UCLA
UCLA Electronic Theses and Dissertations

Title

Exploring a Hypothetical Giga-Library of Synthesizable Macrocyclic Composites for the Identification of New Ligands for Protein Surfaces

Permalink

https://escholarship.org/uc/item/8753p553

Author

Saha, Ishika

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Exploring a Hypothetical Giga-Library of Synthesizable Macrocyclic

Composites for the Identification of New Ligands for Protein Surfaces

A dissertation submitted in partial satisfaction

of the requirements for the degree Doctor of Philosophy

in Chemistry

by

Ishika Saha

2022

ABSTRACT OF THE DISSERTATION

Exploring a Hypothetical Giga-Library of Synthesizable Macrocyclic

Composites for the Identification of New Ligands for Protein Surfaces

by

Ishika Saha

Doctor of Philosophy in Chemistry

University of California, Los Angeles, 2021

Professor Patrick. G. Harran, Chair

Many actively pursued pharmacological targets are difficult to drug using conventional small molecule therapeutics because they lack conventional binding sites. These so called 'undruggable' targets typically interact with other proteins *via* shallow, solvent exposed interfaces. Peptidomimetic macrocycles have the potential to mediate such systems because the embedded peptide can mimic native protein structure and recognition elements, and the ring structures contribute to structural preorganization, lowering entropic penalties upon target binding. Compounds of this type having precise shapes and drug-like character are coveted, but are relatively difficult to synthesize. Our lab has developed methods to synthesize shape-defined macrocycles from small linear peptides. These experiments run as processes, wherein designed templates react incrementally with unprotected oligomers to form composite products. The resulting compounds retain molecular recognition elements in the oligomer, yet display that functionality as part of stable polycyclic structures. Our experimental work is based on proteinogenic amino acids and the reactivity of their nucleophilic side chains. However, using unnatural amino acids, the hypothetical scope of the chemistry becomes vast and far outpaces the capacity of our experimental format. Here, we describe the development of a computational

rendering of our experimental platform, **C**omposite **P**eptide **M**acrocycle **G**enerator (CPMG). This open-source platform simulates our multi-step reaction chemistry using a large, tailored monomer set. We have used the algorithms to anticipate product outcomes of >2 billion processing sequences. We have further developed software to generate three- dimensional structures for each product. Every library member has feature constraints meant to increase the probability of it being bioavailable. We discuss efforts to merge our experimental and computational abilities into a single, iterative workflow to discover new macrocyclic ligands for challenging protein targets. We describe new computational tools and techniques to allow rapid, flexible docking of conformationally dynamic ligands onto multiple protein targets. We describe experiments to validate predictions by synthesizing novel arene amino acids and engaging them in macrocyclizations to generate previously unknown ring systems. We show how the interplay between calculations and synthesis can offer insight into molecular properties and applications.

The dissertation of Ishika Saha is approved.

Neil K. Garg

Kendall N. Houk

Hosea M. Nelson

Patrick G. Harran, Committee Chair

University of California, Los Angeles

2022

This dissertation is dedicated to my grandmother,

Gauri Saha, who inspired this journey,

and to my parents, Neena and Suvamoy Saha,

who give me courage and support every step of the way.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Over the last five years, I have had the opportunity to work closely with several individuals in synthesis, computational chemistry, computer science, and biochemistry. The work described in this dissertation would not have been possible without their efforts, and the knowledge they shared with me. The opportunity to bridge synthesis and computation and make scientific discoveries together has been a great privilege.

I am indebted to Prof. Patrick Harran, under whose mentorship I learned to ground my research in first principles, but fearlessly seek new methods to drive it. I learned that being able to effectively communicate one's research is just as important as being able to perform it. The work described here stems from over a decade of discoveries made by several scientists in his lab, and I am glad to have been a part of that story.

I am deeply grateful to Prof. Kendall Houk for introducing me to a world of computational chemistry. Without his guidance and knowledge, the platforms developed in this work would not exist. I was fortunate to work with Eric Dang and Dr. Dennis Svatunek in his group, both of whom made key contributions to the computational frameworks described in Chapter 1. I am grateful to Prof. Stefano Forli at Scripps Research Institute, whose guidance enabled the docking studies described in Chapter 2. Gabriella Cooper helped in the indolizin-yl amino acid syntheses described in Chapter 3. Dr. David Strugatsky ran the assays described in Chapter 4, and Dr. Emily Murzinski worked with me in synthesizing many of the compounds described in that chapter.

Thank you to my lab mates for their camaraderie, and for sharing their insight the countless times my research hit a wall. Dr. Hui Ding, in particular, provided guidance on experiments and analyses throughout my time in graduate school. No question was ever too big or small, and I learned a great deal from him.

A special thank you to my friends, Dayanni and Jessica, for sharing in all the highs and lows of these years with me.

CURRICULUM VITAE

Education

M.Chem   University of Durham, England, UK, 2013.

Publications

Murzinski, E. S.; **Saha, I**.; Ding, H.; Strugatsky, D.; Hollibaugh, R. A.; Liu, H.; Tontonoz, P.; Harran, P. G. In Search of Small Molecules That Selectively Inhibit MBOAT4. *Molecules,* **2021**, *26* (24), 7599-7576.

Saha, I; Harran, P.G. Virtual Screening for Chemists. *ACS InFocus.* **2021**. (*https://pubs.acs.org/doi/book/10.1021/acsinfocus.7e5001*)

Saha, I.; Dang, E. K.; Svatunek, D.; Houk, K. N.; Harran, P. G. Computational Generation of an Annotated Gigalibrary of Synthesizable, Composite Peptidic Macrocycles. *Proc. Natl. Acad. Sci.* **2020**, *117* (40), 24679–24690.

Saha, I.; Baxendale, I. R.; Baumann, M. Unprecedented Alkene Transposition in Phthalate–Amino Acid Adducts. *Synlett* **2018**, *29* (20), 2648–2654.

Presentations

Saha, I. Exploring a Hypothetical GigaLibrary for Novel Structures and Functions, CUP XXI, OpenEye, Santa Fe, NM, March 9th, 2022.

Saha, I. Computational Generation of a Gigalibrary of Composite Peptidic Macrocycles, 6th Annual Genentech Graduate Student Symposium in Chemical Research, Genentech, San Francisco, CA, May 10th, 2021.

Awards

|	
Summer Mentored Research Fellow, *University of California, Los Angeles*	June 2021
Dissertation Year Fellow, *University of California, Los Angeles*	June 2021
Ralph and Charlene Bauer Awardee, *University of California, Los Angeles*	May 2021
Genentech Graduate Student Symposium Awardee, *Genentech, San Francisco*	May 2021
Christopher S. Foote Senior Fellow, *University of California, Los Angeles*	Jan. 2020
Saul Winstein Fellow – *University of California, Los Angeles*	Sept. 2019
Audree Fowler Fellow; shortlisted – *University of California, Los Angeles*	Nov. 2018

Introduction

Macrocyclic small molecules are expanding the repertoire of lead compounds for drug discovery.[1] Their ring structures scaffold extended pharmacophores and also offset entropic costs for receptor binding.[2–4] Macrocycles have been identified as enzyme inhibitors, as GPCR agonists/antagonists, as modulators of protein–protein interactions (PPIs), and as inhibitors of enzymes with solvent-exposed or cryptic binding sites.[5–7] In certain instances, macrocyclic structures have been shown to better discriminate between proteins bearing close structural homology.[5] Synthetic macrocycles are being advanced to clinical testing with increasing frequency.[8]

Peptide-derived macrocycles are especially attractive for targeting protein surfaces, because the embedded peptide can mimic native protein structure and recognition elements.[9,10] Myriad techniques exist to prepare cyclic peptides, and powerful biosynthetic platforms have emerged that generate cyclopeptides by the billions.[11–16] Screening these collections has enabled the discovery of protein ligands that prove useful in many areas of research. In spite of this promise, 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.[17,18] The majority of cyclic peptides above a kilodalton in mass fail to passively permeate cell membranes. There remain great opportunities to discover new, smaller peptidyl structures that bind target proteins tightly and selectively, and also have useful pharmacological properties.

High-resolution structures, mutagenesis, and computational studies of protein interaction surfaces show that binding determinants are often localized, despite relatively large contact areas and variable secondary structure at PPI interfaces.[19–22] Aromatic amino acid side chains are prevalent at these locations and contribute significantly to affinity and selectivity.[23] Molecular shape and conformational dynamics are also decisive.[24] Based on these observations, over the

past decade, our lab has developed synthetic methods to amalgamate small peptides with designed scaffolding reagents to create composite macrocycles having diverse ring connectivities and embedded heterocyclic motifs (Fig. 1A).[25–34] Aromatic side chains are directly engaged during 2 or 3 step processing sequences. We have characterized hundreds of examples[25–34] and have shown that product structure and properties can depart markedly from those of the starting oligomer (e.g. Fig. 1B). Our experimental work is based on proteinogenic amino acids and the reactivity of their nucleophilic side chains. However, by exploiting abiotic building blocks, we can vastly expand the potential scope of this chemistry far beyond our capacity for synthesis. This dissertation documents our approach to capitalizing on this new and promising chemical space.
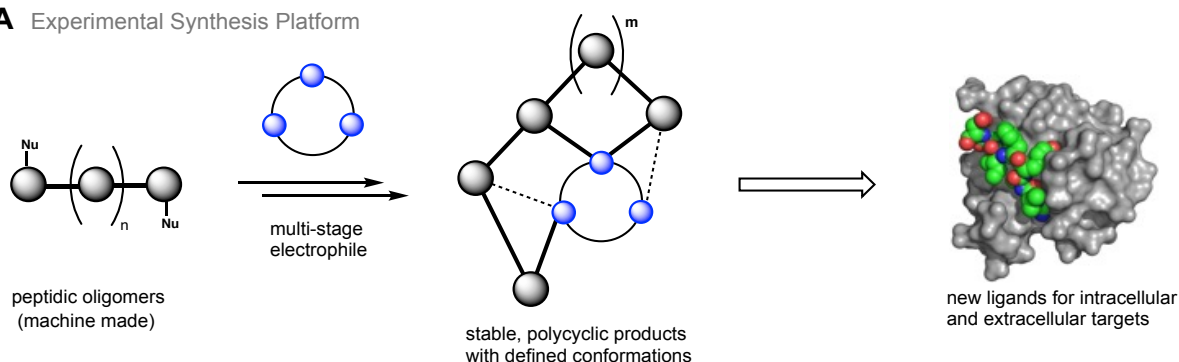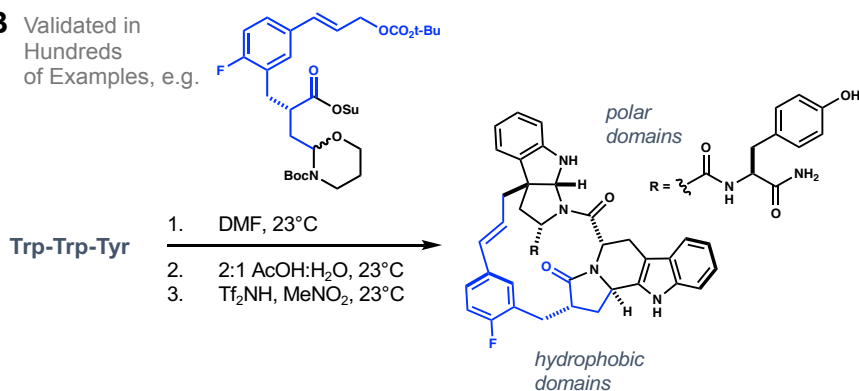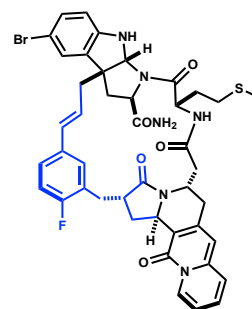
Figure 1. A. We have developed synthetic methods to amalgamate designed scaffolding reagents with peptidyl oligomers to afford macrocyclic composites. The scope of the chemistry is broad and **B.** it can markedly alter the shape and properties of starting materials. **C.** Based on established reactivity patterns, we have written software to predict > 2 billion hypothetical product outcomes.

Ongoing advances in synthesis, molecular biology, and computer technology are enabling high-speed, high-fidelity calculations for *in-silico* drug design and discovery. Although virtual

screening has been an aspirational field for decades, new discoveries are being made at an ever-increasing rate. The first reports of successful lead identification through screening mega- and gigascale small molecule libraries were described in 2019–2020.[35,36] Improvements in protein–ligand interaction modeling, the power of computational resources and new technologies for high-resolution protein structure determination provide a bright outlook for virtual screening – both as a complement to experimental high throughput screening and, increasingly, as a cheaper, faster, and more accessible alternative. As synthetic chemists, our own journey into this field was motivated by a desire to systematically probe chemical space defined by products emanating from in-house experimental procedures. The number of hypothetical products that could be accessed by this methodology far outpaced our experimental capacity. As a surrogate, we developed a computational rendering of the platform (Fig. 1C); and what began as a thought experiment emerged as a promising new resource for protein ligand discovery.

Along those lines, Chapter 1 describes our development of the computational platform, **C**omposite **P**eptide **M**acrocycle **G**enerator (CPMG), to simulate our synthetic methodology (e.g. Fig. 2), on a scale comparable to output numbers of biosynthetic libraries.[31] We describe our development of computer algorithms to predict outcomes of more than 2 billion processing sequences, as well as software to generate accurate three-dimensional structures for each product. Every library member has feature constraints that increase the probability that these new substances will be bioavailable: molecular weight < 1200 Da, total polar surface area < 200 Å^2, <10 rotatable bonds, and fewer than five hydrogen bond donors. Chapter 2 details our efforts to integrate the giga-library into a large-scale virtual screening pipeline for filtering and prioritizing macrocyclic molecules against protein targets of interest.[37] Protein surfaces are computationally analyzed to determine promising sites for ligand binding. Using molecular docking tools, molecules from the giga-library are virtually screened against these sites and prioritized on the basis of both anticipated bioactivity and pharmacological properties. Chapter 3 documents our efforts to experimentally test the veracity of the computational platform. We probe the synthetic

tractability of a subset of virtually generated molecules – from their input building block form in the virtual database through incorporation into final macrocyclic composites. Through this work, we hope to demonstrate that while both experimental and computational methods are invaluable for drug discovery, combining these tools into an integrated platform with multiple feedback points can circumvent the bottlenecks of each approach in isolation.
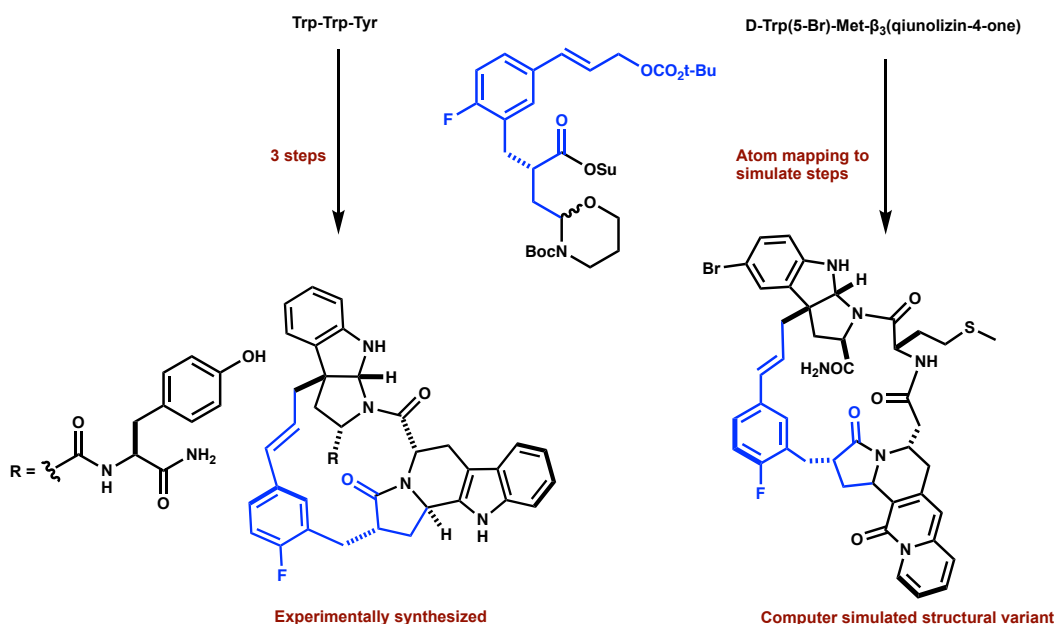
Figure 2. CPMG tests the generality of the chemistry employed in our experimental platform and allows us to broaden the scope of product outcomes. Over 400 building blocks were systematically permuted and processed using reactivity rules derived from in house reaction sequences to generate a library of peptidomimetic macrocycles.

Finally, the work described in chapter 4 documents our efforts in a separate area of interest – identifying small molecule mimics for selectively targeting the Ghrelin O-Acyl Transferase (GOAT) enzyme.[38] Ghrelin is a 28-residue peptide hormone produced by stomach P/D1 cells located in oxyntic glands of the fundus mucosa. Post-translational octanoylation of its Ser-3 residue, catalyzed by GOAT is essential for the binding of the hormone to its receptor in target tissues.[39–41] Physiological roles of acyl ghrelin include the regulation of food intake, growth hormone secretion from the pituitary, and inhibition of insulin secretion from the pancreas.[42–45] Inhibition of GOAT has thus emerged as a promising avenue for addressing metabolic disorders such as Type II diabetes. We describe the synthesis and activities of small lipopeptidomimetics

that inhibit GOAT in vitro. These molecules compete directly for substrate binding. We further describe the synthesis of heterocyclic inhibitors that compete at the acyl coenzyme A binding site (Fig. 3).
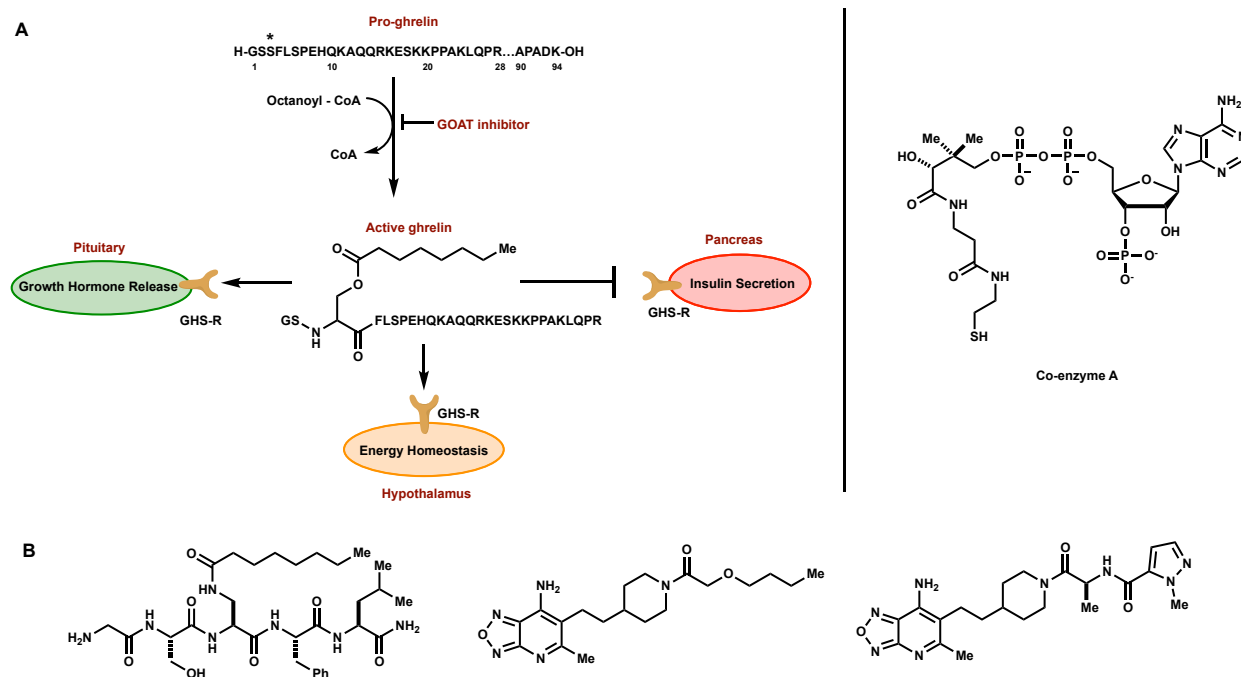
Figure 3. A. GOAT-catalyzed octanoylation of pro-ghrelin is a critical part of ghrelin maturation. Exogenously produced circulating ghrelin acts at its receptor in the hypothalamus, where it is believed to modulate energy homeostasis. Ghrelin also acts in the pancreas to suppress insulin response to blood glucose. Systemic inhibition of GOAT is expected to reduce circulating levels of acyl-ghrelin levels, leading to improved insulin response to glucose challenge. GOAT is weakly inhibited by active acyl ghrelin, and more potently by the structures in panel **B**, developed in our lab.

References

(1)   Driggers, E. M.; Hale, S. P.; Lee, J.; Terrett, N. K. The Exploration of Macrocycles for Drug Discovery — an Underexploited Structural Class. *Nat. Rev. Drug Discov.* **2008**, *7*, 608.

(2)   Giordanetto, F.; Kihlberg, J. Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? *J. Med. Chem.* **2014**, *57* (2), 278–295.

(3)   Arkin, M. R.; Wells, J. A. Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing towards the Dream. *Nat. Rev. Drug Discov.* **2004**, *3* (4), 301–317.

(4)   Villar, E. A.; Beglov, D.; Chennamadhavuni, S.; Porco, J. A. J.; Kozakov, D.; Vajda, S.; Whitty, A. How Proteins Bind Macrocycles. *Nat. Chem. Biol.* **2014**, *10* (9), 723–731.

(5)   Egbert, M.; Whitty, A.; Keserű, G. M.; Vajda, S. Why Some Targets Benefit from beyond Rule of

Five Drugs. *J. Med. Chem.* **2019**, *62* (22), 10005–10025.

(6)   Yudin, A. K. Macrocycles: Lessons from the Distant Past{,} Recent Developments{,} and Future Directions. *Chem. Sci.* **2015**, *6* (1), 30–49.

(7)   Marsault, E.; Peterson, M. L. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *J. Med. Chem.* **2011**, *54* (7), 1961–2004.

(8)   Stotani, S.; Giordanetto, F. Overview of Macrocycles in Clinical Development and Clinically Used. *Practical Medicinal Chemistry with Macrocycles*. August 18, 2017, pp 411–499.

(9)   Gao, M.; Cheng, K.; Yin, H. Targeting Protein-Protein Interfaces Using Macrocyclic Peptides. *Biopolymers* **2015**, *104* (4), 310–316.

(10)   London, N.; Movshovitz-Attias, D.; Schueler-Furman, O. The Structural Basis of Peptide-Protein Binding Strategies. *Structure* **2010**, *18* (2), 188–199.

(11)   Gartner, Z. J.; Tse, B. N.; Grubina, R.; Doyon, J. B.; Snyder, T. M.; Liu, D. R. DNA-Templated Organic Synthesis and Selection of a Library of Macrocycles. *Science* **2004**, *305* (5690), 1601–1605.

(12)   Tse, B. N.; Snyder, T. M.; Shen, Y.; Liu, D. R. Translation of DNA into a Library of 13 000 Synthetic Small-Molecule Macrocycles Suitable for in Vitro Selection. *J. Am. Chem. Soc.* **2008**, *130* (46), 15611–15626.

(13)   Giebel, L. B.; Cass, R.; Milligan, D. L.; Young, D.; Arze, R.; Johnson, C. Screening of Cyclic Peptide Phage Libraries Identifies Ligands That Bind Streptavidin with High Affinities. *Biochemistry* **1995**, *34* (47), 15430–15435.

(14)   Kawakami, T.; Ohta, A.; Ohuchi, M.; Ashigai, H.; Murakami, H.; Suga, H. Diverse Backbone-Cyclized Peptides via Codon Reprogramming. *Nat. Chem. Biol.* **2009**, *5* (12), 888–890.

(15)   Passioura, T.; Katoh, T.; Goto, Y.; Suga, H. Selection-Based Discovery of Druglike Macrocyclic Peptides. *Annu. Rev. Biochem.* **2014**, *83* (1), 727–752.

(16)   Katoh, T.; Sengoku, T.; Hirata, K.; Ogata, K.; Suga, H. Ribosomal Synthesis and de Novo Discovery of Bioactive Foldamer Peptides Containing Cyclic β-Amino Acids. *Nat. Chem.* **2020**, *12* (11), 1081–1088.

(17)   Ahlbach, C. L.; Lexa, K. W.; Bockus, A. T.; Chen, V.; Crews, P.; Jacobson, M. P.; Lokey, R. S.

Beyond Cyclosporine A: Conformation-Dependent Passive Membrane Permeabilities of Cyclic Peptide Natural Products. *Future Med. Chem.* **2015**, *7* (16), 2121–2130.

(18)   Over, B.; Matsson, P.; Tyrchan, C.; Artursson, P.; Doak, B. C.; Foley, M. A.; Hilgendorf, C.; Johnston, S. E.; Lee IV, M. D.; Lewis, R. J.; et al. Structural and Conformational Determinants of Macrocycle Cell Permeability. *Nat. Chem. Biol.* **2016**, *12*, 1065.

(19)   Clackson, T.; Wells, J. A. A Hot Spot of Binding Energy in a Hormone-Receptor Interface. *Science* **1995**, *267* (5196), 383–386.

(20)   Wells, J. A.; McClendon, C. L. Reaching for High-Hanging Fruit in Drug Discovery at Protein–Protein Interfaces. *Nature* **2007**, *450*, 1001.

(21)   Bullock, B. N.; Jochim, A. L.; Arora, P. S. Assessing Helical Protein Interfaces for Inhibitor Design. *J. Am. Chem. Soc.* **2011**, *133* (36), 14220–14223.

(22)   Watkins, A. M.; Arora, P. S. Anatomy of β-Strands at Protein–Protein Interfaces. *ACS Chem. Biol.* **2014**, *9* (8), 1747–1754.

(23)   Bogan, A. A.; Thorn, K. S. Anatomy of Hot Spots in Protein Interfaces. *J. Mol. Biol.* **1998**, *280* (1), 1–9.

(24)   Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Thornton, J. M. Shape Variation in Protein Binding Pockets and Their Ligands. *J. Mol. Biol.* **2007**, *368* (1), 283–301.

(25)   Rose, T. E.; Curtin, B. H.; Lawson, K. V; Simon, A.; Houk, K. N.; Harran, P. G. On the Prevalence of Bridged Macrocyclic Pyrroloindolines Formed in Regiodivergent Alkylations of Tryptophan. *Chem. Sci.* **2016**, *7* (7), 4158–4166.

(26)   Lawson, K. V; Rose, T. E.; Harran, P. G. Template-Constrained Macrocyclic Peptides Prepared from Native, Unprotected Precursors. *Proc. Natl. Acad. Sci.* **2013**, *110* (40), E3753--E3760.

(27)   Rose, T. E.; Lawson, K. V; Harran, P. G. Large Ring-Forming Alkylations Provide Facile Access to Composite Macrocycles. *Chem. Sci.* **2015**, *6* (4), 2219–2223.

(28)   Zhao, H.; Negash, L.; Wei, Q.; LaCour, T. G.; Estill, S. J.; Capota, E.; Pieper, A. A.; Harran, P. G. Acid Promoted Cinnamyl Ion Mobility within Peptide Derived Macrocycles. *J. Am. Chem. Soc.* **2008**, *130* (42), 13864–13866.

(29)   Curtin, B. H.; Manoni, F.; Park, J.; Sisto, L. J.; Lam, Y.-H.; Gravel, M.; Roulston, A.; Harran, P. G.

Assembly of Complex Macrocycles by Incrementally Amalgamating Unprotected Peptides with a Designed Four-Armed Insert. *J. Org. Chem.* **2018**, *83* (6), 3090–3108.

(30)   Lawson, K. V; Rose, T. E.; Harran, P. G. Template-Induced Macrocycle Diversity through Large Ring-Forming Alkylations of Tryptophan. *Tetrahedron* **2013**, *69* (36), 7683–7691.

(31)   Saha, I.; Dang, E. K.; Svatunek, D.; Houk, K. N.; Harran, P. G. Computational Generation of an Annotated Gigalibrary of Synthesizable, Composite Peptidic Macrocycles. *Proc. Natl. Acad. Sci.* **2020**, *117* (40), 24679–24690.

(32)   Tsunemi, T.; Bernardino, S. J.; Mendoza, A.; Jones, C. G.; Harran, P. G. Syntheses of Atypically Fluorinated Peptidyl Macrocycles through Sequential Vinylic Substitutions. *Angew. Chemie Int. Ed.* **2020**, *59* (2), 674–678.

(33)   Sisto, L. J.; Harran, P. G. Sulfane Transalkylations and Metal Catalyzed Allylic Substitutions for the Synthesis of Composite Macrobicyclic Peptides. *Tetrahedron Lett.* **2020**, *61* (24), 151986.

(34)   Sisto, L. J.; Harran, P. G. Syntheses of Hybrid Cyclopeptidyl [n]Sulfanes by Internal Alkyl Group Exchange. *Tetrahedron Lett.* **2020**, *61* (24), 151985.

(35)   Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; et al. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229.

(36)   Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; et al. An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature* **2020**, *580* (7805), 663–668.

(37)   Saha, I.; Harran, P. G. Virtual Screening for Chemists. *ACS In Focus*. American Chemical Society May 12, 2021, p 1.

(38)   Murzinski, E. S.; Saha, I.; Ding, H.; Strugatsky, D.; Hollibaugh, R. A.; Liu, H.; Tontonoz, P.; Harran, P. G. In Search of Small Molecules That Selectively Inhibit MBOAT4. *Molecules* . 2021.

(39)   Kojima, M.; Hosoda, H.; Date, Y.; Nakazato, M.; Matsuo, H.; Kangawa, K. Ghrelin Is a Growth-Hormone-Releasing Acylated Peptide from Stomach. *Nature* **1999**, *402* (6762), 656–660.

(40)   Yang, J.; Brown, M. S.; Liang, G.; Grishin, N. V; Goldstein, J. L. Identification of the Acyltransferase That Octanoylates Ghrelin, an Appetite-Stimulating Peptide Hormone. *Cell* **2008**, *132* (3), 387–396.

(41)   Yang, J.; Zhao, T.-J.; Goldstein, J. L.; Brown, M. S. Inhibition of Ghrelin O-Acyltransferase (GOAT) by Octanoylated Pentapeptides. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (31), 10750–10755.

(42)   Wortley, K. E.; del Rincon, J.-P.; Murray, J. D.; Garcia, K.; Iida, K.; Thorner, M. O.; Sleeman, M. W. Absence of Ghrelin Protects against Early-Onset Obesity. *J. Clin. Invest.* **2005**, *115* (12), 3573–3578.

(43)   Wren, A. M.; Small, C. J.; Abbott, C. R.; Dhillo, W. S.; Seal, L. J.; Cohen, M. A.; Batterham, R. L.; Taheri, S.; Stanley, S. A.; Ghatei, M. A.; et al. Ghrelin Causes Hyperphagia and Obesity in Rats. *Diabetes* **2001**, *50* (11), 2540–2547.

(44)   Dezaki, K.; Sone, H.; Koizumi, M.; Nakata, M.; Kakei, M.; Nagai, H.; Hosoda, H.; Kangawa, K.; Yada, T. Blockade of Pancreatic Islet-Derived Ghrelin Enhances Insulin Secretion to Prevent High-Fat Diet-Induced Glucose Intolerance. *Diabetes* **2006**, *55* (12), 3486–3493.

(45)   Kurashina, T.; Dezaki, K.; Yoshida, M.; Sukma Rita, R.; Ito, K.; Taguchi, M.; Miura, R.; Tominaga, M.; Ishibashi, S.; Kakei, M.; et al. The Beta-Cell GHSR and Downstream CAMP/TRPM2 Signaling Account for Insulinostatic and Glycemic Effects of Ghrelin. *Sci. Rep.* **2015**, *5*, 14041.

1. Computational Generation of an Annotated Gigalibrary of Synthesizable, Composite Peptidic Macrocycles

1.1 Introduction

Our lab seeks to identify novel 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.[16-21] 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 (a.k.a Tsuji-Trost reactions), and *N*-acyliminium ion-mediated cyclizations.[16-21] 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 endo-pyrroloindolines **3** and **5**.

10
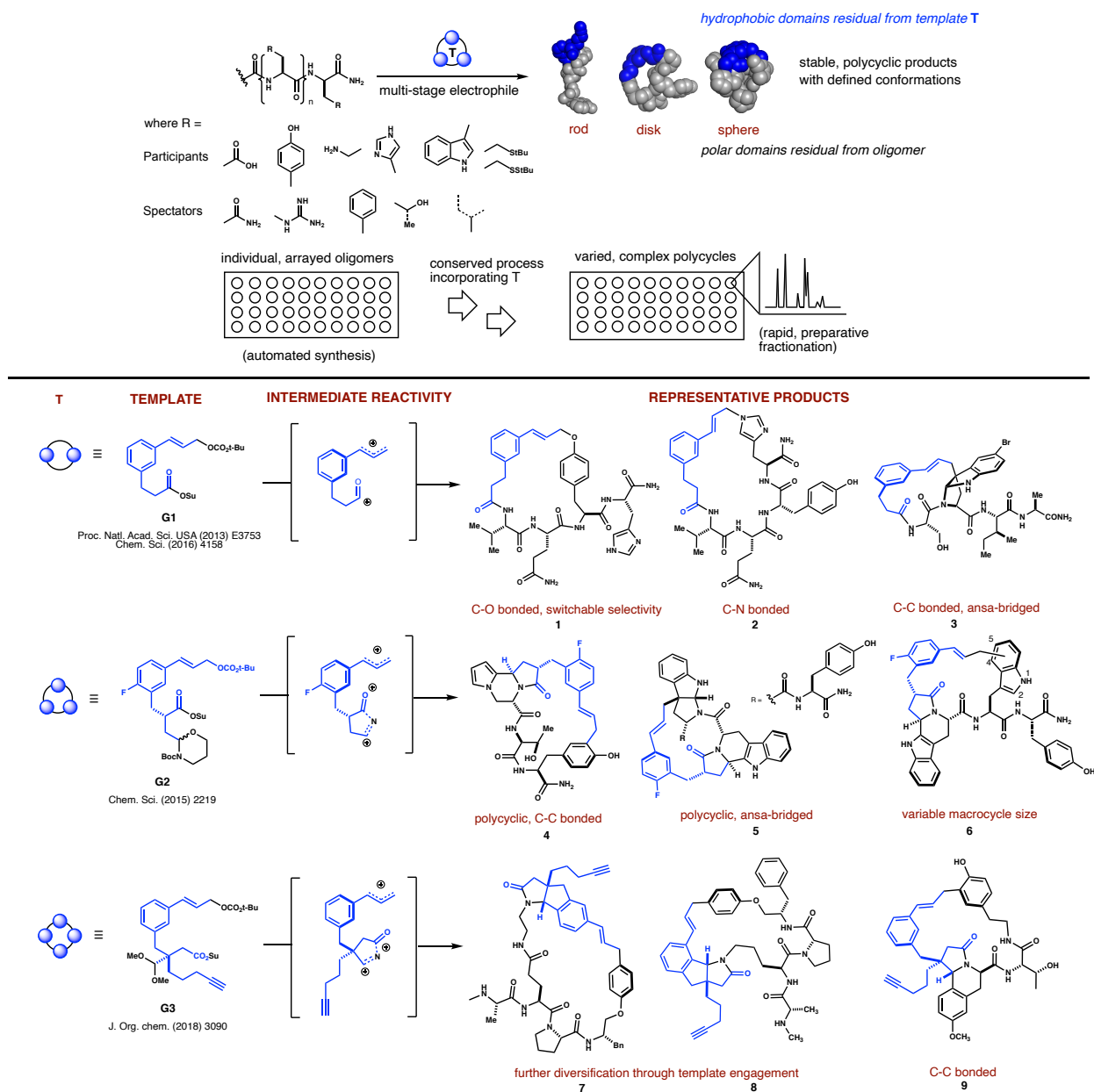
Figure 1. Generalized schematic and representative outcomes of our synthesis platform. Templates indicated in blue and oligomers indicated in grey.

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 β^2- and β^3- 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 (**C**omposite **P**eptide **M**acrocycle **G**enerator), which we have used to generate an *in-silico* library of >2 billion composite macrocycles resultant from multi step reaction sequences. We have also adapted conformational search methods[22] to create Confbuster++, which is able to generate three-dimensional conformations of library members rapidly. The ultimate goal for this work is to enable virtual ligand discovery for diverse structurally characterized protein targets.
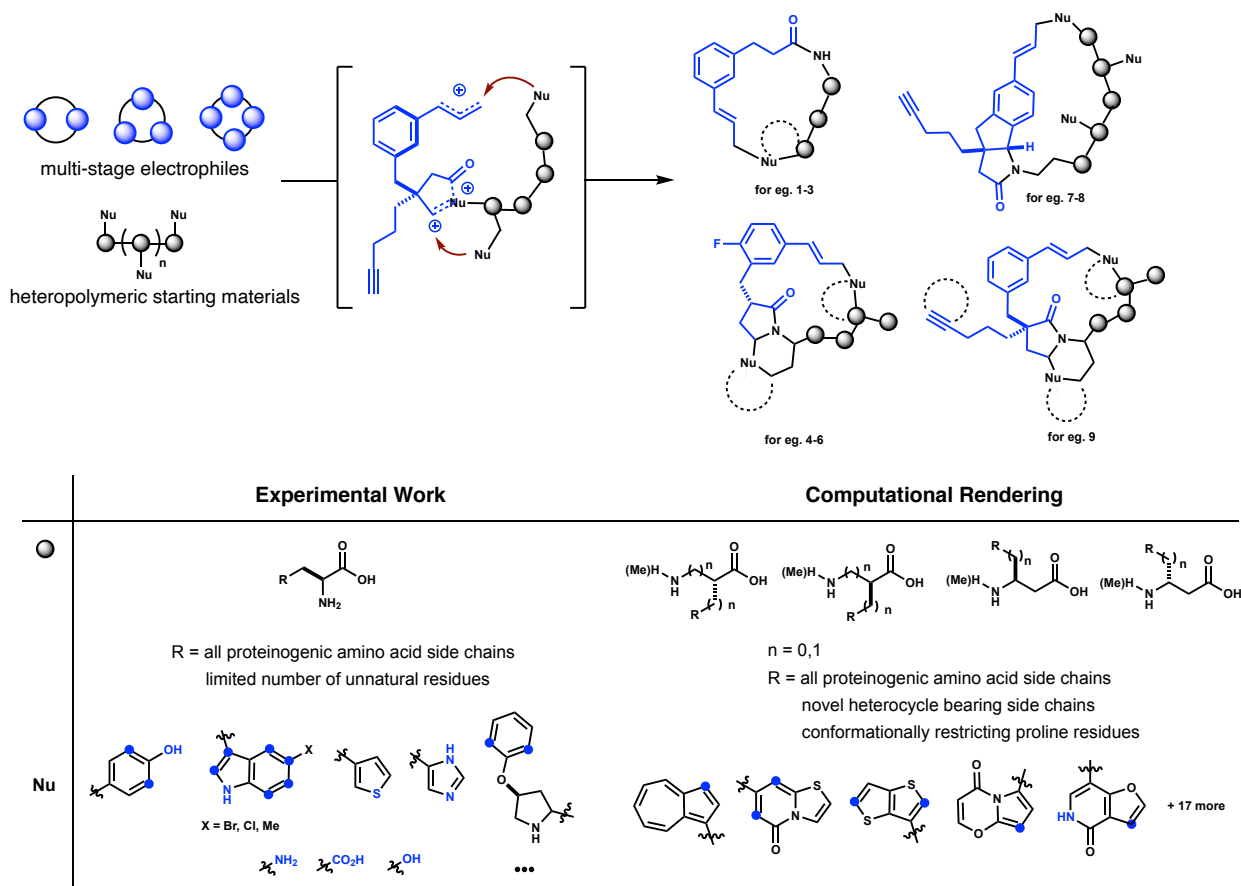
Figure 2. A computational rendering of our synthesis platform.

1.2 Library generation and conformational analysis

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

(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[23] (see Methods for details).

The platform is described in detail in the following sections, and a general workflow is schematized in 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 **G1-3**. 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. 5).

1.2.1 Building CPMG

Heterocycles **9**-**25** in Fig. 3 were extracted from the VEHICLe (virtual exploratory heterocyclic library) database built by Pitt *et al.*[24] This database contains a set of 24847 aromatic ring systems generated using a random forest-based method, of which over 3000 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 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 optimizations using ωB97X-D/def2TZVP (Gaussian 16 RevA.03)[25-26] were performed on all structures having multiple tautomeric forms. The most energetically favorable tautomeric state was included in subsequent

13

Formatted Building Blocks

Known Conformationally Restricting Proline Analogs
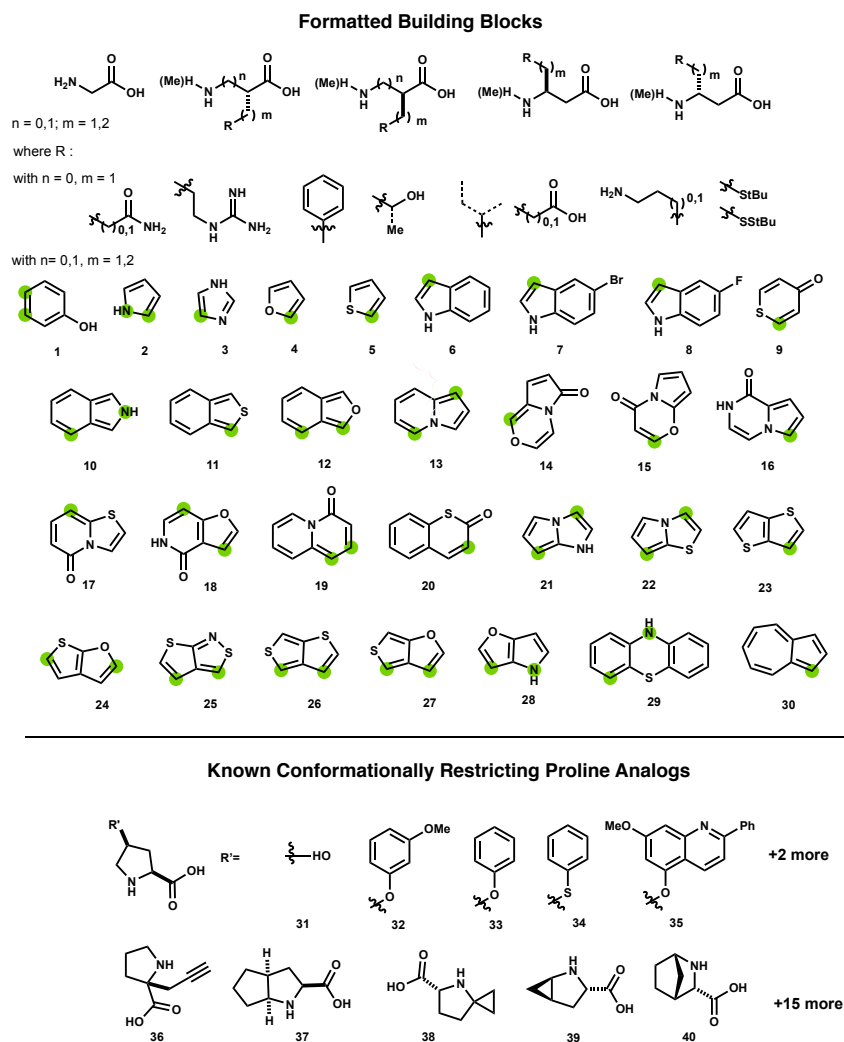
Figure 3. Substrate library (see Experimental Appendix, Fig. S1, for full list). Green dots indicate sites of incorporation into amino acids.

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 Experimental Appendix, Fig. S1 for full list), were added to the set of building blocks. Peptides of varying length (3 to 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 3 heterocyclic side chains. Additionally, the *C*-terminal carboxyl group of all trimers and tetramers

14

was capped with an *N*-ethyl-R unit (R = **1-30** in Fig. 3). Pentamers were excluded from this last step in anticipation of molecular weight cutoffs.
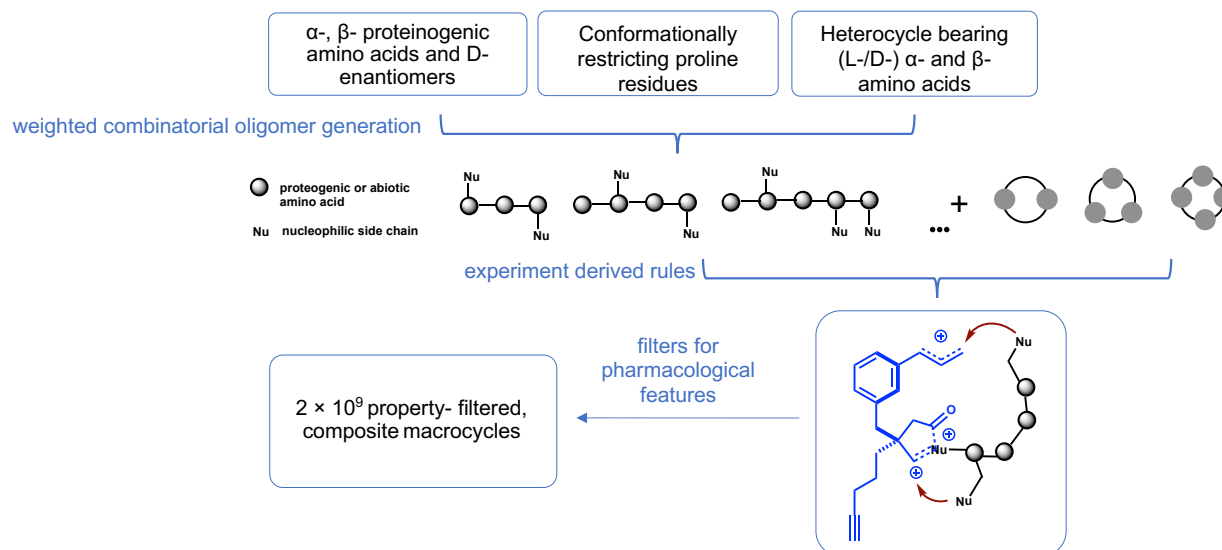
Figure 4. Schematized workflow for CPMG.

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 **G1-3** 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 opensource program, RegioSQM, published by Jorgensen *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.[27] We

analyzed heterocycles **9-30** 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 was fully consistent with experimental observations where available. Pictet–Spengler products were generated by allowing bond formations at RegioSQM predicted nucleophilic sites exactly 6 atoms away in **G2, (**e.g. **4-6** in Fig. 1) or 5 or 6 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 (eg. **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.[28] 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.[29] Computed values were incorporated into CMPG 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.[13-15] Scanning mono-*N*-methylation of secondary amides was applied to all generated macrocycles and only structures having MW ≤ 1200 g mol^{-1}, nrot ≤ 10 and TPSA ≤ 200 Å^2 were retained.

1.2.2 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 three-dimensional structures.[30] 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 towards generating *cis*- alkenes when applied to cyclic alkenes, regardless of the input two-dimensional 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,[31-32] for filtering out three-dimensional 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[33] *et al.* 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, sidechains 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,[34] and Barbeau *et al.*'s opensource ConfBuster.[22] 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.[35-39] 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.[35-38] A tool for an extensive conformer search is CREST.[39] This software package provides a conformer search at a higher level of theory using the GFN2-xTB tight binding DFT functional[40] 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 root mean square deviation (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
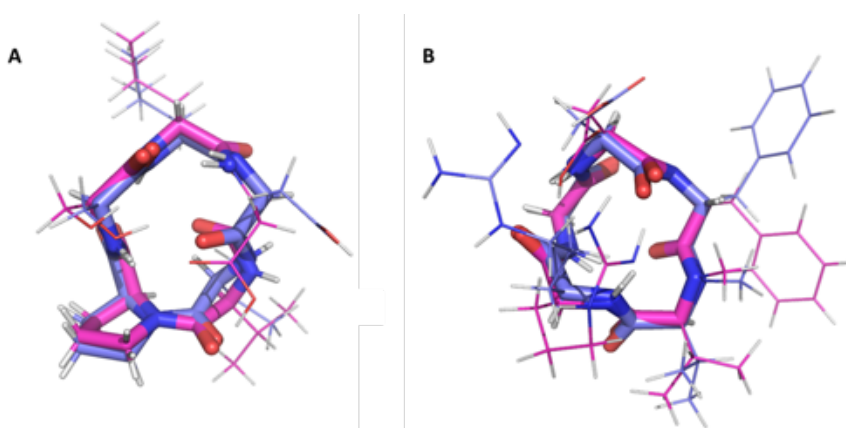
Figure 5. Overlay of the lowest energy conformer found by ConfBuster++ *(blue)* and CREST *(pink)* for **A.** cyclo-(Pro-Ser-leu-Asp-Val) and **B.** cyclo-(Arg-Gly-Asp-phe-([N-Me]Val)). The macrocycle backbone structures are highlighted.

published method based on molecular mechanic force fields[38] wherein ConfBuster++ is much less computationally expensive (see Experimental Appendix, Fig.S4).

Execution times for ConfBuster++ were benchmarked on UCLA's 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 4500 jobs, where each job was allocated a parallel environment of 8 shared memory processing units and 25 GB of RAM. 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, Fig. 9). In comparison, the extensive conformer search in CREST took 96 and 48 CPU hours for cyclo-(Pro-Ser-leu-Asp-Val) and cyclo-(Arg-Gly-Asp-phe-([N-Me]Val)), respectively.

1.3 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.[12,13,41] Although clear guidelines in this regard are yet to established, shape diversity of small molecule libraries has been cited as a fundamental indicator of functional diversity.[42] 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. 6B). This is in contrast to other

virtual databases and experimental datasets in the literature, wherein there exists a preponderance of rod-shaped molecules.[43-46] Our macrocycles occupy unexplored regions of chemical space and present opportunities for identifying novel ligands.
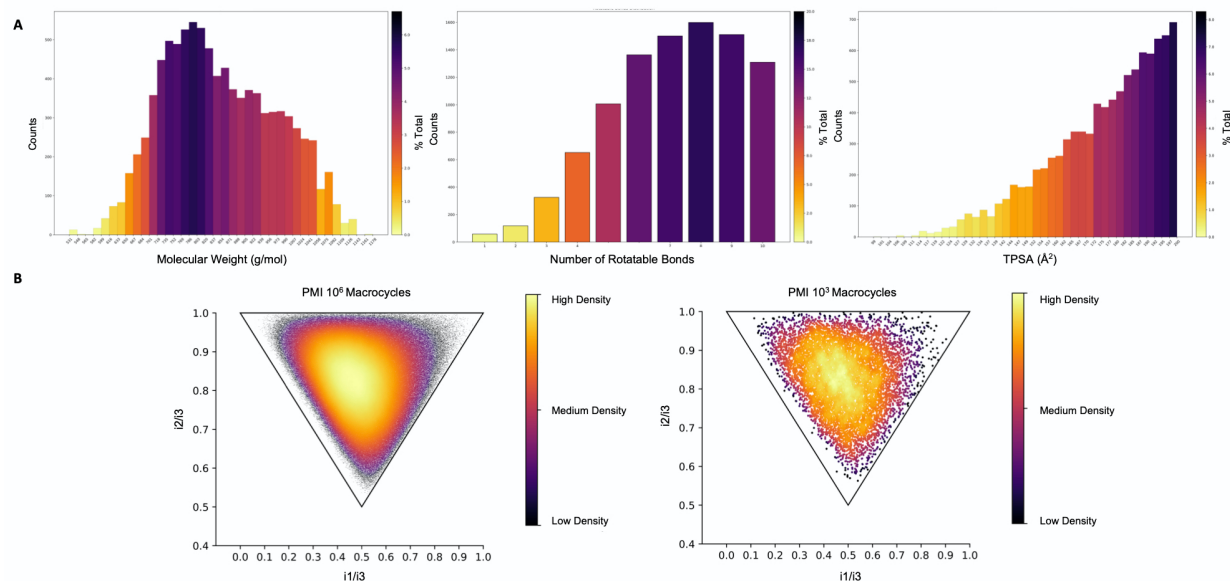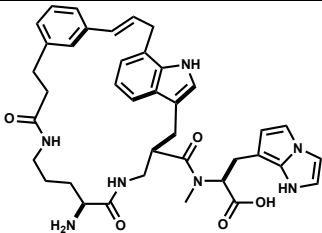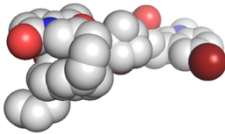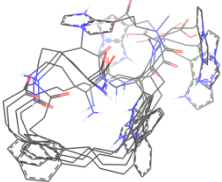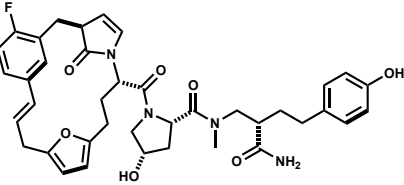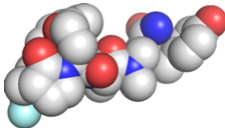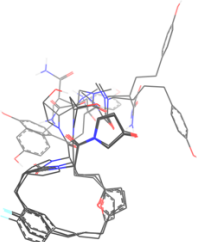

Figure 6. A. Histograms showing distributions of *(left to right)* molecular weight, total polar surface area and rotatable bonds across the subset of 10,000 structures; **B.** *(left to right)* PMI plot for 1 million randomly chosen structures; PMI plot for algorithmically selected subset of 10,000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

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. Towards this, we conducted principal component analysis (PCA) on the 1 million structures using five three-dimensional 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. 6B 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 **G1-3.** 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. 6A.

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.[12,13,41] 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.

Table 1. Representative structures at vertices of the PMI triangle; 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 5kcal mol^{-1} of the lowest energy conformer and over 0.5 RMSD of each other) are listed under 'Shape'.

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			
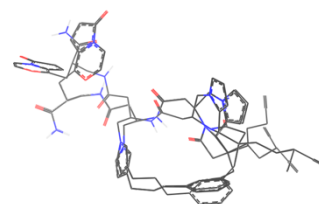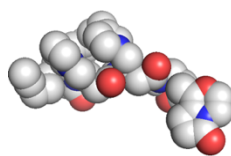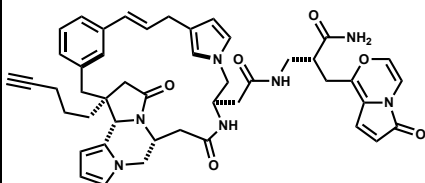
R3 795.9 g/mol 40 conf			
D1 816.0 g/mol 13 conf			
D2 728.8 g/mol 5 conf			
D3 882.1 g/mol 5 conf			
S1 762.9 g/mol 24 conf			
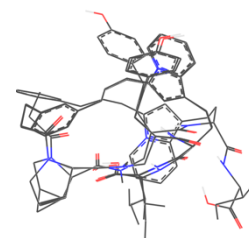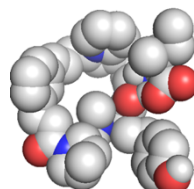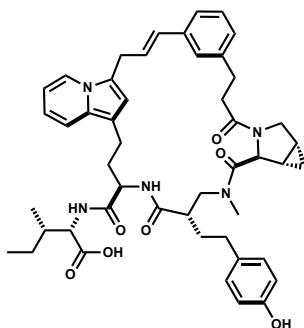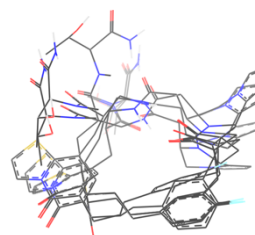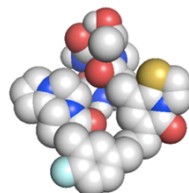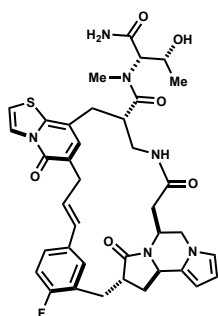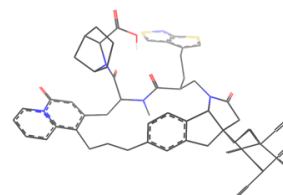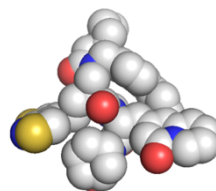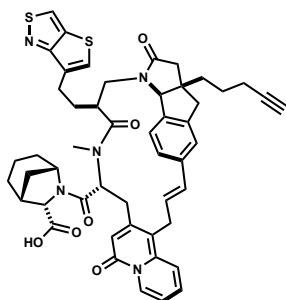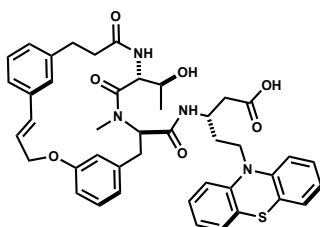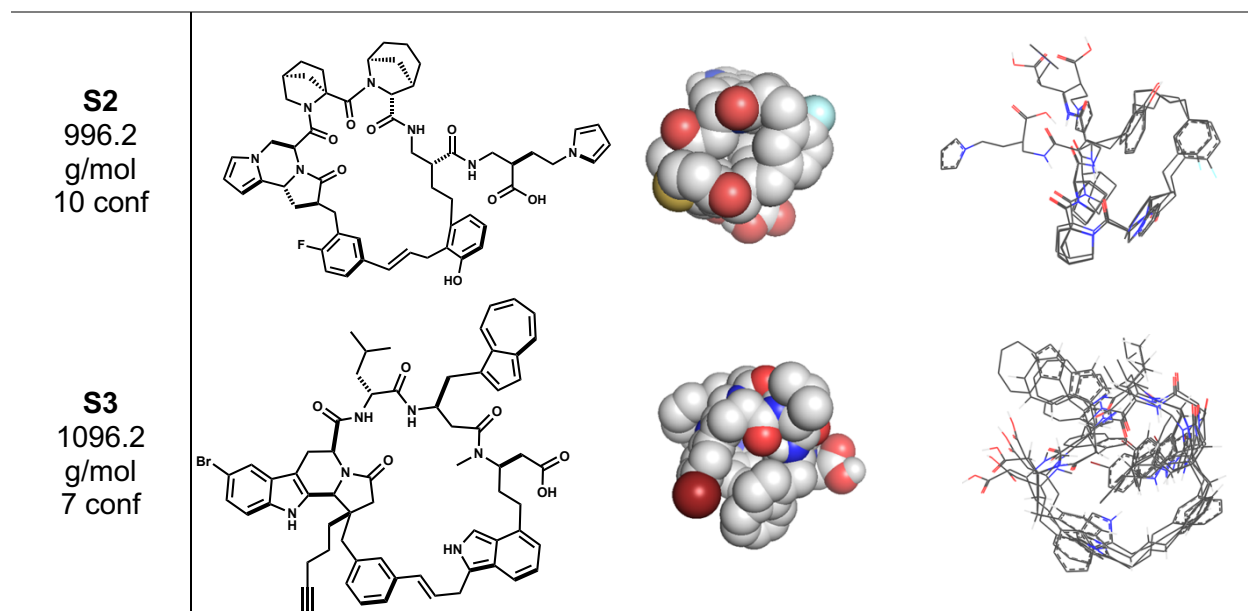
The larger objective of this work is to use constrained macrocycle libraries as a resource for virtual screening. For those studies to be successful, predicted binding interactions will need to be experimentally validated. Our prior experimental results suggest a majority of macrocyclization reactions simulated by CPMG will proceed.[16-21] 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 distance between reacting positions in acyclic precursors.[18,20] Heterocycles shown in Fig. 3 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[27] 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
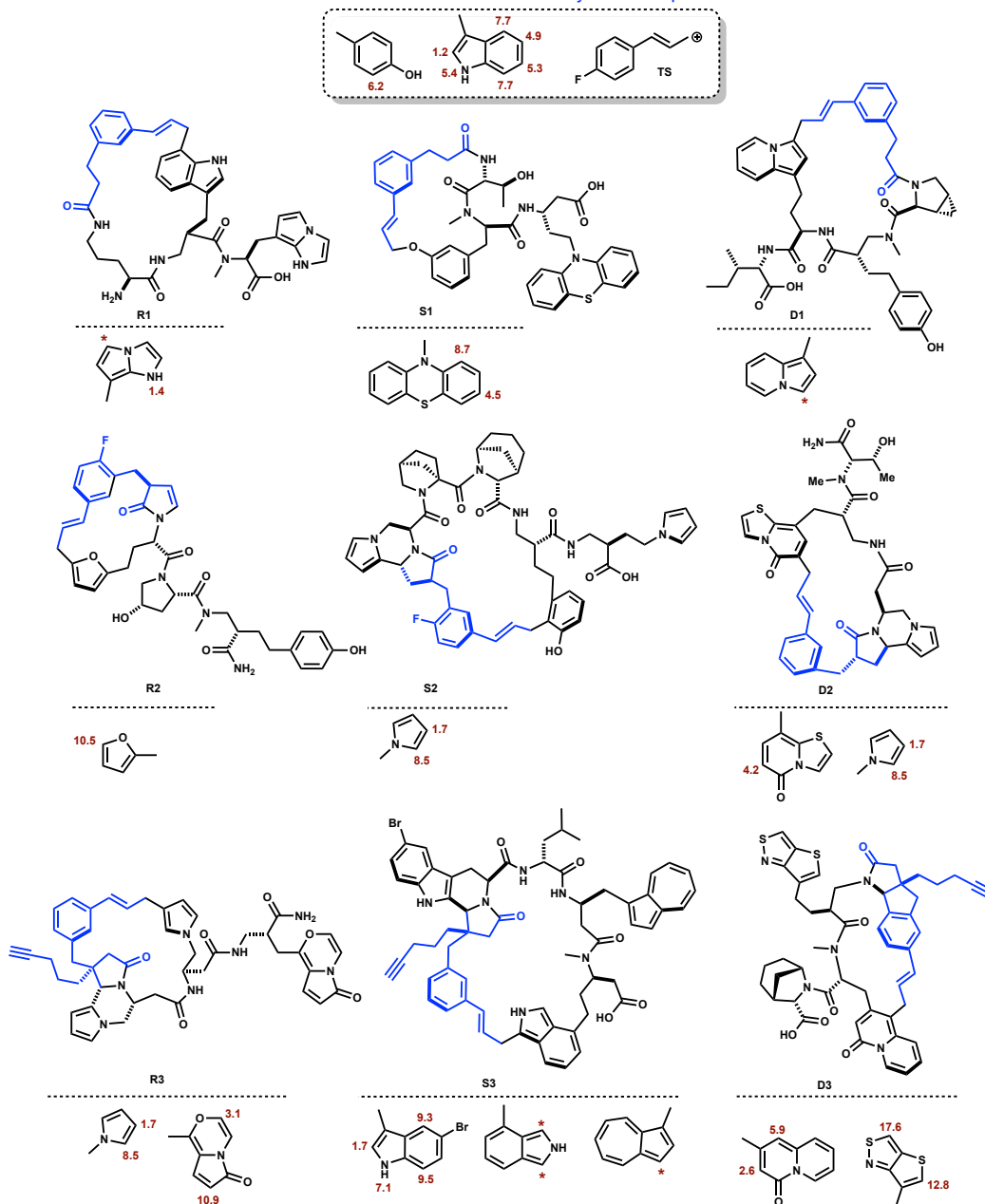
Figure 7. Calculated free energies of activation (ΔG‡) for the reaction of the indicated heterocyclic positions with the allyl cation, **TS**. ΔG‡ values in kcal/mol. Reaction sites with ΔG‡ values comparable to those calculated for phenol and indole are likely participants in ring forming electrophilic 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.

24

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 be 4.2 kcal/mol (Fig. 7), 2 kcal/mol lower than for the reaction with phenol at the ortho position (see Experimental Appendix, Fig. S5, Table S1). 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 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 CMPG library to be similarly accessible and manipulated.
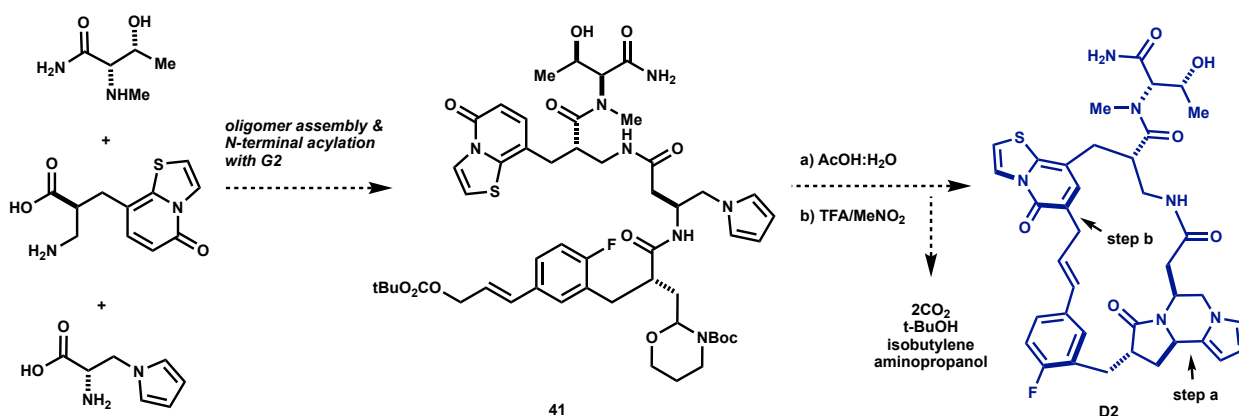
Scheme 1. CMPG generates complex structures that can be synthesized readily from amino acid monomers and functionalized templates.

25

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, 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.

1.4. Conclusions and outlook

There is a trend towards increasing complexity in small molecule drug discovery research.[42] 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 ($\alpha/\beta^2/\beta^3$, L-/D-, methylene/ethylene-appended) of each. Oligopeptides were combinatorially generated, resulting in 2,020,794,198 macrocyclic structures of multi-step

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.[47] 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.[47-49] In recent years, deep learning models for ligand docking, scoring functions, and virtual screening have also emerged.[50-51] 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.

1.5. 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.git 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,

27

referred to as **G1-G3** in this manuscript, we will refer to the **R**DKit **r**eaction **t**emplate as RRT in the following text.

1.5.1 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 (Experimental Appendix Fig. S2), 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 in Section 1.2.1. 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 Experimental Appendix Fig. S1), along with the linkage motifs (Experimental Appendix Fig. S2) were input to the SideChain 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 (Experimental Appendix Fig. S3), were subsequently passed to the MonomerGenerator component which produces a monomer for each combination of sidechain 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 in Section 1.2.1.

28

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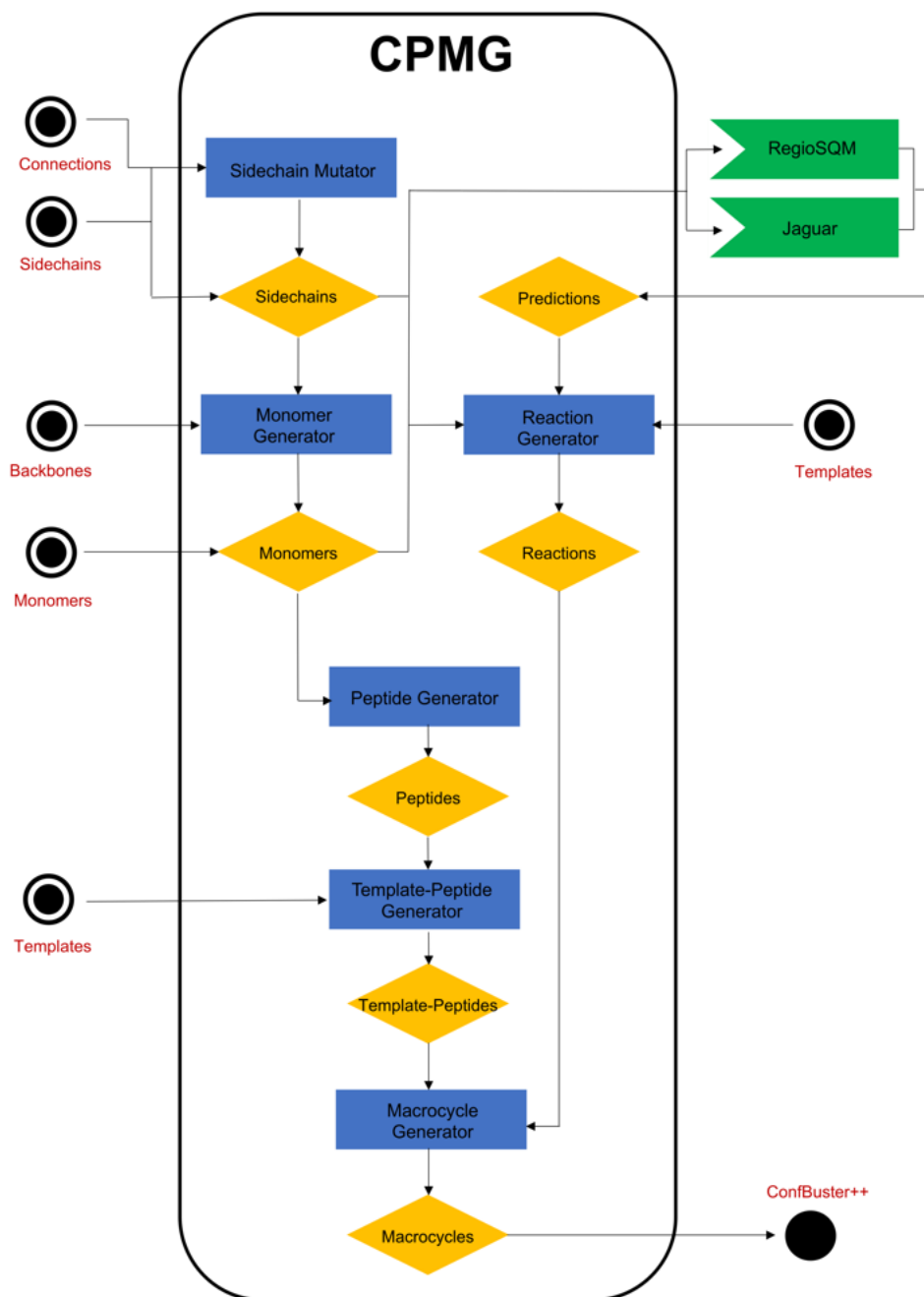
Figure 8. Graphical representation of CPMG and its components.

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 3 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 3 cinnamyl template-peptides. Peptide and template-peptide generation was performed together in an array of 3 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 mono-methylation 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 ≥ 1200 Da, number of rotatable bonds ≥ 10, or TPSA ≥ 200 $Å^2$. Macrocycle generation was performed with 1500 job job array (500 jobs per peptide length) where each job was allocated an 8-slot parallel environment and 12 GB of RAM.

1.5.2 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).

Algorithm 1 Main algorithm for generating low
energy macrocycle conformers

Inputs: $M, N_r, N_g, D_{min}, E_{max}$
1    $bonds[] \leftarrow$ **find_cleavable_bonds**(M)
2   **for** $bond$ **in** $bonds[]$ **do**
3   **cleave_bond**$(M, bond)$
4        $dihedrals[] \leftarrow$ **find_dihedrals**(M)
5        $conformers_r[] \leftarrow$ **rotate_dihedrals**$(M, dihedrals[], N_r)$
6   **for** $i = 0$ **to** N_r **do**
7   **reform_bond**$(conformers_{r,i})$
8            $conformers_g[] \leftarrow$ **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           $conformers_k[] \leftarrow$ **filter_conformers**$(conformers_g[], D_{min}, E_{max})$
13   **end for**
14   **end for**
15   **return** $conformers_k[]$

Figure 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 the number of conformations to generate via the genetic algorithm, D_{min} the minimum RMSD between conformers, E_{max} the maximum energy difference between the lowest and highest energy conformers.

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– 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 feed 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^{-1} 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 4500 jobs (1500 jobs per length of peptide) where each job was allocated an 8-slot parallel environment and 9.6 GB of RAM.

1.5.3 CREST.

Conformer searches in CREST Version 2.7.1[39] were performed using the iMTD-GC workflow in combination with the GFN2-xTB tight binding DFT functional[40] as implemented in XTB Version 6.2.[52] Default settings were used except for an energy window of 5 kcal/mol.

1.5.4 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 Spherocity 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.

1.6 References

(1)   Driggers, E. M.; Hale, S. P.; Lee, J.; Terrett, N. K. The Exploration of Macrocycles for Drug Discovery — an Underexploited Structural Class. *Nat. Rev. Drug Discov.*, **7**, 608 (2008).

(2)   Giordanetto, F.; Kihlberg, J. Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? *J. Med. Chem.*, **57**, 278–295 (2014).

(3)   Arkin, M. R.; Wells, J. A. Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing towards the Dream. *Nat. Rev. Drug Discov.*, **3**, 301–317 (2004).

(4)   Wrighton, N. C *et al.* Small Peptides as Potent Mimetics of the Protein Hormone Erythropoietin. *Science,* **273**, 458-463 (1996).

(5)   Livnah, O *et al.* Functional Mimicry of a Protein Hormone by a Peptide Agonist: The EPO Receptor Complex at 2.8 Å. *Science,* **273,** 464-471 (1996).

(6)   Gartner, Z. *et al.* DNA-Templated Organic Synthesis and Selection of a Library of Macrocycles. *Science*, **305**, 1601–1605 (2004).

(7)   Tse, B. N.; Snyder, T. M.; Shen, Y.; Liu, D. R. 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)   Giebel, L. B. *et al.* Screening of Cyclic Peptide Phage Libraries Identifies Ligands That Bind Streptavidin with High Affinities. *Biochemistry*, **34**, 15430–15435 (1995).

(9)   Kawakami, T.; Ohta, A.; Ohuchi, M.; Ashigai, H.; Murakami, H.; Suga, H. Diverse Backbone-Cyclized Peptides via Codon Reprogramming. *Nat. Chem. Biol.*, **5**, 888–890 (2009).

(10)   Passioura, T.; Katoh, T.; Goto, Y.; Suga, H. Selection-Based Discovery of Druglike Macrocyclic Peptides. *Annu. Rev. Biochem.*, **83**, 727–752 (2014).

(11)   Vinogradov, A. A.; Yin, Y.; Suga, H. Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges. *J. Am. Chem. Soc.*, **141**, 4167–4181 (2019).

(12)   Witek, J *et al.* Interconversion Rates between Conformational States as Rationale for the Membrane Permeability of Cyclosporines. *ChemPhysChem*, **18**, 3309–3314 (2017).

(13)   Matsson, P.; Kihlberg, J. How Big Is Too Big for Cell Permeability? *J. Med. Chem.*, **60**, 1662–1664 (2017).

(14)   Giordanetto, F.; Kihlberg, J. Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? *J. Med. Chem.*, **57**, 278–295 (*2*014).

(15)   Over, B *et al.* Structural and Conformational Determinants of Macrocycle Cell Permeability. *Nat. Chem. Biol.*, **12**, 1065 (2016).

(16)   Zhao, H. *et al.* Acid Promoted Cinnamyl Ion Mobility within Peptide Derived Macrocycles. *J. Am. Chem. Soc.*, **130**, 13864–13866 (2008).

(17)   Lawson, K. V; Rose, T. E.; Harran, P. G. Template-Constrained Macrocyclic Peptides Prepared from Native, Unprotected Precursors. *Proc. Natl. Acad. Sci.*, **110**, E3753--E3760, (2013).

(18)   Lawson, K. V; Rose, T. E.; Harran, P. G. Template-Induced Macrocycle Diversity through Large Ring-Forming Alkylations of Tryptophan. *Tetrahedron*, **69**, 7683–7691 (2013).

(19)   Rose, T. E.; Lawson, K. V; Harran, P. G. Large Ring-Forming Alkylations Provide Facile Access to Composite Macrocycles. *Chem. Sci.*, **6**, 2219–2223 (2015).

(20)   Rose, T. E. *et al*. On the Prevalence of Bridged Macrocyclic Pyrroloindolines Formed in Regiodivergent Alkylations of Tryptophan. *Chem. Sci.*, **7**, 4158–4166 (2016).

(21)   Curtin, B. H. *et al.* Assembly of Complex Macrocycles by Incrementally Amalgamating Unprotected Peptides with a Designed Four-Armed Insert. *J. Org. Chem.*, **83**, 3090–3108 (2018).

(22)   Barbeau, X., Vincent, A. T., & Lagüe, P. ConfBuster: Open-Source Tools for Macrocycle Conformational Search and Analysis. *Journal of Open Research Software*, **6**, 1 (2018).

(23)   Landrum G. RDKit: Open-source cheminformatics; http://rdkit.org.2016. Last accessed 04.10.2020.

(24)   Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R. Heteroaromatic Rings of the Future. *J. Med. Chem.*, **52**, 2952–2963 (2009).

(25)   Chai, J.-D.; Head-Gordon, M. Long-Range Corrected Hybrid Density Functionals with Damped Atom-Atom Dispersion Corrections. *Phys. Chem. Chem. Phys.*, **10**, 6615–6620 (2008).

(26)   Weigend, F.; Ahlrichs, R. 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).

(27)   Kromann, J. C.; Jensen, J. H.; Kruszyk, M.; Jessing, M.; Jørgensen, M. Fast and Accurate Prediction of the Regioselectivity of Electrophilic Aromatic Substitution Reactions. *Chem. Sci.*, **9**, 660–665 (2018).

(28)   Miyaura, N *et al. Cross-Coupling Reactions: A Practical Guide*; Topics in Current Chemistry; Springer Berlin Heidelberg, 2003.

(29)   Bochevarov, A. D. *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).

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

(31)   Rdkit. "EmbedMolecule Not Respecting Double-Bond Stereochemistry in Rings · Issue #1852 · Rdkit/Rdkit." *GitHub*, http://github.com/rdkit/rdkit/issues/1852. Last accessed 04/10/2020.

(32)   https://github.com/rdkit/rdkit/issues/1852. Last accessed 06/16/2020.

(33)   Jacobson, M. P. *et al.* A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins*, **55**, 351–367 (2004).

(34)   Sindhikara, D. *et al.* Improving Accuracy, Diversity, and Speed with Prime Macrocycle Conformational Sampling. *J. Chem. Inf. Model.*, **57**, 1881–1894 (2017).

(35)   Viles, J. H. *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**, 352−62 (1996).

(36)   Marelli, U. K. *et al.* Receptor-Bound Conformation of Cilengitide Better Represented by Its Solution-State Structure Than the Solid-State Structure. *Chem. - Eur. J.,* **20**, 14201−6 (2014).

(37)   Moruno, C. M. M. M.; Doedens, L.; Frank, A. O.; Kessler, H. Enhancement of Receptor Selectivity of Cilengitide by Multiple N- Methylation. *J. Pept. Sci.*, **16**, 45−46 (2010).

(38)   Kamenik, A. S.; Lessel, U.; Fuchs, J. E.; Fox, T.; Liedl, K. R. Peptidic Macrocycles - Conformational Sampling and Thermodynamic Characterization. *J. Chem. Inf. Model.*, **58**, 982–992 (2018).

(39)   Grimme, S. 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).

(40)   Bannwarth, C.; Ehlert, S.; Grimme, S. 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).

(41)   Witek, J. *et al.* Rationalization of the Membrane Permeability Differences in a Series of Analogue Cyclic Decapeptides. *J. Chem. Inf. Model.*, **59**, 294–308 (2019).

(42)   Galloway, W. R. J. D.; Isidro-Llobet, A.; Spring, D. R. Diversity-Oriented Synthesis as a Tool for the Discovery of Novel Biologically Active Small Molecules. *Nat. Commun.*, **1**, 80 (2010).

(43)   Brown, D.G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* 2016, *59* (10), 4443–4458. PMID: 26571338

(44)   Sauer, W. H. B.; Schwarz, M. K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *J. Chem. Inf. Comput. Sci.* 2003, *43* (3), 987–1003. PMID: 12767158.

(45)   Morrison, C. N.; Prosser, K. E.; Stokes, R. W.; Cordes, A.; Metzler-Nolte, N.; Cohen, S. M. Expanding Medicinal Chemistry into 3D Space: Metallofragments as 3D Scaffolds for Fragment-Based Drug Discovery. *Chem. Sci.* 2020, *11* (5), 1216–1225. PMID: 23117010

(46) Carles, F.; Bourg, S.; Meyer, C.; Bonnet, P. PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials. *Molecules* **2018**, *23* (4), 908.

(47) Lyu, J. *et al.* Ultra-Large Library Docking for Discovering New Chemotypes. *Nature*, **566**, 224–229 (2019).

(48) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.*, **31**, 455–461 (2009).

(49) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.*, **53**, 1893–1904 (2013).

(50) Lenselink, E. B. *et al.* Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminform.*, **9**, 45 (2017).

(51) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.*, **57**, 942–957 (2017).

(52) https://doi.org/10.5281/zenodo.3712016. Last accessed on 04/04/20.

2. Virtual Screening and Prioritization of CPMG Structures

2.1 Placing virtual screening in the context of experimental screening platforms

Broadly speaking, two main ligand screening strategies exist – system-based screening and target-based screening. System-based screens monitor how compounds elicit cellular responses in disease relevant models, while target-based methods interrogate the physical association of compounds with discrete protein targets.[1] An analysis of 113 first-in-class drugs between 1999 and 2014 found that 70% arose from target-based screens, while 30% arose from system-based methods.[2] The success of the target-based approach has been attributed in part to the success of large-scale operations such as random or high throughput screens (HTS), fragment-based screens and DNA-encoded libraries (DEL).

Since its advent in the 1990s, HTS has become a bedrock technology for drug discovery. Small molecule libraries are routinely screened for activity against targets without a need for *a priori* knowledge of biological properties or mechanism of action. HTS assays are typically performed in microtiter plates in 96-, 384-, or 1536-well formats.[3] A biological entity of interest is incubated with a unique compound in each well, and analyzed. Robotics and automation technologies allow for handling of large numbers of compounds, enabling data collection at a rate of ~100,000 points per day. An estimated 30% of clinical candidates originate from such empirical screens.[4] Examples include DPP IV inhibitor sitagliptin (Januvia) and factor Xa inhibitor rivaroxaban (Xarelto), to name just two.[5] While previously limited to lead discovery in large pharmaceutical firms, these facilities are now increasingly used for basic and applied research in academic institutions as well.[6]

While clinical candidates are rarely identified directly from HTS, these efforts seed chemistry optimizations and SAR (structure activity guided relationships) studies. HTS offers the unique advantage that it can be used on a broader range of targets than approaches that require upfront knowledge of the target or ligand. Intuitively, this suggests that the success of a screening

campaign hinges on library size, quality (chemical purity) and, importantly, the structural diversity and novelty of library members. Early compound collections for HTS were assembled mainly from pharmaceutical companies' internal activities, as well as acquisitions of largely non-exclusive, commercially available compound sets.[5] The advent of combinatorial chemistry accelerated the growth of libraries, but these methods tended to produce large libraries with limited diversity. Given that novel or therapeutically challenging targets often necessitate new chemotypes, an emphasis on incorporating library members from underexplored or unprecedented chemical space has emerged.[7,8]

Consider, however, that the number of possible small organic molecules is estimated to be 10^{60}.[9] To put that number in context, the CAS database – one of the world's largest – annotates ~ 10^8 known compounds, and just a fraction of those have been synthesized. Thus, even billions of molecules constitute a tiny fraction of potentially bioactive structures, and brute screening is intractable. Moreover, physical compound collections for HTS must be sourced, housed, distributed, and screened, all of which require equipment, buildings, people, and laboratory supplies. In fact, the cost of developing a new drug is roughly $2 billion, and lead identification and optimization contribute approximately a quarter of that amount.[10] This raises the question – given the practical constraints of time and resources, what is the optimum composition of a screening collection?

In February 2017, the Danish biopharmaceutical company Nuevolution announced it had created a library of 40 trillion discrete molecules for screening. This number may seem implausible, and in the context of HTS, it would be. But all of Nuevolution's compounds were stored in a single Eppendorf tube using DNA-encoding technology and could be handled by one person for screening. In terms of throughput, DNA-encoded libraries (DEL) have revolutionized screening. Since being described as a 'thought experiment' by Lerner and Brenner over 25 years ago, DEL screens have recently become a mainstay in drug discovery.[11] The most common way of using DEL treats the DNA like a bar code. The idea is to tag chemical building blocks with

unique DNA sequences and divide these into wells on a plate where they can react with other small organic molecules (Fig. 1). Thereafter, a second sequence of DNA is ligated to the resulting products, thus encoding the reactions that occurred. This process can be repeated to create millions of compounds in just a few iterations. The final molecules are pooled and combined with a biological target of interest, such as an enzyme, and anything that doesn't bind to the target is washed away. After several iterations of washing, the remaining binders are analyzed by their DNA tags *via* PCR amplification.[12] Other forms of DEL harness the specificity of DNA base-pairing to additionally direct reactivity. The most successful application of the latter approach is towards synthesizing and identifying macrocyclic peptide lead compounds.[13]
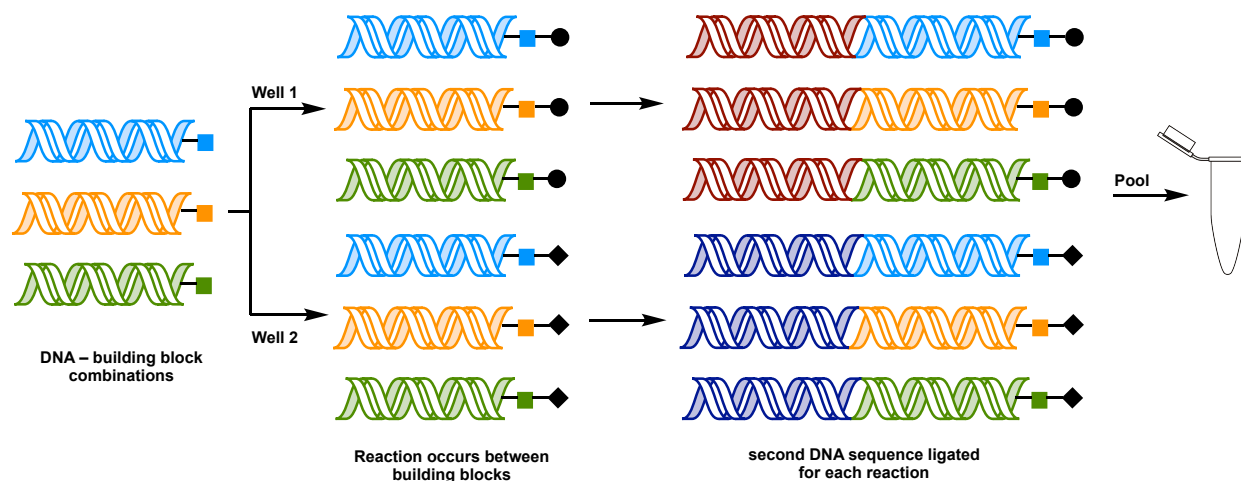
Figure 1. DNA encoding technology allows small molecules to be tagged with unique DNA sequences. These tags act as 'bar codes' for the rapid transformation of bringers in high throughput screens.

Despite DEL's advantages, the technology is not without its challenges. For instance, in the absence of biochemical assays, target binding does not necessarily imply that hits will show desired activity. Another significant bottleneck is that any chemistry used to create a DEL must be operable in an aqueous environment, and at pH > 4. The most common chemistries that can support high yielding, DNA-compatible reactions include amidations, $S_N Ar$ displacements, Suzuki-Miyaura couplings, reductive aminations, Fmoc-removal and sulfonylations.[12] Ring-closing metatheses and cross coupling reactions have been achieved, but these typically require large amounts of catalysts.[14,15] While ongoing efforts will undoubtedly expand the repertoire of DNA-

compatible reaction chemistry, existing constraints limit the types of structurally intricate molecules that can be prepared and evaluated in this format.

The size and complexity of chemical space thus make computational tools particularly attractive. An *in-silico* library not only eliminates expenses associated with maintaining real compound collections, but provides other advantages as well. For instance, the physiochemical properties of molecules can dictate their ability to not only bind to targets, but also passively permeate cell membranes and resist metabolic degradation. These properties can be extensively annotated in virtual libraries. Using these annotations, compounds having empirically derived, 'drug-like' structural features can be readily identified and prioritized. Not surprisingly, several large pharmaceutical firms now incorporate *in-silico* screens into their discovery pipelines. In the context of academic research, *in-silico* screening exercises overcome the limitations of scale. Medicinal chemistry campaigns typically require extensive resources to synthesize, analyze and assay lead compounds. As a result, synthetic platforms designed to generate potentially bioactive molecules often remain underexplored. Virtual, albeit hypothetical, molecules present the opportunity to visualize chemical space and prioritize synthesis without the associated costs. This chapter documents our efforts in deploying the CPMG giga-library (Ch. 1) in virtual screening exercises and creating prioritization filters to identify new ligands for structurally characterized proteins.

2.2 Comparing CPMG to other virtual libraries

Several pharmaceutical companies maintain in-house virtual libraries, but to the best of our knowledge, these are proprietary. The first such database was created by Pfizer in enumerating a collection of 10^{16} virtual compounds (PGVL).[16] The most recent and also the largest existing virtual library was presented by Merck KGaA. The so-called MASSIV space, if one could enumerate it, offers a dizzying 10^{20} molecular possibilities using in-house synthetic procedures.[17] In comparison, the number of approved drugs in DrugBank is just over 10,000.

Lilly,[18] Boehringer-Ingelheim[19] and AstraZeneca are other companies that also have their own virtual libraries. In 2019, several big pharmaceutical firms, including those mentioned above, announced the MELLODDY project, which would allow participating firms to gain access to each other's screening libraries for building and training predictive models.[20] Although privacy and control over all intellectual property remains with individual firms, the adoption of *in-silico* drug discovery methods by big pharma on the scale of the MELLODDY project is a testament to the approach's advantages.

Public databases play a crucial role in computer-aided discovery because they provide data for building knowledge-based models that facilitate the design of new, drug-like molecules. PubChem, ChEMBL, NCI, ChemSpider, DrugBank, Scifinder and Reaxys, among many others, contain large repositories of chemical information such as structural and physical properties of compounds, experimental procedures, reaction pathways and biochemical data. Importantly, and in contrast to the other virtually enumerated libraries, the vast majority of structures included in these databases have been previously synthesized and characterized. These databases are thus especially important in AI-driven applications that attempt to score synthetic tractability. One database in particular – DUD-E (Database of Useful Decoys) – is worth noting as a useful tool for benchmarking the performance of new docking or screening protocols. DUD-E is freely available and contains 102 protein targets, 22,886 known active ligands and 50 decoys for each active ligand. The protein targets within this set include GPCRs, kinases, proteases and nuclear receptors.[21] Its main purpose is to serve as a benchmarking dataset to ensure that new docking methods can distinguish known active molecules from decoys.

Increasingly, libraries containing hypothetical but synthetically tractable molecules derived from reaction chemistry are becoming available to the public. The first and largest of these in terms of ready-to-dock 3D structures is ZINC.[22] ZINC was launched in 2005 as a database of 3.5 million small molecule compounds compiled from vendor catalogs. At the time of this work, ZINC has just under a billion structures, and the majority of these are commercially available. A

significant fraction of these also have associated three-dimensional structures that are ready to dock. The growth of ZINC was led by the Shoichet and Irwin labs at UCSF who, in a recent study, generated new structures by importing several thousands of building blocks from Enamine, a Ukraine-based chemical vendor.[23] Although not strictly speaking a virtual database provider, Enamine provides a so-called 'library on demand' service. Their molecular building blocks, which are available online, can, on request, be subjected to time-tested reaction sequences to generate products that are delivered to customers in approximately 3 weeks. If enumerated, the chemical space of the Enamine catalogue (referred to as Enamine REAL Space) would be about 14 billion compounds. Less than 3% of the compounds in the Enamine space are commercially available elsewhere, and their success rate of synthesis is reported to be about 80%. The company also maintains the Enamine 'REAL database,' which is a smaller catalogue of 1.36 billion fully enumerated molecules. Another comparable library is SCUBIDOO, containing 21 million products.[24] These molecules were generated from approximately 8000 building blocks in 58 robust reactions by combining a maximum of two blocks in each case. These generated structures are not commercially available but are synthetically tractable.

Another approach, adopted by the Chemical Space Project, is to enumerate all possible molecules that obey the rules of valence. Using this idea, the Raymond group at the University of Berne developed the generated database (GDB), which in its first version contained 26.4 million molecules based on 11 heavy atoms.[25] As of the most recent release, GDB-17 contains 166 billion molecules made up of 17 heavy atoms.[26] The algorithms for GDB proceed by (i) exhaustively enumerating molecular skeletons, (ii) eliminating those with excessive strain, (iii) refining the remaining ones by enumerating bond orders and atom types, including heavy atoms such as C, N, O, S and halogens and (iv) eliminating molecules that violate rules of valence or contain unstable bonds. Portions of the GDB have been used for virtual screening, but the database has also found utility in other applications. A recently published machine-learning model for predicting molecular quantum properties used a portion of the GDB as its training set.[27] This study

exemplifies that while many virtual chemical libraries were built with virtual screening applications in mind, they possess a wide range of utilities beyond their intended use.

Given the deficits of docking software, a large percentage of hits are expected to be false positives, and thus ease of access to virtually enumerated molecules is critical to the success of large screening exercises. While many chemical libraries reasonably prioritize this aspect of library building, the issue can present itself as a double-edged sword. Prioritizing ease of access often results in library molecules being low in molecular weight and planar. This presents challenges for mediating therapeutic targets such as protein-protein interactions (PPIs), which have shallow pockets or crevices with large hydrophobic areas. Molecules that disrupt these systems tend to be larger and more lipophilic than traditional small molecule drugs and often fall into the category of bRo5 (beyond the rule of 5 – *vide infra*) compounds.[28–30] The remaining virtual libraries, including one designed in our own lab (CPMG),[31] enumerate this part of chemical space (Table 1).[32–34]

Table 1. Existing Virtual Chemical Libraries

Category	Database	Composition	Utility
Proprietary	Lilly PLC 2016 Boehr-Ing. BICLAIM 2012 Pfizer PGVL 2008 AstraZeneca AZ Space Merck MASSIV 2018		Not publicly available
Public repositories	CAS ChEMBL PubChem ChemSpider DrugBank DUD-E	Many, if not most, of these structures have been previously synthesized and/or reported in the literature	Enable synthetic/biochemical research Provide data for building training and test sets, which in turn can be used for benchmarking new methods Enable virtual screening
Public; most molecules commercially available	Chemical vendors (Sigma-Aldrich, etc.)	Mix of precedented and unprecedented, lower	Enabling ultra large virtual screening exercises, where a high

	Enamine REAL Database Enamine REAL Space WuXi AppTec LabNetwork Collection ZINC15	molecular weight compounds	number of false positives is likely Provide access to make on demand libraries
Public; molecules may be commercially unavailable	GDB17 CPMG (constrained peptidomimetic macrocycles) CH/PMUNK (heterocycles, drug like molecules) V1M (macrolides) CycloPs (cyclic peptides)	Majority are unprecedented but synthetically tractable structures; occupy both Ro5 and bRo5 chemical space	Provide access to new chemical entities for virtual screening May prove to be useful for mediating challenging therapeutic targets without known binders

While synthetic accessibility remains a priority, the compounds in this category are inherently larger and often more structurally complex on account of the targets they are designed to bind. For instance, whereas many of the databases mentioned earlier are compiled by combining small molecule fragments in one- or two-step reaction procedures (such as the majority of final products in the MW ~ 300 Da chemical space), CPMG anticipates macrocyclic products of multi-step reaction sequences, leading to compounds that on average have molecular weights of ~ 800 Da. Despite being larger, these compounds arise from oligopeptidic precursors resultant from simple amide bond linkages of amino acid precursors and derivatives. Final products are afforded by engaging aromatic side chains on the polyamides using well-established protocols such as Friedel Crafts alkylation, metal-catalyzed allylic substitutions and *N*-acyliminium ion-mediated cyclizations. Structurally rigidifying features such as polycyclic motifs and *ansa*-bridges are built into the molecules as a direct consequence of the reaction mechanism. The resulting macrocycles retain ancillary polar groups, yet can achieve a useful balance of cell permeability and aqueous solubility.[35–40] These features are expected to improve target binding and the ability to passively transverse lipid membranes and resist proteolytic degradation.[41,42]

The databases listed in the last category are young, and their utility remains to be demonstrated; however, they may provide a crucial, albeit specific, role in virtual discovery screening. Examining PMI plots of available molecular databases reveals that a preponderance of rod-shaped molecules exists in most databases, but a library like CPMG covers underexplored parts of chemical space (Fig. 2). In a similar vein, empirical analysis of orally administered FDA approved drugs suggests that bro5 compounds having molecular weight <1000 Da, total polar surface area <250 $Å^2$ and fewer than 20 rotatable bonds are more likely to be bioavailable.[28,41–43] Given that virtual libraries offer the unique opportunity to annotate and prioritize molecules in light of such guidelines, the vast majority of CPMG structures have molecular weights between 700-850 Da, 6-10 rotatable bonds and a maximum polar surface area of 200$Å^2$. With these advantages in mind, the remainder of this chapter documents our efforts in deploying CPMG as a novel resource for driving protein-ligand discovery.
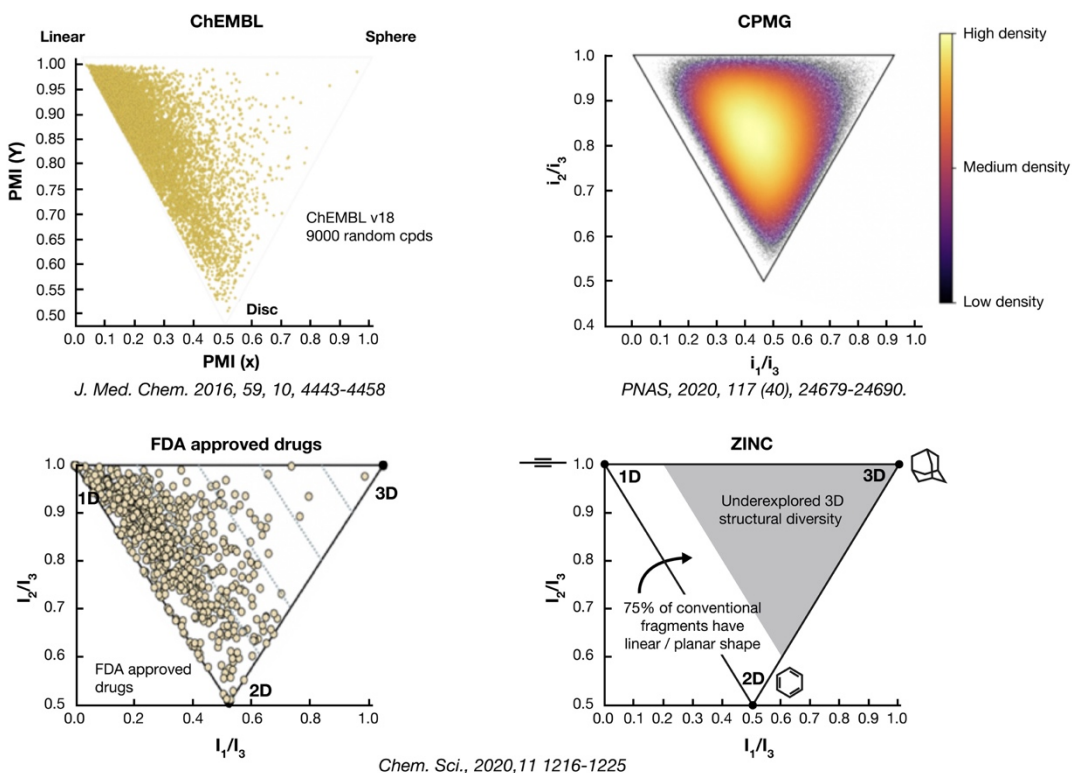

Figure 2. An analysis of PMI plots for existing datasets reveals a preponderance of molecules occupying the rod-disk axis, likely due to an overrepresentation of Ro5 compliant molecules in these datasets. Libraries incorporating greater numbers of bRo5 compounds, such as CPMG, can provide access to underexplored parts of chemical space.

2.3 The docking paradigm

Docking algorithms model the interactions between small molecules and proteins at the atomic level and thereby estimate the energy of binding. Docking results can be used to infer bioactivity, to characterize the behavior of small molecules in binding sites and to elucidate biochemical processes. The docking process involves two basic steps: sampling (or pose prediction) and scoring. The first step predicts ligand configuration and conformation in the protein binding site(s), while the second step calculates binding affinity values for the predicted pose(s).[44–46] It is worth noting that while docking algorithms are able to assess binding interactions, they cannot distinguish agonists from antagonists, and the term 'inhibitor' is often used loosely in the literature to indicate small molecule binders. Similarly, protein targets are often referred to as 'receptors,' independent of their true biological identity.

For many protein targets, binding site locations can be gauged by comparing a target protein to a family of proteins in the PDB (Protein Data Bank) that share similar functions and are co-crystallized with other ligands. However, 'blind docking' may also be performed on targets without known active sites. Many computational methods can locate putative ligand binding sites on protein surfaces using geometric criteria to find clefts and surface depressions.[47] This issue is discussed in greater details in section 2.5. Fig. 3 shows the typical workflow employed in our docking studies. Regardless of whether the active site was known, the basic inputs of our screening pipelines were a target structure – either experimentally resolved or computationally modelled – and a library of small molecules. Structure preparation involved assigning the correct charges and torsions, among other features, to both the target protein and ligand molecules. This is crucial for producing high quality data and occurs outside the main docking program. (Pre-processing details for docking are documented in scripts provided in the Experimental Appendix.) Following processing, each compound in the input library was virtually docked into a defined binding site by sampling various poses to achieve optimal ligand-protein complementarity of steric
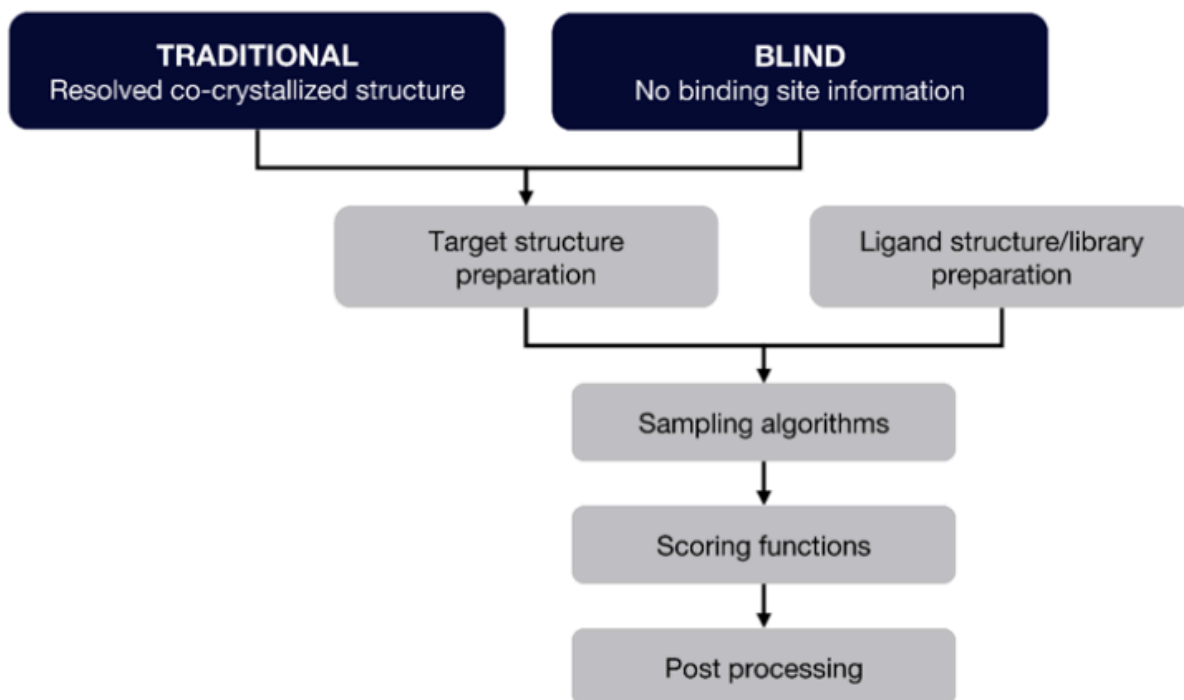
Figure 3. Core workflow of screening pipelines employed in this chapter.

and physiochemical factors. Depending on the docking program used, a mathematical algorithm then quantitatively evaluated the fitness between the docked compound and target to generate a score. The scoring function discriminates favorable interactions between the ligand and protein from unfavorable interactions – in other words, binders from inactive compounds. In any docking workflow, two conflicting requirements must be met – the desire for accuracy and the desire for computational efficiency. As a result, scoring functions are designed to estimate, rather than calculate, the exact binding affinity by adopting various assumptions and simplifications. Scoring functions can be broadly classified into force-field-based (FF), empirical and knowledge-based methods and the studies below employ all three in different implementations (*vide infra*). In recent years, consensus scoring methods have also become popular, whereby programs combine a few, or several, different scores to assess ligand poses.[48] A hit is identified when a pose scores well under several different scoring schemes. Adoption of this strategy has been demonstrated to

improve library enrichments,[49] and the utility of this approach is highest when terms in the different scoring functions are uncorrelated.

Docking is computationally challenging because there are many ways of interfacing two molecular surfaces. For instance, even as rigid bodies, combing all patches of the surface of a protein with all patches of a second protein takes 10^7 trials.[50] While ligands are typically much smaller, the computational problem becomes equally profound when considering the virtual screening of billions of compounds. With six degrees of translation and rotation and multiple conformational degrees of freedom for each entity, implementing full conformational flexibility in docking is challenging. With computational constraints in mind, the earliest elucidation of the protein-ligand binding mechanism was the *lock-and-key* theory proposed by Fischer, in which the ligand fits into the target as a key.[51] Accordingly, both the ligand and protein target are treated as rigid bodies, thereby significantly limiting the search space. The rigid body approximation has been justified in some cases by comparing protein structures in their bound and unbound states, where only side chains in the active site are found to change conformations. Additionally, ligand flexibility under this model can be addressed by using a set of pre-computed ligand conformations.[52] However, since proteins, and especially enzymes, show conformational flexibility, Koshland proposed a more sophisticated *induced fit* theory that treats both the ligand and protein as flexible entities.[53] Koshland's model suggests that a protein's active site is continually reshaped by interactions with the ligand during the binding process. In line with this, typical docking workflows assume that the ligand binding pose is near the protein-ligand complex's global energy minimum.[54] Since this ligand binding pose is the position of the ligand co-crystallized with the respective protein, a co-crystal structure of a protein complex in the PDB is often a suitable starting point for docking programs.

It is worth pointing out differences between the induced fit model and the models that dominate biochemical literature. In the latter, a protein is assumed to pre-exist in a number of energetically similar conformations. Under this assumption, biochemists define *conformational*

selection as the process by which a ligand selectively binds one of the conformers. As a result of binding, the percentage of that conformer in the overall protein population increases, creating the observed effect.[55] Conversely, *conformational induction* is used to describe a process where a ligand converts the protein into a conformation that it would not normally adopt in its unbound state. Importantly, the induced fit and biochemical models are equivalent from a thermodynamic point of view. However, the conformational selection and induction models may be more conceptually useful when considering and exploiting the phenomena encountered during drug design.[56] For instance, consider the energy landscape depicted in Fig. 4, where a protein exists in two conformations – P and P*, with an interconversion energy of $\Delta G_{convert}$. Although the
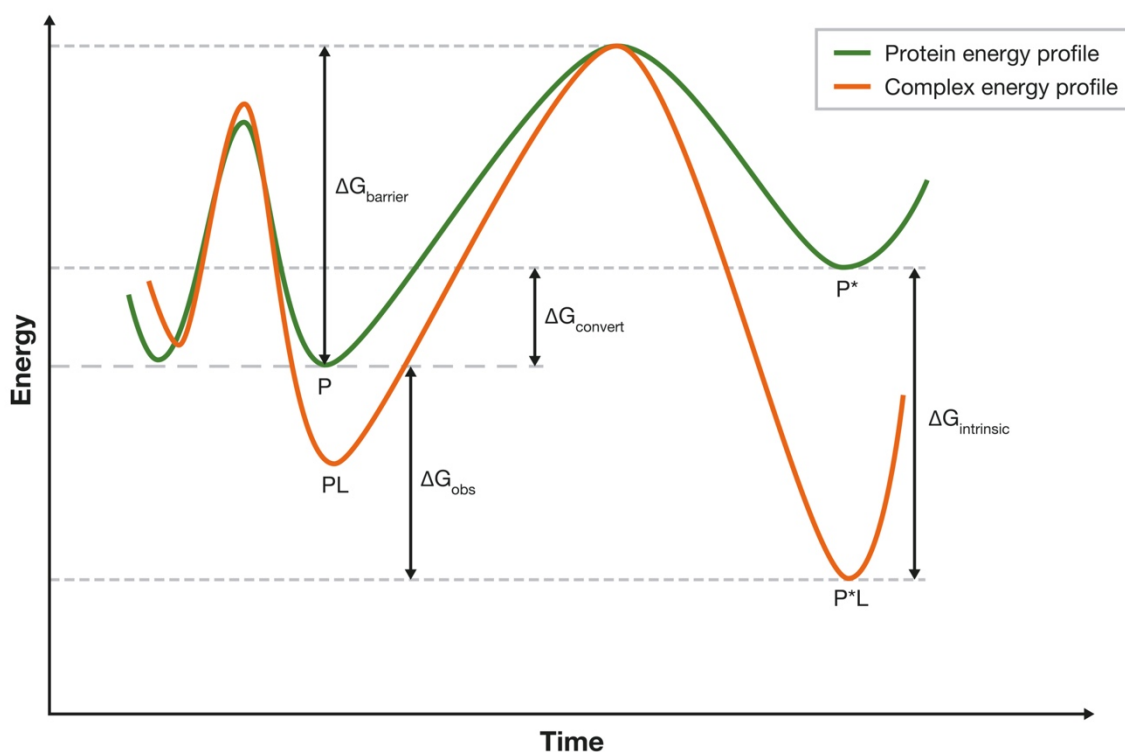
Figure 4. Demonstrative conformational landscape of a protein-ligand binding event for a protein that exists in two energetically distinct conformational states – P and P* (green energy profile). Although state P* is higher in energy than state P of the unbound protein, the protein ligand complex P*L is more stable than the complex PL (orange energy profile). Since the conformational state P* accommodates the ligand at the global energy minimum in the bound state, it is not unreasonable to expect that the protein would prefer to bind other ligands in a similar conformation. It would thus make the most sense to start docking calculations against this protein in state P* rather than in state P. This in turn would make the binding affinity predicted by the docking program more reliable. For this reason, a structure of a target protein co-crystallized with a bound ligand is often a good place to start when undertaking a docking exercise. These structures are readily available in databases like the PDB.

unbound state P* has a higher free energy, it may offer greater scope of interaction with a ligand L. As an example, P* may have hydrophobic areas that are solvent-exposed in the unbound state but offer the potential for favorable interactions with hydrophobic moieties in a suitable ligand L. Thus, interaction of the protein with L leads to conformational induction wherein a complex P*L is generated that has a lower energy than that of complex PL. In the real world, slow binding kinetics may be observed, given that P* is higher in energy than P, and an energy barrier $\Delta G_{barrier}$ must be surmounted before binding can take place.[56] In terms of docking, it makes sense to start a calculation for this protein with the structure in its conformational state P* rather than in the state P (as the induced-fit model would also suggest).

The next few sections document our efforts in docking CPMG ligands. All the docking studies described in this chapter treat the targets as rigid bodies. In addition, we used a combination of rigid and flexible approaches to model the macrocyclic ligands for estimating protein-ligand binding. We detail these efforts for four proteins – a clathrin adaptor protein and HMD2 (human double minute-2 protein) as test systems, and MTHFR (5,10-methylenetetrahydrofolate reductase) and GIPr (glucose-dependent insulinotropic polypeptide receptor).

2.4 Developing a rapid conformational tool to perform initial screening exercises

Traditional molecular docking implementations, such as AutoDock Vina[57] were designed to analyze freely rotating small molecule compounds. When presented with input macrocyclic ligands, these programs mark all macrocyclic bonds to be non-rotatable, and the backbones consequently remain unperturbed during conformational sampling in protein binding sites. However, it is now well established that in most cases, macrocyclic ring conformations play key roles in protein-ligand binding. One can circumvent this limitation in docking by seeding calculations with multiple, higher energy macrocyclic conformations to implicitly capture backbone flexibility. In recent years, methods have emerged to treat the macrocyclic backbone as fully

flexible during docking[58] (*vide infra*). Due to the relative simplicity of the implicit flexibility approach, we opted to initially establish a large virtual screening pipeline for CPMG using this method as an approximation. To enable large scale virtual screening with this approach, we would need to not only generate conformations for several thousands of ligands, but also generate several (higher energy) conformations per input structure. Adapting traditional conformational sampling algorithms to perform large-scale macrocyclic conformer generation is complicated by torsional complexities. We found that even sophisticated tools in the current Schrödinger Maestro suite could not perform this task within the constraints of the computational resources available to us. Given that our RDKit implementation of ConfBuster++ (CB++, Ch. 1) had greatly accelerated runtimes on the UCLA Hoffman2 sever compared to competing methods, we opted to adapt this program to suit our screening needs. We first looked to re-optimize program parameters to not only have optimal runtimes, but also generate reliable sets of low and higher energy conformations per input macrocycle.

The veracity of docking using an implicit flexibility approach to model macrocyclic conformations is largely dependent on the quality of the input ligand conformation. There is an important caveat to this approach. Ligand interactions with binding site residues play key roles in stabilizing bound macrocyclic conformations, such that these conformations may not otherwise be sampled in the ligand's unbound state. For freely rotating small molecules, this is often a surmountable issue, such that the bound conformation is accessed when the ligand undergoes conformational sampling in the environment of the binding site during docking. In the absence of backbone perturbations for macrocycles, this may not always hold true. In addition, solvent effects are not considered during the dihedral sampling stage in CB++, thereby obviating any consideration of dielectric effects during backbone optimization. Despite these limitations, our priority was to create a fully integrated screening pipeline from CPMG through docking, and the implicit flexibility approach was a convenient approximation. Subsequent efforts with explicit flexibility models are described later in this chapter.
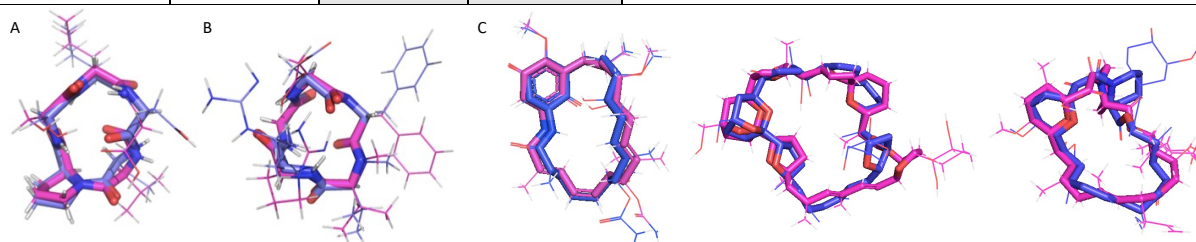
A list of the key parameters used for conformational analysis in CB++ and their default and optimized values are shown in Table 2. These 'default' parameters were identified during benchmarking studies detailed in Ch.1 for reliably reproducing results of other conformational search algorithms in the shortest run times. However, at the time, we had not attempted to benchmark those parameters to any crystallographically resolved ligand structures. Now, for our re-optimization studies, we chose to use the X-ray resolved structures of both the unbound form and bound form of the macrocyclic FK506 ligand to its target FKBP12 (FK506-binding protein-12) as a model system. Although FK506 is not a peptidic macrocycle, this system has been thoroughly studied and X-ray structures of the unbound and bound form were readily available in the CCDC (Cambridge Crystallographic Data Center) and PDB (Protein Data Bank) respectively. Our goal was to identify optimal parameters for generating a conformation population that contained at least one conformer that overlayed well with the bound and unbound form of the test-ligand in each case. Achieving reasonable overlays against both crystal forms would theoretically imply that we were adequately sampling both stable and higher energy conformations. The extent of alignment was quantified in each case by calculating root mean square deviations (RMSDs) between the backbone atoms of the respective overlaid structures. A key consideration while optimizing parameters was to ensure that runtimes were not significantly increased as a result of the changes. A complete list of attempted parameter optimizations is available in the Experimental Appendix (Table S1 and Fig. S1), but key observations are discussed here. We found that increasing -r or -m increased runtimes from less than 1 min to over 2 hours without greatly improving observed RMSD values. On the other hand, changing -ff, -eps, -e or -rmsd moderately affected runtimes but did not significantly lower RMSD values. The best results (Table 2, RMSDs in entries x and xi) were obtained when -n and -N parameters were increased as shown, and the scoring method in the genetic algorithm was changed from 'energy' to 'rmsd'. In general, scoring evaluates the fittest conformers to retain from a conformation population by applying the chosen force field. Scoring through 'energy' simply identifies the next lowest energy conformer, whereas

using RMSD in the scoring option retains the conformer$_{i+1}$ with the highest RMSD to conformer$_i$.

Typically, using the energy option generates the most stable, low energy conformations while the RMSD option generates a more diverse set of conformations. Using this option with an increased number of both (i) backbone conformers generated during rotamer search, and (ii) resulting macrocyclic conformers passed on to the genetic algorithm allowed us to obtain much improved RMSDs for both bound and unbound states of the test ligand. Improvements in RMSD for the unbound structure were greater than for the bound structure (Table 2, entries x and xi). As discussed earlier, this is not surprising when optimizing the ligand structure in absence of the protein.

Table 2. Re-optimized parameters for conformer generation in CB++ for integration with virtual screening. RMSD values obtained with default vs. optimized parameters are shown in rows (x) and (xi).

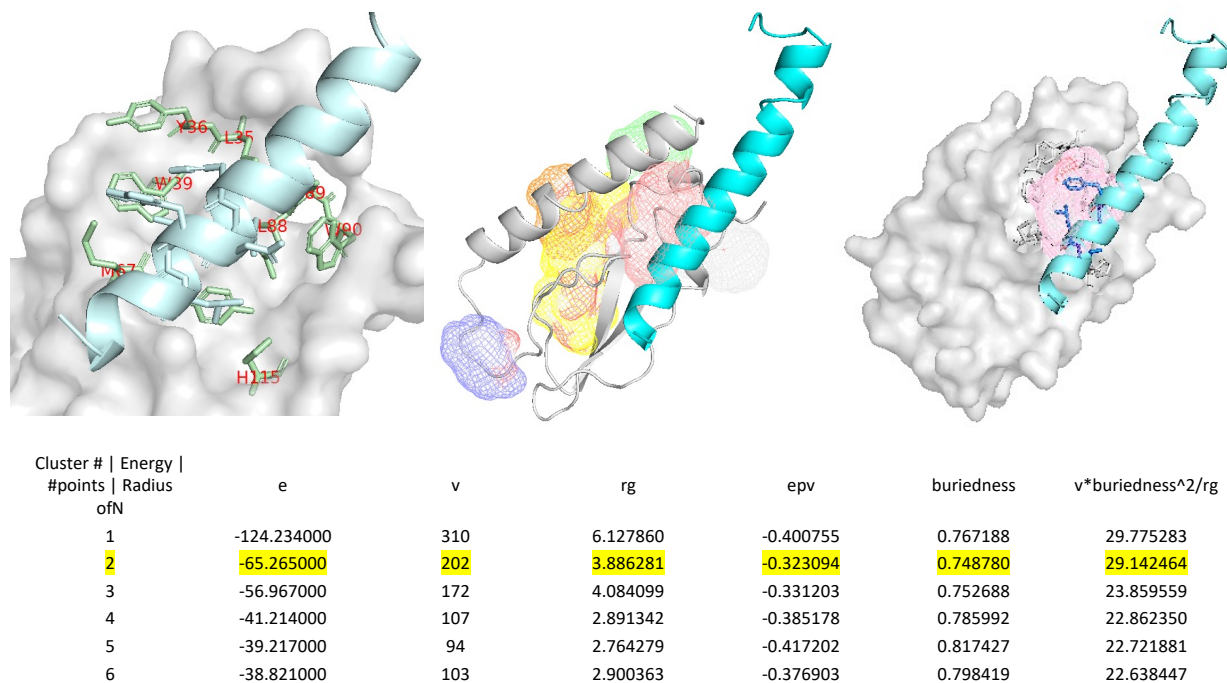| entry	argument	default	optimized	definition
(i)	-r	1	1	Number of times each linearized macrocycle is subjected to the sequence of random embedding, rotor search and genetic algorithm sampling
(ii)	-m	3	3	Number of conformers embedded per cut
(iii)	-n	15	30	Number of conformers generated by rotamer search per cut
(iv)	-N	5	15	Number of clash free conformers where the bond is reformed and passed on to the genetic algorithm
(v)	-ff	MMFF94s	MMFF94s	Force field used during minimization
(vi)	-eps	0	0	Dielectric constant applied during minimization
(vii)	-e	5	5	Maximum energy difference between the lowest and highest energy conformation
(viii)	-rmsd	0.5	0.5	Minimum RMSD threshold that two conformers must be apart to be retained in the final population
(ix)	-s	energy	rmsd	Score to use in the genetic algorithm
Lowest RMSD among generated conformations relative to X-ray structures of:				
(x)	Unbound structure	1.4 Å	0.9 Å	
(xi)	Bound structure	1.1 Å	0.8 Å	

Figure 5. Overlay of conformers from ConfBuster++ (blue) and CREST (pink; A and B) or crystal structures (pink; C) for **(A)** cyclo-(Pro-Ser-lue-Asp-Val); **(B)** cyclo-(Arg-Gly-Asp-phe-([N-Me]Val); **(C)** (left to right) geldanamycin, pectenotoxin, FK-506.

We also evaluated the ability of these re-optimized parameters in CB++ to reproduce the macrocyclic backbones of other X-ray resolved, or CREST generated structures (Fig. 5) and found RMSDs below 1 Å in each case. With these parameters in hand, we were able to generate sets of as many as 50 lowest energy conformations per macrocycle and were poised to perform initial macrocycle docking calculations against protein targets of interest. Following preliminary docking studies with CB++, we later refined the implicit flexibility strategy using new explicit flexibility models in collaboration with Prof. Stefano Forli at the Scripps Research Institute (section 2.6).
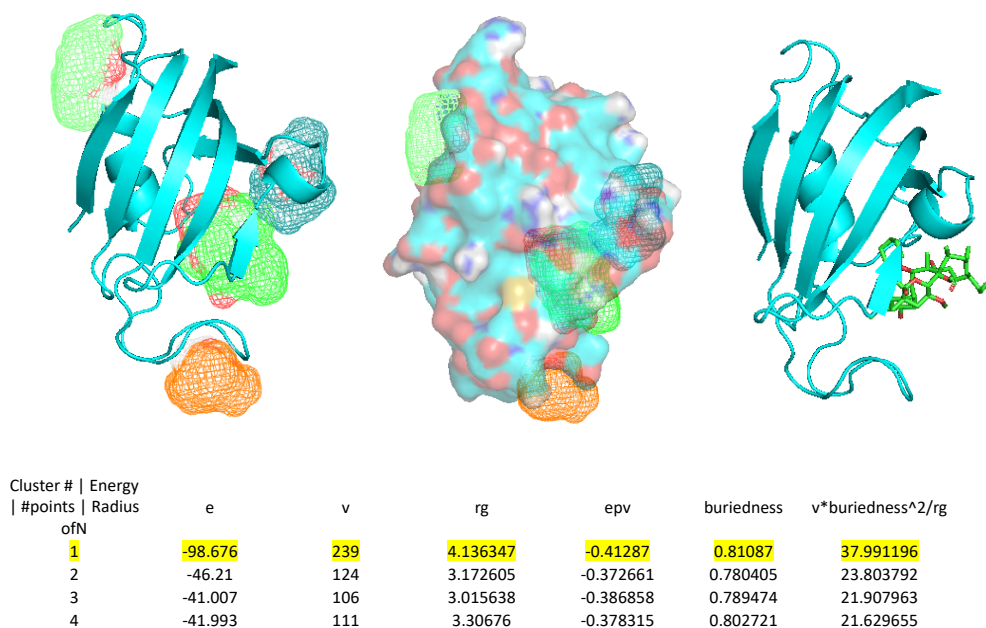
2.5 Protein structure analysis for identification of promising ligand binding sites

One of the targets we were interested in screening against was the protein-protein complex of the glucose-dependent insulinotropic polypeptide (GIP) and its receptor (GIPr) (Fig. 6).[59] Incretins released in response to food intake potentiate insulin secretion from pancreatic β-cells after oral glucose ingestion. This response is signalled, in part, by the GIP hormone through binding and activation of its cognate class II G-coupled protein receptor – GIPr. Since this so-called incretin effect is lost or significantly reduced in patients with type 2 diabetes, this system has attracted considerable attention for use in antidiabetic therapy.[59] The binding interaction between GIP and GIPr is greatly facilitated by the α -helical form of GIP, encompassing multiple hydrophobic interactions and hydrogen bonding interactions along the length of the polypeptide (key residues highlighted in leftmost panel of Fig. 6). The binding surface on GIPr is consequently shallow, devoid of any spatially distinct 'pockets', and much too large to incorporate in its entirety for molecular docking calculations. At the time of this work, there were also no known small molecule binders for this target. Our first step was thus to identify a small, discrete area along the surface to be used to define a binding site for docking calculations, bearing residues that might participate in favorable enthalpic interactions with a putative ligand.

Cluster # | Energy | #points | Radius ofN	e	v	rg	epv	buriedness	v*buriedness^2/rg
1	-124.234000	310	6.127860	-0.400755	0.767188	29.775283
2	-65.265000	202	3.886281	-0.323094	0.748780	29.142464
3	-56.967000	172	4.084099	-0.331203	0.752688	23.859559
4	-41.214000	107	2.891342	-0.385178	0.785992	22.862350
5	-39.217000	94	2.764279	-0.417202	0.817427	22.721881
6	-38.821000	103	2.900363	-0.376903	0.798419	22.638447

Figure 6. AutoSite results for GIPr (PDB: 2qkh) with GIPr in grey and GIP in blue (left to right) Key interacting residues highlighted in red; cartoon representation of GIPr with AutoSite generated clusters indicated in mesh; co-crystallized structure of the protein-ligand complex with the key cluster indicated in mesh in pink; (bottom table) AutoSite scoring of predicted high affinity clusters on the protein surface. Columns from left to right: cluster index; energy e; number of grid points v; radius of gyration rg; energy per grid point epv; buriedness = fill's buried surface area/fill's total surface area; empirical composite score = v*buriedness^2/rg; highlighted row indicates key cluster identified.

Many computational methods can seed 'blind docking' exercises by locating putative ligand binding sites on protein surfaces using geometric criteria to find clefts and surface depressions.[47] For instance, the program Surfnet identifies internal cavities and pockets of proteins by placing virtual spheres into solvent accessible spaces between protein atoms.[60] Other geometry-based methods include Cast, Ligsite, Pocket and PocketPicker.[61–64] Energy based methods, such as AutoSite,[65] have also emerged, relying on potentials generated by probe atoms or chemical moieties to define ligand binding sites. Interactions with probe atoms can be condensed and clustered to generate pharmacophoric points that provide more detailed information about the physiochemical properties in a binding site. We began our efforts using AutoSite on the same test case as in section 2.4 – FKBP12, the results for which are shown in Fig. 7. For FKBP12, AutoSite identifies four potential binding sites, which are rank ordered as a

Cluster # | Energy | #points | Radius ofN	e	v	rg	epv	buriedness	v*buriedness^2/rg
1	-98.676	239	4.136347	-0.41287	0.81087	37.991196
2	-46.21	124	3.172605	-0.372661	0.780405	23.803792
3	-41.007	106	3.015638	-0.386858	0.789474	21.907963
4	-41.993	111	3.30676	-0.378315	0.802721	21.629655

Figure 7. AutoSite results for FKBP12 (PDB: 2fke) with FKBP in blue and FK506 in stick representation in green (left to right) target protein with AutoSite generated clusters indicated in mesh; surface representation of the same, demonstrating buriedness of generated clusters; co-crystallized structure of the protein-ligand complex; (bottom table) AutoSite scoring of predicted high affinity clusters on the protein surface. Columns from left to right: cluster index; energy e; number of grid points v; radius of gyration rg; energy per grid point epv; buriedness = fill's buried surface area/fill's total surface area; empirical composite score = v*buriedness^2/rg; highlighted row indicates key cluster found at the ligand binding site.

function of the number of identified high affinity points and buriedness per cluster. As seen in Fig. 7, the site on FKB12 where the ligand is co-crystallized aligns very well with the highest ranked cluster. For a second model system to test the performance of AutoSite, we chose the surface of one of our proteins of interest – MTHFR, for which a co-crystallized structure bound to its native ligand, S-adenosylmethionine is available. This enzyme is involved in folate and methionine metabolism and its disregulation in disease states is of interest to our collaborator Jared Rutter (Utah / HHMI).[66] We were pleased to observe correct identification of the known allosteric site by AutoSite (data in Experimental Appendix, Fig. S2). Having established AutoSite performance against both test systems, we now sought to analyze the desired target GIPr with AutoSite, the results for which are shown in Fig. 6. We were pleased to find the second highest ranked cluster encompassed the residues believed to form key interactions for the binding of GIP to GIPr (indicated in sticks on the left and rightmost panel of Fig. 6). Using the coordinates of this discrete

three-dimensional cluster as a definition for a binding site, we could now move forward with docking calculations on this target.

2.6 Demonstrating the ability to find useful lead candidates for targets with known binders

To identify candidate ligands for proteins having few or no known small molecule binders, such as MTHFR and GIPr, we needed to first perform proof-of-principle experiments using proteins for which reliable structural data with known binders was available. This information would be used to drive the search, benchmark results, and also perform comparative assessments of two computational pipelines: **(i)** CPMG/ConfBuster++ pipeline using implicit macrocycle flexibility (section 2.4) in AutoDock Vina[57] on UCLA's Hoffman2 computing cluster, and **(ii)** CPMG/AutoDock-GPU[67] with explicitly flexible macrocycles[68] available on the Scripps Garibaldi cluster (in collaboration with Prof. Stefano Forli). The latter algorithm proceeds by identifying a macrocyclic bond to break, obtaining the linear form of the molecule to be docked, exploring conformations for that structure, and finally reconstituting the macrocyclic form. Importantly, these steps occur in the environment of the protein binding site, such that protein-ligand interactions are taken into account during conformational sampling. This strategy is a refinement on the first protocol in that it may grant access to stabilized conformations that the ligand may not otherwise adopt in the absence of the protein. Although fully flexible dockings are typically more computationally expensive, the GPU accelerated program enables these calculations in excellent runtimes (roughly 15,000 ligands in 6-8 hours).

Our first test target, HDM2, is an endogenous inhibitor of the p53 tumor suppressor protein.[69] Small molecule inhibition of the HDM2/p53 interface has been pursued for cancer therapy.[69,70] For our experiment, we used the complex of HDM2 with a cyclic β-hairpin inhibitor that mimics a p53 epitope (PDB:2axi, Fig. 8A). Our second test target, clathrin, plays a central role as a mechanical scaffold in biological processes involving budding vesicles.[71,72] For this

system, we focused on the protein-protein interface between clathrin and a peptide derived from amphiphysin 1, an accessory protein responsible for the regulation of clathrin aggregation.[73] (PDB:1utc, Fig. 8B). Our third and fourth targets were MTHFR (PDB:6fcx, Fig. 8C) and GIPr (PDB:2qkh). This target set presented multiple challenges, from large interaction surface areas (HDM2, clathrin, GIPr) to deep allosteric pockets (MTHFR). For each target we followed the same basic protocol, composed of steps representing the core computational pipeline described earlier in Fig.3. First, for each target, the protein surface was analyzed by means of the AutoSite software to characterize the binding pocket (shape, volume) and its features (HB acceptor/donor, hydrophobic). The results of the pocket analysis were used to drive the docking calculations, which were performed with both (i) pre-generated ligand conformers (using CPMG/ConfBuster++ and AutoDock Vina) and (ii) fully flexible ligands (using AutoDockGPU).
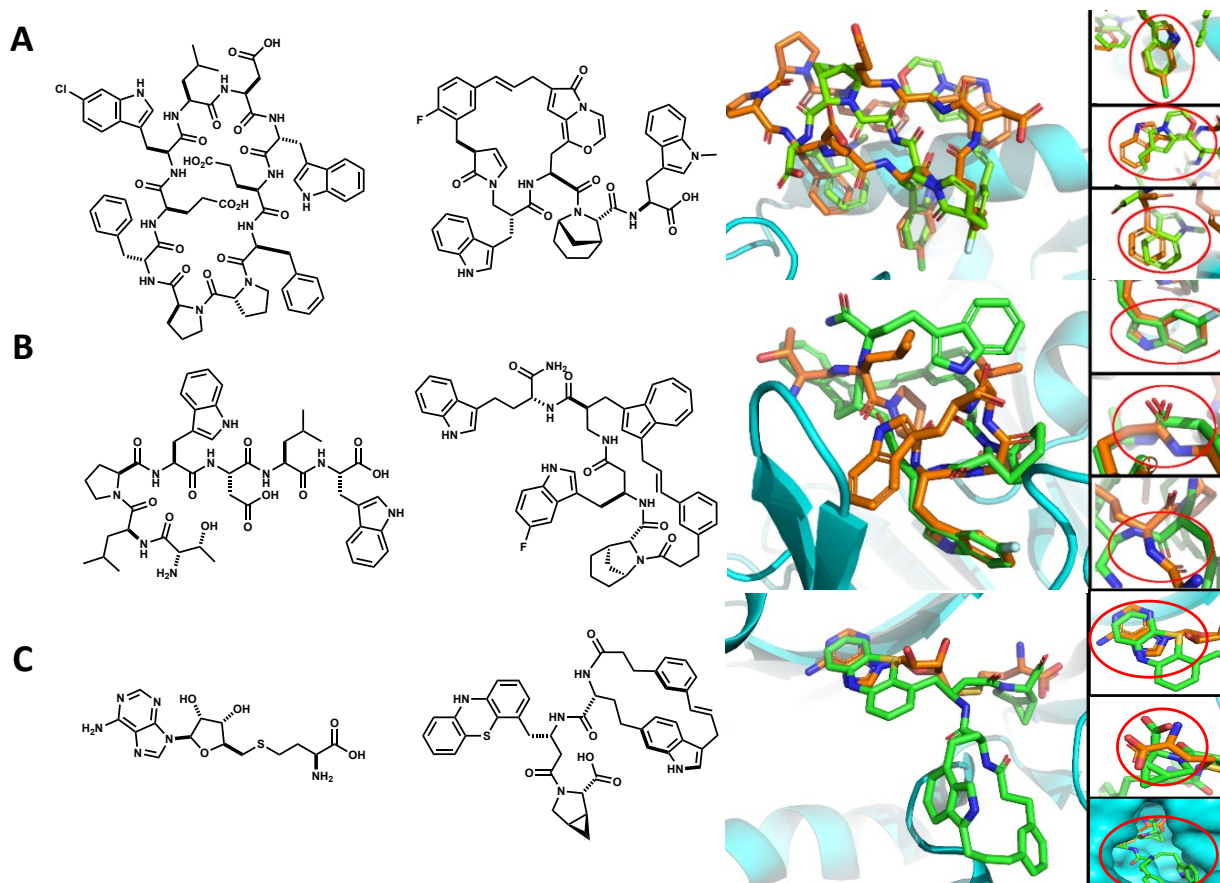
Figure 8. (A) HDM2 (PDB:2axi); (B) clathrin (PDB:1utc); (C) MTHFR (PDB:6fcx). *(right to left)* Known binders of the respective targets. SMARTS patterns based on these structures were used to generate pre-filtered ligand sets from CPMG in each case; top scoring CPMG ligands identified from the pre-filtered sets; overlays of the known binders with the top scoring ligand in each case.

59

One approach to large scale virtual screening is to first enrich ligand libraries on the basis of native or known binders. Accordingly, CPMG subsets were generated for each protein on the basis of the structural and chemical nature of their respective known binders. HDM2, clathrin, and MTHFR were chosen to be our positive control experiments for assessing the impact of ligand enrichment on docking results. For this, we implemented substructure matching in the CPMG library based on SMILES or SMARTS patterns, physiochemical property criteria and three-dimensional shape criteria of known binders (complete filtering criteria discussed in the Experimental Appendix, Tables S3, S5-6 and Fig. S3). On this basis, a set of macrocycles (10,000) harboring at least one endocyclic indole were selected for both HDM2 and clathrin. For MTHFR, a set of 15,000 macrocyles containing the 2-amino-N-phenethylacetamide motif were selected. Our fourth target, GIPr, was chosen to be our negative control experiment. In this case, instead of tailoring a subset based on the known binder (GIP), we compiled a set of randomly selected CPMG macrocycles. To incorporate maximum structural diversity in our random selection process, we ensured that all macrocyclic backbone outcomes of our chemistry (Fig. 9A), were included in equal numbers within the set. For this, we implemented substructure searches using the eight patterns shown in Fig. 9B. Since we also wanted to demonstrate the feasiblity of performing a larger scale screen on our server, we compiled a set of 100,000 ligands for this last
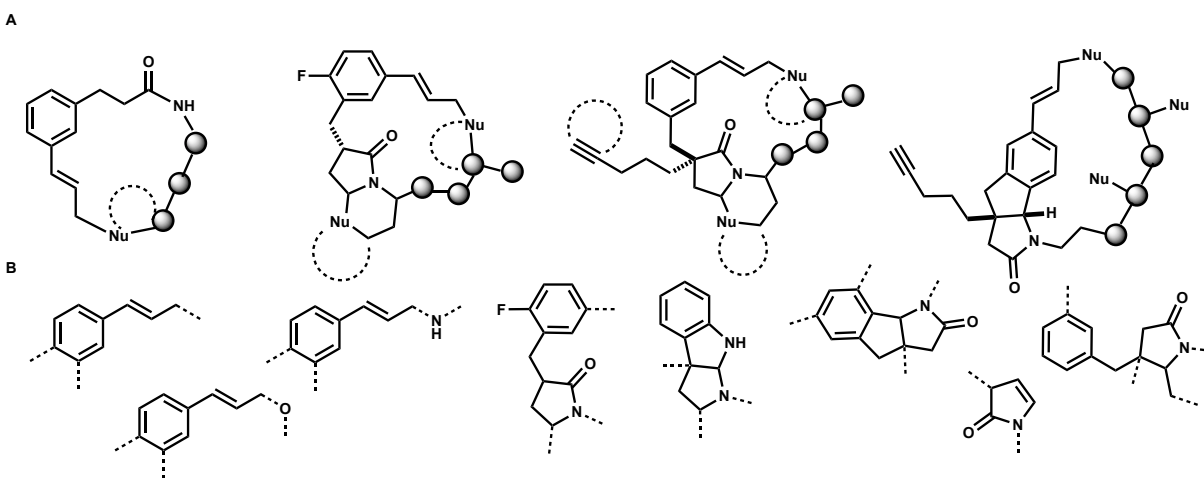
Figure 9. (A) General structural prototypes contained in CPMG; **(B)** scaffolding motifs queried from the CPMG to generate initial macrocycle set for GIPr docking.

60

experiment – resulting in approx. 1 million unique conformations (~10 conformers per macrocycle). Molecular docking was then performed on each set using AutoDock Vina and AutoDock-GPU within the two pipelines described above (configuration files for docking detailed in Experimental Appendix, Table S4).

Our preliminary results highlight important driving forces in the identification of potential target binders from the gigalibrary. First, the presence of opportune chemical features in the ligands is crucial for mapping onto known interactions and important target surface characteristics. Overlays of ligands with high docking scores against the known binders for our three positive control targets are shown in Fig. 8. As shown, pre-filtering of our gigalibrary on the basis of known binders allowed us to identify ligands capable of closely matching the structural features deemed essential for the binding of known inhibitors/substrates. For HDM2, clathrin, and MTHFR targets, dockings identified top scoring poses that were able to recapitulate important structural features of the known ligands found in the PDB complexes, including the key recognition patterns essential for binding. This is a very encouraging result considering the low chemical similarity of the ligands, but most importantly, the very small fraction of the gigalibrary considered in these pilot dockings ($5x10^{-5}$%). Moreover, without using constraints or biases, the dockings were able to match indole-based features of the known ligands at protein-protein interfaces (i.e.: HDM2, clathrin; Fig. 8A-B) with indole-based features in the macrocycles, as well as hydrogen bonding patterns. Interestingly, in both cases, macrocycles bearing two exocyclic indoles were scored highest. For the MTHFR binding site (Fig. 8C), the adenine and the acidic portion of S-adenosyl-L-homocysteine bound in the enzyme active site were matched, respectively, by phenothiazine and carboxylate side chains on the macrocycle, while its backbone engaged a hydrophobic pocket implicated in increased activity of known inhibitors.[74]

In contrast to the positive controls described above, no satisfactory matches were found between the best docked macrocycles and the GIP residues responsible for binding interactions with GIPr. This outcome was expected, and confirms the importance of the library pre-screen. However, increasing the scale of virtual screening should have the same effect as manual library enrichment. In the former case, ligand enrichment should occur naturally by sheer dint of numbers. This was confirmed in the case of GIPr (Fig. 10) by comparing the top 100 docking scores from screens of ~10,000 randomly selected CPMG molecules (in orange) and ~800,000 randomly selected molecules (in blue). As seen in in Fig. 10, the larger set not only had higher scores, but also a larger standard deviation in the data set. This is an important consideration for targets such as GIPr, where little is known about the active pharmacophore, and even a library of 100,000 ligands is not large enough to find hits.

We also found macrocyclic conformation to be essential for placing binding features in their optimal position to interact effectively with receptor functionality. Comparing flexible macrocycle docking results to implicit flexible docking showed that the sampling of ligand conformations in the presence of the protein faciliated subtle, but consistent structural perturbations that resulted in higher docking scores. The flexible docking
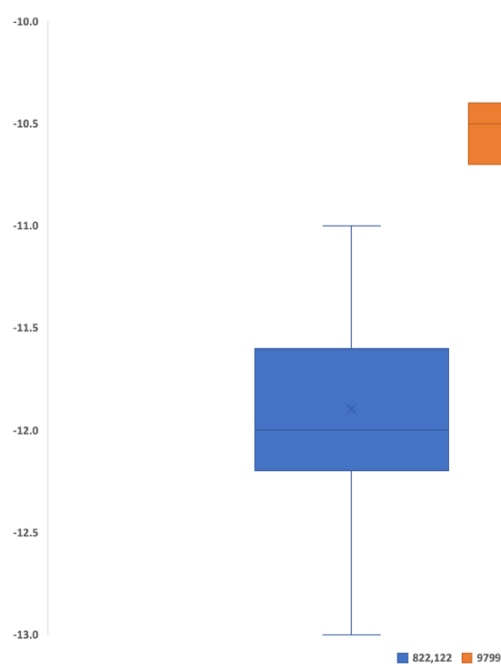
Figure 10. Box and whisper plots demonstrating library enrichment in the top 100 scores for a ligand set of 822,122 structures (blue) relative to a set of 9799 structures (orange). Y-axis represents predicted binding affinities in kcal/mol. Lines in boxes represent first quartile, median and third quartile from top to bottom respectively; x indicates the mean in each case; whiskers indicate minimum and maximum values in each dataset; dots indicated outliers.

protocol routinely involved conformers that were not sampled in implicit flexibility docking.

2.7 Implementation of property filters for anticipated membrane permeability to prioritize hits from virtual screening

In 2012, it was reported that the number of new drugs approved per billion US dollars spent in research and development (R&D) had halved roughly every 9 years since 1950 .[75] This trend is somewhat surprising given the technological advances that were made in R&D during that period, including the advent of combinatorial chemistry, DNA sequencing, HTS platforms and the entire fields of biotechnology and computational drug design. While the observed decline has been linked to several underlying issues, it has been partly attributed to the fact that chemical libraries used in screening exercises incorporate a limited number of chemotypes – and particularly those that are deemed to have acceptable ADME (absorption, distribution, metabolism and excretion) characteristics. This is an important point, exemplifying how a ligand must not only exhibit high affinity but must also be usefully bioavailable in order to be a promising drug lead.

In the early 2000s, having surveyed the structures and physiochemical properties of over 20,000 drugs, Lipinski *et al.* proposed the Ro5 (Rule of 5) for oral bioavailability, stating that *"Poor absorption or permeation is more likely when there are more than 5 H-bond donors* (HBD)*, 10 H-bond acceptors* (HBA)*, the molecular weight is greater than 500, and the calculated logP is greater than 5."*[76] Based on this, small molecule ligands fall into two categories – Ro5 compliant or bRo5 (beyond the Rule of 5). In order to increase the likelihood of obtaining favorable physiochemical and pharmacokinetic properties, medicinal chemists seek to identify small molecule leads that are compliant with the Ro5. However, stepping outside the Ro5 chemical space can enhance compound diversity and present opportunities for developing drugs against "difficult" targets such as proteases, G-protein coupled receptors (GPCRs) and protein-protein interactions (PPIs).[30,77] Most bRo5 compounds reside between two molecular weight extremes – larger than traditional

drug molecules but smaller than biologics, and capable of accessing the pharmacological properties of both. As interest in bRo5 compounds has increased, new and more flexible guidelines for ligand design have emerged. One study noted that while ADME may be a hurdle with increased molecular size and complexity, the limits for oral bioavailability can be extended to approximately MW ≤ 1000 Da, −2 ≤ clogP ≤ 10, HBD ≤ 6, HBA ≤ 15, PSA ≤ 250 $Å^2$ and nrot ≤ 20.[77] The same study noted that at the time of publication in 2014, 182 bRo5 compounds were approved drugs, and 303 were being evaluated in clinical trials. Most of these could be further classified as belonging to a few specific chemical classes – erythronolides, leucomycins, rifamycins, asomycins, rapamycins, cyclosporins, azoles and taxanes. An analysis revealed a significant fraction of the remaining compounds to be peptides and peptidomimetic and macrocyclic variants of the same.

In light of this, we sought to identify CPMG macrocycles that not only matched active pharmacophores, but also had promising physiochemical properties for achieving passive permeation. Previous empirical analyses indicate macrocycles having MW < 1000 Da, TPSA < 250 $Å^2$, clogP < 10, and fewer than 5 hydrogen bond donors are more likely to be bioavailable.[42,77,78] Importantly, it has been noted that all passively permeable FDA approved macrocycles having MW > 500 g/mol have a ratio of TPSA:MW ≥ 0.2.[79]

We first confirmed that implementation of these property filters would not significantly restrict ligand space, and that at least several thousands of structures would be taken forward to docking (data in Experimental Appendix, Table S5). Thereafter, we carefully constructed a second subset of molecules from the CPMG gigalibrary to be screened against MTHFR, wherein the selection criteria were based not only on matching the pharmacophore to the known binder, but also on property filters anticipating the ability to passively permeate cell membranes. While the MTHFR ligand library detailed in section 2.6 was derived from the amino-*N*-phenethylacetamide motif, this time, the implemented pharmacophore identified structures having motifs that could mimic the adenosyl ring of the known binder (Fig. 11, full details in Experimental Appendix, Table

S3), resulting in an initial set of over 3 million structures. To maximize the likelihood of top binders being passively permeable, all structures in this initial set were additionally filtered to have the following properties – MW < 800 g/mol, TPSA <150 Å, 0.2 < TPSA/MW <0.3, hydrogen bond donors (HBD) ≤ 5. Our final set comprised 53,577 unique molecules (full details in Experimental Appendix, Table S6).

We virtually screened this set of compounds using the flexible macrocycle protocol described above, and then filtered the results looking for ligands capable of establishing interactions with key residues within the receptor binding site. To this extent, docking results were rapidly filtered by properties (energy, ligand efficiency) as well as residue interactions (HB, vdW, etc.), based on the metadata collected by the AutoDock-GPU docking engine during the simulation. This tool was used to identify 200 promising macrocycles (full list in Experimental Appendix, Table S7), having diverse and heterogeneous chemical patterns, to mimic the interaction patterns of known ligands to engage the binding site. One of the best hits is shown in Fig. 11, wherein the three key pharmacophoric features of the S-adenosyl-L-homocysteine (SAH) complexed in the MTHFR structure is matched. This ligand is also capable of further engaging the site beyond the SAH site, occupying the hydrophobic accessory pocket toward P371, which is likely to further increase affinity and selectivity. Efforts are currently underway to experimentally synthesize the top scoring hits from this screen in order to test the structures for both passive
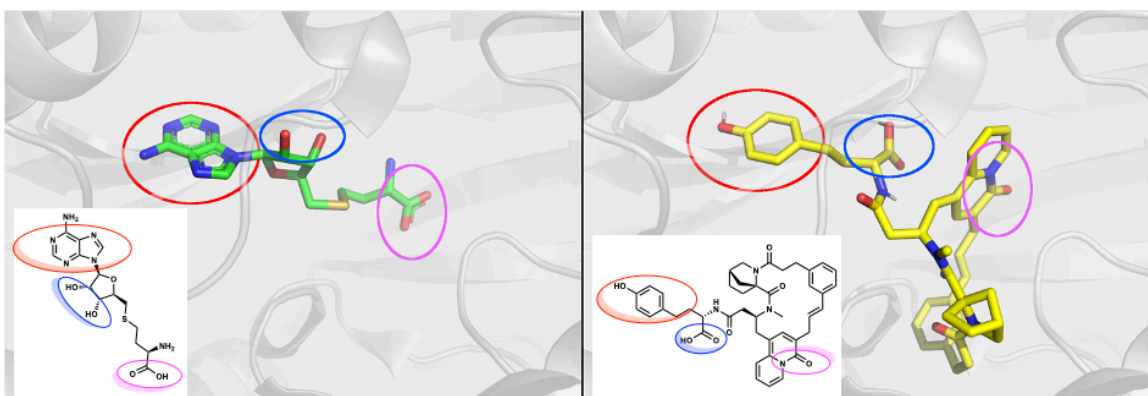
Figure 11. (left to right) Known SAH binder docked in the allosteric site of MTHFR; CPMG macrocycle docked to the same site wherein key interactions are recapitulated by the new binder, as indicated by the circled and color coded functional groups.

permeation and bioactivity. The results from that study will be used to further refine the filtering protocols described in this chapter.

2.8 Conclusions and outlook

In this chapter, we have documented several proof-of-concept experiments that will enable future large-scale virtual screening exercises using CPMG. We have successfully executed pipelines for virtual screening using both an implicit ligand flexibility model and a fully flexible ligand model, wherein each employs a sequence of CPMG → conformational analysis → protein structure analysis → molecular docking → hit prioritization. Importantly, both options are amenable to running screens in the order of several hundreds of thousands and several millions in the case of AutoDock GPU. Through these exercises, we have demonstrated (i) the ability to perform large-scale screens of ligands from CPMG, (ii) the flexible body approach to be far superior to the implicit flexibility model, and (iii) the importance of ligand enrichment in hit identifcation. Despite these advances, there is still much to explore. All of the studies detailed in this chapter employed a rigid body approach to protein flexibility. This may have a significant impact on the number of false negatives that are generated. Moreover, explicit water molecules were excluded from any consideration during these studies. While this should not have had a great impact for the specific targtes studied here, it may become an important consideration when evaluating increasingly diverse proteins. In general, scoring and reliably ranking macrocyclic ligands continues to be a highly underexplored area, and scores scale poorly in relation to molecular mass and rotatable bonds. Since larger molecules can form more hypothetical interactions in binding sites, they tend to generate better scores. On the other hand, the entropic penalty for the immobilization of rotatable bonds means flexible molecules should score lower than more rigid ones, but entropic penalties are often not adequately included. The internal strain of a ligand pose is generally approximated using a single unbound conformation as the reference, which limits the accuracy of enthalpic and entropic losses on binding. The preparation and nature

66

of active sites also affects prediction quality. Binding to active sites that have large hydrophobic areas can be better approximated because methods based on shape complementarity are well developed and implicitly take hydrophobic effects into account. In contrast, hydrophilic sites or binding events that involve metallo-enzymes are less promising targets for docking exercises. Electrostatic interactions are scored by simple Coulombic terms, which have a tendency to grossly overestimate interactions. On the flip side, a large contributor to the hydrophobic effect is desolvation, which is often underscored relative to other scoring terms. Docking programs also occasionally identify promiscuous binders that ultimately have low utility as drug leads.[80] Despite these shortcomings, the methods discussed in this chapter already extend more traditional approaches to drug design, and provide a basis for future large-scale screens of complex molecules.

2.9 References

(1)   Mullard, A. The Phenotypic Screening Pendulum Swings. *Nat. Rev. Drug Discov.* **2015**, *14* (12), 807–809.

(2)   Eder, J.; Sedrani, R.; Wiesmann, C. The Discovery of First-in-Class Drugs: Origins and Evolution. *Nat. Rev. Drug Discov.* **2014**, *13* (8), 577–587.

(3)   Lundblad, R. L. Drug Design; Bradshaw, R. A., Stahl, P. D. B. T.-E. of C. B., Eds.; Academic Press: Waltham, 2016; pp 135–140.

(4)   Brown, D. G.; Boström, J. Where Do Recent Small Molecule Clinical Development Candidates Come From? *J. Med. Chem.* **2018**, *61* (21), 9442–9468.

(5)   Follmann, M.; Briem, H.; Steinmeyer, A.; Hillisch, A.; Schmitt, M. H.; Haning, H.; Meier, H. An Approach towards Enhancement of a Screening Library: The Next Generation Library Initiative (NGLI) at Bayer — against All Odds? *Drug Discov. Today* **2019**, *24* (3), 668–672.

(6)   Mayr, L. M.; Bojanic, D. Novel Trends in High-Throughput Screening. *Curr. Opin. Pharmacol.* **2009**, *9* (5), 580–588.

(7)   Galloway, W. R. J. D.; Isidro-Llobet, A.; Spring, D. R. Diversity-Oriented Synthesis as a Tool for the

Discovery of Novel Biologically Active Small Molecules. *Nat. Commun.* **2010**, *1* (1), 80.

(8)   Gerry, C. J.; Schreiber, S. L. Chemical Probes and Drug Leads from Advances in Synthetic Planning and Methodology. *Nat. Rev. Drug Discov.* **2018**, *17* (5), 333–352.

(9)   Gorse, A.-D. Diversity in Medicinal Chemistry Space. *Current Topics in Medicinal Chemistry*. 2006, pp 3–18.

(10)   DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *J. Health Econ.* **2016**, *47*, 20–33.

(11)   Brenner, S.; Lerner, R. A. Encoded Combinatorial Chemistry. *Proc. Natl. Acad. Sci.* **1992**, *89* (12), 5381–5383.

(12)   Goodnow, R. A.; Dumelin, C. E.; Keefe, A. D. DNA-Encoded Chemistry: Enabling the Deeper Sampling of Chemical Space. *Nat. Rev. Drug Discov.* **2017**, *16* (2), 131–147.

(13)   Gartner, Z. J.; Tse, B. N.; Grubina, R.; Doyon, J. B.; Snyder, T. M.; Liu, D. R. DNA-Templated Organic Synthesis and Selection of a Library of Macrocycles. *Science* **2004**, *305* (5690), 1601–1605.

(14)   Monty, O. B. C.; Nyshadham, P.; Bohren, K. M.; Palaniappan, M.; Matzuk, M. M.; Young, D. W.; Simmons, N. Homogeneous and Functional Group Tolerant Ring-Closing Metathesis for DNA-Encoded Chemical Libraries. *ACS Comb. Sci.* **2020**, *22* (2), 80–88.

(15)   Lu, X.; Roberts, S. E.; Franklin, G. J.; Davie, C. P. On-DNA Pd and Cu Promoted C-N Cross-Coupling Reactions. *Medchemcomm* **2017**, *8* (8), 1614–1617.

(16)   Hu, Q.; Peng, Z.; Sutton, S. C.; Na, J.; Kostrowicki, J.; Yang, B.; Thacher, T.; Kong, X.; Mattaparti, S.; Zhou, J. Z.; et al. Pfizer Global Virtual Library (PGVL): A Chemistry Design Tool Powered by Experimentally Validated Parallel Synthesis Information. *ACS Comb. Sci.* **2012**, *14* (11), 579–589.

(17)   Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discov. Today* **2019**, *24* (5), 1148–1156.

(18)   Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J. The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *J. Chem. Inf. Model.* **2016**, *56* (7), 1253–1266.

(19)   Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. Searching Fragment Spaces with Feature Trees. *J. Chem. Inf. Model.* **2009**, *49* (2), 270–279.

(20)   MELLODDY https://www.melloddy.eu.

(21)   Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594.

(22)   Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337.

(23)   Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; et al. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229.

(24)   Chevillard, F.; Kolb, P. SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. *J. Chem. Inf. Model.* **2015**, *55* (9), 1824–1835.

(25)   Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131* (25), 8732–8733.

(26)   Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864–2875.

(27)   Huang, B.; von Lilienfeld, O. A. Quantum Machine Learning Using Atom-in-Molecule-Based Fragments Selected on the Fly. *Nat. Chem.* **2020**.

(28)   Giordanetto, F.; Kihlberg, J. Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? *J. Med. Chem.* **2014**, *57* (2), 278–295.

(29)   Arkin, M. R.; Wells, J. A. Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing towards the Dream. *Nat. Rev. Drug Discov.* **2004**, *3* (4), 301–317.

(30)   Egbert, M.; Whitty, A.; Keserű, G. M.; Vajda, S. Why Some Targets Benefit from beyond Rule of Five Drugs. *J. Med. Chem.* **2019**, *62* (22), 10005–10025.

(31)   Saha, I.; Dang, E. K.; Svatunek, D.; Houk, K. N.; Harran, P. G. Computational Generation of an Annotated Gigalibrary of Synthesizable, Composite Peptidic Macrocycles. *Proc. Natl. Acad. Sci.* **2020**, *117* (40), 24679–24690.

(32)   Humbeck, L.; Weigang, S.; Schäfer, T.; Mutzel, P.; Koch, O. CHIPMUNK: A Virtual Synthesizable Small-Molecule Library for Medicinal Chemistry, Exploitable for Protein–Protein Interaction Modulators. *ChemMedChem* **2018**, *13* (6), 532–539.

(33)   Duffy, F. J.; Verniere, M.; Devocelle, M.; Bernard, E.; Shields, D. C.; Chubb, A. J. CycloPs: Generating Virtual Libraries of Cyclized and Constrained Peptides Including Nonnatural Amino Acids. *J. Chem. Inf. Model.* **2011**, *51* (4), 829–836.

(34)   Zin, P. P. K.; Williams, G.; Fourches, D. Cheminformatics-Based Enumeration and Analysis of Large Libraries of Macrolide Scaffolds. *J. Cheminform.* **2018**, *10* (1), 53.

(35)   Rose, T. E.; Lawson, K. V; Harran, P. G. Large Ring-Forming Alkylations Provide Facile Access to Composite Macrocycles. *Chem. Sci.* **2015**, *6* (4), 2219–2223.

(36)   Lawson, K. V; Rose, T. E.; Harran, P. G. Template-Constrained Macrocyclic Peptides Prepared from Native, Unprotected Precursors. *Proc. Natl. Acad. Sci.* **2013**, *110* (40), E3753--E3760.

(37)   Curtin, B. H.; Manoni, F.; Park, J.; Sisto, L. J.; Lam, Y.-H.; Gravel, M.; Roulston, A.; Harran, P. G. Assembly of Complex Macrocycles by Incrementally Amalgamating Unprotected Peptides with a Designed Four-Armed Insert. *J. Org. Chem.* **2018**, *83* (6), 3090–3108.

(38)   Rose, T. E.; Curtin, B. H.; Lawson, K. V; Simon, A.; Houk, K. N.; Harran, P. G. On the Prevalence of Bridged Macrocyclic Pyrroloindolines Formed in Regiodivergent Alkylations of Tryptophan. *Chem. Sci.* **2016**, *7* (7), 4158–4166.

(39)   Lawson, K. V; Rose, T. E.; Harran, P. G. Template-Induced Macrocycle Diversity through Large Ring-Forming Alkylations of Tryptophan. *Tetrahedron* **2013**, *69* (36), 7683–7691.

(40)   Zhao, H.; Negash, L.; Wei, Q.; LaCour, T. G.; Estill, S. J.; Capota, E.; Pieper, A. A.; Harran, P. G. Acid Promoted Cinnamyl Ion Mobility within Peptide Derived Macrocycles. *J. Am. Chem. Soc.* **2008**, *130* (42), 13864–13866.

(41)   Over, B.; Matsson, P.; Tyrchan, C.; Artursson, P.; Doak, B. C.; Foley, M. A.; Hilgendorf, C.; Johnston, S. E.; Lee IV, M. D.; Lewis, R. J.; et al. Structural and Conformational Determinants of Macrocycle Cell Permeability. *Nat. Chem. Biol.* **2016**, *12*, 1065.

(42)   Matsson, P.; Kihlberg, J. How Big Is Too Big for Cell Permeability? *J. Med. Chem.* **2017**, *60* (5), 1662–1664.

(43) Witek, J.; Mühlbauer, M.; Keller, B. G.; Blatter, M.; Meissner, A.; Wagner, T.; Riniker, S. Interconversion Rates between Conformational States as Rationale for the Membrane Permeability of Cyclosporines. *ChemPhysChem* **2017**, *18* (23), 3309–3314.

(44) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the Development of Universal, Fast and Highly Accurate Docking/Scoring Methods: A Long Way to Go. *Br. J. Pharmacol.* **2008**, *153 Suppl* (Suppl 1), S7–S26.

(45) Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Aided. Drug Des.* **2011**, *7* (2), 146–157.

(46) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions. *Proteins* **2002**, *47* (4), 409–443.

(47) Simon, Z.; Vigh-Smeller, M.; Peragovics, A.; Csukly, G.; Zahoránszky-Kohalmi, G.; Rauscher, A. A.; Jelinek, B.; Hári, P.; Bitter, I.; Málnási-Csizmadia, A.; et al. Relating the Shape of Protein Binding Sites to Binding Affinity Profiles: Is There an Association? *BMC Struct. Biol.* **2010**, *10*, 32.

(48) Pujadas, G.; Vaque, M.; Ardevol, A.; Blade, C.; Salvado, M. J.; Blay, M.; Arola, J. F.-L. and L. Protein-Ligand Docking: A Review of Recent Advances and Future Perspectives. *Current Pharmaceutical Analysis*. 2008, pp 1–19.

(49) Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; et al. An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature* **2020**, *580* (7805), 663–668.

(50) Cherfils, J.; Janin, J. Protein Docking Algorithms: Simulating Molecular Recognition. *Curr. Opin. Struct. Biol.* **1993**, *3* (2), 265–269.

(51) Fischer, E. Einfluss Der Configuration Auf Die Wirkung Der Enzyme. *Berichte der Dtsch. Chem. Gesellschaft* **1894**, *27* (3), 2985–2993.

(52) Shoichet, B. K.; Kuntz, I. D. Protein Docking and Complementarity. *J. Mol. Biol.* **1991**, *221* (1), 327–346.

(53) KOSHLAND, D. E. J. CORRELATION OF STRUCTURE AND FUNCTION IN ENZYME ACTION. *Science* **1963**, *142* (3599), 1533–1541.

(54) Sulimov, V. B.; Kutov, D. C.; Sulimov, A. V. Advances in Docking. *Curr. Med. Chem.* **2019**, *26* (42),

7555–7580.

(55)   Hammes, G. G.; Chang, Y.-C.; Oas, T. G. Conformational Selection or Induced Fit: A Flux Description of Reaction Mechanism. *Proc. Natl. Acad. Sci.* **2009**, *106* (33), 13737–13741.

(56)   Teague, S. J. Implications of Protein Flexibility for Drug Discovery. *Nat. Rev. Drug Discov.* **2003**, *2* (7), 527–541.

(57)   Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem. 31* (2), 455–461.

(58)   Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **2021**, *61* (8), 3891–3898.

(59)   Parthier, C.; Kleinschmidt, M.; Neumann, P.; Rudolph, R.; Manhart, S.; Schlenzig, D.; Fanghänel, J.; Rahfeld, J.-U.; Demuth, H.-U.; Stubbs, M. T. Crystal Structure of the Incretin-Bound Extracellular Domain of a G Protein-Coupled Receptor. *Proc. Natl. Acad. Sci.* **2007**, *104* (35), 13942–13947.

(60)   Laskowski, R. A. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *J. Mol. Graph.* **1995**, *13* (5), 323–330.

(61)   Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Sci.* **1998**, *7* (9), 1884–1897.

(62)   Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins. *J. Mol. Graph. Model.* **1997**, *15* (6), 359–363.

(63)   Levitt, D. G.; Banaszak, L. J. POCKET: A Computer Graphics Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino Acids. *J. Mol. Graph.* **1992**, *10* (4), 229–234.

(64)   Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: Analysis of Ligand Binding-Sites with Shape Descriptors. *Chem. Cent. J.* **2007**, *1*, 7.

(65)   Ravindranath, P. A.; Sanner, M. F. AutoSite: An Automated Approach for Pseudo-Ligands Prediction-from Ligand-Binding Sites Identification to Predicting Key Ligand Atoms. *Bioinformatics* **2016**, *32* (20), 3142—3149.

(66)   Froese, D. S.; Huemer, M.; Suormala, T.; Burda, P.; Coelho, D.; Guéant, J.-L.; Landolt, M. A.; Kožich, V.; Fowler, B.; Baumgartner, M. R. Mutation Update and Review of Severe

Methylenetetrahydrofolate Reductase Deficiency. *Hum. Mutat.* **2016**, *37* (5), 427–438.

(67)   Santos-Martins, D.; Solis-Vasquez, L.; Tillack, A. F.; Sanner, M. F.; Koch, A.; Forli, S. Accelerating AutoDock4 with GPUs and Gradient-Based Local Search. *J. Chem. Theory Comput.* **2021**, *17* (2), 1060–1073.

(68)   Forli, S.; Botta, M. Lennard-Jones Potential and Dummy Atom Settings to Overcome the AUTODOCK Limitation in Treating Flexible Ring Systems. *J. Chem. Inf. Model.* **2007**, *47* (4), 1481—1492.

(69)   Fasan, R.; Dias, R. L. A.; Moehle, K.; Zerbe, O.; Vrijbloed, J. W.; Obrecht, D.; Robinson, J. A. Using a β-Hairpin To Mimic an α-Helix: Cyclic Peptidomimetic Inhibitors of the P53–HDM2 Protein–Protein Interaction. *Angew. Chemie Int. Ed.* **2004**, *43* (16), 2109–2112.

(70)   Chène, P. Inhibiting the P53–MDM2 Interaction: An Important Target for Cancer Therapy. *Nat. Rev. Cancer* **2003**, *3* (2), 102–109.

(71)   Brodsky, F. M.; Chen, C.-Y.; Knuehl, C.; Towler, M. C.; Wakeham, D. E. Biological Basket Weaving: Formation and Function of Clathrin-Coated Vesicles. *Annu. Rev. Cell Dev. Biol.* **2001**, *17* (1), 517–568.

(72)   Conner, S. D.; Schmid, S. L. Regulated Portals of Entry into the Cell. *Nature* **2003**, *422* (6927), 37–44.

(73)   Robinson, M. S.; Bonifacino, J. S. Adaptor-Related Proteins. *Curr. Opin. Cell Biol.* **2001**, *13* (4), 444–453.

(74)   Bezerra, G. A.; Holenstein, A.; Foster, W. R.; Xie, B.; Hicks, K. G.; Bürer, C.; Lutz, S.; Mukherjee, A.; Sarkar, D.; Bhattacharya, D.; et al. Identification of Small Molecule Allosteric Modulators of 5,10-Methylenetetrahydrofolate Reductase (MTHFR) by Targeting Its Unique Regulatory Domain. *Biochimie* **2021**, *183*, 100–107.

(75)   Scannell, J. W.; Blanckley, A.; Boldon, H.; Warrington, B. Diagnosing the Decline in Pharmaceutical R&D Efficiency. *Nat. Rev. Drug Discov.* **2012**, *11* (3), 191–200.

(76)   Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings1PII of Original Article: S0169-409X(96)00423-1. The Article Was Originally Published in

Advanced Drug Delivery Reviews 23 (1997) 3. *Adv. Drug Deliv. Rev.* **2001**, *46* (1), 3–26.

(77)   Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chem. Biol.* **2014**, *21* (9), 1115–1142.

(78)   Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623.

(79)   Whitty, A.; Zhong, M.; viderengo, L.; Beglov, D.; Hall, D. R.; Vajda, S. Quantifying the Chameleonic Properties of Macrocycles and Other High-Molecular-Weight Drugs. *Drug Discov. Today* **2016**, *21* (5), 712–717.

(80)   Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3* (11), 935–949.

3. Experimentally Validating the Computational Veracity of CPMG

3.1 Introduction

The building blocks in CPMG were chosen to function as heterocyclic surrogates for the side chains of tyrosine and tryptophan. This choice was supported by calculated free energies of activation ($\Delta G^{\ddagger}$) [(ωB97X-D-SMD(methanol)/6-31G(d)) in Gaussian16 RevA.03], which indicated that reaction at the heterocyclic sites predicted by RegioSQM would occur via lower or similar activation barriers relative to those calculated for phenol and indole.[1] Based on this, we anticipated novel CPMG heterocycles to readily participate in macrocyclizations analogous to known participants. However, in the interest of computational runtimes, we had only evaluated core heterocycles to predict site reactivity in CPMG – the effects of conformational dynamics and surrounding structure on competing internal reaction rates during peptidyl cyclizations remained to be determined. Past experimental data showed that tyrosine and tryptophan side chains react rapidly with the cinnamyl cation, regardless of positioning within the peptide relative to the appended template (i.e G1-G3).[2] Overall, we expected a wealth of new reactivity and selectivity data to emerge from experiments designed to validate predictions in Ch.1 and Ch. 2.

All structures from CPMG originate from three types of monomers – (i) proteinogenic α-amino acids and their enantiomers, (ii) known, conformation-restricting proline analogs (see Ch.1), and (iii) heterocycles **1-30 (**Fig. 1) formulated into a matrix of α, β^2 and β^3 amino acids along with their respective one carbon homologs and enantiomers. Proteinogenic amino acids, their enantiomers and a subset of proline analogs are commercially available. The remaining proline analogs can be synthesized in optically active form using reported procedures.[3–7] Our initial synthetic efforts were geared towards identifying a general synthetic route for accessing our set of novel heterocycle-containing amino acids.
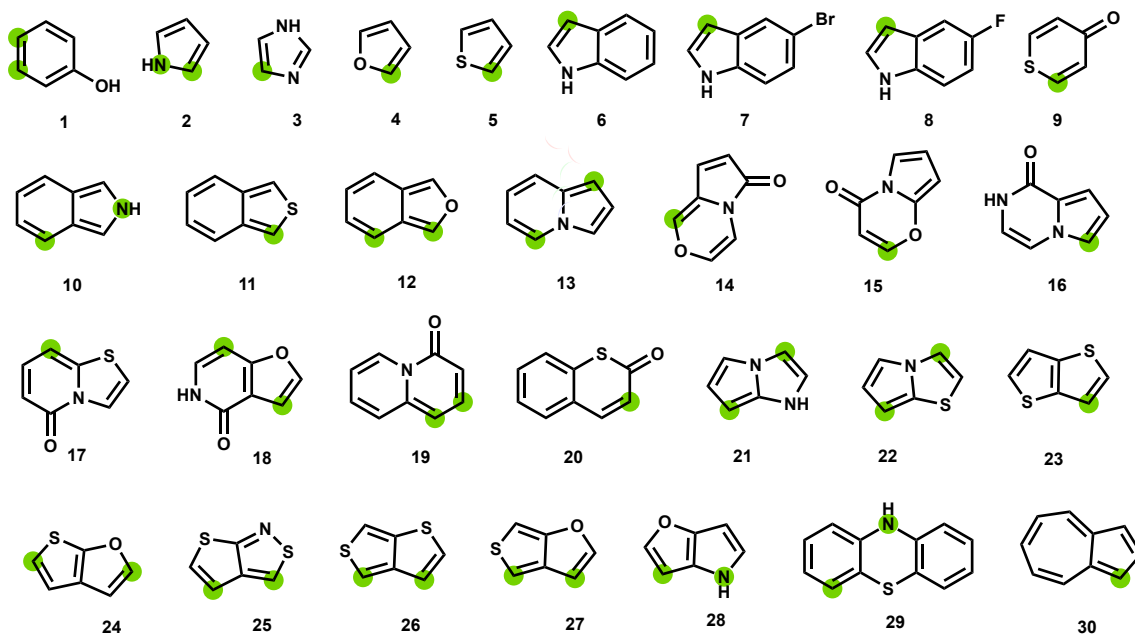
Figure 1. Heterocyclic building blocks in CPMG; green dots indicate sites of incorporation into amino acids.

3.2 General plan for monomer assembly

The synthesis of novel arene amino acids through palladium catalyzed cross coupling of serine derived organozinc reagents with aromatic electrophiles has been reported in the literature (Scheme 1A).[8–10] This approach has been particularly successful in accessing Boc-protected phenylalanine and pyridylalanine analogs and has been extended by several groups to the synthesis of other heterocyclic derivatives.[11–14] The success of this strategy relies on (i) the proper activation of zinc for efficient formation of the organozinc reagent, **32**, and (ii) access to an electrophile bearing the desired arene. We hypothesized this strategy would be amenable to numerous CPMG heterocycles, utilizing the corresponding heterocyclic mono halide in each case. L-iodoalanine, **31**, can be readily synthesized in various *N*-protected forms on multi-gram scale. For on scale solid phase peptide synthesis (SPPS), carbamates including Boc, Fmoc and Cbz groups have been developed to protect the amino group of amino acids. Among these, Fmoc-groups are the most widely used in SPPS, including in our own lab. We therefore envisioned employing the Fmoc-protected iodoalanine derivative, **36**, in our cross-coupling reactions to

eventually enable on-scale SPPS of the prepared monomers. Since Fmoc- groups tend to be more labile than their Boc-protected analogs, we focused our initial efforts on optimizing a route towards a relatively simple model compound **37** (Scheme 1C), wherein the aromatic electrophile, 3-bromothiophene would be commercially available.
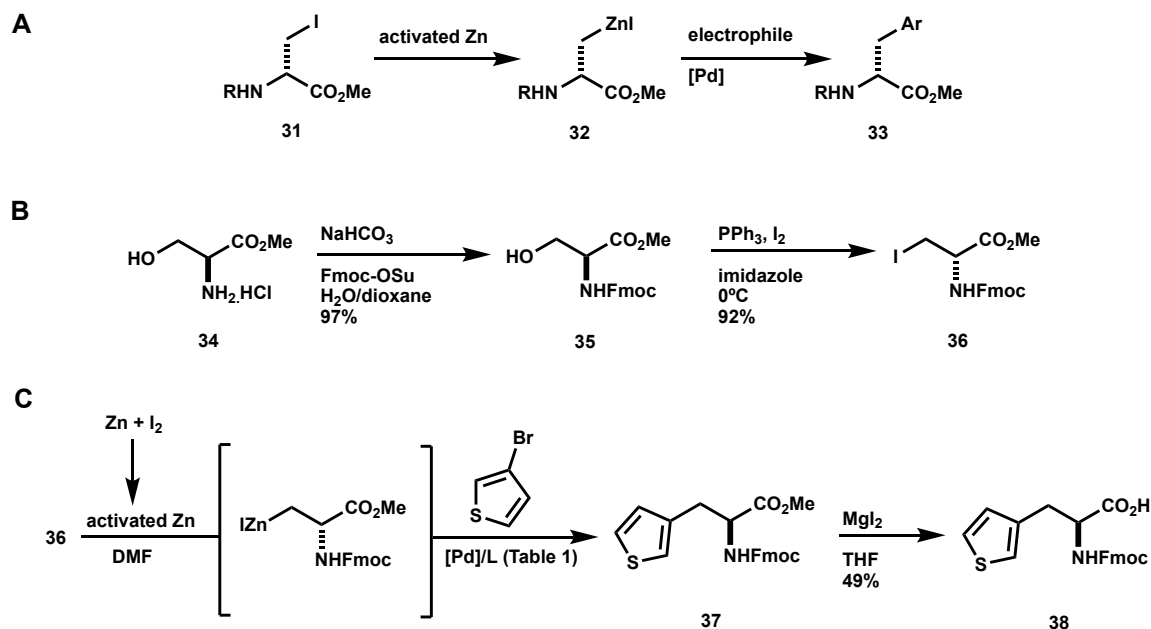
Scheme 1. **A.** General route for [Pd] catalyzed Negishi cross couplings of serine derived organozinc reagents with aromatic electrophiles to generate arene amino acids, R=Boc, Fmoc, CBz, Bn; **B.** Fmoc-L-iodoalanine-methyl ester is readily accessible from L-serine methyl ester hydrochloride; **C.** Cross coupling and ester deprotection for a model system.

Fmoc-protected L-iodoalanine methyl ester was prepared in high yield and on multi-gram scale in two steps from commercially available L-serine methyl ester hydrochloride (Scheme 1B).[15] Thereafter, formation of the desired organozinc reagent could be achieved using commercial zinc dust either activated sequentially with 1,2-dibromoethane and TMSCl in dry DMF, or with zinc dust activated using iodine in the absence of solvent. The insertion reaction with activated zinc was performed in dry DMF, wherein iodoalanine **36** was sonicated in the solvent for several minutes to fully solubilize it. Although complete zinc reagent formation was observed using either method of zinc activation, the reaction was much faster when zinc was activated with iodine. With the organozinc compound in hand, we looked to evaluate catalytic systems for performing the desired Negishi cross coupling reaction (Table 1).

Recently, Pd[P(t-Bu)$_3$]$_2$ catalyzed Negishi cross couplings between organozinc reagents and (hetero)aryl bromides using bulk water with NaCl additive have been reported.[16] It is believed that the addition of NaCl leads to a so-called "halide effect" in creating a Pd-ate complex that is significantly more reactive towards aryl halides due to improved rates of oxidate addition. However, these initial coupling conditions failed to produce our desired cross coupling product (Table 1, entry 1), and altering the reaction stoichiometry, temperature and time led to a very moderately improved yield (entry 2). The lack of product formation was in part due to the formation of Fmoc-L-Ala-OMe, which arises from protonation of the organozinc intermediate, and is well documented in the literature. To obtain an improved synthesis of **37**, we sought to alter the reaction conditions. The bulky biaryl ligand SPhos, has been shown to improve Pd cross-coupling reactions in both Suzuki-Miyaura reactions as well as Negishi type reactions.[8,17,18] The reason for this efficiency has been ascribed to the proclivity of SPhos to donate electron-density to the intermediate Pd(0) complex, thereby facilitating the rate of oxidative addition during cross coupling. It has also been proposed that such ligands cause a reasonable amount of highly reactive monoligated palladium to form. Indeed, employment of this ligand dramatically improved the reaction yield (entry 4). We found the bulky P(t-Bu)$_3$ groups on the Pd catalyst to be equally critical to the success of the reaction. Other Pd catalysts (entries 3, 5-8), caused reaction yield to suffer significantly in each case. These screens were easily quantifiable by crude NMR (Fig. 2) from integrating the multiplets at 6.83 and 6.94 ppm (arising from thiophene protons in the product) relative to the doublet at 7.77 ppm (arising from Fmoc protons in both starting material and product). Despite the moderate yield in the best case for this test system, we had successfully demonstrated engaging **36** with our model heteroaryl electrophile and decided to move forward with these results.

Table 1. Catalyst screens for Negishi cross coupling of **36** (1 eq.) with 3-bromo thiophene (1 eq.) in the presence of activated zinc (2 eq.). 3 mol% [Pd] and 10 mol% L unless otherwise indicated.

Entry	[Pd] Catalyst	Additive L	Solvent	Temperature, Time	Crude yield[d]
1[a]	Pd[(P(t-Bu)$_3$]$_2$	NaCl	H$_2$O/DMF	rt, 20 seconds	0%

2[b]	Pd[(P(t-Bu)$_3$]$_2$	NaCl	H$_2$O/DMF	50 °C for 5 hours, rt o/n	13%
3[c]	Pd(PPh$_3$)$_2$Cl$_2$	–	DMF	50 °C for 5 hours, rt o/n	0%
4	Pd[(P(t-Bu)$_3$]$_2$	SPhos	DMF	50 °C for 5 hours, rt o/n	56%
5	Pd(dba)$_2$	SPhos	DMF	50 °C for 5 hours, rt o/n	20%
6	Pd(PPh$_3$)$_4$	SPhos	DMF	50 °C for 5 hours, rt o/n	6%
7	Pd(OAc)$_2$	SPhos	DMF	50 °C for 5 hours, rt o/n	2%
8	[PdCl(C$_3$H$_5$)]$_2$	SPhos	DMF	50 °C for 5 hours, rt o/n	23%

[a] **36** (3 eq.), 2.5 mol% [Pd], NaCl (1 eq.); [b] **36** (1.3 eq.), 8 mol% [Pd], NaCl (1 eq.); [c] 5 mol% [Pd]; [d] Yield estimated from crude NMR (see Fig. 2)

Although Fmoc- groups are widely utilized in peptide chemistry, they can be unstable to conventional ester cleavage in basic conditions. The recent development of MgI$_2$-assisted protocols as an alternative to conventional ester deprotection methodologies has extended the orthogonal flexibility of this protecting group strategy.[19] MgI$_2$-assisted dealkylation has the added benefit of avoiding
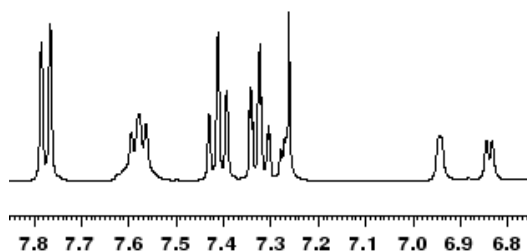
Figure 2. Yields in Table 1 were estimated from ^{1}HNMR spectra of crude mixtures by integrating the signals at 6.83 and 6.94 ppm (arising from thiophene protons in the product) relative to the doublet at 7.77 ppm (arising from Fmoc protons in both starting material and product).

potential epimerization at the alpha carbon, which often presents as an issue during base promoted deprotection. Given these advantages, we attempted literature conditions of MgI$_2$ under microwave irradiation in THF to dealkylate the methyl ester in **37**. To our delight, we successfully isolated the deprotected amino acid, **38**, in moderate yield, and preserving the Fmoc-group. We also demonstrated the reaction proceeds without loss of yield when performed at reflux in EtOAc, thereby increasing the accessibility and scalability of the protocol. Having successfully demonstrated a cross coupling and deprotection strategy for our model system, we looked to reproduce these results for a set of CPMG heterocycles (**39-42** in Fig. 3) – first, by accessing
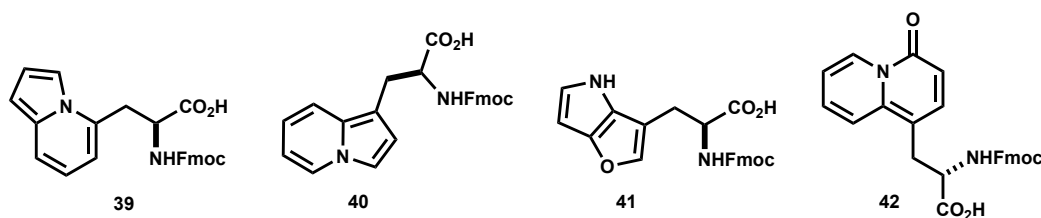

Figure 3. Amino acids chosen for performing initial proof-of concept experiments.
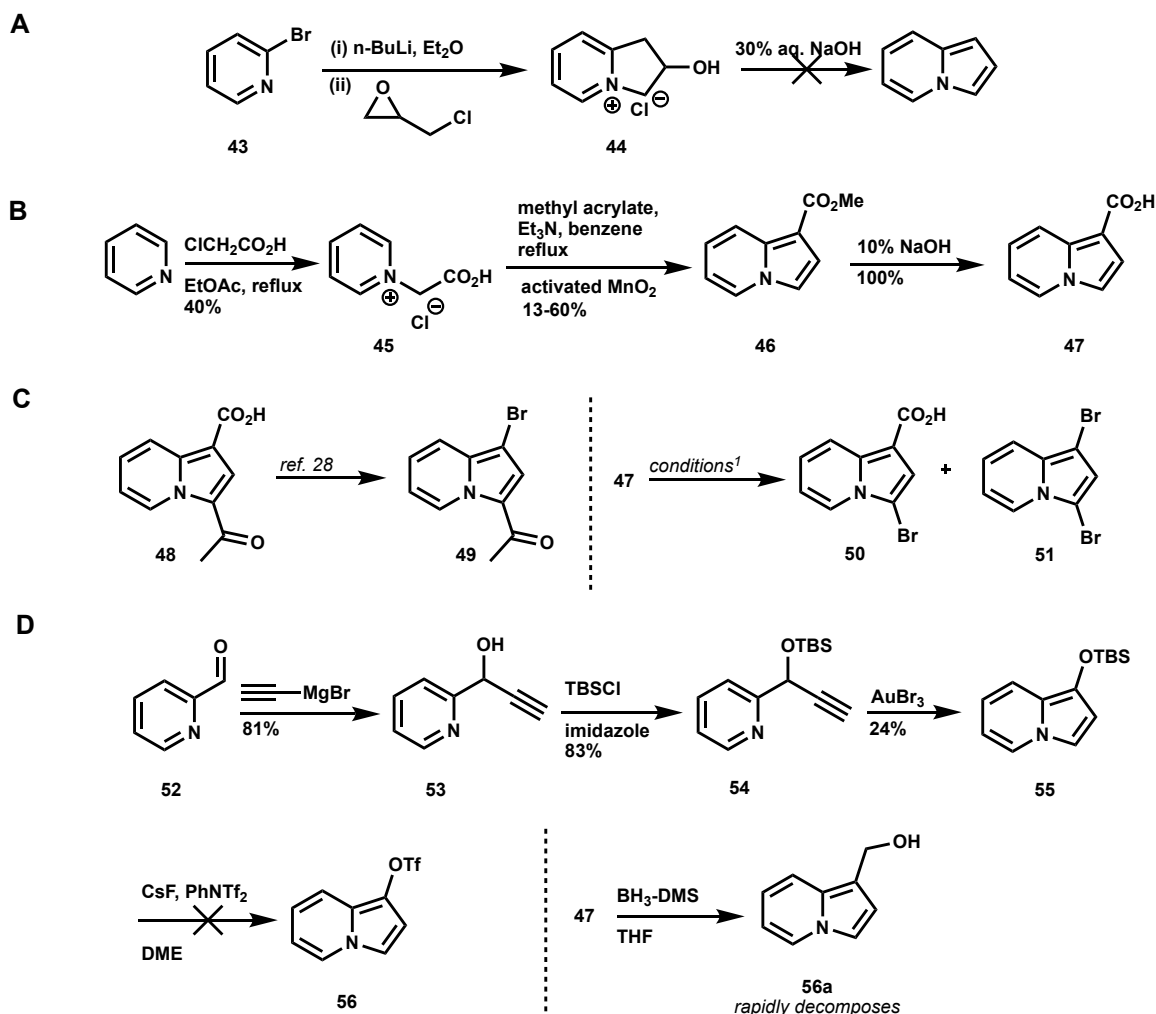
these motifs in their respective mono halide forms.

3.3 Synthetic routes towards (indolizin-5-yl) and (indolinzin-1-yl) amino acids 39 and 40 and their engagement in macrocyclization

Indolizines are key intermediates in the synthesis of biindolizines, cyclophones, cyclazines and alkaloids.[20] The Chichibabin reaction has been conventionally used for their synthesis, but this strategy requires special precursors, and necessarily generates 2-substituted indolizines.[21,22] Indolizines bearing substitution exclusively at C-5 (as in **39**) or C-1 (as in **40**) are challenging to access because of the π-density of the heterocycle, which is highest at C-3. For instance, the reaction of a model system, 1-methylindolizine, with a model cinnamyl cation TS (Fig. 4) is barrierless and entropically controlled at C-3, relative to the corresponding free energies of activation ($\Delta G^{\ddagger}$) for phenol and indole, shown in Fig. 4 [(ωB97X-D-SMD(methanol)/6-31G(d)) in Gaussian16 RevA.03].[1] In theory, 3-unsubstituted indolizines would facilitate several electrophilic additions at C-3 – including electrophilic macrocyclizations, which are of most interest to us – but the existing few methods for preparing these compounds suffer from low yields and unscalable or tedious procedures.[23,24]
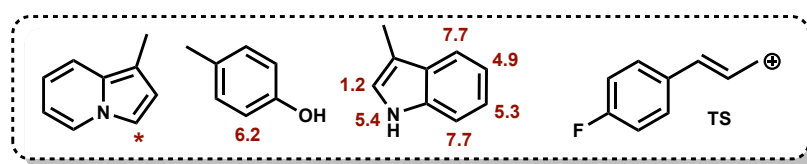
Figure 4. Calculated free energies of activation ($\Delta G^{\ddagger}$/kcal mol^{-1}) for the reaction of the indicated aromatic positions with the model cinnamyl cation, **TS**. Asterisk (*) denotes a site where DFT calculations (ωB97X-D-SMD(methanol)/6–31G(d)) indicate barrierless, entropically controlled reactions.

We began our studies towards **39** and **40** with a reported protocol, where 2-bromopyridine reacts with epichlorohydrin to first form the dihydro-indolizinium salt **44** (Scheme 2A).[25] Although this step proceeded as reported, the subsequent dehydration to afford the desired unsubstituted

A

43 → **44** (i) n-BuLi, Et₂O; (ii) epoxide/CH₂Cl → 30% aq. NaOH ✗ →

B

pyridine → ClCH₂CO₂H, EtOAc, reflux, 40% → **45** → methyl acrylate, Et₃N, benzene reflux, activated MnO₂, 13-60% → **46** (CO₂Me) → 10% NaOH, 100% → **47** (CO₂H)

C

48 → ref. 28 → **49** | **47** → conditions[1] → **50** + **51**

D

52 → ≡—MgBr, 81% → **53** (OH) → TBSCl, imidazole, 83% → **54** (OTBS) → AuBr₃, 24% → **55** (OTBS)

CsF, PhNTf₂, DME ✗ → **56** (OTf) | **47** → BH₃-DMS, THF → **56a** (OH) *rapidly decomposes*

Scheme 2. Efforts to furnish the indolizine core. The fully substituted system could not be accessed by the route from 2-bromopyridine shown in **A**, but the C-1 substituted acid, **47** is readily accessed *via* the route shown in **B**; **C.** Literature reports shown **49** can be accessed from **48**, but similar conditions failed to effect the desired bromo-decarboxylation on **47**, affording **50** and **51** instead; [1]Attempted conditions are as follows – NBS (1 eq.), DMF; NBS (1 eq.), NaHCO₃ (3 eq.), DMF; K₃PO₄ (1 eq.) with N(nBu)₄Br₃ (1 eq.), MeCN; PdCl₂ (0.05 eq.), Ag₂CO₃ (0.5 eq.), CuBr₂ (1 eq.), THF; **D. 55** was synthesized using reported procedures, but attempted conversion of **55** to triflate **56** resulted in a mixture of non-isolable compounds, likely due to the instability of leaving groups at C-1 on the ring; subjecting **47** to reducing conditions affords the alcohol **56a**, but this compound rapidly decomposes.

heterocycle did not yield any product in our hands. We then turned to another reported protocol exploiting pyridinium *N*-methylides to furnish the indolizine core.[26] First, pyridine reacts with chloroacetic acid to afford the corresponding pyridinium salt, **45** (Scheme 2B). This salt undergoes a decarboxylative 1,3 dipolar cycloaddition with methyl acrylate in the presence of MnO₂ to furnish the indolizine core in **46**. Our key finding in this step was that effective activation of MnO₂ to remove adsorbed water is critical to the success of the reaction.[27] When commercial MnO₂ was

used straight from the bottle, the desired product was not observed. This dramatically changed when MnO_2 was first subjected to vigorous azeotropic removal of water in benzene over several hours. In multiple attempts, we obtained yields ranging from 13-60% in the cycloaddition reaction, despite our best efforts to reproduce exact conditions each time. We surmise that the success of the reaction hinges upon the extent of water removal from MnO_2, which is highly sensitive to experimental handling. With our indolizine core in hand, we envisioned performing decarboxylation-halogenation sequences to afford both C-3 and C-5 substituted indolizines. Towards this, we first saponified the methyl ester to afford the corresponding acid, **47**, in quantitative yield. Our subsequent efforts to furnish each regioisomer are detailed in the following two sections.

3.3.1 (indolinzin-1-yl) amino acid

Procedures for the decarboxylative halogenation of **48** (Scheme 2C) using NBS have been reported in the litetature.[28] Unfortunately, subjecting the 3-unsubstituted system, **47**, to NBS in DMF afforded only **50**, bearing the desired halide at C-1 but with the carboxylic acid still present. When NBS in the presence of $NaHCO_3$ was used, some amount of the dibrominated species **51** was observed, but the major product of the reaction was still **50**. Attempting different brominating conditions[29] such as K_3PO_4 with $N(nBu)_4Br_3$ yielded largely the same results. We then attempted a Hunsdiecker-like bromodecarboxylation of **47** using a $PdCl_2$ catalyst with Ag_2CO_3 additive and $CuBr_2$ as the bromination source.[30] This method has been shown to successfully bromodecarboxylate several electron rich arenecarboxylic acids, including analogs of methylanisole and indole. Once again, however, starting material was fully consumed to afford **50** as the single product of the reaction. Finally, subjecting the fully unsubstituted indolizine (*vide infra*) to NBS generated a product with a mass corresponding to a tribrominated indolizine. Looking to identify other potential cross coupling partners, we performed an Au(III) catalyzed conversion of silyl ether protected propargylpyridine **54** into **55** (Scheme 2D).[24] We wondered if

transforming the silyl protecting group in **55** to the corresponding triflate would allow us to perform our desired cross coupling reaction. Unfortunately, subjecting **55** to CsF and PhNTf$_2$ generated a dense mixture of unidentifiable compounds. We hypothesized that the nitrogen centered lone pair might destabilize any functional group at C1 bearing a leaving group, possibly through the formation of a highly reactive indolizinium intermediate. This theory was supported by other observations. When **47** was subjected to either DIBAL or borane reductions, TLC monitoring indicated full consumption of starting material to a single compound. Although this major product was characterized to be alcohol **56a**, the compound decomposed rapidly during work up and characterization, and the intermediate aldehyde was never observed.

Since all attempts at functionalizing C-1 had failed, we looked to alter our original strategy by revisiting the 1,3 dipolar cycloaddition and
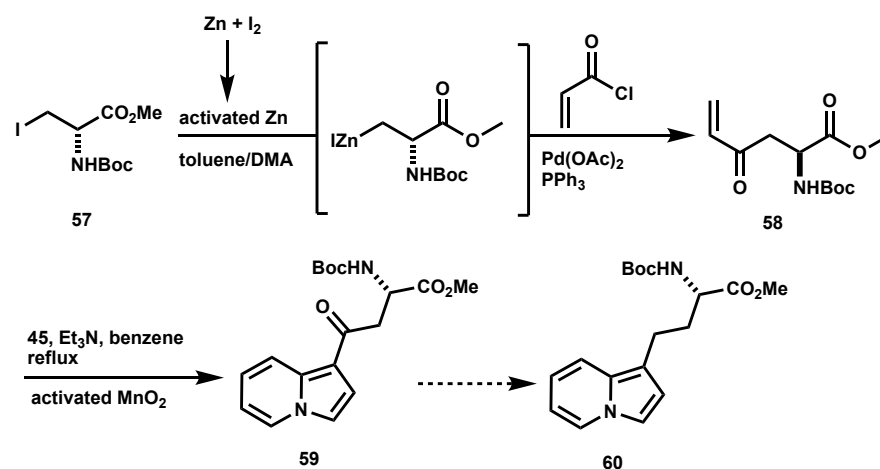
Scheme 3. Modified route to indolizin-1-yl amino acid.

preinstalling the amino acid functionality in that step. For instance, enone **58** (Scheme 3) has been reported in the literature and can be prepared through Negishi cross coupling of acryloyl chloride with Boc-iodoalanine **57**. To our delight, the resulting cross coupled product, **58**, successfully underwent oxidative 1,3 dipolar cycloaddition with pyridinium chloride **45** and activated MnO$_2$ to afford our desired indolizine regioisomer. Efforts are currently underway to selectively reduce the ketone in **59** to generate the one-carbon homologue, **60**, of our desired α-monomer, **40**. Both **40** and **60** are in fact present in CPMG, and efforts are underway to determine an appropriately modified route for accessing **40**.

3.3.2 (indolinzin-5-yl) amino acid

After a few attempts at decarboxylation using conventional methods,[30–32] we were pleased to find that a suspension of the indolizine carboxylic acid in ethanolamine at reflux cleanly affords the decarboxylated heterocycle, **61** (Scheme 4), as the major product within an hour. When subjected to n-BuLi, this fully unsubstituted heterocycle generates the most stable anion at C-5, which can be trapped with iodine to generate 5-iodoindolizine, **62**, in high yield.[33,34] The regiochemistry in this step was confirmed by ^{1}HNMR where the signal corresponding to the C-5 proton on the unsubstituted heterocycle, having the highest shift, is absent after iodination. 5-iodoindolizene was subjected to the highest yielding cross coupling conditions identified in section 3.2, (Table 1, entry 4), wherein we were delighted to observe formation of the desired amino acid monomer, **63**, but in low yield (Table 2, entry 1). Encouraged by these results, we performed a catalyst screen, seeking to improve the efficiency of the reaction for scale up (Table 2).
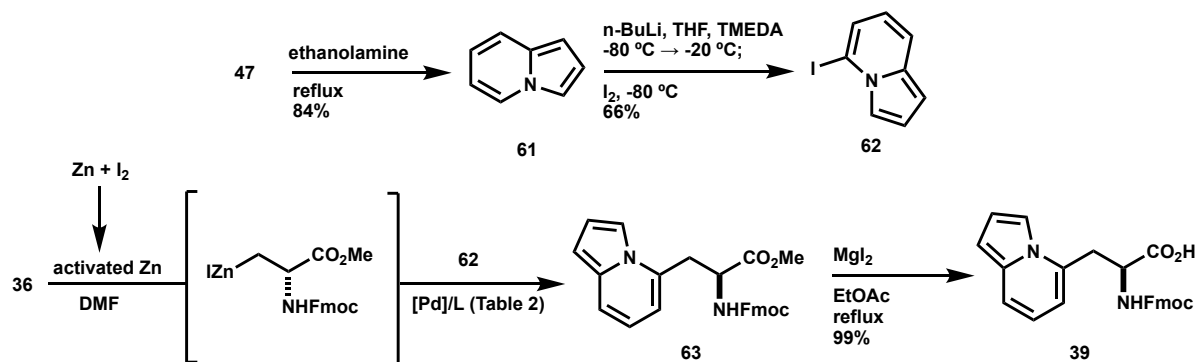
Scheme 4. 5-iodoindolizine is synthesized from **47** in two steps, and subsequently undergoes Pd-catalyzed cross-coupling with **36** derived organozinc reagent to afford the desired inolizin-5-yl amino acid.

Since Pd[P(tBu)$_3$)]$_2$ is highly air and moisture sensitive, yield was improved when we switched to a new bottle of the catalyst (Table 2, entry 1, 6). However, it was still far less efficient for the cross-coupling reaction with **62**, compared to reaction with 3-bromothiophene. The best catalysts for 5-iodoindolizine were found to be Pd$_2$(dba)$_3$ (entry 2) and Pd(OAc)$_2$ (entry 4). It has been reported that Heck reactions with Pd$_2$(dba)$_3$ can be improved by using solid-liquid phase transfer conditions with tetrabutyl ammonium chloride as an additive,[35] however we did not

observe any enhancements for this system (entry 7). Since Pd(OAc)$_2$ is bench stable and relatively cheap, we opted to move forward with this catalyst, and screened higher loadings with varying catalyst-ligand ratios (entries 9-11). We found a 1:3 [Pd]:Sphos ratio to be most effective, and obtained our highest yield with a 10 mol% catalyst loading (entries 2, 11). With these conditions, we were able to synthesize our desired monomer, **63**, on multi gram scale. The methyl ester was deprotected using the MgI$_2$ assisted dealkylative protocol detailed above to the corresponding amino acid, **39**, now ready for incorporation into a peptide.

Table 2. Catalyst screens for Negishi cross coupling of **36** (1 eq.) with 5-iodo indolizine (1 eq.) in the presence of activated zinc (2 eq.), 3 mol% [Pd] and 10 mol% L unless otherwise indicated; all reactions were run in DMF at 50 °C for 5 hours and subsequently stirred at room temperature overnight.

Entry	[Pd] catalyst	Additive L	Isolated Yield
1	Pd[P(tBu)$_3$)]$_2$	Sphos	13%
2	Pd$_2$(dba)$_3$	Sphos	41%
3	Pd(PPh$_3$)$_4$	Sphos	Negligible
4	Pd(OAc)$_2$	Sphos	40%
5	[PdCl(C$_3$H$_5$)]$_2$	Sphos	24%
6	Pd[P(tBu)$_3$)]$_2$	Sphos	24%
7	Pd$_2$(dba)$_3$	Sphos; (n-Bu)$_4$NCl[a]	17%
8[36]	[Pd(μ-I)(PCy$_2$tBu)]$_2$	–	2%
9	Pd(OAc)$_2$[b]	Sphos[b]	13%
10	Pd(OAc)$_2$[b]	Sphos[c]	22%
11	Pd(OAc)$_2$[b]	Sphos[d]	58%

[a] 1 eq; [b] 10 mol%; [c] 20 mol%; [d] 33 mol%.

We were eager to study the reactivity of these novel arenes in macrocyclizations, relative to monomers like Tyr and Trp which have been studied in great depth in our lab.[2,37,38] Towards our first peptide sequence, **64**, **39** was coupled with L-tyrosinamide hydrochloride and Boc-Ala-OH using standard solution phase peptide synthesis procedures (Scheme 5). Interestingly, all compounds containing the indolizine core are highly fluorescent in solution. These compounds are typically vibrantly colored in solution and often change color with solvent. The tripeptide, **64**, was coupled to **G1** template and the product was carefully purified before the next step. Thereafter, macrocyclization of **65** with 5%v TFA in MeNO$_2$ was monitored by TLC and HPLC. Starting material was fully consumed with formation of a single new product with the desired mass

Scheme 5. Acid promoted internal cinnamylation of an indolizine containing peptide, **65**; complete experimental detail for solution phase peptide synthesis included in the experimental appendix.

of **66** within 30 minutes. The reaction was worked up using standard procedures and purified by prep-HPLC to afford the product as a purple solution, which was lyophilized to a brown solid. The ^{13}CNMR spectrum for the isolated compound revealed a non-symmetric Ph ring (Fig. 5), strongly suggesting the identity of the structure to be **66a** rather than **66b**. This was confirmed by full structural assignment using HSQC-, HMBC- and COSY- based NMR connectivity maps. Key HMBC correlations are indicated in Fig. 5 (full structural assignment in the Experimental Appendix). Given the high nucleophilic reactivity of C-3 on indolizine, particularly relative to phenol, we surmised that the positioning of the indolizine within the peptide in this case makes

macrocyclization highly entropically disfavored. This can be analogized to a reactivity trend we have observed for Trp – even though ΔG‡ of reaction (Fig. 4) for a model indole system, 3-methylindole, is more favorable at C-2 than at any other site on the ring, we have never observed
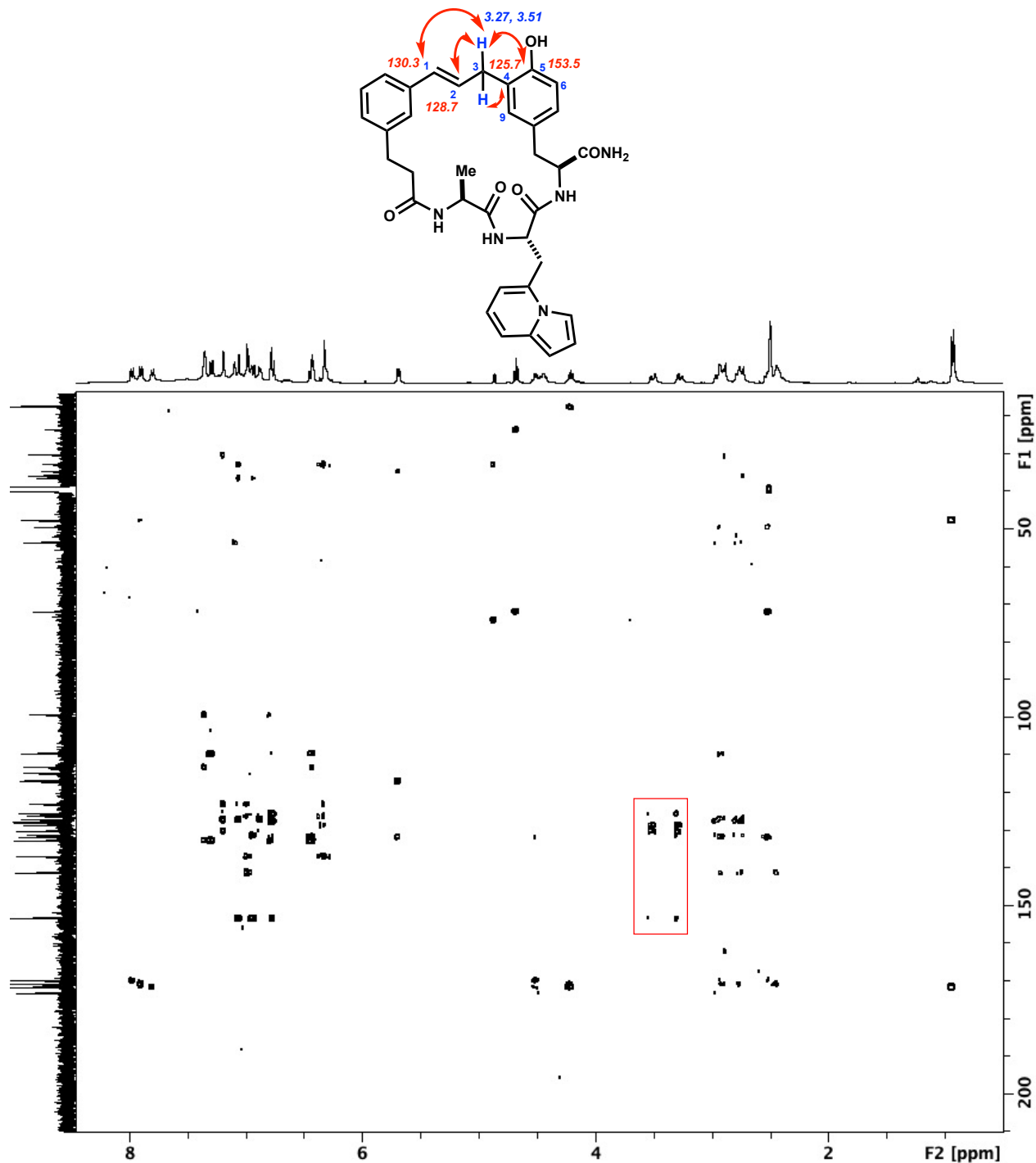
Figure 5. ^{1}H-^{13}C HMBC shows clear correlations between protons *H3* to *C1, C2, C4 and C5.* These correlations, along with the non-symmetric nature of the^{13}C signals arising from the phenyl ring, confirm engagement of the Tyr residue in cinnamylation.

macrocyclization at this position. Experiments are currently underway to study whether a switched order peptide, **67**, having a terminal indolizine would be more favored to macrocycle through engagement of indolizine (**68a)** rather than phenol (**68b**).

In addition to acid promoted macrocyclizations, our lab has previously demonstrated metal catalyzed cinnamylations of aromatic nucleophiles with our templates. In these reactions, palladium catalysis generates a metal stabilized cinnamyl cation, which can be trapped by oxygen or nitrogen nucleophiles to generate the corresponding carbon-heteroatom bonded macrocycle. As expected, in the presence of Pd[(PPh)$_3$]$_4$, **65** macrocyclizes at the phenolic oxygen to generate macrocyclic ether **69** (Scheme 6). In this case, the phenolic carbon signals are symmetric and 2-D NMR correlation data including HSQC, HMBC and COSY allow us to confidently assign the structure (full assignment in Experimental Appendix). Our lab has previously demonstrated that regioisomeric product outcomes during macrocyclization can be altered using structural preorganization in the starting material. In such cases, subjecting a macrocyclic ether such as **69** to acidic conditions, typically promotes a rearrangement of the C-O bond to a C-C bonded macrocycle *ortho* to phenol. This is a useful feature of the chemistry, and often allows us to bias reaction outcome to favor a specific macrocycle in cases where acid promoted macrocyclization affords a regioisomeric product distribution.[2] Since indolizine does not compete with phenol in the formation of **66a** versus **66b**, we would simply expect rearrangement of **69** back to **66a**. To confirm this, **69** was subjected to standard conditions of 15 eq. TFA in DCM. Typically, such
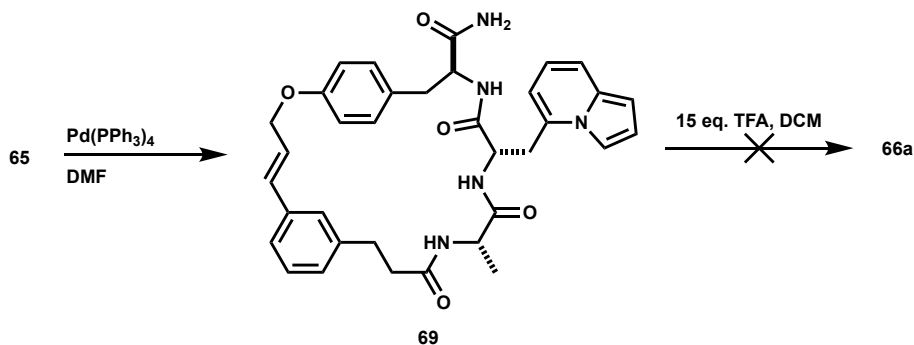
Scheme 6. Palladium catalyzed C-O bond formation by engagement of Tyr in **65** to generate **66**. Subjecting this macrocyclic ether to the indicated acidic conditions does not promote the expected rearrangement to the corresponding *ortho* C-C bonded compound.

88

rearrangement reactions are complete within 2 hours, but we were surprised to find that starting material was still present after 6 hours, and there was no appreciable formation of any new compounds. Although this was unexpected, we have shown previously that varying the acidic conditions for macrocyclization can also alter reaction outcome.[2] Experiments are underway to establish whether the use of stronger acidic conditions promotes the expected rearrangement from **69** back to **66a**.

3.4 Synthetic route towards (furopyrrol-3-yl) amino acid 41

The synthesis of fully unsubstituted 4*H*-furo[3,2-*b*]pyrrole has been previously reported.[32] The precedented route was attractive because the choice of starting material, in this case 4-bromofurfural, would allow us to set the regiochemistry of the desired monomer, **41**. Accordingly, ethyl 2-bromoacetate was first reacted with NaN$_3$ to generate ethyl 2-azidoacetate, **71** (Scheme 7). The azide was reacted with 4-bromofurfural, which is both commercially available and synthesizable from furfural, to generate a single stereoisomer of intermediate **72**.[39] This intermediate can subsequently undergo a nitrene C-H insertion under thermal conditions to afford the fused pyrrole, **73**. Care should be taken in all three steps, particularly when scaling up, as both azides and the highly reactive nitrene intermediate are potentially explosive. The resulting furopyrrole carboxylate was saponified to the corresponding acid **74**, which upon decarboxylation would afford our desired cross coupling partner.
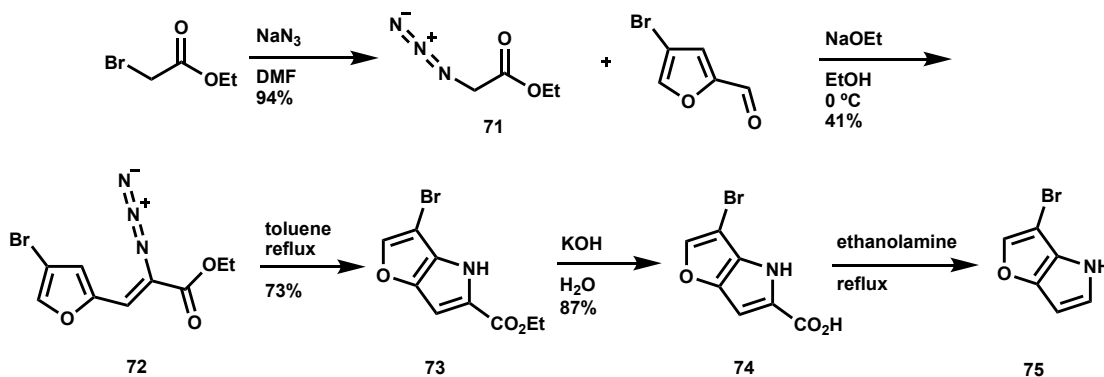
Scheme 7. Accessing 4-bromo furopyrrole.

89

Decarboxylation in ethanolamine at reflux was complete within a few minutes and the reaction was rapidly worked up at low temperatures. The product is a pale purple solid upon isolation and the structure was confirmed by ^{1}HNMR to be the desired heterocycle, **75**. However, consistent with previous reports,[32] this compound is highly unstable and disintegrated into black tar within the few minutes required to run a ^{13}CNMR, wherein decomposition over time was evident by ^{1}HNMR. Nevertheless, reports in the literature suggest that certain *N*-protected forms of the heterocycle are bench stable.[40] It is possible that this monomer will need to be formulated in an *N*-protected form and subjected to late stage deprotection. We sought to use an orthogonal protecting group to Fmoc to facilitate ease of subsequent deprotection. Towards this, we attempted to Teoc- protect the pyrrole nitrogen on **74** using NaH (2 eq.), but no reaction was observed after several hours, suggesting that a stronger base is needed to deprotonate the pyrrolic nitrogen. Experiments are underway to establish the most viable route in this regard.

3.5 Synthetic route towards (quinolizinon-1-yl) amino acid 42

We began our route towards iodo-quinolizinone with diethyl 2-(ethoxymethylene)malonate, which is a cheap, commercially available reagent. Lithiated 2-picoline undergoes a 1,4 addition with this starting material to generate intermediate **76** (Scheme 8B). A precedent in patent literature reported obtaining the quinolizinone core from intermediate **76** when it was subjected to a mixture of polyphosphoric acid and acetic acid.[41] Since polyphosphoric acid is tedious to handle, we attempted a modified reaction by heating **76** at high temperatures in xylenes in the presence of a catalytic amount of acetic acid. We were delighted to find that the desired heterocyclic core, **77**, was furnished in significantly higher yield than the reported protocol.

At this stage, we had two options – saponification and decarboxylation at C-3 followed by halogenation at C-1, or vice versa. We had previously calculated $\Delta G^{\ddagger}$ for the reaction of a model system, 2-methylquinolizinone, with a model cinnamyl cation TS (Scheme 8A) to be 2.6 kcal mol[-]

[1] at the unsubstituted C-3 atom, making this site more reactive [(ωB97X-D-SMD(methanol)/6-31G(d)) in Gaussian16 RevA.03] than the unsubstituted C-1 site, with $\Delta G^{\ddagger} = 5.9$ kcal mol[-1]. Based on this, and our experience with the 1-substituted indolizine, we opted to first halogenate at C-1, prior to attempting decarboxylation at C-3 to avoid running into regioselectivity issues while installing the halide.
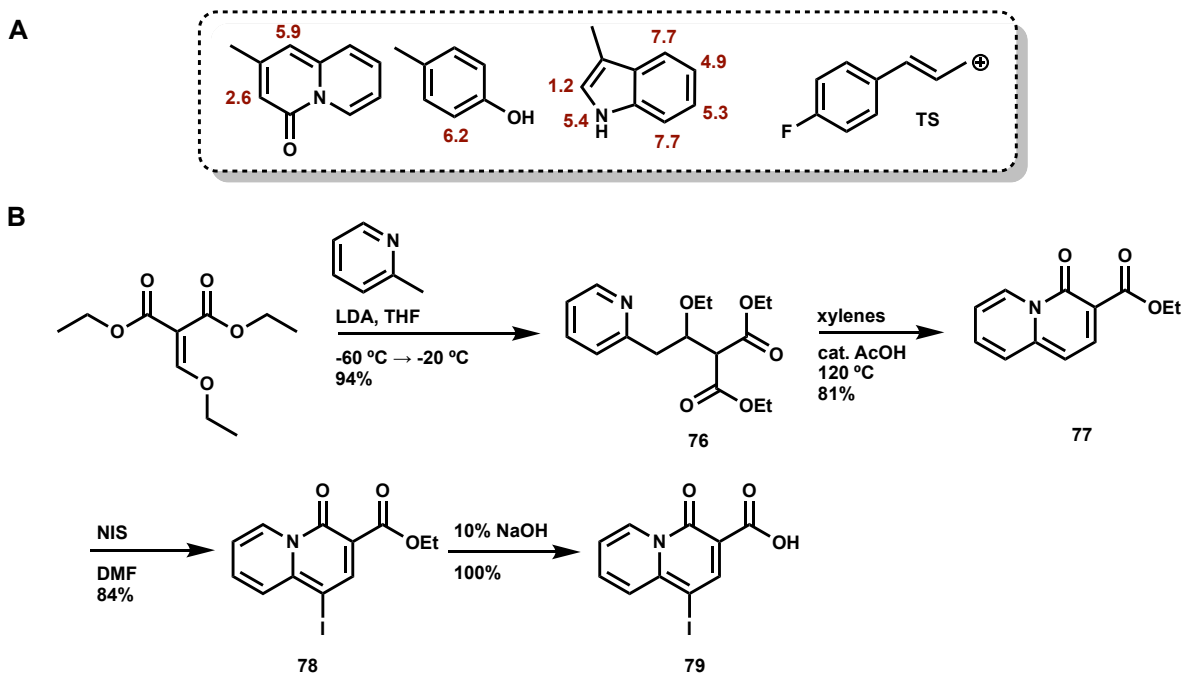
Scheme 8. A. Calculated free energies of activation ($\Delta G^{\ddagger}$/kcal mol[-1]) for the reaction of the indicated aromatic positions with the model cinnamyl cation, TS; **B.** Efforts towards accessing 1-iodo quinolizinone.

Along those lines, **77** was subjected to NIS, resulting in **78,** which was further saponified to generate the desired iodo-quinolizinone carboxylic acid, **79**. Unfortunately, subjecting this heterocycle to ethanolamine even at room temperature generated a non-separable mixture of at least three quinolizinone molecules, based on ^{1}HNMR data. We hypothesize that at least one of the unwanted side products arises from cleavage of the weak C-I bond on the heterocycle. This should be a surmountable issue – for instance, switching to the corresponding bromide should produce fewer undesired side reaction pathways. Alternatively, the order of events may be revisited to perform decarboxylation at C-3 on **77** prior to halogenation, followed by separation of resulting regiosisomers. Cross coupling at the iodide to the serine derived organozinc reagent,

prior to decarboxylation, is also conceivable. Experiments are underway to establish the most viable route in this regard.

3.6 Conclusions and outlook

Overall, we have begun to validate the computational strategy that was employed to create CPMG. Although the heterocyclic building blocks comprising the substrate library in CPMG were strategically chosen to participate as functional isosteres of indole and phenol, inevitably, there were underlying assumptions. Many of the chosen heterocycles had previously been reported only as subunits within larger polycyclic motifs or drug molecules, and little was known about their synthetic tractability in isolation or in their respective amino acid forms. Along those lines, fully validating CPMG requires demonstrating the ability to – (i) formulate the building block heterocycles in their desired amino acid forms, (ii) demonstrate the resulting monomers to be stable to standard coupling, deprotection and macrocyclization procedures, and (iii) engage the heterocycles during macrocyclization as anticipated. In this work, we have begun to expound on some of these ideas for four CPMG monomers and are encouraged by the results. We are close to demonstrating all four amino acids to be synthetically tractable, and have successfully demonstrated a full monomer to macrocycle route employing the indolizin-5-yl amino acid. Template engagement with these heterocycles remains to be demonstrated. However, we are encouraged by experimental observations to date – in particular, sites of experimental electrophilic addition reactions correlate fully with CPMG anticipated sites of reactivity. There may yet be cases where experimental data requires modifications to CPMG code. For instance, in the case of the furopyrrole monomer, **41**, it may be necessary to modify the input heterocycle to be in an *N*-protected form to avoid stability issues during synthesis. It may also become necessary to incorporate rules or estimates for strain energy into CPMG to allow it to distinguish between products such as **66a** and **66b**, and thereby accurately anticipate experimental product distributions. Overall, the viability of using CPMG as a resource for protein-ligand discovery will

continue to grow as we refine our understanding of the relative reactivities of arenes **1**-**30**. These experiments will also enable future testing for next stage validations of anticipated protein-ligand interactions. As computer-assisted drug discovery becomes increasingly pervasive in medicinal chemistry, we are uniquely positioned to establish a feedback loop between experiments and computations. We hope these efforts will culminate in a highly refined predictive model.

3.7 References

(1)   Saha, I.; Dang, E. K.; Svatunek, D.; Houk, K. N.; Harran, P. G. Computational Generation of an Annotated Gigalibrary of Synthesizable, Composite Peptidic Macrocycles. *Proc. Natl. Acad. Sci.* **2020**, *117* (40), 24679–24690.

(2)   Lawson, K. V; Rose, T. E.; Harran, P. G. Template-Induced Macrocycle Diversity through Large Ring-Forming Alkylations of Tryptophan. *Tetrahedron* **2013**, *69* (36), 7683–7691.

(3)   Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Thornton, J. M. Shape Variation in Protein Binding Pockets and Their Ligands. *J. Mol. Biol.* **2007**, *368* (1), 283–301.

(4)   Grygorenko, O. O.; Artamonov, O. S.; Palamarchuk, G. V; Zubatyuk, R. I.; Shishkin, O. V; Komarov, I. V. Stereoselective Synthesis of 2,4-Methanoproline Homologues. *Tetrahedron: Asymmetry* **2006**, *17* (2), 252–258.

(5)   Shinisha, C. B.; Sunoj, R. B. Bicyclic Proline Analogues as Organocatalysts for Stereoselective Aldol Reactions: An in Silico DFT Study. *Org. Biomol. Chem.* **2007**, *5* (8), 1287–1294.

(6)   Petter, R. C. Synthesis of a New Analog of Proline. *Tetrahedron Lett.* **1989**, *30* (4), 399–402.

(7)   Kondaiah, G. C. M.; Vivekanandareddy, M.; Reddy, L. A.; Anurkar, S. V; Gurav, V. M.; Ravikumar, M.; Bhattacharya, A.; Bandichhor, R. Asymmetric Synthesis of (S,S,S)-2-Aza-Bicyclo-[3.3.0]-Octane-3-Carboxylic Acid Benzyl Ester: Formal Synthesis of Ramipril. *Synth. Commun.* **2011**, *41* (8), 1186–1191.

(8)   Ross, A. J.; Lang, H. L.; Jackson, R. F. W. Much Improved Conditions for the Negishi Cross-Coupling of Iodoalanine Derived Zinc Reagents with Aryl Halides. *J. Org. Chem.* **2010**, *75* (1), 245–248.

(9)   Deboves, H. J. C.; Montalbetti, C. A. G. N.; Jackson, R. F. W. Direct Synthesis of Fmoc-Protected Amino Acids Using Organozinc Chemistry: Application to Polymethoxylated Phenylalanines and 4-Oxoamino Acids. *J. Chem. Soc.{,} Perkin Trans. 1* **2001**, No. 16, 1876–1884.

(10)   Jackson, R. F. W.; Wishart, N.; Wood, A.; James, K.; Wythes, M. J. Preparation of Enantiomerically Pure Protected 4-Oxo .Alpha.-Amino Acids and 3-Aryl .Alpha.-Amino Acids from Serine. *J. Org. Chem.* **1992**, *57* (12), 3397–3404.

(11)   Tabanella, S.; Valancogne, I.; F. W. Jackson, R. Preparation of Enantiomerically Pure Pyridyl Amino Acids from Serine. *Org. Biomol. Chem.* **2003**, *1* (23), 4254–4261.

(12)   Greulich, H.; Kaplan, B.; Mertins, P.; Chen, T.-H.; Tanaka, K. E.; Yun, C.-H.; Zhang, X.; Lee, S.-H.; Cho, J.; Ambrogio, L.; et al. Functional Analysis of Receptor Tyrosine Kinase Mutations in Lung Cancer Identifies Oncogenic Extracellular Domain Mutations of ERBB2. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (36), 14476–14481.

(13)   Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. Searching Fragment Spaces with Feature Trees. *J. Chem. Inf. Model.* **2009**, *49* (2), 270–279.

(14)   Čapek, P.; Pohl, R.; Hocek, M. A Facile and Efficient Synthesis of (Purin-6-Yl)Alanines. *J. Org. Chem.* **2004**, *69* (23), 7985–7988.

(15)   Ozturk, S.; Forneris, C. C.; Nguy, A. K. L.; Sorensen, E. J.; Seyedsayamdost, M. R. Modulating OxyB-Catalyzed Cross-Coupling Reactions in Vancomycin Biosynthesis by Incorporation of Diverse d-Tyr Analogues. *J. Org. Chem.* **2018**, *83* (13), 7309–7317.

(16)   Dilauro, G.; Azzollini, C. S.; Vitale, P.; Salomone, A.; Perna, F. M.; Capriati, V. Scalable Negishi Coupling between Organozinc Compounds and (Hetero)Aryl Bromides under Aerobic Conditions When Using Bulk Water or Deep Eutectic Solvents with No Additional Ligands. *Angew. Chemie Int. Ed.* **2021**, *60* (19), 10632–10636.

(17)   Barder, T. E.; Walker, S. D.; Martinelli, J. R.; Buchwald, S. L. Catalysts for Suzuki−Miyaura Coupling Processes: Scope and Studies of the Effect of Ligand Structure. *J. Am. Chem. Soc.* **2005**, *127* (13), 4685–4696.

(18)   Hudson, A. S.; Caron, L.; Colgin, N.; Cobb, S. L. A Direct Method for the Synthesis of Orthogonally Protected Furyl- and Thienyl- Amino Acids. *Amino Acids* **2015**, *47* (4), 779–785.

(19) Berthet, M.; Martinez, J.; Parrot, I. MgI2-Chemoselective Cleavage for Removal of Amino Acid Protecting Groups: A Fresh Vision for Peptide Synthesis. *Pept. Sci.* **2017**, *108* (2), e22908.

(20) Sadowski, B.; Klajn, J.; Gryko, D. T. Recent Advances in the Synthesis of Indolizines and Their π-Expanded Analogues. *Org. Biomol. Chem.* **2016**, *14* (33), 7804–7828.

(21) Kostik, E. I.; Abiko, A.; Oku, A. Chichibabin Indolizine Synthesis Revisited: Synthesis of Indolizinones by Solvolysis of 4-Alkoxycarbonyl-3-Oxotetrahydroquinolizinium Ylides. *J. Org. Chem.* **2001**, *66* (8), 2618–2623.

(22) Kakehi, A.; Ito, S.; Watanabe, K.; Kitagawa, M.; Takeuchi, S.; Hashimoto, T. Preparation of New Nitrogen-Bridged Heterocycles. Synthesis and Some Reactions of 2,3-Dihydroindolizin-2-One Derivatives. *J. Org. Chem.* **1980**, *45* (25), 5100–5104.

(23) Tominaga, Y.; Shiroshita, Y.; Kurokawa, T.; Gotou, H.; Matsuda, Y.; Hosomi, A. Synthesis of Cycl[3.2.2]Azine and Benzo[g]Cycl[3.2.2]Azine Derivatives by Use of the [2 + 8] Cycloaddition Reaction of Indolizines and Dimethyl Acetylenedicarboxylate. *J. Heterocycl. Chem.* **1989**, *26* (2), 477–487.

(24) Seregin, I. V; Gevorgyan, V. Gold-Catalyzed 1,2-Migration of Silicon, Tin, and Germanium En Route to C-2 Substituted Fused Pyrrole-Containing Heterocycles. *J. Am. Chem. Soc.* **2006**, *128* (37), 12050–12051.

(25) Flitsch, W.; Gerstmann, E. Cyclische Verbindungen Mit Heterobrückenatomen, IV. Einfache Synthesen von Indolizin Und Einigen Chinoliziniumsalzen. *Chem. Ber.* **1969**, *102* (4), 1309–1311.

(26) Zhang, L.; Liand, F.; Sun, L.; Hu, Y.; Hu, H. A Novel and Practical Synthesis of 3-Unsubstituted Indolizenes. *Synthesis (Stuttg).* **2000**, No. 12, 1733–1737.

(27) Goldman, I. M. Activation of Manganese Dioxide by Azeotropic Removal of Water. *J. Org. Chem.* **1969**, *34* (6), 1979–1981.

(28) Chen, P.; Chaikuad, A.; Bamborough, P.; Bantscheff, M.; Bountra, C.; Chung, C.; Fedorov, O.; Grandi, P.; Jung, D.; Lesniak, R.; et al. Discovery and Characterization of GSK2801, a Selective Chemical Probe for the Bromodomains BAZ2A and BAZ2B. *J. Med. Chem.* **2016**, *59* (4), 1410–1424.

(29) Quibell, J. M.; Perry, G. J. P.; Cannas, D. M.; Larrosa, I. Transition-Metal-Free Decarboxylative

Bromination of Aromatic Carboxylic Acids. *Chem. Sci.* **2018**, *9* (15), 3860–3865.

(30)   Peng, X.; Shao, X.-F.; Liu, Z.-Q. Pd(II)-Catalyzed Bromo- and Chlorodecarboxylation of Electron-Rich Arenecarboxylic Acids. *Tetrahedron Lett.* **2013**, *54* (24), 3079–3081.

(31)   Pizzorno, M. T.; Albonico, S. M. Novel Synthesis of 5,6,7,8-Tetrahydroindolizines. *J. Org. Chem.* **1977**, *42* (5), 909–910.

(32)   Michael Welch; Robert S. Phillips. Improved Syntheses of [3,2-b]- and [2,3-b]-Fused Selenolo- and Thienopyrroles, and of Furo[3,2-b]Pyrrole. *Heterocycl. Commun.* **1999**, *5* (4), 305–310.

(33)   Blake, James F.; Burgess, Laurence E.; Chicarelli, Mark Joseph; Christensen, James Gail; Cook, Adam; Fell, Jay Bradford; Fischer, John P.; Marx, Matthew Arnold; Mejia, Macedonio J.; Savechenkov, P. et al. 5,6,7,8-Tetrahydropyrido[3,4-d]Pyrimidine Derivatives as KRas G12C Inhibitors and Their Preparation. US20190144444, 2018.

(34)   Blake, James F.; Burgess, Laurence E.; Chicarelli, Mark Joseph; Christensen, James Gail; Fell, Jay Bradford; Fischer, John P.; Gaudino, John J.; Hicken, Erik James; Hinklin, Ronald Jay; Lee, M. R. et al. Preparation of Substituted Pyrido[3,4-d]Pyrimidine Compounds as KRas G12C Inhibitors. US20180072723, 2018.

(35)   Jeffery, T. Palladium-Catalysed Vinylation of Organic Halides under Solid–Liquid Phase Transfer Conditions. *J. Chem. Soc.{,} Chem. Commun.* **1984**, No. 19, 1287–1289.

(36)   Kundu, G.; Opincal, F.; Sperger, T.; Schoenebeck, F. Air-Stable PdI Dimer Enabled Remote Functionalization: Access to Fluorinated 1,1-Diaryl Alkanes with Unprecedented Speed. *Angew. Chemie Int. Ed.* **2022**, *61* (1), e202113667.

(37)   Rose, T. E.; Lawson, K. V; Harran, P. G. Large Ring-Forming Alkylations Provide Facile Access to Composite Macrocycles. *Chem. Sci.* **2015**, *6* (4), 2219–2223.

(38)   Lawson, K. V; Rose, T. E.; Harran, P. G. Template-Constrained Macrocyclic Peptides Prepared from Native, Unprotected Precursors. *Proc. Natl. Acad. Sci.* **2013**, *110* (40), E3753--E3760.

(39)   Chiarello, J.; Joullié, M. M. Synthetic Routes to Cristatic Acid and Derivatives. *Tetrahedron* **1988**, *44* (1), 41–48.

(40)   Taylor, M. S.; Ruch, T. R.; Hsiao, P.-Y.; Hwang, Y.; Zhang, P.; Dai, L.; Huang, C. R. L.; Berndsen, C. E.; Kim, M.-S.; Pandey, A.; et al. Architectural Organization of the Metabolic Regulatory Enzyme

Ghrelin O-Acyltransferase. *J. Biol. Chem.* **2013**, *288* (45), 32211–32228.

(41) Dey, Fabian; Ding, Xiao; Hu, Yimin; Liu Yongqiang; Shen, Hong, Shi, Hougang; Tan Xuefei; Zhou, C. Novel Oxoquinolizine Compounds for the Treatment and Prophylaxis of Bacterial Infection. WO 2019/228940 A1, 2019.

4. In Search of Small Molecules that Selectively Inhibit MBOAT4

4.1 Introduction

Ghrelin is a 28-residue lipopeptide (Fig. 1A) discovered by Kojima and co-workers as the endogenous ligand for the growth hormone secretagogue receptor (GHS-r).[1] It was found that the octanoylation of Ser3 was required for ghrelin's endocrine activities. At the time, the enzyme responsible for this modification was unknown. Later, Yang hypothesized this atypical lipidation was likely performed by a member of the membrane-bound O-acyl transferase (MBOAT) family of enzymes.[2] The MBOATs catalyze the lipidation of a variety of substrates including phospholipids, neutral lipids, and proteins using saturated and unsaturated acyl coenzyme-A derivatives as acyl donors.[3] In 2008, Yang et al. utilized a candidate cloning approach to demonstrate that MBOAT 4, now termed ghrelin O-acyl transferase (GOAT), the only MBOAT capable of catalyzing proghrelin octanoylation.[2] Gutierrez used a candidate gene silencing approach to independently reach the same conclusion for human GOAT.[4] Short-interfering RNAs (siRNAs) were produced and the effects of silencing individual MBOAT genes were observed. MS/MS fragmentation analyses of GOAT acylated ghrelin showed that the acylation is on Ser-3, identical to acyl ghrelin produced in the stomach.

Active ghrelin (hereafter referred to as ghrelin) was shown to induce adiposity in rodents and food intake in both rodents and humans.[5,6] Ghrelin was also found to play a role in glucose homeostasis by inhibiting insulin secretion from the pancreas.[7,8] Levels of active ghrelin in the blood peak during fasting and decrease after meals.[9,10] Ghrelin is primarily produced by peripheral tissues, rather than the central nervous system.[11] In addition, no other hormone is known to be octanoylated.[12] Given the biological specificity of GOAT and ghrelin, it was thought that, unlike GHS-r ligands, an inhibitor of GOAT would not need to penetrate the brain to cause pharmacological effects. The potential therapeutic effect of suppressing circulating ghrelin levels

fueled efforts to identify inhibitors of the enzyme. Such molecules could clarify the roles of ghrelin in feeding, weight gain and glucose homeostasis, both in normal physiology and in disease.
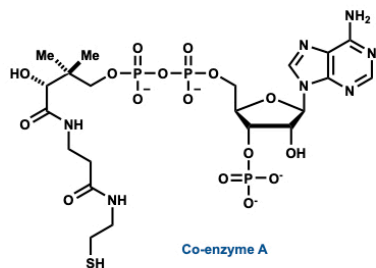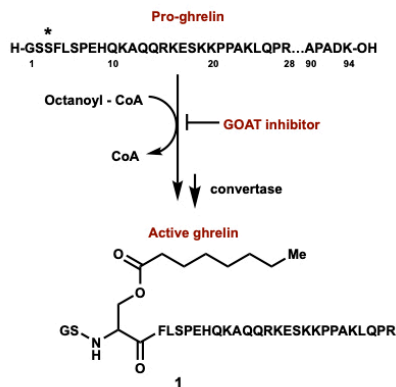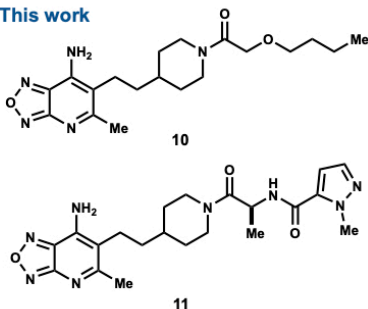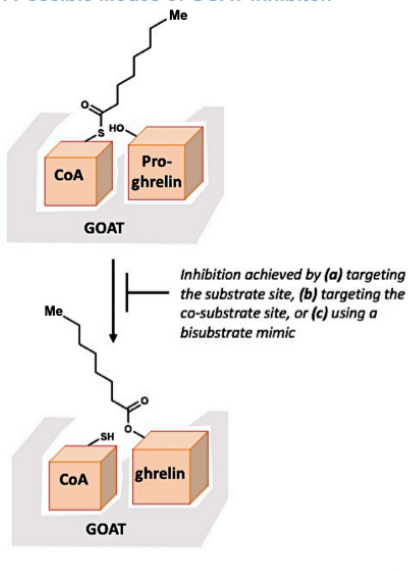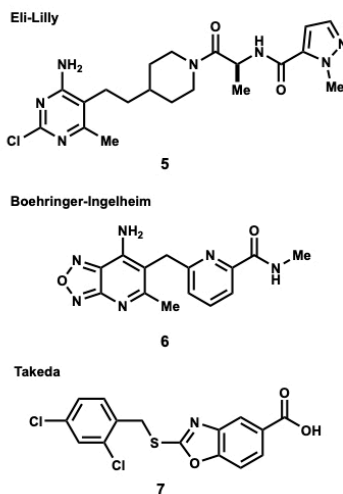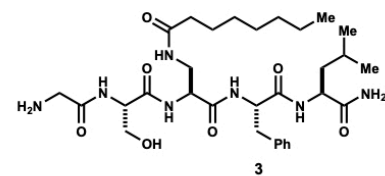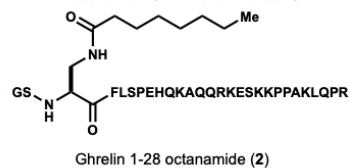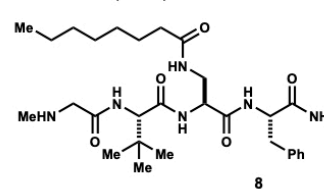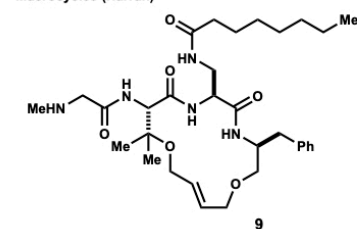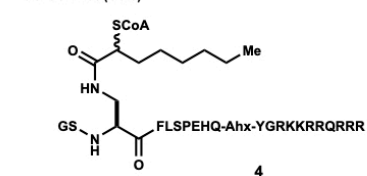
Figure 1. A. GOAT-catalyzed octanoylation of pro-ghrelin is a critical part of ghrelin maturation. GOAT is weakly inhibited by its reaction product, acyl ghrelin (**1**), and more potently by analog **2**, wherein Ser-3 is replaced by diaminopropionic acid (DAP). Systemic inhibition of GOAT is expected to reduce circulating acyl-ghrelin levels, leading to improved insulin response to glucose challenge. **B**. GOAT inhibition can be achieved by targeting either the substrate or co-substrate site, or by targeting both sites with a "bisubstrate" mimic. **C-E**. Known inhibitors of GOAT.

A general model of GOAT transmembrane architecture was proposed, which describes eleven membrane-spanning helix domains and two intramembrane domains.[13,14] Although

integral membrane proteins constitute an essential portion of the proteome, they are often difficult to purify and analyze because of their hydrophobicity and reliance on lipid–bilayer interactions for stability and activity. Since a purified form of GOAT has yet to be described, medicinal chemistry efforts lack the benefit of high-resolution structural data and detailed binding models. Crude membrane preparations from GOAT expressing cells are used as a source of enzymatic activity for inhibitor assays. Despite this challenge, several groups reported GOAT inhibitors (Fig. 1B and Fig. 1C–E).[15,17] Shortly after identifying GOAT, Yang et al. demonstrated the enzyme was inhibited by its reaction product **1**, and more potently by esterase resistant octanamide analog **2** (IC_{50} = 200 nM).[18] A truncated peptide (**3**) consisting of the first five N-terminal ghrelin residues also inhibited GOAT, albeit five-fold less potently (IC_{50} = 0.1 μM). Barnett et al. developed 'bi-substrate' mimic GO-CoA-Tat (**4**, IC_{50} = 3 mM). This inhibitor incorporated the first ten N-terminal residues of active ghrelin, which is S-linked to a coenzyme A molecule, and fused to a Tat sequence to promote endocytosis.[19] We found GO-CoA-Tat less potent than **2** in vitro, but it was reported to reduce circulating ghrelin levels in vivo, improve glucose tolerance and limit weight gain in mice fed a high-fat diet. These findings were attributed to effects on ghrelin signaling because they were not observed in ghrelin knockout mice. Eli Lilly, Boehringer Ingelheim and Takeda recently patented more drug-like, heterocyclic inhibitors of GOAT.[20-22] Takeda showed that their inhibitor **7** competed at the CoA site on GOAT, presumably by mimicking elements of an adenine base. It is likely that amino pyrimidine **5** and oxadiazolopyridine **6** function similarly.

Whereas GO-CoA-Tat exploits the ability of a polyarginine sequence to promote active cellular uptake, we sought a ghrelin peptidomimetic showing passive membrane permeability. There was precedent for this outcome; for example, the marketed drug Aliskiren was designed based on the peptide sequence of renin.[23] Nonetheless, we faced a difficult problem. Rather than elaborating a screening hit to a (typically larger) end-product, an effective GOAT inhibitor would need to be smaller, less polar, and metabolically stable relative to the parent peptide. We sought to identify and enhance key binding interactions while removing superfluous functionality. Herein,

we report peptidomimetic GOAT inhibitors, including **8** and **9**, and describe their in vitro performance. We also outline the synthesis of novel hybrid heterocyclic GOAT inhibitors based on the structures of prototypes **5** and **6**.

4.2 Results and Discussion

Building on data reported by Brown and Goldstein,[18] we began our studies by evaluating a series of modified truncation peptides ranging from 3–8 amino acids in length (Fig. 2). We discovered that mono methylating the N-terminus and replacing Ser2 with t-butyl glycine increased the in vitro potency of 5-mer 3 by 20-fold (see **13**, Fig. 2). The corresponding 4-mer **8** was even more effective. It showed an IC_{50} of 14 nM when assayed in membrane fractions. The phenyl ring in **8** was important. Analogs of **8** with either histidine or tyrosine at P4 were inactive, as was three residue peptide **12**.

Peptide	Residues	R	IC_{50} (nM)
12	3	$CONH_2$	> 1000
8	4	$F-NH_2$	14
13	5	$FL-NH_2$	50
14	6	$FLS-NH_2$	62
15	7	$FLSP-NH_2$	96
16	8	$FLSPK-NH_2$	38

9a (*E* isomer) = 173 nM
9b (*Z* isomer) = 61 nM

17a (*E* isomer) = 48 nM
17b (*Z* isomer) = 62 nM
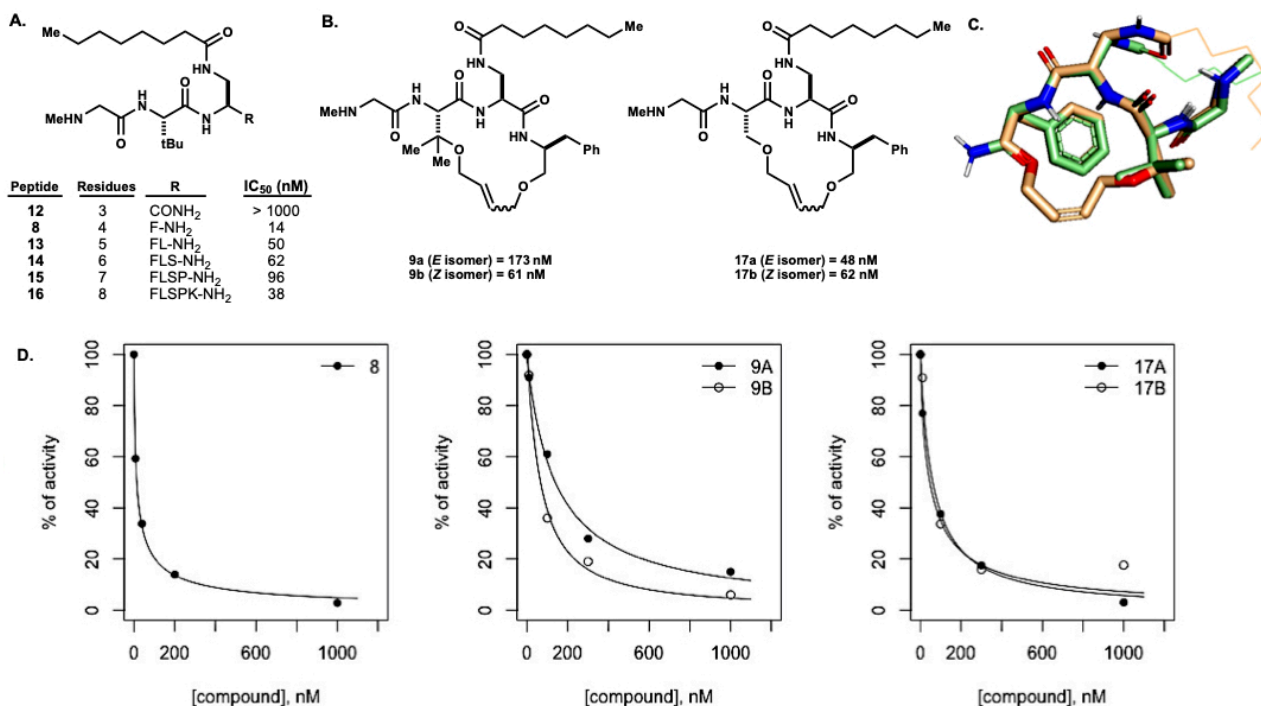
Figure 2. (A) Ghrelin truncation peptide series and respective IC_{50} values. Enzyme activity was quantified using the method described by Yang et al.[18] (see Experimental Appendix); **(B)** Macrocycles derived from the parent peptides and respective IC_{50} values; **(C)** Structural overlay of **9b** (orange) against **8** (green), octanamide chains are indicated as line bonds for clarity; **(D)** In vitro inhibition of proghrelin octanoylation mediated by GOAT using the shown compounds. The experimental data were fit to a 4-parameter logistic

101

model, (Bottom − Top)/(1 + (I/IC$_{50}$)^n) + Top [24] using DYNAFIT software.[25] The Top was defined as reaction conditions without inhibitor and the Bottom was defined as reaction conditions with full inhibition by 5 µM concentration of **2**. Top and Bottom conditions were included in every experiment.

Peptides chains longer than three amino acids are generally not taken up into cells passively. However, their cyclic counterparts can show an improved stability, pharmacokinetics and, if their structure facilitates internal shielding of the polar surface area, useful levels of cell permeability.[26] We used Schrödinger's Macromodel software to conduct molecular modeling studies, using AMBER in a continuum solvation model (octanol) to inform the design of hypothetical macrocyclic analogs of compound **8**, which was our best performing tetrapeptide. During these studies, we were pleased to find that the peptide backbone of the macrocyclic analog **9b** overlaid almost perfectly with that of **8** (Fig. 2C). The root mean square deviation (RMSD) between the two structures was 0.06 Å at the α-carbon atoms for the lowest energy conformations of both molecules. We ignored variations of the octanamide chain in these analyses as this group was expected to be highly flexible. Based on our observations, we synthesized cyclic variants **9** and **17** using catalyzed ring-closing metatheses of bis-allyl ether precursors, (Fig. 2, full details in Experimental Appendix), resulting in mixtures of separable alkene geometric isomers. The ability of these compounds to inhibit GOAT activity in vitro was comparable to their linear peptide counterparts. Moreover, enzymatic reaction rates measured while varying substrate concentrations in the presence of a fixed concentration of **17a** suggested that the compound inhibited GOAT by competing for the peptide substrate. With both linear and cyclic inhibitors in hand, we began an extensive program to synthesize and assay analogs, wherein the octanamide side chain, the P4 aromatic moiety, N-methylation patterns, ring substituents and connectivity (for **9/17**) and stereochemistry were varied in search of molecules that inhibited GOAT potently in vitro, which could block acyl ghrelin secretion from an engineered insulinoma cell line (INS-1, see SI). Over time hundreds of compounds were evaluated.[27,28] While several of those components inhibited GOAT activity in membrane fractions in the 50–200 nM range, none were superior to prototypic peptide **8** and none, including **8**, showed significant activity in cell culture. Actually, the

best performing oligomer in cell culture was lipopentapeptide **13** (Fig. 2). Unfortunately, that molecule showed poor pharmacokinetics and a low oral bioavailability in mice.

During this time, our lab had begun developing methods to scaffold peptidic macrocycles with octafluorocyclopentene through vinylic substitutions of fluorine atoms with nucleophilic side chains.[29] The judicious introduction of fluorine into a molecule can productively influence conformation, intrinsic potency, membrane permeability, metabolic pathways, and pharmacokinetic properties. We thus aimed to explore the effects of fluorination in the context of our macrocyclic inhibitors through the synthesis of **18-22** (Fig. 3), wherein the peptidic backbone would retain the active pharmacophore identified from our peptide screen, but the macrocycles would have varying ring sizes and connectivities. We expected these fluorinated compounds to exhibit very different pharmacological properties compared to the previously synthesized compounds. Unfortunately, while some of the compounds in this series did indeed appear to have much improved cellular activities, all were found to be inactive in vitro. We surmised the apparent cellular activities arose from undesired off target effects.
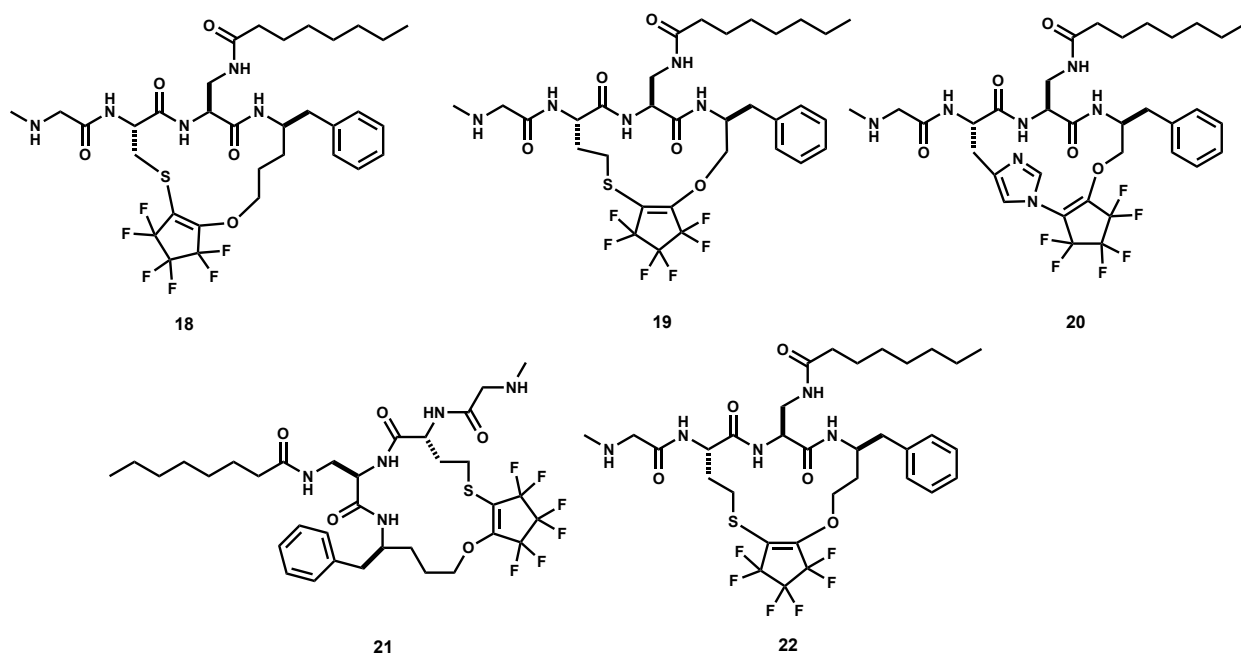
Figure 3. Macrocycles scaffolded by a fluorinated moiety (full experimental details documented in Experimental Appendix).

At this stage of the work, pharmaceutical companies began publishing patent literature on GOAT inhibitors discovered through high-throughput screening. Takeda reported data indicating a series of benzoxazole carboxylates (including **7** in Fig. 1D), which inhibited the enzyme by competing at its co-enzyme A binding site.[21] Based on their structures, we believed it likely the aminopyrimidines reported by Eli Lilly[20] and the oxadiazolopyridines described by Boehringer Ingelheim[22] functioned similarly. Because numerous metabolic enzymes utilize co-enzyme A derivatives for catalysis, there was the risk of the unanticipated off-target effects for inhibitors of this kind. Nonetheless, Eli Lilly reported that compound **5** was orally bioavailable and inhibited ghrelin production in vivo at doses that did not cause observable adverse side effects.

We sought a hybrid molecule that would contain an oxadiazolopyridine linked to an octanoyl motif through the piperidinyl ethyl scaffold used in Eli Lilly compound **5** (and also a feature of certain ACAT inhibitors).[30] We first synthesized compounds **5** and **6** in house and tested them in our own assays. We found the membrane fraction IC_{50} = 88 nM and 64 nM, as well as INS-1 cellular IC_{50} = 670 nM and 540 nM, respectively. Having roughly confirmed literature activities, we targeted new hybrids **10** and **11** wherein the oxadiazolopyridine would replace the aminopyrimidine in **5** and, in **10**, its dipeptidyl segment would be replaced by an octanoyl unit, a feature essential for activity in our previous peptidomimetic efforts.

Amino-iodinated oxadiazolopyridine **24** (Scheme 1) was synthesized from amino cyano oxadiazole **23** and ethyl acetoacetate, as described by Boehringer Ingelheim. Sonogashira coupling of **24** with 4-ethynyl piperidine **25** (prepared via Ohira–Bestmann homologation of the corresponding aldehyde) gave chromophore **26**. The attempted saturation of the alkyne in this molecule by hydrogenation over PtO_2 gave mainly *cis*-alkene **27** alongside a by-product that lacked an oxadiazole ring, presumably formed via N-O bond reduction. The attempted hydrogenation over various other heterogeneous and homogeneous catalysts gave similar results. Only when catalytic $Pd(OH)_2/C$ was used (EtOH, 1 atm $H_2(g)$), were small amounts of alkane **28** formed in a mixture that predominately contained **27**. Treatment of the crude mixture

with TFA followed by acylation with butylated glycolic acid **29** allowed pure target **10** to be isolated

(along with alkene congener **30**) following column chromatography on silica gel, albeit in low yield.
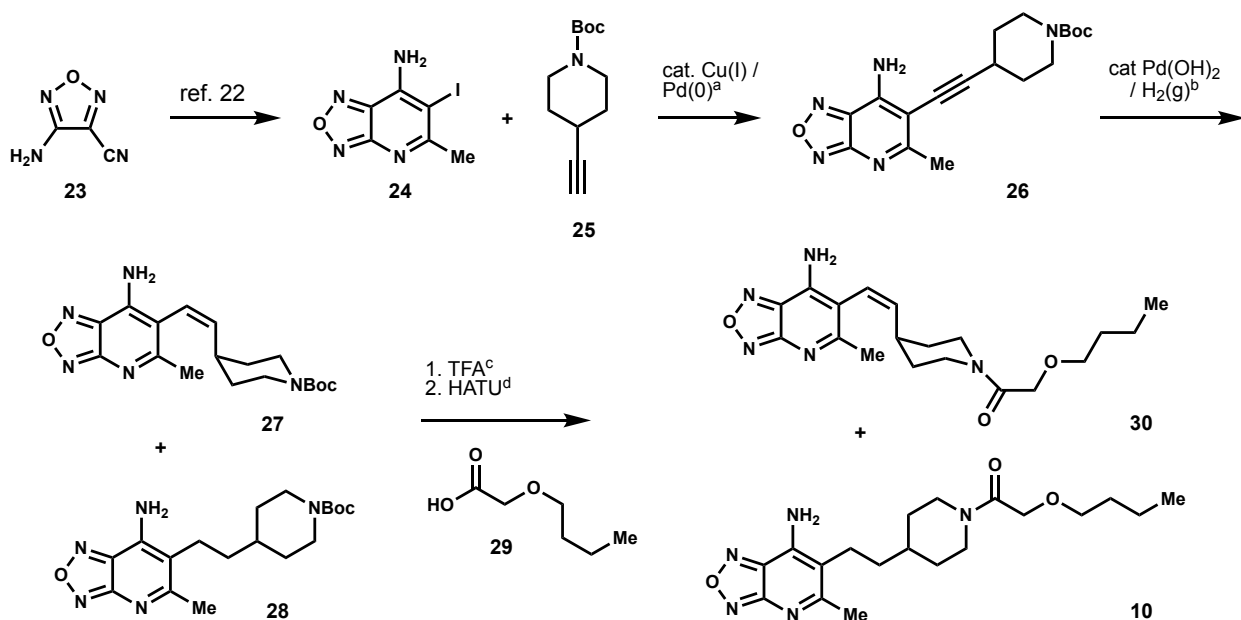
Scheme 1. (a) **24** (1 eq.), **25** (1.5 eq), CuI (0.1 eq), (PPh₃)₂PdCl₂ (0.2 eq), Et₃N (0.25M), 80 °C, 51%; (b) Pd(OH)₂ (20% on carbon) (0.3 eq.), EtOH (0.1M), H₂ atm.; (c) CH₂Cl₂ (0.04M), trifluoroacetic acid (TFA) (10% *v/v*); (d) **29** (1.1 eq), HATU (1.1 eq.), DIPEA (3 eq.), DMF (0.2 M), 32% 25 over three steps, 1% 10 over three steps.
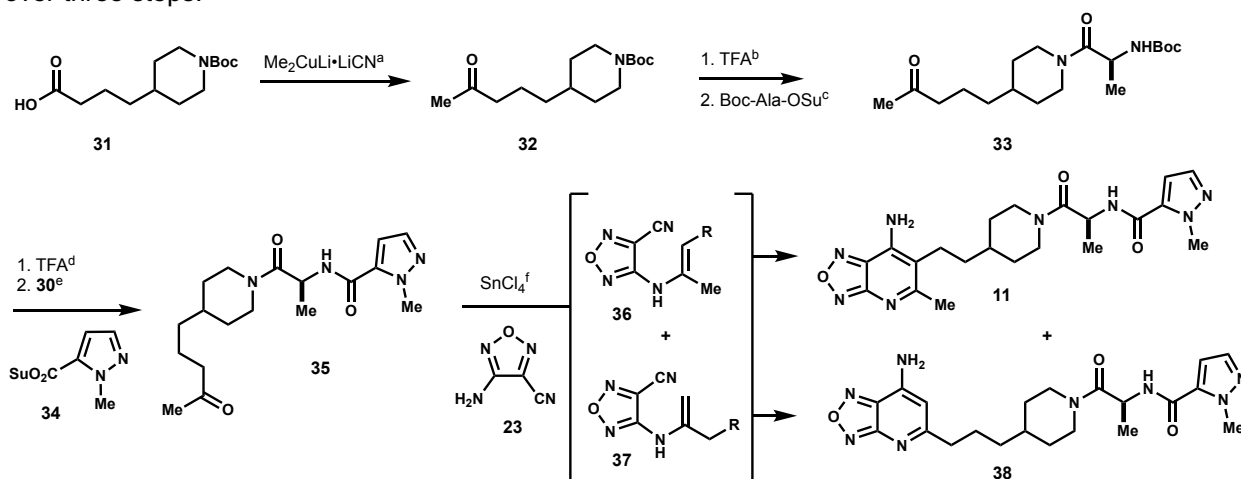
Scheme 2. (a) CuCN (5 eq.), Et₂O (0.6 M), MeLi (1.6M in Et₂O; 10 eq.), 0 °C; 27 (1 eq.), Et₂O (0.2 M), 71%; (b) CH₂Cl₂ (0.07 M), trifluoroacetic acid (TFA) (10% *v/v*), 0 °C; (c) Boc-Ala-OSu (1.1 eq.), DIPEA (3.1 eq.), MeCN (0.2 M), 0 °C, 50% over two steps; (d) CH₂Cl₂ (0.07 M), trifluoroacetic acid (TFA) (10% *v/v*), 0 °C; (e) 30 (1 eq.), DIPEA (3 eq.), MeCN (0.2 M), 0 °C, 72% over two steps; (f) **23** (1 eq.), SnCl₄ (2 eq.), toluene (0.2 M), reflux.

To avoid the problematic hydrogenation of **26**, we developed a different route to access

target **11**, a path also applicable to **10**. In this sequence, the oxadiazolopyridine unit is installed

late via annulation onto an intermediate already at the desired oxidation state. We required a protected 5-piperidinyl-2-pentanone for this purpose. Boc derivative **32** (Scheme 2) was known, although it was prepared via a 3-step sequence beginning with a relatively costly starting material.[31] We instead developed a one-step synthesis of **32** from commercial acid **31** using Me$_2$CuLi.LiCN, as described by Posner and Genna.[32,32a] Multiple grams of **32** were prepared readily using this method. Degradation of the carbamate in **32** with TFA and acylation of the incipient amino ketone salt with Boc-L-Ala-OSu produced piperidyl amide **33**. Further N-terminal extension with N-Me pyrazole derivative **34** afforded compound **35**. Structure **35** contains the dipeptide segment of Eli Lilly compound **5** tethered to a methyl ketone handle from which varied heterocycles can derive. To prepare target oxadiazolopyridine **11**, we condensed **35** with cyanofurazan **23**[33] using SnCl$_4$ as promoter. This construction was originally developed by Vasil'ev et al.[34] and, in the current example, affords oxadiazolopyridine **11**, directly following aqueous workup. A second regioisomer (**38**) was also isolated from the reaction (dr = 1.5:1), a result we interpret in terms of the competing formation of regioisomeric enamine intermediates **36/37** that cyclize onto the pendant nitrile. Subsequent tautomerization would afford oxadiazolopyridines.

Synthetic compounds **10**, **11**, alkene **30** and regiosiomer **38** were tested in membrane fractions containing GOAT and in ghrelin-secreting INS-1 cells. As shown in Fig. 4, both **30** and **38** were significantly impaired, likely a result of their altered geometries. However, both **10** and **11** inhibited membrane GOAT as potently as did **5** and **6.** Notably, compound **11** blocked ghrelin secretion from INS-1 cells 4-fold more potently than **5**. To date, it is the most potent compound we have tested in our cellular assay. Because **10** is much less active in cells, yet it inhibits GOAT in membrane fractions just as well as **11**, we speculate that the adenine mimcry provided by the oxadiazolopyridine is key, with a linked functionality contributing mainly to stability and cellular uptake, or lack thereof.
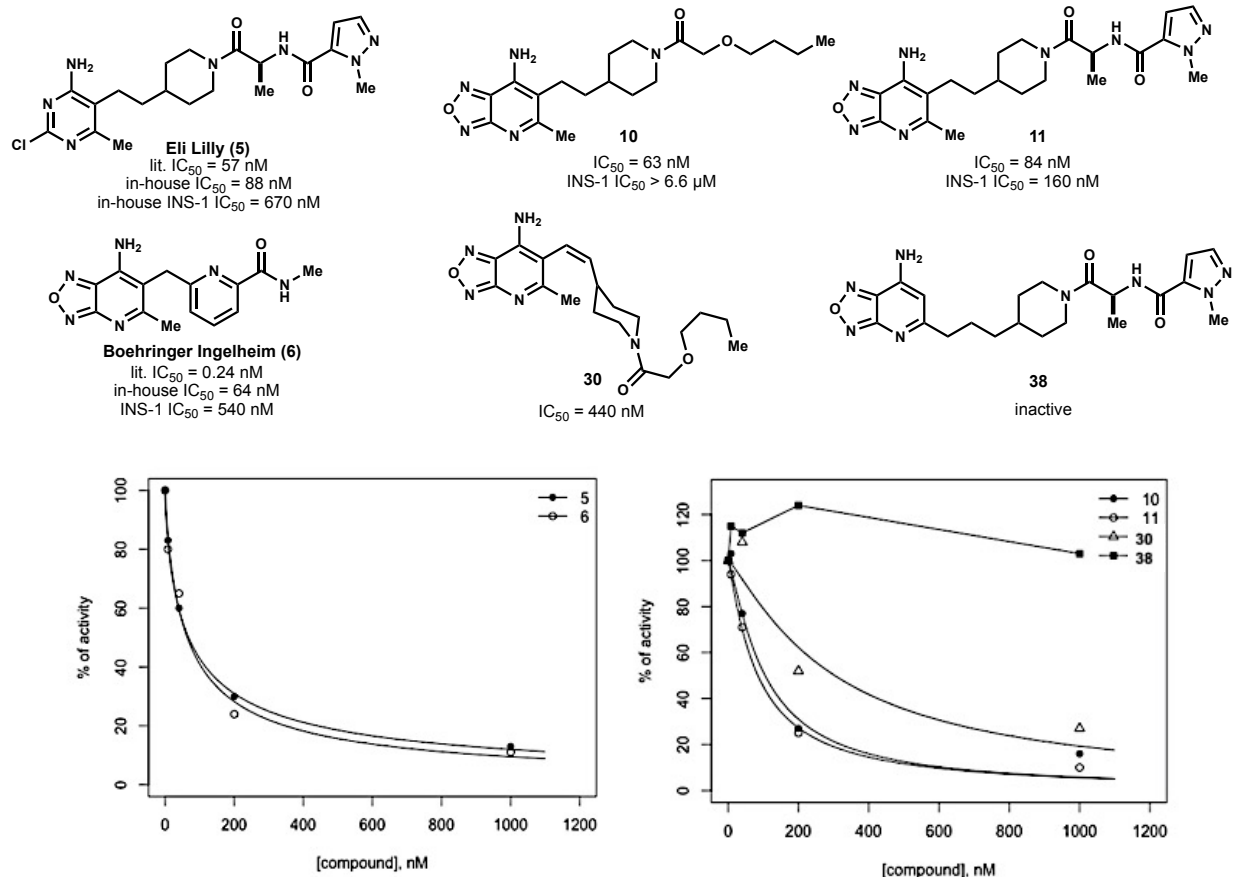
Figure 4. *(top)* In vitro and cellular activities of heterocyclic analogues (see Experimental Appendix for assay descriptions); *(bottom)* In vitro inhibition of proghrelin octanoylation mediated by GOAT using the shown compounds. The experimental data was fit to a 4-parameter logistic model, (Bottom − Top)/(1 + $(I/IC_{50})^n$) + Top[24] using DYNAFIT software[25]. The Top was defined as reaction conditions without inhibitor and the Bottom was defined as reaction conditions with full inhibition by 5 µM concentration of **2**. Top and Bottom conditions were included in every experiment.

4.3 Conclusions

Ghrelin signaling continues to be an active area of basic biology research and drug discovery. The discovery of GOAT and its role in maturing ghrelin in the stomach provided an opportunity for pharmacological intervention without accessing the central nervous system. These efforts were largely empirical due to a lack of structural data for the system. However, we have learned a great deal. Peptidomimetic inhibitors that compete for substrate binding can perform well in vitro but, thus far, their utility in cell culture and in animals has been limited. The high-throughput screens developed in the private identified compounds uniformly appear to complete

for binding at the co-enzyme site. Several of these molecules advanced through pre-clinical developments and into human trials as therapy for diabetes type II, Prader-Willi syndrome and alcohol use disorder. We developed a six-step synthesis of heterocyclic GOAT inhibitors based on a hybrid design that utilizes a piperidinyl ethyl linker and a late-stage oxadiazolopyridine, forming annulation. The potent activity of novel product **11** both in vitro and in cells indicates this approach has the potential to identify new GOAT inhibitors with increasingly refined properties.

4.4 References

(1)   Kojima, M.; Hosoda, H.; Date, Y.; Nakazato, M.; Matsuo, H.; Kangawa, K. Ghrelin Is a Growth-Hormone-Releasing Acylated Peptide from Stomach. *Nature* **1999**, *402*, 656–660.

(2)   Yang, J.; Brown, M. S.; Liang, G.; Grishin, N. V; Goldstein, J. L. Identification of the Acyltransferase That Octanoylates Ghrelin, an Appetite-Stimulating Peptide Hormone. *Cell* **2008**, *132*, 387–396.

(3)   Masumoto, N.; Lanyon-Hogg, T.; Rodgers, U. R.; Konitsiotis, A. D.; Magee, A. I.; Tate, E. W. Membrane Bound O-Acyltransferases and Their Inhibitors. *Biochem. Soc. Trans.* **2015**, *43*, 246–252.

(4)   Gutierrez, J. A.; Solenberg, P. J.; Perkins, D. R.; Willency, J. A.; Knierman, M. D.; Jin, Z.; Witcher, D. R.; Luo, S.; Onyia, J. E.; Hale, J. E. Ghrelin Octanoylation Mediated by an Orphan Lipid Transferase. *Proc. Natl. Acad. Sci.* **2008**, *105*, 6320–6325.

(5)   Wortley, K. E.; del Rincon, J.-P.; Murray, J. D.; Garcia, K.; Iida, K.; Thorner, M. O.; Sleeman, M. W. Absence of Ghrelin Protects against Early-Onset Obesity. *J. Clin. Invest.* **2005**, *115*, 3573–3578.

(6)   Wren, A. M.; Small, C. J.; Abbott, C. R.; Dhillo, W. S.; Seal, L. J.; Cohen, M. A.; Batterham, R. L.; Taheri, S.; Stanley, S. A.; Ghatei, M. A.; et al. Ghrelin Causes Hyperphagia and Obesity in Rats. *Diabetes* **2001**, *50*, 2540–2547.

(7)   Dezaki, K.; Sone, H.; Koizumi, M.; Nakata, M.; Kakei, M.; Nagai, H.; Hosoda, H.; Kangawa, K.; Yada, T. Blockade of Pancreatic Islet-Derived Ghrelin Enhances Insulin Secretion to Prevent High-Fat Diet-Induced Glucose Intolerance. *Diabetes* **2006**, *55*, 3486–3493.

(8)   Kurashina, T.; Dezaki, K.; Yoshida, M.; Sukma Rita, R.; Ito, K.; Taguchi, M.; Miura, R.; Tominaga, M.; Ishibashi, S.; Kakei, M.; et al. The Beta-Cell GHSR and Downstream CAMP/TRPM2 Signaling Account for Insulinostatic and Glycemic Effects of Ghrelin. *Sci. Rep.* **2015**, *5*, 14041.

(9)   Cummings, D. E.; Purnell, J. Q.; Frayo, R. S.; Schmidova, K.; Wisse, B. E.; Weigle, D. S. A Preprandial Rise in Plasma Ghrelin Levels Suggests a Role in Meal Initiation in Humans. *Diabetes* **2001**, *50*, 1714–1719.

(10)   Tschöp, M.; Wawarta, R.; Riepl, R. L.; Friedrich, S.; Bidlingmaier, M.; Landgraf, R.; Folwaczny, C. Post-Prandial Decrease of Circulating Human Ghrelin Levels. *J. Endocrinol. Invest.* **2001**, *24*, RC19-21.

(11)   Stengel, A.; Goebel, M.; Wang, L.; Taché, Y.; Sachs, G.; Lambrecht, N. W. G. Differential Distribution of Ghrelin-O-Acyltransferase (GOAT) Immunoreactive Cells in the Mouse and Rat Gastric Oxyntic Mucosa. *Biochem. Biophys. Res. Commun.* **2010**, *392*, 67–71.

(12)   Romero, A.; Kirchner, H.; Heppner, K.; Pfluger, P. T.; Tschop, M. H.; Nogueiras, R. GOAT: The Master Switch for the Ghrelin System? *Eur. J. Endocrinol.* **2010**, *163*, 1–8.

(13)   Taylor, M. S.; Ruch, T. R.; Hsiao, P.-Y.; Hwang, Y.; Zhang, P.; Dai, L.; Huang, C. R. L.; Berndsen, C. E.; Kim, M.-S.; Pandey, A.; et al. Architectural Organization of the Metabolic Regulatory Enzyme Ghrelin O-Acyltransferase. *J. Biol. Chem.* **2013**, *288*, 32211–32228.

(14)   Campaña, M. B.; Irudayanathan, F. J.; Davis, T. R.; McGovern-Gooch, K. R.; Loftus, R.; Ashkar, M.; Escoffery, N.; Navarro, M.; Sieburg, M. A.; Nangia, S.; et al. The Ghrelin O Acyltransferase Structure Reveals a Catalytic Channel for Transmembrane Hormone Acylation. *J. Biol. Chem.* **2019**, *294*, 14166–14174.

(15)   Iyer, M. R.; Wood, C. M.; Kunos, G. Recent Progress in the Discovery of Ghrelin O-Acyltransferase (GOAT) Inhibitors. *RSC Med. Chem.* **2020**, *11*, 1136–1144.

(16)   Garner, A. L.; Janda, K. D. A Small Molecule Antagonist of Ghrelin O-Acyltransferase (GOAT). *Chem. Commun.* **2011**, *47*, 7512–7514.

(17)   Darling, J. E.; Zhao, F.; Loftus, R. J.; Patton, L. M.; Gibbs, R. A.; Hougland, J. L. Structure–Activity Analysis of Human Ghrelin O-Acyltransferase Reveals Chemical Determinants of Ghrelin Selectivity and Acyl Group Recognition. *Biochemistry* **2015**, *54*, 1100–1110.

(18)   Yang, J.; Zhao, T.-J.; Goldstein, J. L.; Brown, M. S. Inhibition of Ghrelin O-Acyltransferase (GOAT) by Octanoylated Pentapeptides. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 10750–10755.

(19)   Barnett, B. P.; Hwang, Y.; Taylor, M. S.; Kirchner, H.; Pfluger, P. T.; Bernard, V.; Lin, Y.; Bowers, E. M.; Mukherjee, C.; Song, W.-J.; et al. Glucose and Weight Control in Mice with a Designed Ghrelin O-Acyltransferase Inhibitor. *Science (80-.).* **2010**, *330*, 1689–1692.

(20)   Galka, C. S.; Hembre, E. J.; Honigschmidt, N. A.; Keding, S. J.; Martinez-Grau, M. A.; Plaza, G. R.; Rubio, A.; Smith, D. L. Ghrelin O-Acyl Transferase Inhibitors. WO 2016/168225 A1, April 2016.

(21)   Yoneyama-Hirozane, M.; Deguchi, K.; Hirakawa, T.; Ishii, T.; Odani, T.; Matsui, J.; Nakano, Y.; Imahashi, K.; Takakura, N.; Chisaki, I.; et al. Identification and Characterization of a New Series of Ghrelin O-Acyl Transferase Inhibitors. *SLAS Discov. Adv. Sci. Drug Discov.* **2017**, *23*, 154–163.

(22)   Godbout, C.; Trieselmann, T.; Vintonyak, V. Oxadiazolopyridine Derivatives for Use as Ghrelin O-Acyl Transferase (GOAT) Inhibitors. US 2018/0037594 A1, February 2018.

(23)   Jensen, C.; Herold, P.; Brunner, H. R. Aliskiren: The First Renin Inhibitor for Clinical Treatment. *Nat. Rev. Drug Discov.* **2008**, *7*, 399–410.

(24)   Sebaugh, J. L. Guidelines for Accurate EC50/IC50 Estimation. *Pharm. Stat.* **2011**, *10*, 128–134.

(25)   Kuzmic, P. Program DYNAFIT for the Analysis of Enzyme Kinetic Data: Application to HIV Proteinase. *Anal. Biochem.* **1996**, *237*, 260–273.

(26)   Yudin, A. K. Macrocycles: Lessons from the Distant Past{,} Recent Developments{,} and Future Directions. *Chem. Sci.* **2015**, *6*, 30–49.

(27)   Murzinski, E. Development of Potent, Small Molecule Inhibitors of Ghrelin O-Acyltransferase (GOAT). 2020. Available online: https://escholarship.org/uc/item/9fq6b1r5. (accessed on 12/14/2021)

(28)   Hollibaugh, R. Defining a Minimal Pharmacophore to Selectively Inhibit MBOAT4 (Ghrelin O-Acyl Transferase), 2016.

(29)   Tsunemi, T.; Bernardino, S. J.; Mendoza, A.; Jones, C. G.; Harran, P. G. Syntheses of Atypically Fluorinated Peptidyl Macrocycles through Sequential Vinylic Substitutions. *Angew. Chemie Int. Ed.* **2020**, *59* (2), 674–678.

(30)   He, S.; Hong, Q.; Lai, Z.; Yang, D. X.; Ting, P. C.; Kuethe, J. T.; Cernak, T. A.; Dykstra, K. D.; Sperbeck, D. M.; Wu, Z.; et al. Discovery of a Potent and Selective DGAT1 Inhibitor with a Piperidinyl-Oxy-Cyclohexanecarboxylic Acid Moiety. *ACS Med. Chem. Lett.* **2014**, *5*, 1082–1087.

(31)   Bennett, D. J.; Brnardic, E. J.; Han, Y.; Huang, C.; Liverton, N. J.; Meng, Z.; Rudd, M. T.; Stachel, S. J.; Tempest, P.; Wai, J.; et al. Piperidine Derivatives as Liver X Receptor B Agonists, Compositions, and Their Use. U.S. Patent 10,894,775, 19 Jan. 2021.

(32)   Genna, D. T.; Posner, G. H. Cyanocuprates Convert Carboxylic Acids Directly into Ketones. *Org. Lett.* **2011**, *13*, 5358–5361.

(a)   **27** is commercially available for $20/gram from Combi-Blocks.

(33)   Pagoria, P. F.; Zhang, M.-X.; Zuckerman, N. B.; DeHope, A. J.; Parrish, D. A. Synthesis and Characterization of Multicyclic Oxadiazoles and 1-Hydroxytetrazoles as Energetic Materials. *Chem. Heterocycl. Compd.* **2017**, *53*, 760–778.

(34)   Vasil"ev, L. S.; Sheremetev, A. B.; Khoa, N. K.; Dem"yanets, Z. K.; Dmitriev, D. E.; Dorokhov, V. A. Reactions of Cyanofurazans with β-Dicarbonyl Compounds. *Russ. Chem. Bull.* **2001**, *50*, 1280–1286.

Experimental Appendices

Chapter 1 Experimental Appendix

Comparison of ConfBuster++ to literature methods. We compared conformers generated by ConfBuster++ the macrocycles cyclo-(Pro-Ser-leu-Asp-Val) and cyclo-(Arg-Gly-Asp-phe-([N-Me]Val)) (also known as cilengitide) and compared the average inter-proton distances from the resulting conformers to experimental NOE NMR data,[35-37] as well as those generated by the molecular dynamics approach presented by Kamenik *et al.* (Fig. S4).[38] Following the approach used by Kamenik *et al.,* we further made a distinction between *cis-* and *trans-* states of cyclo-(Pro-Ser-leu-Asp-Val), where the *cis-* state is characterized by the Val-Pro peptide bond having a dihedral angle of $\leq 90°$ and *trans-* otherwise. However, we note that the conformers generated by ConfBuster++ were generated in vacuum without any ionized residues, while the those generated by Kamenik *et al.* were done in water with ionized residues. This is on account of the fact that ConfBuster++ makes use of OpenBabel's genetic algorithm, which at the time of writing, does not allow the dielectric constant to be changed. All conformers produced by ConfBuster++ for a given structure are within 5 kcal mol^{-1} of the lowest energy conformer and have greater than 0.5 Å root mean square deviation (RMSD) from each other.

For the cyclo-(Pro-Ser-leu-Asp-Val) conformers generated by ConfBuster++, the *trans-* to *cis-* ratio is 1/2, whereas Kamenik *et al.* achieved a ratio of 1/3, which is closer to the experimentally determined ratio of 1/4. Fig. S4 shows the average inter-proton distances for *cis-*cyclo-(Pro-Ser-leu-Asp-Val) and *trans-*cyclo-(Pro-Ser-leu-Asp-Val) conformers (panels A and B respectively) generated by ConfBuster++ in comparison to those generated by Kamenik *et al.* and the experimental NOE data. For the *cis-* conformers, we see that two average inter-proton distances generated by ConfBuster++ are outside the bounds determined by the NOE data, which is not unreasonable considering these two distances involve the beta hydrogens on the aspartate residue, which would be most affected by the difference in the sidechain carboxyl ionization state and dielectric. In comparison, one inter-proton distance generated by Kamenik *et al.* is outside the NOE bounds. For the *trans-* conformers, we see that neither ConfBuster++ or Kamenik *et al.*'s method produced any conformers with inter-proton distances outside the predicted boundaries of the NOE data. Across all proton pairs, the inter-atomic distances for *cis-* conformers generated by ConfBuster++ deviated from the NOE data by an average 0.5 Å, whereas Kamenik *et al.*'s deviated by 0.4 Å . For *trans-* conformers, ConfBuster++ deviated by an average of 0.6 Å and Kamenik *et al.*'s method by 0.3 Å.A similar comparison between ConfBuster++ and Kamenik *et al.*'s method on cilengitide is summarized in panel C of Fig. S4. Experimental NOE data for cilengitide only specified an upper bound on the inter-proton distances. Four out of the nine average inter-proton distances of the cilengitide conformers produced by ConfBuster++ exceeded this upper bound, however two of these four only exceed the upper bound by 0.1 Å or less. In comparison, Kamenik *et al.*'s method produced conformers with an average inter-proton distance that exceeded the upper bound only once. Due to the inconsistency in solvent environment with the method discussed above, the ability of Confbuster++ to generate low energy macrocycle conformers was additionally assessed in comparison to conformers generated by CREST,[39] as discussed in Results and Discussion.

Finally, since ConfBuster++ is open-source, we have the freedom to extend and modify it as more resources may become available to us. For instance, ConfBuster++ currently only runs on shared memory nodes, however, it can be extended to operate in a distributed environment as well, which would help increase its scalability. Other changes that could be made to ConfBuster++ to increase its scalability are random sampling of cleavable bonds rather than full

enumeration, and implementing tree search algorithms for the rotate_dihedrals() step (see Methods, Fig.9).

Reactivity of thiazolopyridin-5-one in EAS was confirmed using DFT calculations (ωB97X-D-SMD(methanol)/6-31G(d)) in Gaussian16 RevA.03. Activation free energies were calculated for the reaction between thiazolopyridin-5-one and allyl cation **1** and compared to the reaction barrier of phenol with **1** (Figure S5). Entropies were corrected using quasi-harmonic correction by setting all frequencies below 100 cm^{-1} to 100 cm^{-1}. Stationary points were confirmed by either having no or exactly one (in case of transition states) imaginary frequency. Table S1 shows calculated energies for all relevant structures followed by cartesian coordinates for all relevant structures.
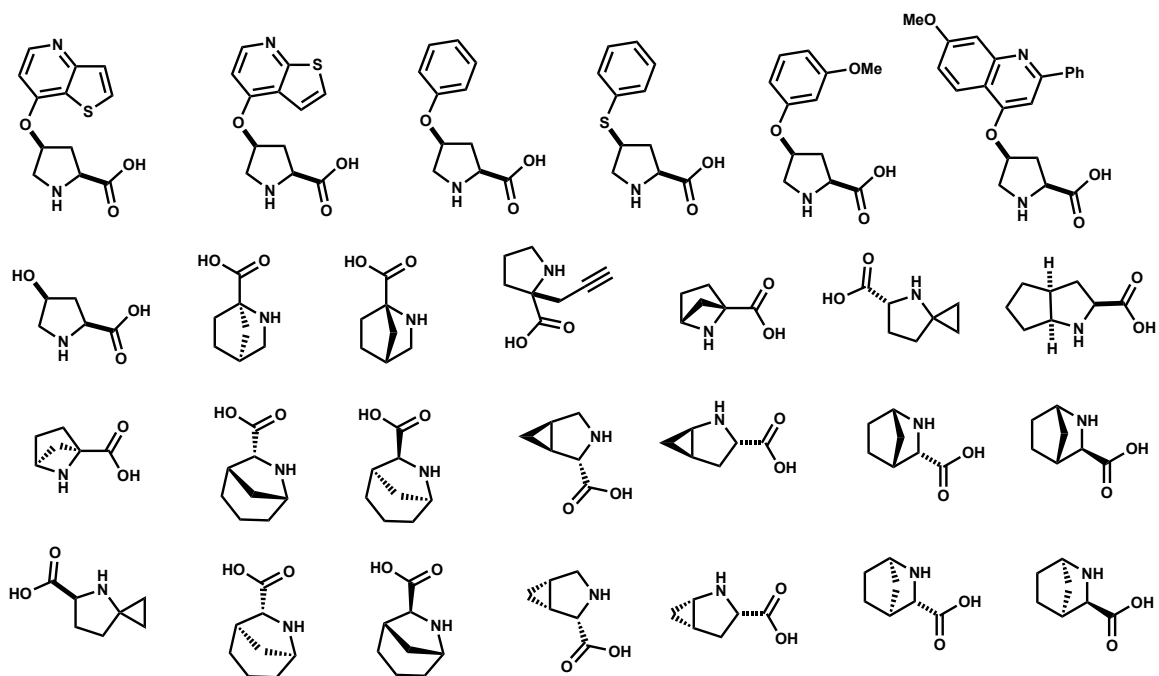
Figure S1. Full list of proline analogs.
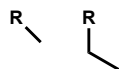
Figure S2. Connection moieties used in CPMG where R is from the set of heterocycles/prolines in Ch.1 Fig. 3 and Fig. S1. (a) The methyl connection. b) The ethyl connection.
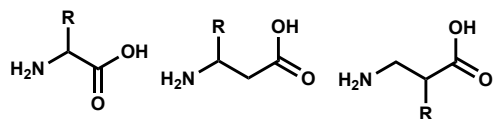
Figure S3. Amino acid backbones used in CPMG for monomer generation where R is from the set of heterocycles/prolines in Fig. 3 and Fig. S1 (left to right) α amino acid backbone, β^3 amino acid backbone, β^2 amino acid backbone.

A.

Cis cyclo-(Pro-Ser-leu-Asp-Val)

Index	Atom1	Atom2
1	Ser HN	Ser Hb
2	Ser HN	Ser Hb
3	Ser HN	Val Ha
4	Ser HN	Pro Ha
5	Leu HN	Ser Ha
6	Leu HN	Leu Ha
7	Asp HN	Leu Ha
8	Asp HN	Asp Ha
9	Asp HN	Asp Hb
10	Asp HN	Asp Hb
11	Val HN	Asp HN
12	Val HN	Asp Ha
13	Val HN	Asp Hb
14	Val HN	Asp Hb
15	Val HN	Val Ha
16	Val HN	Val Hb
17	Val Ha	Val Hb
18	Val Ha	Pro Ha

Trans cyclo-(Pro-Ser-leu-Asp-Val)

Index	Atom1	Atom2
1	Ser HN	Ser Ha
2	Leu HN	Ser Ha
3	Asp HN	Leu Ha
4	Asp HN	Asp Ha
5	Asp HN	Asp Hb
6	Asp HN	Asp Hb
7	Val HN	Asp Ha
8	Val HN	Asp Hb
9	Val HN	Asp Hb
10	Val HN	Val Hb

Cilengitide

Index	Atom1	Atom2
1	Arg HN	Arg Ha
2	Arg HN	Gly HN
3	Arg HN	Val Ha
4	Gly HN	Arg Ha
5	Asp HN	Asp Ha
6	Asp HN	Phe HN
7	Phe HN	Asp Ha
8	Phe HN	Phe Ha
9	Phe HN	Val Ha

Figure. S4. Comparison between inter-proton distances generated by ConfBuster++ and the molecular dynamics method presented by Kamenik et al. Upper and lower bounds of NOE inter-proton distances are depicted by the shaded area. Tables list the indices given to each proton pair. A. Data for cis-cyclo-(Pro-Ser-leu-Asp-Val) B. Data for trans-cyclo-(Pro-Ser-leu-Asp-Val). C. Data for cilengitide. Lowest energy conformations for each macrocycle displayed on the right of each panel.
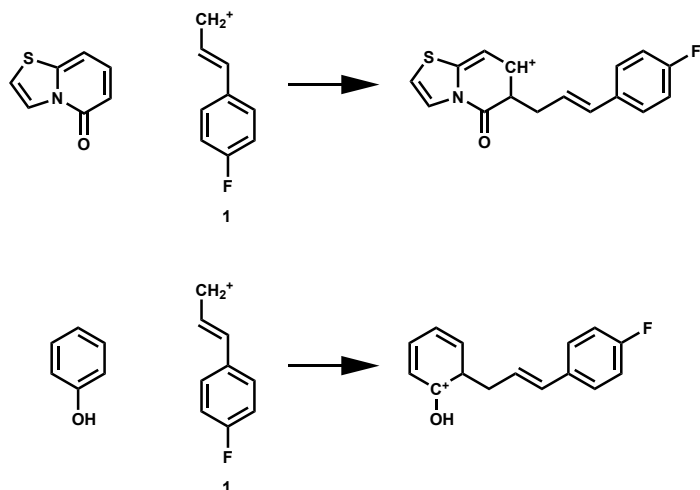
Figure S5. Investigated reactions.

Table S1. Calculated energies.

	E (hartree)	H (hartree)	G (hartree)
Thiazolopyridin-5-one	-797.772853	-797.656245	-797.696617
Phenol	-307.367587	-307.256540	-307.291349
1	-447.254128	-447.099572	-447.142266
Transition state: thiazolopyridin-5-one + X	-1245.044820	-1244.772772	-1244.832163
Transition state: phenol + X	-754.637120	-754.368674	-754.423717

Cartesian Coordinates
thiazolopyridin-5-one
```
C    1.772188000000   -1.656118000000   -0.000088000000
C    2.323742000000   -0.396441000000    0.000096000000
C    1.518799000000    0.779346000000    0.000009000000
C   -0.412849000000   -0.745370000000    0.000082000000
O    1.908469000000    1.956108000000   -0.000210000000
H    3.397781000000   -0.249592000000    0.000182000000
H    2.428844000000   -2.521418000000    0.000277000000
C    0.381469000000   -1.864432000000   -0.000095000000
H   -0.051926000000   -2.856614000000   -0.000057000000
N    0.127766000000    0.519008000000    0.000173000000
C   -2.083032000000    1.094430000000   -0.000042000000
H   -2.988689000000    1.684961000000   -0.000203000000
C   -0.823957000000    1.544627000000    0.000172000000
H   -0.480857000000    2.568144000000    0.000166000000
S   -2.157839000000   -0.652728000000   -0.000043000000
```

phenol

C	0.931589000000	0.000039000000	-0.017430000000
C	0.240283000000	1.209809000000	-0.011437000000
C	-1.152849000000	1.205652000000	0.005341000000
C	-1.852413000000	-0.000061000000	0.014283000000
C	-1.152777000000	-1.205698000000	0.005295000000
C	0.240385000000	-1.209756000000	-0.011478000000
H	0.798865000000	2.141275000000	-0.022317000000
H	-1.692445000000	2.148687000000	0.009843000000
H	-2.938507000000	-0.000079000000	0.025489000000
H	-1.692260000000	-2.148798000000	0.009766000000
H	0.798979000000	-2.141214000000	-0.022402000000
O	2.315250000000	0.000158000000	-0.089487000000
H	2.678066000000	-0.001041000000	0.808073000000

1

C	-4.190349000000	-0.053083000000	0.000258000000
H	-5.020371000000	-0.753510000000	0.000351000000
H	-4.430899000000	1.007515000000	0.000491000000
C	-2.909221000000	-0.483539000000	-0.000058000000
H	-2.686644000000	-1.545216000000	-0.000210000000
C	-1.867993000000	0.479695000000	-0.000085000000
H	-2.179678000000	1.523331000000	-0.000042000000
C	-0.486507000000	0.240736000000	-0.000183000000
C	0.071108000000	-1.069055000000	-0.000175000000
C	0.383692000000	1.365917000000	-0.000138000000
C	1.434207000000	-1.241843000000	-0.000037000000
H	-0.572434000000	-1.941318000000	-0.000273000000
C	1.749769000000	1.199567000000	0.000003000000
H	-0.042405000000	2.364464000000	-0.000205000000
C	2.245469000000	-0.103708000000	0.000081000000
H	1.889062000000	-2.225602000000	-0.000020000000
H	2.434473000000	2.039723000000	0.000058000000
F	3.558650000000	-0.275279000000	0.000206000000

Transition state: thiazolopyridin-5-one + **1**

C	-2.550300000000	-1.138990000000	1.590692000000
C	-3.354340000000	-0.874625000000	0.485797000000
C	-3.149736000000	0.287084000000	-0.339237000000
C	-1.128562000000	0.669172000000	1.000686000000
O	-3.854327000000	0.672078000000	-1.266806000000
H	-4.278887000000	-1.416403000000	0.323349000000
C	-2.088583000000	-2.253593000000	-1.063271000000
H	-2.894501000000	-2.121878000000	-1.778769000000
H	-2.183385000000	-3.075730000000	-0.360622000000
C	-0.876463000000	-1.611929000000	-1.239047000000
H	-0.790459000000	-0.866392000000	-2.023946000000
C	0.159269000000	-1.841353000000	-0.346630000000
H	-0.007523000000	-2.599122000000	0.416815000000
C	1.449731000000	-1.210884000000	-0.334857000000
C	2.344529000000	-1.546928000000	0.701720000000
C	1.842645000000	-0.247388000000	-1.288542000000
C	3.591999000000	-0.954333000000	0.787483000000
H	2.047014000000	-2.280829000000	1.445119000000
C	3.091058000000	0.343716000000	-1.219551000000

H	1.174992000000	0.039912000000	-2.093535000000
C	3.940827000000	-0.021038000000	-0.180909000000
H	4.289220000000	-1.201788000000	1.580058000000
H	3.412090000000	1.080690000000	-1.947178000000
F	5.144528000000	0.554688000000	-0.110623000000
H	-2.799456000000	-1.977310000000	2.233163000000
C	-1.393812000000	-0.414036000000	1.836475000000
H	-0.733757000000	-0.638808000000	2.664753000000
N	-1.962959000000	1.007439000000	-0.025635000000
C	-0.388789000000	2.628321000000	-0.309876000000
H	0.151028000000	3.478205000000	-0.703840000000
C	-1.535609000000	2.109361000000	-0.767899000000
H	-2.132653000000	2.437270000000	-1.605587000000
S	0.206104000000	1.761622000000	1.080095000000

Transition state: phenol + **1**

C	-2.677164000000	-1.577366000000	0.725654000000
C	-3.419321000000	-0.602776000000	0.023537000000
C	-3.215597000000	0.771237000000	0.319851000000
C	-2.156058000000	1.154716000000	1.143696000000
C	-1.351856000000	0.173329000000	1.700452000000
H	-1.961255000000	2.207623000000	1.324075000000
H	-4.313624000000	-0.881985000000	-0.525271000000
C	-2.146421000000	-0.545565000000	-1.788731000000
H	-2.892471000000	0.057054000000	-2.298116000000
H	-2.236659000000	-1.620287000000	-1.912048000000
C	-0.896336000000	0.013782000000	-1.482629000000
H	-0.781389000000	1.090354000000	-1.558086000000
C	0.115408000000	-0.781139000000	-1.011778000000
H	-0.082379000000	-1.850052000000	-0.944340000000
C	1.446687000000	-0.367029000000	-0.619585000000
C	2.359169000000	-1.357505000000	-0.216634000000
C	1.855641000000	0.979854000000	-0.608059000000
C	3.645517000000	-1.026692000000	0.181333000000
H	2.050868000000	-2.399037000000	-0.216498000000
C	3.136545000000	1.324977000000	-0.213206000000
H	1.170342000000	1.767258000000	-0.903569000000
C	4.007742000000	0.312414000000	0.173637000000
H	4.358261000000	-1.781615000000	0.494367000000
H	3.467407000000	2.357642000000	-0.197592000000
F	5.246674000000	0.646671000000	0.557311000000
H	-2.906204000000	-2.625630000000	0.562934000000
C	-1.619125000000	-1.198891000000	1.519936000000
H	-0.999153000000	-1.939054000000	2.013948000000
H	-0.507876000000	0.473516000000	2.314603000000
O	-4.039153000000	1.641241000000	-0.273531000000
H	-3.781688000000	2.554168000000	-0.066810000000

Chapter 2 Experimental Appendix

Parameter Reoptimization in ConfBuster++

Table S1. Parameter optimizations in CB++ using FK506 ligand as a model system, as described in section 2.4 of Ch.2. Table abbreviations are as follow – t, runtime; -r, number of times each linearized macrocycle is subjected to the sequence of random embedding, rotor search and genetic algorithm sampling; -m, number of conformers embedded per cut; -n, number of conformers generated by rotamer search per cut; -n, number of clash free conformers where the bond is reformed and passed on to the genetic algorithm; -ff, force field used during minimization; -eps dielectric constant applied during minimization; -e, maximum energy difference between the lowest and highest energy conformation; -rmsd, minimum rmsd threshold that two conformers must be apart to be retained in the final population; -s, score to use in the genetic algorithm. RMSD values listed are for all generated conformers per row wherein RMSDs are calculated relative to a co-crystallized structure of FK506 at the macrocyclic ring atoms (freely rotating side chains not considered, see Fig. S1).

t	-r	-m	-n	-N	-ff; -eps	-e; -rmsd	RMSD	-s
1 min	1	3	15	5	MMFF94s	5.0; 0.5	2.173, 2.625, 2.516, 2.738, 2.540, 2.774, 2.918	energy
>2 hrs	5	50	15	50	MMFF94s	5.0; 0.5	2.557, 2.588, 2.546, 2.566, 2.792, 2.751	energy
1 min	1	3	15	5	MMFF94; 78.4	5.0; 0.5	2.599, 2845	energy
2 min	1	3	30	15	MMFF94s	5.0; 0.5	2.497, 2.155, 2.155, 2.151, 2.171, 2.484, 2.435	energy
3 min	1	3	50	30	MMFF94s	5.0; 0.5	2.423, 2.589, 2.347, 2.273, 2.583, 2.386, 2.706, 2.715, 2.911, 2.778, 2.774, 2.130, 2.639, 2.486, 2.394, 2.374,	energy

							2.879, 2.332	
4 min	1	3	50	30	MMFF94s	5.0; 1.0	2.426, 2.286, 2.500, 2.296, 2.156, 2.505, 2.526, 2.796, 2.491, 2.184, 2.597, 2.165	energy
2 min	1	3	30	15	MMFF94s	10.0; 0.5	2.775, 2.705, 2.988, 2.624, 2.532, 2.529, 2.435, 2.018, 2.453, 2.306, 2.642, 2.651, 2.091, 2.213	
2 min	1	3	30	30	MMFF94s	5.0; 0.5	2.435, 2.933	energy
2 min	1	3	30	15	MMFF94s	5.0; 0.2	2.617, 2.362, 2.277, 2.403, 2.608	rmsd
2 min	1	3	30	15	MMFF94s	10; 0.5	2.305, 2.594, 2.258, 2.405, 2.548, 2.778, 2.049, 2.273	rmsd
2 min	1	3	50	30	MMFF94s	5.0; 0.5	2.406, 2.185, 2.706, 2.561, 2.406, 2.178, 2.178, 2.543, 2.580, 2.720, 2.832, 2.662	rmsd
6 min	1	3	50	15	MMFF94s	5.0; 0.5	1.991, 2.285, 2.202, 2.629	rmsd
4 min	1	3	50	15	MMFF94s	5.0; 0.5	2.420, 2.639,	rmsd

							2.795, 2.638	
3 min	1	3	50	5	MMFF94s	5.0; 0.5	2.312, 2.334, 2.893, 2.886	rmsd
3 min	1	5	30	15	MMFF94s	5.0; 0.5	2.197, 2.463, 2.570	rmsd
3 min	1	5	30	15	MMFF94s	5.0; 0.5	2.641, 2.454, 2.661, 2.740, 2.603	rmsd
4 min	1	5	50	5	MMFF94s	10.0; 0.5	2.727, 2.128, 2.461, 2.162, 2.235, 2.356	rmsd
3 min	1	3	30	15	MMFF94s	15.0; 0.5	2.469, 2.750, 2.460, 2.439, 2.747, 2.635, 2.650, 2.431, 2.718, 2.686, 2.185, 2.814, 2.830, 2.637, 2.531, 2.426, 2.771, 2.458, 2.726, 2.482, 2.609, 2.715, 2.686, 2.537	rmsd
3 min	1	3	30	5	MMFF94s	5.0; 1.0	2.224, 3.074, 2.822, 2.851	rmsd
3 min	1	3	50	5	MMFF94s	5.0; 1.0	2.757, 2.499, 2.805, 2.462, 2.533, 2.640	rmsd
3 min	1	3	30	5	MMFF94s	5.0; 2.0	2.519, 2.649, 2.477, 2.246, 2.915	rmsd
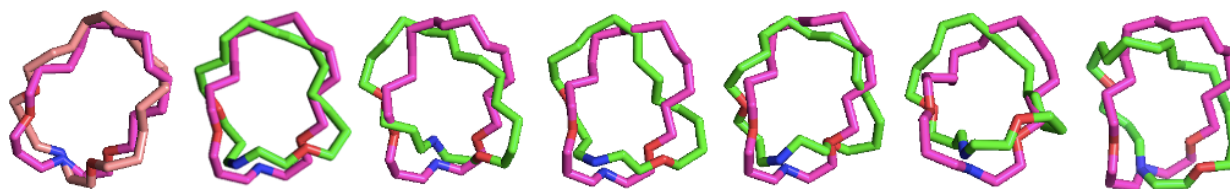
Figure S1. Representative backbone overlays for conformers generated with CB++ using parameters displayed in Table 1, row 1; pink backbone represents co-crystallized conformation of FK506 with FKBP-12, downloaded from the PDB (Protein Data Bank), orange backbone represents crystallized conformation of the unbound ligand, downloaded from the CCDC (Cambridge Crystallographic Data Center), green backbones represent CB++ conformers.

MTHFR Structure Analysis
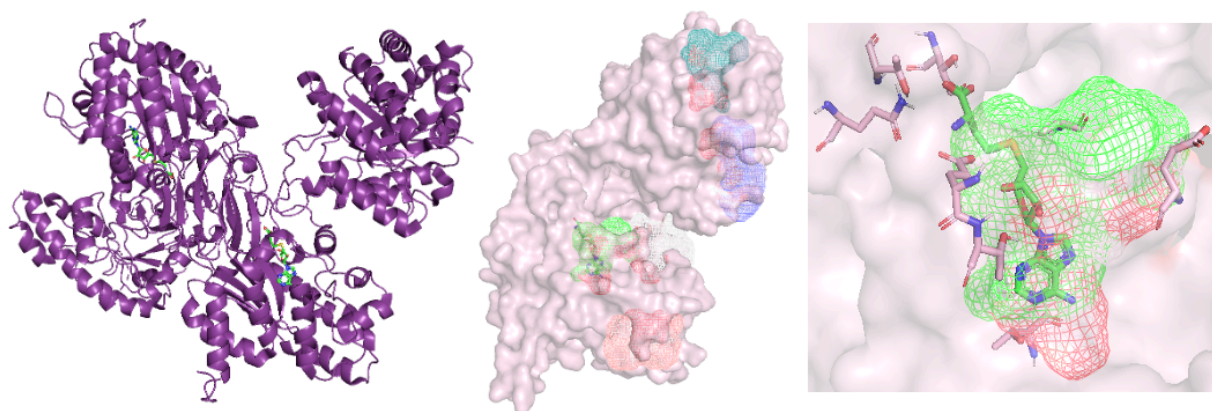
Cluster # | Energy | #points | Radius ofN	e	v	rg	epv	buriedness	v*buriedness^2/rg
1	-136.551000	299.000000	4.719103	-0.456692	0.865772	47.491802
2	-164.925000	370.000000	5.267753	-0.445743	0.821173	47.363713
3	-191.648000	426.000000	6.426949	-0.449878	0.819188	44.480749
4	-136.885000	266.000000	4.805483	-0.514605	0.886323	43.483936
5	-151.297000	295.000000	5.359018	-0.512871	0.871224	41.782711

Figure S2. AutoSite results for MTHFR (PDB: 6fcx) with MTHFR in purple and *S*-adenosyl homocysteine (SAH) in stick representation in green (left to right) MTHFR dimer; surface representation of MTHFR with AutoSite generated clusters indicated in mesh; co-crystallized structure of the protein-ligand complex with the key cluster indicated in mesh in green and key protein residues in pink; (bottom table) AutoSite scoring of predicted high affinity clusters on the protein surface. Columns from left to right: cluster index; energy e; number of grid points v; radius of gyration rg; energy per grid point epv; buriedness = fill's buried surface area/fill's total surface area; empirical composite score = v*buriedness^2/rg; highlighted row indicates key cluster found at the ligand binding allosteric site.

Filter Implementations

Table S3. SMILES string-based filtering for GIPr and MTHFR respectively for the screens described in the indicated sections of Ch.2.

Target	Known Binder	SMILES filter
MTHFR – 1 (section 2.6)	S-adenosyl-L-homocysteine	NCC(NCCC1=CC=CC=C1)=O
MTHFR – 2 (section 2.7)		O=C1N2C(C=CC=C2)=CC=C1 or O=C1C2=CC=CN2C=CN1
GIPr (section 2.6)	GIP	OC/C=C/C1=CC=CC=C1 NC/C=C/C1=CC=CC=C1 CC/C=C/C1=CC=CC=C1 O=C1CC=CN1 C1(C(CCN2)C2N3)=C3C=CC=C1 O=C1CC2CC3=CC=CC=C3C2N1 O=C1CC(CC2=CC=CC=C2)CN1 O=C1CC(CC2=CC=CC=C2)C(C)N1 O=C1NCCC1CC2=CC=CC=C2F

Table S4. Configuration parameters used for running Autodock Vina against the listed targets within the screens described in section 2.6 of Ch. 2; exhaustiveness = 8 in each case; grid size in Å.

Target	PDB	Center_x	Center_y	Center_z	Size_x	Size_y	Size_z
FKBP-12	2fke	23.200	28.760	6.900	20	20	15
GIPr	2qkh	18.742	4.962	43.351	20	20	20
MTHFR	6fcx	-29.997	22.154	-4.575	22	22	24

SMARTS patterns in Fig. S3 are as follows –

P1 = [#8]=[#6](-[#7]-[#6]-[#6]-[cR1]1[cR1][nR1]c2c1[cR1][cR1][cR1][cR1]2)-[#6]-[#6]-[cR1]3[cR1][nR1]c4c3[cR1][cR1][cR1][cR1]4

P2 = [#8]=[#6](-[#6]-[#6]-[#6]-[cR1]1[cR1][nR1]c2[cR1][cR1][cR1][cR1]c12)-[#7]-[#6]-[#6]-[#6](-[#7]-[#6]-[#6]-[#6]-[cR1]3[cR1][nR1]c4[cR1][cR1][cR1][cR1]c34)=[#8]

P3 = [#8]=[#6](-[#7]-[#6]-[#6](-[#7]-[#6]-[#6](-[#7]-[#6]-[#6]-[cR1]1[cR1][nR1]c2[cR1][cR1][cR1][cR1]c12)=[#8])=[#8])-[#6]-[#6]-[cR1]3[cR1][nR1]c4[cR1][cR1][cR1][cR1]c34

P4 = [#8]=[#6](-[#7]-[#6]-[#6](-[#7]-[#6]-[#6](-[#7]-[#6]-[#6]-c1cnc2ccccc12)=[#8])=[#8])-[#6]-[#6]-c3cnc4ccccc34; P5= [#8]=[#6](-[#6]-[#6]-[#6]-c1cnc2ccccc12)-[#7]-[#6]-[#6]-[#6](-[#7]-[#6]-[#6]-[#6]-c3cnc4ccccc34)=[#8]

Figure S3. A. Demonstrating the application of a SMARTS pattern filter (top left) for a corresponding bis-indole containing peptide sequence (top right). The R1 designations in the SMARTS string requires both indoles to be exocyclic. Accordingly, **I1** is retained but **I2** (indole is endocyclic) and **I3** (indole is part of a polycycle) are filtered. **B**. SMARTS patterns generated based on known ligands for constructing ligand libraries screened against HDM2 (known ligand top left) and AP2 (known ligand top right) in section 2.6 of Ch. 2. All patterns **P1**-**P4** maintain the required bond distance between the indole rings in both known binders (shown in red within the respective binders); **P2** and **P4** identifies peptides containing β^2- and β^3-amino acids, whereas **P3** and **P5** identifies peptides with α- amino acids. Using R1 designations as in **P2** and **P3** identifies macrocycles having only exocyclic indoles, whereas **P4** and **P5** identifies macrocycles containing endocyclic and/or exocyclic indoles.

Table S5. Studying CPMG spread within property filter variations; MW = molecular weight, TPSA = total polar surface area, HBD = hydrogen bond donors. All properties calculated with RDKit (see scripts at the end of this appendix). Duration refers to the amount of time the job was run before it was terminated, wherein running the jobs for longer would identify a larger number of compounds in each case from the giga-library.

Properties	Length	Duration	Number of compounds identified
$0.2 \leq m \leq 0.3$	3mers	3h	7926
MW ≤ 800		24h	60,593
TPSA ≤ 150	4mers	3 h	0
HBD ≤ 5	5mers	3 h	0
$0.2 \leq m \leq 0.3$	3mers	3h	46,841
MW ≤ 960	4mers	3h	11,487
TPSA ≤ 180	5mers	3h	555
No HBD filter			
$0.2 \leq m \leq 0.3$	3mers	3h	80,425
MW ≤ 1200	4mers	3h	40,831
TPSA ≤ 200	5mers	3h	11,569
HBD ≤ 5			

Table S6. Property and pharmacophoric filters applied for constructing the ligand library against MTHFR for the screen detailed in section 2.7 of Ch.2. The final ligand library for this screen was compiled using the common structures from sets 1A and 2A(ii), comprising a total of 53,577 ligands.

Criteria	Set	CPMG search space	Number identified	Number searched
0.2 < TPSA/MW < 0.3 MW <800 TPSA <150 HBD ≤ 5	1A	Macrocycles derived from trimer peptides	354,876	At least 418,104,951
With regioisomer/stereoisomer restrictions	1B	Macrocycles derived from tetramer peptides	27,111	At least 715,789,823
Structures containing any of the following heterocycles:	2A (i)	Macrocycles derived from trimer peptides	1,948,297	At least 418,104,792
	2A (ii)	With only **16** or **19**	865,274	
	2B (i)	Macrocycles derived from tetramer peptides	1,048,576	At least 849,503,407

MTHFR Screen Results

Table S7. Top 200 hits from MTHFR screen described in section 2.7 of Ch.2.

ID	SMILES
50k3_1_12	Cn(cc(CC[C@@H](C(N1[C@@H](C[C@@H]2C[C@@H]21)C(N)=O)=O)NC([C@@H]3CNC(CCc4cc(/C=C/Cc(cc5C3)c6c5ccccc6)ccc4)=O)=O)n7c8ccc7)c8=O
50k3_1_216	CN1C([C@H]2[C@@H]3CCC[C@H](N2C(CCc4cc(/C=C/Cc(cc5)n6c5[nH]cc6C[C@@H]1CC(N[C@@H](C(O)=O)Cc(cc7=O)cc8n7cccc8)=O)ccc4)=O)C3)=O
50k3_10_144	Cn(ccn1c2ccc1C[C@@H](C(NCC(N)=O)=O)NC(C[C@@H]3CCc(cc4C/C=C/c5cc(C[C@H]6C=CN3C6=O)c(F)cc5)c7c4ccccc7)=O)c2=O
50k3_10_198	CN([C@@H](C(O)=O)CCc1nc[nH]c1)C([C@@H]2C[C@H]3C[C@H]3N2C(C[C@@H]4Cc(cc(C/C=C/c5cc(C[C@H]6C=CN4C6=O)c(F)cc5)c7=O)c8n7cccc8)=O)=O
50k3_10_212	CN([C@@H](C(N)=O)Cc(cc1)c2c1ccccc2)C([C@H]3[C@@H]4CC[C@H](N3C(C[C@@H]5Cc(co6)c7c6ccn(C/C=C/c8cc(CCC(N5)=O)ccc8)c7=O)=O)C4)=O
50k3_11_8	CN(C([C@@]12CC[C@H](C2)CN1C([C@@H]3CCc4c(C/C=C/c5cc(CCC(NC3)=O)ccc5)cccc4)=O)=O)C[C@@H](C(N)=O)CCc(occ6)c7n6c(=O)cc7
50k3_11_20	O=C1CCc2cc(/C=C/Cn(cc(C[C@@H](NC([C@@H](CN1)Cc3ccccc3)=O)C(N4[C@@H]5CCC[C@H]([C@@H]4C(O)=O)C5)=O)n6c7ccc6)c7=O)ccc2
50k3_11_64	NC([C@@H]1CCC2(N1C([C@H]3Cc(cc4)n5c4c(=O)n(C/C=C/c6cc(CCC(N[C@H](CC(N3)=O)CCc(cc7)c8c7ccccc8)=O)ccc6)cc5)=O)CC2)=O
50k3_11_67	Cn(ccn1c2cc(C/C=C/c3cc(CC4)ccc3)c1C[C@@H](CC(N5[C@H](CCC56CC6)C(N7[C@H]8CC[C@@H]([C@H]7C(O)=O)C8)=O)=O)NC4=O)c2=O
50k3_11_170	CC([C@@H](N(C([C@@H]1[C@@H]2CC[C@H](N1C(C[C@@H]3CCc(cn(C/C=C/c4cc(CC(N3)=O)ccc4)c5=O)c6c5cco6)=O)C2)=O)C)C(NCCc7[nH]ccc7)=O)C
50k3_12_30	CN(C([C@@H]1Cc(occ2)c3n2c(=O)c(C/C=C/c4cc(CCC(N5C[C@@H]6C[C@@H]6[C@H]5C(N1)=O)=O)ccc4)c3)=O)C[C@@H](C(O)=O)Cc(cc7)c8n7cccc8
50k3_12_80	NC([C@]12CC[C@@H](N2C([C@H](CNC([C@H]3CCc4cc(C/C=C/c5cc(CCC(N3)=O)ccc5)ccc4)=O)CCc(ccc6=O)c7n6cccc7)=O)C1)=O
50k3_12_88	NC(C[C@@H](NC([C@@H]1[C@H]2C[C@H]2CN1C(C[C@@H]3CCc(cc(C/C=C/c4cc(C[C@H]5C=CN3C5=O)c(F)cc4)c6=O)c7n6cccc7)=O)=O)Cc8occc8)=O
50k3_12_94	C#CC[C@@]12CCCN1C(C[C@@H](NC(CCc3cc(/C=C/Cc4c(C[C@@H](NC2=O)C(N)=O)cccc4)ccc3)=O)CCc(cc5=O)cc6n5ccccc6)=O
50k3_12_95	CC([C@@H](N(C(C[C@@H](NC([C@@H]1Cc(cc2C/C=C/c3cc(CCC(N1)=O)ccc3)c4n2cccc4)=O)CCc(ccc5=O)c6n5cccc6)=O)C)C(N)=O)C
50k3_100_36	C#CCCC[C@@]12CC(N3CC(N([C@@H](Cc4cc(C/C=C/c5cc(C1)c([C@H]32)cc5)ccc4)C(N[C@@H](C(N)=O)CCc(occ6)c7n6c(=O)cc7)=O)C)=O)=O
50k3_100_45	CN([C@@H](CC(N)=O)CCc1occc1)C([C@@H]2[C@H]3CC[C@@H](N2C(C[C@@H]4Cc(cc5n6cccc5)c(C/C=C/c7cc(CCC(N4)=O)ccc7)c6=O)=O)C3)=O
50k3_100_56	O=C([C@@H]1C[C@@H](CN1C([C@@H]2Cc(cn(C/C=C/c3cc(CCC(N4[C@H](C(N2)=O)C[C@H]5C[C@H]54)=O)ccc3)c6=O)n7c6ccc7)=O)Oc8ccccc8)O
50k3_100_58	Cn(ccn1c2cc(C/C=C/c3cc(CC4)ccc3)c1C[C@@H](C(N5[C@@H](CCC56CC6)C(N[C@@H](Cc(cc7)c8c7ccccc8)CC(N)=O)=O)=O)NC4=O)c2=O
50k3_100_112	CN([C@@H](Cc(cc1)c2n1cc[nH]2)CC(N3[C@@H](C[C@@H]4C[C@@H]43)C(O)=O)=O)=O)C([C@@H]5Cc(cc6n7cccc6)c(C/C=C/c8cc(CCC(N5)=O)ccc8)c7=O)=O
50k3_100_189	NC([C@H](CNC([C@H]1[C@H]2CC[C@@H](N1C([C@@H]3CNC(CCc4cc(/C=C/Cc(c(C3)c5=O)c6n5cccc6)ccc4)=O)=O)C2)=O)CCc7ccccc7)=O
50k3_100_239	CN([C@@H](C(N1[C@@H](C[C@@H]2C[C@@H]21)C(O)=O)=O)Cc(c[nH]c3=O)n4c3ccc4)C([C@@H]5CNC(CCc6cc(/C=C/Cc(cc7C5)n8c7cccc8)ccc6)=O)=O
50k3_101_22	CC(C[C@@H](N(C(C(C[C@@H]1CCc(cn(C/C=C/c2cc(CCC(NC[C@@H](C(N1)=O)CCc3cccc3)=O)ccc2)c4=O)n5c4ccc5)=O)C)C(N)=O)C
50k3_101_155	CN([C@@H](Cc1ccccc1)CC(N2[C@H](C[C@H]2C(O)=O)O)=O)C(C[C@@H]3CCc(cc4n5cccc4)c(C/C=C/c6cc(CCC(N3)=O)ccc6)c5=O)=O

128

50k3_1 02_183	CN([C@@H](C(N)=O)Cc(cc1)c2c1ccccc2)C([C@@H]3[C@@H]4CCC[C@H](N3C([C@@H]5Cc(cn(C/C=C/c6cc(CCC(N5)=O)ccc6)c7=O)c8c7cco8)=O)C4)=O
50k3_1 03_42	NC([C@@H]1CCC2(N1C([C@]34CC[C@@H](C4)CN3C(C[C@@H]5Cc(cn(C/C=C/c6cc(C[C@H]7C=CN5C7=O)c(F)cc6)c8=O)c9c8cco9)=O)=O)CC2)=O
50k3_1 03_68	CN([C@@H](C(N1[C@H]2CC[C@@H]([C@@H]1C(O)=O)C2)=O)CCc3cc(O)ccc3)C(C[C@@H]4Cc(cc5n6cccc5)c(C/C=C/c7cc(CCC(N4)=O)ccc7)c6=O)=O
50k3_1 03_127	CN([C@@H](Cc(cc1)c2c1ccccc2)CC(N3C[C@H]4C[C@H]4[C@H]3C(N)=O)=O)C([C@@H]5Cc(cn(C/C=C/c6cc(CCC(N5)=O)ccc6)c7=O)c8c7cco8)=O
50k3_1 03_188	CN(C([C@@H]1Cc(oc(C/C=C/c2cc(CCC(N1)=O)ccc2)c3)c4n3c(=O)cc4)=O)C[C@@H](C(N5C[C@H]6C[C@H]6[C@H]5C(N)=O)=O)CCc(cc7)c8c7ccccc8
50k3_1 03_200	CN1C[C@@H](CCc(cn(C/C=C/c2cc(CCC(N3[C@H](C1=O)C[C@H]4C[C@H]43)=O)ccc2)c5=O)c6c5cco6)C(N7C[C@H](C[C@H]7C(N)=O)Oc8ccccc8)=O
50k3_1 03_233	Cn(ccn1c2cc(C/C=C/c3cc(CC4)ccc3)c1C[C@@H](CC(N5C[C@@H]6C[C@@H]6[C@H]5C(NC[C@@H](C(N)=O)Cc(cc7)c8c7ccccc8)=O)=O)NC4=O)c2=O
50k3_1 04_51	O=C([C@@H](NC([C@@H]1C[C@H]2C[C@H]2N1C([C@@H]3CCc(cc(C/C=C/c4cc(C[C@H]5C=CN3C5=O)c(F)cc4)c6=O)c7n6cccc7)=O)=O)Cc(c[nH]8)n9c8ccc9)O
50k3_1 04_190	NC([C@@]12CC[C@H](N2C([C@H]3[C@H]4CC[C@@H](N3C(C[C@H]5Cc(cc(C/C=C/c6cc(CCC(N5)=O)ccc6)c7=O)c8n7cccc8)=O)C4)=O)C1)=O
50k3_1 04_200	O=C(C[C@@H](NC([C@H]1CCC2(N1C([C@@H]3CCc4cc(OC/C=C/c5cc(CCC(NC3)=O)cc5)ccc4)=O)CC2)=O)Cc(cc6)cc7n6cccc7)O
50k3_1 05_113	CC([C@@H](N(C([C@@H](NC(C[C@@H]1Cc(ccc2)n3c2ccc3C/C=C/c4cc(CCC(N1)=O)ccc4)=O)CCc(cc5=O)cc6n5cccc6)=O)C)C(N)=O)C
50k3_1 05_146	O=C([C@@H](Cc(cc1=O)cc2n1cccc2)CNC([C@@H]3Cc4c(C/C=C/c5cc(CCC(N6C[C@H]7CC[C@]6(C(N3)=O)C7)=O)ccc5)cccc4)=O)O
50k3_1 06_10	NC([C@]12CC[C@@H](N2C([C@@H]3CCC4(N3C([C@@H]5CNC(CCc6cc(/C=C/Cc(cc(C5)c7n8cccc7)c8=O)ccc6)=O)=O)CC4)=O)C1)=O
50k3_1 06_29	O=C1CCc2cc(/C=C/Cn(cc(C[C@@H](N1)CC(N3[C@@H]([C@H]4CCC[C@@H]3C4)C(N5[C@H]6CCC[C@@H]([C@@H]5C(O)=O)C6)=O)=O)c7c8cco7)c8=O)ccc2
50k3_1 06_39	CN(C([C@@H]1Cc(co2)c3c2ccn(C/C=C/c4cc(CCC(N5[C@H](C(N1)=O)CCC56CC6)=O)ccc4)c3=O)=O)C[C@@H](C(N)=O)Cc(cc7)c8c7ccccc8
50k3_1 06_174	CN(C([C@@H]1Cc(cc2n3cccc2)c(C/C=C/c4cc(CCC(N1)=O)ccc4)c3=O)=O)C[C@@H](C(N5C[C@@H]6CC[C@]5(C6)C(O)=O)=O)Cn(cc7)c8c7occ8
50k3_1 06_184	NC([C@@H]1CCC2(N1C([C@@H]3[C@H]4C[C@H]4CN3C(C[C@H]5Cc(cc(C/C=C/c6cc(CCC(N5)=O)ccc6)c7=O)c8n7cccc8)=O)=O)CC2)=O
50k3_1 06_223	NC([C@H]1[C@H]2CC[C@@H](N1C([C@@H]3[C@H]4C[C@H]4CN3C(C[C@@H]5Cc(cc6=O)oc7n6cccc7C/C=C/c8cc(C[C@H]9C=CN5C9=O)c(F)cc8)=O)=O)C2)=O
50k3_1 09_17	NC([C@@]12CC[C@H](N2C([C@H]3CCc(cc4)n5c4c(=O)n(C/C=C/c6cc(CCC(NC[C@H](C(N3)=O)Cc(cc7)c8c7ccccc8)=O)ccc6)cc5)=O)C1)=O
50k3_1 09_44	O=C1CCc2cc(/C=C/Cn(cc(C[C@@H](N1)CC(N3[C@@H]([C@H]4CC[C@@H]3C4)C(NC[C@@H](C(O)=O)Cc5ccccc5)=O)=O)n6c7ccc6)c7=O)ccc2
50k3_1 09_107	O=C(C[C@@H](NC(CNC(C[C@@H]1Cc(cc(C/C=C/c2cc(CCC(N1)=O)ccc2)c3=O)c4n3cccc4)=O)=O)CCc(cc5)c6c5ccccc6)O
50k3_1 10_20	CN([C@@H](CC(O)=O)CCc(c[nH]1)c2c1cccc2)C([C@@H]3CCc(cc4)n5c4c(=O)n(C/C=C/c6cc(CCC(N7[C@@H]8CC[C@@]7(C(N3)=O)C8)=O)ccc6)cc5)=O
50k3_1 10_57	NC(C[C@@H](NC([C@@H]1[C@H]2CCC[C@@H](N1C([C@@H]3Cc(cn(C/C=C/c4cc(CCC(N3)=O)ccc4)c5=O)n6c5ccc6)=O)C2)=O)Cc(cc7)c8c7ccccc8)=O
50k3_1 10_140	CN(C([C@@H]1CCc(occ2)c3n2c(=O)c(C/C=C/c4cc(CCC(N5C[C@H]6C[C@H]6[C@H]5C(N1)=O)=O)ccc4)c3)=O)C[C@@H](C(N)=O)Cc(cc7)c8c7ccccc8
50k3_1 11_105	Cn(cc1n2c3ccc2C/C=C/c4cc(CCC(N[C@@H](C1)CC(N5C[C@@H]6C[C@@H]6[C@H]5C(N[C@@H](Cc(cc7)c8c7ccccc8)CC(N)=O)=O)=O)=O)ccc4)c3=O
50k3_1 11_199	C#CC[C@]1(CCCN1C(C[C@@H]2Cc(cc3n4cccc3)c(C/C=C/c5cc(CCC(N2)=O)ccc5)c4=O)=O)C(N6[C@@H]7CC[C@]6(C7)C(N)=O)=O
50k3_1 11_215	CN1C([C@H]2[C@@H]3CC[C@H](N2C(CCc4cc(/C=C/Cn(cc(C[C@@H]1CC(N5C[C@H](C[C@H]5C(N)=O)Oc6ccccc6)=O)c7c8cco7)c8=O)ccc4)=O)C3)=O

50k3_1 12_173	CN([C@@H](CC(N1C[C@H](C[C@H]1C(O)=O)O)=O)CCc(cc2=O)cc3n2cccc3)C([C@@H]4 CCc5c(C/C=C/c6cc(CCC(N4)=O)ccc6)cccc5)=O
50k3_1 12_202	Cn(ccn1c2ccc1CC[C@@H](C(N)=O)NC([C@@H]3CCC4(N3C([C@@H]5Cc(cc6C/C=C/c7c c(CCC(N5)=O)ccc7)c8c6ccccc8)=O)CC4)=O)c2=O
50k3_1 13_52	CN(C([C@@H]1Cc(cn(C/C=C/c2cc(CCC(N3[C@@H]4CC[C@H]([C@@H]3C(N1)=O)C4)= O)ccc2)c5=O)c6c5cco6)=O)C[C@@H](C(N)=O)Cc(cc7)c8c7ccccc8
50k3_1 13_241	NC(C[C@@H](NC([C@@]12CC[C@H](C2)CN1C(C[C@@H]3Cc4ccc(C/C=C/c5cc(CCC(N3)=O)ccc5)cc4)=O)=O)Cc(cc6=O)cc7n6cccc7)=O
50k3_1 14_45	CN([C@@H](C(O)=O)CCc(ccc1)n2c1ccc2)C([C@@H]3[C@H]4CC[C@@H](N3C([C@@H] 5Cc(co6)c7c6ccn(C/C=C/c8cc(CCC(N5)=O)ccc8)c7=O)=O)C4)=O
50k3_1 14_150	CN([C@@H](Cc(c[nH]1)c2c1cccc2)CC(O)=O)C(C[C@@H]3Cc(cn(C/C=C/c4cc(CCC(N5[C @H]6CC[C@@H]([C@H]5C(N3)=O)C6)=O)ccc4)c7=O)n8c7ccc8)=O
50k3_1 14_156	NC([C@@H]1CCc(cc(CC=Cc2cc(CCC(N3[C@H]4CC[C@]3(C(N5[C@H]6CCC[C@@H]([C @@H]5C(N1)=O)C6)=O)C4)=O)ccc2)c7=O)c8n7cccc8)=O
50k3_1 14_235	Cn(cc1n2c3cc(C/C=C/c4cc(CCC(NC[C@@H](CC1)C(NC[C@@H](C(N5[C@@H]([C@H]6C CC[C@@H]5C6)C(N)=O)=O)Cc7ccccc7)=O)=O)ccc4)c2)c3=O
50k3_1 16_52	Cn(ccn1c2c(C/C=C/c3cc(CC4)ccc3)cc1CC[C@@H](C(N5[C@@H]6CC[C@]5(C6)C(N[C@ @H](C(N)=O)Cc(cc7)c8c7ccccc8)=O)=O)NC4=O)c2=O
50k3_1 16_75	NC([C@H]1[C@@H]2CC[C@H](N1C(C[C@@H]3Cc(cc(C/C=C/c4cc(CCC(N5C[C@@H]6C[C@@H]6[C@H]5C(N3)=O)=O)ccc4)c7=O)c8n7cccc8)=O)C2)=O
50k3_1 16_224	CN(C([C@@H]1Cc2ncn(C/C=C/c3cc(CCC(N1)=O)ccc3)c2)=O)C[C@@H](C(N4[C@H]5CC[C@@]4(C5)C(O)=O)=O)CCc(cc6=O)cc7n6cccc7
50k3_1 18_20	Cn1cc(C[C@@H](C(N)=O)CNC([C@H]2[C@@H]3CC[C@H](N2C([C@@H]4Cc(cc(C/C=C/ c5cc(CCC(N4)=O)ccc5)c6=O)c7n6cccc7)=O)C3)=O)c8c1cccc8
50k3_1 18_231	CN([C@@H](Cc(ccc1=O)c2n1cccc2)CC(N3[C@@H]4CCC[C@H]([C@@H]3C(O)=O)C4)= O)C([C@@H]5CNC(CCc6cc(/C=C/Cn7cc(C5)nc7)ccc6)=O)=O
50k3_1 19_91	CN([C@@H](C(N1CCC[C@@H]1C(O)=O)=O)Cc(c[nH]2)c3c2cccc3)C([C@@H]4CCc(cn(C/ C=C/c5cc(CCC(NC4)=O)ccc5)c6=O)n7c6ccc7)=O
50k3_1 19_93	NC([C@H]1CCc(cc2)n3c2c(=O)n(C/C=C/c4cc(CCC(NC[C@H](C(N5[C@@H]6CC[C@H]([C @@H]5C(N1)=O)C6)=O)Cc(cc7)c8c7ccccc8)=O)ccc4)cc3)=O
50k3_1 20_68	CN([C@@H](C(O)=O)CCc(cc1)c2n1cccc2)C([C@@H]3C[C@H]4C[C@H]4N3C([C@@H]5 Cc(occ6)c7n6c(=O)c(C/C=C/c8cc(CCC(N5)=O)ccc8)c7)=O)=O
50k3_1 20_78	CN(C([C@@H]1Cc2ncn(C/C=C/c3cc(CCC(N1)=O)ccc3)c2)=O)C[C@@H](C(N4[C@@H]5C CC[C@H]([C@@H]4C(O)=O)C5)=O)CCc(cc6=O)cc7n6cccc7
50k3_1 21_157	NC([C@H](NC([C@@H]1CCn2c(C/C=C/c3cc(CCC(N4[C@@H]([C@H]5CCC[C@@H]4C5) C(NC1)=O)=O)ccc3)ccc2)=O)CCc(cc6=O)cc7n6cccc7)=O
50k3_1 22_31	Cn1c(C[C@@H](NC([C@@H]2[C@@H]3C[C@@H]3CN2C(CCc4cc(/C=C/C5)ccc4)=O)=O) CC(NC[C@@H](C(O)=O)Cc(cc6=O)cc7n6ccccc7)=O)c5cc1
50k3_1 22_191	CN([C@@H](C(N)=O)CCc(cc1)c2c1ccccc2)C([C@@H]3Cc(occ4)c5n4c(=O)c(C/C=C/c6cc(CCC(N7[C@@H]8CC[C@H]([C@@H]7C(N3)=O)C8)=O)ccc6)c5)=O
50k3_1 23_24	CN([C@@H](Cc(cc1)c2c1ccccc2)CC(N)=O)C([C@@H]3CNC([C@@]45CC[C@H](N5C(CC c6cc(/C=C/Cn(ccc7c8c(C3)co7)c8=O)ccc6)=O)C4)=O)=O
50k3_1 23_103	CN(C([C@@H]1C[C@@H]2C[C@@H]2N1C([C@@H]3CCc(cn(C/C=C/c4cc(CCC(NC3)=O) ccc4)c5=O)c6c5cco6)=O)=O)C[C@@H](C(N)=O)CCc7ccccc7
50k3_1 23_217	C[C@@H](C(N[C@@H](C(N)=O)Cc(ccc1=O)c2n1cccc2)=O)NC([C@@H]3CCc(cc4C/C=C/ c5cc(C[C@H]6C=CN3C6=O)c(F)cc5)c7n4cccc7)=O
50k3_1 25_52	O=C(N[C@@H](C(O)=O)Cc(cc1)c2n1cccc2)C[C@@H]3Cc(cn(C/C=C/c4cc(CCC(N5[C@@ H]([C@H]6CC[C@@H]5C6)C(N3)=O)=O)ccc4)c7=O)n8c7ccc8
50k3_1 25_82	CN(C([C@@H]1Cc2c(C/C=C/c3cc(CCC(N1)=O)ccc3)ccc(O)c2)=O)C[C@@H](C(N4[C@H](CCC45CC5)C(O)=O)=O)Cc(cc6=O)cc7n6cccc7
50k3_1 25_133	O=C1CCc2cc(/C=C/Cn(cc(CC[C@@H](CN1)C(N3C[C@@H]4C[C@@H]4[C@H]3C(NC[C @@H](C(O)=O)Cc(cc5)c6n5ccccc6)=O)=O)n7c8ccc7)c8=O)ccc2
50k3_1 25_184	CN(C([C@@H]1Cc(cn(C/C=C/c2cc(CCC(N1)=O)ccc2)c3=O)c4c3cco4)=O)C[C@@H](C(N5[C@H]6CC[C@@H]([C@H]5C(N)=O)C6)=O)Cc(cc7)c8c7ccccc8

130

50k3_1 26_111	CN([C@@H](Cc(cc1=O)cc2n1cccc2)CC(N3[C@H]4CCC[C@@H]([C@H]3C(O)=O)C4)=O)C([C@@H]5Cc(cc6C/C=C/c7cc(CCC(N5)=O)ccc7)c8n6cc[nH]8)=O
50k3_1 26_224	Nc(=O)c1c[C@]23C/C=C/c4cc(CCC(N5[C@H](C(N[C@@H](C(N1[C@@H]2Nc6c3cc(F)cc6)=O)CCc(cc7=O)cc8n7cccc8)=O)C[C@H]9C[C@H]95)=O)ccc4
50k3_1 27_17_	CN([C@@H](C(O)=O)Cc(cc1)c2n1cccc2)C([C@@H]3[C@H]4CC[C@@H](N3C(C[C@@H]5Cc(cn(C/C=C/c6cc(CCC(N5)=O)ccc6)c7=O)c8c7cco8)=O)C4)=O
50k3_1 27_79	CN([C@@H](C(O)=O)CCc1ccccc1)C([C@H]2[C@H]3CCC[C@@H](N2C(C[C@@H]4Cc(c[nH]c5=O)n6c5cc(C/C=C/c7cc(CCC(N4)=O)ccc7)c6)=O)C3)=O
50k3_1 27_97	Cn(ccn1c2ccc1C[C@@H](NC([C@@H]3CNC(CCc4cc(/C=C/Cc(cc5C3)c6c5ccccc6)ccc4)=O)=O)CC(N7C[C@@H]8C[C@@H]8[C@H]7C(N)=O)=O)c2=O
50k3_1 27_121	Cn(cc1CC[C@@H](C(N)=O)NC([C@@H]2C[C@@H]3C[C@@H]3N2C(C[C@@H]4Cc(cc5=O)c(C/C=C/c6cc(CCC(N4)=O)ccc6)c7n5cccc7)=O)=O)c8c1cccc8
50k3_1 28_54	O=C1CCc2cc(/C=C/Cn(cc(C[C@@H](CN1)C(N3[C@H]4CC[C@@H]([C@H]3C(NC[C@@H](C(O)=O)Cc5ccccc5)=O)C4)=O)n6c7ccc6)c7=O)ccc2
50k3_1 28_123	Cn1c(C[C@@H](NC([C@H]2[C@H]3CCC[C@@H](N2C([C@@H]4CCc(cc5n6cccc5)c(C/C=C/c7cc(CCC(NC4)=O)ccc7)c6=O)=O)C3)=O)CC(N)=O)ccc1
50k3_1 28_142	CN([C@@H](C(N1C[C@@H]2CC[C@]1(C2)C(N)=O)=O)CCc(cc3=O)oc4n3ccc4)C([C@@H]5CCc6c(C/C=C/c7cc(CCC(NC5)=O)ccc7)cccc6)=O
50k3_1 28_173	CN([C@@H](CC(N1[C@@H]([C@H]2CC[C@@H]1C2)C(N)=O)=O)CCc(ccc3=O)c4n3cccc4)C([C@@H]5CNC(CCc6cc(/C=C/Cc7oc(C5)cc7)ccc6)=O)=O
50k3_1 29_107	CN([C@@H](C(N1[C@@H]([C@H]2CCC[C@@H]1C2)C(O)=O)=O)Cc(ccc3=O)c4n3cccc4)C(C[C@@H]5Cc6cc(C/C=C/c7cc(CCC(N5)=O)ccc7)c(O)cc6)=O
50k3_1 29_166	NC([C@@H](NC([C@@H]1Cc(cc2n3cccc2)c(C/C=C/c4cc(CCC(N5[C@@H]6CC[C@H]([C@@H]5C(N1)=O)C6)=O)ccc4)c3=O)=O)Cc(ccc7)n8c7ccc8)=O
50k3_1 29_195	CN1C(C[C@@H](NC(CCc2cc(/C=C/Cc3oc(C[C@@H]1CC(N4[C@H](CCC45CC5)C(N)=O)=O)cc3)ccc2)=O)CCc(cc6=O)cc7n6ccccc7)=O
50k3_1 29_196	NC([C@H]1[C@H]2CC[C@@H](N1C([C@@H]3[C@@H]4C[C@@H]4CN3C([C@H]5CCc(cc6=O)c(C/C=C/c7cc(CCC(NC5)=O)ccc7)c8n6cccc8)=O)=O)C2)=O
50k3_1 30_31	NC([C@H](NC(C[C@@H]1Cc2c(C/C=C/c3cc(CCC(N4[C@H](C(N1)=O)C[C@@H]5CCC[C@@H]54)=O)ccc3)cccc2)=O)Cc(cc6=O)cc7n6cccc7)=O
50k3_1 30_83	Cn(ccc1c2c(CC[C@@H](NC(CCc3cc(/C=C/C4)ccc3)=O)CC(N5[C@H]6CCC[C@@H]([C@H]5C(N7[C@@H](C[C@@H]8CCC[C@@H]87)C(O)=O)=O)C6)=O)c4o1)c2=O
50k3_1 30_191	CN(C([C@@H]1[C@H]2CC[C@@H](N1C([C@@H]3CCc(cc(C/C=C/c4cc(CCC(N3)=O)ccc4)c5=O)c6n5cccc6)=O)C2)=O)C[C@@H](C(O)=O)Cc(c[nH]7)n8c7ccc8
50k3_1 30_215	NC([C@H]1[C@H]2CC[C@H](N1C(C[C@@H]3Cc(cc(C/C=C/c4cc(CCC(N5[C@H](C(N3)=O)C[C@H]6C[C@H]65)=O)ccc4)c7=O)c8n7cccc8)=O)C2)=O
50k3_1 30_246	O=C([C@@H]1CCc(c[nH]c2=O)n3c2ccc3C/C=C/c4cc(C[C@H]5C=CN1C5=O)c(F)cc4)N[C@@H](Cc6ccccc6)CC(N7[C@@H](C[C@H]8C[C@H]87)C(O)=O)=O
50k3_1 31_14	CN([C@@H](C(N)=O)CCc(cc1)c2c1ccccc2)C([C@@H]3CCc(cn(C/C=C/c4cc(CCC(N5C[C@H]6C[C@H]6[C@H]5C(N3)=O)=O)ccc4)c7=O)c8c7cco8)=O
50k3_1 31_115	C[C@@H](N(C(C[C@@H]1CCc(cc2)n3c2c(=O)n(C/C=C/c4cc(CCC(N1)=O)ccc4)cc3)=O)C)C(N[C@@H](CC(N)=O)CCc(cc5)c6c5ccccc6)=O
50k3_1 31_161	CN([C@@H](CC(O)=O)CCc1ccccc1)C([C@]23CC[C@@H](N3C([C@@H]4CCc(c(C/C=C/c5cc(CCC(N4)=O)ccc5)c6)n7c6c(=O)[nH]cc7)=O)C2)=O

Code Snippets

The following scripts demonstrate filter implementation on the CPMG library, structure analysis, submitting job arrays for conformer generation with CB++ and submitting job arrays for parallelized docking with AutoDock Vina. These may be used as general templates and adapted to perform similar tasks.

```python
### filter by smiles or smarts

import cpmg.repository as repository
from rdkit import Chem
from rdkit.Chem import rdMolDescriptors
import random

def load_macrocycles(length, mol_type='macrocycle'):
    repo = repository.create_repository_from_string(mol_type)
    for mol in repo.find({'length': length}):
        yield mol

pattern_sm=['smiles']
#for smarts use pattern_sm=Chem.MolFromSmarts['smarts']

patterns=[]
for smiles in pattern_sm:
    m=Chem.MolFromSmiles(smiles)
    patterns.append(m)

###for smarts use:
#for smarts in pattern_sm:
#   m=Chem.MolFromSmarts(smarts)
#   patterns.append(m)

print(patterns)

a=1        #a and b represent different datasets
b=1
n="%d % k"     #k=desired output number
m="%d % k"     #split up dataset

matched=[Chem.MolFromSmiles('CCCCCCCCCC'), Chem.MolFromSmiles('CCCCCCCCC(C)=O')]
#arbitrary entries for first if loop

if __name__ == "__main__":
    length=3              # change to access trimers, tetramers, or pentamers
    try:
        for macrocycle in load_macrocycles(length):
            m = Chem.MolFromSmiles(macrocycle['kekule'])
```

```python
        mw=Chem.rdMolDescriptors.CalcExactMolWt(m)
        mw_lastmatch=Chem.rdMolDescriptors.CalcExactMolWt(matched[-1])
        mw_seclastmatch=Chem.rdMolDescriptors.CalcExactMolWt(matched[-2])
            #filter stereoisomers, regioisomers and terminal acid/amide congeners
        if (m.HasSubstructMatch(matched[-1])
        or int(mw)==int(mw_lastmatch) or int(mw)==int(mw_seclastmatch)
        or int(mw)-int(mw_lastmatch)==0 or int(mw_lastmatch)-int(mw)==0
        or int(mw)-int(mw_lastmatch)==1 or int(mw_lastmatch)-int(mw)==1
        or int(mw)-int(mw_seclastmatch)==0 or int(mw_seclastmatch)-int(mw)==0
        or int(mw)-int(mw_seclastmatch)==1 or int(mw_seclastmatch)-int(mw)==1
        or int(mw)-int(mw_lastmatch)==13 or int(mw_lastmatch)-int(mw)==13):
            continue
        else:
            if m.HasSubstructMatch(patterns[0]) and a<m:
                a+=1
                matched.append(m)
            elif m.HasSubstructMatch(patterns[1]) and b<m:
                b+=1
                matched.append(m)
            elif a==m and b==m:
                if length==5:
                    raise ValueError
                else:
                    print('max reached', length)
                    length+=1
                    print('length changed', length)
                    a=1
                    b=1
            elif a==m or b==m:
                continue
    except ValueError:
        pass

print("a=",a, "b=",b)
print(length)

txtfile='/path/to/output.txt'
for mol in matched:
    smiles=Chem.MolToSmiles(mol)
    with open(txtfile, 'a') as f:
        f.write(smiles+"\n")

### filter by property

import cpmg.repository as repository
from rdkit import Chem
from rdkit.Chem import rdMolDescriptors, Descriptors
```

133

```python
import csv

def load_macrocycles(length, mol_type='macrocycle'):
    repo = repository.create_repository_from_string(mol_type)
    for mol in repo.find({'length': length}):
        yield mol


csv_file = "/path/to/output.csv"
csv_columns = ['slope', 'smiles', 'mw', 'tpsa', 'hbd']
n=0

matched=[Chem.MolFromSmiles('CCCCCCCCCC'), Chem.MolFromSmiles('CCCCCCCCC(C)=O')]
#arbitrary entries for first if loop

if __name__ == "__main__":
    length=3        # change to access trimers, tetramers, or pentamers
    with open(csv_file, 'w') as csvfile:
        writer=csv.DictWriter(csvfile, fieldnames=csv_columns)
        for macrocycle in load_macrocycles(length):
            n+=1
            m = Chem.MolFromSmiles(macrocycle['kekule'])
            mw = Chem.rdMolDescriptors.CalcExactMolWt(m)
            mw_lastmatch=Chem.rdMolDescriptors.CalcExactMolWt(matched[-1])
            mw_seclastmatch=Chem.rdMolDescriptors.CalcExactMolWt(matched[-2])
            tpsa = Descriptors.TPSA(m, includeSandP=True)
            slope = tpsa/mw
            if slope<0.2 or slope>0.3:
                continue
                #filter stereoisomers, regioisomers and terminal acid/amide congeners
            elif (m.HasSubstructMatch(matched[-1])
            or int(mw)==int(mw_lastmatch) or int(mw)==int(mw_seclastmatch)
            or int(mw)-int(mw_lastmatch)==0 or int(mw_lastmatch)-int(mw)==0
            or int(mw)-int(mw_lastmatch)==1 or int(mw_lastmatch)-int(mw)==1
            or int(mw)-int(mw_seclastmatch)==0 or int(mw_seclastmatch)-int(mw)==0
            or int(mw)-int(mw_seclastmatch)==1 or int(mw_seclastmatch)-int(mw)==1
            or int(mw)-int(mw_lastmatch)==13 or int(mw_lastmatch)-int(mw)==13
            or int(mw)-int(mw_seclastmatch)==13 or int(mw_seclastmatch)-int(mw)==13):
                continue
            else:
                hbd=Chem.Lipinski.NumHDonors(m)
                if (int(mw) <= 800 and tpsa <= 150 and hbd <=5):
                    print(n)
                    matched.append(m)
                    writer.writerow({'slope':slope, 'smiles':macrocycle['kekule'], 'mw':mw, 'tpsa':tpsa, 'hbd':hbd})
                else:
                    continue
print(n)
```

```python
csvfile.close()

### generate properties for a structure

from __future__ import print_function
from rdkit import Chem
from rdkit.Chem import rdMolDescriptors, Descriptors, Crippen, Lipinski

while True:

    smiles=input("input smiles: ")
    m=(Chem.MolFromSmiles(smiles))

    if smiles=="":
        break

    else:
        print("Rotatable bonds=",
            Chem.rdMolDescriptors.CalcNumRotatableBonds(m,strict=True))
        print("MW=",Chem.rdMolDescriptors.CalcExactMolWt(m))
        print("TPSA=", Descriptors.TPSA(m, includeSandP=True))
        print("TPSA/MW=", (Descriptors.TPSA(m,
includeSandP=True)/Chem.rdMolDescriptors.CalcExactMolWt(m)))
        print("logP=", Chem.Crippen.MolLogP(m))
        print("hbd=", Chem.Lipinski.NumHDonors(m))
        print("hba=", Chem.Lipinski.NumHAcceptors(m))

### running an array of CB++ jobs (re-optimized parameters from section 2.4 used here)

#!/bin/bash
#$ -cwd
#$ -N cbpp
#$ -o joblogD.$TASK_ID.txt
#$ -j y
#$ -l h_data=10G,h_rt=24:00:00
#$ -pe shared 8
#$ -t 1-n:1

set[1]=/path/to/file/with/smiles.txt
out[1]=/path/to/output.pdb

set[n]=/path/to/file/with/smiles.txt
out[n]=/path/to/output.pdb

echo "Job $JOB_ID.$SGE_TASK_ID started on:   " `hostname -s`
echo "Job $JOB_ID.$SGE_TASK_ID started on:   " `date `
echo " "
```

```bash
. /u/local/Modules/default/init/modules.sh
module load python/anaconda3
module load openbabel/2.4.1
source activate rdkit
export TMP_DIR="/path/to/tmp/"
export PYTHONPATH="/path/to/ConfBusterPlusPlus-master/":
/path/to/ConfBusterPlusPlus-master/bin/confbpp.sh --smiles $(cat ${set[$SGE_TASK_ID]}) -r 1 -m 3 -N 5
-n 15 -e 5 -o ${out[$SGE_TASK_ID]}

echo "Job $JOB_ID.$SGE_TASK_ID ended on:   " `hostname -s`
echo "Job $JOB_ID.$SGE_TASK_ID ended on:   " `date `
echo " "
```

```python
### generate shape metrics for a conformer

from __future__ import print_function
from rdkit import Chem
from rdkit.Chem import AllChem, Descriptors3D
import glob

pmis=[]
x=[]
y=[]
num=0

for pdb in glob.glob('/path/to.conformers/*.pdb'):
    num+=1
    m=Chem.MolFromPDBFile(pdb)
    i1=Chem.Descriptors3D.PMI1(m)
    i2=Chem.Descriptors3D.PMI2(m)
    i3=Chem.Descriptors3D.PMI3(m)

    if i3!=0:
        x=i1/i3
        y=i2/i3
        new_m={'num':num,'i1/i3':x,'i2/i3':y}
        pmis.append(new_m)
    else:
        print("no conformer at", num)


import csv
csv_columns=['num','i1/i3','i2/i3']

with open('/path/to/output.csv', 'w') as file:
    writer=csv.DictWriter(file, fieldnames=csv_columns)
```

```python
    writer.writeheader()
    for data in pmis:
        writer.writerow(data)
```

convert pdb to pdbqt with obabel

```bash
#!/bin/bash
#$ -cwd
#$ -N convert
#$ -o joblog.$JOB_ID
#$ -j y
#$ -l h_data=1G,h_rt=24:00:00
#$ -pe shared 1

. /u/local/Modules/default/init/modules.sh
module load python/anaconda3
module load openbabel/2.4.1

for dir in */ ; do
    cd "$dir"
    obabel *.pdb -opdbqt -O .pdbqt -p 7.4 -m
    cd ..
done
```

split multi-conformer pdbqt files into individual input files

```bash
#!/bin/bash
#$ -cwd
#$ -N split_pdbqt
#$ -o joblog.$JOB_ID
#$ -j y
#$ -l h_data=1G,h_rt=24:00:00
#$ -pe shared 1

for dir in */ ; do
    for f in /path/to/$dir*.pdbqt; do
        /path/to/autodock_vina_1_1_2_linux_x86/bin/vina_split --input $f
    done
done
```

vina.py – running vina with parallelization

```python
import time, sys
from multiprocessing import Pool
from subprocess import run
import glob
from os.path import join, splitext, basename
```

```python
from datetime import datetime

#the values for --cpu and pool processes were chosen based on literature precedent[1]

def f(lig):
    ligname = "%s" % (splitext(lig)[0])
    run(['/path/to/vina',
    '--cpu', '4',
    '--receptor', '/path/to/receptor.pdbqt',
    '--config', '/path/to/config.txt',
    '--ligand', lig,
    '--log', ligname+'.txt'])

if __name__ == "__main__":
    k=sys.argv[1]
    print("this job is running on folder path/%s" % k)
    pool = Pool(processes=5)          # start worker processes
    ligs = glob.glob("/path/to/ligand/library/%s/*.pdbqt" % k)

    pool.map(f, ligs)                 # run for each ligand
```

###submitting an array of jobs for running docking calculations in parallel

```bash
#!/bin/bash
#$ -cwd
#$ -N vina_run
#$ -o joblog.$JOB_ID.$TASK_ID.txt
#$ -j y
#$ -l h_data=8G,h_rt=24:00:00
#$ -pe shared 8
#$ -t 1-n:1

set[1]="$set_1"
set[n]="$set_n"

echo "Job $JOB_ID.$SGE_TASK_ID started on:   " `hostname -s`
echo "Job $JOB_ID.$SGE_TASK_ID started on:   " `date `
echo " "

. /u/local/Modules/default/init/modules.sh
module load python/anaconda3
python /path/to/vina.py ${set[$SGE_TASK_ID]}

echo "Job $JOB_ID.$SGE_TASK_ID ended on:   " `hostname -s`
echo "Job $JOB_ID.$SGE_TASK_ID ended on:   " `date `
echo " "
```

```
### compiling scores from docking

import sys
import glob
import csv
import numpy as np
import pandas as pd
import matplotlib
matplotlib.use('agg')
import matplotlib.pyplot as plt

data = []
not_out_files = []
values=[]
directories = glob.glob("/path/to/docked/output/files/*/")

for directory in directories:
    file_names = glob.glob("%s/*.pdbqt" % directory)
    print ('Found', len(file_names), 'pdbqt files in', directory)
    for file_name in file_names:
        file = open(file_name)
        lines = file.readlines()
        file.close()
        try:
            line = lines[1]
            score = float(line.split(':')[1].split()[0])
            datapoint={'score':score, 'file':file_name}
            data.append(datapoint)
            values.append(score)
        except:
            not_out_files.append(file_name)

csv_file = "/path/to/output/scores.csv"
csv_columns = ['score', 'file']

with open(csv_file, 'w') as csvfile:
    writer=csv.DictWriter(csvfile, fieldnames=csv_columns)
    for point in data:
        writer.writerow(point)

print("Total number of docked conformers = ", len(data))

df = pd.read_csv("/path/to/output/scores.csv")
data = df.iloc[:, 0]
bins = np.arange(min(data), max(data), 0.2)
print("Highest affinity = ", min(data))
print("Lowest affinity = ", max(data))
```

```
plt.hist(data, bins=bins, color='blue', alpha=0.5, edgecolor='black')
plt.title('Scores for run2')
plt.ylabel('Frequency')
plt.xlabel('Scores')
plt.xlim([min(data)-2, max(data)+2])
plt.savefig('/path/to/output/scores.png', dpi = 300)
```

References

(1)   Jaghoori, M. M.; Bleijlevens, B.; Olabarriaga, S. D. 1001 Ways to Run AutoDock Vina for Virtual Screening. *J. Comput. Aided. Mol. Des.* **2016**, *30* (3), 237–249.

Chapter 3 Experimental Appendix

Analytical Methods. NMR spectra were recorded on Bruker Advance spectrometers (300 MHz, 400 MHz, 500 MHz) and are reported as δ values in ppm relative to CDCl$_3$ (calibrated to 7.26 ppm in ^{1}H NMR and 77.16 ppm in ^{13}C NMR, unless otherwise indicated). Splitting patterns are abbreviated as follows: singlet (s), doublet (d), triplet (t), quartet (q), multiplet (m), broad (br), and combinations thereof. Column chromatography was conducted on silica gel 60 (240–400 mesh) purchased from Sillicycle. Thin-layer chromatography (TLC) was performed using pre-coated, glass-backed plates (silica gel 60 PF254, 0.25 mm) and visualized using a combination of UV and potassium permanganate staining. HPLC analyses were carried out using an Agilent 1200 HPLC system equipped with an Agilent Quadrupole 6130 ESI-MS detector. Mobile phase was prepared with 0.1% TFA.

Methyl (((9*H*-fluoren-9-yl)methoxy)carbonyl)-*L*-serinate (35). Methyl *L*-serinate hydrochloride (2.06 g, 13.29 mmol) and sodium bicarbonate (2.75 g, 32.7 mmol) were dissolved in water (30 mL) and cooled to 0 °C. A solution of Fmoc-succinimide (4.40 g, 13.06 mmol) in 1,4 dioxane (30 mL) was added over 30 minutes. The ice bath was subsequently removed, and the mixture was stirred for another hour. 15 mL water was added, and the solution was extracted with ethyl acetate (x3). The combined organics were washed with 0.2 M HCl (15 mL), brine (15 mL), dried of MgSO$_4$, filtered, and concentrated *in vacuo*. Recrystallization in CH$_2$CL$_2$ afforded the product as a white solid (4.33 g, 97%); R$_f$ 0.2 (1:1 Hex: EtOAc). ^{1}H NMR (400 MHz, CDCl$_3$) δ 7.77 (d, *J* = 7.5 Hz, 2H), 7.67-7.55 (m, 2H), 7.41 (t, *J* = 7.4 Hz, 2H), 7.32 (t, *J* = 4.6 Hz, 2H), 5.77-5.59 (m, 1H), 4.52-4.35 (m, 3H), 4.23 (t, *J* = 6.8 Hz, 1H), 4.07-3.87 (m, 2H), 3.80 (s, 3H).
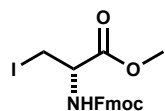
Methyl (*S*)-2-((((9*H*-fluoren-9-yl)methoxy)carbonyl)amino)-3-iodopropanoate (36). PPh$_3$ (0.88 g, 3.4 mmol) and imidazole (0.29 g, 4.3 mmol) were dissolved in dry CH$_2$CL$_2$ (9 mL) under argon and cooled on an ice bath. Iodine (0.91 g, 3.6 mmol) was added and the mixture was stirred at room temperature for 10 minutes. The resulting yellow suspension was once again cooled on an ice bath, and **35** (0.85 g, 2.4 mmol) in dry CH$_2$CL$_2$ (24 mL) was added. The mixture was allowed to stir over the ice bath for 2 hours. The reaction mixture was filtered and washed with water (x2), sodium thiosulfate (x2) and brine (x1), dried over MgSO$_4$ and concentrated. Recrystallization in EtOH/ether afforded the product as a white solid. (0.99 g, 92%); R$_f$ 0.8 (1:1 Hex: EtOAc). ^{1}H NMR (400 MHz, CDCl$_3$) δ 7.77 (d, *J* = 7.4 Hz, 2H), 7.61 (d, *J* = 7.5 Hz), 7.41 (t, *J* = 7.7 Hz, 2H), 7.37-7.29 (m, 2H), 5.65 (d, *J* = 6.9 Hz, 1H), 4.64-4.53 (m, 1H), 4.50-4.32 (m, 1H), 4.25 (t, *J* = 7.2 Hz, 1H), 3.83 (s, 3H), 3.61 (d, *J* = 3.4 Hz, 2H); ^{13}C NMR (100 MHz, CDCl$_3$) δ 169.7, 155.4, 143.7, 143.6, 141.3, 127.8, 127.1, 125.1, 120.1, 67.4, 54.1, 53.2, 47.1, 7.4.

Methyl (*S*)-2-((((9*H*-fluoren-9-yl)methoxy)carbonyl)amino)-3-(thiophen-3-yl)propanoate (37). *A. Zinc insertion.* Zinc dust (63 mg, 0.95 mmol) and iodine (4 mg, 14 μmol) were added to a flask equipped with a stir bar. The flask was evacuated, heated over a Bunsen burner and flushed with argon 3 times. The iodide derivative **36** (216 mg, 0.48 mmol) was dissolved in anhydrous DMF (~1.3 cm^3/mmol) and transferred *via* syringe to the reaction mixture, which was previously cooled to 0 °C. The reaction was stirred at room temperature and monitoring the reaction by TLC (3:2 Hex:EtOAc) showed full consumption of starting material in 90 minutes. *B. Pd Catalyzed Cross Coupling.* Pd[P(t-Bu)$_3$)]$_2$ (8 mg, 14 μmol) and SPhos ligand (20 mg, 48 μmol) were weighed into a flame dried flask in a glove bag under argon atmosphere. The flask was kept under an argon atmosphere and 3-

bromothiophene (0.04 ml, 0.48 mmol) was added to it. The zinc insertion product from part A was syringed into this flask under argon atmosphere. The flask was heated at 50 °C for 5 hours and subsequently stirred at room temperature overnight. The reaction mixture was cooled to room temperature, diluted with EtOAc and washed with water to remove DMF. The organic layer was dried over $MgSO_4$, filtered, concentrated, and purified by column chromatography (4:1 Hex: EtOAc) to afford the product as a yellow oil (94 mg, 48%); R_f 0.2 (4:1 Hex: EtOAc). ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.77 (d, J = 7.5 Hz, 2H), 7.58 (t, J = 6.2 Hz, 2H), 7.41 (t, J = 7.5 Hz, 2H), 7.32 (tt, J = 7.4, 1.4 Hz, 2H), 7.28-7.25 (m, 1H), 6.94 (br s, 1H), 6.83 (d, J = 4.6 Hz, 1H), 4.70-4.59 (m, 1H), 4.51-4.43 (m, 1H), 4.43-4.32 (m, 1H), 4.26-418 (m, 1H), 3.74 (s, 3H), 3.24-3.10 (m, 2H); ^{13}C NMR (125 MHz, $CDCl_3$) 172.0, 155.7, 144.0, 143.8, 141.5, 135.9, 128.4, 127.9, 127.2, 126.2, 125.2, 123.0, 120.1, 67.0, 54.4, 52.6, 47.3, 32.9; HRMS (ESI) calculated for $C_{23}H_{21}NO_4S$ [M+H]$^+$ 408.1264, found 408.1264.
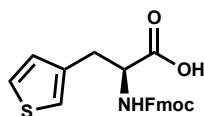
(S)-2-(((((9H-fluoren-9-yl)methoxy)carbonyl)amino)-3-(thiophen-3-yl)propanoic acid (38). MgI_2 (334 mg, 1.2 mmol) and **37** (49 mg, 0.12 mmol) were weighed into a microwave vial equipped with a stir bar in a glove bag under argon atmosphere. Anhydrous THF (2.0 mL) was added to the vial and the sealed reactor was heated at 120 °C for 1 hour under microwave irradiation with stirring. A $Na_2S_2O_3$ aq. solution (0.1 M) was added and the resulting mixture was extracted with EtOAc (x3). The organic layer was washed with brine, dried over $MgSO_4$, filtered, and concentrated *in vacuo*. Purification by column chromatography (15:1 CH_2CL_2:MeOH) afforded the desired product (23 mg, 49%); R_f 0.1 (15:1 CH_2CL_2:MeOH). ^{1}H NMR (400 MHz, MeOD-d_4) δ 7.78 (d, J = 7.4 Hz, 2H), 7.64-7.56 (m, 2H), 7.38 (t, J = 7.5 Hz, 2H), 7.33-7.25 (m, 3H), 7.06 (br s, 1H), 6.97 (d, J = 4.8 Hz, 1H), 4.42-4.30 (m, 2H), 4.28-4.20 (m, 1H), 4.17 (t, J = 6.8 Hz, 1H), 3.20 (dd, J = 14.3, 4.3 Hz, 1H), 3.02 (dd, J = 14.3, 8.6 Hz, 1H); IR (CH_2Cl_2) v 3320*br*, 2920*w*, 1700*s*, 1600*m*, 1520*m*, 1440*s*, 1420*s* cm^{-1}; HPLC/MS [M+H]$^+$ 394.1.
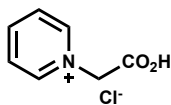
1-(carboxymethyl)pyridin-1-ium chloride (45). A mixture of pyridine (8.10 mL, 100 mmol) and chloroacetic acid (9.45 g, 100 mmol) in EtOAc (20 mL) was stirred at reflux for 2 hours. After cooling to ambient temperature, the formed precipitate was filtered, washed with EtOAc and dried to afford the product as a pale-yellow solid (6.86 g, 40%). ^{1}H NMR (400 MHz, DMSO-d_6) δ 9.06 (dd, J = 6.7, 1.3 Hz 2H), 8.68 (tt, J = 7.8, 1.3 Hz, 1H), 8.22 (td, J = 6.5, 1.0 Hz 2H), 5.58 (s, 2H).
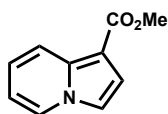
Methyl indolizine-1-carboxylate (46). *A. Activation of MnO_2 by azeotropic removal of water.* MnO_2 (15 g, 0.173 mol) and benzene (90 mL) were added to a 250 mL flask fitted with a Dean-Stark trap, condenser, and magnetic stirrer bar. Water was removed by vigorous azeotropic distillation with stirring for over 6 hours until separation of water was deemed to be complete. The apparatus was disassembled, and the flask was placed under argon atmosphere. Yields are optimal when the reagent is used immediately in the next step, but it may be kept under argon atmosphere in a desiccator for a few days without significant loss of activity. *B. Oxidative cycloaddition.* Pyridinium chloride, **45**, (3.81 g, 22 mmol) was added to the flask containing activated MnO_2, kept under argon atmosphere. Benzene (90 mL) and Et_3N (3.4 mL, 24 mmol) were added to the flask. A condenser was fitted to the flask and the mixture was heated at reflux. Monitoring by TLC (5:1 Hex:EtOAc) indicated complete consumption of starting material after 4 hours. After cooling to room temperature, the mixture was filtered over celite and washed with acetone. The combined filtrates were evaporated to give a residue which was purified by column chromatography to afford the product as a yellow oil (1.88 g, 42%); R_f 0.4 (5:1 Hex:EtOAc). ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.17 (d, J = 9.1 Hz, 1H),

7.99 (dt, J = 6.9, 1.1 Hz, 1H), 7.23 (dd, J = 6.8, 2.9 Hz, 1H), 7.04 (ddd, J = 9.2, 6.8, 1.1 Hz, 1H), 6.70 (td, J = 6.8, 1.2 Hz, 1H), 3.89 (s, 3H); ^{13}C NMR (100 MHz, CDCl$_3$) δ 165.4, 135.8, 126.0, 122.3, 119.9, 116.1, 113.7, 112.4, 103.6, 50.9.
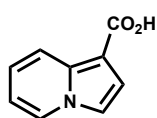
Indolizine-1-carboxylic acid (47). A suspension of indolizine carboxylate, **46** (1.53 g, 8.7 mmol) in 10% NaOH solution (108 mL) was stirred at 50 °C until starting material was fully consumed, as monitored by TLC. The reaction mixture was neutralized to pH=1 with conc. HCl and extracted with EtOAc. The organic layer was washed with brine, dried over MgSO$_4$, filtered and concentrated *in vacuo* to afford the product as a grey-blue solid (1.40 g, 100%); R$_f$ 0.1 (5:1 Hex:EtOAc). ^{1}H NMR (400 MHz, CDCl$_3$) δ 8.21 (d, J = 9.1 Hz, 1H), 8.02 (dt, J = 6.9, 1.1 Hz, 1H), 7.38 (d, J = 3.0 Hz, 1H), 7.25 (d, J = 3.5 Hz, 1H), 7.10 (ddd, J = 9.0, 6.7, 1.1 Hz, 1H), 6.75 (td, J = 6.7, 1.1 Hz, 1H).
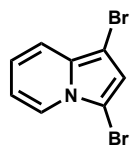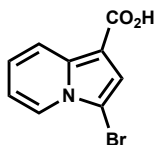
3-bromoindolizine-1-carboxylic acid (50) and 1,3-dibromoindolizine (51). To a slurry of **46** (192 mg, 1.19 mmol) in anhydrous DMF (12 mL) was added NaHCO$_3$ (301 mg, 3.58 mmol) and N-bromo succinimide (223 mg, 1.25 mmol). The mixture was stirred at room temperature until starting material was fully consumed, as indicated by TLC. The solution was partitioned between water and EtOAc. The extracted organics were washed with water and brine, dried over MgSO$_4$, filtered and concentrated. Purification by column chromatography afforded **x** (R$_f$ 0.1, 5:1 Hex:EtOAc) and **x'** (R$_f$ 0.8 Hex:EtOAc). **x** ^{1}H NMR (400 MHz, CDCl$_3$) δ 8.26 (dt, J = 9.0, 1.1, 1H), 8.15 (dt, J = 7.0, 1.0 Hz, 1H), 7.37 (s, 1H), 7.17 (ddd, J = 15.8, 7.8, 1.0 Hz, 1H), 6.91 (td, J = 6.9, 1.2 Hz, 1H); HRMS (ESI) calculated for C$_9$H$_6$BrNO$_2$ [M+H]$^+$ 239.9655, found 239.9647. **x'** ^{1}H NMR (400 MHz, CDCl$_3$) δ 7.94 (d, J = 7.1 Hz, 1H), 7.36 (d, J = 9.1 Hz, 1H), 6.83 (s, 1H), 6.80 (ddd, J = 15.7, 6.6, 0.8 Hz, 1H), 6.65 (td, J = 6.3, 1.1 Hz, 1H); HRMS (ESI) calculated for C$_8$H$_5$Br$_2$N [M+H]$^+$ 273.8862, found 273.8865.
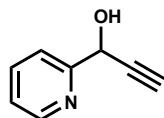
1-(pyridin-2-yl)prop-2-yn-1-ol (53). To a flame dried flask was added freshly titrated ethynylmagnesium bromide (13.75 ml, 0.4 M in THF). The solution was cooled to –78 °C and a solution of 2-picolyl aldehyde (0.48 mL, 5 mmol) was added dropwise with stirring. The solution was stirred at –78 °C for 1 hour, then allowed to warm to room temperature and stirred for another hour. The reaction was poured into sat. NH$_4$Cl (30 mL) and extracted with EtOAc. The organic phase was washed with sat. NH$_4$Cl and brine, dried over MgSO$_4$ and concentrated. The residue was purified over silica gel with 1:1 Hex:EtOAc to afford the product as a yellow oil (0.54 g, 81%); R$_f$ 0.3 (9:1 Hex:EtOAc). ^{1}H NMR (400 MHz, CDCl$_3$) δ 8.57 (d, J = 4.5 Hz, 1H), 7.76 (td, J = 7.7, 1.7 Hz, 1H), 7.52 (dd, J = 7.9, 0.5 Hz), 7.30-7.27 (m, 1H), 5.49 (d, J = 2.2, 1H), 2.57 (d, J = 2.3 Hz, 1H).
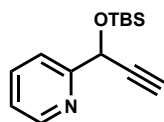
2-(1-((*tert*-butyldimethylsilyl)oxy)prop-2-yn-1-yl)pyridine (54). A solution of **53** (0.54 g, 4.05 mmol), in DCM (1 mL) was added to a stirred solution of TBSCl (0.61 g, 4.05 mmol) and imidazole (0.30 g, 4.46 mmol) in DCM (4.7 mL). The reaction was stirred at room temperature for 2 hours. The solvent was removed *in vacuo* and the product was purified by column chromatography to afford a yellow oil (0.83 g, 83%); R$_f$ 0.9 (9:1 Hex:EtOAc). ^{1}H NMR (400 MHz, CDCl$_3$) δ 8.58-8.53 (m, 1H), 7.74 (td, J = 7.7, 1.8 Hz, 1H), 7.63 (d, J = 7.9 Hz, 1H), 7.24-7.18 (m, 1H), 5.56 (d, J = 2.2 Hz, 1H), 2.53 (d, J = 2.3 Hz, 1H), 0.95 (s, 9H), 0.21 (s, 3H), 0.15 (s, 3H).
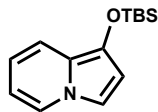
1-((*tert*-butyldimethylsilyl)oxy)indolizine (55). AuBr₃ (29 mg, 0.19 mmol) was weighed out under an argon atmosphere. A solution of **54** (270 mg, 1.09 mmol) in anhydrous toluene (2.4 mL) was added and the reaction was stirred at 50 °C under an argon atmosphere. Starting material was not fully consumed but TLC monitoring indicated no change between 2 and 3 hours. The solvent was concentrated and the reaction was purified by column chromatography (9:1 Hex:EtOAc) to afford 33% recovered starting material, and the product as a yellow oil (65 mg, 24%); R_f 0.8 (9:1 Hex:EtOAc) ^{1}H NMR (400 MHz, CDCl₃) δ 7.64 (d, *J* = 7.0 Hz, 1H), 7.23 (d, *J* = 9.1 Hz, 1H), 6.98 (d, *J* = 2.8 Hz, 1H), 6.44-6.35 (m, 1H), 6.33 (d, *J* = 2.7 Hz, 1H), 6.26 (t, *J* = 6.6 Hz, 1H), 1.01 (s, 9H), 0.18 (s, 6H).
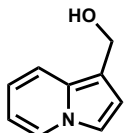
indolizin-1-ylmethanol (56a). 47 (188 mg, 1.16 mmol) was dissolved in dry THF (3.5 mL) under argon. Borane-DMS complex (0.58 mL, 2M in THF) was added dropwise by syringe. The reaction was refluxed for 3 hours. After cooling to room temperature, the mixture was poured into 2N NaOH solution and stirred for 0.5 hour. The mixture was extracted twice with EtOAc. The combined organics were washed with brine, dried over MgSO₄ and concentrated to afford the product (62 mg, 36%). The compound is not bench-stable and rapidly decomposes into an indigo oil. ^{1}H NMR (400 MHz, CDCl₃) δ 7.85 (d, *J* = 7.7 Hz, 1H), 7.62 (6.3 Hz, 1H), 7.3 (d, *J* = 9.0 Hz), 7.24-7.20 (m, 1H), 6.60-6.55 (m, 1H), 6.43-6.40 (m, 1H), 4.29 (s, 2H).
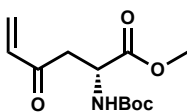
methyl (*R*)-2-((*tert*-butoxycarbonyl)amino)-4-oxohex-5-enoate (58). Boc-iodo- *L*-alanine methyl ester, **57**, was prepared by the same procedure as for **36**. *A. Zinc insertion.* Zn dust (294 mg, 4.50 mmol) was weighed into a flask and flame dried under vacuum. After cooling to room temperature, the flask was back filled with argon and evacuated three times. Dry toluene (1 mL) and TMSCl (12 µL, 94 µM) were added to the flask and the solution was stirred for 30 minutes. A solution of **57** (494 mg, 1.50 mmol) in a mixture of dry toluene (1 mL) and dimethyl acetamide (0.4 mL) was added to the flask dropwise. Consumption of starting material was monitored by TLC (2:1 ether:EtOAc). *B. Pd Catalyzed Cross Coupling.* The zinc insertion product was syringed into a flask containing acryloyl chloride (0.16 mL, 2 mmol), Pd(OAc)₂ (22 mg, 0.1 mmol) and PPh₃ (53 mg, 0.02 mmol) under argon and the reaction was stirred for 3 hours. The mixture was partitioned between sat. NH₄Cl and EtOAc. The extracted organics were washed with brine, dried over MgSO₄ and concentrated. Purification by column chromatography (100% Hex → 15% EtOAc in Hex) afforded the product as a yellow oil (130 mg, 34%); R_f 0.6 (20:3 Hex:EtOAc). ^{1}H NMR (400 MHz, CDCl₃) δ 6.33 (dd, *J* = 17.7, 10.1 Hz, 1H), 6.24 (dd, *J* = 17.7, 1.4 Hz, 1H), 5.92 (dd, *J* = 10.1, 1.4 Hz, 1H), 5.50 (d, *J* = 8.2 Hz, 1H), 4.63-4.49 (m, 1H), 3.73 (s, 3H), 3.14 (dd, *J* = 18.0, 4.2 Hz, 1H), 3.14 (dd, *J* = 18.0, 4.2 Hz, 1H), 1.44 (s, 9H).
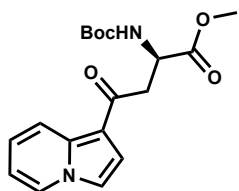
methyl (*R*)-2-((*tert*-butoxycarbonyl)amino)-4-(indolizin-1-yl)-4-oxobutanoate (59). MnO₂ was activated as described for **46**. Enone **58** (130 mg, 0.50 mmol), **45** (43 mg, 0.25 mmol) and Et₃N (38 µL, 0.28 mmol) were quickly added directly into a flask containing activated MnO₂ (174 mg, 2.0 mmol) in benzene (25 mL). The reaction was refluxed until starting material was fully consumed, as indicated by TLC. The reaction was cooled to room temperature, filtered over celite, concentrated and purified by column chromatography (20% → 50% EtOAc in Hex) to afford the product as yellow oil (20 mg, 23%); R_f 0.5 (1:1 Hex:EtOAc). ^{1}H NMR (400 MHz, CDCl₃) δ 8.38 (d, *J* = 9.0 Hz, 1H), 8.02 (dt, *J* = 6.8, 1.1 Hz, 1H), 7.21 (d, *J* = 3.0 Hz, 1H), 7.18-7.12 (m, 2H), 6.79 (td, *J* = 6.8, 1.2 Hz, 1H), 5.77 (d, *J* = 8.8 Hz, 1H), 4.69-4.60 (m, 1H), 3.75 (s, 3H), 3.66 (dd, *J* = 17.2, 4.3 Hz, 1H), 3.41 (dd, *J* = 17.2, 4.1 Hz, 1H), 1.43 (s, 9H).

144
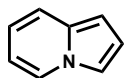
Indolizene (61). 47 (2.35 g, 14.6 mmol) is dissolved in ethanolamine (607 mL) and brought to reflux. Upon consumption of starting material, the reaction is quenched by pouring over ice-cold water. The resulting solution is extracted with ether (x3), the organics are washed with brine, dried over MgSO$_4$ and concentrated to afford the product as a yellow crystalline solid (1.43 g, 84%). The compound is taken forward without further purification. ^{1}H NMR (500 MHz, CDCl$_3$) δ 7.91 (d, J = 7.0 Hz, 1H), 7.37 (d, J = 9.0 Hz, 1H), 7.31 (br s, 1H), 6.79 (br s, 1H), 6.64 (ddd, J = 9.0, 6.5, 1.0 Hz, 1H), 6.47-6.39 (m, 2H); ^{13}C NMR (125 MHz, CDCl$_3$) δ 132.7, 125.1, 119.1, 116.6, 113.4, 112.3, 110.0, 98.6.
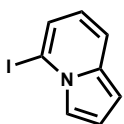
5-iodoindolizine (62). To a stirred solution of indolizine, **61**, (1.47 g, 12.5 mmol) and TMEDA (2.06 mL, 13.8 mmol) in anhydrous THF (42 mL) was added freshly titrated n-BuLi (2.27 M solution in hexanes, 6 mL) dropwise. The reaction flask was warmed to –20 °C and maintained at that temperature for 2 hours. The mixture was cooled back down to –80 °C and a solution of I$_2$ (3.17 g, 12.5 mmol) in anhydrous THF (22 mL) was added. The mixture was warmed to room temperature and treated with sat. NH$_4$Cl. The organic layer was separated, and the aqueous layer was extracted with hexanes. The combined organics were dried over MgSO$_4$, concentrated and purified by column chromatography (100% hexanes) to afford the product as a green oil (2.0 g, 66%); R$_f$ 0.6 (hexanes); ^{1}H NMR (500 MHz, CDCl$_3$) δ 7.59-7.55 (m, 1H), 7.41 (d, J = 8.9 Hz, 1H), 7.04 (dd, J = 6.8, 1.0 Hz, 1H), 6.88 (dd, J = 3.9, 2.9 Hz, 1H), 6.69 (dd, J = 6.69, 1.5 Hz, 1H), 6.41 (dd, J = 8.9, 6.8 Hz, 1H); ^{13}C NMR (125 MHz, CDCl$_3$) δ 133.5, 122.8, 119.0, 117.9, 117.1, 113.2, 101.8, 86.2. HRMS (ESI) calculated for C$_8$H$_7$IN [M+H]$^+$ 243.9618, found 243. 9617.
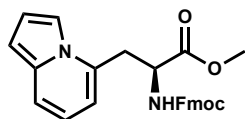
Methyl (*S*)-2-((((9*H*-fluoren-9-yl)methoxy)carbonyl)amino)-3-(indolizin-5-yl)propanoate (63). *A. Zinc insertion.* Zinc dust (1.07 g, 16.5 mmol) and iodine (63 mg, 30 μmol) were added to a flask equipped with a stir bar. The flask was evacuated, heated over a Bunsen burner and flushed with argon 3 times. The iodide derivative **36** (3.7 g, 8.23 mmol) was dissolved in anhydrous DMF (~1.3 cm^3/mmol) and transferred *via* syringe to the reaction mixture, which was previously cooled to 0 °C. The reaction was stirred at room temperature and monitoring the reaction by TLC (3:2 Hex:EtOAc) showed full consumption of starting material in 90 minutes. *B. Pd Catalyzed Cross Coupling.* **62** (2 g, 8.23 mmol), Pd(OAc)$_2$ (185 mg, 0.82 mmol) and SPhos ligand (1.11 g, 2.7 mmol) was added to a flask equipped with a stir bar. The zinc insertion product from part A was syringed into this flask under argon atmosphere. The flask was heated at 50 °C for 5 hours and subsequently stirred at room temperature overnight. The reaction mixture was cooled to room temperature, diluted with EtOAc and washed with water to remove DMF. The organic layer was dried over MgSO$_4$, filtered, concentrated, and purified by column chromatography (3:2 Hex: EtOAc) to afford the product as a greenish-white solid foam (2.12 mg, 58%); R$_f$ 0.3 (4:1 Hex: EtOAc). ^{1}H NMR (400 MHz, MeOD-d_4) δ 7.75 (d, J = 7.6 Hz, 2H), 7.57 (t, J = 7.0 Hz, 2H), 7.43 (br s, 1H), 7.39-7.30 (m, 3H), 7.30-7.21 (m, 2H), 6.80 (dd, J = 3.8, 2.8 Hz, 1H), 6.59 (dd, J = 8.9, 6.6 Hz, 1H), 6.44 (dd, J = 3.9, 1.0 Hz, 1H), 6.38 (d, J = 6.5 Hz, 1H), 4.71 (dd, J = 14.9, 9.6 Hz, 1H), 4.29-4.22 (m, 1H), 4.19-4.09 (m, 1H), 3.70 (s, 3H), 3.42 (dd, J = 15.0, 5.0 Hz, 1H), 3.20 (dd, J = 15.0, 9.7 Hz, 1H); ^{13}C NMR (100 MHz, MeOD-d_4) δ 173.5, 158.3, 145.1, 142.5, 135.1, 132.9, 128.7, 128.1, 126.2, 126.1, 120.9, 118.9, 117.6, 114.8, 111.6, 110.4, 100.9, 68.0, 52.9, 51.8, 50.9, 35.5. HRMS (ESI) calculated for C$_{27}$H$_{24}$N$_2$O$_4$ [M+H]$^+$ 441.1808, found 441.1803.
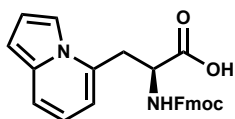
(*S*)-2-((((9*H*-fluoren-9-yl)methoxy)carbonyl)amino)-3-(indolizin-5-yl)propanoic acid (39). MgI$_2$ (3.43 g, 12.3 mmol) and **63** (544 mg, 1.2 mmol) were measured into a flask equipped with a stir bar in a glove bag under argon atmosphere. EtOAc (20 mL) was added to the flask, which was fitted

with a condenser and heated to reflux until starting material was fully consumed, as monitored by TLC. A $Na_2S_2O_3$ aq. solution (0.1 M) was added and the resulting mixture was extracted with EtOAc (x3). The organic layer was washed with brine, dried over $MgSO_4$, filtered, and concentrated *in vacuo*. The product was afforded as a green oil and used directly in the next step without further purification (0.52 g, 99%); R_f 0.6 (15:1 CH_2CL_2:MeOH). ^{1}H NMR (400 MHz, MeOD-d_4) δ 7.73 (d, *J* = 7.6 Hz, 2H), 7.55 (t, *J* = 6.7 Hz, 2H), 7.37-7.30 (m, 2H), 7.30-7.18 (m, 4H), 6.80-6.72 (m, 1H), 6.53 (dd, *J* = 8.9, 6.6 Hz, 1H), 6.45-6.38 (m, 2H), 4.56 (dd, *J* = 9.4, 4.0 Hz, 1H), 4.22 (dd, *J* = 9.4, 6.1 Hz, 1H), 4.13-4.02 (m, 2H), 3.66-3.55 (m, 1H), 3.54-3.44 (m, 1H), 3.10 (dd, *J* = 15.0, 9.5 Hz, 1H); ^{13}C NMR (100 MHz, MeOD-d_4) δ 172.9, 158.3, 145.2, 145.2, 142.5, 134.9, 134.4, 128.7, 128.1, 128.1, 126.3, 126.2, 120.8, 117.6, 114.2, 111.2, 100.4, 67.8, 54.0, 47.6, 36.8. HRMS (ESI) calculated for $C_{26}H_{22}N_2O_4$ [M+H]+ 427.1652, found 427.1644.
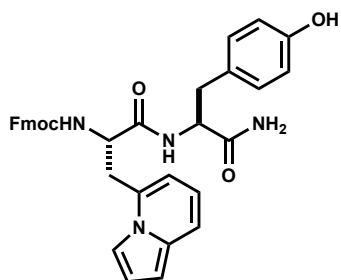
(9H-fluoren-9-yl)methyl ((S)-1-(((S)-1-amino-3-(4-hydroxyphenyl)1-oxopropan-2-yl)amino)-3-(indolizin-5-yl)-1-oxopropan-2-yl)carbamate (S1). 63 (472 mg, 1.1 mmol), *L*-tyrosinamide hydrochloride (264 mg, 1.2 mmol) and HBTU (455 mg, 1.2 mmol) were measured into a flask. Anhydrous DMF (3.3 mL) and DIPEA (0.4 mL, 2.3 mmol) were added, and the resulting mixture was stirred at room temperature until starting material was fully consumed, as monitored by TLC. The reaction mixture was partitioned between EtOAc and sat. $NaHCO_3$. The organic phase was washed with sat. $NaHCO_3$ (x2), sat. NH_4Cl (x3), brine (x1), dried over $MgSO_4$, filtered and concentrated to afford the product as a white solid (0.63 g, 97%); R_f 0.6 (15:1 CH_2CL_2:MeOH). MS (ESI) *m/z* calculated for $C_{35}H_{33}N_4O_5$ [M+H]+ 589.2, found 589.3.
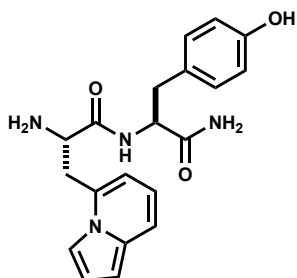
(S)-2-amino-N-((S)-1-amino-3-(4-hydroxyphenyl)-1-oxopropan-2-yl)-3-(indolizin-5-yl)propenamide (S2). S1 (627 mg, 1.07 mmol), octanethiol (1.85 mL, 10.65 mmol), DBU (0.18 mL, 1.17 mmol) and THF (14 mL) were stirred in a flask at room temperature until starting material was fully consumed, as monitored by TLC. The mixture was concentrated *in vacuo* and purified by column chromatography (100% $CHCl_3$ → 10% MeOH in $CHCl_3$) to afford the product as a white solid (380 mg, 97%); R_f 0.5. MS (ESI) *m/z* calculated for $C_{20}H_{23}N_4O_3$ [M+H]+ 367.2, found 367.2.
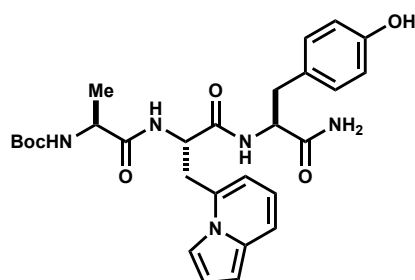
***tert*-butyl((S)-1-(((S)-1-(((S)-1-amino-3-(4-hydroxyphenyl)-1-oxopropan-2-yl)amino)-3-(indolizin-5-yl)-1-oxopropan-2-yl)amino)-1-oxopropan-2-yl)carbamate (S3). S2** (300 mg, 0.82 mmol), Boc-Ala-OH (187 mg, 0.90 mmol) and HBTU (342 mg, 0.90 mmol) were measured into a flask. Anhydrous DMF (3 mL) and DIPEA (0.16 mL, 0.90 mmol) were added, and the reaction mixture was stirred at room temperature until starting material was fully consumed, as indicated by TLC monitoring. The reaction mixture was partitioned between EtOAc and sat. $NaHCO_3$. The organic phase was washed with sat. $NaHCO_3$ (x2), sat. NH_4Cl (x3), brine (x1), dried over $MgSO_4$, filtered and concentrated. Purification by column chromatography (100% CH_2Cl_2 → 10% MeOH in CH_2Cl_2) afforded the product as a pale yellow solid (0.32 g, 72%); R_f 0.6 (15:1 CH_2CL_2:MeOH). ^{1}H NMR (400 MHz, MeOD-d_4) δ 9.16 (s, 1H), 8.16 (d, *J* = 8.0 Hz, 1H), 7.97 (d, *J* = 7.8 Hz, 1H), 7.51 (s, 1H), 7.35 (d, *J* = 8.9 Hz, 1H), 7.28 (br s, 1H), 7.08 (br s, 1H), 6.99 (d, *J* = 8.5 Hz, 2H), 6.90 (d, *J* = 6.9 Hz 1H), 6.82 (dd, *J* = 3.9, 2.8 Hz, 1H), 6.68-6.62 (m, 1H), 6.62-6.57

(m, 1H), 6.45 (dd, J = 3.9, 1.3 Hz, 1H), 6.34 (d, J = 6.6 Hz, 1H), 4.78-4.64 (m, 1H), 4.41-4.29 (m, 1H), 3.96-3.80 (m, 1H), 3.22-3.15 (m, 1H), 3.05 (dd, J = 15.6, 9.3 Hz, 1H), 2.88 (dd, J = 14.2, 5.4 Hz, 1H), 2.73 (dd, J = 13.9, 8.8 Hz, 1H), 1.36 (s, 9H), 1.08 (d, J = 7.0 Hz, 3H). MS (ESI) m/z calculated for $C_{38}H_{36}N_5O_6$ [M+H]$^+$ 538.3, found 538.3.
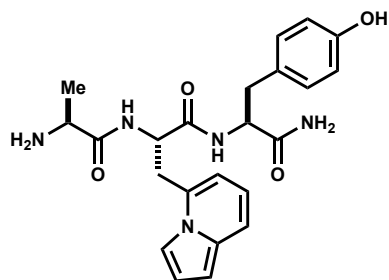
(S)-N-((S)-1-amino-3-(4-hydroxyphenyl)-1-oxopropan-2-yl)-2-((S)-2-aminopropanamido)-3-(indolizin-5-yl)propenamide (64). **S3** (318 mg, 0.59 mmol) and TFA (1.4 mL, 17.7 mmol) in CH$_2$Cl$_2$ (6.8 mL) was stirred at room temperature until the starting material was fully consumed, as monitored by TLC. Upon completion, the reaction mixture was concentrated *in vacuo* and used directly in the next step without further purification.
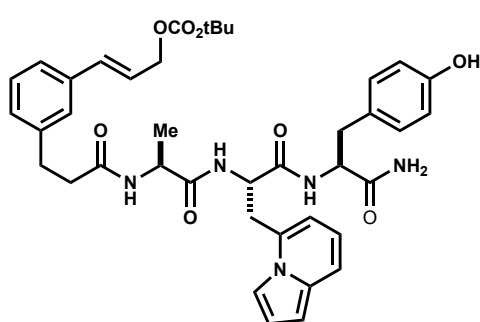
(E)-3-(3-(3-(((S)-1-(((S)-1-(((S)-1-amino-3-(4-hydroxyphenyl)-1-oxopropan-2-yl)amino)-3-(indolizin-5-yl)-1-oxopropan-2-yl)amino)-1-oxopropan-2-yl)amino)-3-oxopropyl)phenyl)allyl *tert*-**butyl carbonate (65).** **64** (0.59 mmol), **G1** (238 mg, 0.59 mmol) and anhydrous DMF (6 mL) were added to a flask. DIPEA (0.41 mL, 2.36 mmol) was added dropwise and the resulting mixture was stirred at room temperature until starting material was fully consumed, as indicated by TLC. The reaction mixture was concentrated *in vacuo*, the residue was redissolved in MeCN/CHCl$_3$ 1:3 and purified by column chromatography (100% CHCl$_3$ → 10% MeOH in CHCl$_3$) to afford the product as a bright pink-purple solid (387 mg, 90%); R$_f$ 0.9 (15:1 DCM:MeOH). ^{1}H NMR (400 MHz, DMSO-d_6) δ 9.17 (s, 1H), 8.28 (d, J = 7.7 Hz), 8.05 (d, J = 7.0 Hz, 1H), 7.91 (d, J = 8.3 Hz, 1H), 7.50 (br s, 1H), 7.40-7.33 (m, 1H), 7.30-7.20 (m, 4H), 7.13-7.06 (m, 2H), 7.05-6.97 (m, 2H), 6.82 (dd, J = 13.9, 4.9 Hz, 1H), 6.68-6.57 (m, 4H), 6.46 (dd, J = 3.8, 1.2 Hz, 1H), 6.38-6.27 (m, 2H), 4.72-4.62 (m, 2H), 4.41-4.30 (m, 1H), 4.25-4.17 (m, 1H), 3.49-3.39 (m, 1H), 3.22-3.12 (m, 1H), 3.05 (dd, J = 15.9, 9.5 Hz, 1H), 2.92 (dd, J = 13.9, 4.9 Hz, 1H), 2.82-2.74 (m, 3H), 2.44-2.36 (m, 2H), 1.43 (s, 9H), 1.11 (d, J = 7.12 Hz, 3H). MS (ESI) m/z calculated for $C_{40}H_{48}N_5O_8$ [M+H]$^+$ 726.35, found 726.36.
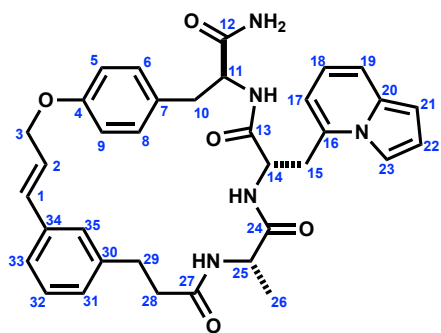
(11S,14S,17S,E)-14-(indolizin-5-ylmethyl)-11-methyl-9,12,15-trioxo-2-oxa-10,13,16-triaza-1(1,4),6(1,3)-dibenzenacyclooctadecaphan-4-ene-17-carboxamide (69). **65** (127 mg, 0.18 mmol) is dissolved in anhydrous DMF and the solution is sparged with argon for 15 minutes. Pd(PPh$_3$)$_4$ is added and the solution is stirred at room temperature until starting material is fully consumed, as determined by TLC. The reaction is partitioned between sat. NH$_4$Cl and EtOAc. The organic phase is washed with brine, dried over MgSO$_4$, filtered, and concentrated. Purification by column chromatography (100% CHCl$_3$ → 10% MeOH in CHCl$_3$) affords the product as a blue solid (56 mg, 53%); R$_f$ 0.3 (10:1 CHCl$_3$:MeOH). MS (ESI) calculated for $C_{35}H_{38}N_5O_5$ [M+H]$^+$ 608.3, found 608.3.

(500 MHz, DMSO-d_6, 298K)

	^{13}C	^{1}H	Key correlations
1	136.0	6.58-6.72 (m, 4H)	HMBC → 3, 33/35
2	122.3	6.30-6.48 (m, 2H)	HMBC → 3
3	76.6	4.59-4.72 (m, 3H)	HMBC → 2, 5/9
			COSY → 1, 2
4	155.8	–	HMBC → 5/9, 6/8
5	114.9	6.58-6.72 (m, 4H)	COSY → 6/8
6	130.1	7.02 (d, J = 8.5 Hz, 2H)	HMBC → 10
7	127.8	–	HMBC → 5/9, 10, 11
8	130.1	7.02 (d, J = Hz, 2H)	HMBC → 10
9	114.9	6.58-6.72 (m, 4H)	
10	36.6	2.94 (dd, J = 14.0, 4.8 Hz, 1H), 2.71-2.83 (m, 3H)	HMBC → 6/8
11	54.2	4.31-4.42 (m, 1H)	COSY → 10
12	172.7	–	HMBC → 11
13	170.2	–	HMBC → 11, 14
14	49.7	4.59-4.72 (m, 3H)	COSY → 15
15	33.7	3.19 (dd, J = 15.8, 4.3 Hz, 1H)	HMBC → 17
		3.05 (dd, J = 15.7, 9.4 Hz, 1H)	
16	132.4	–	HMBC → 15
17	109.5	6.30-6.43 (m, 2H)	
18	116.7	6.58-6.72 (m, 4H)	
19	117.2	7.36 (d, J = 8.9 Hz, 1H)	HMBC → 16, 17
20	132.9	–	HMBC → 19, 21
21	99.6	6.43-6.49 (dd, J = 3.9, 1.2 Hz, 1H)	
22	113.6	6.82 (dd, J = 6.6, 1.0 Hz, 1H)	HMBC → 21, 20
23	109.7	7.50 (s, 1H)	HMBC → 21, 16
24	172.6	–	HMBC → 14, 25, 26
25	48.4	4.16-4.27 (m, 1H)	HMBC → 26
26	17.8	1.12 (d, J = 7.0 Hz, 3H)	
27	171. 4	–	HMBC → 29
28	36.5	2.36-2.47 (m, 2H)	
29	30.8	2.71-2.83 (m, 3H)	HMBC → 31/35
30	141.8	–	HMBC → 29, 28
31	126.6	7.18-7.32 (m, 4H)	
32	128.6	7.18-7.32 (m, 4H)	
33	124.3	7.18-7.32 (m, 4H)	
34	135.8	–	HMBC → 1, 2, 3
35	128.3	7.08-7.15 (m, 2H)	
5 x N**H**	–	8.29 (d, J = 7.8 Hz, 1H), 8.06 (d, J = 6.9 Hz, 1H), 7.92 (d, J = 8.2 Hz, 1H), 7.18-7.32 (m, 4H), 7.08-7.15 (m, 2H)	
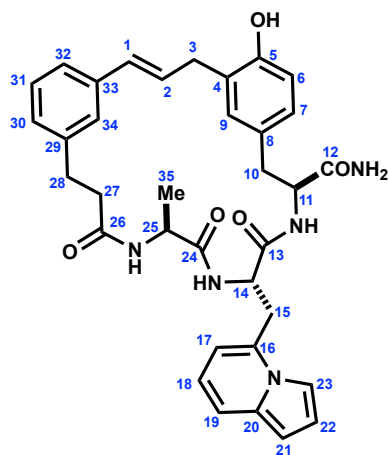
(7S,10S,13S,E)-5⁶-hydroxy-10-(indolizin-5-ylmethyl)-13-methyl-9,12,15-trioxo-8,11,14-triaza-1,5(1,3)-dibenzenacycloheptadecaphan-2-ene-7-carboxamide (66a).
65 (164 mg, 0.23 mmmol) was suspended in nitromethane (46 mL) under argon and treated with 5%v TFA (2.3 mL) at room temperature. The reaction was allowed to stir until full consumption of starting material was observed by TLC. The reaction mixture was partitioned between sat. NaHCO$_3$ and EtOAc. The organic phase was washed with sat. NaHCO$_3$ (x2), sat. NH$_4$Cl (x3), brine (x1), dried over MgSO$_4$, filtered and concentrated. The residue was redissolved in DMF and purified by preparative HPLC {Waters Sunfire C18, 19x250 mm, 5μm; 20-88% (MeCN+0.1%v TFA)/(H$_2$O+0.1%TFA) 2-9 min} to give the product as a beige solid (60 mg, 43%), which turns bright purple when dissolved in most common organic solvents; R$_f$ 0.5 (15:1 CH$_2$Cl$_2$:MeOH). MS (ESI) calculated for C$_{35}$H$_{38}$N$_5$O$_5$ [M+H]$^+$ 608.3, found 608.3.

(500 MHz, DMSO-d_6, 298K)

	^{13}C	^{1}H	Key correlations
1	130.3	6.25-6.37 (m, 2H)	HMBC → 3, 32, 33, 34
2	128.7		
3	33.0	3.51 (dd, J = 15.3, 5.4 Hz, 1H)	HMBC → 1, 2, 4, 5
		3.27 (dd, J = 15.0, 6.1 Hz, 1H)	
4	125.7	–	HMBC → 3, 9
5	153.5	–	HMBC → 3
6	127.5	6.91-7.01 (m, 3H)	HMBC → 5, 8
7	131.4	7.04-7.13 (m, 2H)	HMBC → 8
8	131.9	–	HMBC → 9, 10, 11
9	115.0	6.71 – 6.82 (m, 2H)	HMBC → 4
10	34.7	2.36-2.48 (m, 3H)	HMBC → 8
		2.85-3.00 (m, 3H)	
11	49.6	4.35-4.58 (m, 2H)	HMBC → 10
12	170.0	–	HMBC → 11
13	170.9	–	HMBC → 14
14	53.7	4.35-4.58 (m, 2H)	
15	36.6	2.36-2.48 (m, 3H)	HMBC → 14
		2.85-3.00 (m, 3H)	
16	132.8	–	HMBC → 19/21, 23
17	109.8	5.69 (d, J = 6.6 Hz, 1H)	
18	116.8	6.38-6.47 (m, 2H)	
19	117.0	7.30 (d, J = 8.9 Hz, 1H)	
20	127.6	–	HMBC → 15
21	99.5	6.38-6.47 (m, 2H)	HMBC → 22, 23
22	113.4	6.71 – 6.82 (m, 2H)	HMBC → 23
23	109.7	7.32 – 7.41 (m, 2H)	HMBC → 21
24	173.3	–	
25	47.8	4.10-4.29 (m, 1H)	
26	171.7	–	
27	35.9	2.36-2.48 (m, 3H)	HMBC → 28, 29

		2.69-2.83 (m, 2H)	
28	30.5	2.69-2.83 (m, 2H) 2.85-3.00 (m, 3H)	HMBC → 29
29	141.3	–	
30	127.3	6.91-7.01 (m, 3H)	
31	128.1	6.91-7.01 (m, 3H)	
32	123.2	6.85-6.90 (m, 1H)	HMBC → 1/2
33	137.1	–	HMBC → 3, 30, 31
34	126.4	7.19-7.24 (m, 2H)	HMBC → 1/2, 30, 31
35	17.6	0.93 (d, J = 6.8 Hz, 3H)	HMBC → 25
5 x N**H** 1 x O**H**	–	7.04-7.13 (m, 2H), 7.19-7.24 (m, 2H), 7.32 – 7.41 (m, 2H), 7.80 (d, J = 8.2 Hz, 1H), 7.90 (d, J = 7.7 Hz, 1H), 7.97 (d, J = 8.4 Hz, 1H)	
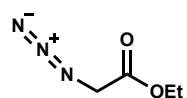
ethyl 2-azidoacetate (71). Ethyl bromoacetate (2 mL, 18.08 mmol) was dissolved in anhydrous DMF (32 mL) and cooled to 0 °C. NaN$_3$ was added slowly, and the reaction mixture was allowed to warm to room temperature. The mixture was stirred at room temperature until starting material was fully consumed, as indicated by TLC. Water was added and the reaction was extracted with EtOAc (x3). The organic phase was washed with brine, dried over MgSO$_4$, filtered and concentrated to afford the product as a colorless oil (2.2 g, 94%); R$_f$ 0.1 (9:1 Hex:EtOAc). ^{1}H NMR (400 MHz, CDCl$_3$) 4.26 (q, J = 7.1 Hz, 2H), 3.85 (s, 2H), 1.31 (t, J = 7.2 Hz, 3H); ^{13}C NMR (100 MHz, CDCl$_3$) 168.4, 62.0, 50.5, 14.2.
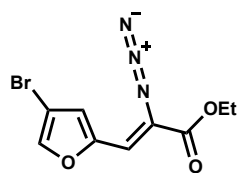
ethyl (Z)-2-azido-3-(4-bromofuran-2-yl)acrylate (72). *A. Preparation of NaOEt.* Dry ethanol (13.2 mL) was added to a flame dried flask equipped with a stir bar, and the flask was cooled to 0 °C under argon. Na metal was added to the cooled solution and the mixture was stirred at 0 °C under argon until all the Na metal dissolved. The solution was allowed to warm to room temperature and was used immediately in the next step. *B. Nitrene insertion into bromofurfural.* To a solution of bromofurfural (750 mg, 4.26 mmol) and ethyl azidoacetate (2.2 g, 17.04 mmol) at 0 °C was added the freshly prepared solution of NaOEt. The solution was stirred at 0 °C until the starting material was fully consumed, as indicated by TLC. The reaction was poured into ice cold NH$_4$Cl and extracted with ether. The organic phase was washed with brine, dried and concentrated. Purification by column chromatography (9:1 Hex:EtOAc) afforded the product as a pale yellow solid (500 mg, 41%); R$_f$ 0.6 (9:1 Hex:EtOAc). ^{1}H NMR (400 MHz, CDCl$_3$) 7.46 (d, J = 0.7 Hz, 1H), 7.16 (t, J = 0.6 Hz, 1H), 6.75 (d, J = 0.5 Hz), 4.35 (q, J = 7.1 Hz, 2H), 1.38 (t, J = 7.2 Hz, 3H).
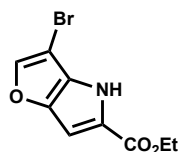
methyl 3-bromo-4H-furo[3,2-b]pyrrole-5-carboxylate (73). 72 (500 mg, 1.75 mmol) in toluene (15 mL) was added dropwise to refluxing toluene (45 mL) in a two necked flask fitted with a condenser. The mixture was stirred at reflux until starting material was fully consumed, as indicated by TLC. The reaction mixture was cooled to room temperature and concentrated *in vacuo*. The crude mixture was purified by column chromatography (9:1 Hex:EtOAc) to afford the product as a white solid, which gradually turns yellow on exposure to air (330 mg, 73%); R$_f$ 0.4 (9:1 Hex:EtOAc). ^{1}H NMR (400 MHz, CDCl$_3$) 8.83 (br s, 1H), 7.50 (s, 1H), 6.80 (d, J = 1.8 Hz), 4.37 (q, J = 7.1 Hz, 2H), 1.38 (t, J = 7.2 Hz, 3H).
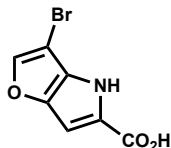
3-bromo-4*H*-furo[3,2-*b*]pyrrole-5-carboxylic acid (74). 73 (330 mg, 1.28 mmol) was added to a solution of KOH (194 mg, 3.45 mmol) in water (6.5 mL) and heated to 40 °C. A few drops of MeOH were added to wash **x** off the walls of the flask and into the aqueous solution. Over the course of the reaction, the insoluble starting material dissolved into the aqueous solution as the potassium salt. When no more starting material was evident, the reaction was cooled and MeOH was removed *in vacuo*. The remaining solution was extracted with ether (x2), then acidified with ice cold 5M HCl to pH=2. This precipitated the product, which was extracted with ether (x3), washed with water and brine, dried over MgSO$_4$, filtered, and concentrated to afford the product as a yellow solid (257 mg, 87%). When exposed to air, the solid gradually changes color to purple, but no structural change is evident by NMR. ^{1}H NMR (400 MHz, MeOD-*d$_4$*) δ 7.63 (s, 1H), 6.74 (s, 1H); ^{13}C NMR (100 MHz, MeOD-*d$_4$*) 163.4, 147.7, 146.0, 128.0, 125.2, 96.4, 88.2, 84.9, 24.4.
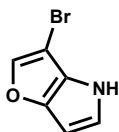
3-bromo-4*H*-furo[3,2-*b*]pyrrole (75). 74 (40 mg, 0.17 mmol) was dissolved in ethanolamine (7 mL) and the solution was brought to reflux. The reaction was quenched by pouring over ice cold water after 15 minutes at reflux. The product was rapidly extracted into cold ether, washed with cold water and brine, and concentrated in a cold bath to afford the product as a purple solid. Extreme care is required during post reaction handling as the compound decomposes into black tar within a few minutes at room temperature and exposed to air. Yield not available due to decomposition. ^{1}H NMR (400 MHz, MeOD-*d$_4$*) δ 7.39 (d, *J* = 1.2 Hz, 1H), 6.80 (dd, *J* = 4.2, 1.9 Hz, 1H), 6.05 (d, *J* = 3.1 Hz, 1H). HMRS (ESI) calculated for C$_6$H$_5$BrNO [M+H]$^+$ 185.9549, found 185.9539.
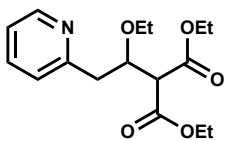
diethyl 2-(1-ethoxy-2-(pyridin-2-yl)ethyl)malonate (76). *A. Preparation of LDA solution.* Anhydrous THF (12.1 mL) and diisopropylamine (3.76 mL, 26.62 mmol) were added under argon to a flame dried flask equipped with a stir bar. The reaction mixture was cooled to –78 °C, after which *n*-BuLi (10.5 mL, 2.3 M in hexanes) was added dropwise *via* syringe. The resulting mixture was stirred for 10 minutes at –78 °C. After this time, the flask was allowed to warm to –60 °C and was used immediately in the next step. *B. Conjugate addition.* To the freshly prepared solution of LDA at –60 °C was added a solution of 2-picoline (2.17 mL, 22 mmol) in anhydrous THF (38 mL) dropwise over 15 minutes under argon. The mixture was stirred at this temperature for 45 minutes before diethyl 2-(ethoxymethylene)malonate (4.67 mL, 23 mmol) was added dropwise. The resulting mixture was stirred at –20 °C until the starting material was fully consumed, as indicated by TLC. The color of the solution gradually changed from red to yellow during this time. The reaction mixture was quenched with sat. NH$_4$Cl (10 mL) and extracted with EtOAc (x2). The combined organics were washed with brine (x2), dried over MgSO$_4$ and concentrated. Purification by column chromatography (5% → 50% EtOAc in hexanes) afforded the product as a yellow oil (6.37 g, 94%); R$_f$ 0.5 (1:1 Hex:EtOAc). ^{1}H NMR (400 MHz, CDCl$_3$) 8.53 (ddd *J* = 4.9, 1.8, 0.9 Hz, 1H), 7.58 (dt, *J* = 7.6, 1.9 Hz, 1H), 7.22 (d, *J* = 7.8 Hz, 1H), 7.12 (ddd, *J* = 7.5, 4.9, 0.1 Hz, 1H), 4.46-4.39 (m, 1H), 4.23-4.13 (m, 4H), 3.61 (d, *J* = 8.1 Hz, 1H), 3.50-3.23 (m, 2H), 3.19-3.01 (m, 2H), 1.28-1.22 (m, 6H), 0.96 (t, *J* = 7.03, 3H). HRMS (ESI) calculated for C$_{16}$H$_{24}$NO$_5$ [M+H]$^+$ 310.1649, found 310.1620.
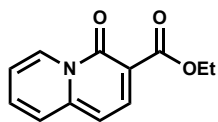
ethyl 4-oxo-4*H*-quinolizine-3-carboxylate (77). A suspension of **76** (6.27 g, 20.3 mmol) in xylenes (35 mL) was mixed with acetic acid (8.5 mL) and the reaction was heated to 120 °C and stirred at this temperature until starting material as fully consumed, as indicated by TLC. After cooling to room temperature, the reaction was diluted with EtOAc and washed with water (x3). The organic phase was washed with brine, dried over MgSO$_4$, filtered, and concentrated to afford the product as a

yellow solid (3.57g, 81%); R$_f$ 0.1 (1:1 Hex:EtOAc). ^{1}H NMR (400 MHz, CDCl$_3$) δ 9.43-9.37 (m, 1H), 8.40 (d, J = 8.4 Hz, 1H), 7.63-7.54 (m, 2H), 7.22-7.16 (m, 1H), 6.65 (d, 8.37, 1H), 4.42 (q, J = 7.1 Hz, 2H), 1.42 (t, J = 1.4 Hz, 3H); ^{13}C NMR δ (100 MHz, CDCl$_3$) 166.3, 155.9, 146.1, 141.1, 133.5, 129.4, 125.4, 116.5, 106.9, 102.3, 60.9, 14.5. HRMS (ESI) calculated for C$_{12}$H$_{12}$NO$_5$ [M+H]$^+$ 218.0812, found 218.0802.
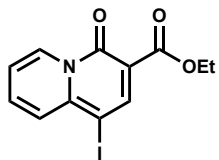
ethyl 1-iodo-4-oxo-4*H*-quinolizine-3-carboxylate (78). To a solution of **77** (3.57 g, 16.4 mmol), in anhydrous DMF (71 mL) was added *N*-iodosuccinimide (9.22 g, 41 mmol). The mixture was stirred at 40 °C until starting material was fully consumed, as indicated by TLC. The reaction mixture was poured into 10% aq. Na$_2$SO$_3$ (700 mL) and stirred for an additional 30 minutes. A yellow precipitate formed which was collected by filtration, washed with 5% aq. NaSO$_3$ and water. The filter cake was redissolved in DCM. MgSO$_4$ was added to remove water from the filter cake and the resulting solution was refiltered and concentrated to afford the product as a yellow solid (4.7 g, 84%); R$_f$ 0.15 (1:1 Hex:EtOAc). ^{1}H NMR (400 MHz, CDCl$_3$) δ 9.40 (d, J = 7.2 Hz, 1H), 8.73 (s, 1H), 8.0 (d, J = 8.9 Hz, 1H), 7.76-7.68 (m, 1H), 7.24 (td, J = 7.0, 2.5 Hz, 1H), 4.38 (q, J = 7.1 Hz, 2H), 1.39 (t, J = 7.1 Hz, 3H). ^{13}C NMR (100 MHz, CDCl$_3$) δ 164.8, 155.2, 150.2, 144.8, 136.1, 130.6, 129.5, 117.0, 108.3, 62.5, 61.2, 14.5.
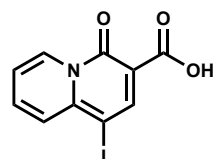
1-iodo-4-oxo-4*H*-quinolizine-3-carboxylic acid (79). A suspension of **78** (343 mg, 1 mmol) in 10% NaOH solution (12.4 mL) was stirred at 50 °C until starting material was fully consumed, as indicated by TLC. The reaction mixture was neutralized to pH=1 with conc. HCl and extracted with CH$_2$Cl$_2$. The organic phase was washed with brine, dried over MgSO$_4$, filtered, and concentrated to afford the product as a yellow solid (314 mg, 100%). ^{1}H NMR (400 MHz, CDCl$_3$) δ 13.62 (br s, 1H), 9.43-9.39 (m, 1H), 9.05 (s, 1H), 8.25-8.19 (m, 1H), 7.92-7.86 (m, 1H), 7.46 (td, J = 7.0, 1.3 Hz, 1H).
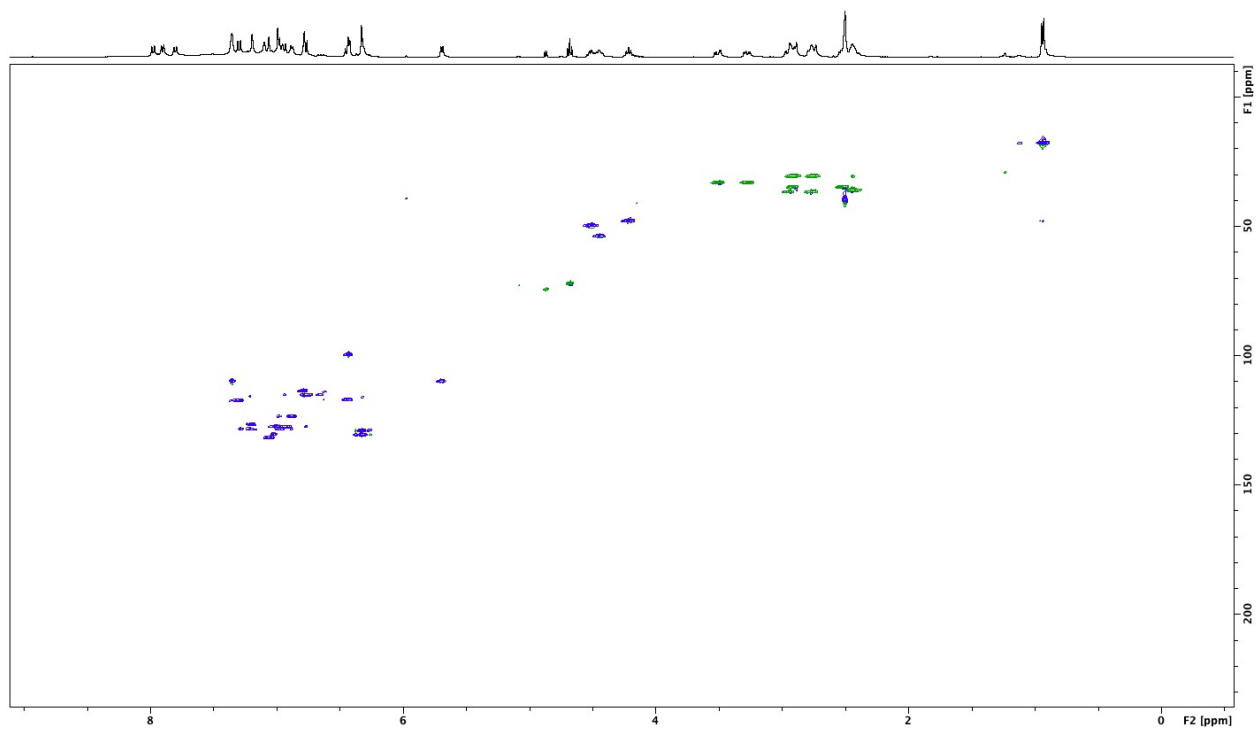
Chapter 3 NMR Spectra

Arrows indicate HMBC correlations

HSQC

HMBC

162

Arrows indicate HMBC correlations

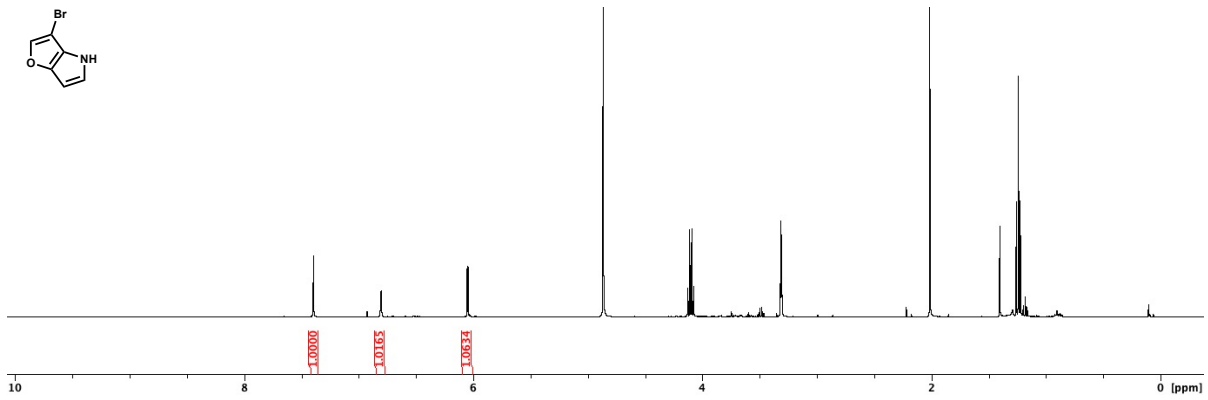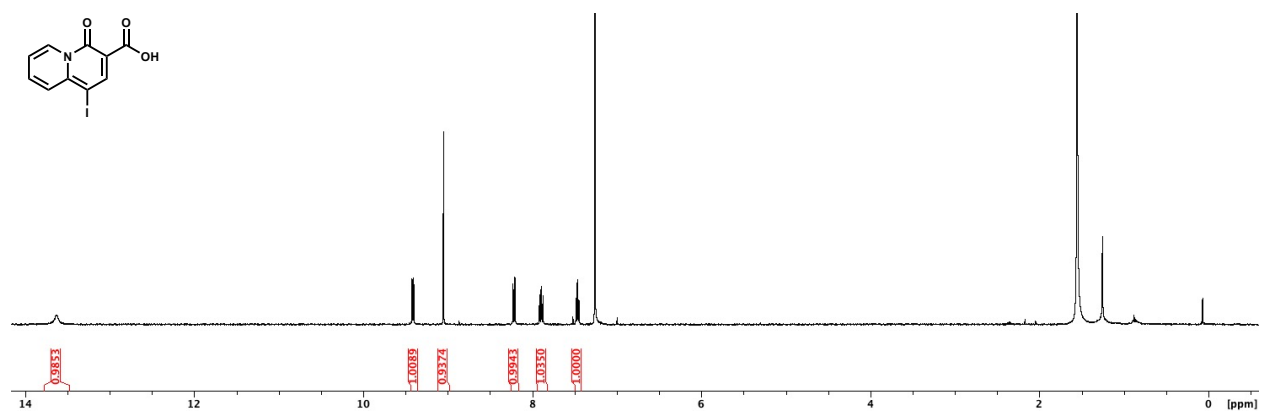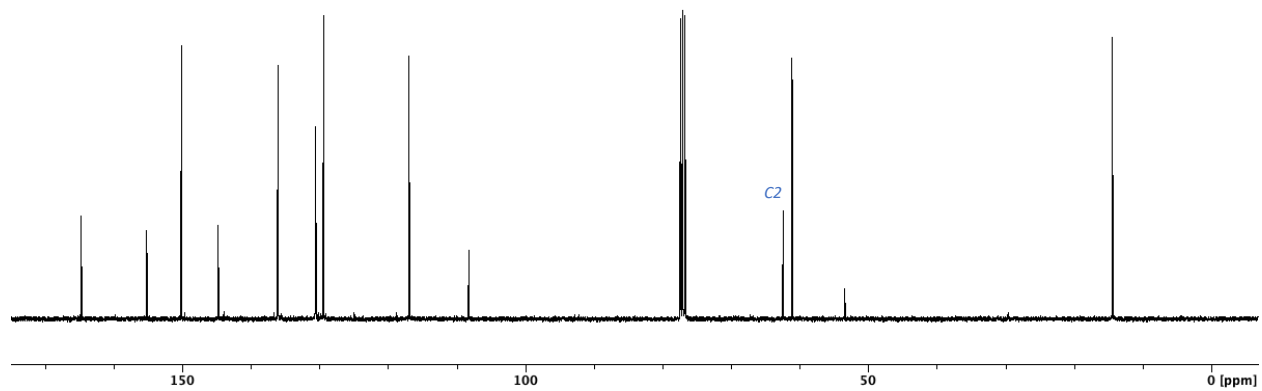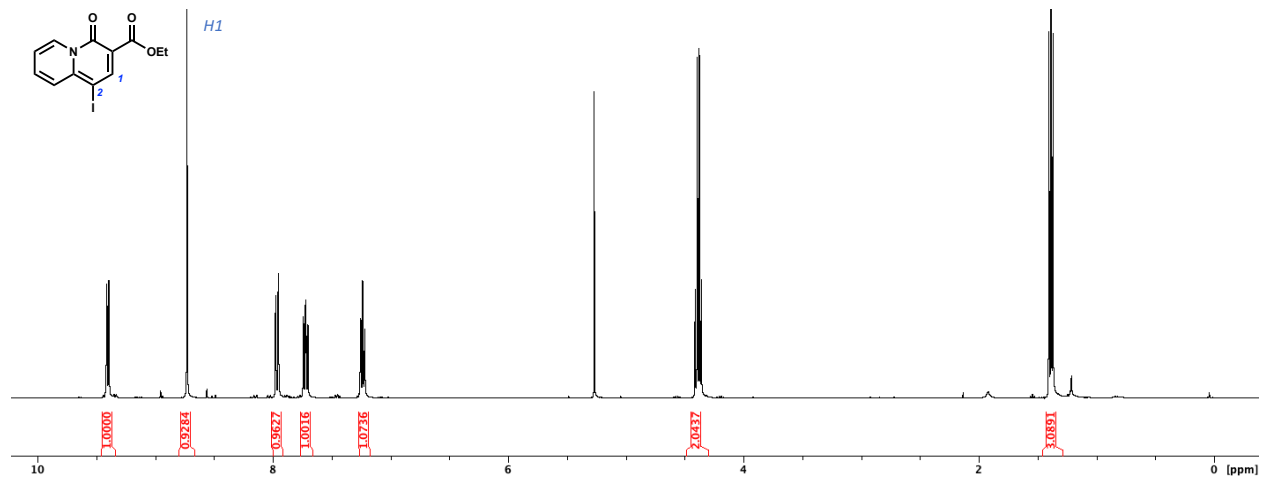HSQC

HMBC

164

165

167
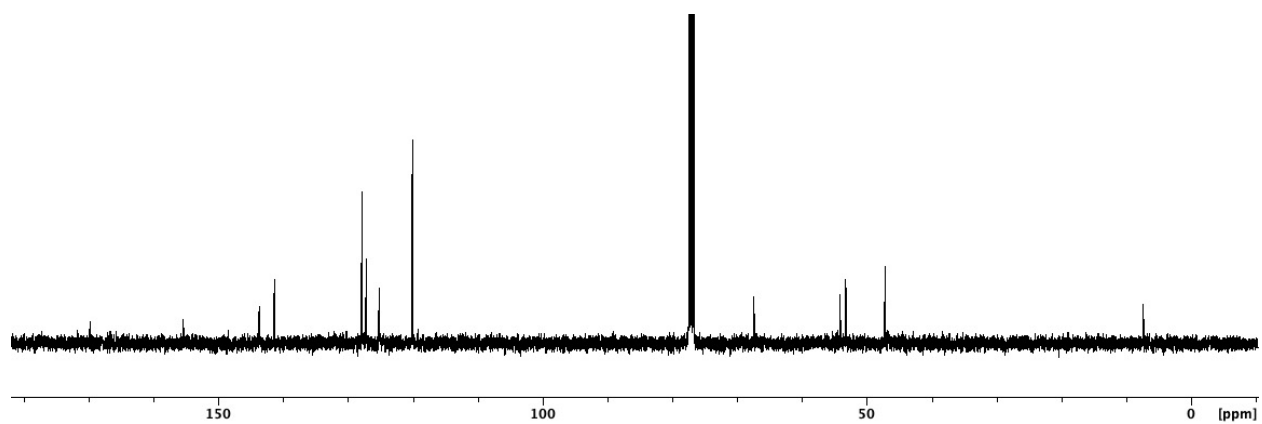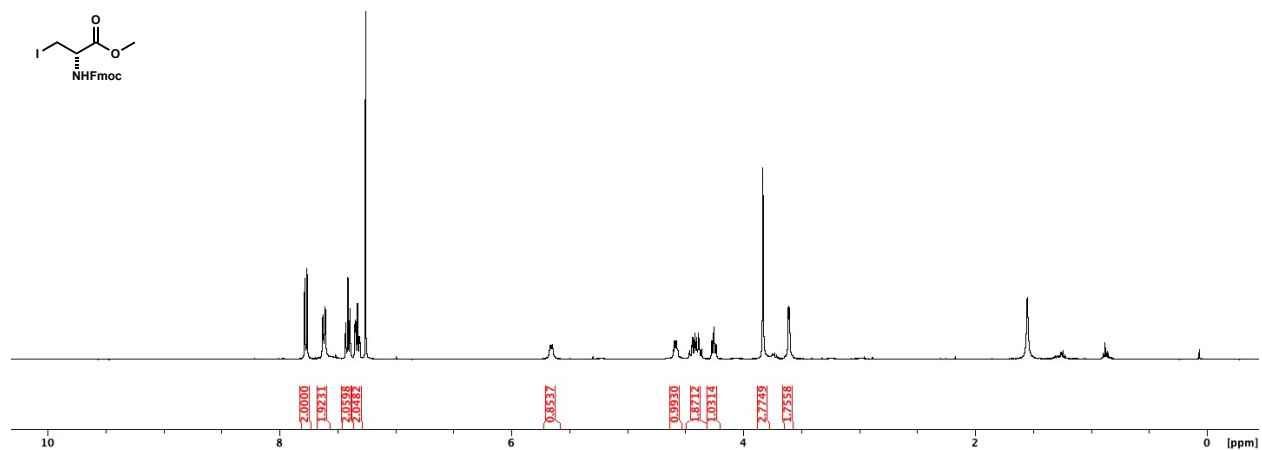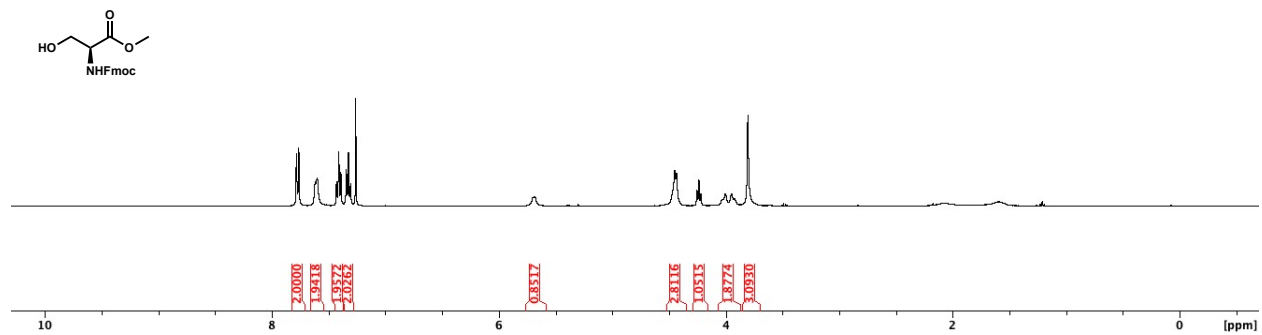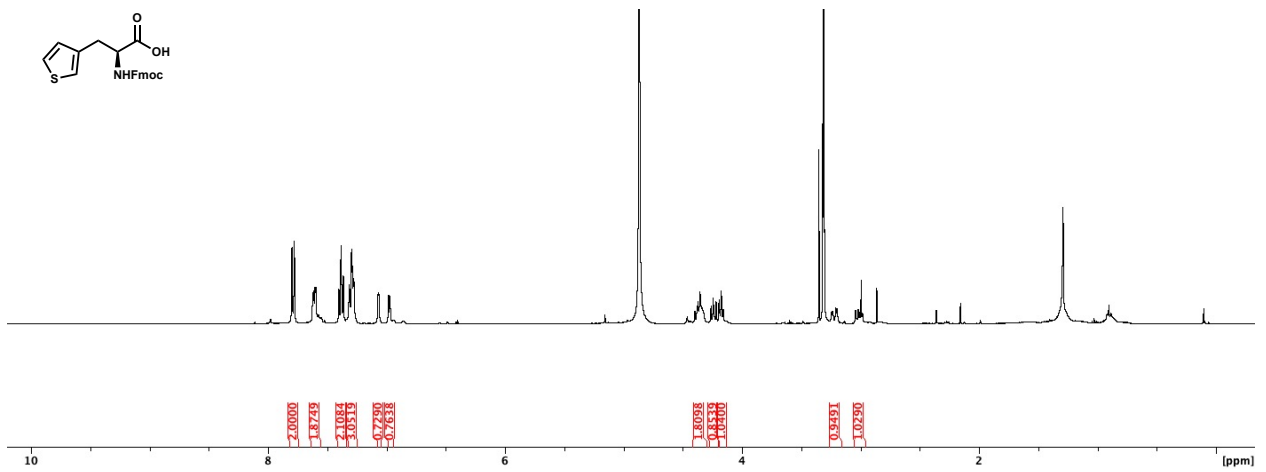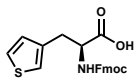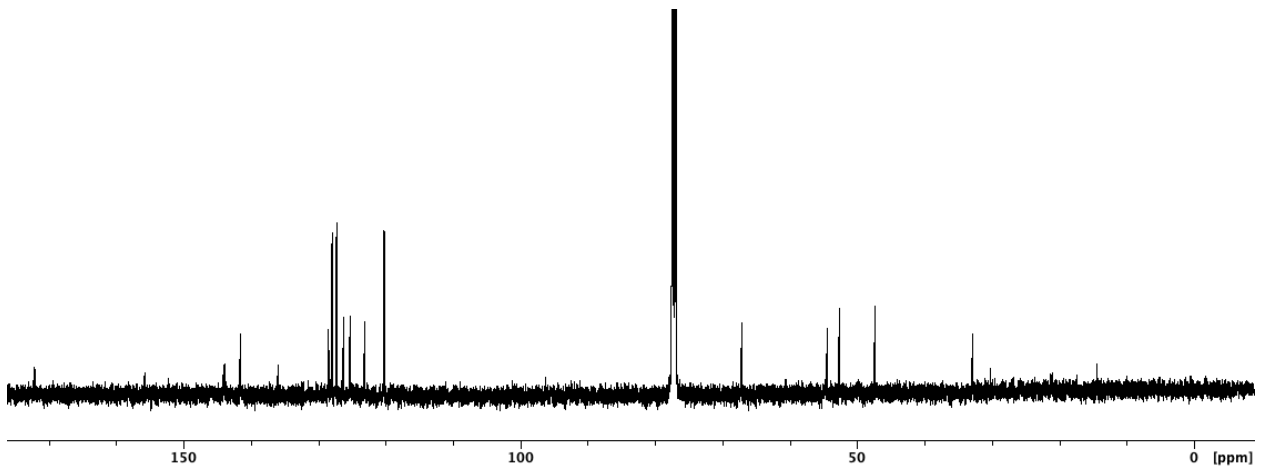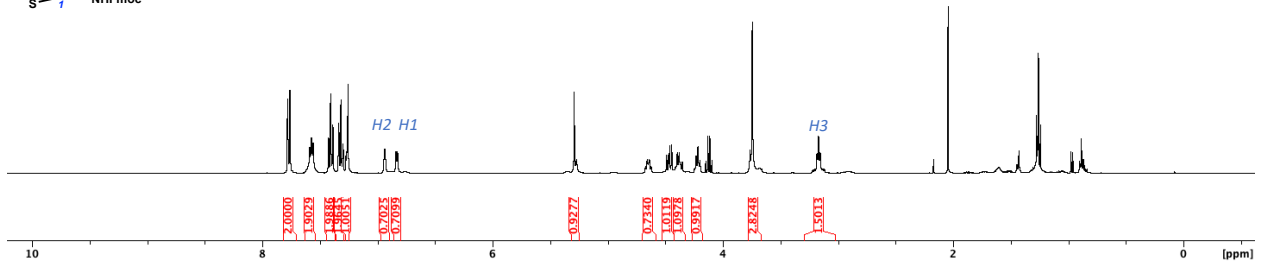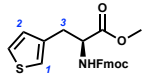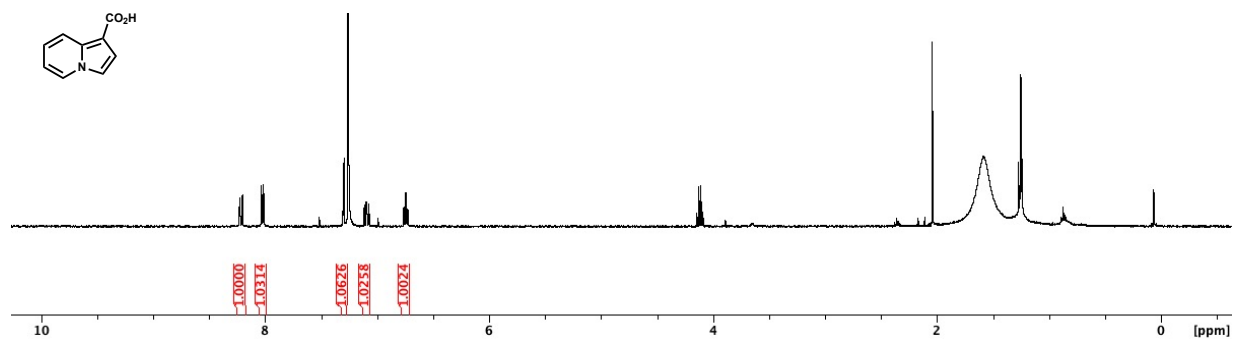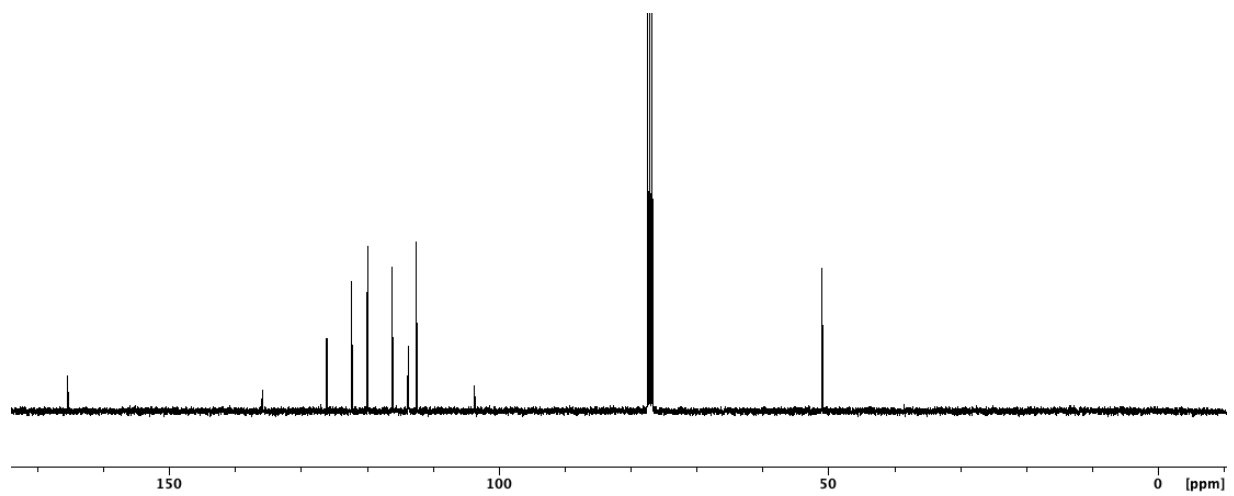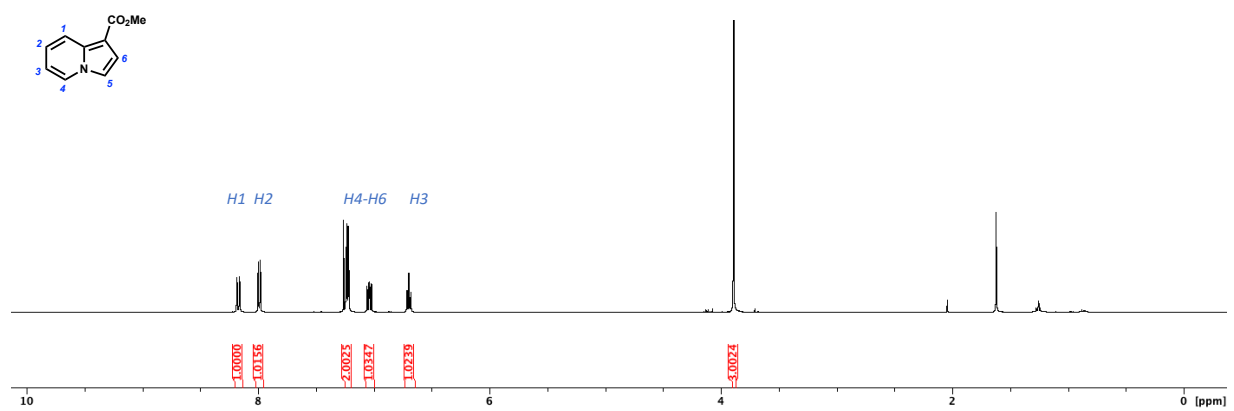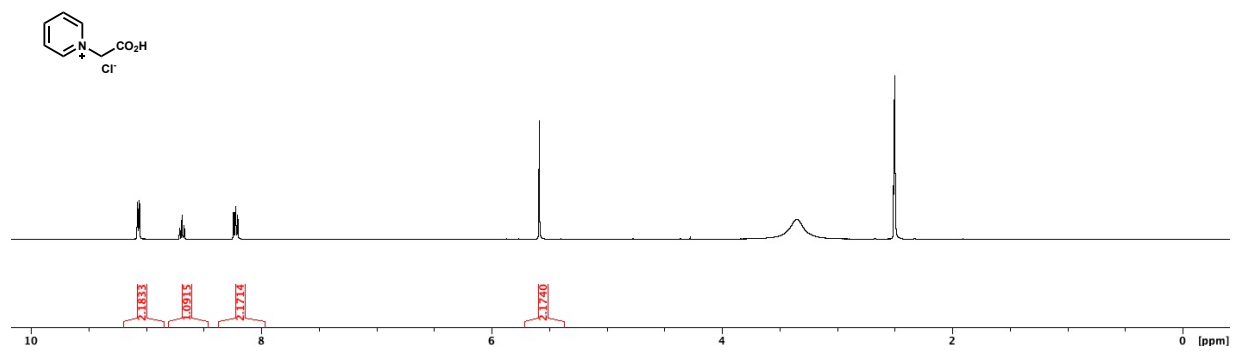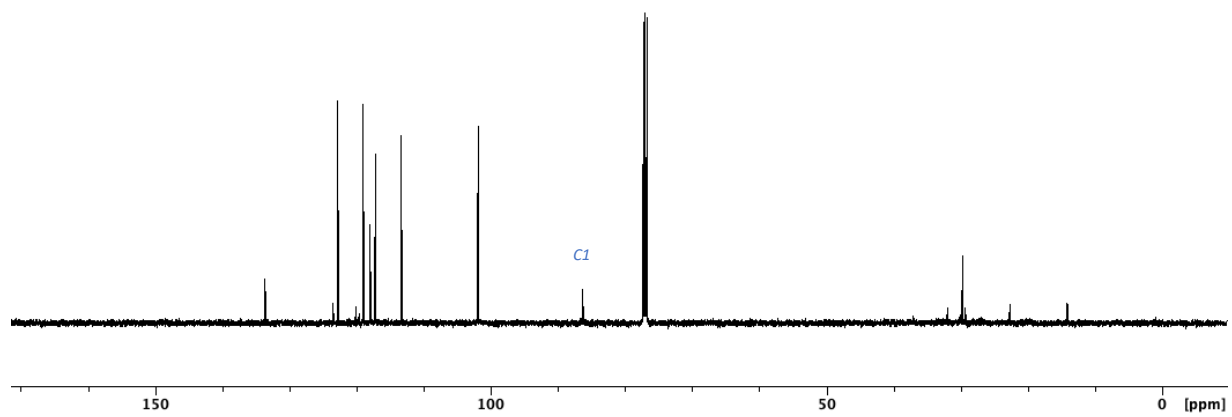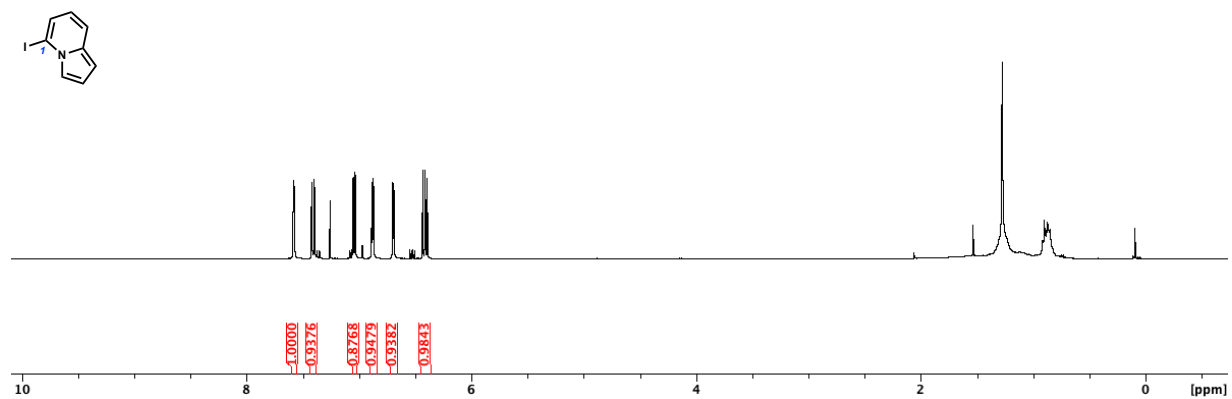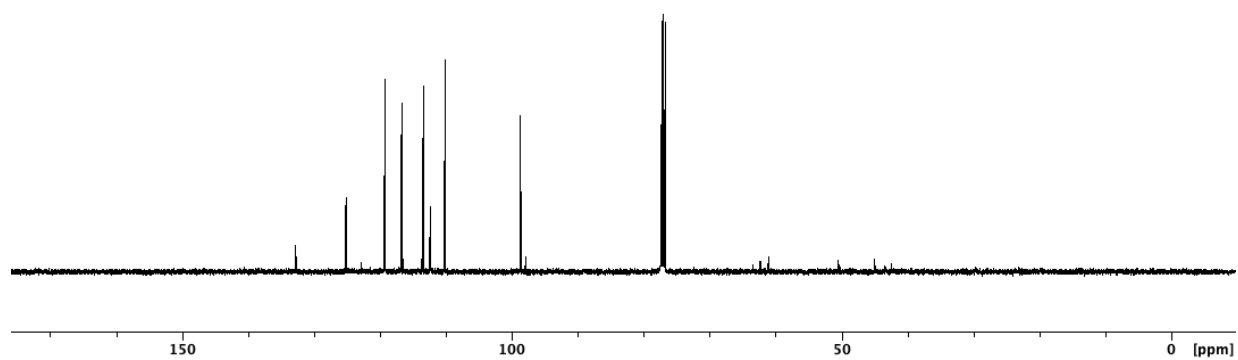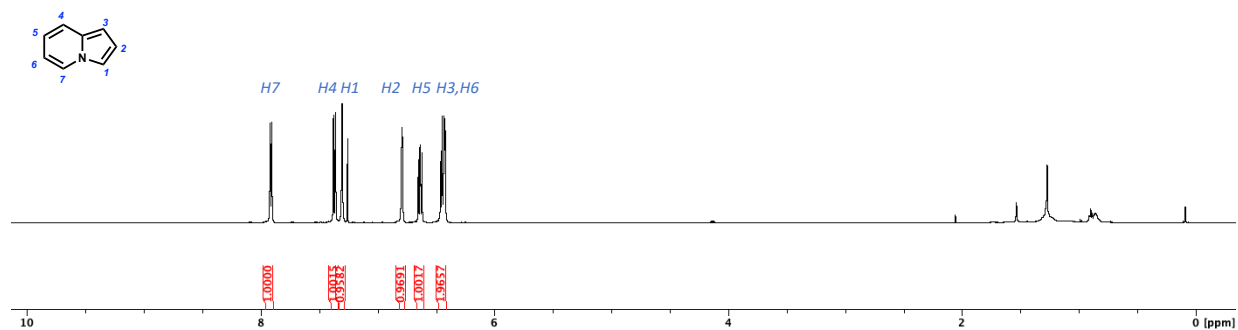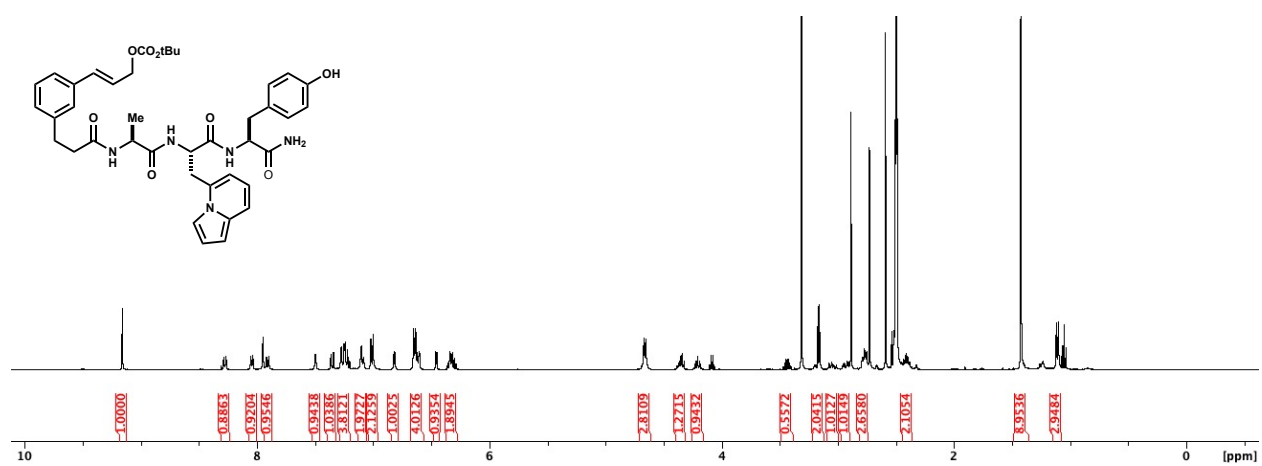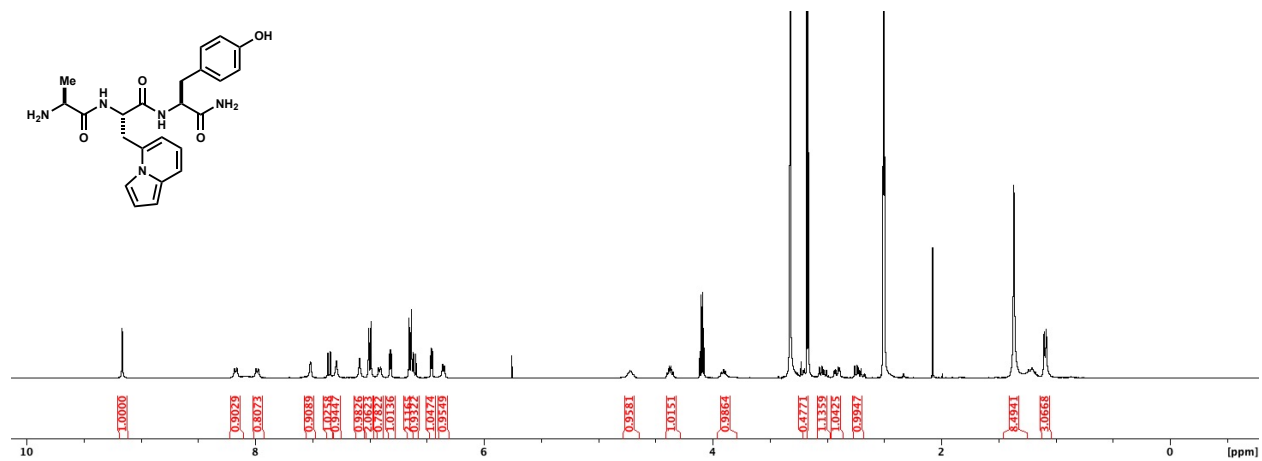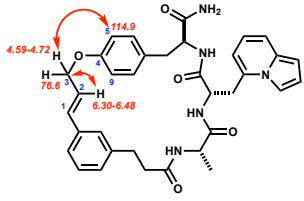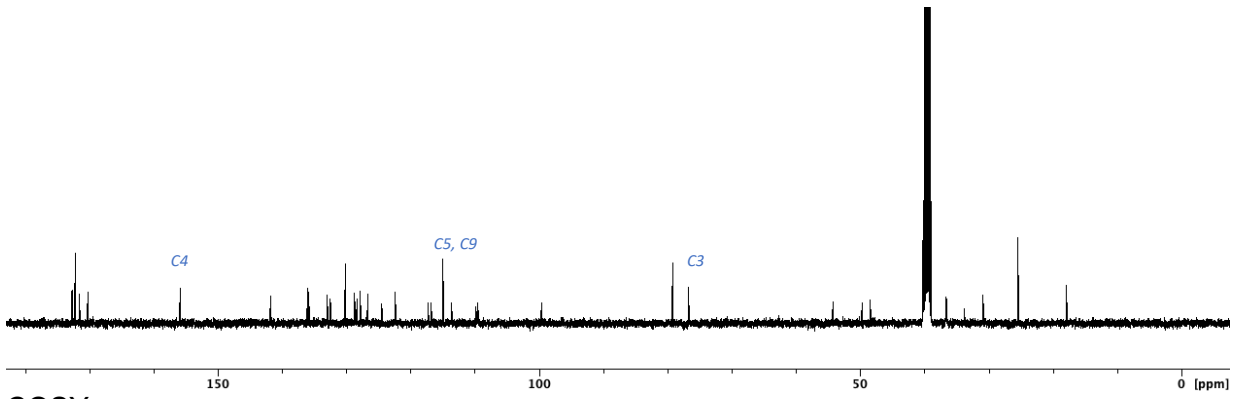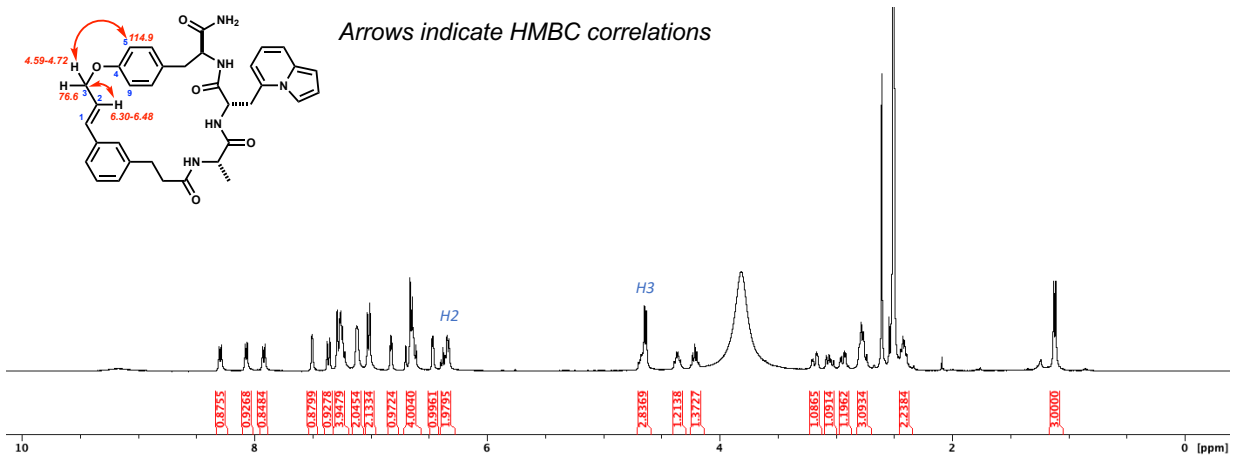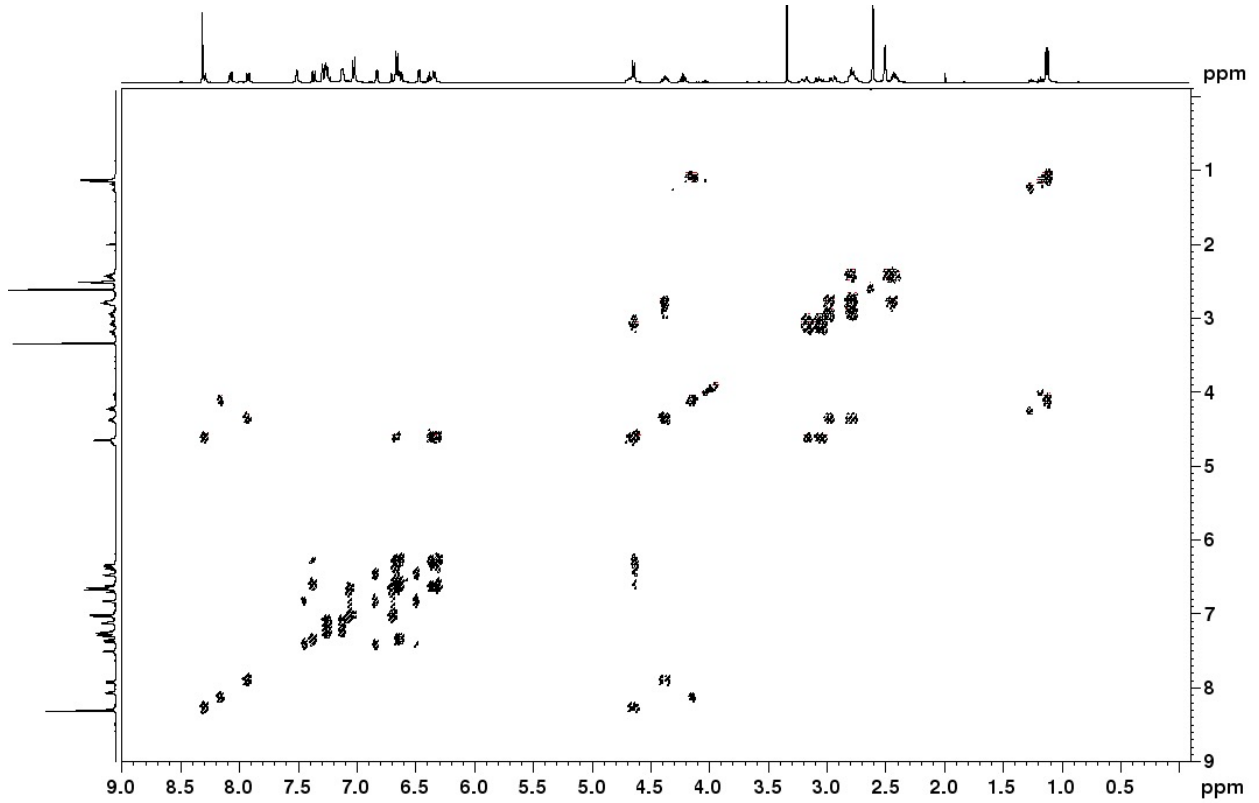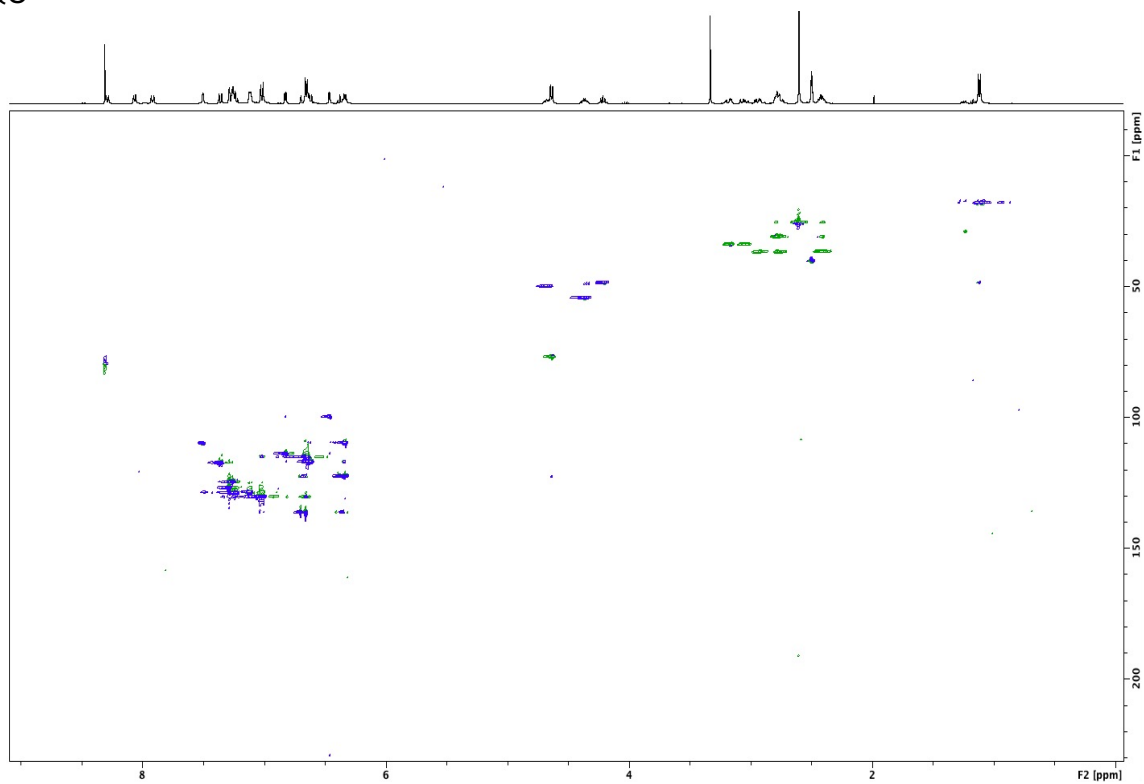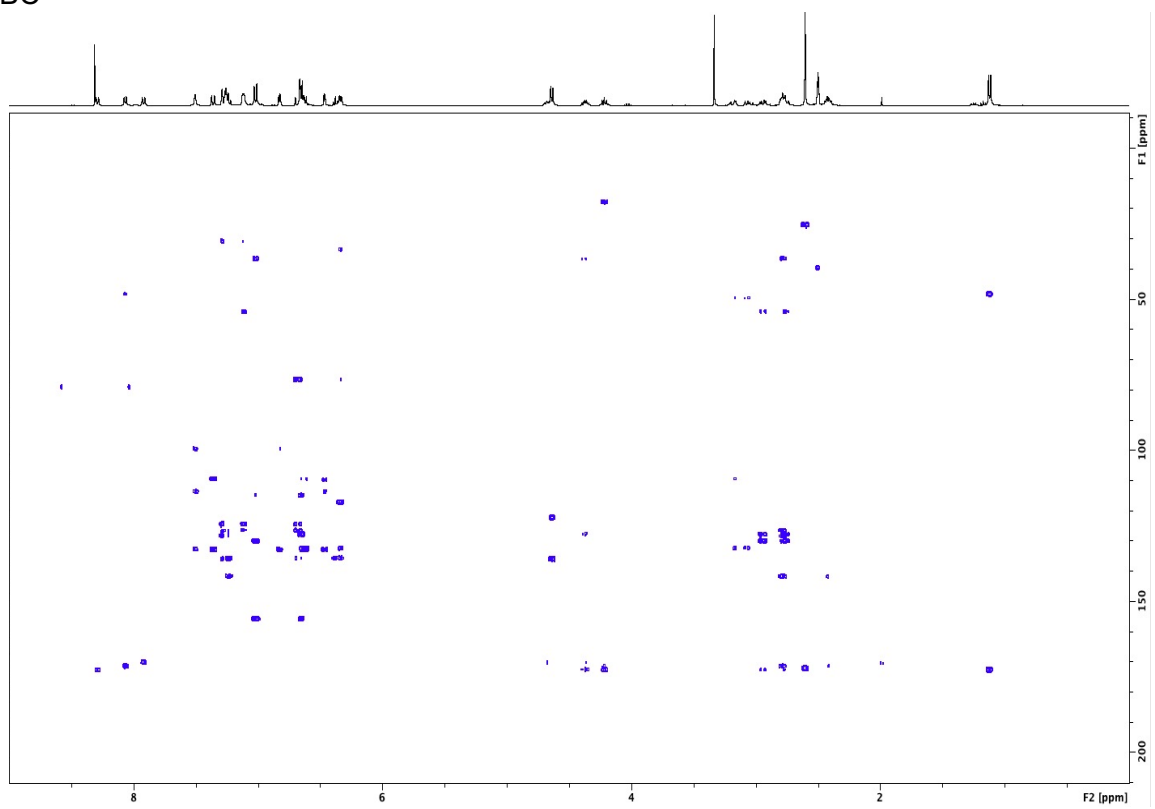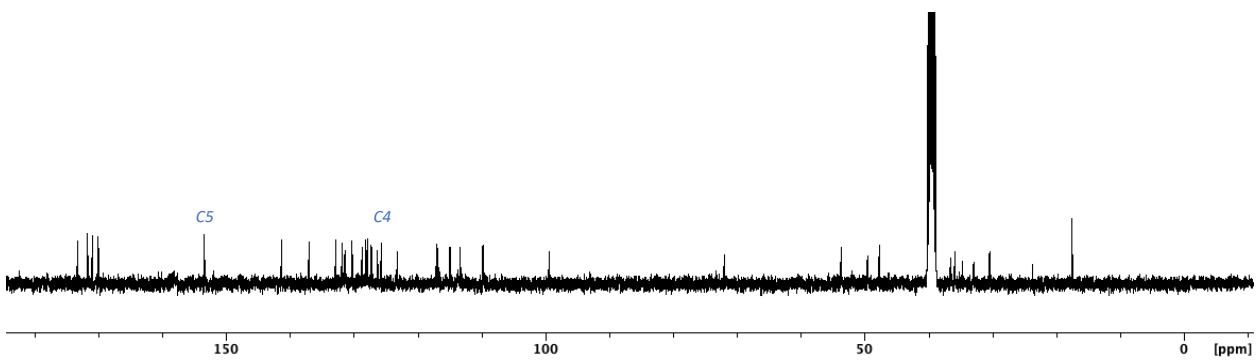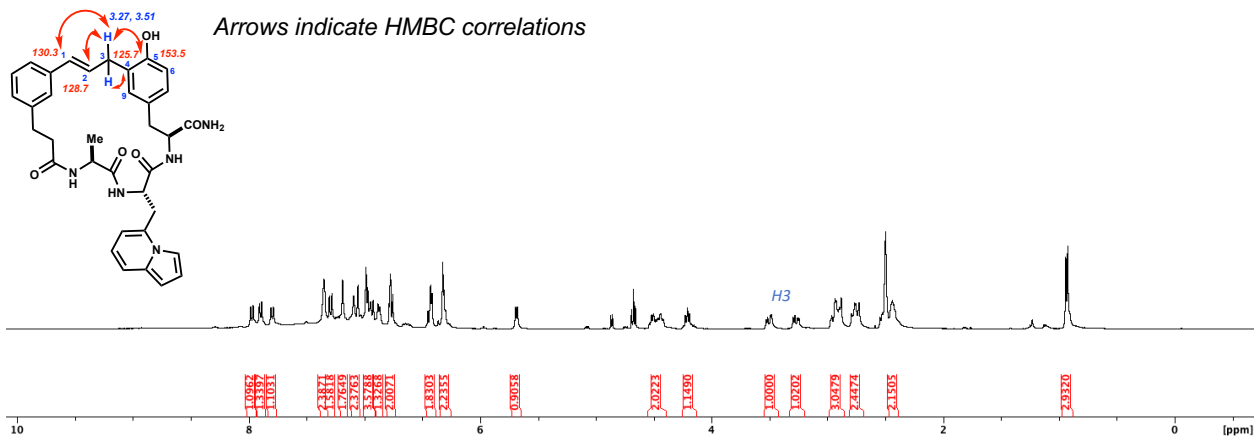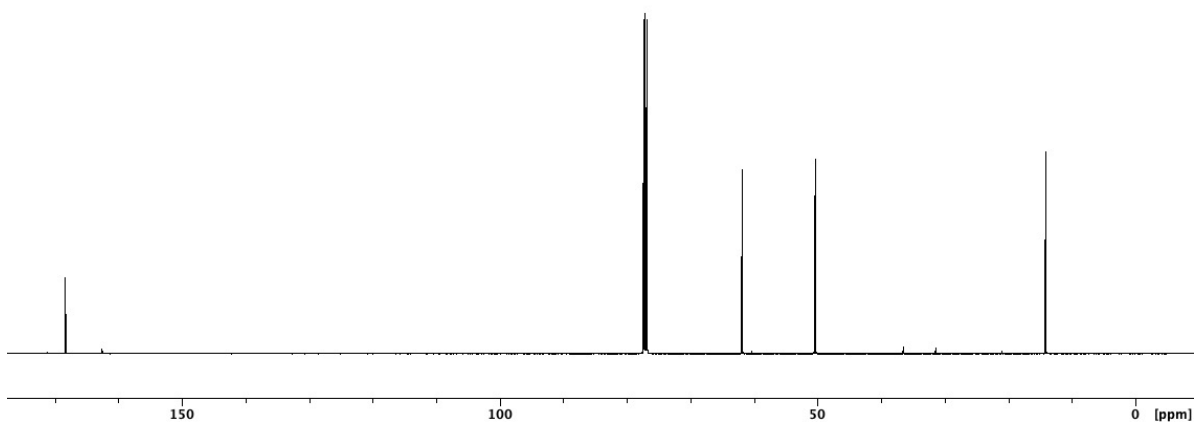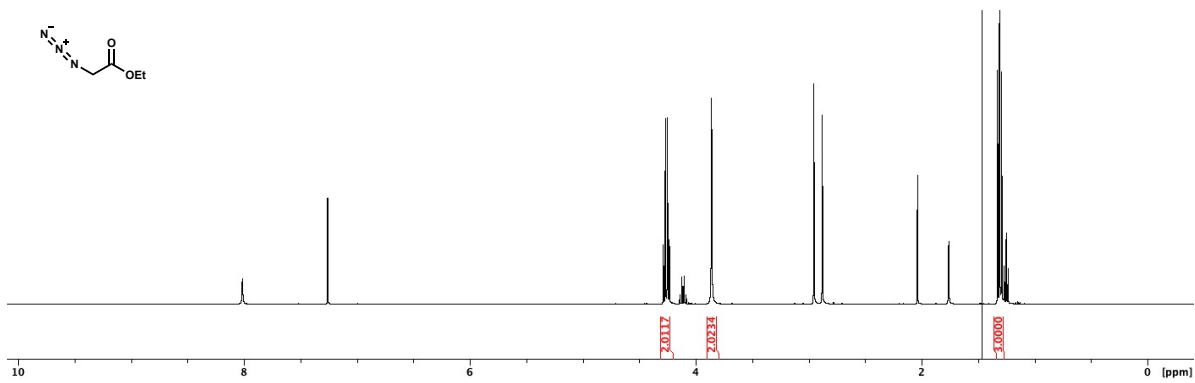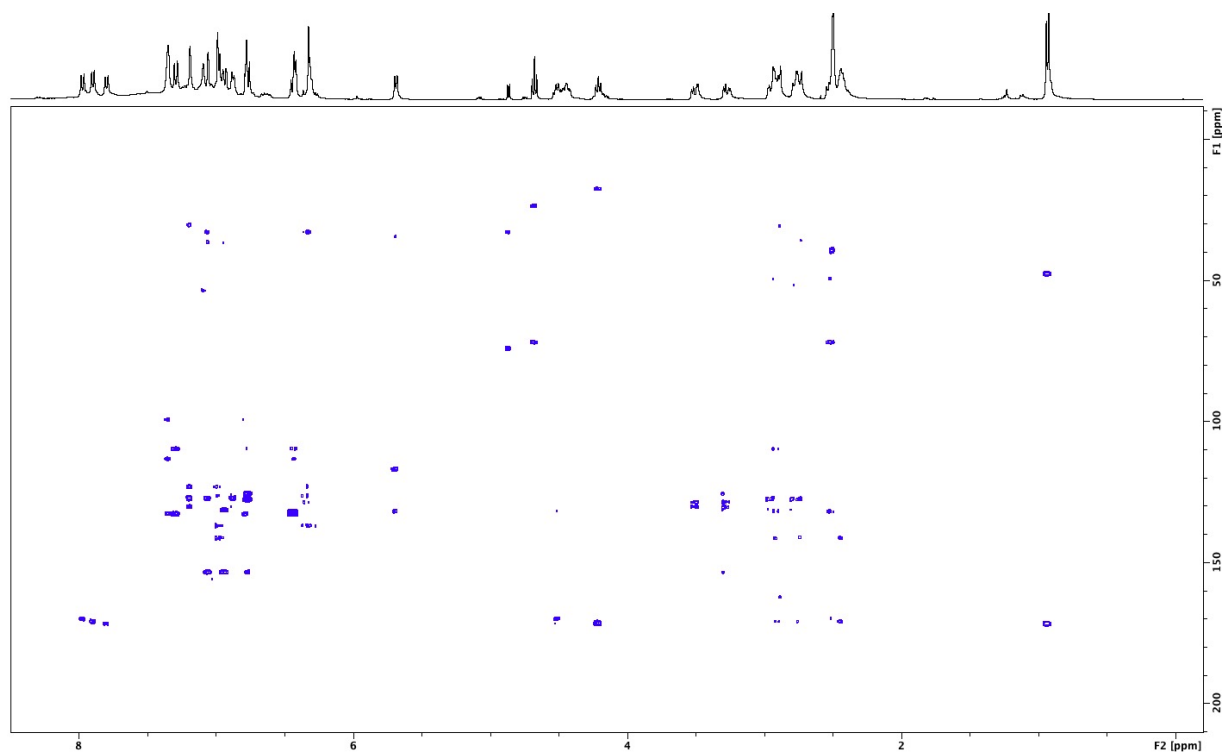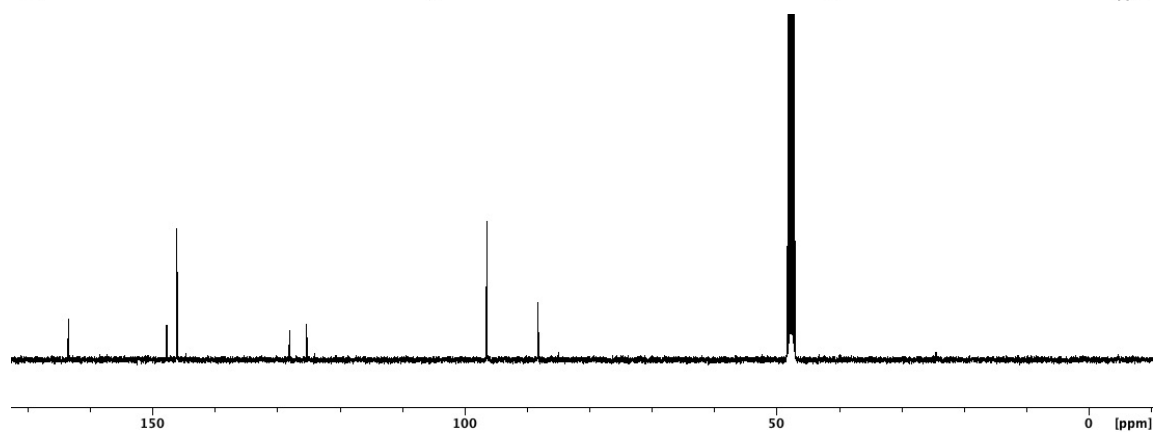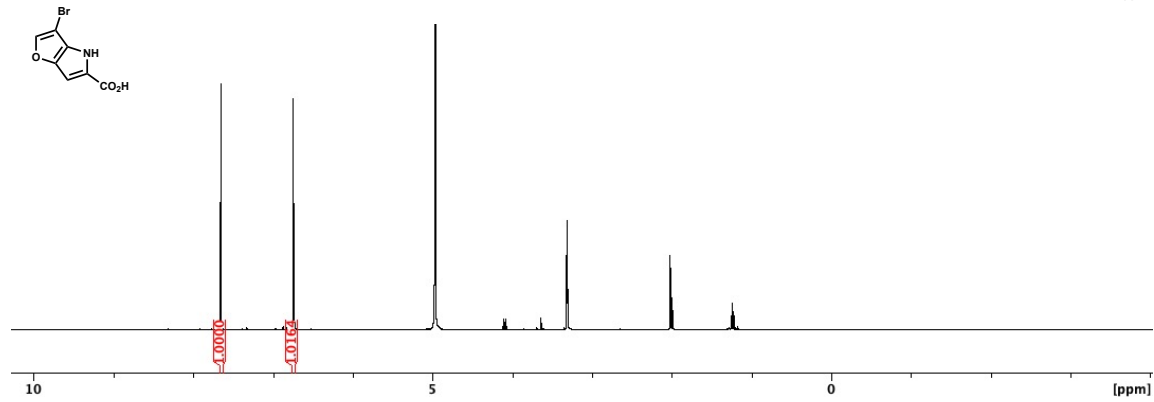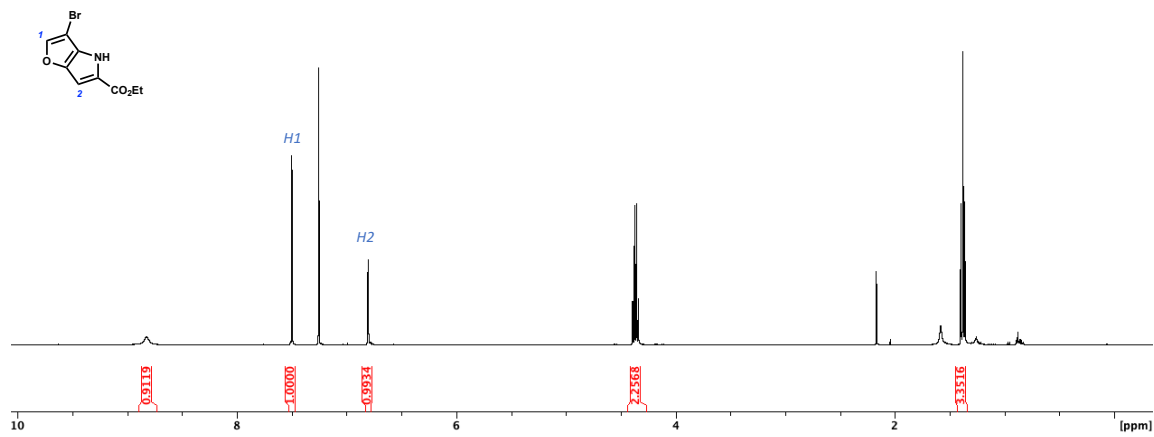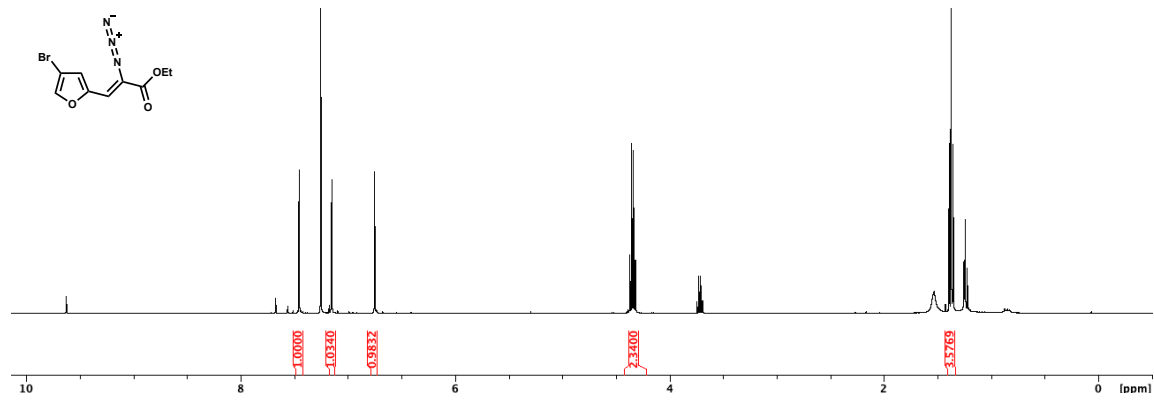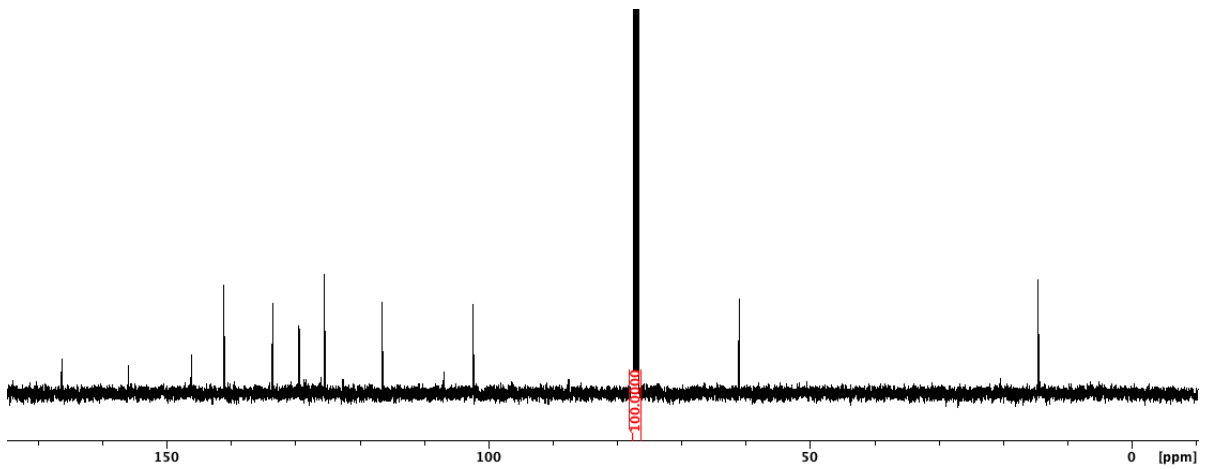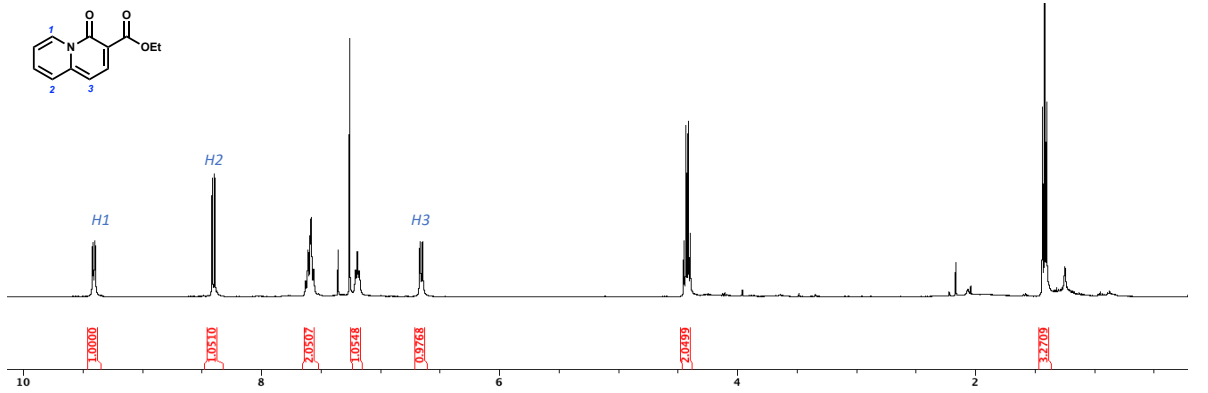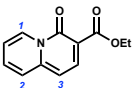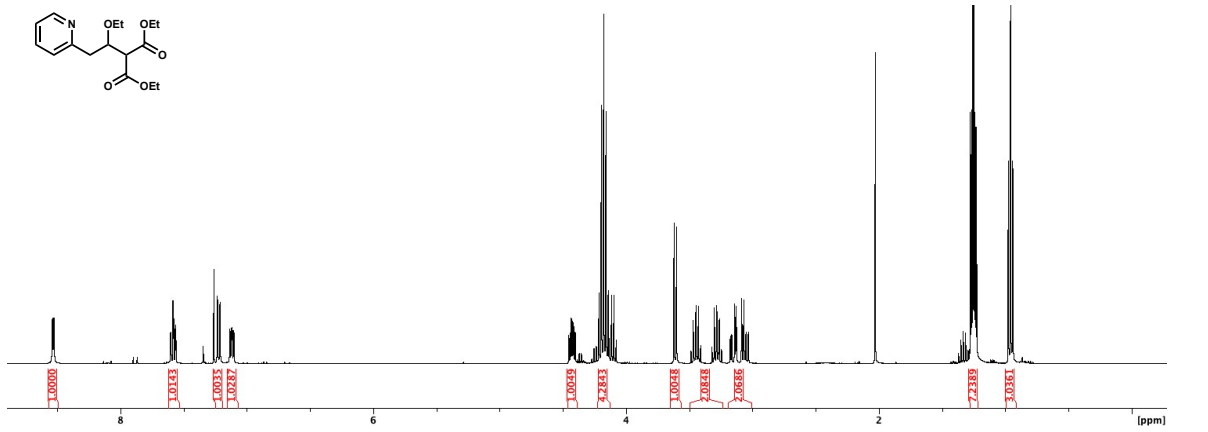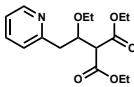
Chapter 4 Experimental Appendix

In vitro Assay. pFastBac1-mouseGOAT and pGEX-GST-proGhrelin8His plasmid encoding for mouse proghrelin, fused to GST with TEV cleavage site to release proghrelin moiety, was a kind gift from the Brown and Goldstein laboratory.[2] The BL21Gold *E. coli* chemically competent cells were transformed with the pGEX-GST-proGhrelin8His plasmid and selected on agar plates with ampicillin. A few colonies were used to inoculate 50 mL LB, and cultures were grown overnight at 37 °C. The next morning, 10 mL of overnight culture was used to inoculate 1L LB. A total of 4 flasks with 1L LB each were inoculated with 10 mL of the overnight culture, and bacterial cultures were grown at 37 °C until they reached OD_{600} ~ 0.6. The cultures were chilled on ice for 30 min and then 0.25 mM of IPTG was added to each culture. The cultures were then moved back to a shaker set to 18 °C and incubated overnight. Cells were harvested and resuspended in 200 mL buffer containing 25 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM EDTA and 0.5 mg/mL lysozyme. The cell suspension was sonicated, and cell debris was removed by centrifugation. The soluble fraction was passed through glutathione resin by gravity. The resin was washed with 75 mL of lysis buffer without lysozyme. The GST-proGhrelin8His protein was eluted with lysis buffer without lysozyme, supplemented with 15 mM reduced glutathione, GSH. About 50 mg of target protein, as evaluated by SDS-PAGE, was eluted. The eluate was supplemented with 1 mM DTT and about 7.5 mg of recombinant GST-TEV protease (see below) was added. The solution was incubated at 16 °C overnight to allow complete release of proghrelin8His from GST. Prior to Ni-NTA chromatography, 5 mM $CaCl_2$ was added, and the solution was spun down at 4,000 x g for 10 min. After elution from Ni-NTA resin, protein solution was dialyzed against buffer containing 10 mM Tris-HCl pH 8.5, 50 mM NaCl, 10% glycerol and 0.01% CHAPS using 3kDa cutoff membrane with 3 buffer exchanges. The protein was quantified by absorbance at A_{280} and qualified by SDS-PAGE. *Generation of GST-TEV protease*. The pMHT vector[2] (gift from Dr. Arbing, UCLA) was linearized by PCR reaction with primers outside of MBP gene: forw- 5′ gcgaccatcctccaaaatcgggagaaagcttgtttaaggggccg 3′ and rev- 5′ caataacctagtataggggacatggttaatttctcctctttaatg 3′. The resulting linear plasmid was used to clone GST gene in-frame at 5′ of TEV gene by in vitro Gibson assembly. The GST gene was prepared by PCR amplification with primers: form 5′ atgtcccctatactaggttattg and rev 5′ cgattttggaggatggtcgc 3′ from pGEX-GST-proGhrelin8His plasmid as a template. The final bacterial expression vector, pGST-TEV was used for expression and purification of GST-TEV fusion protein. BL21Gold chemically competent *E. coli* cells were transformed with pGST-TEV plasmid and entire transformation mixture was used to inoculate 50 mL LB supplemented with kanamycin. Next day, 10 mL of the overnight culture was used to inoculate 1L of LB. Four flasks with 1L LB each were inoculated with 10 mL of the overnight culture, and bacterial cultures were grown at 37 °C till OD600 ~ 0.8. was reached. After adding 1mM IPTG, culture was allowed to grow for 4 h, and cells were harvested by centrifugation. The cell pellet was resuspended in 200 mL lysis buffer containing 25 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM EDTA, 5% glycerol, 0.5 mg/mL lysozyme and 1 mM DTT. Cell suspension after freeze–thaw cycle was sonicated and cellular debris was removed by centrifugation. The soluble fraction was bound to glutathione resin by gravity. The resin was washed with 50 mL washing buffer (lysis buffer without lysozyme and DTT) and fusion GST-TEV protein was eluted with 25 mL washing buffer supplemented with 15 mM reduced glutathione (GSH). The elution was fractionated into 5 fractions and after SDS-PAGE/Coomassie analysis, fractions containing most of the GST-TEV protein were combined and dialyzed against buffer containing 25 mM Tris-HCl pH 8.0, 150 mM NaCl and 50% glycerol using 3kDa cutoff membrane with 3 buffer exchanges. The total yield was roughly 8 mg per 1L bacterial culture. *The baculovirus expression of mouse GOAT*. DH10Bac *E. coli* strain (ThermoFisher) was transformed with pFastBac1-mouseGOAT plasmid, and Bacmid DNA from 12 white colonies was analyzed by DNA sequencing and PCR reaction. All clones were transfected to sf9 cells and level

of mGOAT expression was compared between clones from cells producing P2. The clone with the highest mGOAT expression level was used to produce P3. For membrane isolation, sf9 cells were seeded at 2–4x10^6 /mL in total 1L volume of sf-900 serum-free medium (ThermoFisher) and infected with P3 virus. At day 2 post-infection, cells were harvested and resuspended in 40 mL of buffer containing 50 mM NaPi pH 7.2, 150 mM NaCl, 1 mM EDTA, 100 µM bis (4-nitrophenyl) phosphate, 2.5 µg/mL aprotinin, 10 µg/mL leupeptin, 10 µg/mL pepstatin A. Cell suspension was briefly sonicated and cell debris was removed by centrifugation at 3000g x 10 min. Membrane fraction was collected from supernatant by centrifugation at 100,000g x 1 h. Membrane pellet was resuspended in storage buffer (50 mM NaPi pH 7.2, 150 mM NaCl and 10% glycerol) and kept at -80 °C. *Acyltransferase assay*: The assay conditions included per 50 µl reaction: 50 µg of total sf9 membranes, various (for modality experiment) or fixed (for IC$_{50}$) concentrations of recombinant proghrelin8His peptide and various concentrations of tested compound (for IC$_{50}$), 100 µM palmitoyl CoA, 50 mM HEPES pH 7.0 and 1 µM [^{3}H] octanoyl CoA (~5.5 dpm/fmol—American Radioactive Chemicals). After incubation of the reaction mixture at 37C for 10 min, tubes were placed on ice and 10 µl of 1M HCl was added to each tube. Following the addition of 740 µl of cold quench buffer (50 mM NaPi pH 7.4, 10 mM Imidazole, 150 mM NaCl, 100 µM bis (4-nitrophenyl) phosphate, 1 mM phenylmethylsulfonyl fluoride and 0.1% Triton), 0.2 mL of 50% Ni-NTA slurry was added to each reaction. Tubes were incubated at 4C for 1 h with rotation to capture proGhrl8His. After washing Ni resin with 40 mM Imidazole, all bound proghrelin8His was eluted with 250 mM Imidazole, and an amount of octanylated proghrelin was assessed with scintillation counting.
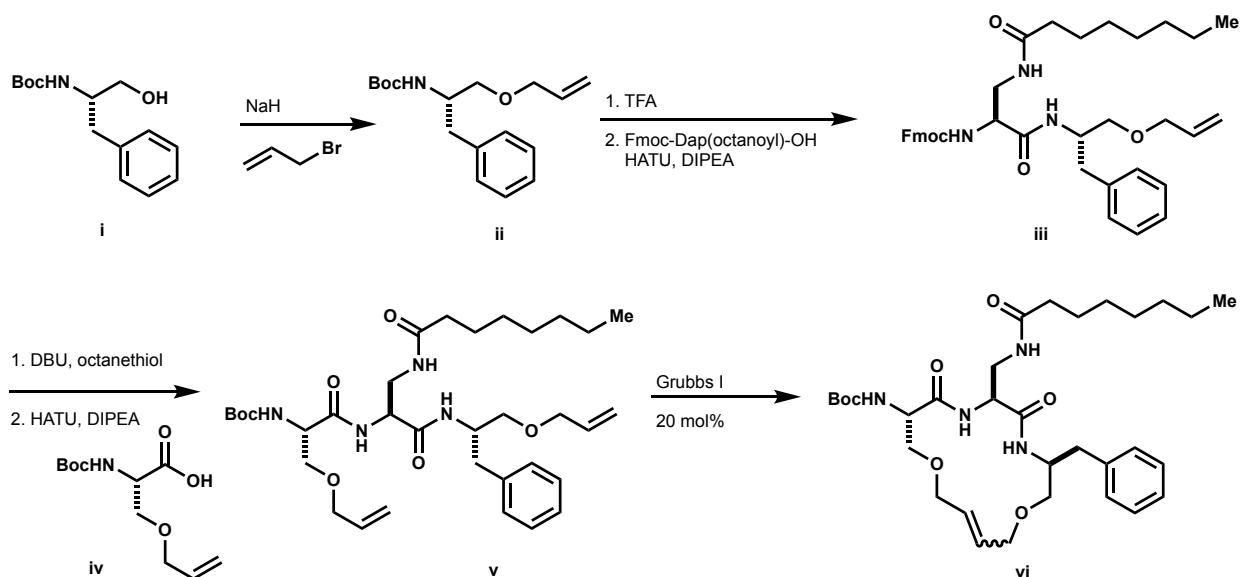
INS1-cellular Assay. *Generation of Recombinant Retrovirus for mGOAT Expression.* The mouse GOAT cDNA with C-terminal HA tag was amplified from pcDNA3.1-mouseGOAT-HA vector (gift from Brown and Goldstein lab [2]) using the following primers: mGOAT_attb1 (forward primer) 5′-gggg*acaagtttgtacaaaaaagcaggct*acc**atggattggctccagctc**—3′(*attb1* recombination site is in italics, 5′ of mGOAT coding sequence is in **bold**, and Kozak coding sequence is underlined) and mGOAT_attb2 (reverse primer) 5′-gggg*accactttgtacaagaaagctgggt*<u>**cta**</u>**agcgtaatctggaacatc** -3′ (attb2 recombination site is in *italics*, 3′ of mGAOT-HA coding sequence is in **bold**, and stop codon is underlined). The PCR product was cloned into donor vector pDONR221 (ThermoFisher) with BP clonase according to manufacturer's instructions. Positive clones were verified by DNA sequencing with M13F and M13R primers. The resulting entry plasmid pDONR221-mGOAT-HA was used to transfer mGOAT-HA cDNA to destination vector pBabe-puro (Addgene cat# 51070)[3] using LR clonase according to manufacturer's instructions. Positive clones were verified by DNA sequencing with pBABE-5 and pBABE-3 primers. The resulting plasmid, pBabe-puro-mGOAT-HA, was used to generate retrovirus. The retrovirus was packed in 293T PhoE (Phoenix-ECO AT0CC® CRL-3214™) cells and used to infect INS-1 cells. The stable INS/GOAT cells were selecting on 1 µg/mL puromycin and GOAT expression was confirmed with immunoblot of membrane fraction isolated from antibiotic resistant culture using anti-HA antibody.

Generation of Lentivirus for Expression of Ghrelin. The cDNA for mouse preproghrelin was subcloned from pcDNA3.1-preproghrelin (gift from Brown and Goldstein lab)[1] into pULTRA vector (Addgene cat# 24129)[4] with *XbaI* and *BamHI* restriction enzymes. Using of these restriction sites for cloning will result in the creation of bi-cistronic expression of ghrelin along with EGFP to facilitate identification of positive cells by fluorescence. The above restriction sites were engineered via PCR reaction using the following primers: ghrl_pultra_forw 5′-taccgagctc*tctaga***atgctgtcttcaggc** -3′ (5′ of preproghrelin coding sequence is in **bold**; *XbaI* site is in *italics*) and ghrl_pultra_rev 5′-agcggccgc*ggatcc*<u>**tta**</u>**cttgtcagctggc** -3′ (3′ of preproghrelin coding sequence is in **bold**, stop codon is underlined, and *BamHI* site is *italics*). The recombinant lentivirus encoding preproghrelin cDNA was packed in Lenti-X 293T cells (Takara cat# 632180) by co-transfection of pUltra-EGFP-mouse-preproghrelin, with packaging encoding plasmid pCMV

ΔR8.2 (Addgene cat# 12263) and envelope encoding plasmid pCMV-VSV g (Addgene cat# 8454). The INS/GOAT cells were infected with recombinant lentivirus to generate INS/GOAT/GHRL cell line. After few passages, the population of cells with puromycin resistance and ~90% fluorescence was saved for cell-based assay.

Cellular assay. INS/GOAT/GHRL cell line was routinely cultured in RPMI medium supplemented with 10% FBS, 1% Pen/Strep, 10 mM HEPES pH 7.2 and 50 µM 2-mercaptoethanol. Day 0: One 10 cm dish at full confluency was used to seed one 96-well plate. Day 1: Growth media was removed and cells were washed with PBS prior to adding fresh growth media as above but without 2-mercaptoethanol. The serial dilutions of tested compound at 50x concentration were prepared in vehicle containing growth media without 2-mercaptoethnol supplemented with 6% DMSO and were added to cells in duplicates. Day 2: 10 µL of growth media were removed from each well and amount of secreted acyl-ghrelin was measured with ELISA (Cayman cat# 10006307).

Analytical Methods. NMR spectra were recorded on Bruker Advance spectrometers (300 MHz, 400 MHz, 500 MHz) and are reported as δ values in ppm relative to CDCl$_3$ (calibrated to 7.26 ppm in ^{1}H NMR and 77.16 ppm in ^{13}C NMR, unless otherwise indicated). Splitting patterns are abbreviated as follows: singlet (s), doublet (d), triplet (t), quartet (q), multiplet (m), broad (br), and combinations thereof. Column chromatography was conducted on silica gel 60 (240–400 mesh) purchased from Sillicycle. Thin-layer chromatography (TLC) was performed using pre-coated, glass-backed plates (silica gel 60 PF254, 0.25 mm) and visualized using a combination of UV and potassium permanganate staining. HPLC analyses were carried out using an Agilent 1200 HPLC system equipped with an Agilent Quadrupole 6130 ESI-MS detector. Mobile phase was prepared with 0.1% TFA.

Scheme S1. Synthetic route towards **9A** and **9B**.

***tert*-butyl (*S*)-(1-(allyloxy)-3-phenylpropan-2-yl)carbamate** (**ii,** Scheme S1). To a solution of **i** (1.0 g, 3.98 mmol) in DMF (22 mL) at 0 °C, NaH was added (60% in mineral oil, 365 mg, 9.12 mmol) followed by allyl bromide (0.77 mL, 9.72 mmol). The resulting mixture was stirred at room temperature for 4 h. The reaction was quenched with saturated aqueous NH$_4$Cl and extracted with EtOAc (×2). The organics were dried (MgSO$_4$), filtered, and concentrated under reduced pressure. The resultant oil was purified by column chromatography using 9:1 Hex:EtOAc as eluents to afford **ii** (526 mg, 45%) as a clear oil. ^{1}H NMR (500 MHz, CDCl$_3$) δ 7.28 (q, *J* = 4.9 Hz,

2H), 7.21 (d, *J* = 7.4 Hz, 3H), 5.91 (m, 1H), 5.27 (q, *J* = 6.3 Hz, 1H), 5.19 (q, *J* = 4.0 Hz, 1H), 4.86 (s, 1H), 3.96 (m, 2H), 3.34 (m, 2H), 2.87 (m, 2H), 1.42 (s, 9H). HPLC/MS MH⁺ 192.1 (-Boc).

(9H-fluoren-9-yl) methyl((S)-1-(((S)-1-(allyloxy)-3-phenylpropan-2-yl)amino)-3-octanamido-1-oxopropan-2-yl)carbamate (iii). To a solution of **ii** (1 eq.) in CH_2Cl_2 (0.4 M) was added TFA (10% *v/v*). The resulting solution was stirred at room temperature until completion as determine by LCMS. The reaction mixture was concentrated under reduced pressure and the resultant TFA salt was used without further purification. The resultant residue (1.1 eq.), Fmoc-Dap(octanoyl)-OH (1 eq.) and HATU (1.1 eq.) was suspended in DMF (0.3 M). To the suspension DIPEA was added (2.5 eq.) and the resulting solution was stirred at room temperature until completion, as determined by LCMS. Upon completion, the reaction mixture was diluted with EtOAc and washed with saturated aqueous NH_4Cl (×2), water (×3), saturated aqueous $NaHCO_3$ (×1), and brine (×1). The organics were dried over $MgSO_4$, filtered, and concentrated under reduced pressure to afford **iii** (178 mg, 95%). ¹H NMR (500 MHz, $CDCl_3$) δ 7.77 (d, *J* = 7.5 Hz, 2H), 7.60 (d, *J* = 7.2 Hz, 2H), 7.40 (t, *J* = 7.5 Hz, 2H), 7.32 (m, 2H), 7.25 (d, *J* = 14.7 Hz, 2H), 7.19 (q, *J* = 3.5 Hz, 3H), 6.97 (d, *J* = 7.8 Hz, 1H), 6.41 (d, *J* = 4.3 Hz, 1H), 6.12 (s, 1H), 5.84 (m, 1H), 5.22 (d, *J* = 17.1 Hz, 1H), 5.12 (d, *J* = 10.3 Hz, 1H), 4.37 (t, *J* = 8.9 Hz, 1H), 4.31 (m, 2H), 4.21 (t, *J* = 7.1 Hz, 2H), 3.94 (s, 2H), 3.69 (d, *J* = 11.6 Hz, 1H), 3.49 (t, *J* = 7.0 Hz, 1H), 3.37 (s, 2H), 2.87 (m, 2H), 2.13 (t, *J* = 7.2 Hz, 2H), 1.74 (s, 1H), 1.60 (s, 2H), 1.26 (t, *J* = 5.2 Hz, 8H), 0.85 (t, *J* = 7.0 Hz, 3H). ¹³C NMR (126 MHz, $CDCl_3$) δ 175.37, 169.47, 143.74, 141.29 (d, *J* = 2.4 Hz), 137.81, 134.40, 129.30, 128.44, 127.80, 127.15 (d, *J* = 2.7 Hz), 126.53, 125.20, 120.03, 117.20, 72.14, 69.90, 67.52, 56.73, 50.35, 47.07, 42.15, 37.54, 36.51, 31.67, 29.24, 29.01, 25.66, 22.62, 14.06. HPLC/MS MH⁺ 626.1.

Boc-Ser(O-allyl)-OMe. To a solution of Boc-Ser-OMe (1.28 g, 5.85 mmol) in THF (11.7 mL) a solution of allyl ethyl carbonate (1.54 mL, 11.7 mmol), allyl palladium chloride dimer (43 mg, 0.12 mmol) and PPh_3 (138 mg, 0.53 mmol) was added dropwise to THF (5.85 mL). The resulting reaction mixture was refluxed under argon overnight. The reaction mixture was concentrated under a reduced pressure and the resultant residue was purified by column chromatography using 9:1 Hex : EtOAc as eluents to afford the desired product (1.24 g, 82%) as a clear oil. ¹H NMR (500 MHz, $CDCl_3$) δ 5.82 (m, 1H), 5.37 (d, *J* = 8.1 Hz, 1H), 5.23 (m, 1H), 5.17 (d, *J* = 10.4 Hz, 1H), 4.42 (t, *J* = 4.4 Hz, 1H), 3.97 (t, *J* = 5.7 Hz, 2H), 3.84 (q, *J* = 4.1 Hz, 1H), 3.75 (s, 9H), 3.64 (q, *J* = 4.2 Hz, 1H). ¹³C NMR (126 MHz, $CDCl_3$) δ 171.23, 155.51, 134.06, 117.40, 79.99, 72.24, 69.94, 53.99, 52.48, 28.32. HPLC/MS MH⁺ 1601.1 (-Boc).

Boc-Ser(O-allyl)-OH (iv). To a solution of Boc-Ser(*O*-allyl)-OMe (1.24 g, 4.78 mmol) in a mixture of THF/MeOH/H_2O (1:1:1, 75 mL) LiOH•H_2O (502 mg, 11.96 mmol) was added and the reaction was stirred at room temperature until completion as determined by TLC. THF and MeOH were removed under reduced pressure and the resulting solution was partitioned between 1N HCl (70 mL) and CH_2Cl_2 (150 mL). The aqueous layer was extracted with CH_2Cl_2 (×2) and the combined organics were dried over $MgSO_4$, filtered, and concentrated under reduced pressure to afford the desired product, **iv** (1.1 g) as a colorless oil, which was used without further purification. ¹H NMR (500 MHz, $CDCl_3$) δ 5.85 (m, 1H), 5.40 (d, *J* = 7.9 Hz, 1H), 5.25 (m, 1H), 5.19 (q, *J* = 3.9 Hz, 1H), 4.45 (t, *J* = 3.9 Hz, 1H), 4.01 (d, *J* = 5.7 Hz, 2H), 3.90 (d, *J* = 4.5 Hz, 1H), 3.67 (q, *J* = 4.4 Hz, 1H).

***tert*-butyl ((6S,9S,12S)-6-benzyl-9-(octanamidomethyl)-8,11-dioxo-4,14-dioxa-7,10-diazaheptadeca-1,16-dien-12-yl)carbamate (v)**. To a solution of the Fmoc-protected amine **iv** (1 eq.) and octanethiol (10 eq.) in THF (0.1 M) DBU (1.1 eq.) was added. The resultant solution was stirred at room temperature for 15 min and then concentrated under reduced pressure. The resultant residue was purified by column chromatography. The resultant residue (1.1 eq.), **iv** (1 eq.) and HATU (1.1 eq.) was suspended in DMF (0.3 M). To the suspension was added DIPEA

(2.5 eq.) and the resulting solution was stirred at room temperature until completion, as determined by LCMS. Upon completion, the reaction mixture was diluted with EtOAc and washed with saturated aqueous NH₄Cl (×2), water (×3), saturated aqueous NaHCO₃ (×1), and brine (×1). The organics were dried over MgSO₄, filtered, and concentrated under reduced pressure to afford **v** (195 mg). ^{1}H NMR (500 MHz, CDCl₃) δ 7.98 (d, J = 4.7 Hz, 1H), 7.28 (t, J = 3.8 Hz, 2H), 7.20 (q, J = 3.4 Hz, 3H), 7.08 (d, J = 8.6 Hz, 1H), 6.01 (s, 1H), 5.87 (m, 2H), 5.42 (d, J = 4.9 Hz, 1H), 5.28 (m, 1H), 5.24 (m, 1H), 5.19 (dd, J = 1.5, 10.5 Hz, 1H), 5.16 (dd, J = 1.5, 10.5 Hz, 1H), 4.40 (m, 1H), 4.35 (t, J = 3.8 Hz, 1H), 4.23 (t, J = 4.9 Hz, 1H), 3.98 (m, 4H), 3.78 (q, J = 4.8 Hz, 1H), 3.65 (q, J = 5.7 Hz, 1H), 3.58 (q, J = 4.9 Hz, 1H), 3.46 (m, 1H), 3.38 (m, 2H), 2.94 (q, J = 7.0 Hz, 1H), 2.77 (q, J = 7.2 Hz, 1H), 2.01 (q, J = 4.9 Hz, 1H), 1.64 (s, 6H), 1.56 (s, 1H), 1.46 (s, 9H), 1.26 (s, 8H), 0.87 (t, J = 6.9 Hz, 3H). ^{13}C NMR (126 MHz, CDCl₃) δ 175.55, 170.73, 134.72, 133.96, 129.31, 129.16, 128.37, 126.38, 117.78, 117.02, 72.29, 72.26, 72.16, 70.33, 69.43, 55.64, 54.95, 50.17, 41.78, 37.50, 36.29, 31.66, 29.22, 28.98, 28.34, 25.57, 22.61, 14.05. HPLC/MS MH⁺ 631.1.

tert-butyl ((3*S*,6*S*,9*S*)-3-benzyl-6-(octanamidomethyl)-5,8-dioxo-1,11-dioxa-4,7-diazacyclopentadec-13-en-9-yl)carbamate (**vi**). The diene, **v** (158 mg, 0.25 mmol) was dissolved in CHCl₃ (0.01 M) and sparged with argon for 10 min, and then a solution of Grubbs I (20 mol %) in CHCl₃ (0.02 M) was added. The resultant solution was stirred under argon until completion as determined by LCMS. Column purification afforded **vi** (134 mg) as a mixture of alkene isomers. ^{1}H NMR (500 MHz, CDCl₃) δ 7.25 (m, 2H), 7.18 (m, 3H), 5.83 (m, 1H), 5.49 (dd, J = 8.5, 32.0 Hz, 1H), 5.20 (m, 1H), 4.33 (m, 2H), 4.12 (m, 1H), 3.97 (m, 2H), 3.85 (q, J = 5.7 Hz, 1H), 3.77 (m, 1H), 3.69 (q, J = 5.1 Hz, 1H), 3.57 (m, 2H), 3.44 (m, 1H), 3.31 (m, 2H), 2.91 (m, 1H), 2.76 (m, 1H), 2.25 (s, 1H), 2.13 (q, J = 7.2 Hz, 1H), 2.04 (m, 1H), 1.90 (d, J = 12.7 Hz, 3H), 1.82 (s, 5H), 1.71 (s, 2H), 1.56 (m, 2H), 1.50 (s, 1H), 1.42 (m, 11H), 1.23 (s, 13H), 0.85 (t, J = 4.1 Hz, 3H). ^{13}C NMR (126 MHz, CDCl₃) δ 176.40, 175.77, 175.58, 175.52, 171.65, 170.67, 170.57, 169.48, 169.24, 169.00, 168.77, 155.24, 155.18, 138.20, 129.37, 129.31, 129.15, 128.43, 128.42, 128.38, 128.34, 126.47, 126.44, 117.74, 72.23, 72.13, 71.19, 70.31, 70.22, 69.23, 69.46, 69.34, 69.02, 68.69, 67.29, 67.18, 56.44, 54.34, 50.95, 41.74, 37.45, 36.37, 36.32, 35.54, 35.06, 31.68, 31.66, 29.30, 29.27, 29.01, 28.33, 26.95, 26.86, 26.34, 26.32, 26.13, 22.61, 14.06. HPLC/MS MH⁺ 603.1.

N-(((3*S*,6*S*,9*S*)-9-(2-aminoacetamido)-3-benzyl-5,8-dioxo-1,11-dioxa-4,7-diazacyclopentadec-13-en-6-yl)methyl)octanamide (**9**). To a solution of **ii** (1 eq.) in CH₂Cl₂ (0.4 M) was added TFA (10% v/v). The resulting solution was stirred at room temperature until completion as determine by LCMS. The reaction mixture was concentrated under reduced pressure and the resultant TFA salt was used without further purification. The resultant residue (1.1 eq.), Boc-Gly-OH (1 eq.) and HATU (1.1 eq.) was suspended in DMF (0.3 M). DIPEA (2.5 eq.) was added to the suspension and the resulting solution was stirred at room temperature until completion, as determined by LCMS. Upon completion, the reaction mixture was diluted with EtOAc and washed with saturated aqueous NH₄Cl (×2), water (×3), saturated aqueous NaHCO₃ (×1), and brine (x1). The organics were dried over MgSO₄, filtered, and concentrated under reduced pressure to afford Boc-**vi** (178 mg, 95%). To a solution of Boc-**vi**. (1 eq.) in CH₂Cl₂ (0.4 M), TFA (10% v/v) was added. The resulting solution was stirred at room temperature until completion, as determined by LCMSvto afford **9** (50 mg) which was purified by preparative HPLC to separate the alkene isomers. HRMS m/z: [M + H]⁺ Calculated for C₃₂H₅₂N₅O₉ 602.3918; Found 602.3871. **Z-9:** ^{1}H NMR (500 MHz, MeOD) δ 8.35 (d, J = 7.3 Hz, 1H), 7.26 (d, J = 6.6 Hz, 3H), 7.18 (m, 2H), 5.79 (m, 2H), 4.46 (d, J = 2.3 Hz, 1H), 4.38 (m, 1H), 4.17 (d, J = 14.1 Hz, 1H), 4.03 (m, 3H), 3.83 (dd, J = 3.2, 9.8 Hz, 1H), 3.76 (dd, J = 8.1, 14.3 Hz, 1H), 3.54 (dd, J = 5.2, 10.5 Hz, 1H), 3.48 (q, J = 5.2 Hz, 1H), 3.35 (dd, J = 2.2, 10.5 Hz, 1H), 3.19 (dd, J = 2.4, 14.2 Hz, 1H), 2.92 (dd, J = 6.5, 13.0 Hz, 1H), 2.78 (dd, J = 8.7, 13.4 Hz, 1H), 2.20 (m, 2H), 1.93 (m, 2H), 1.85 (m,

2H), 1.75 (s, 1H), 1.59 (t, *J* = 6.7 Hz, 2H), 1.42 (d, *J* = 12.2 Hz, 2H), 1.31 (m, 11H), 0.88 (t, *J* = 6.7 Hz, 3H). ^{13}C NMR (126 MHz, MeOD) δ 172.09, 170.08, 166.86, 138.15, 130.16, 129.01, 128.04, 127.50, 126.12, 69.47, 69.22, 69.02, 66.63, 57.29, 57.19, 55.21, 52.59, 41.30, 41.25, 40.60, 36.47, 35.67, 34.96, 34.47, 31.48, 29.01, 28.76, 26.41, 26.32, 25.82 (d, *J* = 2.9 Hz, 1C), 25.75, 25.65, 22.30. ***E*-9**: ^{1}H NMR (500 MHz, DMSO) δ 7.35 (d, *J* = 8.4 Hz, 1H), 7.28 (t, *J* = 7.4 Hz, 2H), 7.24 (t, *J* = 4.1 Hz, 2H), 7.20 (m, 1H), 5.87 (m, 2H), 4.57 (t, *J* = 5.4 Hz, 1H), 4.35 (m, 2H), 4.10 (dd, *J* = 4.7, 12.2 Hz, 2H), 4.01 (dd, *J* = 5.7, 11.4 Hz, 1H), 3.94 (dd, *J* = 6.3, 12.7 Hz, 1H), 3.77 (m, 2H), 3.72 (s, 2H), 3.47 (m, 2H), 3.37 (dq, *J* = 3.2, 9.5 Hz, 2H), 2.92 (q, *J* = 6.5 Hz, 1H), 2.82 (q, *J* = 7.4 Hz, 1H), 2.19 (m, 2H), 1.89 (m, 1H), 1.59 (m, 2H), 1.42 (m, 1H), 1.31 (s, 9H), 0.89 (t, *J* = 6.9 Hz, 3H). HPLC/MS MH$^+$ 560.3.
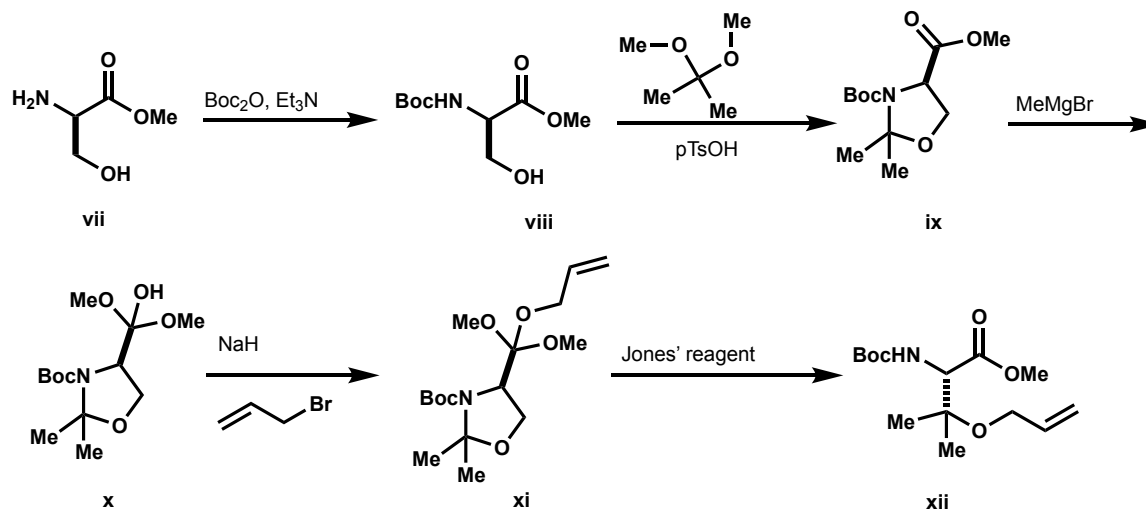
Scheme S2. Synthesis of starting material towards **17A** and **17B**.

Boc-*D*-Ser-OMe (**viii,** Scheme S2). To a solution of **vii** (1.0 g, 8.39 mmol) and Et$_3$N (2.51 mL, 18.04 mmol) in THF (22 mL) at 0 °C, a solution of Boc$_2$O was added (1.81 g, 8.31 mmol) in THF (7 mL) dropwise over 20 min. The resulting solution was warmed to room temperature, stirred at that temperature overnight, and then stirred at 50 °C for 3 h. The solvent was removed under reduced pressure and the resultant residue was taken up in ether (15 mL) and saturated aqueous NaHCO$_3$ (20 mL). The aqueous phase was extracted with ether (×3). The organics were combined, dried over MgSO$_4$, filtered, and concentrated under a reduced pressure to afford **viii** (1.61 g, 87%) as a colorless oil which was used without further purification. ^{1}H NMR (500 MHz, CDCl$_3$) δ 5.46 (s, 1H), 4.38 (s, 1H), 3.95 (m, 1H), 3.89 (m, 1H), 3.77 (s, 3H), 2.46 (t, *J* = 6.2 Hz, 1H), 1.44 (s, 9H).

*3-(tert-butyl) 4-methyl (R)-2,2-dimethyloxazolidine-3,4-dicarboxylate (*i*x*): To a solution of **viii** (1.67 g, 7.33 mmol) in CH$_2$Cl$_2$ (9 mL) at 0 °C, 2,2-dimethoxypropane (4.50 mL, 36.66 mmol) and pTsOH (140 mg, 0.73 mmol) were added. The resulting mixture was warmed to room temperature and then stirred at room temperature overnight. The reaction mixture was then poured into saturated aqueous NaHCO$_3$ (10 mL) and extracted with ether (×3). The organics were combined and washed with NaHCO$_3$ and brine, dried over MgSO$_4$, filtered, and concentrated under reduced pressure. The residue was purified by column chromatography using 1:1 Hex:EtOAc as eluents to afford **ix** (1.63 g, 86%) as a colorless oil. ^{1}H NMR (500 MHz, CDCl$_3$) δ 4.48 (dd, *J* = 2.4, 6.8 Hz, 0.4H), 4.37 (dd, *J* = 3.0, 6.8 Hz, 0.6H), 4.13 (m, 1H), 4.03 (m, 1H), 3.75 (s, 3H), 1.66 (s, 1.89H), 1.63 (s, 1.42H), 1.52 (s, 1.77H), 1.49 (s, 5.11H), 1.40 (s, 5.29H).

***tert*-butyl(*R*)-4-(2-hydroxypropan-2-yl)-2,2-dimethyloxazolidine-3-carboxylate** (**x**). To a solution of **ix** (1.63 g, 6.29 mmol) in THF (50 mL) at −20 °C MeMgBr (1.95 M in Et$_2$O, 10.8 mL, 21.13 mmol) was added dropwise. The resulting mixture was stirred at 0 °C for 4 h. Saturated, aqueous NH$_4$Cl was added to the reaction mixture to quench the reaction, and then extracted with EtOAc (×3). The organics were combined, washed with brine (×2), dried over MgSO$_4$, filtered, and concentrated under reduced pressure. The residue was purified by column chromatography using 3:1 Hex:EtOAc as eluents to afford **x** (982 mg, 60%). ^{1}H NMR (500 MHz, CDCl$_3$) δ 4.00 (m, 2H), 3.79 (s, 1H), 1.59 (s, 3H), 1.50 (s, 12H), 1.17 (d, *J* = 7.0 Hz, 6H). HPLC/MS MH$^+$ 186.2 (-Boc +Na).

***tert*-butyl (*R*)-4-(2-(allyloxy)propan-2-yl)-2,2-dimethyloxazolidine-3-carboxylate** (**xi**). To a solution of **x** (289 mg, 1.12 mmol) in DMF (5.2 mL) at 0 °C, NaH (60% in mineral oil, 90 mg, 2.24 mmol) was added, followed by allyl bromide (0.11 mL, 1.25 mmol). The resulting solution was warmed to room temperature, stirred for 1 h, and cooled back to 0 °C. The reaction was quenched with the addition of saturated aqueous NH$_4$Cl and organics were extracted with EtOAc (×3). The organics were combined, washed with brine, dried over MgSO$_4$, filtered, and concentrated under reduced pressure. The resultant residue was purified by column chromatography using 3:1 Hex : EtOAc to afford **xi** (89 mg). ^{1}H NMR (500 MHz, CDCl$_3$) δ 5.88 (m, 1H), 5.25 (m,1H), 5.09 (d, 1H), 4.20 (d, *J* = 7.7 Hz, 1H), 3.94 (m, 2H), 3.87 (m, 1H), 1.60 (s, 3H), 1.49 (s, 3H), 1.48 (s, 9H), 1.22 (s, 3H), 1.17 (s, 3H). ^{13}C NMR (126 MHz, CDCl$_3$) δ 145.18, 136.00, 115.46, 80.16, 62.87, 29.71, 28.35. HPLC/MS MH$^+$ 200.2 (-Boc).

(*S*)-3-(allyloxy)-2-((*tert*-butoxycarbonyl)amino)-3-methylbutanoic acid (**xii**). To a solution of **xi** (80 mg, 0.27 mmol) in acetone (3 mL), Jones' reagent (2.5 M in H$_2$O, 0.16 mL, 0.40 mmol) was added at 0 °C. The resulting mixture was warmed to room temperature, and then stirred at this temperature overnight. To the reaction mixture, celite (100 mg) and isopropanol (0.5 mL) were added, and the resulting precipitate was filtered through off through a plug of celite. The filtrate was adjusted to pH 9 with aqueous NaHCO$_3$, and then concentrated under reduced pressure. The aqueous layer was washed with ether (×2) and acidified to pH 3 with citric acid. The resulting solution was extracted with EtOAc (×3) and the combined extracts were washed with brine (×2), dried over MgSO$_4$, filtered, and concentrated under reduced pressure to afford **xii** (50 mg, 68%), which was used without further purification. ^{1}H NMR (500 MHz, CDCl$_3$) δ 5.89 (m, *J* = 5.5 Hz, 1H), 5.30 (q, *J* = 6.2 Hz, 1H), 5.20 (q, *J* = 3.8 Hz, 1H), 4.38 (s, 1H), 4.04 (m, 2H), 1.44 (s, 9H), 1.35 (s, 3H), 1.24 (s, 3H). ^{13}C NMR (126 MHz, CDCl$_3$) δ 172.36, 56.00, 133.90, 117.54, 80.37, 78.67, 63.60, 28.26, 22.64, 21.11.

***N*-(((3*S*,6*S*,9*S*)-9-(2-aminoacetamido)-3-benzyl-10,10-dimethyl-5,8-dioxo-1,11-dioxa-4,7-diazacyclopentadec-13-en-6-yl)methyl)octanamide** (**17**). Macrocycles **17** were synthesized in the same manner as **9** with the modified residue substituted in where necessary. **Z-17**: ^{1}H NMR (500 MHz, MeOD) δ 8.06 (q, *J* = 3.7 Hz, 1H), 7.64 (m, 1H), 7.54 (t, *J* = 4.0 Hz, 1H), 7.27 (t, *J* = 7.3 Hz, 2H), 7.23 (t, *J* = 4.1 Hz, 2H), 7.19 (m, 1H), 5.79 (m, 2H), 4.56 (s, 1H), 4.49 (m, 1H), 4.02 (m, 2H), 3.94 (t, *J* = 5.3 Hz, 2H), 3.90 (t, *J* = 3.8 Hz, 1H), 3.74 (q, *J* = 13.5 Hz, 2H), 3.62 (m, 1H), 3.45 (q, *J* = 4.2 Hz, 1H), 3.36 (q, *J* = 4.6 Hz, 1H), 3.34 (s, 1H), 2.97 (q, *J* = 7.1 Hz, 1H), 2.82 (q, *J* = 7.0 Hz, 1H), 2.14 (t, *J* = 7.6 Hz, 2H), 1.57 (m, 2H), 1.43 (s, 2H), 1.29 (m, 9H), 1.27 (s, 3H), 1.24 (s, 3H), 0.89 (t, *J* = 7.0 Hz, 3H). ^{13}C NMR (126 MHz, MeOD) δ 175.69, 170.13, 169.79, 165.86, 138.42, 131.02, 128.93, 128.09, 127.22, 126.12, 77.02, 69.39, 67.88, 60.51, 60.33, 54.35, 53.18, 41.01, 40.15, 35.83, 35.70, 31.51, 29.00, 28.80, 25.47, 22.62, 22.31, 18.77, 13.02. **E-17**: ^{1}H NMR (500 MHz, MeOD) δ 7.25 (m, 5H), 5.82 (m, 2H), 4.50 (s, 1H), 4.41 (t, *J* = 5.4 Hz, 1H), 4.24 (q, *J* = 5.6 Hz, 1H), 4.17 (q, *J* = 6.3 Hz, 1H), 4.05 (m, 2H), 3.96 (q, *J* = 5.5 Hz, 1H), 3.73 (d, *J* = 4.3 Hz, 2H), 3.52 (t, *J* = 4.7 Hz, 2H), 3.43 (q, *J* = 4.4 Hz, 1H), 3.34 (q, *J* = 4.2 Hz, 1H), 2.97 (m, 1H), 2.78

(q, J = 7.5 Hz, 1H), 2.18 (m, 2H), 1.60 (t, J = 7.2 Hz, 2H), 1.36 (s, 3H), 1.30 (m, 11H), 0.89 (t, J = 6.9 Hz, 3H). ^{13}C NMR (126 MHz, MeOD) δ 176.54, 169.93, 138.18, 131.29, 129.02, 128.82, 128.13, 126.14, 77.43, 68.71, 65.98, 60.13, 58.53, 55.14, 51.81, 40.56, 36.52, 35.57, 31.49, 29.04, 28.84, 25.54, 22.47, 22.33, 21.13, 13.03. HPLC/MS MH$^+$ 588.3.
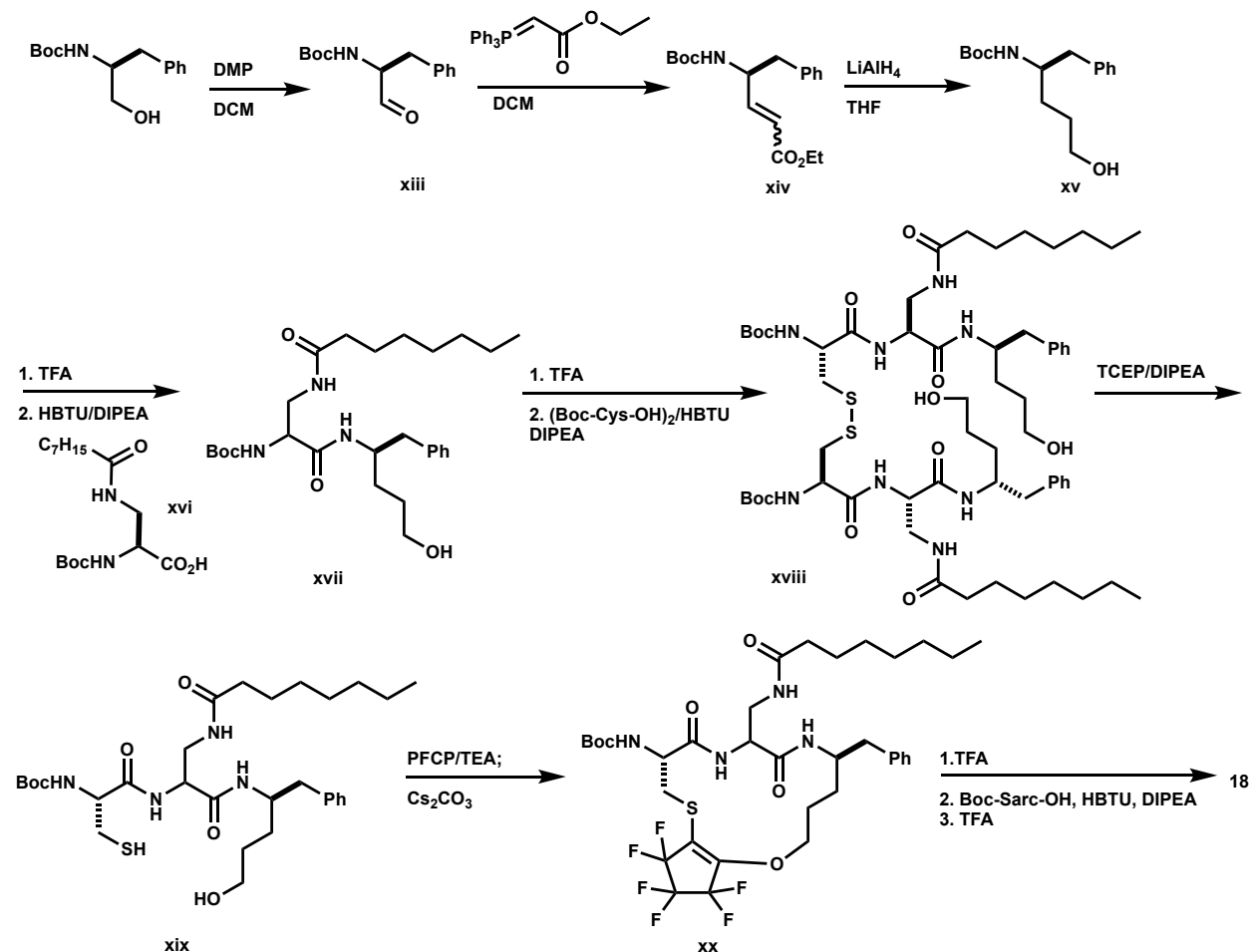
Scheme S3. Synthetic route towards **18**.

tert-butyl (S)-(1-oxo-3-phenylpropan-2-yl)carbamate (xiii, Scheme S3**).** A solution of *N*-Boc-(L)-phenylalaninol (1.51 g, 6 mmol) in CH$_2$Cl$_2$ (6 mL) was added dropwise over three minutes to a solution of DMP (2.80 g, 6.6 mmol) in CH$_2$Cl$_2$ (18 mL). The reaction was allowed to stir at room temperature and monitored by TLC. Et$_2$O (12 mL) and a 25% solution of Na$_2$S$_2$O$_3$ in saturated aq. NaHCO$_3$ was added. After stirring for 5 minutes, Et$_2$O (12 mL) was added again. The solution was partitioned, and the organic phase washed with saturated aq. NaHCO$_3$ (×2), H$_2$O (×2). The combined organic layers were dried over MgSO$_4$ and concentrated *in vacuo* to afford a white solid. (1.33 g, 88%) ^{1}H NMR (400 MHz, CDCl$_3$) δ 9.63 (s, 1H), 7.33-7.23 (m, 3H), 7.18-7.16 (m, 2H), 5.07-5.03 (br s, 1H), 4.51-4.35 (m, 1H), 3.11 (d, J = 6.5 Hz, 2H), 1.43 (s, 9H).

ethyl (S)-4-((tert-butoxycarbonyl)amino)-5-phenylpent-2-enoate (xiv). To an ice cooled solution of **xiii** (1.33 g, 5.54 mmol) in dry CH$_2$Cl$_2$ (55 mL) was added Ph$_3$PCHCO$_2$Et (2.51 g, 7.20 mmol). The reaction mixture was stirred at room temperature and monitored by TLC. After

completion, the solvent was evaporated and the residue was purified by flash chromatography over silica gel using 8:2 Hex:EtOAc to furnish a white solid as a mixture of *E* and *Z* isomers. (1.49 g, 84%) ^{1}H NMR (400 MHz, CDCl$_3$) δ 7.32-7.25 (m, 2H), 7.23-7.21 (m, 1H), 7.18-7.16 (m, 2H), 6.90 (dd, *J*=15.7, 5.0 Hz, 1H), 5.85 (dd, *J*=15.7, 1.6 Hz, 1H), 4.61 (br s, 1H), 4.51 (br s, 1H), 4.16 (q, *J* = 7.2 Hz, 2H), 2.94-2.83 (m, 2H), 1.39 (s, 9H), 1.28 (t, *J* = 7.2 Hz, 3H).

tert-butyl (*R*)-(5-hydroxy-1-phenylpentan-2-yl)carbamate (xv). LiAlH$_4$ (0.53 g, 14 mmol) was added portion wise to a stirred solution of **xiv** (1.49 g, 4.68 mmol) in anhydrous THF (50 mL) at 0 °C. The reaction was allowed to stir at 0 °C and monitored by TLC. The mixture was quenched by addition of EtOAc followed by water at 0 °C. The aqueous layer was extracted with EtOAc and the organic layers were dried over anhydrous Na$_2$SO$_4$. After concentration, the crude product was purified over silica gel using a gradient eluent system of 7:3 to 1:1 Hex:EtOAc to afford a colorless oil. (0.189 g, 15%). ^{1}H NMR (400 MHz, CDCl$_3$) δ 7.33-7.24 (m, 2H), 7.24-2.16 (m, 3H), 3.84 (br s, 1H), 3.63 (t, *J* = 4.1 Hz, 2H), 2.77 (d, *J* = 6.5 Hz, 2H), 1.69-1.54 (m, 4H), 1.40 (s, 9H).

(*S*)-2-((*tert*-butoxycarbonyl)amino)-3-octanamidopropanoic acid (xvi). To a solution of octanoic acid (34.7 mmol, 1.0 eq) in CH$_2$Cl$_2$ (15 mL) and MeCN (10 mL) is added *N*-hydroxy succinimide (1.0 eq) and DCC (1.0 eq). The reaction is allowed to stir at room temperature and monitored by TLC. After completion, the mixture is concentrated *in vacuo*. To the NHS ester (8.3 mmol, 1.0 eq) is added a solution of *N*-Boc-DAP-OH (1.2 eq.) in 65% EtOH and Et$_3$N (2.5 eq). The reaction is allowed to reflux for 20 hours after which it is concentrated *in vacuo*, diluted with EtOAc and washed with 1M HCl (30 mL×2), and saturated aq. NaCl (25 mL×1). The organic layer is dried over MgSO$_4$, filtered and concentrated *in vacuo* (2.7 g, 98%).

tert-butyl (1-(((*R*)-5-hydroxy-1-phenylpentan-2-yl)amino)-3-octanamido-1-oxopropan-2-yl)carbamate (xvii). **xv** (190 mg, 0.7 mmol) is subjected to TFA (1.56 mL, 20 mmol) in CH$_2$Cl$_2$ (8 mL) and the reaction is stirred until starting material is fully consumed. TFA is removed under reduced pressure to afford the free amine, which is carried forward without further protection. The TFA salt of the free amine was combined with **xvi** (254 mg, 0.77 mmol), HBTU (292 mg, 0.77 mmol) and DIPEA (0.37 mL, 2.1 mmol) in DMF (2.5 mL). After consumption of the starting material is observed, the reaction is diluted with EtOAc and sequentially washed with saturated aq. solutions of NH$_4$Cl, NaHCO$_3$ and NaCl. The organic layers are combined, dried and concentrated *in vacuo*. The product is obtained as a white solid after purification over silica gel using a gradient eluent system of 1.25% to 3% MeOH in CHCl$_3$ (305 mg, 89%).

di-tert-butyl((11*S*,14*R*,19*R*,22*S*)-11,22-bis(((*R*)-5-hydroxy-1-phenylpentan-2-yl)carbamoyl)-8,13,20,25-tetraoxo-16,17-dithia-9,12,21,24-tetraazadotriacontane-14,19-diyl)dicarbamate (xviii). **xvii** (288 mg, 0.6 mmol) is subjected to TFA (1.34 mL, 18 mmol) in CH$_2$Cl$_2$ (7 mL) and the reaction is stirred until starting material is fully consumed. TFA is removed under reduced pressure to afford the free amine, which is carried forward without further protection. The TFA salt of the free amine is dissolved in DMF (0.6 mL). To this solution is added (Boc-Cys-OH)$_2$ (106 mg, 0.24 mmol), HBTU (224 mg, 0.59 mg), and DIPEA (0.32 mL, 1.8 mmol). The reaction is allowed to stir at room temperature and is monitored by TLC. After consumption of the starting material is observed, the reaction is diluted with EtOAc and sequentially washed with saturated aq. solutions of NH$_4$Cl, NaHCO$_3$ and NaCl. The organic layers are combined, dried and concentrated *in vacuo*. The product is obtained as a white solid after purification over silica gel using a gradient eluent system of 1.25% to 3% MeOH in CHCl$_3$ (127 mg, 20%).

tert-butyl((2*R*)-1-((1-(((*R*)-5-hydroxy-1-phenylpentan-2-yl)amino)-3-octanamido-1-oxopropan-2-yl)amino)-3-mercapto-1-oxopropan-2-yl)carbamate (xix). TCEP.HCl (27 mg,

0.093 mmol) followed by DIPEA (78 µL, 0.42 mmol) is added to a solution of **xviii** (100 mg, 0.084 mmol) in DMF (0.42 mL). The reaction is allowed to stir at room temperature and monitored by TLC. After consumption of the starting material, the mixture is diluted with EtOAc, washed with saturated aq. NaHCO$_3$ and brine. The organic layers are dried over MgSO$_4$ and carried forward to the next step immediately without further purification.

tert-butyl((3R,9R)-9-benzyl-14,14,15,15,16,16-hexafluoro-6-(octanamidomethyl)-4,7-dioxo-3,4,5,6,7,8,9,10,11,12,15,16-dodecahydro-2H,14H-cyclopenta[b][1]oxa4thia[8,11]diazacyclo pentadecin-3-yl)carbamate (xx). To a solution of **xix** (47 mg, 0.079 mmol) in DMF (8 mL) was added PFCP (0.40 mL, 1M in MeCN) at 0 °C, followed by triethylamine (16 µL, 0.12 mmol). The mixture was stirred under argon at 0 °C and the reaction was monitored by TLC. After consumption of the starting material, excess PFCP was removed by concentrating *in vacuo*. The mixture was cooled back down to 0 °C and Cs$_2$CO$_3$ (39 mg, 0.12 mmol) was added. The reaction was allowed to stir at 0 °C with TLC monitoring. After completion, the reaction was diluted with EtOAc and washed with brine. The organic layer was dried over Na$_2$SO$_4$ and concentrated *in vacuo.* The reaction mixture was purified using flash chromatography over silica gel with an eluent system of 1 to 5% MeOH in CHCl$_3$ to afford diastereomeric compounds having the same desired product mass (0.029 mmol and 0.013 mmol, 53% combined yield). HRMS (ESI) calculated for C$_{39}$H$_{49}$F$_6$N$_4$O$_6$S [M+H]$^+$ 767.3272, found 767.3248.

N-(((3R,6S,9R)-9-benzyl-14,14,15,15,16,16-hexafluoro-3-(2-(methylamino)acetamido)-4,7-dioxo-3,4,5,6,7,8,9,10,11,12,15,16-dodecahydro-2H,14H-cyclopenta[b][1]oxa4thia[8,11]diazacyclopentadecin-6-yl)methyl)octanamide (18). The major diastereoisomer **xix** (10 mg, 0.013 mmol) was subjected to deprotection with TFA (0.05 mL, 0.65 mmol) in CH$_2$Cl$_2$ (0.1 mL) until starting material was fully consumed on TLC. TFA was removed under reduced pressure and the crude product was dissolved in DMF (0.1 mL). Boc-Sarc-OH (6 mg, 0.032 mmol), HBTU (12 mg, 0.032 mmol) and DIPEA (0.03 mL, 0.15 mmol) were added and the reaction mixture was stirred until starting material was fully consumed by TLC. The reaction was diluted with EtOAc, washed with brine, dried over MgSO$_4$, filtered and concentrated. The final product was obtained by HPLC purification {Sunfire C18 column, 20-80% ACN in H$_2$O, 2-8 minutes, 20 mL/min} as two compounds- an orange solid and a white solid having the same desired product mass. (4mg and 1 mg respectively, 23% combined yield over three steps). Following NMR data for major diastereomer. ^{1}H NMR (500 MHz, MeOD) δ 7.29-7.25 (m, 2H), 7.25-2.16 (m, 3H), 4.72-4.67 (m, 1H), 4.43-4.32 (m, 2H), 4.24-4.17 (m, 1H), 4.08-4.00 (m, 1H), 3.90-3.79 (m, 2H), 3.63-3.54 (m, 1H), 3.46-3.38 (m, 1H), 3.15-3.12 (m, 1H), 3.12-3.09 (m, 1H), 2.89-2.81 (m, 1H), 2.80-2.76 (m, 1H), 2.75 (s, 3H), 2.20-2.13 (m, 2H), 1.85-1.73 (m, 2H), 1.73-1.62 (m, 2H), 1.62-1.52 (m, 2H), 1.36-1.23 (m, 8H), 0.92-0.85 (m, 3H); ^{19}F NMR (376 MHz, MeOD) −105.7 (d, J=250 Hz), −106.4 (d, J = 250 Hz) −113.8 (d, J = 259 Hz), −115.3 (d, J = 259 Hz), −130.0 (d, J = 239 Hz), −131.0 (d, J = 239 Hz); HRMS (ESI) calculated for C$_{33}$H$_{46}$F$_6$N$_5$O$_5$S [M+H]$^+$: 738.3118, found 738.3149.

N-(((3S,6S,9S)-3-benzyl-13,13,14,14,15,15-hexafluoro-9-(2-(methylamino)acetamido)-5,8-dioxo-2,3,4,5,6,7,8,9,10,11,14,15-dodecahydro-13H-cyclopenta[b][1]oxa4thia[9,12]diazacyclotetradecin-6-yl)methyl)octanamide (19). 19 was synthesized using solution phase peptide synthesis analogous to **18** with the appropriately modified peptidic sequence (L-phenylalaninol and N-Boc-L-homocystine are commercially available). ^{1}H NMR (500 MHz, MeOD) δ 7.34-7.28 (m, 2H), 7.28-7.20 (m, 3H), 5.37-5.28 (m, 1H), 4.52-4.47 (m, 1H), 4.47-4.42 (m, 1H), 4.30-4.19 (m, 2H), 3.84-3.77 (m, 1H), 3.77-3.70 (m, 1H), 3.56-3.46 (m, 1H), 3.43-3.37 (m, 1H), 3.14-3.01 (m, 2H), 3.00-2.89 (m, 2H), 2.70 (s, 3H), 2.22-2.16 (m, 2H), 2.08-1.99 (m, 2H), 1.64-1.55 (m, 2H), 1.33-1.26 (m, 8H), 0.92-0.87 (m, 3H); ^{19}F

NMR (376 MHz, MeOD) δ −106.6 (d, J = 250 Hz), −108.1 (d, J = 250 Hz), −108.7 (d, J = 262 Hz), −110.3 (d, J = 262 Hz); HRMS (ESI) calculated for $C_{34}H_{46}F_{43}N_5O_5S$ [M+H]$^+$: 724.2961, found 724.2985.

***N*-(((5*S*,8*S*,11*S*,*Z*)-5-benzyl-2^3,2^3,2^4,2^4,2^5,2^5-hexafluoro-11-(2-(methylamino)acetamido)-7,10-dioxo-1^1*H*-3-oxa-6,9-diaza-1(1,4)-imidazola-2(1,2)-cyclopentanacyclododecaphan-2^1-en-8-yl)methyl)octanamide (20).** **20** was synthesized using solution phase peptide synthesis analogous to **18** with the appropriately modified peptidic sequence (L-phenylalaninol and Boc-L-His are commercially available). ^{1}H NMR (500 MHz, MeOD) δ 8.00 (s, 1H), 7.44 (s, 1H), 7.37-7.29 (m, 2H), 7.29-7.20 (m, 3H), 4.78 (dd, J = 11.7, 4.9 Hz, 1H), 4.58-4.48 (m, 1H), 4.42-4.30 (m, 2H), 4.28-4.16 (m,1H), 3.83 (q, J = 15.9 Hz, 2H), 3.44 (d, J = 6.6 Hz, 2H), 3.21 (dd, J = 14.8, 8.4 Hz, 1H), 3.13-3.04 (m, 2H), 2.96 (dd, J = 13.7, 7.8 Hz, 1H), 2.76 (s, 3H), 2.17 (t, J = 7.6 Hz, 2H), 1.63-1.52 (m, 2H), 1.35-1.22 (m, 8 H), 0.88 (t, J = 13.7 Hz, 3H). ^{19}F NMR (376 MHz, MeOD) −109.1 (d, J = 249 Hz), −110.4 (d, J = 263 Hz), −111.0 (d, J = 253 Hz), −112.7 (d, J = 259 Hz), −129.2 (d, J = 240 Hz), −130.2 (d, J = 240 Hz).
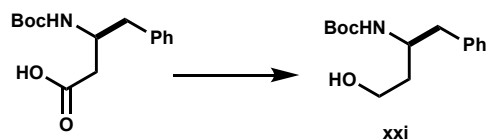
Scheme S4. Synthesis of starting material towards **22**.

***tert*-butyl (*S*)-(4-hydroxy-1-phenylbutan-2-yl)carbamate (xxi,** Scheme S4). At 0 °C, a solution of Boc-homo-β-phenylalanine (100 mg, 0.36 mmol), in anhydrous THF (0.75 mL) was treated with ethylchloroformate (0.04 mL, 0.394) and Et₃N (0.05 mL, 0.394 mmol). After 45 min, the cloudy reaction mixture was filtered over celite and rinsed with THF. In another flask, I₂ (45 mg, 0.18 mmol) was added to a suspension of NaBH₄ (16 mg, 0.43 mmol) in anhydrous THF (0.9 mL), and the resulting mixture was cooled to 0 °C. After 10 minutes, the filtrate containing the mixed anhydride was transferred into the NaBH₄ solution. The mixture was allowed to warm to room temperature and stirred for 12 hours. Water was added and the mixture was extracted with EtOAc. The combined organics were dried over MgSO₄, filtered and concentrated. The crude oil was purified over silica with 3:1 Hex:EtOAc to yield the product (70 mg, 67%).^{1}H NMR (400 MHz, MeOD) δ 7.26-7.20 (m, 2H), 7.20-7.10 (m, 3H), 3.88-3.75 (m, 1H), 3.60-3.50 (m, 2H), 2.72 (d, J = 7.0 Hz, 2H), 1.76-1.66 (m, 1H), 1.59-1.44 (m, 1H), 1.34 (s, 9H).

***N*-(((5*S*,8*R*,11*R*)-5-benzyl-15,15,16,16,17,17-hexafluoro-11-(2-(methylamino)acetamido)-7,10-dioxo-2,3,4,5,6,7,8,9,10,11,12,13,16,17-tetradecahydro-15*H*-cyclopenta[*b*][1]oxa[4]thia[9,12]diazacyclohexadecin-8-yl)methyl)octanamide (21).** **21** was synthesized using solution phase peptide synthesis analogous to **18** with the appropriately modified peptidic sequence (**xv** was synthesized as described above, and *N*-Boc-L-homocystine is commercially available). MS (ESI) *m/z* calculated for $C_{34}H_{47}F_6N_5O_5S$ [M+H]$^+$ 752.3, found 752.3. **21** was submitted for testing without further characterization.

***N*-(((4*S*,7*S*,10*S*)-4-benzyl-14,14,15,15,16,16-hexafluoro-10-(2-(methylamino)acetamido)-6,9-dioxo-3,4,5,6,7,8,9,10,11,12,15,16-dodecahydro-2*H*,14*H*-cyclopenta[*b*][1]oxa[4]thia[9,12]diazacyclopentadecin-7-yl)methyl)octanamide (22).** **22** was synthesized using solution phase peptide synthesis analogous to **18** with the appropriately modified peptidic sequence (**xxi** was synthesized as described above, and *N*-Boc-L-homocystine

is commercially available). MS (ESI) *m/z* calculated for $C_{33}H_{45}F_6N_4O_5S$ [M+H]$^+$ 738.3, found 738.3. **22** was submitted for testing without further characterization.

***tert*-butyl 4-((7-amino-5-methyl-[1,2,5]oxadiazolo[3,4-b]pyridin-6-yl)ethynyl)piperi-dine-1-carboxylate (26)**: **23**, **24** and **25** were prepared using methods described in the literature.[5,6] A solution of **24** (1.35 g, 4.89 mmol), **25** (1.45 g, 7.15 mmol), $(PPh_3)_2PdCl_2$ (687 mg, 0.98 mmol), and CuI (93 mg, 0.49 mmol) in Et_3N (20 mL) was degassed with argon for 15 min and then heated at 80 °C for 24 h. The mixture was then stirred at room temperature for 2 days, diluted with EtOAc and filtered through a pad of celite. The filter cake was rinsed with EtOAc. The organics were washed with brine (×2), dried over $MgSO_4$, filtered, and concentrated under a reduced pressure. The resultant residue was purified via column chromatography using Hex:EtOAc as eluents to afford **26** (889 mg, 51%). ^{1}H NMR (500 MHz, DMSO) δ 3.63 (m, 2H), 3.15 (s, 2H), 2.94 (m, 1H), 2.53 (s, 3H), 1.83 (m, 2H), 1.61 (m, 2H), 1.38 (s, 9H). HPLC/MS MH$^+$ 358.3.

***tert*-butyl (Z)-4-(2-(7-amino-5-methyl-[1,2,5]oxadiazolo[3,4-b]pyridin-6-yl)vinyl)piperidine-1-carboxylate (28)**: To a solution of **26** (250 mg, 0.70 mmol) in EtOH (7 mL) under argon $Pd(OH)_2$ was added (20% on carbon, 150 mg, 0.21 mmol). Argon was removed and reaction was conducted under an H_2 atmosphere and stirred at room temperature. Upon completion, as determined by LCMS, the reaction mixture was diluted with EtOH and filtered through a pad of celite to remove the catalyst. The filter cake was rinsed with EtOH and the filtrate was concentrated under reduced pressure to afford **27** and **28** (276 mg crude) which were identified by HPLC/MS analysis. The crude mixture was used in the next step without further purification.

***(Z)-1-(4-(2-(7-amino-5-methyl-[1,2,5]oxadiazolo[3,4-b]pyridin-6-yl)vinyl)piperidin-1-yl)-2-butoxyethan-1-one (30)** and **1-(4-(2-(7-amino-5-methyl-[1,2,5]oxadiazolo[3,4-b]pyridin-6-yl)ethyl)piperidin-1-yl)-2-butoxyethan-1-one (10)**: To the crude mixture of **27** and **28** (252 mg, 0.70 mmol) in CH_2Cl_2 (16 mL) TFA was added (10% *v/v*, 1.61 mL, 21.0 mmol). The resulting mixture was stirred at room temperature until completion as determined by LCMS. Upon completion, the reaction mixture was concentrated under a reduced pressure to afford the TFA salts, which were immediately suspended with (0.70 mmol), n-butoxyacetic acid (0.10 mL, 0.77 mmol), and HATU (293 mg, 0.77 mmol) in DMF (3 mL). DIPEA (0.37 mL, 2.10 mmol) was added to the suspension, and the resulting mixture was stirred at room temperature until completion, as determined by LCMS. Upon completion, the reaction was diluted with EtOAc and washed with saturated, aqueous NH_4Cl (×2), water (×2); saturated, aqueous $NaHCO_3$ (×2); and brine (×2). The organics were dried over $MgSO_4$, filtered, and concentrated under reduced pressure. The residue was purified by column chromatography using Hex:EtOAc as eluents to afford **35** (300 mg) and **10** (10 mg) as yellow residues. **30**: ^{1}H NMR (500 MHz, MeOD) δ 7.89 (s, 2H), 6.15 (d, *J* = 10.9 Hz, 1H), 5.87 (dd, *J* = 9.7, 10.7 Hz, 1H), 4.39 (d, *J* = 13.0 Hz, 1H), 4.12 (q, *J* = 18.1 Hz, 2H), 3.88 (s, 1H), 3.85 (d, *J* = 12.8 Hz, 1H), 3.51 (t, *J* = 6.6 Hz, 1H), 3.46 (q, *J* = 6.7 Hz, 2H), 2.98 (s, 1H), 2.94 (d, *J* = 12.0 Hz, 1H), 2.85 (d, *J* = 0.6 Hz, 1H), 2.80 (s, 1H), 2.58 (t, *J* = 11.7 Hz, 1H), 2.44 (s, 3H), 2.26 (m, 1H), 1.57 (m, 4H), 1.38 (m, 5H), 0.92 (t, *J* = 7.4 Hz, 3H). ^{13}C NMR (126 MHz, MeOD) δ 169.65, 168.58, 158.30, 141.64, 140.78, 139.02, 120.46, 109.81, 78.08, 71.12, 70.88, 69.36, 44.37, 41.23, 37.48, 36.17, 31.33, 31.25, 23.91, 18.89, 12.79. HRMS *m/z*: [M + H]$^+$ Calculated for $C_{19}H_{28}N_5O_3$ 374.2192; Found 374.2159. **10**: ^{1}H NMR (500 MHz, CDCl$_3$) δ 5.13 (br s, 2H), 4.62 (d, *J* = 13.2 Hz, 1H), 4.13 (d, *J* = 12.3 Hz, 2H), 4.00 (d, *J* = 13.6 Hz, 1H), 3.49 (q, *J* = 7.1 Hz, 2H), 3.03 (t, *J* = 12.0 Hz, 1H), 2.62 (s, 3H), 2.59 (d, *J* = 8.5 Hz, 1H), 1.87 (t, *J* = 12.8 Hz, 2H), 1.66 (m, 1H), 1.58 (m, 2H), 1.49 (m, 2H), 1.37 (m, 2H), 1.24 (s, 2H), 1.21 (s, 1H), 1.20 (s, 1H), 0.91 (t, *J* = 7.4 Hz, 3H). ^{13}C NMR (126 MHz, CDCl$_3$) δ 170.01, 167.86, 157.97, 139.23, 138.21, 114.14, 71.29, 70.72, 45.33, 42.13, 36.51, 34.10, 32.81, 31.67, 29.71, 25.37, 25.00, 24.02, 19.28, 13.88. HRMS *m/z*: [M + H]$^+$ Calculated for $C_{19}H_{30}N_5O_3$ 376.2349; Found 376.2315.

***tert*-butyl 4-(4-oxopentyl)piperidine-1-carboxylate** (**32**): To a flask containing CuCN (1.65 g, 18.42 mmol), dry Et$_2$O (30 mL) was added. The resulting suspension was cooled to 0 °C and MeLi (1.6 M in Et$_2$O, 23 mL, 36.85 mmol) was added dropwise. After addition, the resulting mixture was stirred at 0 °C for 5 min before the dropwise addition of a solution of **31** (1.00 g, 3.68 mmol) in Et$_2$O (18.4 mL). The resulting reaction mixture was warmed to room temperature over an hour and then stirred at room temperature overnight. The reaction was quenched with saturated aqueous NH$_4$Cl, and then extracted with CH$_2$Cl$_2$ (×2). The organics were dried over MgSO$_4$, filtered, and concentrated under reduced pressure to afford **32** (703 mg, 71%) as a colorless oil. ^{1}H NMR (500 MHz, CDCl$_3$) δ 4.05 (d, *J* = 13.2 Hz, 2H), 2.65 (dt, *J* = 2.6, 12.8 Hz, 2H), 2.40 (t, *J* = 7.4 Hz, 2H), 2.12 (s, 3H), 1.63 (m, 2H), 1.58 (m, 2H), 1.44 (s, 9H), 1.36 (m, 1H), 1.21 (m, 2H), 1.05 (m, 2H). ^{13}C NMR (126 MHz, CDCl$_3$) δ 208.99, 154.90, 79.20, 43.99, 43.82, 35.98, 35.91, 32.07, 29.93, 28.48, 20.88. HPLC/MS MH$^+$ 170.2 (-Boc).

***tert*-butyl (*S*)-(1-oxo-1-(4-(4-oxopentyl)piperidin-1-yl)propan-2-yl)carbamate** (**33**): To a solution of **32** (200 mg, 0.74 mmol) in CH$_2$Cl$_2$ (10 mL) at 0 °C TFA was added (10% *v/v*, 1.14 mL, 14.85 mmol). The resulting solution was stirred at 0 °C until completion as determined by TLC. The reaction mixture was concentrated under reduced pressure and the resultant residue was immediately used. To a solution of the TFA salt (188 mg, 0.74 mmol) and Boc-Ala-OSu (203 mg, 0.71 mmol) in MeCN (3 mL) DIPEA was added (0.39 mL, 2.22 mmol) at 0 °C. The resulting mixture was stirred at 0 °C for 2 h, diluted with EtOAc and washed with saturated aqueous NH$_4$Cl and H$_2$O (×3). The organics were dried over MgSO$_4$, filtered, and concentrated under reduced pressure. Residue was purified by column chromatography using 1:1 Hex:EtOAc as eluents to afford **33** (122 mg, 50%) as a colorless oil. ^{1}H NMR (500 MHz, CDCl$_3$) δ 5.59 (s, 1H), 4.56 (m, 2H), 3.83 (t, *J* = 12.4 Hz, 1H), 3.00 (m, 1H), 2.57 (m, 1H), 2.42 (t, *J* = 6.7 Hz, 2H), 2.13 (s, 3H), 1.77 (m, 2H), 1.59 (t, *J* = 3.8 Hz, 2H), 1.49 (m, 2H), 1.43 (d, *J* = 4.0 Hz, 9H), 1.28 (dd, *J* = 6.9, 8.5 Hz, 1H), 1.10 (m, 3H).

(*S*)-1-methyl-*N*-(1-oxo-1-(4-(4-oxopentyl)piperidin-1-yl)propan-2-yl)-1*H*-pyrazole-5-carboxamide (**35**): To a solution of **33** (122 mg, 0.36 mmol) in CH$_2$Cl$_2$ (5 mL) at 0 °C, TFA was added (10% *v/v*, 0.55 mL, 7.16 mmol). The resulting mixture was stirred at 0 °C until completion as determined by TLC. The reaction mixture was concentrated under a reduced pressure and the resultant residue was used immediately without purification. The deprotected material (0.36 mmol) and **34** (80 mg, 0.36 mmol) were suspended in MeCN (1.5 mL) at 0 °C. DIPEA was added (0.19 mL, 1.08 mmol) to the suspension, and the resulting reaction mixture was stirred at 0 °C for 2 h and diluted with EtOAc. Organics were washed with saturated, aqueous NH$_4$Cl and H$_2$O (×3), dried over MgSO$_4$, filtered, and concentrated under reduced pressure. The residue was purified by column chromatography using 1:1 Hex:EtOAc as eluents to afford **35** (90 mg, 72%). (Rotational isomers present.) ^{1}H NMR (500 MHz, CDCl$_3$) δ 7.44 (d, *J* = 1.5 Hz, 1H), 7.28 (s, 1H), 6.60 (q, *J* = 2.6 Hz, 1H), 4.97 (q, *J* = 6.6 Hz, 1H), 4.56 (t, *J* = 12.9 Hz, 1H), 4.18 (s, 3H), 3.87 (m, 1H), 3.06 (m, 1H), 2.62 (m, 1H), 2.43 (q, *J* = 7.1 Hz, 2H), 2.13 (d, *J* = 3.7 Hz, 3H), 1.81 (m, 2H), 1.58 (m, 2H), 1.52 (m, 1H), 1.40 (q, *J* = 5.5 Hz, 3H), 1.24 (t, *J* = 18.2 Hz, 2H), 1.11 (m, 2H). ^{13}C NMR (126 MHz, MeOD) δ 208.73, 170.13, 169.94, 158.81, 137.51, 106.76, 60.41, 45.90, 45.52, 45.41, 43.64 (d, *J* = 5.3 Hz, 1C), 42.85, 42.51, 39.26, 36.06, 35.89, 35.69, 35.62, 32.66, 32.47, 31.71 (d, *J* = 6.5 Hz, 1C), 29.98, 20.74, 19.42, 18.91, 14.21.

(*S*)-*N*-(1-(4-(2-(7-amino-5-methyl-[1,2,5]oxadiazolo[3,4-b]pyridin-6-yl)ethyl)pip-eridin-1-yl)-1-oxopropan-2-yl)-1-methyl-1*H*-pyrazole-5-carboxamide (**11**) and **(*S*)-*N*-(1-(4-(3-(7-amino-[1,2,5]oxadiazolo[3,4-*b*]pyridin-5-yl)propyl)pip-eridin-1-yl)-1-oxopropan-2-yl)-1-methyl-1H-pyrazole-5-carboxamide** (**38**): To a suspension of **35** (90 mg, 0.26 mmol) and **23** (28 mg, 0.26

mmol) in toluene (0.2M) tin(IV) chloride was added (2 eq). The resulting mixture was stirred at room temperature for 30 min, and then refluxed for 18 h. The reaction was cooled to room temperature and concentrated under reduced pressure. The resultant residue was dissolved in CH_2Cl_2 and washed with saturated aqueous $NaHCO_3$ (×1). The aqueous layer was extracted with CH_2Cl_2 (×3). The organics were combined, washed with brine (×1), dried over $MgSO_4$, filtered, and concentrated under reduced pressure. The residue was purified by column chromatography using 5% MeOH in CH_2Cl_2 as eluents to afford **11** and **38**. HRMS *m/z*: [M + H]+ Calculated for $C_{21}H_{29}N_5O_6$ 441.2363; Found 441.2313. **11**: ^{1}H NMR (500 MHz, MeOD) δ 7.43 (d, *J* = 2.0 Hz, 1H), 6.83 (s, 1H), 4.97 (m, 1H), 4.47 (m, 1H), 4.06 (d, *J* = 5.6 Hz, 1H), 4.01 (d, *J* = 10.5 Hz, 1H), 3.12 (m, 1H), 2.65 (m, 1H), 2.46 (q, *J* = 7.0 Hz, 2H), 2.10 (d, *J* = 3.6 Hz, 3H), 1.78 (m, 2H), 1.55 (m, 3H), 1.36 (dd, *J* = 7.1, 13.9 Hz, 3H), 1.24 (m, 3H), 1.07 (m, 2H). ^{13}C NMR (126 MHz, MeOD) δ 210.31, 170.8, 164.8, 137.12, 107.16, 45.69, 45.36, 45.22, 42.74, 42.55, 42.19, 37.73, 35.64, 35.43, 35.35, 35.23, 32.35, 32.21, 31.53, 31.43, 28.29, 20.32, 16.46, 15.98. **38**: ^{1}H NMR (500 MHz, MeOD) δ 7.42 (q, *J* = 2.4 Hz, 1H), 6.83 (m, 1H), 6.29 (d, *J* = 4.5 Hz, 1H), 4.97 (m, 1H), 4.46 (m, 1H), 4.05 (s, 3H), 4.02 (s, 1H), 3.13 (m, 1H), 2.67 (m, 3H), 2.53 (d, *J* = 9.5 Hz, 1H), 1.88 (q, *J* = 12.7 Hz, 1H), 1.75 (m, 3H), 1.59 (s, 1H), 1.46 (q, *J* = 8.0 Hz, 1H), 1.35 (dd, *J* = 7.1, 13.0 Hz, 3H), 1.25 (s, 2H), 1.08 (m, 2H). ^{13}C NMR (126 MHz, MeOD) δ 174.48, 159.60, 137.10, 107.15, 100.26, 45.23, 42.56, 38.87, 37.72, 35.53, 35.38, 32.20, 31.53, 31.43, 25.78, 23.05, 16.44, 15.95.

References

(1)   Yang, J., Zhao, T.-J., Goldstein, J. L. & Brown, M. S. Inhibition of ghrelin O-acyltransferase (GOAT) by octanoylated pentapeptides. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10750–10755 (2008).
(2)   Yang, J., Brown, M. S., Liang, G., Grishin, N. V & Goldstein, J. L. Identification of the acyltransferase that octanoylates ghrelin, an appetite-stimulating peptide hormone. *Cell* **132**, 387–396 (2008).
(3)   Greulich, H. *et al.* Functional analysis of receptor tyrosine kinase mutations in lung cancer identifies oncogenic extracellular domain mutations of ERBB2. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 14476–81 (2012).
(4)   Lou, E. *et al.* Tunneling nanotubes provide a unique conduit for intercellular transfer of cellular contents in human malignant pleural mesothelioma. *PLoS One* **7**, e33093 (2012).
(5)   Godbout, C.; Trieselmann, T.; Vintonyak, V. Oxadiazolopyridine Derivatives for Use as Ghrelin O-Acyl Transferase (GOAT) Inhibitors. US 2018/0037594 A1, February 2018.
(6)   Pagoria, P. F.; Zhang, M.-X.; Zuckerman, N. B.; DeHope, A. J.; Parrish, D. A. Synthesis and Characterization of Multicyclic Oxadiazoles and 1-Hydroxytetrazoles as Energetic Materials. *Chem. Heterocycl. Compd.* **2017**, *53*, 760–778.