Title

Exhaustive Conformational Sampling of Complex Fused Ring Macrocycles Using Inverse Kinematics

Journal

Journal of Chemical Theory and Computation, 12(9)

ISSN

1549-9618

Authors

Coutsias, Evangelos A
Lexa, Katrina W
Wester, Michael J
et al.

Publication Date

2016-09-13

DOI

10.1021/acs.jctc.6b00250

Peer reviewed

HHS Public Access

Author manuscript

J Chem Theory Comput. Author manuscript; available in PMC 2017 September 13.

Published in final edited form as:

J Chem Theory Comput. 2016 September 13; 12(9): 4674–4687. doi:10.1021/acs.jctc.6b00250.

Exhaustive Conformational Sampling of Complex Fused Ring Macrocycles Using Inverse Kinematics

Evangelos A. Coutsias[*,†], **Katrina W. Lexa**[*,‖], **Michael J. Wester**[¶], **Sara N. Pollock**[§], and **Matthew P. Jacobson**[‖]

[†]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York 11794, United States

[¶]New Mexico Center for Spatiotemporal Modeling of Cell Signaling, University of New Mexico, Albuquerque, New Mexico 87131, United States

[§]Department of Mathematics, Texas A&M University, College Station, Texas 77843, United States

[‖]Department of Pharmaceutical Chemistry, University of California in San Francisco, San Francisco, California 94107, United States

Abstract

Natural product and synthetic macrocycles are chemically and topologically diverse. An efficient, accurate, and general method for generating macrocycle conformations would enable structure-based design of macrocycle drugs or host–guest complexes. Computational sampling also provides insight into transiently populated states, complementing crystallographic and NMR data. Here, we report a new algorithm, BRIKARD, that addresses this challenge through computational algebraic geometry and inverse kinematics together with local energy minimization. BRIKARD is demonstrated on 67 diverse macrocycles with structural data, encompassing various ring topologies. We find this approach enumerates diverse structures with macrocyclic RMSD < 1.0 Å to the experimental conformation for 85% of our data set in contrast to success rates of 67–75% with other approaches, while for the subset of 21 more challenging compounds in the data set, these rates are 57% and 10–29%, respectively. Because the algorithm can be efficiently run in parallel on many processors, exhaustive conformational sampling of complex cycles can be obtained in minutes rather than hours: with a 40 processor implementation protocol, BRIKARD samples the conformational diversity of a potential energy landscape in a median of 1.3 minutes of

[*]Corresponding Authors: evangelos.coutsias@stonybrook.edu, katrina.lexa@merck.com.

Present Addresses

(K.W.L.) Merck & Co., Inc. 2000 Galloping Hill Rd. Kenilworth, New Jersey 07033, United States.

(S.N.P.) Department of Mathematics and Statistics, Wright State University, Dayton, Ohio 45435, United States.

Author Contributions

K.W.L., E.A.C. contributed equally. K.W.L., M.P.J., E.A.C. planned the study. K.W.L., E.A.C. designed and executed the simulations and performed the analysis. E.A.C., M.J.W., S.N.P. designed the numerical algorithms. All authors contributed to writing the paper.

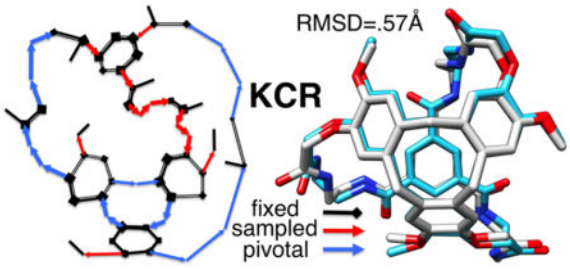
The authors declare the following competing financial interest(s): M.P.J. is a consultant to Schrodinger LLC, which developed or licensed some of the software used in this work.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.6b00250. Coordinates for the best-matching structures for all 67 compounds in this study (ZIP) Details on structure generation, sampling performance of MMBS vs BRIKARD using a parallel protocol, RMSD calculations for symmetrically arranged compounds, outlines of the IK algorithms including a numerical example of R6B6, and all tables and figures referenced in the text (PDF)

wallclock time, much faster than 3.1–10.3 hours necessary with current programs. By rigorously testing BRIKARD on a broad range of scaffolds with highly complex ring systems, we push the frontiers of macrocycle sampling to encompass multiring compounds, including those with more than 50 ring atoms and up to seven interlaced flexible rings.

Graphical Abstract



1. INTRODUCTION

Macrocycles straddle the space between small molecules and macromolecules, possessing complex and frequently flexible three-dimensional structures but relatively modest molecular weight (typically <1200). In the context of drug discovery, macrocycles can modulate challenging targets such as protein–protein interactions while obtaining passive membrane permeability and oral bioavailability similar to small molecule drugs. Other macrocycles are cavitands, capable of carrying guest molecules of a certain size and regulating their exchange, which is of considerable interest for controlling chemical reaction rates and stabilizing reactive intermediates. In principle, novel macrocycles could be designed using conformational sampling, but the inherent challenges of generating near-native structures for large, constrained cycles have rendered successful application of this approach difficult.

Conformational sampling algorithms developed for small molecules perform poorly for large and complex macrocycles, in part due to an inability to efficiently sample conformations consistent with ring closure.[1] In principle, protein loop sampling algorithms can be adapted to macrocycle sampling because they solve a similar ring closure problem.[2] In practice, such approaches are limited, because they rely on knowledge of energetics and conformational preferences of standard amino acids and thus are not easily generalized to treat chemically diverse macrocycles. The difficulty in sampling macrocycles is often compounded by intricate molecular topology: beyond the primary macrocyclic ring, which frequently includes large (>20) numbers of rotatable bonds, additional linkages often connect distant atoms along the ring, and they are often braided by smaller rings, sharing atoms with the main chain; see Figure 1. The connectivity topology is particularly labyrinthine for the cavitands, as they involve many interconnected yet flexible rings. Exploring the allowed conformations of long chains with multiple connectivity constraints presents considerable computational challenges, in particular, how can we thoroughly sample the potential energy surface without falling into kinetic traps?

Some success has been seen recently with low-mode-based methods, where high-temperature Molecular Dynamics (MD) or Monte Carlo searches of torsional space are combined with perturbations along the low-energy eigenvectors of local energy minima. These methods include LowModeMD,[3] MD/LLMOD implemented as MacroModel's macrocycle baseline search (MMBS) by Watts et al.,[4] and the optimized sampling protocols from Chen and Foloppe:[5] CF-LowModeMD and CF-MMBS. Despite some success, these low-mode approaches are hampered by computational costs ranging from several hours to days of wallclock time and a tendency for kinetic trapping to occur, particularly for complex ring scaffolds.

Structural computations of multiring systems with possible interdependence among the rings has been addressed by geometrical techniques from Distance Geometry (DG).[6,7] The DG techniques are capable of completely covering conformation space, but the computational cost grows steeply with the number of Degrees of Freedom (DoF),[8] especially for large, flexible cycles. This was demonstrated by Labute, who found that DG-based methods did not cover the potential energy surface of complex cycles as accurately as LowModeMD.[3]

Inverse Kinematics (IK) is another class of geometric methods that differs from DG in that IK focuses on torsional rather than distance DoF. IK methods have been previously applied to structures with multiple rings through iterative minimization of a multiring-closure objective.[9] However, no fully algebraic approach has been attempted to date. A distinguishing feature of fully algebraic IK methods is that, by explicitly and rigorously accounting for the constraints imposed by ring closure, they provide for substantially lower dimensional descriptions than general DG methods, resulting in superior efficiency and speed. As a simple example, conformational sampling for cyclooctane can be reduced to a two-dimensional search.[10,11]

In the present study, we reveal an IK algorithm for sampling multiconstraint macrocycle systems. Combining techniques from kinematics and computational algebraic geometry,[12,13] such as the theory of resultants in multivariate polynomial systems, our algorithm, BRIKARD ('Builder for Recursive Inverse Kinematic Assembly and Ring Design', which we named after Raoul Bricard, an originator of the modern theory of mathematical linkages[14]), can generate macrocycle conformations satisfying a wide range of structural constraints: amide bonds, cysteine bridges, five- and six-atom rings fused to the macrocycle, cages, lassos, etc. Energy minimization may be employed to capture intramolecular interactions.

The IK/algebraic approach we follow is enumerative: a recursive breadth-first construction produces all sterically feasible ring combinations that are consistent with a given set of values of the sampled torsions. Thus, BRIKARD explores conformational space in parallel. For every sampled torsion, points on all alternative branches of shape space are generated, in sharp contrast to the current trajectory-based searches that explore a single conformation at a time, and may not be able to find all alternatives even after extensive simulation. Our method is tested against MMBS, CF-MMBS, and CF-LowModeMD, achieving lower average Root-Mean-Square Deviation (RMSD), 2 orders of magnitude increase in speed

using our standard protocol, and success even in cases for which the currently available sampling methods fail.

2. RESULTS

2.1. Diversity of Macrocycle Space

The main objective of this study was to develop a general-purpose method that is capable of quickly sampling near-native conformations of macrocycles. We examined all 30 macrocycles in Chen and Foloppe[5] as well as the most challenging compounds identified in other studies[1,3,4] (Table S1). We augmented this collection with another 25 compounds which together represent a broader spectrum of interesting macrocycles used as chemical probes, catalysts, or pharmaceutical leads. To determine the extent to which our set covered a wide range of cyclic scaffolds, we examined common descriptors of molecular topology (Table S2). The cyclomatic number is defined in graph theory as the number of edges that must be removed from a cyclic system to break all cycles and therefore can be used to deduce the number of rings on rings in a macrocycle. The average cyclomatic number of all macrocycles in our set is 4.8, versus 3.7 in the Chen set and 3.3 in the Watts set. The average number of bonds in a ring system was 43.2 in our set, compared with 30.4 for the Chen set and 33.0 for the Watts set. As with a cyclomatic number, a larger number of bonds in the ring indicates a more complex ring. According to both of these common metrics, the increased diversity and complexity of our data set is significant (p-value ≤ 0.01) by an independent *t*-test. Putting these metrics in the context of common macrocycles, epothilone B (EPB) is a small ring system, with a cyclomatic number of 3 and 23 bonds in the ring system. Cyclosporine (CSK) is a large, single-ring system with a cyclomatic number of 1 and 33 bonds in the ring system, while the molecular cage cryptophane (CRP) is a highly fused system, with a cyclomatic number of 10 and 85 bonds in the ring systems. In combination, these descriptors illustrate the broad diversity of our collection of macrocycles. The SI includes a full list of each macrocycle used in this study, together with sampling results for each method (Tables S1–S5).

An accurate sampling approach should be capable of identifying conformations within 1.0 Å RMSD to the ring atoms of an experimental structure, ideally within 0.5 Å for smaller macrocycles. This level of accuracy enables the use of these conformations in structure-based design programs where near-native structures are necessary for modeling biophysical properties. In this study, we assessed the performance of BRIKARD with the currently available programs for macrocycle sampling across our data set. This comparison was done in multiple ways: 1) evaluation of the performance of each program run on a single processor, starting from a single input conformation, 2) evaluation of the performance of each program run using its published protocol compared with BRIKARD run as we feel the algorithm should be, and 3) evaluation of the performance of each program run in the same manner as BRIKARD, with multiple short sampling searches in parallel (see the SI: Parallelization Comparison). Keeping the number of sampling iterations the same across all algorithms, we found that BRIKARD more frequently sampled regions of the conformational space that were close to the crystal structure relative to CF-LowModeMD and the MMBS approaches.

BRIKARD is fast and accurate in a single-processor mode starting from a single seed compared to results for MMBS and CF-LowModeMD run in an identical manner. This comparison was done for 26 macrocycles randomly selected from our total collection of compounds under study in this paper (Figure 2). For this subset of macrocycles, BRIKARD locates 46% within the strict success metric of 0.5 Å ring RMSD. BRIKARD without minimization locates 21%, MMBS locates 19%, CF-MMBS locates 35%, and CF-LowModeMD locates 23%. The efficiency of BRIKARD alone is clear when run on a single processor, requiring a fraction of the time the other programs require to explore the potential energy surface of each compound. Minimization as a serial process after running BRIKARD shifts the time scale into the region occupied by MMBS, CF-MMBS, and CF-LowModeMD. However, BRIK-ARD does not need to run as a single process, which is one of its great advantages. The IK approach allows a complete redesign of the geometry by wide sampling of all torsions subject to exactly enforceable closure constraints. As a result, each iteration is truly independent from the ones before and after, while the nontorsional information for a structure is inherited from the input conformation. Subsequent structures may be minimized, and the input, or seed, conformation may be arbitrarily changed. We decided to use two seed conformations to increase the information diversity from our initial structures. Our discussion throughout the remainder of this study will focus on the results obtained running all programs according to their optimum/published procedures.

For this study, each conformation generated by BRIKARD was minimized using PLOP.[15,16] This energy minimization improved sampling performance, with increased computational cost. Analyzing the performance of BRIKARD sampling alone, we found substantial performance gains and decent reproduction of the experimental state. At a threshold of 1.0 Å ring RMSD to the experimental conformation, raw BRIKARD samples near-native conformations as well as the other programs, even without minimization of the structural ensemble (Table 1). However, energy minimization was crucial for achieving the threshold of 0.5 Å ring RMSD (Figure 3).

Among the 67 compounds used for this study, we identified 21 “hard” cases, based on at least one of the methods failing to achieve sampling below 1.5 Å ring RMSD to the experimental structure. For these compounds, BRIKARD succeeded on 18/21 using energy minimization, while without minimization that number was 16/21. For the other methods, success rates for CF-LowModeMD were 5/21, for MMBS 6/21 and for CF-MMBS 8/21. The hard compounds with the best RMSD for each method are listed in Table 2, while Figure 3(a) plots the best RMSD for each compound vs wallclock time. The remaining compounds which we characterize as “easy” show reasonable success rates for all methods (Table S5); however, as can be seen in Figure 3(b), BRIKARD with energy minimization does show a small accuracy gain together with a substantial computational speedup. Figure S1 shows a cumulative plot of RMSD vs wallclock time for the entire data set.

Beyond an improvement in accuracy, a major feature of BRIKARD lies in its highly efficient random sampling and high scalability. To assess the relationship between the number of sampling iterations requested and RMSD to the experimental structure, we examined subsets of the conformational ensemble found for each compound. Subsets containing 1, 5, 10, 25, 50, 75, 100, 250, or 500 conformers were selected randomly from the total conformational

ensemble; this process was repeated 1000 times for each subset size. The linearity between the lowest RMSD and subset size was used to judge sampling efficiency, defined as the number of iterations required to observe the lowest RMSD conformer. As anticipated, we found the sampling efficiency for the more complex macrocycles to be more affected by number of sampling iterations, whereas smaller macrocycles reached the asymptote of effective sampling rapidly, requiring fewer iterations to reach maximally efficacious sampling (Figure 4). In contrast to BRIKARD which can immediately find a novel conformation within a single iteration, LowMode-based sampling methods require many iterations to fully sample the native ensemble and therefore cannot be efficiently parallelized to the same extent. When asked for a limited number of sampling iterations, instead of jumping into a new state, LowMode methods will only sample conformations among the local minima close to the starting point requested on a single processor (see the SI, Section 2).

Since BRIKARD sampling is memoryless, sampling of a single compound can be distributed over multiple processors with calculation time decreasing linearly as the number of processors used increases. In this study, a total of 40 individual processors were used, each performing 250 sampling iterations. We wanted to further explore this idea of sampling efficiency by comparing BRIKARD with LowMode methods when they are actually implemented to run in parallel. MOE uses multithreading to distribute LowModeMD searches over available processors, although it is still a single sampling trajectory being explored. The MMBS approach was not published with a parallel implementation, so for the purposes of comparison we examined MMBS and BRIKARD results when both are run using two seed conformations to start and 40 processors total. These results from running both BRIKARD and MMBS in the same fashion indicate that although these programs could be run in parallel in principle, it is generally not advantageous to run low mode methods in this manner. A detailed comparison is available in the SI, Section 2.

When each method is employed according to its optimal protocol, BRIKARD is 27 times faster than CF-LowModeMD, 59 times faster than MMBS, and 89 times faster than CF-MMBS for an equivalent number of steps, based on the median of the maximum time for parallel BRIKARD. If only a single processor was used, the median cumulative time for BRIKARD would still be 1.4 times faster than CF-LowModeMD, 3.1 times faster than MMBS, and 4.7 times faster than CF-MMBS. Importantly, CF-LowModeMD is run as a multithreaded process by default, thus it behaves as a partially parallel program. Timing information from CF-LowModeMD run on a single thread indicates the process uses between 3 and 4 threads. As a consequence, the actual CPU-hour cost for a single CF-LowModeMD run is between 1.4–4 times greater than that indicated by the wallclock time.

Many macrocycles can adopt different conformations depending on their environment (solvent, temperature, whether they are bound to another molecule), and thus it is important to identify all low-energy conformations. In addition to the high accuracy and speed of BRIKARD, our method also generates ensembles of substantial diversity. A measure used to characterize performance of a macrocycle sampling method is eccentricity (ecc), which is found from the ratio of the minimum and average of the principal moments of inertia (PMI)

$$\mathrm{ecc} = 1 - \mathrm{PMI}_{\mathrm{min}} / \mathrm{PMI}_{\mathrm{avg}}$$

where ecc describes a shape ranging from completely linear (ecc = 1) to spherical (ecc = 0). Previous studies[4] found the average ecc of more challenging macrocycles to be significantly different from their experimental ecc. We found the conformer shapes predicted by BRIKARD match the known structure substantially better than the other algorithms, with the average eccs from BRIKARD having an $r^2 = 0.49$ to experiment. The single outlier was BIX, which has a linear-like eccentricity (0.86) in the crystallographic "imploded" state, but an average eccentricity of 0.25 from all BIX conformations sampled with BRIKARD and of 0.35 within the low-energy set. Without BIX, the overall correlation between experimental and average calculated ecc improves to $r^2 = 0.63$. By contrast, the other methods sampled conformer shapes further from the experimental state, with correlations of $r^2 = 0.24$ ($r^2 = 0.29$ without BIX) for MMBS and $r^2 = 0.35$ ($r^2 = 0.45$ without BIX) for CF-LowModeMD. The correlation between the average eccentricity by compound for each method relative to the crystallographic eccentricity is given in SI Figure S2.

2.2. Kinematic Sampling Retrieves the Complete NMR Ensemble

In addition to examining results over the entire data set, we were interested in understanding performance in the context of a few particularly challenging cases. First, we asked whether BRIKARD could sample native conformations of IP01 and IP02, nonapeptides cyclized via a side chain-side chain disulfide linkage. The disulfide linkage hinders application of peptide-based loop-sampling approaches, while the 28 ring atoms contribute to significant conformational diversity. cyclo(CLLRMRSIC), or IP01, was identified as a micromolar ICAM-1 inhibitor and later optimized to improve binding affinity, yielding cyclo-(CLLRMKSAC), IP02.[17,18] Interestingly, the less potent nonapeptide, IP01, was found by NMR and MD to interconvert between four major conformations, while IP02 stably occupied a single conformation. This experimental observation suggested the higher affinity of IP02 may result from its conformational rigidity leading to a smaller entropic loss.[19] Thus, we examined BRIKARD's utility for uncovering differences in the conformational ensembles of closely related macrocycles.

For IP01, BRIKARD generated low-energy conformations within 0.83–0.97 Å of all four major conformations identified by NMR[17] in 0.14 wallclock hours. The torsional angles of residue 4 (ARG) are distinct for each NMR state, so we use them to visualize the extent of conformational sampling in Figure 5. The four major conformations of IP01 are well-represented within the conformational ensemble generated by BRIKARD; however, two of the four (states C and D) dominate the low-energy ensemble (Figure 5a–b). Among these two lower-energy conformations, one conformer is similar (state C, ring RMSD = 1.33 Å) to the NMR conformation identified for the more potent analogue IP02. Sampling of IP02 revealed a narrower range of conformational variability compared with IP01 and low-energy conformations 0.93 Å from the NMR state. During the conformational search, BRIKARD explored all known NMR states for both peptides, in addition to alternate conformations.

In contrast, sampling of IP01 and IP02 with MMBS and CF-MMBS required substantial time, 22–26.62 wallclock hours each, and every conformation generated was >3 Å in RMSD to all NMR conformers. Only CF-LowModeMD generated conformations with accuracy approaching that of BRIKARD, albeit with substantially greater computational time, requiring 7.67 wallclock hours to sample conformations with lowest RMSDs of 0.73–1.18 Å to the NMR structures of IP01 and 1.21 Å to IP02. This outcome is reflected in the torsions for residue 4 (Figure 5c–f), wherein CF-LowModeMD samples the native torsions for IP01 but not IP02, and MMBS fails to sample them for either cyclic peptide.

2.3. Sampling with BRIKARD Is Unaffected by High Energy Barriers to Conformational Change

A central goal in developing BRIKARD was to establish a robust sampling method for studying fused ring systems, which are very difficult for other current methods due to their high topological complexity combined with significant steric constraints. Synthetic cage molecules, developed to selectively sense and sequester various ions, represent perhaps the most challenging systems to sample. Cryptophane is a canonical example of a cage macrocycle, where two symmetrical sets of interlocked cyclotriveratrylene rings are connected by three ether chains. The CSD crystal structure RAYFED depicts a more complex analogue of cryptophane E, with six disordered ether chains connecting the two ring systems and an antimony anion in the cage center.

BRIKARD required a maximum of 1.15 wallclock hours per run to identify 65 unique conformations within 10 kcal/mol of the lowest energy and sampled within 0.67 Å of the crystal structure of the cryptophane E analog. None of the other sampling methods tested generated conformations similar to the crystal structure, identifying only conformations with >4.4 Å RMSD. In the case of MMBS and CF-MMBS, the generated conformations were all highly similar to the initial conformation, indicating an inability to escape local energy minima.

In addition to conformations similar to the crystallographic structure, BRIKARD also identified two other major low-energy conformational states, which we speculate could represent preferred conformations in the unbound state (Figure 6). Upon examining crystal structures available for similar compounds, we found several other cryptophane systems that adopted conformations similar to the alternate states found by BRIKARD (CSD reference codes OJIVUZ, OJITUX, and BIMXUR). Conformational sampling of cryptophane (±)-anti-1 with BRIKARD reproduced the crystallographic observations of an “imploded” cryptophane state by Mough et al.[20] while also sampling the expanded state, despite the reported ~24 kcal/mol barrier to this conformational switch. Through the study of these molecular cages, we found that BRIKARD is capable of identifying the experimentally relevant conformations consistent with all ring constraints, without becoming trapped in spurious local minima or requiring substantial calculation time.

2.4. Pushing the Limits of Computationally Tractable Macrocycles

Our ultimate goal in developing the kinematic sampling approach for macrocycles was to address pharmaceutically relevant compounds well beyond the limits of the current best-in-

class approaches. As macrocycle size increases beyond 20–25 rotatable bonds (eq 1), the sheer scale of the search problem overwhelms these methods.[4] Stochastic searches become trapped in deep local energy minima and fail to produce reliable structures that approximate the experimental structure. QN7 is a bicyclic 17-residue peptide, able to bind urokinase-type plasminogen activator at high nanomolar affinity by mimicking the binding interactions of a small protein.[21] QN7 sits at the limits of what we believe a macrocycle sampling algorithm must be able to handle in order to be a reliable tool for predicting biophysically relevant macrocycle conformations. The only available structure of QN7 is a PDB structure at 1.90 Å resolution, with good density for the macrocycle residues in contact with the protein interface, but missing density for four of the solvent-exposed residues.

BRIKARD samples within 1.66 Å of the experimental structure in 0.43 wallclock hours. Restricting the RMSD calculation to the macrocycle residues in contact with the protein interface improves the fit to 1.40 Å, indicating that BRIKARD yields a reasonable starting conformation for structure-based design applications. In comparison, the closest RMSDs to the crystallographic structure for CF-LowModeMD, MMBS, and CF-MMBS were 3.83, 3.36, and 6.12 Å respectively, with corresponding computation times of 0.43, 49.1, and 68.8 wallclock hours. At the time of this study, QN7 was among the largest cyclic peptides available in the literature with a reasonable crystal structure. These results for QN7 indicate that cyclic systems comprised of approximately 50 ring atoms sit at the current boundary for BRIKARD and are beyond the scope of other macrocycle sampling tools.

3. METHODS

3.1. Data Set

A set of 67 diverse compounds covering a wide spectrum of macrocyclic scaffolds, including polyketides, nonribosomal peptides, and cavitands, was compiled to develop and evaluate BRIKARD. Initial coordinates of the experimental structure for each compound were downloaded from the Cambridge Structural Database (CSD),[22] if available. Otherwise, structural data was downloaded from the PDB.[23] To enable an assessment of conformational plasticity under the influence of a binding partner, two compounds were included that are present in both the CSD and the PDB: soraphen (S1A) and tacrolimus (BKF). In situations where multiple entries were available for a single compound, the best resolution crystal was selected for inclusion in our set. In a few instances, NMR data exists for cyclic peptide macrocycles that offered a compelling opportunity to judge BRIKARD's ability to sample all conformations within the experimental ensemble.[17,18] Several compounds from previous work on macrocycle sampling[1] do not have an experimental structure, but the lowest-energy conformation sampled by their method was published. These minimum-energy conformations are used as the sole structural information for AP2, HP3, and MP4. The full set of compounds included in this study are listed in the Supporting Information.

3.2. Conformational Sampling

MMBS was performed using the optimal protocol published by Watts et al. CF-MMBS was performed using the optimal settings for macrocycles according to Chen and Foloppe using the MD/LLMOD method. CF-LowModeMD sampling was performed using the improved

protocol identified by Chen and Foloppe, where the electrostatics treatment was changed from distance dielectric to generalized Born and the energy cutoff was increased from 7 to 15 kcal/mol.

In contrast to these protocols, which are based on diversity resulting from MD trajectories, BRIKARD's approach to sampling is purely geometrical. All rotatable bonds are sampled. Rotatable bonds belonging to open side chains are sampled randomly, except for side chains of standard amino acids that are sampled using the Dunbrack backbone dependent rotamer library.[24] Rotatable bonds on rings are sampled in a manner consistent with the ring closure constraints.

The number of rotatable bonds (RTB) sampled by BRIKARD is given by

$$\mathrm{RTB}=N_{\mathrm{free}}+\sum_{i\in\mathscr{R}}(N_i-6-B_i^{\mathrm{rigid}}-B_i^{\mathrm{shared}})+\sum_{k\in\mathscr{F}}B_k^{\mathrm{shared}}+R_5 \qquad (1)$$

where N_{free} is the number of free torsions not in rings, excluding torsions whose only effect is to rotate methyl hydrogens, $\mathscr{R}$ is the set of flexible (nonaromatic) six or higher membered rings, N_i is the number of bonds on ring i, and B_i^{rigid} are rigid bonds and B_i^{shared} are rotatable bonds shared with (an)other ring(s). To avoid multiple counts, shared bonds are removed from the first sum and added back in the second, counted separately once. $\mathscr{F}$ is the set of ring fusion segments. R_5 is the number of nonplanar five-membered rings. As discussed in the SI, Section 6, five-membered rings have a one-parameter set of shapes, and the driver torsion has a limited range, e.g., proline χ_5 angle ranges between ~±30°; the shape space is topologically a circle with two alternative shapes (maximum and minimum pucker) for each permissible value of χ_5. The method for five-membered rings in Ho et al.[25] is implemented here in a robust, fully polynomial formulation. For six-membered rings, the TLC algorithm (see the SI, Section 7.1) is used, and at most four (rigid) states are possible (RTB = 0 with no driver torsions). Rings involving $N \geq 6$ rotatable bonds possess up to N–6 continuous DoF. When N – 6 of the rotatable torsions are set to some value, up to 16 distinct sets of values for the remaining 6 torsions that close the ring are possible. Consequently, 6 among the torsions around a ring system need to be set aside for loop closure and may not be counted as independently rotatable. Not all combinations of the drivers may give a solution, but determining the regions of existence (workspace of the molecular linkage system) a priori is a hard, open problem.

In sampling protein backbones and side chains, abundant information on structural propensities, such as Ramachandran preferences for the backbone[26] and side chain rotamers,[24] is available to limit the sampling space.[27–30] Such general information is not often available for macrocycles in which case a wide-range scan of all available DoF is used.

Each solution may be subjected to local energy minimization using the multiscale truncated Newton algorithm in PLOP[15,16] with the OPLS2005 force field (also used by both Chen et al.[5] and Watts et al.[4]) and a generalized Born implicit water model. After minimization, the

diversity of the ensemble may be substantially reduced, with a few dominant clusters emerging for the smaller rings.

3.3. Data Set Comparisons

The set of molecules used in this study is listed in Table S1, together with the codes that we use to identify each compound in all subsequent tables and in the text. Table S2 lists molecular descriptors of size and flexibility, including cyclomatic number, number of rotatable bonds (the number of degrees of freedom sampled by BRIKARD as given by eq 1), molecular weight, and eccentricity. Crystallographic additives and binding partners were removed, leaving only the compound of interest. Dihedral angles were perturbed, and the resulting conformations were minimized in MacroModel to obtain starting conformations at least 2.5 Å from the crystallographic state. This approach was followed to retain atomic naming conventions and simplify root-mean-square calculations; however, input structures can also be generated from a SMILES string. As explained in detail in the Supporting Information, two low-energy conformations, Seed A and Seed B, from this initial ensemble were selected as inputs for BRIKARD. We found that using multiple starting states provided an advantage over use of a single input conformation. Analysis of the sampling results indicated that neither seed offered a privileged initial state for any compound.

Twenty independent BRIKARD runs were performed in parallel for each input conformation, each requiring 250 sampling steps and starting from a different random seed. This resulted in a total of 10,000 sampling iterations, the same number requested in MMBS and CF-LowModeMD, distributed across 40 independent processors. This number of processors was settled upon as a good balance of parallelism and subsampling efficiency. OPLS2005 force field parameters were assigned for each subunit involved in the macrocycle using the hetgrp_ffgen utility in Schrödinger. This is the same force field used by MMBS and CF-MMBS, which facilitates fair comparison of the minimized conformations. Seed B was used as the initial seed for all MOE/MMBS runs. In the Supporting Information, we show results from running MMBS using the same protocol performed with BRIKARD (Seeds A and B each used to initialize 20 independent, 250-iteration runs). The output conformations were minimized by the Protein Local Optimization Program (PLOP)[15,16] using the default Generalized Born solvation model, with each set of 250 structures from a single BRIKARD run minimized within a single PLOP run. Total wallclock time required for each run is detailed in Table S3. The global energy minimum was identified among the 10,000 minimized conformations and used to determine the energy cutoff for retaining conformations within 10 or 15 kcal/mol. These cutoffs were chosen because MMBS uses a 10 kcal/mol filter for reducing the total number of conformations generated, and CF-MMBS and CF-LowModeMD use a 15 kcal/mol filter. Table S4 provides the total number of conformations generated by each method for every compound and includes the number of conformations within the 10 or 15 kcal/mol filter window for BRIKARD. To contrast effective sampling around near-native states by these different techniques, full details of IP01 and IP02 ARG 4 torsion angle populations are given for CF-LowModeMD, MMBS, and BRIKARD overall and filtered to the low energy structures (10 or 15 kcal/mol from the minimum) in Figure S2.

One metric by which performance of BRIKARD was evaluated was the RMSD of the sampled backbone conformation relative to experimental data. The RMSD to the experimental structure of all atoms along the major ring of the macrocycle was calculated using Chimera[31] and BRIK-ARD's frmsd code.[32] Only the ring atoms are used to compare to the experimental structure since the side chains, especially the polyketide tails, may adopt multiple low-energy conformations in solution. Unique conformations were judged by clustering with a heavy-atom RMSD threshold of 0.25 Å. For symmetric compounds composed of identical subunits (such as the cryptophane analogs), correct RMSD calculation required considering all possible alternative alignments. These ring RMSD results are provided in Table S5. The success of each method in sampling structures close to the native structure is illustrated in Figure 7. The cumulative fraction shows the percentage of native-like structures identified across the entire compound set as a function of the criteria for success. Using a ring RMSD cutoff of 0.5 Å, BRIKARD samples a native-like conformation for 58.2% of the data set, while at a loose cutoff of 2.0 Å, BRIKARD samples a native-like conformation for 100% of the data set.

The diversity of the conformational ensemble generated by each method was assessed by calculating the radius of gyration and the eccentricity. This enabled comparison to the performance of the current best-in-class methods: CF-Low-ModeMD,[3] MD/LLMOD, implemented as MacroModel's macrocycle baseline search (MMBS) by Watts et al.,[4] and CF-MMBS by Chen and Foloppe.[5] For every sampled conformer, we calculated the ring and heavy-atom RMSD to the experimentally derived structures obtained from the PDB, CSD, and/or NMR.

4. METHOD OVERVIEW

BRIKARD models a polycyclic molecular system as interlaced chains of linked rotors or linkages. Such systems are assumed to have fixed lengths and angles but generally flexible torsional DoF. Given a closed molecular chain (a ring) or a chain with fixed ends (a loop), the problem of IK is to find all values of the torsions consistent with the closure constraints, referred to as the problems of Ring or Loop Closure (RC or LC, respectively). These two problems are essentially the same, and we shall use the terms interchangeably. In the kinematics literature,[12] the problem is referred to as a 6R-6 bar linkage, with closure expressed as a set of polynomials that involve a set of six torsions, or pivots, used to close the ring. The remaining torsions are free torsions or drivers. This formulation enables algebraic expansion of all possible states wherein the ring successfully meets closure conditions.

Previous applications of IK to molecular modeling focused mainly on the LC problem, relying on two fixed frames in space, e.g., two protein residues separated by at least two other residues along the backbone, and the goal is to find the conformational ensemble of the protein chain that traverses between those two fixed frames. Go and Scheraga's[33] approach, one of the earliest applications of IK to molecular modeling, results in transcendental ring closure equations which are solved by an iterative method that may miss certain solutions unless exceeding care is taken. Subsequently, algebraic formulations were developed, paralleling work in robotics. Algebraic formulations of ring closure result in systems of

polynomials which allow determination of all the roots reliably, using standard polynomial real root locator algorithms.[34–38] However, algebraic IK formulations for cyclic molecules have been limited to simple rings.[38] An application of IK to systems with multiple constraints by Thorpe and Lei[9] relied on iterative solution of the multiconstraint system, again offering no guarantee that all solutions to a given system may be found.

Alternative LC schemes rely on generating initial conformers which are then used as seed points for exploring their neighborhood in shape space by perturbing various DoF and then relaxing strains. This approach is shared by purely geometrical methods (DG,[6,7] CCD,[39] and other iterative closures[2,29,40]). MD methods similarly proceed from a starting structure that must be generated via some minimization process, again not guaranteeing completeness. The search then proceeds by following a continuous trajectory that is generated via mechanical or geometrical perturbation, the latter followed by an annealing step. Such motion might fail to leave certain neighborhoods—a phenomenon called kinetic trapping for MD—so that even if one of the neighborhoods in torsion space corresponding to one particular shape may be explored thoroughly, adjacent ones could be completely missed.

Several technical difficulties need to be overcome in order to extend IK to loop/ring closure with fused rings. The order in which various rings are solved (Figure 8) is crucial; torsions belonging to a solved ring must be frozen and properly converted at points of fusion to torsions along the backbones of adjacent rings that are solved in succession.

Figure 8 shows the key elements of this hierarchical closure approach for a typical example of a macrocycle, 13-desmethyl spirolide C (SQX).[41] SQX has a total of 11 nontrivial rotatable bonds and consists of a braided set of small rings on a larger ring system, which is a common theme in many natural product macrocycles. The main backbone is traced in boldface bonds, {1–7} index kinematic loops defining the order of solution, broken lines indicate “gaps” or bonds that must be specified for a “de novo” construction. Arrows indicate nonrigid torsions: red are drivers or free torsions while black are slaved torsions or pivots. Note that for each ring there are alternative sets of pivot values for a given value of the drivers.

In the remainder of this section we outline the key ingredients of the method: Section 4.1 reviews the main ideas in ring closure, while Section 4.2 discusses the issues arising from the presence of interlocked rings. Finally, Section 4.3 discusses topological issues related to exploring the shape space of multicyclic systems.

4.1. The IK Problem and Ring Closure

We present a general mathematical framework for the problem in Inverse Kinematics known as a 6R-6 bar linkage. The problem has been studied extensively in the mechanical engineering/robotics as well as the biochemical literature. However, there remain subtle issues related to the geometry of families of lines.[42] Here we introduce notation, and recapitulate key ideas and useful formulas, that have not previously appeared in this context.

Consider a kinematic chain of $N \geq 6$ rotor links, $\mathbf{b}_i$, $i = 1,\ldots,N$. The chain may be closed or attached to rigid ends. We fix the lengths b_i and pairwise angles θ_i of the links. Then, the

torsions t_i may be chosen to construct realizations of the chain that are consistent with the closure constraints.

Introduce a concerted change of all the torsions between the end points of a loop $\mathbf{t} \rightarrow \mathbf{t} + d\mathbf{t}$ that maintains the invariance of the chain past its ends (i.e., with respect to which the ends of the chain maintain geometrically correct attachments to rigid molecular frames or to each other in the case of a ring). Then, at any point of the chain **R** past the fixed ends, we should have that

$$0=d\mathbf{R}=\sum_{i=1}^{N}\Gamma_i \times (\mathbf{R}-\mathbf{R})dt_i \Rightarrow (\sum_{i=1}^{N}\Gamma_i dt_i) \times \mathbf{R}-(\sum_{i=1}^{N}\Gamma_i \times \mathbf{R}_i dt)=0$$

Since this is true for arbitrary **R** past the end of the flexible part of the linkage (in fact it is sufficient to consider the three consecutive nodes defining the end rigid frame, which coincides with the start frame for a ring), both expressions in parentheses must vanish independently, from which we find

$$Pd\mathbf{t}=\sum_{i=1}^{N}\mathbf{P}_i dt_i=0,\ P:=(\mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_N)\ \text{where}\ \mathbf{P}_i=\begin{pmatrix}\Gamma_i\\ \Gamma_i\times\mathbf{R}_i\end{pmatrix} \quad (2)$$

and Γ_i and $\mathbf{R}_i$ are the unit vectors along the ith rotator axis and its position from some arbitrary reference origin. Basic analysis (the Implicit Function Theorem) guarantees that six of the variables may be expressed as differentiable functions of the remaining ones provided the $6 \times N$ matrix P has full rank. Let their indices be i_k, $k = 1,\ldots,6$; the corresponding variables are called the pivots. The remaining variables, called the drivers, may be indexed by j_k, $k = 1,\ldots,N–6$. Introduce $p_k := t_{i_k}$, $k = 1,\ldots,6$ for the pivots and $q_k := t_{j_k}$, $k = 1,\ldots, N–6$ for the drivers. Then

$$J\begin{pmatrix}dt_{i_1}\\ \vdots\\ dt_{i_6}\end{pmatrix}=-\sum_{i\neq i_{1,..,6}}\mathbf{P}_i dt_i,\ J:=(\mathbf{P}_{i_1}\mathbf{P}_{i_2}\cdots\mathbf{P}_{i_6}) \Rightarrow d\mathbf{p}=-J^{-1}\mathbf{Q}d\mathbf{q} \quad (3)$$

giving the differential of the pivot variables as a function of the differential of the free variables and involving the inverse of the Jacobian. Where that inverse exists, we say that the IK problem is well posed. Where the determinant of the Jacobian vanishes, we have kinematic singularities. Such singularities are not necessarily intrinsic and could be simply a result of the particular choice of pivots and drivers. The columns of the Jacobian J are the Plücker coordinates[12] of the pivot axes. A similar, coordinate-free expression for the 6-membered ring Jacobian appeared in Viquerat et al.[43]

The relations implied by eq 3, giving the pivots in terms of the drivers, are the explicit closure conditions, and they are polynomials in the sines and cosines of the pivots. Let $u_k = \tan(p_k/2)$, $k = 1,\ldots,6$, be the half-tangents of the pivots. The q_i are sampled and set to a fixed value. Then, the 6R-6 bar problem may in principle be formulated as a polynomial system

$$\mathscr{P}_j(\mathbf{u};\mathbf{q},\mathbf{a})=0, \quad j=1,\ldots,6$$

where **a** is the parameter vector of bond lengths and bond angles. Multiple solutions for the pivots are possible for a given set of drivers: choosing values for the q_i results in at most 16 real solution sets $u_{j,k}$, $j = 1,\ldots,6$, $k = 1,\ldots, K \leq 16$.[35,44]

4.1.1. Ring Closure Algorithms—Two different algorithms are employed by BRIKARD for selecting pivots and solving for pivot torsions, TLC and R6B6. The first, TLC (*Triaxial Loop Closure*),[35] requires that the pivots be arranged in 3 coterminal pairs. This arrangement, natural for amino acid chains in which pairs of $\phi - \psi$ torsions are typically free to assume values from the Ramachandran regions, allows for a simple, elegant, and robust formulation of the problem but cannot be directly applied in certain cases where 3 pairs of coterminal pivots may not be available. In such cases, we use the general formulation of the IK problem without a pivot restriction. We have developed a new algorithm, R6B6 (for 6 rotors/6 bar), that allows arbitrary pivots. Among the innovations in R6B6 are its ability to robustly eliminate variables without needing to try different elimination combinations[38] and to arrive at optimally sized polynomial problems of degree 16, as in the original formulation by Lee and Liang,[50] rather than degree 24 as in subsequent implementations that have avoided the original method's defects at the expense of solving a larger system.[45]

In general, TLC is faster than R6B6 by approximately a factor of 2, so it is used whenever applicable, while we employ R6B6 in cases where three pairs of rotors are not available. The two methods are outlined in the SI.

4.2. Ring Fusion

Along an open backbone, closures may proceed independently for braids of nonoverlapping rings, and the set of possible solutions is constructed purely combinatorially. By contrast, if rings are "fused", sharing certain DoF among two or more neighbors, then their shapes become interdependent. BRIKARD's data structure for modeling a polygonal line with multiple interconnecting branches requires a decomposition into a hierarchy of cycles, each with at least 6 additional rotatable bonds so loop closure may be carried out at each step.

Assume there are r rings; then the solution of the l^{th} (with $l \leq r$) ring involves solving the system

$$\mathscr{P}^l_j(\mathbf{u}^l;\mathbf{q}^l,\mathbf{a}^l)=0, \quad j=1,\ldots,6$$

Here, some nonpivot torsions q_i^l may be shared with lower-indexed rings that have already been solved.

When a ring of known structure (i.e., known torsions) is adjoined to a chain, then certain simple relations exist between the ring torsions and the torsions along the chain, assuming given three-dimensional structure at each branchpoint. As both the ring and the chain are oriented (with compatible orientation along shared bonds), we have to distinguish between convergent and divergent branches (Figure 9). We consider first merging a ring with the backbone. The case of two rings attached to the backbone and sharing off-backbone DoF is handled without any special processing by the chain reconstruction algorithm. We have the following:

1. *Converging branches* (Figure 9(a), 10(c))

$$q_i^m = q_1^l + \gamma_1^{l,m}$$

Referring to Figure 9(a), γ_1 is the improper dihedral formed by the planes $R_{i,1}$, R_i, R_{i+1} and R_{i-1}, R_i, R_{i+1} with $l < m$ being ring indices.

2. *Diverging branches* (Figure 9(b), 10(c))

$$q_j^m = q_\kappa^l + \gamma_2^{l,m}$$

Referring to Figure 9(b), γ_2 is the improper dihedral formed by the planes $R_{j-1}, R_j, R_{j,1}$ and R_{j-1}, R_j, R_{j+1}, again with $l < m$ indexing the fused rings.

3. *Converging-Diverging branches: single shared bond* (Figure 10(b))

$$q_j^m = q_1^l + \gamma_1^{l,m} + \gamma_2^{l,m}$$

Here, it must be understood that no torsion that is shared by several rings may be sampled or used as a pivot except for closing the lowest-indexed ring of which it is a part, and, together with all other torsions in that ring, it must be fixed to that value for all calculations involving all subsequent rings that contain it. Hierarchical rules are observed so that a torsion set during a previous step is fixed for all subsequently solved rings in which it enters. For solitary rings with no atoms shared with the backbone (Figure 10(a)), no special reconciliation is necessary.

4.3. Exploring Multiring System Topology

Fixing the driver torsions in a ring system and solving for the pivots, a set of alternative conformations may be found. Therefore, a neighborhood in the space of driver torsions may form the basis for a set of coordinate charts for a collection of diverse conformation neighborhoods. There may or may not exist direct kinematic pathways connecting these states, as these alternative torsion-based coordinate atlases comprised by the union of the set of drivers with all alternative sets of pivots may indeed converge along singularities of the

Jacobian or not at all. This is for example the case for canonical cycloheptane, whose shape space is composed of two disjoint cycles; however, without a priori knowledge of the topological properties of the conformation space, there is no practical way to ensure exhaustive coverage. Indeed, given "witness sets"[13] of points spanning all connected components of the shape space, it is possible to design efficient search algorithms to quickly scan each component.[46] However, guaranteeing the completeness of such sets is in general nontrivial.[11]

BRIKARD employs an algebraic approach to generate and explore all sterically feasible alternatives simultaneously. Figure 11 demonstrates this for the CSD cavitand BIMXUR. In Figure 11(a), the graph depiction shows the four interdependent flexible rings; there are 6 rotatable bonds in chain C that could be set to arbitrary values, although the solvability space is a priori unknown (*workspace* of the IK problem). The 9-membered rings 1 and 2 are capable of 4 shapes each, although only a few of the combinations correspond to realizable overall shapes. Once the backbone (heavy arrows) is selected, a data structure is constructed in which all rings are defined in terms of their branching off points and solved for hierarchically. Once a ring is solved for, its torsions and all other torsions rigidly connected to those are fixed through subsequent calculations.

Figure 11(b) shows a superposition of 53 shapes. These were generated by keeping rings 1 and 2 as well as chain C fixed to their crystal shapes. Under these restrictions, there were 8 possible solutions for chain A and 8 for chain B; in the latter case, one of the shapes was sterically excluded. The crystal structure (shown in Figure 11(c)) corresponds to the combination (a_1, b_1). Of the 56 distinct combinations, 3 are sterically infeasible. BRIKARD's purely algebraic approach finds all solutions reliably, using standard polynomial real root locator algorithms.

A strength of the fully algebraic approach is the ability to enumerate all alternate conformations resulting from a set of driver dihedrals. Consequently, beyond making it possible to design Monte Carlo schemes that obey detailed balance,[37,47–49] algebraic algorithms also make it possible to design successively finer searches around regions of interest with the confidence that all topological components may be found. However, the topological complexity is not a simple product of lower dimensional spaces unless the rings are truly independent, at least as far as choosing the values for the sampled torsions are concerned. When the various rings truly share DoF, we must take into account that each point of the conformational space of a ring of lower order has a "fiber" attached to it on which live the shape-spaces of rings of higher order. These attached spaces have structures that depend nontrivially on the base space. It is well-known that the spaces of molecular rings have complex topologies[10] with possibly disconnected topological components.[8,11,13] So for the same set of values of the sampled torsions, there exists a combinatorially large set of alternative shapes. In addition, steric conflicts upon 3D reconstruction impose restrictions, thus introducing forbidden regions.

BRIKARD employs a simple topological structure: a continuous chain, the "backbone" must trace the graph so it has at least one bond in common with each ring. This simple "interlaced braid" data structure was adequate for sampling all the topologies encountered in this study.

However, it is easy to extend this structure to allow either a branched backbone or rings with stems to the backbone, etc. The key challenge in modeling a multicyclic system with BRIKARD is the preprocessing step: generate a spanning tree and order the rings so each has at least six torsional DoF that have not been set by any previous ring closure. This is not always possible, and there are structures with enough DoF and flexibility that do not strictly fit this criterion (e.g., BPH is sampled by allowing flexibility in amide bonds). Additionally, such examples may be found among highly constrained, small multicyclics. We feel that such rigid structures may be modeled quite adequately by currently available techniques, which however tend to fail for systems of larger, flexible cycles where the kinematic methods in BRIKARD are ideally applicable.

5. DISCUSSION

Our method is the first approach to sampling multicyclic structures that harnesses the power of inverse kinematics to correctly build structural topologies. BRIKARD's novelty rests in the analytical closure of the macrocycle, enabling extremely rapid scanning of the complete potential energy landscape, even in cases of highly complex ring scaffolds.

One of the key strengths of such a fully algebraic approach is that it generates and explores all sterically feasible alternatives simultaneously for a given set of sampled torsions, which prevents entrapment of the entire search due to kinetic energy barriers. Purely geometric methods can miss certain solutions due to their iterative method of solving transcendental loop closure equations.

We found that BRIKARD substantially outperformed the existing best methods for sampling macrocycles; searches required minutes rather than hours or days, the conformations returned were not biased by the starting conformation, and near-native conformations were sampled for topologically complex, fused ring systems. Our results show that BRIKARD can successfully reproduce the known structures of macrocycles solved by NMR, small-molecule crystallography, as well as in complex with a protein structure. Furthermore, we have described the range of topological complexity in which reliable near-native conformations can be produced. Importantly, the experimental and low-energy states do not always correspond, nor do the unbound and bound experimental states. That is, it is not necessarily only the lowest-energy conformations that matter to the exploration of biochemically interesting events like target recognition, host–guest complexation, and permeability. BRIKARD samples these relevant conformations without incurring high computational cost.

Interest in use of macrocycles in drug discovery, catalysis, and other industries continues to grow, but until now there has not been a general method for performing rapid screening of macrocyclic libraries. Application of our search algorithm will provide distinct conformations that can be used as a basis for structure-based projects, such as macrocycle docking and optimization of physiochemical properties.

Further improvements are on the horizon, such as more sophisticated sampling strategies for the torsional DoF, which are now sampled uniformly for nonamino acids. It is well-known

that the sampling efficiency for single-loop systems may be critically enhanced by employing a Jacobian,[49] and we believe that homotopy-based topological clustering methods may also be used to offer additional order-of-magnitude speedups. Combined with the state-of-the-art algorithm for energy minimization we employ, we show that it is now possible to quickly predict the near-native ensemble for diverse, biologically relevant macrocycles, enabling realistic modeling of physiochemical properties and molecular recognition events.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank John K. Prentice for enthusiastic encouragement and motivation and Cristian Bologa for helpful suggestions. Molecular graphics and analyses were performed with the UCSF Chimera package. E.A.C. and M.J.W. were supported in part by NIH grant R01GM090205. K.W.L. thanks the NIGMS for support (GM100619). E.A.C. acknowledges support from the Laufer Center for Physical and Quantitative Biology at Stony Brook University. A portion of this work was carried out at the Center for Advanced Research Computing, University of New Mexico.

NOMENCLATURE

DG	distance geometry
DoF	degrees of freedom
IK	inverse kinematics
MD	molecular dynamics
RMSD	root mean square deviation

References

1. Bonnet P, Agrafiotis DK, Zhu F, Martin EJ. Conformational analysis of macrocycles: finding what common search methods miss. J Chem Inf Model. 2009; 49:2242. [PubMed: 19807090]
2. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins: Struct, Funct Genet. 2004; 55:351–367. [PubMed: 15048827]
3. Labute P. LowModeMD-implicit low-mode velocity filtering applied to conformational search of macrocycles and protein loops. J Chem Inf Model. 2010; 50:792. [PubMed: 20429574]
4. Watts KS, Dalal P, Tebben AJ, Cheney DL, Shelley JC. Macrocycle Conformational Sampling with MacroModel. J Chem Inf Model. 2014; 54:2680–96. [PubMed: 25233464]
5. Chen IJ, Foloppe N. Tackling the conformational sampling of larger flexible compounds and macrocycles in pharmacology and drug discovery. Bioorg Med Chem. 2013; 21:7898. [PubMed: 24184215]
6. Crippen, GM., Havel, TF. Distance Geometry and Molecular Conformation. Vol. 74. Research Studies Press; Taunton, UK: 1988.
7. Crippen, GM. Distance Geometry 2013: Theory, Methods and Applications. Mucherino, A.Lavor, C.Liberti, L., Maculan, N., editors. Springer; NY: 2013. p. 351-361.
8. Porta JM, Ros L, Thomas F, Corcho F, Canto J, Perez JJ. Complete Maps of Molecular Loop Conformational Spaces. J Comput Chem. 2008; 29:144–155.
9. Thorpe MF, Lei M. Macromolecular Flexibility. Philos Mag. 2004; 84:1323–331.

10. Brown WM, Martin S, Pollock SN, Coutsias EA, Watson JP. Algorithmic dimensionality reduction for molecular structure analysis. J Chem Phys. 2008; 129:064118. [PubMed: 18715062]
11. Martin S, Thompson A, Coutsias EA, Watson JP. Topology of Cyclooctane Energy Landscape. J Chem Phys. 2010; 132:234115. [PubMed: 20572697]
12. Hunt, KH. Kinematic Geometry of Mechanisms. Oxford University Press; Oxford, UK: 1990.
13. Sommese, AJ., Wampler, C. The Numerical Solution of Systems of Polynomials Arising in Engineering and Science. World Scientific; Hackensack, NJ: 2005.
14. Bricard, R. Leçons de Cinématique. Vol. 1–2. Gauthier–Villars; Paris, France: 1926.
15. Jacobson MP, Friesner RA, Xiang Z, Honig B. On the Role of the Crystal Environment in Determining Protein Side-chain Conformations. J Mol Biol. 2002; 320:597–608. [PubMed: 12096912]
16. Zhu K, Shirts MR, Friesner RA, Jacobson MP. Multiscale Optimization of a Truncated Newton Minimization Algorithm and Applications to Proteins and Protein-Ligand Complexes. J Chem Theory Comput. 2007; 3:640–648. [PubMed: 26637042]
17. Sillerud LO, Burks EJ, Wester MJ, Brown DC, Vijayan S, Larson RS. NMR-derived model of interconverting conformations of an ICAM-1 inhibitory cyclic nonapeptide. J Pept Res. 2003; 62:97–116. [PubMed: 12895272]
18. Sillerud LO, Burks EJ, Brown WM, Brown DC, Larson RS. NMR solution structure of a potent nonapeptide inhibitor of ICAM-1-mediated adhesion produced by homologous amino acid substitution. J Pept Res. 2004; 64:127–140. [PubMed: 15357668]
19. Khan AR, Parrish JC, Fraser ME, Smith WW, Bartlett PA, James MNG. Lowering the Entropic Barrier for Binding Conformationally Flexible Inhibitors to Enzymes. Biochemistry. 1998; 37:16839–16845. [PubMed: 9836576]
20. Mough ST, Goeltz JC, Holman KT. Isolation and Structure of an “Imploded” Cryptophane. Angew Chem, Int Ed. 2004; 43:5631–5635.
21. Angelini A, Cendron L, Chen S, Touati J, Winter G, Zanotti G, Heinis C. Bicyclic Peptide Inhibitor Reveals Large Contact Interface with a Protease Target. ACS Chem Biol. 2012; 7:817–821. [PubMed: 22304751]
22. Allen FH. The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Crystallogr, Sect B: Struct Sci. 2002; 58:380–388.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
24. Dunbrack RL Jr. Rotamer libraries in the 21st century. Curr Opin Struct Biol. 2002; 12:431–440. [PubMed: 12163064]
25. Ho BK, Coutsias EA, Seok C, Dill KA. The flexibility in the proline ring couples to the protein backbone. Protein Sci. 2005; 14:1011–1018. [PubMed: 15772308]
26. Lovell SC, Davis IW III, Arendall BW, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure Validation by Cα Geometry: ϕ, ψ and Cβ Deviation. Proteins: Struct, Funct Genet. 2003; 50:437–450. [PubMed: 12557186]
27. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods. 2009; 6:551–552. [PubMed: 19644455]
28. Lee J, Lee D, Park H, Coutsias EA, Seok C. Protein loop modeling by using fragment assembly and analytical loop closure. Proteins: Struct, Funct Genet. 2010; 78:3428–3436. [PubMed: 20872556]
29. Ko J, Lee D, Park H, Coutsias EA, Lee J, Seok C. The FALC-Loop web server for protein loop modeling. Nucleic Acids Res. 2011; 39:W210–W214. [PubMed: 21576220]
30. Park H, Lee GR, Heo L, Seok C. Protein Loop Modeling Using a New Hybrid Energy Function and Its Application to Modeling in Inaccurate Structural Environments. PLoS One. 2014; 9:e113811. [PubMed: 25419655]
31. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem. 2004; 25:1605–1612. [PubMed: 15264254]

32. Coutsias EA, Seok C, Dill KA. Using quaternions to calculate RMSD. J Comput Chem. 2004; 25:1849–1857. [PubMed: 15376254]
33. Go N, Scheraga HA. Ring Closure and Local Conformational Deformations of Chain Molecules. Macromolecules. 1970; 3:178–187.
34. Wedemeyer WJ, Scheraga HA. Exact Analytical Loop Closure in Proteins Using Polynomial Equations. J Comput Chem. 1999; 20:819–844.
35. Coutsias EA, Seok C, Jacobson MP, Dill KA. A kinematic view of loop closure. J Comput Chem. 2004; 25:510–528. [PubMed: 14735570]
36. Milgram RJ, Liu G, Latombe JC. On the Structure of the Inverse Kinematics Map of a Fragment of Protein Backbone. J Comput Chem. 2008; 29:50–68. [PubMed: 17542001]
37. Wu MG, Deem MW. Efficient Monte Carlo methods for cyclic peptides. Mol Phys. 1999; 97:559–580.
38. Wu MG, Deem MW. Analytical rebridging Monte Carlo: Application to cis/trans isomerization in proline-containing, cyclic peptides. J Chem Phys. 1999; 111:6625–6632.
39. Canutescu AA, Dunbrack RL Jr. Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci. 2003; 12:963–972. [PubMed: 12717019]
40. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YEA, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popović Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011; 487:545–574. [PubMed: 21187238]
41. Bourne Y, Radic Z, Aráoz R, Talley TT, Benoit E, Servent D, Taylor P, Molgó J, Marchot P. Structural determinants in phycotoxins and AChBP conferring high affinity binding and nicotinic AChR antagonism. Proc Natl Acad Sci U S A. 2010; 107:6076–6081. [PubMed: 20224036]
42. Husty ML, Pfurner M, Schröcker HP. A new and efficient algorithm for the inverse kinematics of the general serial 6R manipulator. Mech Mach Theory. 2007; 42:66–81.
43. Viquerat AD, Hutt T, Guest SD. A plane symmetric 6R foldable ring. Mech Mach Theory. 2013; 63:73–88.
44. Coutsias EA, Seok C, Wester MJ, Dill KA. Resultants and loop closure. Int J Quantum Chem. 2006; 106:176–189.
45. Manocha D, Canny JF. Efficient Inverse Kinematics for General 6R Manipulators. IEEE T Robotic Autom. 1994; 10:648–657.
46. Porta JM, Jaillet L. Exploring the Energy Landscapes of Flexible Molecular Loops using Higher-Dimensional Continuation. J Comput Chem. 2013; 34:234–244. [PubMed: 23015474]
47. Dodd LR, Boone TD, Theodorou DN. A Concerted Rotation Algorithm for Atomistic Monte-Carlo Simulation of Polymer Melts and Glasses. Mol Phys. 1993; 78:961–996.
48. Ulmschneider JP, Jorgensen WL. Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias. J Chem Phys. 2003; 118:4261–4271.
49. Nilmeier J, Hua L, Coutsias EA, Jacobson MP. Assessing protein loop flexibility by hierarchical Monte Carlo sampling. J Chem Theory Comput. 2011; 7:1564–1574. [PubMed: 21743800]
50. Lee HY, Liang CG. Displacement analysis of the general spatial 7-link 7R mechanism. Mech Mach Theory. 1988; 23:219–226.

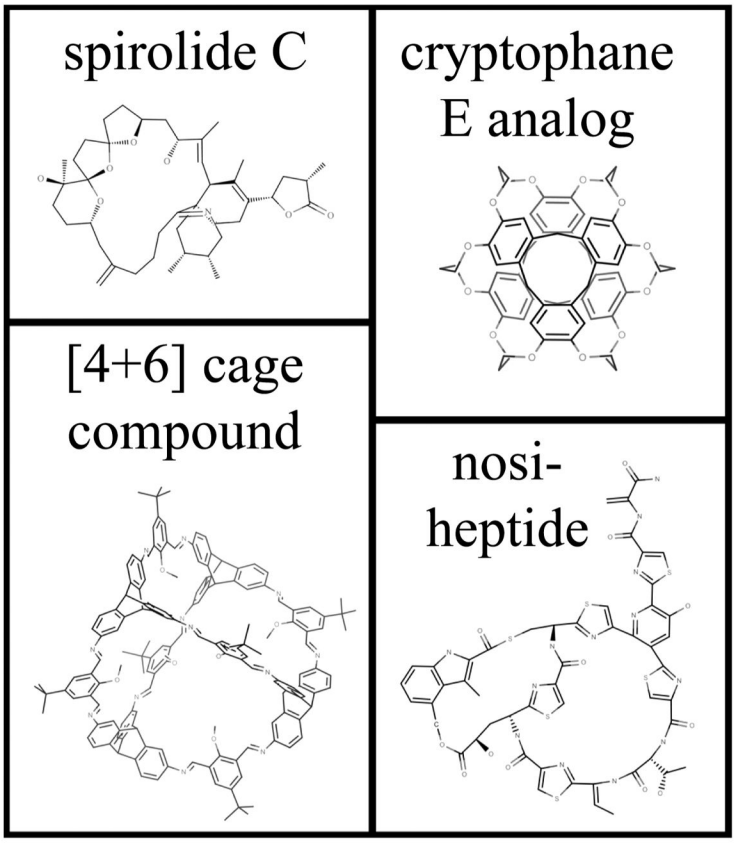


Figure 1.

A subset of the macrocycles contained in our study set, depicting the vast diversity in degrees of freedom and ring topology. Starting at the upper left and proceeding clockwise, these macrocycles are identified in the supplementary tables by the abbreviations WZY, CD4, NOS, and MST, respectively.

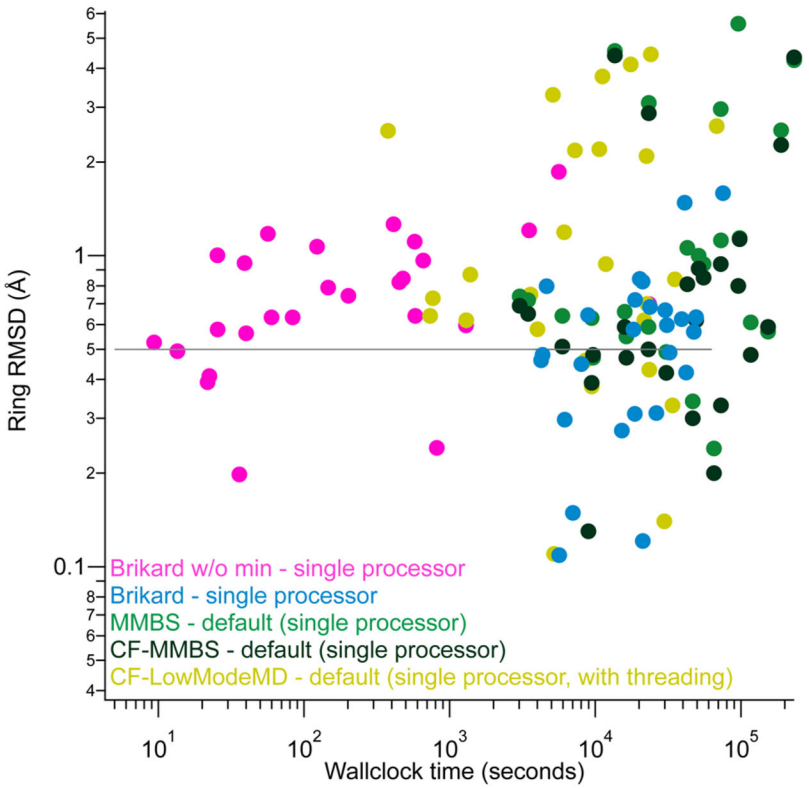


Figure 2.

RMSD of ring atoms to crystallographic coordinates vs run time according to different sampling methods. Each method was run on a single processor to enable comparison with default sampling protocols. CF-LowModeMD was run with multithreading enabled, which is the program default. Without this option, runs take 1.4–4 times more wallclock time.

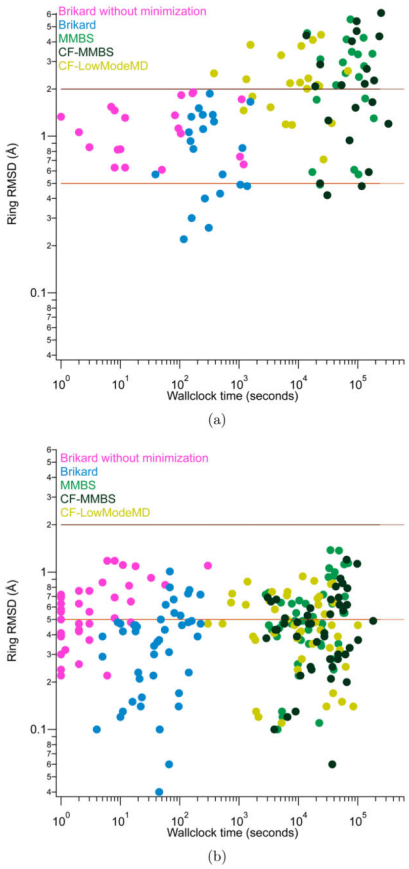


Figure 3.

RMSD of ring atoms to crystallographic coordinates vs run time for (a) hard examples in the data set and (b) easy examples in the data set according to different sampling methods. See the text for the meanings of “hard” and “easy” in this context.

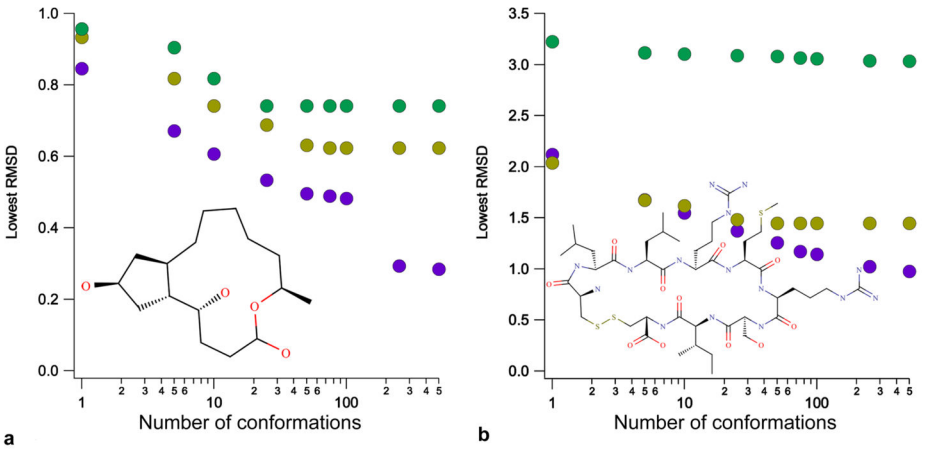


Figure 4.

Illustrative examples of subsampling efficiency for BRIKARD (purple), MMBS (green), and CF-LowModeMD (yellow) given a) simple macrocycle AFB and b) large cyclic peptide IP01.

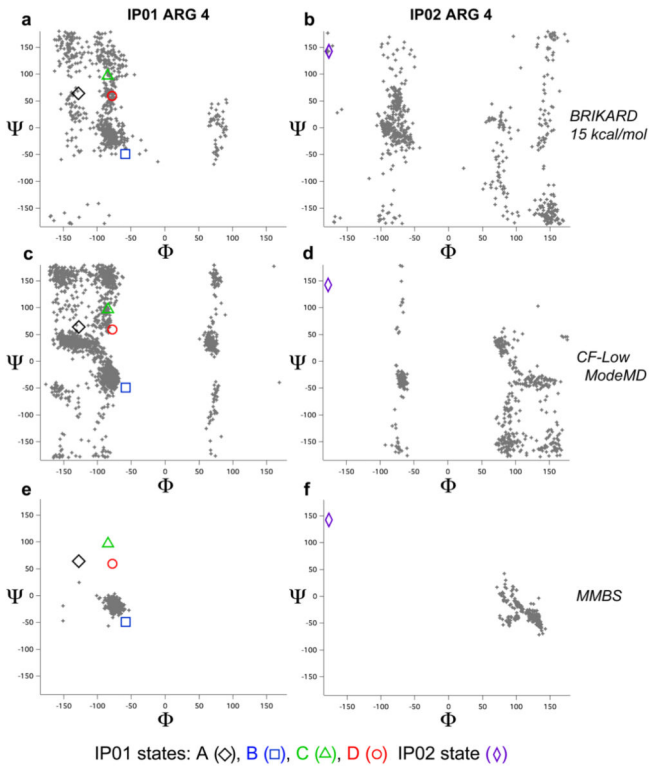


Figure 5.

Arginine 4 torsional angle sampling across the potential energy surfaces of IP01 and IP02 by a–b) BRIKARD restricted to the conformations within 15 kcal/mol of the lowest energy identified, c–d) CF-LowModeMD, and e–f) MMBS.

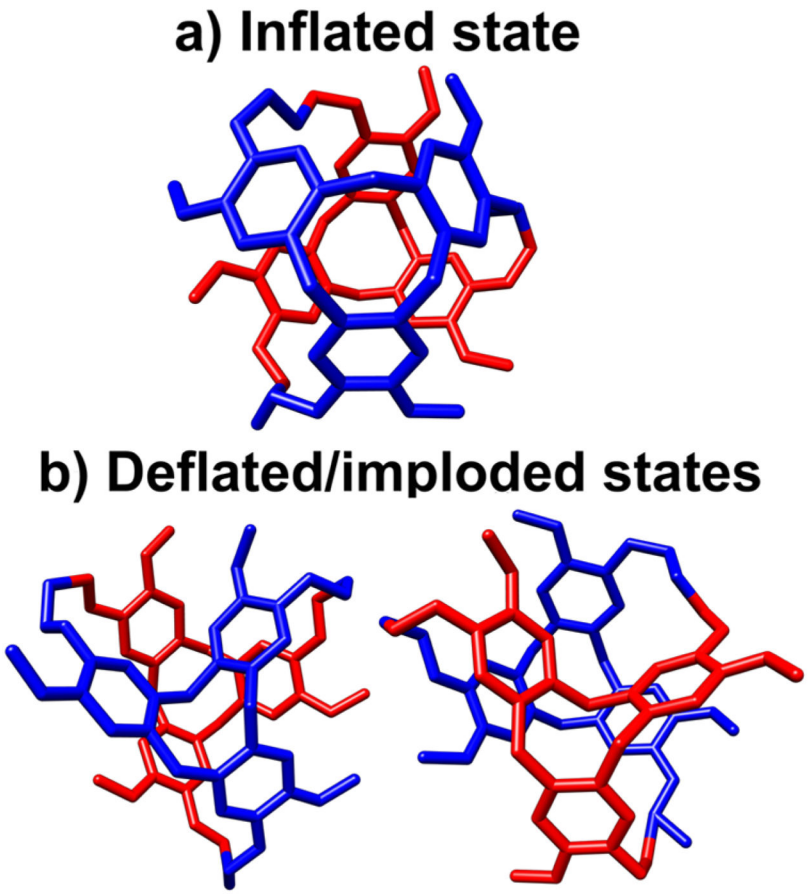


Figure 6.
a) Expanded (cargo-carrying) and b) deflated/imploded (apo) major states of cryptophane E analogue (CD4) present in the crystallographic ensemble.

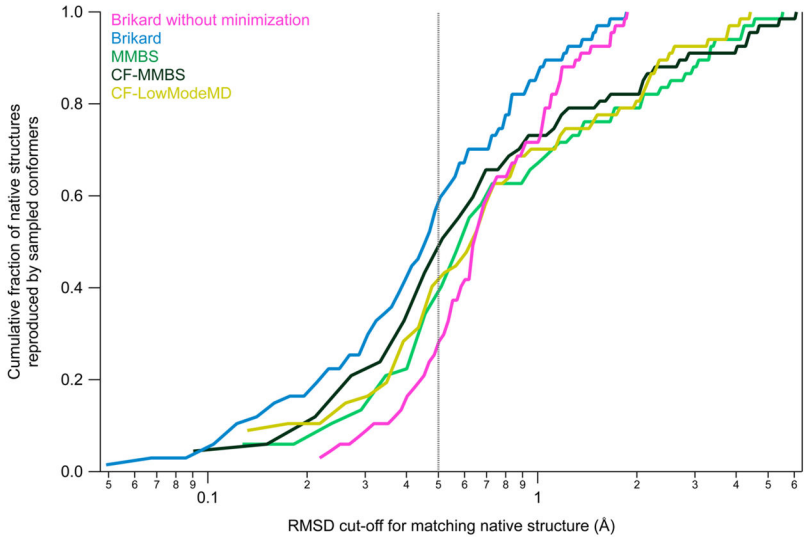


Figure 7.

Frequency of success for each method in sampling the near-native conformation of each structure. The desired RMSD of 0.5 Å to the experimental structure is delineated by a gray dashed line.

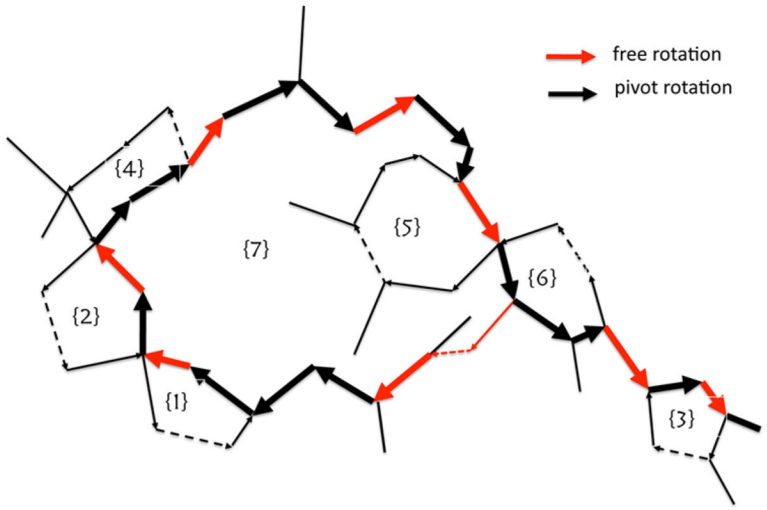


Figure 8.
Depiction of how the ring topology drives ring closure using 13-desmethyl spirolide C (SQX) as an example.

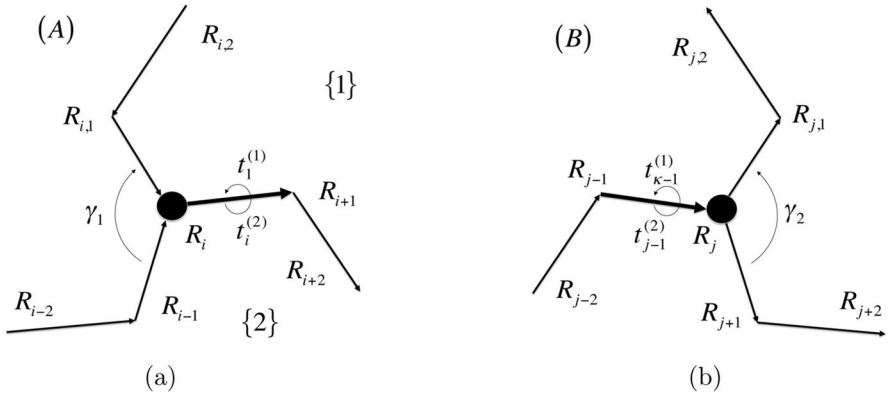


Figure 9.
Torsions at branch points (a), (b).

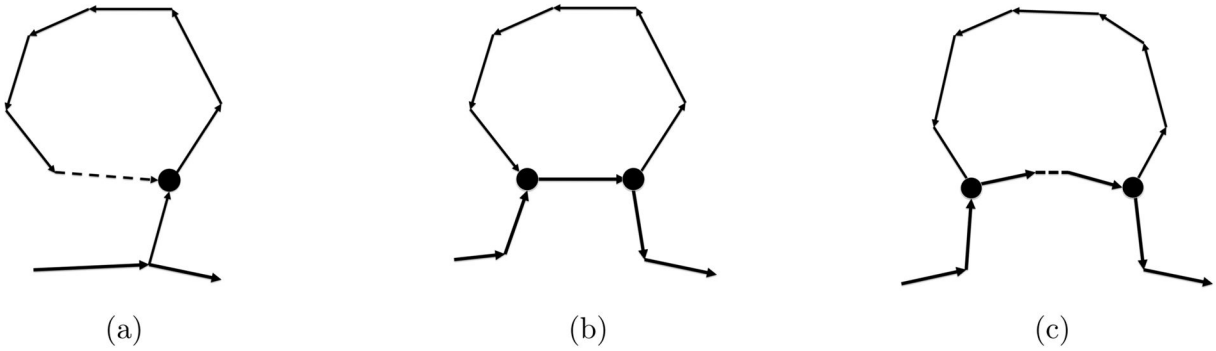


Figure 10.
Ring attachment (a) on a side chain, (b) on backbone, one common bond, and (c) on backbone, several common bonds.

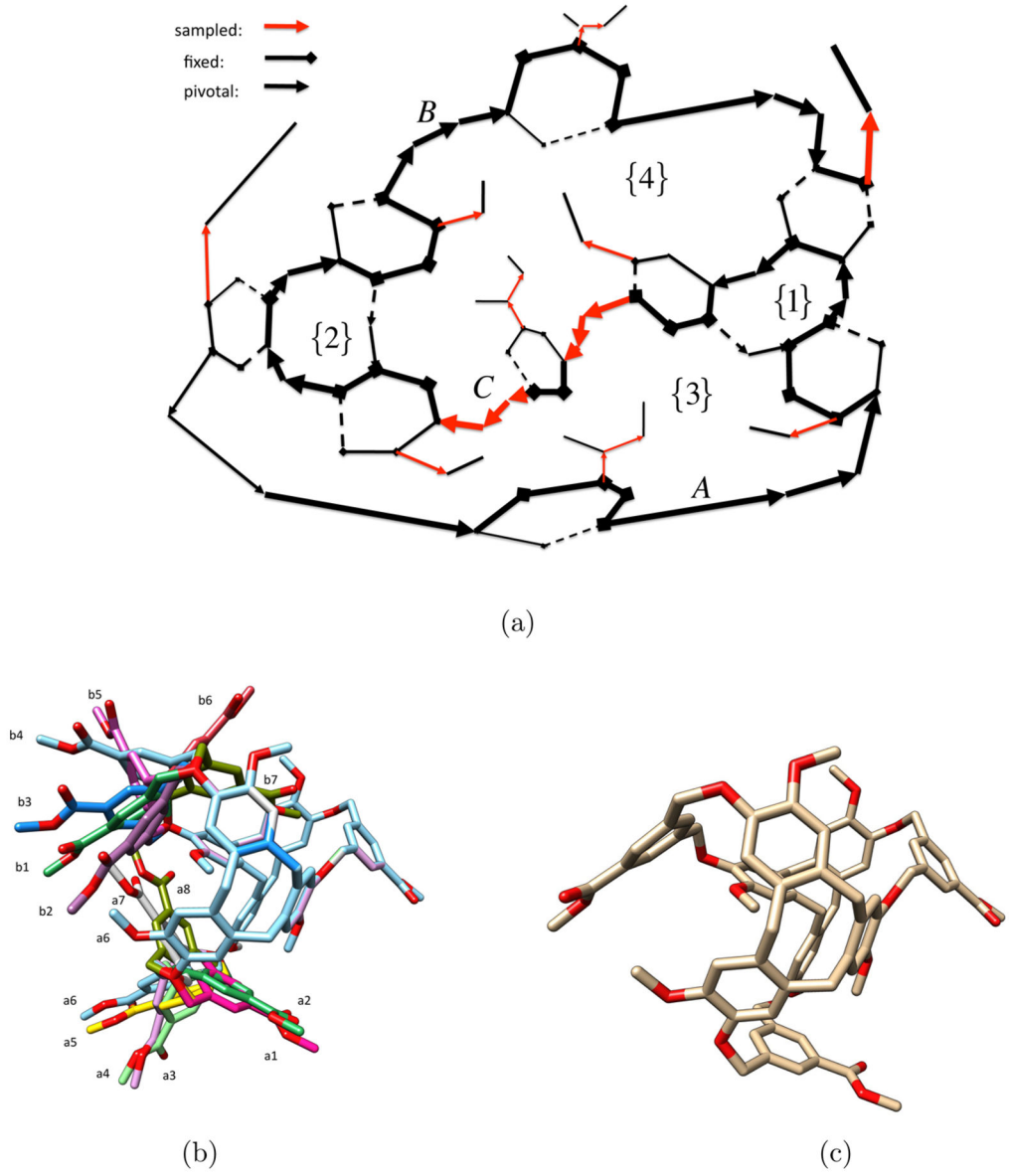


Figure 11.

BIMXUR (a) Graph depiction. Red arrows: sampled torsions. Black arrows: torsions set by ring closure. Rings numbered in the order by which they are solved. (b) Ensemble of alternative shapes keeping chain C and rings 1 and 2 fixed. (c) Crystal structure, corresponding to (a1, b1) in ensemble.

Table 1

Sampling Performance of BRIKARD, with and without Minimization Compared with CF-LowModeMD, MMBS, and CF-MMBS Using the Energy Cutoff Implemented for Each Method Previously[a]

	CF-LowModeMD	MMBS	CF-MMBS	BRIKARD	BRIKARD no Min
1.0 Å	71.6%	67.2%	74.6%	85.1%	71.6%
0.5 Å	40.3%	38.8%	44.8%	58.2%	25.4%
median calculation time (h)	3.067	6.850	10.340	0.116	0.003
standard deviation	5.78	11.52	18.17	0.41	0.33

[a]Accuracy assessed by ring RMSD to the experimental structure; calculation time in wallclock hours.

Table 2

Hard Cases: Ring RMSD to Experimental Structure of Compounds for Which at Least One Method Failed To Achieve an RMSD Less than 1.5 Å[a]

compd	CF-LowModeMD	MMBS	CF-MMBS	BRIKARD	BRIKARD w/o min
1P2	**2.20**	0.86	0.80	0.83	1.15
AAS	**2.20**	0.49	0.42	0.84	1.06
ACT	**2.09**	**2.96**	0.94	1.20	**1.67**
AP2	**3.29**	0.57	0.59	0.40	0.63
BPH	**4.12**	0.59	0.50	0.30	0.32
CD4	**4.44**	**4.55**	**4.40**	0.33	0.67
CRP	**2.61**	**3.10**	**2.87**	0.27	0.61
CSK	**2.52**	0.61	0.48	1.00	1.02
DAP	**2.18**	**4.25**	**4.34**	1.48	**1.67**
HAX	1.19	**2.53**	**2.27**	0.94	1.17
IP1	0.73	**3.03**	**4.23**	0.83	0.83
IP2	1.21	**4.15**	**4.04**	0.93	0.82
KCR	**2.01**	**2.10**	**2.08**	0.57	0.63
MST	**1.53**	1.30	1.20	0.43	1.31
QN7	**3.83**	**3.36**	**6.12**	**1.66**	**1.72**
RFP	0.71	**1.71**	1.26	1.06	1.04
RH9	**1.78**	**2.80**	**1.52**	1.24	1.36
SP1	**2.34**	**2.39**	**2.16**	**1.51**	1.46
SWI	1.46	**1.74**	**1.65**	1.37	**1.83**
TET	**2.31**	**2.12**	**2.12**	**1.87**	**1.88**
TIO	**2.13**	**3.33**	**2.69**	0.26	0.63

[a]Boldface entries.

J Chem Theory Comput. Author manuscript; available in PMC 2017 September 13.