UCLA

UCLA Previously Published Works

Title

Computational generation of an annotated gigalibrary of synthesizable, composite peptidic macrocycles

Permalink

https://escholarship.org/uc/item/9dk4n1d7

Journal

Proceedings of the National Academy of Sciences of the United States of America, 117(40)

ISSN

0027-8424

Authors

Saha, Ishika
Dang, Eric K
Svatunek, Dennis
et al.

Publication Date

2020-10-06

DOI

10.1073/pnas.2007304117

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Computational generation of an annotated gigalibrary of synthesizable, composite peptidic macrocycles

Ishika Saha[a,1], Eric K. Dang[b,1], Dennis Svatunek[a], Kendall N. Houk[a,2], and Patrick G. Harran[a,2]

[a]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095; and [b]Department of Computer Science, University of California, Los Angeles, CA 90095

Contributed by Kendall N. Houk, July 22, 2020 (sent for review April 17, 2020; reviewed by Abigail Doyle, William A. Goddard III, and William L. Jørgensen)

Peptidomimetic macrocycles have the potential to regulate challenging therapeutic targets. Structures of this type having precise shapes and drug-like character are particularly coveted, but are relatively difficult to synthesize. Our laboratory has developed robust methods that integrate small-peptide units into designed scaffolds. These methods create macrocycles and embed condensed heterocycles to diversify outcomes and improve pharmacological properties. The hypothetical scope of the methodology is vast and far outpaces the capacity of our experimental format. We now describe a computational rendering of our methodology that creates an in silico three-dimensional library of composite peptidic macrocycles. Our open-source platform, CPMG (Composite Peptide Macrocycle Generator), has algorithmically generated a library of 2,020,794,198 macrocycles that can result from the multistep reaction sequences we have developed. Structures are generated based on predicted site reactivity and filtered on the basis of physical and three-dimensional properties to identify maximally diverse compounds for prioritization. For conformational analyses, we also introduce ConfBuster++, an RDKit port of the open-source software ConfBuster, which allows facile integration with CPMG and ready parallelization for better scalability. Our approach deeply probes ligand space accessible via our synthetic methodology and provides a resource for large-scale virtual screening.

macrocyclic peptides | reaction product prediction | conformational analysis

Macrocyclic compounds have been identified as enzyme inhibitors, as GPCR (G protein-coupled receptor) agonists and antagonists, inhibitors of protein–protein interactions, and as modulators of various other biological pathways. The ring structures of macrocycles contribute to structural preorganization, lowering entropic penalties for binding to biological targets (1–3). Peptide-derived macrocycles are especially attractive for targeting protein surfaces because the embedded peptide can mimic native protein structure and recognition elements (4, 5). The advent of several high-throughput biosynthetic platforms has transformed macrocyclic peptides into a powerful ligand discovery modality (6–10). However, key challenges remain (11). In the forty years since the discovery of cyclosporine as a membrane-permeable and orally bioavailable drug, efforts to anticipate related structures with comparable properties have had limited success (12). Nevertheless, some general trends have emerged. Empirical analysis of data suggests macrocycles having molecular weight (MW) < 1,000 Da, total polar surface area (TPSA) < 250 Å^2, *c*logP < 10, and fewer than five hydrogen bond donors are more likely to be bioavailable. Molecular shape and conformational dynamics have also been cited as important factors for achieving this end (2,12–14).

Our laboratory seeks to identify structural settings in which macrocycles can retain ancillary polar groups yet achieve a useful balance of cell permeability and aqueous solubility (Fig. 1). We have designed scaffolding reagents that are easily integrated into peptide structure to afford diverse ring connectivities and embedded heterocyclic motifs (15–20). These structural features have been shown to improve target binding, the ability to passively transverse lipid membranes and resist proteolytic degradation (12). Our methodology uses multiply reactive templates, **G1**–**G3**, which are activated in stages to react with unprotected polyamides to form macrocyclic composites (Fig. 1). The methodology was designed to allow systematic alteration of product topology and properties by engaging a broad range of native peptide functional groups in carbon–heteroatom and carbon–carbon bond-forming reactions. Our experimental studies have demonstrated that templates **G1**–**G3** engage aromatic side chains (including but not limited to phenol, indole, and imidazole) on the polyamides to participate in Friedel Crafts alkylation, metal-catalyzed allylic substitutions (also known as Tsuji–Trost reactions), and *N*-acyliminium ion-mediated cyclizations (15–20). For example, macrocyclization under Pd^0-catalysis affords C-O or C-N bonded macrocycles, wherein chemoselectivity is switchable by the addition of Cs_2CO_3 (**1** or **2**). Macrocyclization under acidic conditions generates C-C bonded products (**3**–**9**) via electrophilic aromatic substitution (EAS), many of which also incorporate polycyclic motifs via sequential, diastereoselective *N*-acyliminium ion cyclization of the P1 side chain along with macrocyclization (**4**–**6**, **9**). **G3** is able to itself participate in *N*-acyliminium ion-promoted EAS reactions in the absence of an aromatic side chain at P1 to yield structures such as **7**–**8**. Reaction of **G1** and **G2** with Trp-Trp-Tyr produces *N*-acyliminium ion-mediated bridged endopyrroloindolines **3** and **5**.

Significance

We describe computations to anticipate products of multistep reaction sequences. The work is based on experimental methods developed earlier to amalgamate synthetic scaffolding reagents with small linear peptides. Hybrid products retain molecular recognition elements in the peptide, but display that functionality as part of amphipathic macrocycles having defined conformations and improved pharmacological properties. The hypothetical scope of the chemistry is large and far outpaces the experimental format. To explore the structure space more extensively, we devised algorithms to predict outcomes of more than 2 billion processing sequences. Software was also developed to generate accurate three-dimensional structures for each product. The resultant virtual library is a resource that can be deployed broadly in search of novel ligands for protein receptors.

Author contributions: K.N.H. and P.G.H. designed research; I.S., E.K.D., D.S., and P.G.H. performed research; E.K.D. contributed new reagents/analytic tools; I.S., E.K.D., D.S., K.N.H., and P.G.H. analyzed data; and I.S., E.K.D., K.N.H., and P.G.H. wrote the paper.

Reviewers: A.D., Princeton University; W.A.G., California Institute of Technology; and W.L.J., Yale University.

The authors declare no competing interest.

Published under the PNAS license.

[1]I.S. and E.K.D. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: houk@chem.ucla.edu or harran@chem.ucla.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2007304117/-/DCSupplemental.

First published September 18, 2020.

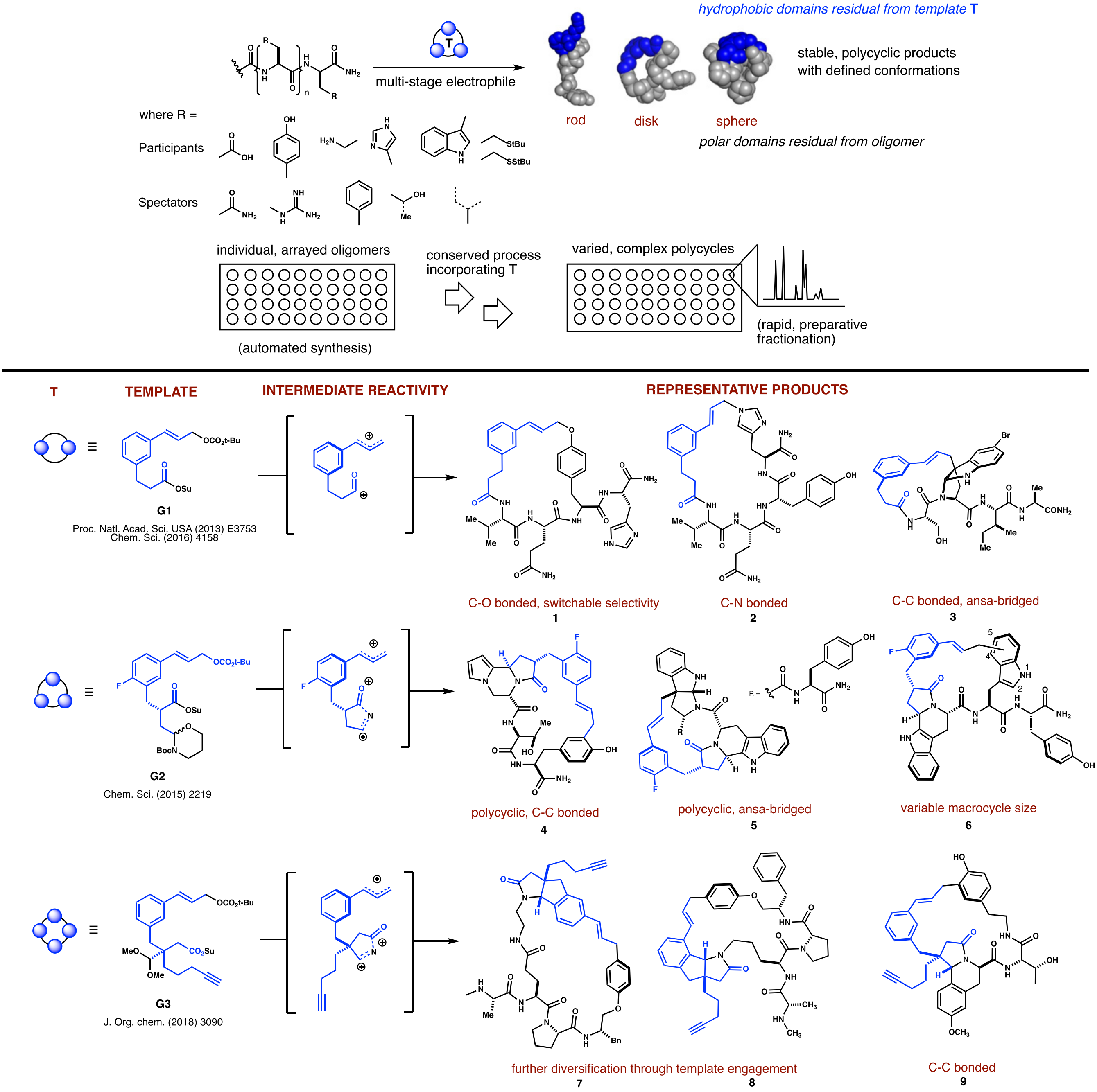

hydrophobic domains residual from template T
multi-stage electrophile
rod
disk
sphere
stable, polycyclic products with defined conformations
polar domains residual from oligomer
where R =
Participants
Spectators
individual, arrayed oligomers
(automated synthesis)
conserved process incorporating T
varied, complex polycycles
(rapid, preparative fractionation)
T
TEMPLATE
INTERMEDIATE REACTIVITY
REPRESENTATIVE PRODUCTS
G1
Proc. Natl. Acad. Sci. USA (2013) E3753
Chem. Sci. (2016) 4158
C-O bonded, switchable selectivity
1
C-N bonded
2
C-C bonded, ansa-bridged
3
G2
Chem. Sci. (2015) 2219
polycyclic, C-C bonded
4
polycyclic, ansa-bridged
5
variable macrocycle size
6
G3
J. Org. chem. (2018) 3090
further diversification through template engagement
7
8
C-C bonded
9

Fig. 1. Generalized schematic and representative outcomes of our synthesis platform. Templates indicated in blue and oligomers indicated in gray.

Within the well-established reactivity patterns exemplified by **G1–G3**, we envisioned a platform, wherein the methodology could be extended to a much broader range of oligomers. These would be assembled not only from α-amino acids, but also β2- and β3-amino acids in both enantiomeric forms. The side chains on these monomers could conceivably harbor diverse, drug-like heterocycles, chosen for their susceptibility to engagements by **G1–G3**. Such alterations could provide property advantages over products of biosynthetic methods which largely arise from natural amino acids. This thought experiment presented a unique challenge—the possible outcomes far outpaced our experimental capabilities. We therefore turned to a computational rendering of our synthesis platform (Fig. 2) to systematically assess the scope of reaction outcomes. Herein we describe our program, CPMG (Composite Peptide Macrocycle Generator), which we have used to generate an in silico library of >2 billion composite macrocycles resultant from multistep reaction sequences. We have also adapted conformational search methods (21) to create Confbuster++, which is able to generate three-dimensional (3D) conformations of library members rapidly. The ultimate goal for this work is to enable virtual ligand discovery for diverse structurally characterized protein targets.

Results and Discussion

Our computational platform consists of two components: 1) CPMG, for generating a library of two-dimensional (2D) macrocycle structures from a set of user-defined building blocks, and

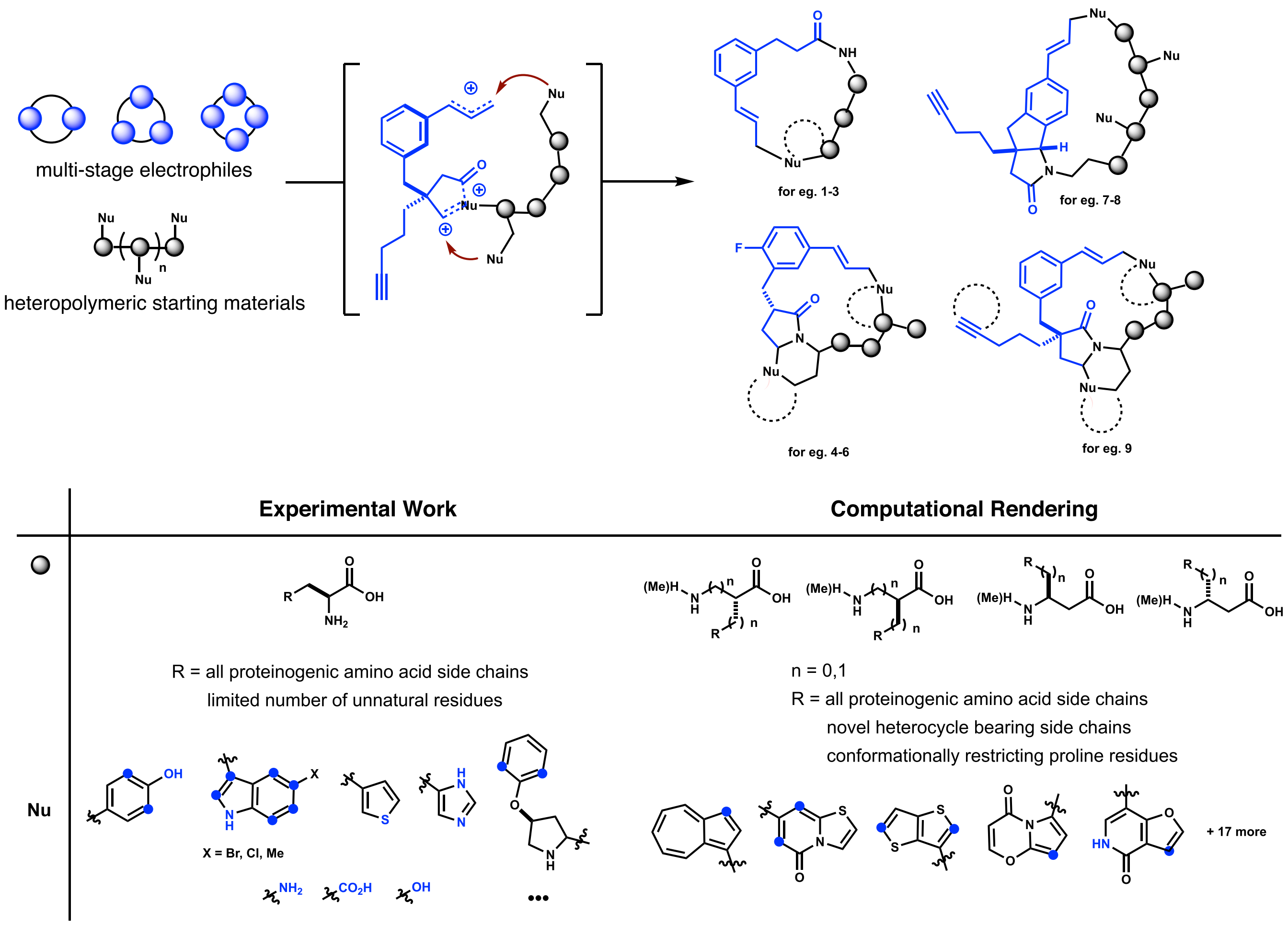

multi-stage electrophiles
Nu
Nu
n
Nu
heteropolymeric starting materials
⊕
Nu
O
⊕
Nu
⊕
Nu
O
NH
Nu
for eg. 1-3
Nu
Nu
Nu
H
N
O
for eg. 7-8
F
O
Nu
N
Nu
for eg. 4-6
O
Nu
N
Nu
for eg. 9
Experimental Work
Computational Rendering
R
O
OH
NH_2
R = all proteinogenic amino acid side chains
limited number of unnatural residues
(Me)H
N
H
n
R
n
O
OH
n = 0,1
R = all proteinogenic amino acid side chains
novel heterocycle bearing side chains
conformationally restricting proline residues
Nu
OH
X
N
H
X = Br, Cl, Me
S
H
N
N
O
N
H
NH_2
CO_2H
OH
•••
S
N
O
O
S
S
O
N
O
O
HN
O
O
+ 17 more

Fig. 2. A computational rendering of our synthesis platform.

2) ConfBuster++, for generating the conformers for each macrocycle. Both of these components are written in Python 3.6.8 and rely primarily on the open-source framework RDKit (22) (see *Methods* for details).

The platform is described in detail in the following sections; for a general workflow, see Fig. 4. In brief, our library is generated using a pool of amino acid derivatives incorporating several drug-like and conformationally restricting motifs (Fig. 3). These building blocks are systematically permuted by CPMG to generate linear oligopeptides that are subsequently bound to templates **G1–G3**. Template-bound sequences are then converted to macrocyclic structures based on rules derived from experimental observations and calculations for site reactivity. The macrocycles are finally filtered according to property criteria and analyzed using Confbuster++ for assessments of shape diversity (Fig. 4).

Library Generation. Heterocycles **9–25** in Fig. 3 were extracted from the VEHICLe (virtual exploratory heterocyclic library) database built by Pitt et al. (23) This database contains a set of 24,847 aromatic ring systems generated using a random forest-based method, of which over 3,000 ring systems were predicted by a decision tree method to be synthetically tractable. Some of these motifs have been experimentally synthesized since the time of publication, but many remain hypothetical. In choosing our heterocycles, we arranged VEHICLe by the number of hits the structures generated in the Beilstein database, based upon their incorporation into drug-like molecules. Heterocycles with a nitrogen-centered lone pair oriented orthogonal to the π plane of the aromatic system were not considered based on the assumption these would protonate under acidic conditions and resist EAS—consistent with our experimental observations. Since many of the chosen VEHICLe heterocycles were not characterized in the literature, DFT (density-functional theory) optimizations using ωB97X-D/def2TZVP (Gaussian 16 RevA.03) (24, 25) were performed on all structures having multiple tautomeric forms. The most energetically favorable tautomeric state was included in subsequent calculations. The set of VEHICLe heterocycles were supplemented with known motifs **26–30** based upon their likely participation in EAS reactions.

Using RDKit, each heterocycle in the final pool was formulated into amino acids by first attaching alkyl linkers (methylene, ethylene) to the atoms highlighted in green in Fig. 3. These congeners were subsequently formulated into L-α-, L-β^2-, and L-β^3-amino acids and their corresponding enantiomers. Proteinogenic amino acids (both D and L forms), along with a set of known, conformationally restricting proline analogs **31–40** (see *SI Appendix*, Fig. S1 for full list), were added to the set of building blocks. Peptides of varying length (3–5 monomer units) were produced from the Cartesian product of the set with the rule that each must contain at least one cyclization-competent nucleophile. Trimers were allowed to harbor at most two heterocyclic side chains, and tetra- and pentamers were allowed to contain no more than three heterocyclic side chains. Additionally, the C-terminal carboxyl group of all trimers and tetramers was capped with an *N*-ethyl-R unit (R = **1–30** in Fig. 3).

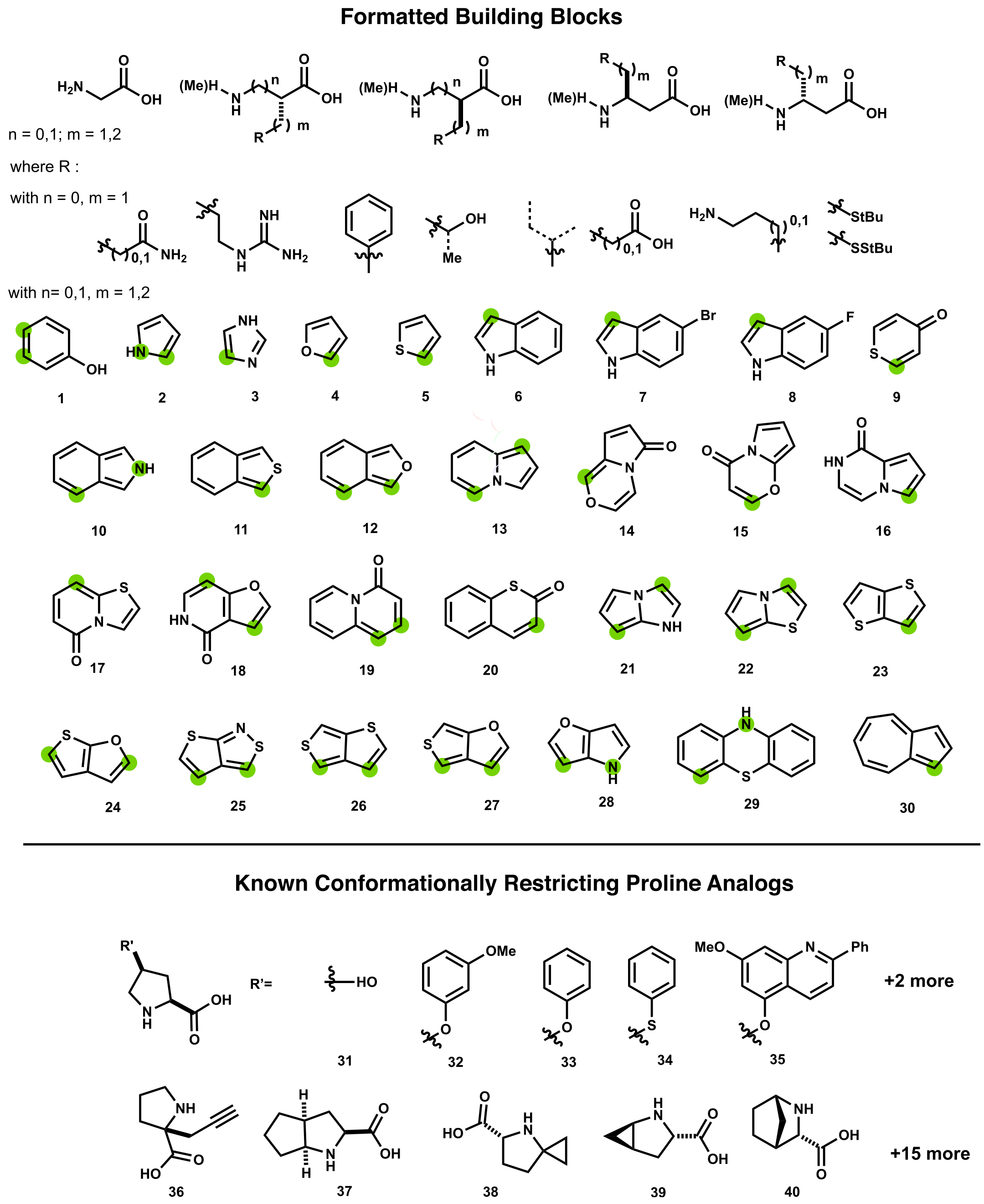


Fig. 3. Substrate library (see *SI Appendix* for full list). Green dots indicate points of incorporation into amino acids.

Pentamers were excluded from this last step in anticipation of molecular weight cutoffs.

The macrocycle enumeration step in CPMG replicates our synthetic processing sequences whereby macrocyclic composites are afforded from Friedel Crafts alkylations, metal-catalyzed allylic substitutions (Tsuji–Trost), and *N*-acyliminium ion-mediated cyclizations. In enumerating macrocyclic products, CPMG first adds templates **G1**–**G3** to each oligopeptide in the substrate library, wherein the templates are atom-mapped to form two bonds with the peptides—1) an amide bond with the peptide N-terminus or amine-bearing side chain and 2) a bond between the distal cinnamyl carbon atom and a nucleophilic side chain. Experimentally—the latter may be formed via Friedel Crafts alkylation or Tsuji–Trost substitution, and the linkage at the amide bond can be further diversified into condensed ring polycycles via *N*-acyliminium ion-mediated cyclizations when using templates **G2** and **G3**.

The in silico enumeration of macrocycles is simplified by the predictable nature of our incremental synthesis. For instance, all C-C bond formations depend entirely on EAS reactivity. The open-source program, RegioSQM, published by Jørgensen et al., is able to identify the most reactive nucleophilic atoms in heterocycles by systematic simulated protonation of all carbon atoms followed by a comparison of the relative free energies of incipient ionic states (26). We analyzed heterocycles **9**–**31** using RegioSQM, and results were incorporated into CPMG as site selectivity predictors. Since regioisomeric outcomes are experimentally observed in oligopeptides having multiple reactive sites, all such outcomes are also allowed in CPMG. Generated data were fully consistent

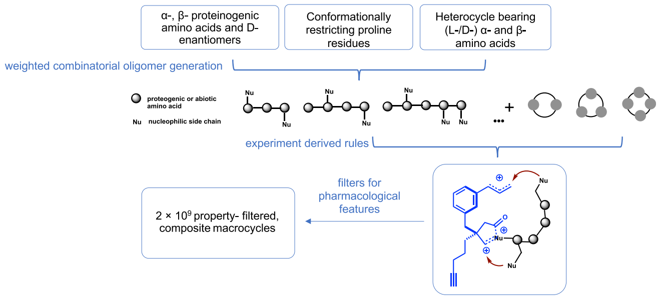

α-, β- proteinogenic amino acids and D-enantiomers
Conformationally restricting proline residues
Heterocycle bearing (L-/D-) α- and β-amino acids
weighted combinatorial oligomer generation
proteogenic or abiotic amino acid
Nu nucleophilic side chain
experiment derived rules
filters for pharmacological features
2 × 10⁹ property- filtered, composite macrocycles

Fig. 4. Schematized workflow for CPMG.

with experimental observations where available. Pictet–Spengler products were generated by allowing bond formations at RegioSQM-predicted nucleophilic sites exactly six atoms away in **G2** (e.g., **4**–**6** in Fig. 1), or five or six atoms away in **G3** (e.g., **7**–**8** or **9**, respectively, in Fig. 1), from the α-carbon of intermediate *N*-acyl iminium ions. In situations where *N*-acyliminium ion capture is not possible in **G2**, we allow formation of unsaturated *N*-alkyl pyrrolidinones. This is not implemented in **G3** as the quaternary center prevents formation of the same. Rules were also encoded for indole-containing oligopeptides to reflect experimental outcomes (e.g., **4** in Fig. 1), wherein *ansa*-bridged pyrroloindolines are formed through main-chain capture of indolenium ion intermediates formed by cinnamylation at the indole 3-position.

Since RegioSQM does not extend to predicting sites of heteroatomic nucleophilicity, compounds containing heteroatomic linkages resultant from allylic substitutions were generated on the basis of computed pK_a values, The underlying assumption is that a nucleophile attached to a sufficiently acidic hydrogen will participate effectively in the Tsuji–Trost catalytic cycle, consistent with literature precedent and in-house data (27). Heterocycles in the substrate library having heteroatom-bound hydrogen atoms were analyzed using the Jaguar module of Schrödinger Maestro in H_2O at *pH* 7.4 (28). Computed values were incorporated into CPMG such that heteroatom-bound hydrogens with $pK_a \leq 13.5$ (in H_2O) were allowed to bond to the distal cinnamyl carbon atom in **G1**–**G3**. To increase the probability of the library being populated with structures having useful pharmacological properties, the initial collection was filtered using guidelines advocated for achieving cellular permeability and oral bioavailability (2,12–14). Scanning mono-*N*-methylation of secondary amides was applied to all generated macrocycles, and only structures having MW $\leq$ 1,200 g mol^{-1}, number of rotatable bonds $\leq$ 10, and TPSA $\leq$ 200 Å^2 were retained.

Conformational Search: ConfBuster++. The identification of shape-diverse molecules necessitated the employment of a rapid conformational search algorithm. The use of traditional conformational sampling algorithms for large-scale macrocyclic conformer generation is challenging due to the torsional complexity of macrocyclic ring architectures. Native RDKit conformer search methods, as demonstrated by Ebejer, have been shown to quickly produce reasonable 3D structures (29). However, the authors note that these are not well suited to macrocycles. Moreover, the native RDKit conformer search methods treat alkenes as isomerizable units and are biased toward generating *cis*-alkenes when applied to cyclic alkenes, regardless of the input 2D structure. This is problematic when applied to our library as experimentally we only observe *trans*-alkenes in the macrocycles generated from **G1**–**G3**. To overcome this deficiency, we implemented the filtration method outlined by Landrum (30) for filtering out 3D structures with inverted double-bond stereochemistries. While this worked for most structures, it failed to produce any conformers for a subset of macrocycles despite several attempts with various native RDKit embedding methods (ETKDGv2 and random coordinate generation). We subsequently shifted our focus to the following family of conformational search algorithms.

An algorithm introduced by Jacobson et al. (31) for predicting protein loop structure has shown success in generating low-energy macrocycle conformers by a different analysis. This method involves cleavage of macrocycle rings into an acyclic form, conformational searches, and then systematic sampling of dihedral angles in order to permit reformation of the bond that was cleaved. Any resulting conformers that have atom clashes or torsions in non-allowed Ramachandran regions are filtered out for cyclic proteins. Thereafter, side-chain conformations are optimized by a similar process before a final energy minimization step. This general conformational sampling method has been adapted into two other variants: Posy et al.'s proprietary MacroModel (32) and Barbeau et al.'s open-source ConfBuster (21). In order to maintain continuity with CPMG, we have ported ConfBuster to an RDKit-based implementation, which we call ConfBuster++. The RDKit implementation of ConfBuster++ allows us to more easily maintain data associated with our macrocycles through the conformational sampling stage, as well as easier parallelization on our cluster. A detailed discussion of the algorithm is included under *Methods*.

To demonstrate the ability of ConfBuster++ to generate low-energy macrocycle conformers it was compared to published conformation search methods (33–37). The two cyclopeptides, cyclo-(Pro-Ser-leu-Asp-Val) and cyclo-(Arg-Gly-Asp-phe-([*N*-Me] Val)) (also known as cilengitide), were used as model systems to maintain consistency with published data (33–36). A tool for an extensive conformer search is CREST (37). This software package provides a conformer search at a higher level of theory using the GFN2-xTB tight-binding DFT functional (38) in combination with a metadynamic sampling and genetic *z*-matrix crossing approach. ConfBuster++ was able to find macrocyclic conformers for both model systems with only small deviation in the backbone when compared to CREST. Overlays of optimized lowest-energy conformers generated by CREST and ConfBuster++ for both model systems are shown in Fig. 5. The backbone rmsd for cyclo-(Pro-Ser-leu-Asp-Val) and cyclo-(Arg-Gly-Asp-phe-([*N*-Me]Val)) are 0.43 and 0.21 Å, respectively. We have further demonstrated the ability of ConfBuster++ to generate correct backbone conformations in comparison to another published method based on molecular mechanic force fields (36), wherein ConfBuster++ is much less computationally expensive (*SI Appendix*, Fig. S4).

Execution times for ConfBuster++ were benchmarked on University of California, Los Angeles' computer cluster using a subset of 10^6 macrocycles randomly selected from our library. The conformational sampling was done in parallel using a job array of 4,500 jobs, where each job was allocated a parallel

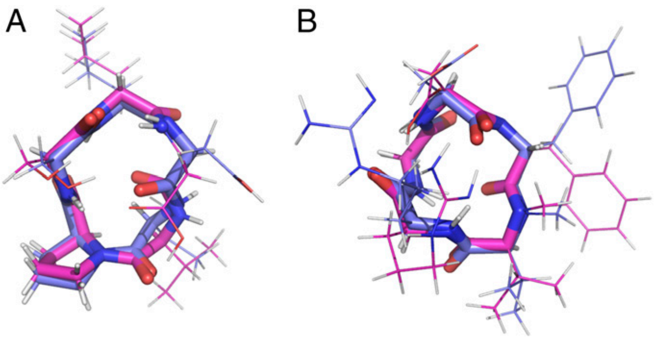


Fig. 5. Overlay of the lowest-energy conformer found by ConfBuster++ (blue) and CREST (pink) for (*A*) cyclo-(Pro-Ser-leu-Asp-Val) and (*B*) cyclo-(Arg-Gly-Asp-phe-([*N*-Me]Val)). The macrocycle backbone structures are highlighted.

environment of eight shared memory-processing units and 25 GB of RAM (random-access memory). As a result, each job processed 222 macrocycles. The average runtime per macrocycle at each peptide length is as follows: 275 s for trimers, 355 s for tetramers, and 407 s for pentamers. Larger macrocycles have longer runtimes due to the increased number of cleavable bonds within the macrocycle ring, which necessitate more iterations of the main algorithm (see *Methods* and Fig. 9). In comparison, the extensive conformer search in CREST took 96 and 48 CPU (central processing unit) hours for cyclo-(Pro-Ser-leu-Asp-Val) and cyclo-(Arg-Gly-Asp-phe-([*N*-Me]Val)), respectively.

Shape Diversity of Library Members. A given biological macromolecule imposes a shape selection for binding partners, and molecules possessing shape complementarity would thus be expected to display higher binding affinities. Molecular shape has been further demonstrated to be a key factor in promoting passive permeation (12, 13, 39). Although clear guidelines in this regard are yet to be established, shape diversity of small-molecule libraries has been cited as a fundamental indicator of functional diversity (40). With future screening applications in mind, we sought to probe the shape diversity within our own library.

In order to conduct shape assessment, we randomly chose a subset of 1 million structures to serve as representatives of the entire library. A commonly used method for measuring molecular space coverage is the calculation of normalized principal moment-of-inertia (PMI) ratios. Upon calculation of these values for the 1 million randomly chosen structures, we were pleased to find coverage of the PMI plot in almost its entirety (Fig. 6*B*). This is in contrast to other virtual databases and experimental datasets in the literature, wherein there exists a preponderance of rod-shaped molecules (41–44). Our macrocycles occupy unexplored regions of chemical space and present opportunities for identifying ligands.

To make visual assessment of some of these structures more practical, we decided to probe a smaller set of 10,000 structures while still retaining this spread. Toward this, we conducted principal component analysis (PCA) on the 1 million structures using five 3D molecular descriptors in RDKit to generate two principal components. Intuitively, a subset of maximally diverse structures, when represented in Euclidean space, would incorporate those that make up the smallest convex shape containing the data. The convex hull algorithm is an efficient algorithm for finding the sets of points that enclose this convex shape. We thus implemented this on the generated PCA data. The PMI plot for the algorithmically chosen 10,000 structures is shown in Fig. 6*B*, wherein, as desired, the spread of the original set has been retained. Table 1 displays examples of structures at each vertex of the triangle and for each template **G1**–**G3**. Descriptions of the PMI, PCA, and convex hull algorithms employed are detailed in *Methods*. Histograms for property distributions within our filters are shown in Fig. 6*A*.

Conformational dynamics are key drivers of passive permeability and target binding. As exemplified by cyclosporine A, "chameleonic" molecules that can alternately shield or expose polar functionality depending on solvent environment may be both membrane-permeable and water-soluble (12, 13, 39). Equally, however, conformational rigidity minimizes entropic costs upon target binding, enabling a molecule to achieve higher affinities (1–3). Thus, an interplay between these often contrasting properties is key to the success of an inhibitor. We were pleased to observe a broad range of conformational rigidity across library members, as illustrated by the structural overlays of the five lowest-energy conformations for each molecule in Table 1. Conformational variations largely arise from side-chain bond rotations rather than deviations in the macrocyclic backbones. Furthermore, as anticipated, structures bearing proline residues (**R2**, **D1**, **D3**, and **S2**) are less flexible than structures lacking the same. The number of conformations observed under our chosen thresholds (within 5 kcal/mol of the lowest-energy conformation and > 0.5 rmsd between conformers) vary based on the number and lengths of side chains.

The larger objective of this work is to use constrained macrocycle libraries as a resource for virtual screening. For those studies to be successful, predicted binding interactions will need to be experimentally validated. Our prior experimental results suggest a majority of macrocyclization reactions simulated by CPMG will proceed (15–20). In a series of studies, we demonstrated that the phenol of tyrosine and the indole of tryptophan react internally with the cinnamyl carbocation—regardless of the

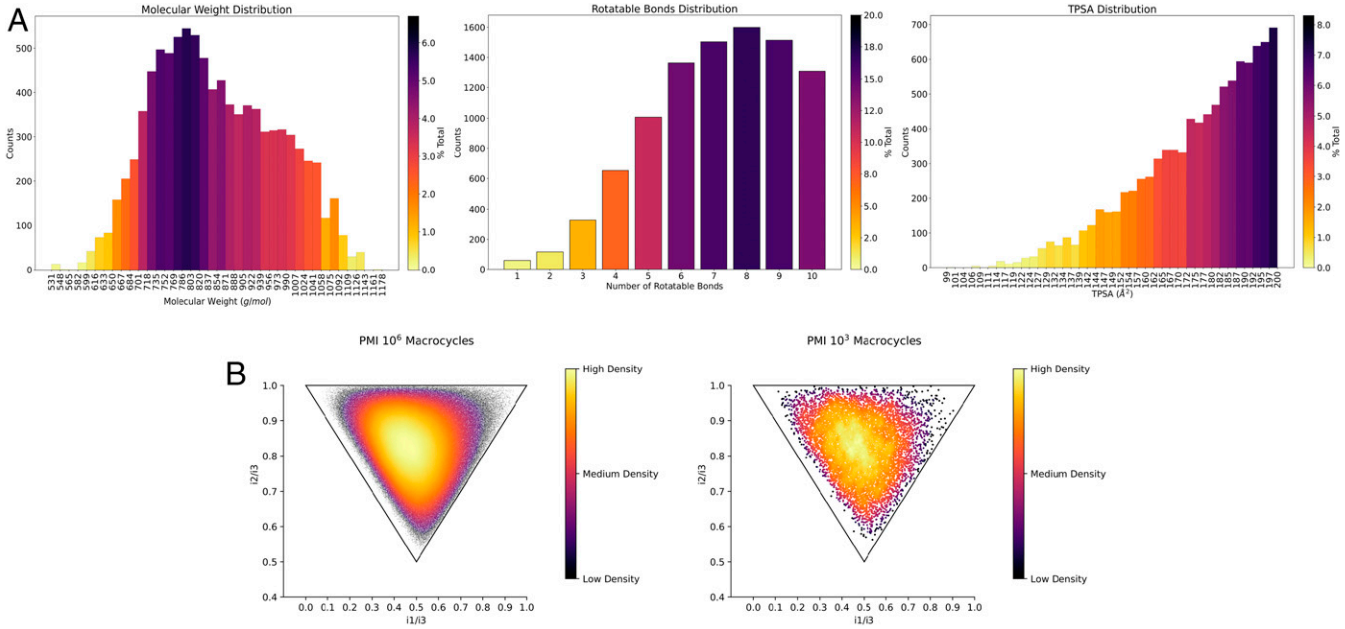
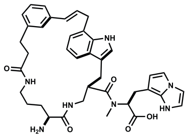
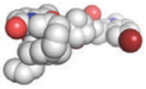
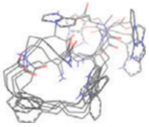
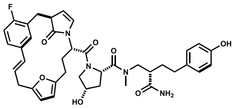
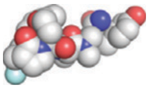
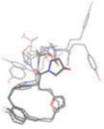
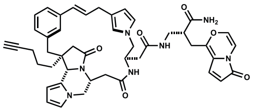
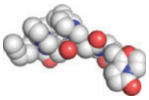
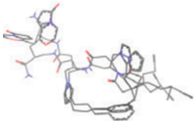
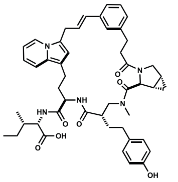
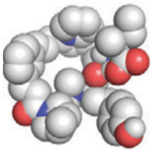
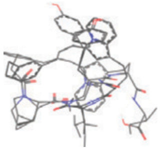
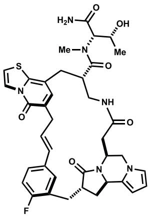
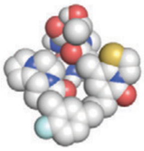
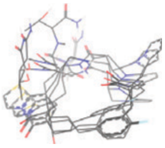
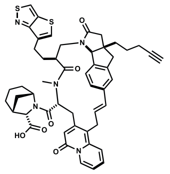
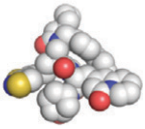
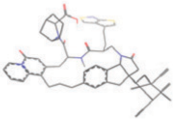
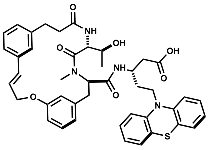
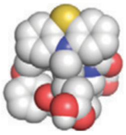
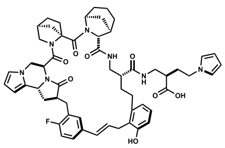
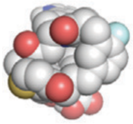
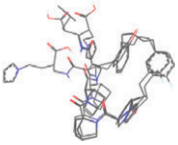
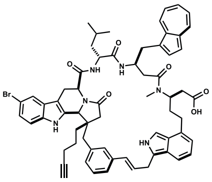
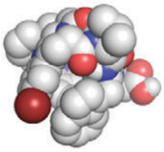
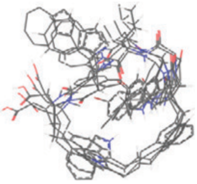


Fig. 6. (*A*) Histograms showing distributions of (*Left* to *Right*) molecular weight, total polar surface area, and rotatable bonds across the subset of 10,000 structures. (*B*) (*Left* to *Right*) PMI plot for 1 million randomly chosen structures; PMI plot for algorithmically selected subset of 10,000 structures.

Table 1. Representative structures at vertices of the PMI triangle

Shape MW/g mol^{-1}	Structure	Space Filling Model	Lowest Energy Conformations
R1 691.8 g/mol 13 conf			
R2 698.8 g/mol 20 conf			
R3 795.9 g/mol 40 conf			
D1 816.0 g/mol 13 conf			
D2 728.8 g/mol 5 conf			
D3 882.1 g/mol 5 conf			
S1 762.9 g/mol 24 conf			
S2 996.2 g/mol 10 conf			
S3 1096.2 g/mol 7 conf			

R, D, and S under “shape” stand for rod, disk, and sphere shapes, respectively. Numbers 1, 2, and 3 under “shape” represent the presence of G1, G2, and G3 templates, respectively, in the shown molecules. Structural overlays shown for the five lowest-energy conformations. Molecular weight (g/mol) and number of conformations (conf) observed within our used ConBuster++ thresholds (within 5 kcal/mol of the lowest-energy conformer and over 0.5 rmsd of each other) are listed under “shape.”

distance between reacting positions in acyclic precursors (17, 19). Heterocycles shown at the top of Fig. 4 were chosen as approximate isosteres of phenol or indole, such that CPMG would be predicting variable structures with the reactivity framework previously established using tyrosine and tryptophan. The macrocycles in Table 1 are revisited in Fig. 7, wherein this comparison is supported by more rigorous computations.

DFT calculations (ωB97X-D-SMD(methanol)/6–31G(d)) in Gaussian16 RevA.03 were performed to quantify the free energy of activation, $\Delta G^{\ddagger}$, of the rate-determining step of EAS between each heterocycle and an allyl cation model reactant, **TS**. From these calculations we find the sites predicted by RegioSQM (26) on almost all heterocycles to be under or within a few kcal/mol of $\Delta G^{\ddagger}$ values observed at reactive sites on phenol and indole. This suggests facile engagement of the heterocycles relative to known participants, as anticipated by CPMG, and thereby makes successful synthesis of library members probable. Moreover, our experimentally studied sequences were assembled almost exclusively from α-amino acids, whereas CPMG further incorporates β-residues. The extra methylene units serve to increase translational degrees of freedom and thereby reduce incipient ring strain during macrocyclization events.

To demonstrate access to library members, consider entry **D2** in Table 1. $\Delta G^{\ddagger}$ of the thiazolopyridinone in **D2** is determined to

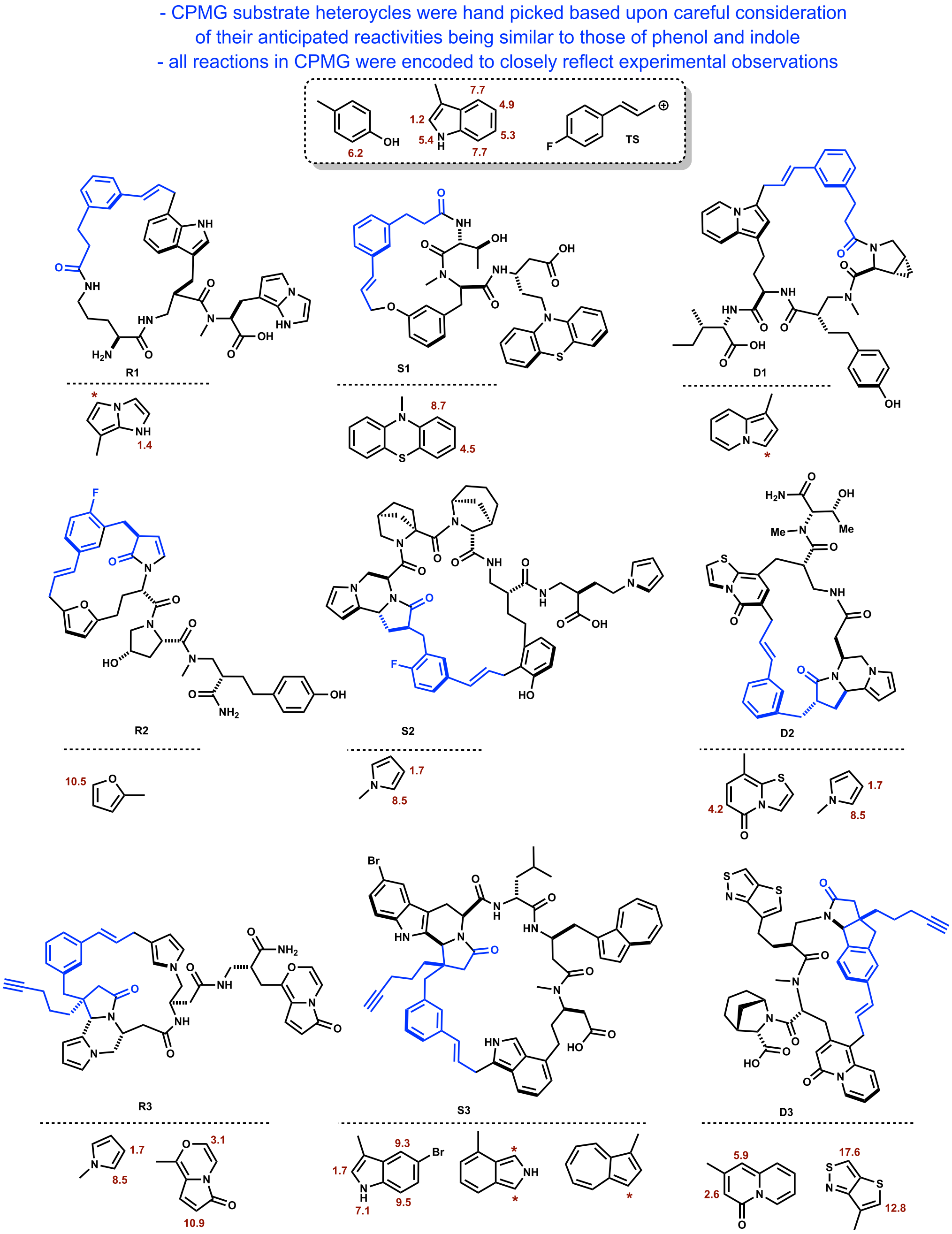
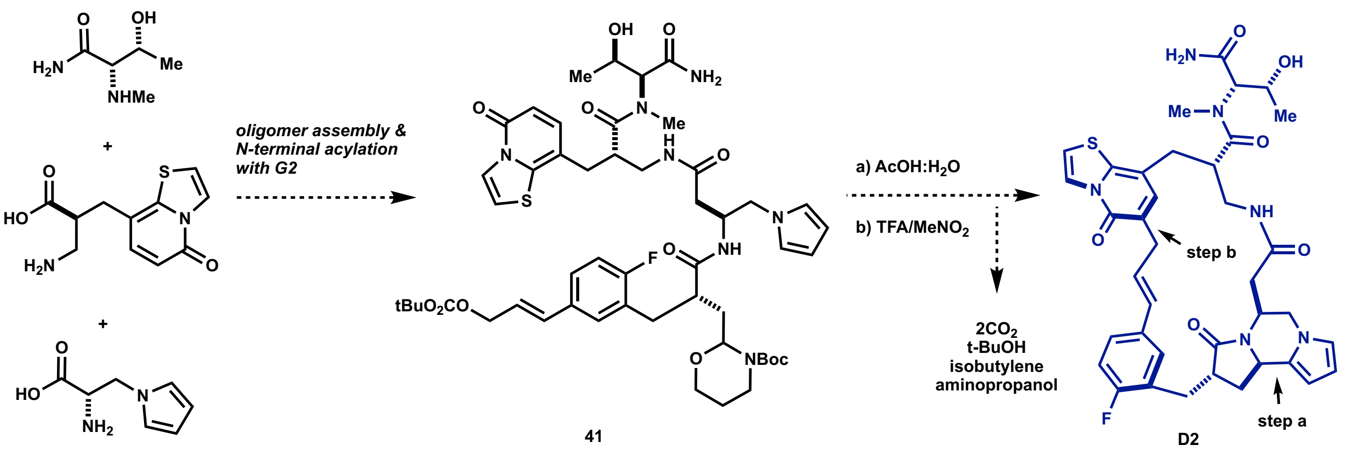

- CPMG substrate heteroycles were hand picked based upon careful consideration
of their anticipated reactivities being similar to those of phenol and indole
- all reactions in CPMG were encoded to closely reflect experimental observations
OH
6.2
7.7
4.9
1.2
5.3
5.4
7.7
F
TS
R1
1.4
S1
8.7
4.5
D1
R2
10.5
S2
1.7
8.5
D2
4.2
1.7
8.5
R3
1.7
8.5
3.1
10.9
S3
9.3
1.7
7.1
9.5
Br
D3
5.9
2.6
17.6
12.8
*

Fig. 7. Calculated free energies of activation ($\Delta G^{\ddagger}$) for the reaction of the indicated heterocyclic positions with the allyl cation, **TS**. $\Delta G^{\ddagger}$ values in kcal/mol. Reaction sites with $\Delta G^{\ddagger}$ values comparable to those calculated for phenol and indole are likely participants in ring-forming electrophilic aromatic substitutions. CPMG oligomers harboring multiple reactive sites result in regioisomeric macrocycles, analogous to experimental outcomes. Asterisks (*) denote sites where DFT calculations (ωB97X-D-SMD(methanol)/6–31G(d)) indicate barrierless, entropically controlled reactions.



Scheme 1. CPMG generates complex structures that can be synthesized readily from amino acid monomers and functionalized templates.

be 4.2 kcal/mol (Fig. 7), 2 kcal/mol lower than for the reaction with phenol at the ortho position (*SI Appendix*, Fig. S5). CPMG thus predicts a structure fully consistent with experimental methods data. **D2** would derive from L-*N*-Me threonine carboxamide, two unnatural amino acids (Scheme 1) and template **G2**. Solid-phase assembly of monomers followed by N-terminal acylation with **G2** would provide **41**. Treatment with aqueous acetic acid would then initiate hydrolytic degradation of the tetrahydrooxazine ring to form an intermediate *N*-acyl iminium ion that could capture the proximal pyrrole. Concentration and subsequent exposure to trifluoroacetic acid (TFA) in MeNO$_2$ would heterolyze the cinnamyl carbonate, and the resultant carbocation would engage the thiazolopyridinone in an electrophilic substitution reaction to afford **D2**. This molecule could be made on milligram scales in three facile steps from a machine-made tripeptide. The structure could be readily reduced, oxidized, and derivatized. Moreover, replacing **G2** with **G3** in the processing sequence would give a modified macrocycle displaying a terminal alkyne for derivatizations and tagging, enabling further analysis of structure activity relationships in binding and functional assays. We expect any member of the CPMG library to be similarly accessible and manipulated.

In the case of **D2**, no regioisomeric macrocycles are anticipated. For macrocycles having multiple reactive sites, we would obtain distributions of products arising from macrocyclizations at each reactive nucleophilic position (as for **D1**, **S2**, **R3**, and **S3** in Fig. 7). C-O linked **S1** would be afforded under Pd catalysis, whereas a C-C bonded regioisomer would be formed under acidic conditions upon engagement of the phenothiazine.

Conclusions and Outlook

There is a trend toward increasing complexity in small-molecule drug discovery research (40). The functions sought for small molecules are increasingly sophisticated. Among diverse chemotypes studied as drug leads, peptidomimetics and cyclic peptides are prominent (1–3, 11). They are obvious candidates for protein binding, and the field has surged with the use of DNA-templated reactions, in vitro biosynthesis, codon-reprogramming, and phage display (6–10). These powerful technologies generate large product libraries. In the case of cyclic peptides, however, they often produce large, conformationally flexible structures with poor pharmacological properties. We have developed a synthetic alternative, wherein small linear peptides are amalgamated with design inserts. Our composite products retain molecular recognition elements in the peptide while displaying that functionality as part of stable, conformationally defined polycyclic structures. The potential scope of the chemistry is enormous, but the experimental format has throughput limitations. To fully explore possible products, we developed CPMG as a computational rendering, wherein our synthetic methodology is simulated on a scale comparable to output numbers of biosynthetic libraries.

Our experimental methods were designed to be general, and CPMG tests the limits of that generality. Monomeric building blocks were created using all natural and 53 unnatural side chains in 12 backbone variations (α/β^2/β^3, L-/D-, methylene/ethylene-appended) of each. Oligopeptides were combinatorically generated, resulting in 2,020,794,198 macrocyclic structures of multistep sequences using designed templates **G1**–**G3**. Using ConfBuster++, we were additionally able to conduct rapid, large-scale conformational analyses.

Fragment-based databases have been utilized in the literature to generate unique small-molecule structures (45). However, we are aware of only a limited number of open-source tools to generate and analyze libraries of compounds at this level of molecular complexity. Moreover, CPMG is a means to augment and focus experiments in an integrated discovery platform—by computationally assessing, within constraints of property filters, which of these 10^9 molecules have potential to selectively interact with target protein surfaces. Recent improvements in both hardware and software make possible high-fidelity, high-speed docking simulations of millions of conformationally dynamic structures. Physics-based scoring functions such as DOCK, AutoDock Vina and smina have demonstrated predictive value (45–47). In recent years, deep learning models for ligand docking, scoring functions, and virtual screening have also emerged (48, 49). Convolutional neural networks have been advocated for protein-ligand scoring due to the number of parameters these systems generate relative to traditional scoring functions. As understanding deepens and docking implementations are refined, CPMG/ConfBuster++ could provide a unique ligand discovery resource for all types of structurally characterized proteins.

Methods

CPMG and ConfBuster++ are both written in Python 3.6.8 and rely primarily on the open-source framework RDKit (release 2019.03.2) (23). Additionally, CPMG incorporates data generated using third-party software RegioSQM and Jaguar, and ConfBuster++ employs OpenBabel 3.0.0. CPMG code can be found at: https://github.com/e-dang/Composite-Peptide-Macrocycle-Generator and ConfBuster++ code can be found at: https://github.com/e-dang/ConfBusterPlusPlus.

A note on reaction implementation in RDKit:

RDKit recognizes and implements reactions based on so-called "reaction templates." To avoid confusing this with our experimental templates, referred to as **G1**–**G3** in this manuscript, we will refer to the RDKit reaction template as RRT in the following text.

CPMG. Virtual library generation was performed using CPMG, which follows the schema outlined in Fig. 8. CPMG requires four types of user-defined building blocks from which all macrocycles derive: electrophilic templates, linking motifs (*SI Appendix*, Fig. S1), amino acid backbones, and heterocycles. **G1**–**G3** in Fig. 1 were used as our set of electrophilic templates, CH_2 and CH_2CH_2 chains were used at linking motifs with α, β^2, β^3 amino acid backbones. Heterocycles were chosen as described under *Results and Discussion*. Optionally, CPMG may accept intact, user-defined monomers to be used for peptide generation. These may or may not participate in reactions with the templates.

The set of selected heterocycles (Fig. 3 and *SI Appendix*, Fig. S1), along with the linkage motifs (*SI Appendix*, Fig. S2), were input to the SideChain

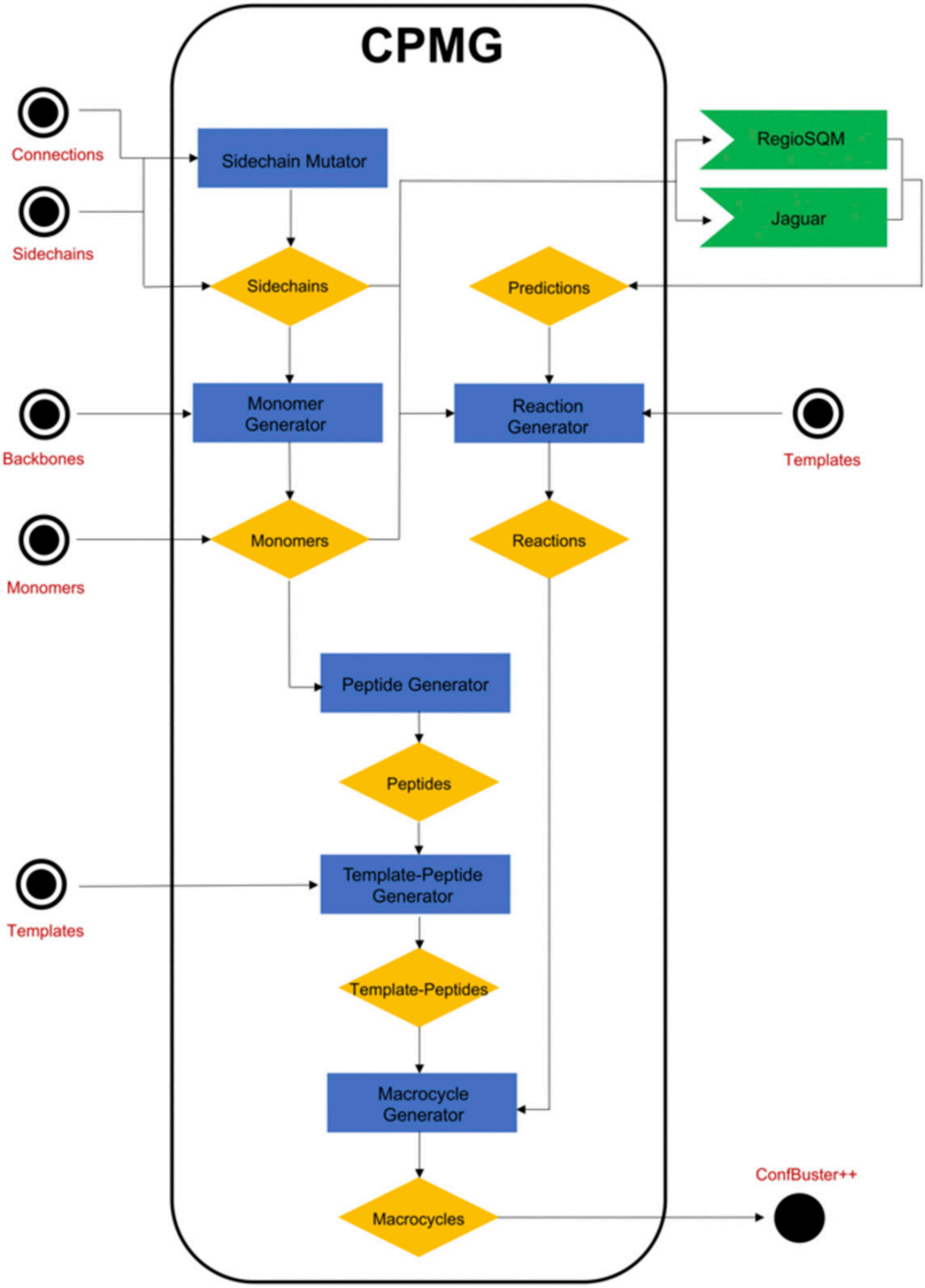


Fig. 8. Graphical representation of CPMG and its components.

Mutator component of CPMG, which not only allows methylene but also any other user-defined linking motifs (Fig. 8). The side chains along with the user-defined set of amino acid backbones (*SI Appendix*, Fig. S3) were subsequently passed to the MonomerGenerator component which produces a monomer for each combination of side-chain and amino acid backbone (Fig. 8). Monomers produced in this step were left with undefined stereochemistry, deferring stereochemical resolution to the macrocycle enumeration step (*vide infra*). The EAS regioselectivity and heteroatomic pK_a values for all aromatic-containing side chains and monomers (both CPMG-generated and user-defined) were incorporated from RegioSQM and Jaguar, respectively. RegioSQM calculations were simulated in $\varepsilon = 35.87$ in order to approximate experimental nitromethane conditions. pK_a values were generated for all acidic hydrogen-containing heterocycles as described under *Results and Discussion*.

Using these predictions, RRTs were generated by CPMG's ReactionGenerator (Fig. 8) for *N*-acyliminium ion capture, EAS, heteroatomic allylic substitution reactions between the cinnamyl electrophile and all predicted nucleophiles. The side-chain mutation, monomer and reaction generation steps were performed together as a single serial job with 1 GB of RAM.

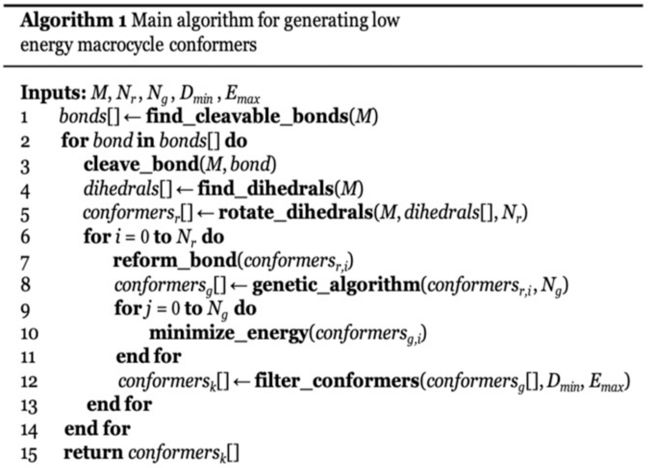
Peptides of specified length (trimer, tetramer, and pentamers) were assembled by the PeptideGenerator (Fig. 8) using the set of monomers. Each monomer was uniformly selected at random for each position of the peptide, with the constraint that all peptides must have at least one monomer capable of participating in an EAS reaction. Additionally, trimers and tetramers were allowed to contain at most two heterocycles, and pentamers were allowed three. The PeptideGenerator was also encoded to generate peptides that are C-terminally "capped" with *N*-ethyl-R units where R was uniformly selected at random (R = **1**–**30**, Fig. 3). Each eligible peptide was duplicated and C-terminally capped in oligopeptides where the maximum number of heterocycles had not been reached.

The set of peptides were then combined with **G1**–**G3,** Fig. 1, via amide linkage in the TemplatePeptideGenerator, forming template-bound oligomers (Fig. 8). This procedure produces at least three cinnamyl template-peptide hybrids for each peptide. However, more can be made if there are any primary or secondary amine-containing side chains (not including guanidine) present on the peptide. This allows for peptides containing residues such as lysine to produce more than three cinnamyl template-peptides. Peptide and template-peptide generation was performed together in an array of three jobs (one job for each peptide length), where each job was allocated an 8-slot parallel environment with 16 GB of RAM.

The cinnamyl-bound oligomers were then input to the MacrocycleGenerator, which applies the relevant RRTs to each cinnamyl-bound oligomer in sequence, producing the set of macrocycles (Fig. 8). Additionally, the MacrocycleGenerator applied monomethylation and carboxyl to amide transformations to each macrocycle and permuted the stereochemistry at each stereocenter forming all combinations of enantiomers. The MacrocycleGenerator then filtered out any resultant macrocycle that had a molecular weight ≥ 1,200 Da, number of rotatable bonds ≥ 10, or TPSA ≥ 200 Å^2. Macrocycle generation was performed with 1,500 job job array (500 jobs per peptide length) where each job was allocated an 8-slot parallel environment and 12 GB of RAM.

ConfBuster++. Conformers for a subset of 1×10^6 randomly selected macrocycles were generated using ConfBuster++. Fig. 9 depicts the pseudocode for the main algorithm in ConfBuster++, which we extracted from ConfBuster. Implementation details have been left as function calls within the pseudocode, as how they are accomplished greatly depends on the molecular representation one is working with (Fig. 9).

The algorithm begins by identifying all cleavable bonds within the macrocyclic ring, where a cleavable bond is defined as any single bond, that when cleaved, will result in a linearized molecule. A constraint requires the single bond to not be between double bonds or two chiral atoms (Fig. 9, line 1). The latter constraint is to prevent stereochemical inversion at those stereocenters. For each cleavable bond, the following sequence of operations are performed on the macrocycle (Fig. 9, line 2). First the bond is cleaved, and the dihedral angles composed of the atoms that were in the macrocycle ring are identified on the resultant linearized molecule (Fig. 9, lines 3–4). These dihedral angles are then rotated systematically, starting from the dihedrals farthest from the cleaved atoms and ending on the dihedrals that contain the cleaved atoms (Fig. 9, line 5). Once the cleaved atoms are brought into a distance between 1.0 and 2.5 Å of each other, the resulting conformation is retained, and the process repeats until N_r = 15 number of conformers are generated in this manner (Fig. 9, line 5). For each of the N_r conformers, the cleaved bond is reformed, and each resultant macrocycle conformer is fed into OpenBabel's genetic algorithm, producing N_g = 5 conformers (Fig. 9, line 8). Each N_g conformer is then subjected to energy minimization using RDKit's MMFF94s force field (Fig. 9, lines 9–10). The resulting conformers are then filtered based on rmsd and energy, where all conformers must have greater than D_{min} = 0.5 Å rmsd and no greater than E_{max} = 5 kcal/mol energy from the lowest-energy conformer (Fig. 9, line 12). The set of all filtered conformers are stored in *conformers*$_k$[], which is eventually returned to client (Fig. 9, line 15). Conformer generation was performed with an array of 4,500 jobs (1,500 jobs per length of peptide) where each job was allocated an 8-slot parallel environment and 9.6 GB of RAM.



Algorithm 1 Main algorithm for generating low energy macrocycle conformers

```
Inputs: $M, N_r, N_g, D_{min}, E_{max}$
1   bonds[] ← find_cleavable_bonds(M)
2   for bond in bonds[] do
3       cleave_bond(M, bond)
4       dihedrals[] ← find_dihedrals(M)
5       $conformers_r$[] ← rotate_dihedrals(M, dihedrals[], $N_r$)
6       for $i = 0$ to $N_r$ do
7           reform_bond($conformers_{r,i}$)
8           $conformers_g$[] ← genetic_algorithm($conformers_{r,i}$, $N_g$)
9           for $j = 0$ to $N_g$ do
10              minimize_energy($conformers_{g,i}$)
11          end for
12          $conformers_k$[] ← filter_conformers($conformers_g$[], $D_{min}$, $E_{max}$)
13      end for
14  end for
15  return $conformers_k$[]
```

Fig. 9. The pseudo code for the main Algorithm in ConfBuster++. M is the macrocycle, N_r is the number of conformers to find via dihedral rotations, N_g is the number of conformations to generate via the genetic algorithm, $D_{\min}$ is the minimum rmsd between conformers, and $E_{\max}$ is the maximum energy difference between the lowest- and highest-energy conformers.

CREST. Conformer searches in CREST Version 2.7.1 (37) were performed using the iMTD-GC workflow in combination with the GFN2-xTB tight-binding DFT functional (38) as implemented in XTB Version 6.2 (50). Default settings were used except for an energy window of 5 kcal/mol.

Algorithms for Analysis. *PCA* was implemented using the Scikit-learn Python library to generate two principal components. The features chosen for conducting PCA were calculated using the RDKit Chem.Descriptors3D module and are as follows: Asphericity, Eccentricity, Inertial Shape Factor, Radius of Gyration, and Sphericity Index. The variance ratio for the amount of variance explained by each component was 57:27. The principal axes in feature space for the aforementioned features across the two components, respectively, were as follows:

[[0.51630555 0.44168347 0.56647178 0.4563838 0.09574887]

[−0.22203754 0.0528527 −0.11022665 0.5093864 −0.82236337]]

Following PCA, the *convex hull algorithm* was applied iteratively to the set of coordinates produced. At each iteration, macrocycles corresponding to the convex hull of the data were added to the set of maximally diverse structures. These points were subsequently removed from the data, and the process was repeated until a set of at least 10,000 structures were chosen by the algorithm. *PMI* plots were generated using the Matplotlib library in Python. Normalized principal moments ratios were calculated in RDKit using the Chem.Descriptors3D module.

Data Availability. All study data are included in the article and *SI Appendix*.

1. E. M. Driggers, S. P. Hale, J. Lee, N. K. Terrett, The exploration of macrocycles for drug discovery—An underexploited structural class. *Nat. Rev. Drug Discov.* **7**, 608–624 (2008).
2. F. Giordanetto, J. Kihlberg, Macrocyclic drugs and clinical candidates: What can medicinal chemists learn from their properties? *J. Med. Chem.* **57**, 278–295 (2014).
3. M. R. Arkin, J. A. Wells, Small-molecule inhibitors of protein-protein interactions: Progressing towards the dream. *Nat. Rev. Drug Discov.* **3**, 301–317 (2004).
4. N. C. Wrighton *et al.*, Small peptides as potent mimetics of the protein hormone erythropoietin. *Science* **273**, 458–464 (1996).

5. O. Livnah *et al*., Functional mimicry of a protein hormone by a peptide agonist: The EPO receptor complex at 2.8 A. *Science* **273**, 464–471 (1996).
6. Z. J. Gartner *et al*., DNA-templated organic synthesis and selection of a library of macrocycles. *Science* **305**, 1601–1605 (2004).
7. B. N. Tse, T. M. Snyder, Y. Shen, D. R. Liu, Translation of DNA into a library of 13,000 synthetic small-molecule macrocycles suitable for in vitro selection. *J. Am. Chem. Soc.* **130**, 15611–15626 (2008).
8. L. B. Giebel *et al*., Screening of cyclic peptide phage libraries identifies ligands that bind streptavidin with high affinities. *Biochemistry* **34**, 15430–15435 (1995).
9. T. Kawakami *et al*., Diverse backbone-cyclized peptides via codon reprogramming. *Nat. Chem. Biol.* **5**, 888–890 (2009).
10. T. Passioura, T. Katoh, Y. Goto, H. Suga, Selection-based discovery of druglike macrocyclic peptides. *Annu. Rev. Biochem.* **83**, 727–752 (2014).
11. A. A. Vinogradov, Y. Yin, H. Suga, Macrocyclic peptides as drug candidates: Recent progress and remaining challenges. *J. Am. Chem. Soc.* **141**, 4167–4181 (2019).
12. J. Witek *et al*., Interconversion rates between conformational states as rationale for the membrane permeability of cyclosporines. *Chem. Phys. Chem.* **18**, 3309–3314 (2017).
13. P. Matsson, J. Kihlberg, How big is too big for cell permeability? *J. Med. Chem.* **60**, 1662–1664 (2017).
14. B. Over *et al*., Structural and conformational determinants of macrocycle cell permeability. *Nat. Chem. Biol.* **12**, 1065–1074 (2016).
15. H. Zhao *et al*., Acid promoted cinnamyl ion mobility within peptide derived macrocycles. *J. Am. Chem. Soc.* **130**, 13864–13866 (2008).
16. K. V. Lawson, T. E. Rose, P. G. Harran, Template-constrained macrocyclic peptides prepared from native, unprotected precursors. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E3753–E3760 (2013).
17. K. V. Lawson, T. E. Rose, P. G. Harran, Template-induced macrocycle diversity through large ring-forming alkylations of tryptophan. *Tetrahedron* **69**, 7683–7691 (2013).
18. T. E. Rose, K. V. Lawson, P. G. Harran, Large ring-forming alkylations provide facile access to composite macrocycles. *Chem. Sci. (Camb.)* **6**, 2219–2223 (2015).
19. T. E. Rose *et al*., On the prevalence of bridged macrocyclic pyrroloindolines formed in regiodivergent alkylations of tryptophan. *Chem. Sci. (Camb.)* **7**, 4158–4166 (2016).
20. B. H. Curtin *et al*., Assembly of complex macrocycles by incrementally amalgamating unprotected peptides with a designed four-armed insert. *J. Org. Chem.* **83**, 3090–3108 (2018).
21. X. Barbeau, A. T. Vincent, P. Lagüe, ConfBuster: Open-source tools for macrocycle conformational search and analysis. *J. Open Res. Softw.* **6**, 1–6 (2018).
22. G. Landrum, RDKit: Open-source cheminformatics; 2016. rdkit.org. Accessed 4 October 2020.
23. W. R. Pitt, D. M. Parry, B. G. Perry, C. R. Groom, Heteroaromatic rings of the future. *J. Med. Chem.* **52**, 2952–2963 (2009).
24. J.-D. Chai, M. Head-Gordon, Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615–6620 (2008).
25. F. Weigend, R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297–3305 (2005).
26. J. C. Kromann, J. H. Jensen, M. Kruszyk, M. Jessing, M. Jørgensen, Fast and accurate prediction of the regioselectivity of electrophilic aromatic substitution reactions. *Chem. Sci. (Camb.)* **9**, 660–665 (2017).
27. N. Miyaura *et al*., *Cross-Coupling Reactions: A Practical Guide; Topics in Current Chemistry*, (Springer Berlin Heidelberg, 2003).
28. A. D. Bochevarov *et al*., Jaguar: A high-performance quantum chemistry software program with strengths in life and materials sciences. *Int. J. Quantum Chem.* **113**, 2110–2142 (2013).
29. J.-P. Ebejer, G. M. Morris, C. M. Deane, Freely available conformer generation methods: How good are they? *J. Chem. Inf. Model.* **52**, 1146–1158 (2012).
30. G. Landrum, EmbedMolecule not respecting double-bond stereochemistry, https://github.com/rdkit/rdkit/issues/1852. Accessed 4 October 2020.
31. M. P. Jacobson *et al*., A hierarchical approach to all-atom protein loop prediction. *Proteins* **55**, 351–367 (2004).
32. D. Sindhikara *et al*., Improving accuracy, diversity, and speed with prime macrocycle conformational sampling. *J. Chem. Inf. Model.* **57**, 1881–1894 (2017).
33. J. H. Viles *et al*., Multiple solution conformations of the integrin-binding cyclic pentapeptide cyclo(-Ser-D-Leu-Asp-Val-Pro-). Analysis of the (phi, psi) space available to cyclic pentapeptides. *Eur. J. Biochem.* **242**, 352–362 (1996).
34. U. K. Marelli *et al*., Receptor-bound conformation of cilengitide better represented by its solution-state structure than the solid-state structure. *Chemistry* **20**, 14201–14206 (2014).
35. C. M. M. M. Moruno, L. Doedens, A. O. Frank, H. Kessler, Enhancement of receptor selectivity of cilengitide by multiple *N*-methylation. *J. Pept. Sci.* **16**, 45–46 (2010).
36. A. S. Kamenik, U. Lessel, J. E. Fuchs, T. Fox, K. R. Liedl, Peptidic macrocycles - conformational sampling and thermodynamic characterization. *J. Chem. Inf. Model.* **58**, 982–992 (2018).
37. S. Grimme, Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *J. Chem. Theory Comput.* **15**, 2847–2862 (2019).
38. C. Bannwarth, S. Ehlert, S. Grimme, GFN2-xTB-An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
39. J. Witek *et al*., Rationalization of the membrane permeability differences in a series of analogue cyclic decapeptides. *J. Chem. Inf. Model.* **59**, 294–308 (2019).
40. W. R. J. D. Galloway, A. Isidro-Llobet, D. R. Spring, Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules. *Nat. Commun.* **1**, 80 (2010).
41. D. G. Brown, J. Boström, Analysis of past and present synthetic methodologies on medicinal chemistry: Where have all the new reactions gone? *J. Med. Chem.* **59**, 4443–4458 (2016).
42. W. H. B. Sauer, M. K. Schwarz, Molecular shape diversity of combinatorial libraries: A prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.* **43**, 987–1003 (2003).
43. C. N. Morrison *et al*., Expanding medicinal chemistry into 3D space: Metallofragments as 3D scaffolds for fragment-based drug discovery. *Chem. Sci. (Camb.)* **11**, 1216–1225 (2020).
44. F. Carles, S. Bourg, C. Meyer, P. Bonnet, PKIDB: A curated, annotated and updated database of protein kinase inhibitors in clinical trials. *Molecules* **23**, 908 (2018).
45. J. Lyu *et al*., Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
46. O. Trott, A. J. Olson, AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
47. D. R. Koes, M. P. Baumgartner, C. J. Camacho, Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).
48. E. B. Lenselink *et al*., Beyond the hype: Deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminform.* **9**, 45 (2017).
49. M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, D. R. Koes, Protein-ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).
50. S. Ehlert *et al*., Semiempirical Extended Tight-Binding Program Package, xtb version 6.2.3. https://doi.org/10.5281/zenodo.3712016. Accessed 4 April 2020.