales exponentially with the number of rotatable bonds, and so this 'brute-force' procedure becomes computationally intractable for molecules that contain more than 5 rotatable bonds.

For molecules with 6 or more rotatable bonds, a divide-and-conquer fragmentation algorithm is used to circumvent this conformational explosion problem. It is based upon the assumption that conformational changes in distant regions of an extended molecule are uncoupled and so can be considered separately. This means that all conformers with a sterically forbidden motif anywhere within the molecule are discarded in one step, rather than this motif being incorporated into a range of related conformers which must then be screened and discarded individually.

The divide-and-conquer algorithm first divides molecules into fragments with 5 or fewer rotatable bonds, by fragmenting at rigid linker junctions, and attaching each rigid linker unit to all neighboring fragments. This process is illustrated in Figure 4. Fragments with more than five rotatable bonds are subdivided as evenly as possible. The torsions of groups of rotatable bonds that are *ortho* in an aromatic ring or *cis* across are double bond are tightly coupled, so such fragments are merged if that does not put them over the size limit of 5 rotatable bonds.

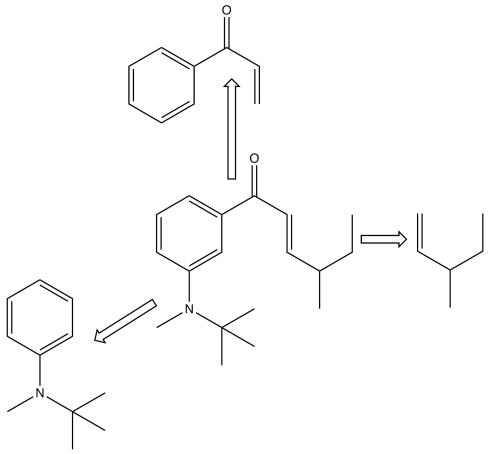


Figure 4. A molecule and the fragments produced from it using the fragmentation algorithm described in the text. In this particular case, each fragment contains a pair of adjacent rotatable bonds but, in general, each fragment may contain up to five adjacent rotatable bonds.

Conformer ensembles for each fragment, or the whole molecule if it has 5 or fewer rotatable bonds, are generated by performing systematic bond rotations. A multi-step process is used to avoid locating multiple conformations in a given basin on the potential energy surface. The first scan is performed with a relatively large step size, and then fine-grained searches are performed in areas of the potential energy surface where no sterically-allowed conformer has yet been found. The number of trial conformations is further reduced by decreasing the maximum torsion angle for rotation about symmetrical bonds, which include symmetric rotors and bonds that are equivalent under nuclear permutational symmetry, as illustrated in Figures 5 and 6.

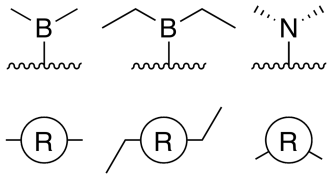


Figure 5. Left: a dimethylboryl group is a symmetric rotor of order 2. Middle: a diethylboryl group is not a symmetric rotor because the ethyl groups connected to the boron atom contain rotatable bonds. Right: a dimethylamino group is not a symmetric rotor because the methyl groups are not equidistant *i.e.* they do not evenly divide the circle in a Newman projection.

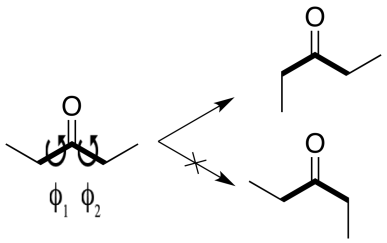


Figure 6. Left: An allowable conformer for pentan-3-one with symmetry-equivalent rotatable bonds in marked in bold. Torsion angles about these bonds are defined as angle made by the terminal methyl group to the plane defined by the rotatable bond and adjacent carbonyl bond. The two different torsion angles are labelled ϕ_1 and ϕ_2, as indicated. Right top: another allowable conformer for pentan-3-one, with $\phi_1 > \phi_2$. Right bottom: a forbidden conformer identical to the right top conformer but with $\phi_1 < \phi_2$.

A symmetric rotor is formed by an atom attached to terminal non-rotatable groups that are equidistant from each other and in the same symmetry class, as illustrated in Figure 5. In these cases, the maximum torsion angle around the bond to the symmetric rotor is divided by the number of attached terminal groups.

Rotations around equivalent bonds lead to automorphically equivalent conformers, as shown on the right-hand side of Figure 6. A bond is defined as equivalent to another bond if the symmetry classes of the atoms involved are equivalent and the two bonds share a common atom. At no stage in the conformer generation process may the dihedral angle of the second equivalent bond exceed that of the first. This ensures that all conformers generated are distinct.

After all fragment conformers have been generated and screened for steric clashes, they are systematically recombined to form molecular conformers, and the ring conformers of each are varied. Because the rotations of bonds in rings are concerted, rotating them in a straightforward fashion is impractical. Instead, UCONGA generates ring conformers using the flip-of-fragments method[19] and a concerted flip-of-fragments process referred to as ring-flipping.

The flip-of-fragments method is a two-step process. In the first step, a two-atom section of the ring with all substituents is reflected through the plane defined by its junction with the rest of the ring. In the second step, each atom in the ring that moved in the first step and all substituents are reflected in the plane of its in-ring bonds. The first step generates all observed conformations for non-macrocyclic rings, and the second step corrects for the inversions of stereochemistry in the first step.

In addition, the ring-flipped versions of all of these conformers are generated by reflecting all bonds around the ring and then correcting the stereochemistry of all substituents. Ring-flipping

can only meaningfully be applied to rings that contain a molecular reflection plane e.g. unsaturated six-membered rings but not unsaturated 5-membered rings.

Because bond-by-bond ring-flipping can generate a large number of conformers, UCONGA provides the option to only ring-flip the original conformer or to keep rings entirely fixed in their original conformation.

Screening criterion

During conformer generation, trial conformations are screened for steric crowding, based upon the scaled sum of their van der Waals radii.[19] If two atoms are closer than the scaled sum of their van der Waals radii, the conformer is rejected as sterically forbidden, otherwise it is accepted. Scaling is required to account for interactions that enable or force atoms into closer proximity with one another than otherwise expected e.g. hydrogen bonding or competing steric crowding / crystal packing forces. A universal scaling factor of 0.9 accounts for hydrogen bonding interactions, but must also be adjusted according to the degree of steric crowding within a molecule. Therefore, the average heavy valence of the molecule is calculated as the average number of bonds per non-terminal heavy atom. As the average heavy valence value increases from 2 (linear alkane) to 4 (infinite diamond), the scaling factor decreases linearly from 0.9 to 0.7.

Analysis

The UCONGA package contains not only a conformer generation algorithm, but also a collection of conformer ensemble analysis capabilities. There are two parts to this analysis: clustering and visualization.

Clustering is performed using the *k*-means algorithm,[20] based on either bounding-box dimensions or trigonometrically-transformed torsion angle vectors. Bounding boxes are aligned along the uniquely-defined molecular axes of inertia with the corresponding distance metric:

$$d_{\mathrm{bb}} = \sqrt{x_1^2 + y_1^2 + z_1^2} \qquad \textbf{Equation 1}$$

where (x_1,y_1,z_1) are the coordinates of the furthest corner of the smallest bounding box originating at (0,0,0) and completely containing the nuclear coordinates of conformer 1.

Root-mean-squared distances between torsion vectors are calculated as:

$$d_{\mathrm{tors}} = \sqrt{\frac{1}{N_{\mathrm{tors}}} \sum_{i=1}^{N_{\mathrm{tors}}} \left(\left(\sin \varphi_{1,i} - \sin \varphi_{2,i} \right)^2 + \left(\cos \varphi_{1,i} - \cos \varphi_{2,i} \right)^2 \right)} \qquad \textbf{Equation 2}$$

where $\varphi_{1,i}$ is the i^{th} torsion angle of conformer 1.

The optimal number of clusters is determined using the Calinski-Harabasz criterion,[21] which is the ratio of the between-cluster variance to the within-cluster variance, corrected for the increase in variance explained with the number of clusters due to overfitting. It is therefore at a maximum for high-quality clustering where the clusters are tight and well-separated.

Torsionally-clustered conformers can be visualized using a parallel coordinates plot with torsion angles on the *x* axis and the values they adopt on the *y* axis. Conformers clustered by bounding-box dimensions can be visualized using a scatterplot of one bounding-box side length against another. Additionally, a matrix of root-mean-squared (RMS) distances between conformers' Cartesian coordinates can be calculated and visualized using a heatmap where the similarity between conformers is represented by the color of elements within the square matrix. The RMS distance is calculated as:

$$d_{\mathrm{RMS}} = \sqrt{\frac{1}{N_{\mathrm{atom}}}\sum_{i=1}^{N_{\mathrm{atom}}}(x_{1,i} - x_{2,i})^2 + (y_{1,i} - y_{2,i})^2 + (z_{1,i} - z_{2,i})^2} \qquad \textbf{Equation 3}$$

where $x_{1,i}$ is the x coordinate for the i^{th} atom of conformer 1.

To ensure minimal RMS distances between conformers are calculated, the molecules are aligned using the Schonemann[22] algorithm if enantiomeric conformers are treated as identical, or the Kabsch[23] algorithm otherwise.

Implementation and Usage

The implementation of UCONGA is designed to make the program platform-independent, portable and easy to use. It is written in Python to avoid the need for compilation or installation. The mandatory external dependencies are NumPy[24] and SciPy[25], high-performance linear algebra and scientific algorithm libraries common among scientific python users. The visualization capabilities of the analysis module rely on MatPlotLib,[26] which builds on NumPy, but this is an optional extra. Three command-line programs are provided;

'UCONGA_generate': generates the trial conformer ensemble and performs screening to select sterically-allowed conformers

'UCONGA_analyse': performs clustering, RMSD calculation, and visualization

'UCONGA_align': aligns multiple conformers for viewing with a molecule viewer

All programs read the cml file format, which can be generated by the free software Avogadro and OpenBabel, and all programs can write cml, xyz, and the geometry portion of a GAMESS, Gaussian, or NWchem input file. This flexibility in output format generation allows for easier visualization or further optimization of the generated or aligned conformers. The source code and instructions for usage are freely available on github.

METHODS

Benchmarking data set

The UCONGA benchmarking data set contains three groups of experimentally-determined structures. The first is the subset of the Astex dataset[27] containing 8 or fewer rotatable bonds. The Astex dataset contains high-quality crystal structures of ligands bound to proteins, and is commonly used for benchmarking conformer-ensemble generation methods. The second is a collection of structures determined by gas-phase electron diffraction (GED), including the structures of dimethylbis(trimethylsilylketyl)silane,[28] tris(*tert*-butyl) sulfurtriimide,[29] bis(tert-butyl)trichlorosilylphosphane,[30] 1,1,2,2-tetrakis(trimethylsilyl)disilane,[31] and 1,1,2-tris(trimethylsilyl)disilane.[32] The third is a collection of small-molecule organometallic crystal structures from the Cambridge Crystallographic Data Centre (CCDC) database. This subgroup contains several substituted metallocenes, with rotatable bonds defined between the metal atom and dummy atoms placed at the centre of each ring. Only two molecules with 9 rotatable bonds could be identified from the CCDC, due to the over-representation of small, rigid molecules within this database.

Between them, this dataset includes a total of 180 molecules of different types and in different chemical environments; organic and inorganic, gas-phase and crystalline, symmetric and asymmetric, sterically crowded and uncrowded. The data set composition is summarized in Table 1, and the coordinates of all reference structures are provided as Supporting Information.

Table 1. Composition of the reference conformer data set using in benchmarking UCONGA, broken down by number of rotatable bonds and data source (GED = gas phase molecular structures from Gas Electron Diffraction data, Astex = protein-bound ligand structures from protein crystallography data, CCDC = small molecule crystal structures from the Cambridge Crystallographic Data Centre database).

	Number of molecules from			
Number of rotatable bonds	GED	Astex	CCDC	Total
1	0	10	0	10
2	0	13	7	20
3	2	13	17	32
4	2	16	32	50
5	1	12	15	28
6	0	8	11	19
7	0	5	7	12
8	0	6	1	7
9	0	0	2	2

Despite this diversity in molecular structure and complexity, there is only a weak relationship between molecular mass and number of rotatable bonds across the data set (Figure 7). This is because most molecules contain a similar number of rigid or semi-rigid rings, irrespective of the number of rotatable bonds.

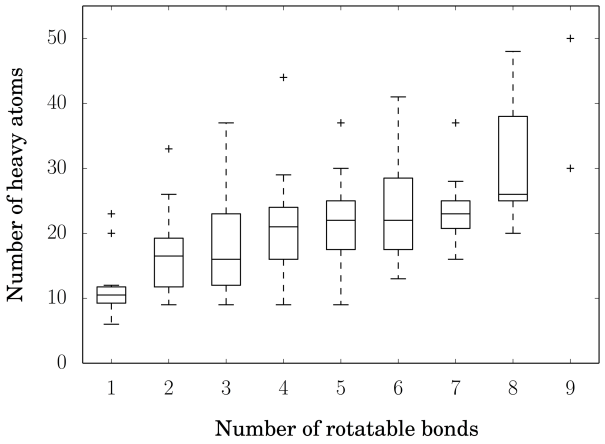


Figure 7. A boxplot summary of the number of heavy (non-hydrogen) atoms per molecule in the benchmarking data set, as a function of number of rotatable bonds. Each box spans the upper to lower interquartile range and contains a line indicating the median. Whiskers extend to the limits of the data, excluding outliers more than 2.5 times the interquartile range from the median. These are represented by crosses.

Benchmarking procedure

Torsion angles for all rotatable bonds were initialized to zero, except for those in rings, which were left at their reference values. Two step sizes were employed during generation of sterically-allowed rotamers; an initial coarse step of 60°, and a second fine step of 30°. The brute-force approach was used to generate trial conformers for molecules with up to 5 rotatable bonds, after which the divide-and-conquer algorithm was used. Alternative ring conformers were generated using only the ring-flip procedure. No subsequent geometry optimization was performed.

Additionally, the divide-and-conquer algorithm was benchmarked against the brute-force algorithm for fragmentable molecules with 4 or 5 rotatable bonds *i.e.* those whose flexible regions are partially or completely separated by rigid linker units.

RESULTS AND DISCUSSION

Conformers of best fit to experimental reference data

UCONGA's ability to generate experimentally-accessible conformers is assessed by comparing all conformers in each sterically-allowed ensemble to an experimental reference structure. The conformer with the lowest root-mean-square deviation (RMSD) in "heavy atom" (non-hydrogen) positions is taken as the conformer of best fit. Aggregated statistical data on all conformers of best fit are illustrated in Figure 8, and the raw data is available as Supporting Information.

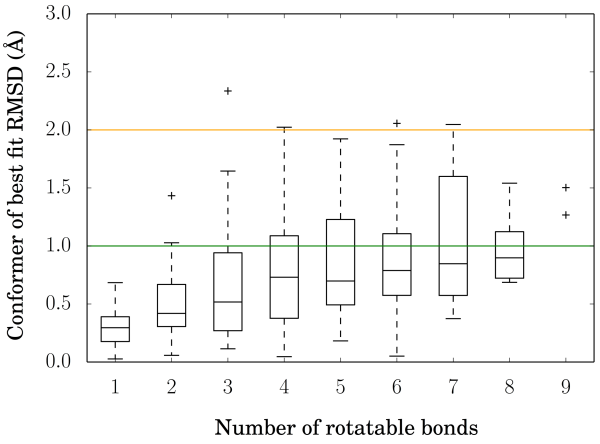


Figure 8. A boxplot summary of the RMSD alignment data for all conformers of best fit to the reference structures, grouped by the number of rotatable bonds. Each box spans the upper to lower interquartile range and contains a line indicating the median. Whiskers extend to the limits of the data, excluding outliers more than 2.5 times the interquartile range from the median, which are represented by crosses.

In line with previous studies involving automated conformer generation processes,[3,10] an RMSD of 1 Å to the reference is considered a good fit, 1-2 Å acceptable and beyond that, poor.

By these cutoffs, 74% (140/180) of the generated ensembles contain at least one conformer that is a good fit to the reference, and a further 22% (46/180) contain an acceptably close conformer.

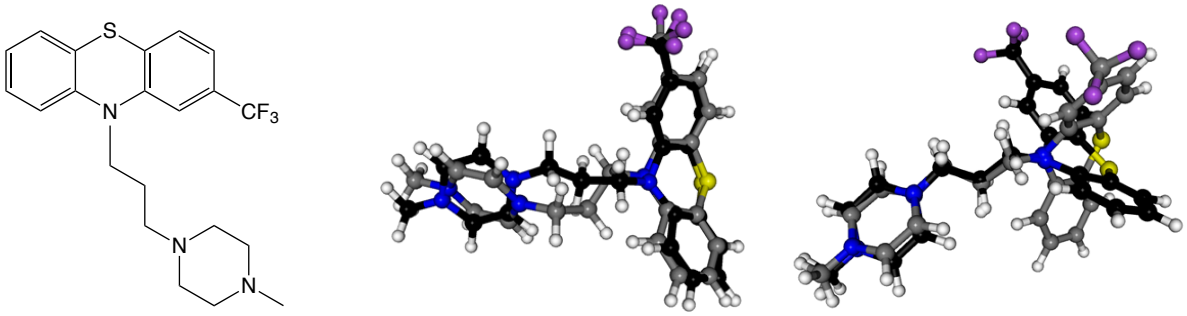
There were only 4 cases in which UCONGA failed to generate any conformers within 2 Å RMSD of the reference. The dimensionality of the search space is apparently not a factor in determining whether this will occur, as each case has a different number of rotatable bonds. Instead, the common factor that links these molecules is that their torsionally-discretized reference conformers – with all torsion angles rounded to their nearest 30 degree increment – would all be rejected as sterically forbidden according to UCONGA's selection criteria.

In most cases, the reference conformers themselves would pass the steric screening test, but their torsionally-discretized analogues do not. This situation arises when a series of subtle adjustments in torsion angles are required to alleviate steric strain within a molecule, typically within sterically crowded molecules. Another case involves a molecule with disparate sterically crowded and sparse regions, resulting in a moderate steric scaling factor that over-penalizes the sterically crowded region, leading to rejection of both the reference conformer and its torsionally-discretized counterpart.

These observations suggest that UCONGA be used with a degree of caution, particularly for completely or partially sterically crowded molecules. Fortunately, sterically crowded molecules tend to have small sterically-allowed conformer ensembles that can be efficiently and effectively generated using the divide-and-conquer algorithm. Therefore, we recommend that the divide-and-conquer approach is always used for sterically crowded molecules, regardless of the number of rotatable bonds. This enables smaller step sizes to be used when generating trial conformers, to more comprehensively explore the torsional space for a given computational cost.

Finally, it should be noted that the RMSD values illustrated in Figure 8 are for conformers directly generated by UCONGA, and are likely to be upper limits to those that would be found after subsequent geometry optimization.

In particular, molecules with rigid, bulky terminal groups stand to improve substantially in RMSD fit to the reference following geometry optimization, as small torsional misalignments of these groups contribute disproportionately to overall RMSDs. For example, conformers of best fit for trifluoperazine according to Cartesian coordinate and dihedral RMSDs are illustrated in Figure 9.



(a) (b) (c)

Figure 9. (a) the chemical structure of trifluoperazine, (b) all-atom alignment to reference for conformer of best fit as identified by Cartesian coordinate RMSD, (c) rotatable-bonds-only alignment to reference for the conformer of best fit as identified by dihedral RMSD.

Computational efficiency

The ability to generate experimentally relevant conformers is the primary performance criterion for conformer generation algorithms, but computational efficiency is also important. While each conformer generation and screening step within UCONGA is fast, the computational run-time grows exponentially when using the brute-force algorithm (Figure 10, $n_{rot} < 6$). This makes brute-

force generation of trial conformers computationally prohibitive for molecules with 6 or more rotatable bonds.

Therefore, the divide and conquer algorithm is used for these molecules to avoid generating multiple trial conformers containing identical sterically-forbidden conformational motifs. Its effectiveness is apparent from the run-time data for $n_{rot} > 5$ presented in Figure 10.

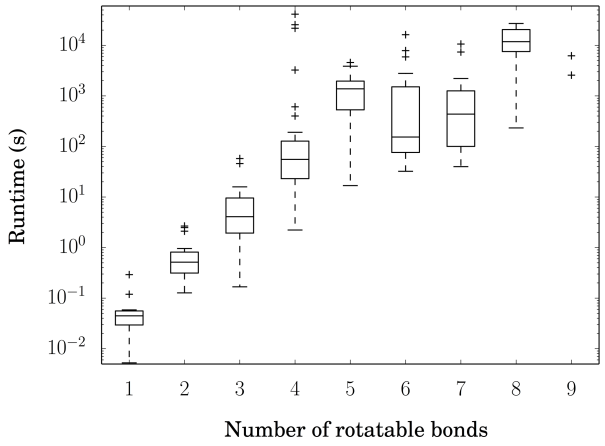
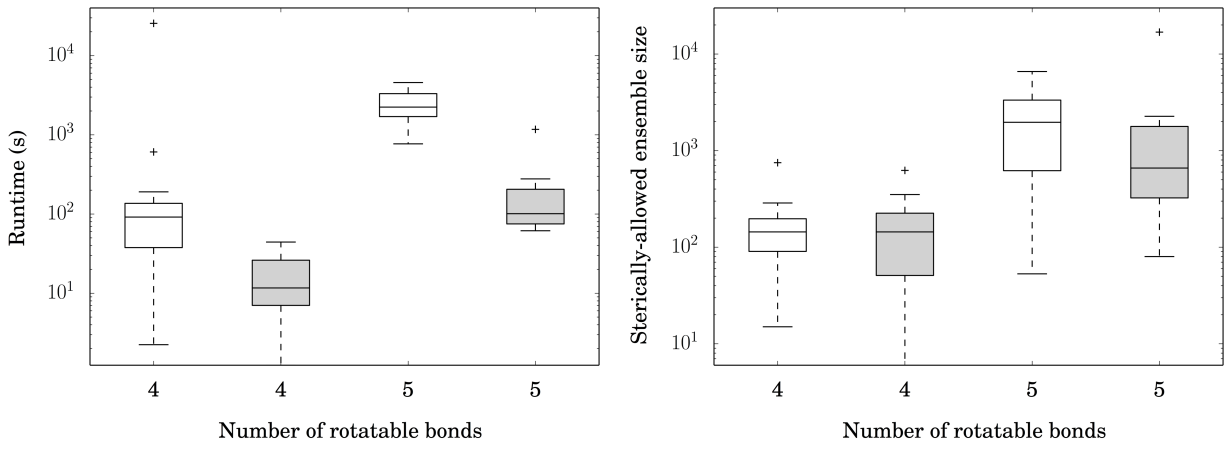


Figure 10. A boxplot summary of time taken to generate and screen all possible trial conformers for all molecules in the benchmarking data set, grouped by the number of rotatable bonds. Each box spans the upper to lower interquartile range and contains a line indicating the median. Whiskers extend to the limits of the data, excluding outliers more than 2.5 times the interquartile range from the median, which are represented by crosses. All calculations were carried out on an iMac-i3 (2010).

While the divide and conquer approach is clearly computationally advantageous, this must not come at the expense of restricting conformational diversity in the final, sterically-allowed conformer ensemble.

Conformer ensembles for 20 fragmentable molecules with 4 and 5 rotatable bonds were generated using both the brute-force and divide-and-conquer algorithms. Comparison of run times and conformer ensemble sizes between approaches are illustrated in Figure 11.



(a) (b)

Figure 11. Boxplot representations of (a) the time taken to generate and screen all possible trial conformers, and (b) the number of sterically-allowed conformers, for each fragmentable molecule with either 4 or 5 rotatable bonds. White boxes = brute force algorithm, grey boxes = divide-and-conquer algorithm. Each box spans the upper to lower interquartile range and contains a line indicating the median. Whiskers extend to the limits of the data, excluding outliers more than 2.5 times the interquartile range from the median, which are represented by crosses. All calculations were carried out on an iMac-i3 (2010).

From Figure 11(a), we observe that the divide-and-conquer algorithm reduces the run-time by just under an order of magnitude, on average, for the molecules with 4 rotatable bonds, and just over an order of magnitude, on average, for molecules with 5 rotatable bonds.

Importantly, though, Figure 11(b) shows that the sterically-screened ensembles are approximately the same size regardless of the approach taken to generate trial conformers. Therefore, it appears that divide-and-conquer method largely discards only sterically forbidden conformers *a priori*, as intended.

To double-check that no important information is lost during the divide-and-conquer process, conformers of best fit are compared between brute-force and divide-and-conquer generated conformer ensembles. Summary statistics are presented in Table 2.

Table 2. Mean, median, and maximum differences in RMSDs (ΔRMSD) for conformers of best fit identified within brute-force and divide-and-conquer generated conformer ensembles.

ΔRMSD (Å)	All	Excluding outlier
Mean	0.22	0.19
Median	0.11	0.10
Maximum	1.05	0.62

In most cases, the divide-and-conquer conformers of best fit are the same as or similar to those identified using the brute-force approach. The outlier is a substituted disaccharide, whose brute-force conformer of best fit (RMSD = 0.35 Å) is much closer to the reference than the one found using the divide-and-conquer algorithm (RMSD = 1.40 Å).

In this case, the fragmentation algorithm generates two substituted sugar units, and an unsubstituted disaccharide fragment. UCONGA detects that the disaccharide is symmetric, and so restricts the range of rotations about the bonds connecting the sugar units. However, the disaccharide is not symmetric within the context of the entire molecule, so the full range of bond rotations should be explored. Problem cases could be avoided by turning off symmetry constraints during the fragment conformer generation step.

The overall similarity between conformer ensemble sizes and conformers of best fit produced by the different approaches gives us confidence that the divide-and-conquer algorithm does not artificially restrict the number or range of sterically allowed conformers. Therefore, the plateau in number of sterically-allowed conformers for molecules with more than 6 rotatable bonds (Figure 12) is not an artefact of using the divide-and-conquer algorithm.

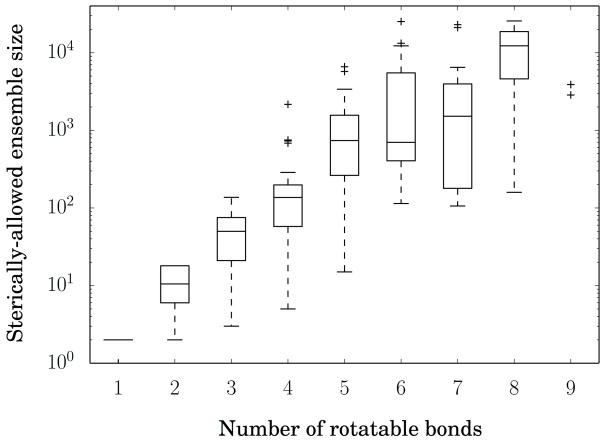


Figure 12. A boxplot representation of the number of sterically-allowed conformers for the benchmarking molecules, grouped by the number of rotatable bonds. Each box spans the upper to lower interquartile range and contains a line indicating the median. Whiskers extend to the limits of the data, excluding outliers more than 2.5 times the interquartile range from the median. These are represented by crosses.

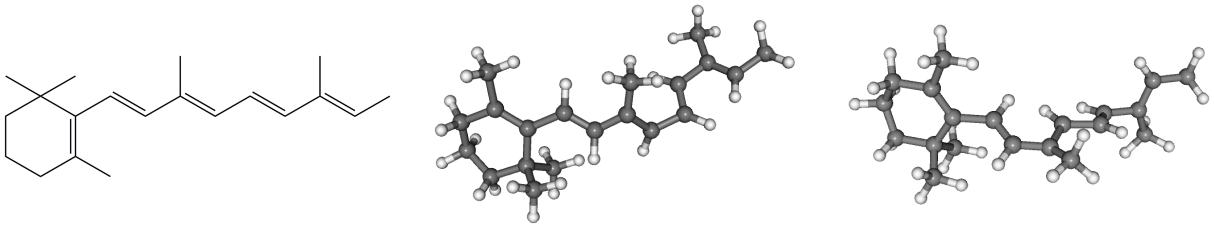
Instead, the most likely explanation is that the average degree of steric crowding is constant for molecules with 1-6 rotatable bonds, but additional steric crowding is introduced by interactions between distant sites for molecules with 7 or more rotatable bonds, restricting the growth of the sterically-allowed conformer ensembles.

Ensemble analysis

Bounding-box dimensions provide a simple but powerful metric of the overall size and shape of conformers. Unlike Cartesian coordinate RMSDs, they are not biased towards aligning large, rigid units. Similarly, torsion-space distances can be misleading, as small variations in bond angles can cause large changes in molecular shapes.

The utility of bounding-box based clustering is most clearly demonstrated by example - analyzing the conformer ensemble of deoxoretinal, whose chemical structure is depicted in Figure 13(a). Because we have prioritized universality, UCONGA does not account for

electronic effects, and thus does not penalize breaking conjugation in this molecule. Therefore, the resulting sterically-allowed conformer ensemble is quite large, consisting of 3240 conformers.



(a) (b) (c)

Figure 13. (a) the chemical structure of deoxoretinal, (b) representative central structure from cluster of folded conformers, (c) representative central structure from cluster of extended conformers.

Bounding-box clustering (Figure 14) reveals two distinct clusters. The more numerous, compact cluster contains conformers with extended side-chains, while the less populated, more dispersed cluster contains folded conformers. The major advantage of clustering is that it is easy to identify representative conformers for visualization and/or subsequent geometry optimization. The conformers closest to the center of each cluster are shown in Figure 13.

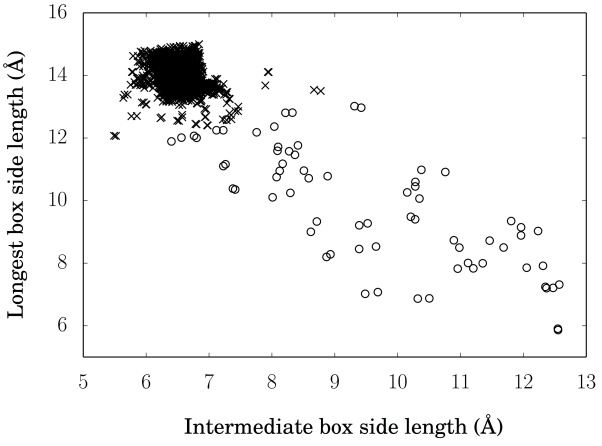


Figure 14. Bounding-box clusters of the deoxoretinal sterically-allowed conformer ensemble, visualized using the two longest box edge distances for each conformer. Different clusters are represented by different markers; o = cluster of folded conformers, x = cluster of extended conformers.

From Figure 14, it is clear that the conformers in the folded cluster would become sterically-forbidden if the molecule was planarized to enforce conjugation, while conformers in the extended cluster would be sterically allowed. Upon geometry optimization, the representative conformer from the extended cluster would readily converge to the fully extended experimentally-observed conformer, while the representative folded conformer would have to undergo large-scale conformational reorganization to reach the experimental reference structure.

This process of clustering and representative conformer selection avoids the computational cost associated with optimizing all sterically-allowed conformers to correct for breaking conjugation, or the need to apply additional constraints to detect and enforce planarity in extended conjugated π systems. We consider that this is an acceptable trade-off, as it enables UCONGA to be used as a truly universal 'black-box' conformer generation tool.

CONCLUSION

The newly developed UCONGA method is capable of generating conformer ensembles that almost always contain conformers within 2 Å RMSD from experimental reference structures, for molecules as diverse as amino acids, metallocenes and sterically hindered asymmetrical disilanes. The divide-and-conquer algorithm implemented within UCONGA is a powerful technique that prevents exponential scaling of run-time with number of rotatable bonds, and enables fine-grained exploration of torsional space in sterically-crowded molecules. The UCONGA method is, to the best of our knowledge, the first fully general conformer generation tool, and we anticipate that it will find widespread application in a range of fields, including non-conventional drug design,[8-11] gas-phase structural chemistry[14] and supramolecular chemistry.[15]

ASSOCIATED CONTENT

Supporting Information. Reference conformers of all molecules in our benchmarking data set (all_structures.xyz) and a complete list of benchmarking results (benchmarking.xlsx).

AUTHOR INFORMATION

Corresponding Author

Deborah Crittenden, Department of Chemistry, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand. E-mail: deborah.crittenden@canterbury.ac.nz

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The UCONGA code is available from http://github.com/NRGunby/UCONGA

ACKNOWLEDGMENTS

N.R.G thanks the University of Canterbury (UC) for a Canterbury Doctoral Scholarship and the Lord Rutherford Memorial Research Fellowship (awarded by UC on behalf of the late Arthur and Agnes Marion Sims). The authors acknowledge the contribution of NeSI high-performance computing facilities to the results of this research. NZ's national facilities are provided by the NZ eScience Infrastructure and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme.

REFERENCES

(1) Klett, J.; Cortés-Cabrera, Á.; Gil-Redondo, R.; Gago, F.; Morreale, A. ALFA: Automatic Ligand Flexibility Assignment. *J. Chem. Inf. Model.* **2014**, *54*, 314–323.

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

(3) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic Generation of Diverse Low-Energy Conformers. *J. Cheminform.* **2011**, *3*, 8.

(4) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534–546.

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

(6) Schwab, C. H. Conformations and 3D Pharmacophore Searching. *Drug Discov. Today Technol.* **2010**, *7*, e245–e253.

(7) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146–1158.

(8) Donadel, O. J.; Martín, T.; Martín, V. S.; Villar, J.; Padrón, J. M. The Tert-Butyl Dimethyl Silyl Group as an Enhancer of Drug Cytotoxicity against Human Tumor Cells. *Bioorganic Med. Chem. Lett.* **2005**, *15*, 3536–3539.

(9) Franz, A. K.; Wilson, S. O. Organosilicon Molecules with Medicinal Applications. *J. Med. Chem.* **2013**, *56*, 388–405.

(10) Bikzhanova, G.; Toulokhonova, I.; Gately, S.; West, R. Novel Silicon-Containing Drugs Derived from the Indomethacin Scaffold: Synthesis, Characterization and Evaluation of Biological Activity. *Silicon Chem.* **2007**, *3*, 209–217.

(11) Van Hattum, A. H.; Pinedo, H. M.; Schlüper, H. M. M.; Hausheer, F. H.; Boven, E. New Highly Lipophilic Camptothecin BNP1350 Is an Effective Drug in Experimental Human Cancer. *Int. J. Cancer* **2000**, *88*, 260–266.

(12) Lloyd, N. C.; Morgan, H. W.; Nicholson, B. K.; Ronimus, R. S. The Composition of Ehrlich's Salvarsan: Resolution of a Century-Old Debate. *Angew. Chem. Int. Ed.* **2005**, *44*, 941–944.

(13) Melnick, J. G.; Yurkerwich, K.; Buccella, D.; Sattler, W.; Parkin, G. Molecular Structures of Thimerosal (Merthiolate) and Other Arylthiolate Mercury Alkyl Compounds. *Inorg. Chem.* **2008**, *47*, 6421–6426.

(14) Mitzel, N. W.; Rankin, D. W. H. SARACEN – Molecular Structures from Theory and Experiment: The Best of Both Worlds. *Dalton Trans.* **2003**, 3650–3662.

(15) Yang, D.; Liang, Y.; Ma, P.; Li, S.; Wang, J.; Niu, J. Ligand-Directed Conformation of Inorganic–Organic Molecular Capsule and Cage. *Inorg. Chem.* **2014**, *53* (6), 3048–3053.

(16) Mantina, M.; Chamberlin, A. C.; Valero, R.; Cramer, C. J.; Truhlar, D. G. Consistent van der Waals Radii for the Whole Main Group. *J. Phys. Chem. A* **2009**, *113*, 5806–5812.

(17) Shelley, C. A.; Munk, M. E. Computer Perception of Topological Symmetry. *J. Chem. Inf. Model.* **1977**, *17*, 110–113.

(18) Razinger, M.; Balasubramanian, K.; Perdih, M.; Munk, M. E. Stereoisomer Generation in Computer-Enhanced Structure Elucidation. *J. Chem. Inf. Model.* **1993**, *33*, 812–825.

(19) Mekenyan, O.; Pavlov, T.; Grancharov, V.; Todorov, M.; Schmieder, P.; Veith, G. 2D-3D Migration of Large Chemical Inventories with Conformational Multiplication. Application of the Genetic Algorithm. *J. Chem. Inf. Model.* **2005**, *45*, 283–292.

(20) Lloyd, S. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory.* **1982**, *28*, 129–137.

(21) Caliński, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat.* **1974**, *3*, 1–27.

(22) Schönemann, P. H. A Generalized Solution Of The Orthogonal Procrustes Problem. *Psychometrika* **1966**, *31*, 1–10.

(23) Kabsch, W. A. Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. Sect. A* **1976**, *32*, 922–923; Kabsch, W. A. Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. Sect. A* **1978**, *34*, 827–828.

(24) van der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.

(25) Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python http://www.scipy.org/ (accessed Mar 26, 2015).

(26) Hunter, J. D. Matplotlib: A 2D Graphic Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.

(27) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A New Test Set for Validating Predictions of Protein-Ligand Interaction. *Proteins Struct. Funct. Genet.* **2002**, *49*, 457–471.

(28) Borisenko, K. B.; Yezhov, R. N.; Gruener, S. V.; Robertson, H. E.; Rankin, D. W. H. Gas-Phase Molecular Structures of Substituted 1,3-Bisketenes: A Challenge for Theory and Experiment. *Inorg. Chim. Acta* **2008**, *361*, 467–472.

(29) Hinchley, S. L.; Trickey, P.; Robertson, H. E.; Smart, B. A.; Rankin, D. W. H.; Leusser, D.; Walfort, B.; Stalke, D.; Bühl, M.; Obrey, S. J. Bis(tert-Butyl)sulfurdiimide, $S(NBu^t)_2$, and Tris(tert-Butyl)sulfurtriimide, $S(NBu_t)^3$: Structures by Gas Electron Diffraction, X-Ray Crystallography and Ab Initio Calculations. *J. Chem. Soc., Dalton Trans.* **2002**, 4607-4616.

(30) du Mont, W. W.; Müller, L.; Martens, R.; Papathomas, P. M.; Smart, B. A.; Robertson, H. E.; Rankin, D. W. H. Intermediates and Products of the Hexachlorodisilane Cleavage of Group 14 Element Phosphanes and Amines – Molecular Structure of Di-Tert-Butyl (trichlorosilyl) Phosphane in the Gas Phase Determined by Electron Diffraction and Ab Initio Calculations. *Eur. J. Inorg. Chem.* **1999**, *1999*, 1381–1392.

(31) Tekuutz, G.; Hassler, K. Molecular Inorganic Silicon Chemistry: Conformational Properties of 1,1,2,2-Tetrakis(trimethylsilyl)disilanes and of Tetrakis(trimethylsilyl) Diphosphine: A Comparative Vibrational

Spectroscopicand Ab Initio Study. In *Organosilicon Chemistry VI: From Molecules to Materials*; 2008; pp 368–372.

(32) Hinchley, S. L.; Smart, B. A.; Morrison, C. A.; Robertson, H. E.; Rankin, D. W. H.; Zink, R.; Hassler, K. 1, 1, 2-Tri-Tert-Butyldisilane, $Bu^t_2HSiSiH_2Bu^t$: Vibrational Spectra and Molecular Structure in the Gas Phase by Electron Diffraction and Ab Initio Calculations. *J. Chem. Soc., Dalton Trans*. **1999**, 2303–2310.